D-SPIN constructs gene regulatory network models from multiplexed scRNA-seq data revealing organizing principles of cellular perturbation response

Jialong Jiang[1,*], Sisi Chen[1,2*], Tiffany Tsou[1,2], Christopher S. McGinnis[3], Tahmineh Khazaei[1], Qin Zhu[3], Jong H. Park[1,2], Inna-Marie Strazhnik[1], John Hanna[4], Eric D. Chow[5,6], David A. Sivak[7], Zev J. Gartner[3,8,9,10], and Matt Thomson[1,2†]

[1]Division of Biology and Biological Engineering, California Institute of Technology, Pasadena, California, 91125, USA.

[2]Beckman Single-Cell Profiling and Engineering Center, California Institute of Technology, Pasadena, CA, 91125, USA.

[3]Department of Pharmaceutical Chemistry, University of California San Francisco, San Francisco, CA, 94143, USA

[4]Department of Pathology, Harvard Medical School and Brigham and Women's Hospital, Boston, MA, 02115,USA

[5]Department of Biochemistry and Biophysics, University of California San Francisco, San Francisco, CA, 94143, USA

[6]Center for Advanced Technology, University of California San Francisco, San Francisco, CA, 94143, USA

[7]Department of Physics, Simon Fraser University, Burnaby, BC V5A 1S6, Canada

[8]Helen Diller Family Comprehensive Cancer Center, San Francisco, CA, 94115, USA

[9]Chan Zuckerberg BioHub, University of California San Francisco, San Francisco, CA, 94143, USA

[10]Center for Cellular Construction, University of California San Francisco, San Francisco, CA, 94143, USA

[*]These authors contributed equally to this work

[†]mthomson@caltech.edu

Abstract

Gene regulatory networks within cells modulate the expression of the genome in response to signals and changing environmental conditions. Reconstructions of gene regulatory networks can reveal the information processing and control principles used by cells to maintain homeostasis and execute cell-state transitions. Here, we introduce a computational framework, D-SPIN, that generates quantitative models of gene-regulatory networks from single-cell mRNA-seq data sets collected across thousands of distinct perturbation conditions. D-SPIN models the cell as a collection of interacting gene-expression programs, and constructs a probabilistic model to infer regulatory interactions between gene-expression programs and external perturbations. Using large perturb-seq and drug-response data sets, we demonstrate that D-SPIN models reveal the organization of cellular pathways, sub-functions of macromolecular complexes, and the logic of cellular regulation of transcription, translation, metabolism, and protein degradation in response to gene knock-down perturbations. D-SPIN can also be applied to dissect drug response mechanisms in heterogeneous cell-populations, elucidating how combinations of immunomodulatory drugs can induce novel cell states through additive recruitment of gene expression programs. D-SPIN provides a computational framework for constructing interpretable models of gene-regulatory networks to reveal principles of cellular information processing and physiological control.

Introduction

The human genome encodes more than $30,000$ genes, but a typical human cell expresses only 5,000 genes at any given time [1]. Gene regulatory networks within cells modulate gene expression based upon environmental cues and the cell's internal state [2, 3, 4, 5]. In gene regulatory networks, transcription factors and associated regulatory proteins interact with one another and gene promoters to activate or repress gene transcription [6, 7, 8]. Gene regulatory networks play a central role in cellular decision-making processes [9, 10, 11, 12]. However, we have limited understanding of the global logic of networks and cell types and underlying principles of cellular information processing.

Global reconstruction of gene regulatory networks in E. coli, yeast, and sea urchin embryos have revealed features of biological information processing, including network modularity, recurring network motifs, and combinatorial logic through the assembly of transcription factor complexes at gene promoters [13, 14, 15, 16, 17, 18, 19]. However, our understanding of the information processing principles of gene regulatory networks is based on an limited number of network reconstructions performed on model organisms over a limited range of physiological conditions. Even in E. coli a large fraction of genes have unknown regulators [20]. In metazoans, gene network models have primarily focused on sub-circuits involved in specific processes like T-cell activation, T-cell fate selection, and embryonic stem cell differentiation [10, 11, 12]. The comparative analysis of gene regulatory network architectures across cell-types, developmental stages of an organism, and species, therefore, remains nascent, and there are very few quantitative models of regulatory networks that can predict the response of a cell to signals, genetic perturbations or therapeutics. Cells simultaneously regulate tens of different processes including transcription, metabolism, protein degradation, differentiation, and how gene regulatory networks globally modulate these processes in parallel in response to environmental conditions remains poorly understood.

Historically, gene regulatory network reconstruction and modeling has been constrained by the number of biochemical or genetic measurements required to reconstruct networks with hundreds to thousands of interacting protein components. Classical biochemical approaches perform bottom-up gene network reconstructions through pairwise binding measurements requiring order-M^2 biochem-

ical assays for a network with M components [21, 22]. Genetics approaches use a top-down strategy that infers gene regulatory network architecture from perturbation-response experiments where genes are knocked down or protein activity inhibited with small molecules. Perturbation of one network component alters gene expression for subsets of genes; through top-down association of perturbation with response, perturbation approaches enable the identification of gene regulators, the organization of regulators into pathways, and the construction of integrated network models [23, 24, 25, 26]. For global analysis of gene regulatory networks, perturbation approaches require knocking out hundreds to thousands of genes while monitoring transcription across thousands of genes. Historically, experimental realization of genome-wide network reconstruction through perturbation has, like biochemistry, been limited by experimental scale when genes were knocked out in bulk assays.

Developments in single-cell genomics and perturbation bar-coding circumvent some of the conventional limitations of perturbation-driven network reconstruction. Perturbation bar-coding approaches, including perturb-seq, click-tags, and multi-seq, allow the transcriptional state of each cell in a cell population to be measured across thousands of different genetic, signaling, and small molecule conditions [27, 28, 29, 30]. These experimental approaches identify the perturbation delivered to each single cell in a population while also measuring changes in the entire transcriptome through single-cell mRNA-seq. Given single-cell perturbation measurements allow interrogation of a cell-population's response to thousands of genetic, signaling, or small molecule perturbations, a core challenge is developing computational methods to integrate data from thousands of such experiments into a gene regulatory network model that can classify perturbations, map the flow of information across a regulatory network, and predict cellular response to novel perturbations.

Here, we introduce a mathematical modeling and network inference framework, D-SPIN (Dimension-reduced Single-cell Perturbation Integration Network), that constructs gene regulatory network models directly from single-cell perturbation-response data. D-SPIN models the cell as a collection of interacting gene-expression programs, and constructs a probabilistic model, known as Markov random fields, to infer regulatory interactions between gene-expression programs and applied perturbations. To enable scaling of D-SPIN to large data sets with millions of single cells, D-SPIN exploits a natural factoring within the mathematical structure of Markov random fields inference to separate the learning problem into two steps, construction of a unified gene regulatory network model and inference of how each perturbation interacts with the gene programs within the unified network. To maximize interpretability of D-SPIN models and computational efficiency, D-SPIN operates on gene-expression programs, rather than single genes, enabling D-SPIN to derive a coarse-grained description of a cellular regulatory network in terms of physiological units—gene-expression programs—and their interactions with applied perturbations. In benchmarking tests, we demonstrate that D-SPIN exhibits state-of-the-art performance on simulated gene network inference tasks.

We apply D-SPIN to construct gene regulatory network models from experimental data sets with thousands of perturbations and millions of cells, including two of the largest single-cell mRNA-seq data sets in existence, a genome-wide Perturb-seq experiment [28] and a new human immune cell drug-response experiment we performed. The integrated regulatory network models reveal global organizing principles of cellular regulation and perturbation responses, including the organizing principles of cellular pathways and networks of pathways used to maintain homeostasis in response to perturbation. Applied to profile the response of human immune cells to combinations of immunomodularity drugs, D-SPIN models remonstrate that drug combinations can generate novel cell states through additive recruitment or superposition of drug-specific gene expression programs. Broadly, D-SPIN provides a computational framework for large-scale modeling and comparative analysis of gene regulatory networks across cell-types, physiological conditions and organisms.

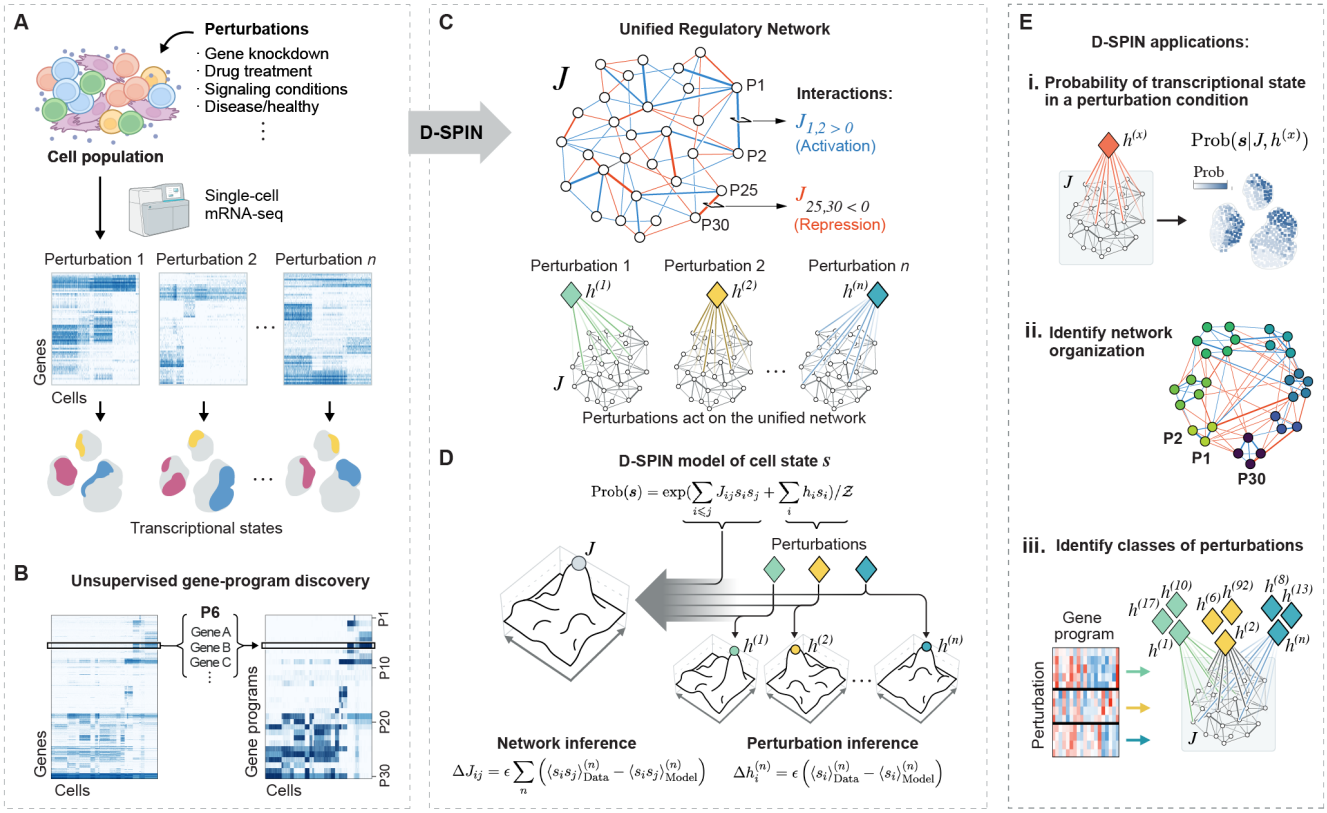


Figure 1: **D-SPIN constructs unified gene regulatory network models from single-cell mRNA-seq data collected across perturbation conditions.** (A) D-SPIN accepts as input single-cell mRNA-seq data from a cell population profiled across a series of different perturbed experimental conditions such as genetic perturbations, drug treatments, signaling conditions, or disease/healthy conditions. (B) To reduce the dimensionality of the network-inference task, D-SPIN extracts gene-expression programs from the data using orthogonal non-negative matrix factorization. (C) D-SPIN then constructs a unified regulatory network $\boldsymbol{J}$, whose edges represent inferred interactions between gene-expression programs, and also infers interactions between the network and applied perturbations. (D) Mathematically, D-SPIN constructs the regulatory network model by inferring a probabilistic model that encodes the probability of each transcriptional state in terms of an interaction matrix $\boldsymbol{J}$ and sample-specific perturbation response vectors $\boldsymbol{h}^{(n)}$. D-SPIN scales to large data sets with millions of cells through an inference maximum-likelihood procedure with paralleled Monte Carlo sampling. At each step of inference, D-SPIN generates model samples for each perturbation condition in parallel, and the use the difference between model and data cross-correlation and mean to update the network weights $\boldsymbol{J}$ and each perturbation response vector $\boldsymbol{h}^{(n)}$. (E) D-SPIN yields a probabilistic model that can be interpreted as a gene regulatory network. The D-SPIN model can be applied to (i) estimate distribution of transcriptional state defined by the regulatory network model under a specific condition, (ii) reveal principles of gene regulatory network organization including modularity and sub-network architecture, and (iii) classify perturbations into groups that reflect biological function.

D-SPIN infers a unified regulatory network model from single-cell perturbation data

Here, we develop mathematical and computational strategies to integrate data across many different single-cell perturbation experiments into a network model that provides qualitative insights into cellular pathways and regulatory strategies. D-SPIN solves the gene regulatory network inference problem by constructing a probabilistic model that encodes the distribution of transcriptional states within a cell population across a set of perturbed conditions (Fig. 1). Specifically, D-SPIN builds a spin-network model or Markov random field, which can be interpreted as an interaction network. Spin-network models were initially introduced in physics to study magnetic materials known as spin glasses whose properties emerge through positive and negative interactions between localized magnetic moments or spins. The models have been generalized and applied to many different systems of interacting elements including neural networks, bird flocks, and proteins [31, 32, 33]. In general, spin-network models consider a network of M interacting elements s_i. The goal of the framework is to construct a model that encodes the probability $P(s_1, \ldots, s_M)$ of each potential configuration of the network of elements based upon inferred interactions between elements, as quantified in a coupling matrix $\boldsymbol{J}$.

In general, D-SPIN can be applied to model networks of individual genes; however, cells express thousands of genes, and gene-level models are both computationally expensive and also difficult to interpret. Therefore, we designed D-SPIN to construct reduced or low-dimensional spin-network models by modeling interactions between gene programs, i.e., co-regulated groups of genes [34, 35, 36]. Cells regulate their transcriptional state by modulating transcription factors that impact sets of genes related to differentiation state, cellular pathways, and core physiological processes. By focusing on gene programs, rather than single genes, D-SPIN generates low-dimensional models of cellular regulation that are interpretable and that can be inferred from data through a computationally efficient inference procedure. Building upon previously published work, D-SPIN identifies co-regulated groups of genes, gene programs, through unsupervised orthogonal nonnegative matrix factorization (oNMF) and phenotypic gene-program annotation (Fig. 1A,B) [37, 38].

Following the extraction of gene programs, D-SPIN applies a Bayesian statistical inference procedure to infer a spin network of regulatory interactions between gene-expression programs and between programs and applied perturbations (Fig. 1C). Regulatory interactions between gene programs are represented in a gene-program interaction matrix $\boldsymbol{J}$ where each entry J_{ij} quantifies the inferred interaction between program i and j. Interactions between the regulatory network and perturbations are encoded in a set of perturbation vectors $\boldsymbol{h}^{(n)}$ where $h_i^{(n)}$ quantifies the interaction of the perturbation n with gene program i.

Mathematically, an energy function $E(\boldsymbol{s})$ computes the effective energy of a given transcriptional state $\boldsymbol{s}$ by balancing the influence of the regulatory network (encoded in $\boldsymbol{J}$) and the perturbation-specific inputs (encoded in each $\boldsymbol{h}^{(n)}$). The resulting probability of a given transcriptional state $\boldsymbol{s}$ is $P(\boldsymbol{s}) = \frac{1}{Z}\exp[-E(\boldsymbol{s}; \boldsymbol{J}, \boldsymbol{h})]$, where

$$E(\boldsymbol{s}; \boldsymbol{J}, \boldsymbol{h}^{(n)}) = -\sum_{ij} J_{ij} s_i s_j - \sum_i h_i^{(n)} s_i$$

and Z is a normalizing constant ensuring all probabilities sum to 1 (SI Text) [39, 40]. The regulatory network accounts for positive ($J_{ij} > 0$) and negative ($J_{ij} < 0$) interactions between gene programs. Perturbations exert a bias on each gene program to increase the likelihood of being on ($h_i > 0$) or off ($h_i < 0$).

Given single-cell gene-expression data collected from a series of perturbation conditions, D-SPIN performs an optimization procedure through gradient ascent to estimate model parameters

$\boldsymbol{J}$ and $\boldsymbol{h}^{(n)}$ given the data (Fig. 1D). For spin-network models, the optimization procedure yields closed-form parameter-update rules

$$\Delta J_{ij} = \epsilon \sum_n \left(\langle s_i s_j \rangle^{(n)}_{\text{Data}} - \langle s_i s_j \rangle^{(n)}_{\text{Model}} \right) \tag{1}$$

$$\Delta h_i^{(n)} = \epsilon \left(\langle s_i \rangle^{(n)}_{\text{Data}} - \langle s_i \rangle^{(n)}_{\text{Model}} \right) \tag{2}$$

where $\langle s_i s_j \rangle^{(n)}_{\text{Data}}$ are gene-program cross correlations computed from data for each of n perturbation conditions, $\langle s_i s_j \rangle^{(n)}_{\text{Model}}$ are gene-program cross correlations computed from the model, and ϵ is a learning rate. At each step of model inference, the procedure samples from the model for each perturbation condition and updates the regulatory network $\boldsymbol{J}$ and each perturbation vector $\boldsymbol{h}$. In general, spin-network model inference is a convex optimization problem, so that the optimization has a single unique solution.

The form of the spin-network model learning rules enabled us to develop a highly efficient and scalable training algorithm for D-SPIN by exploiting intrinsic parallelization. Training of general graphical models requires estimating the gradient function on every data point separately, which is typically computationally extensive. In contrast, in D-SPIN the gradient is pooled for each experimental condition, and only depends on the mean $\langle s_i \rangle$ and cross-correlation $\langle s_i s_j \rangle$ between gene programs in each condition. Consequently, the training of the network can be deployed in parallel, with each computational thread computing the gradient based on the mean and cross-correlation of a single experimental condition. This approach eliminates the need to estimate gradients on every data point separately (i.e., every cell's transcriptional state) and minimizes data communication by only requiring the mean and cross-correlation to be exchanged during parallelization. Therefore, we routinely use hundreds of CPU cores for the optimization, enabling efficient network inference over large data sets. Practically, we estimate the network using a subset of experimental conditions with highest diversity, and utilize the inferred network to identify response vectors of each condition independently (SI Text).

Following inference of $\boldsymbol{J}$ and $\boldsymbol{h}$, D-SPIN models can be analyzed as network models that reveal regulatory connections between cellular processes and identify interactions between applied perturbations and cellular gene-expression programs. (Fig. 1E, ii-iii). D-SPIN models can also be used to predict the distribution of transcriptional states in a cell population under normal and perturbed condition and explain how interactions within a gene network lead to a specific distribution of cell states in an experimental population. The models provide a computational strategy that can bridge geometric approaches like UMAP with logic-based mechanistic modeling approaches that view the cell as a regulatory network with links between cellular processes, pathways, and perturbations. D-SPIN can reveal how perturbations to specific nodes of a network alter the distribution of transcriptional states in a cell population and, thus, alter the rendering of a cell population in UMAP.

D-SPIN achieves state-of-the-art performance on gene regulatory network inference

To evaluate the accuracy of D-SPIN in performing network inference, we first applied D-SPIN to reconstruct and analyze a model of the hematopoietic stem cell (HSC) differentiation network using the network simulation and benchmarking framework BEELINE [41]. The HSC network has two major groups of regulatory genes, the first group includes Pu1 controlling granulocyte/monocyte differentiation and the second group includes Gata1 controlling megakaryocyte/erythrocyte differentiation[42]. BEELINE simulates single-cell transcription data from biologically identified regulatory networks

by generating and simulating ordinary differential equation (ODE) models with stochastic noise. BEELINE also includes existing network inference methods for benchmark accuracy evaluation. We used the BEELINE framework to generate synthetic gene-expression profiles from the HSC network for 22 perturbation conditions encompassing knockdown/activation of each network node individually by turning down/up the transcription rate parameter (Fig. 2B,C). For interpretation of the data we used the biological functions of each transcription factors to classify the single-cell expression profiles into 7 different cell state designations including monocytes (Mono), granulocytes (Gran), erythrocytes (Ery) and megakaryocytes (Mega) [42].

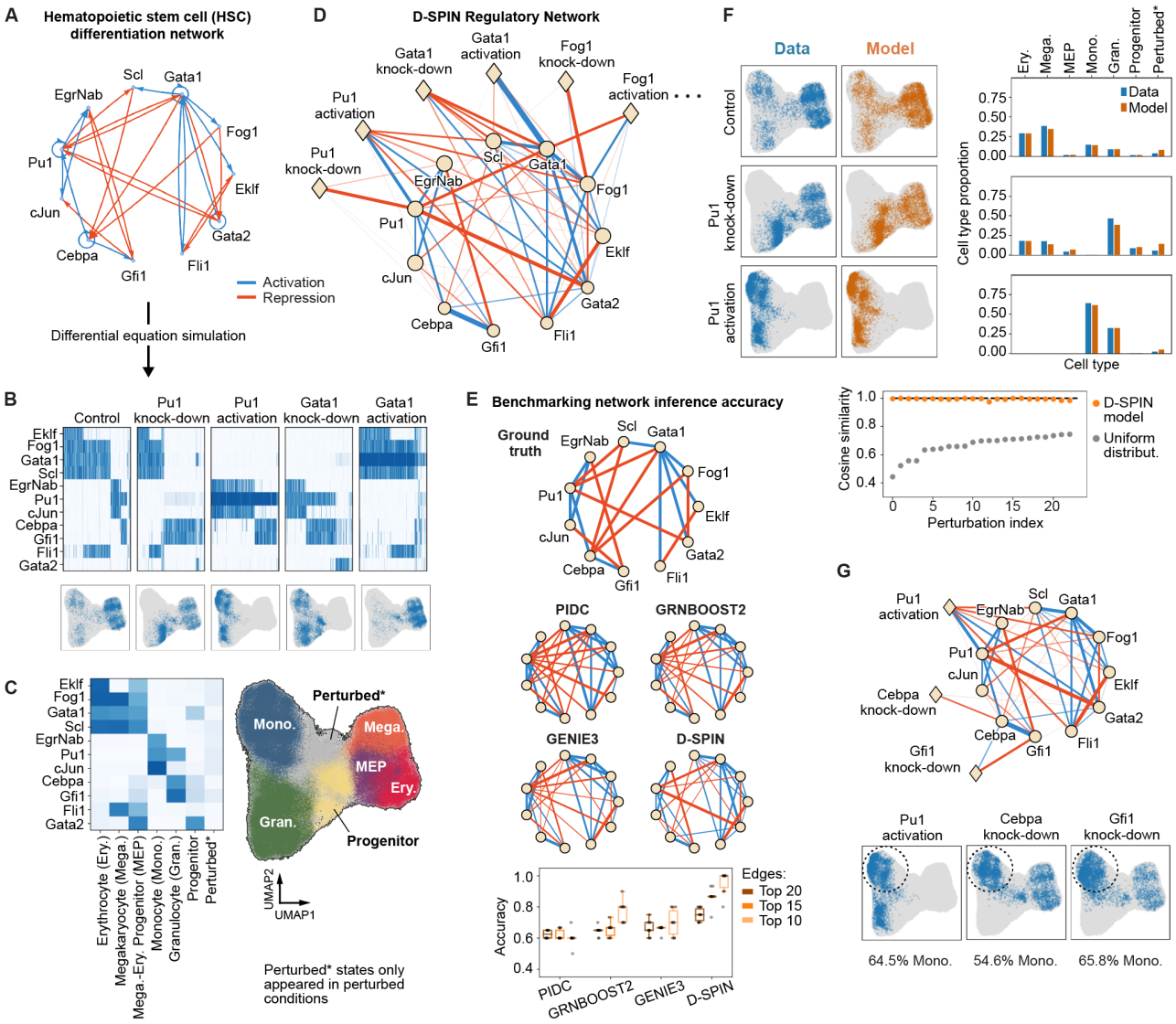


Figure 2: **D-SPIN achieves state-of-the-art performance on network inference benchmarking task, revealing cell-fate control strategies for hematopoietic stem cell differentiation network** (A) HSC regulatory network model [42] contains 11 transcription factors that interact through positive (blue) and negative interactions (red) to modulate the differentiation of HSCs into different cell types. (B) Example simulated gene-expression profile heatmaps generated using published computational model of the network for a series of simulated single-gene knockout and activation conditions. (C) The simulated network generates a distribution of cell states that can be classified by their transcription factor expression into 7 cell states. The heatmap shows transcription factor expression patterns that define each cell state generated across all simulated conditions, visualized as a UMAP. (D) Using the simulated data, D-SPIN infers an integrated regulatory network model that encodes inferred interactions between transcription factors as an interaction graph. D-SPIN also infers perturbation vectors $\boldsymbol{h}$ that estimate how knockdown or activation of transcription factors (e.g., Gata1 knockdown) impacts the regulatory network through up- or downregulation of transcription factors. (E) The D-SPIN model has greater accuracy than state-of-the-art methods [41], as shown through rendering of the network inferred from data by each method as well as a box plot of the number of correctly inferred edges across different batches of simulated gene-expression data. The box plot shows accuracy of top k edges for different values of k where the ground truth network has a total of 20 edges. (F) The D-SPIN model predicts the cell-fate distribution generated through perturbations to the underlying network. UMAP plots of cell state distribution generated for control and Pu1 activation/knockdown shown for simulated data and D-SPIN model. Bar graphs quantify the proportion of HSC-derived cell states in each condition from simulated data and the D-SPIN model. The data and model distribution are highly consistent across all 22 perturbations as quantified by the cosine similarity between cell state distributions of data and model. (G) The model can be used to understand how perturbations to distinct transcription factors can generate similar cell populations. Both Pu1 activation and knockdown of Cebpa or Gfi1 generate increased Mono state by directly activating Mono genes or indirectly through the interaction between Gfi1 and EgrNab.

We applied D-SPIN to reconstruct a gene regulatory network model from the simulated data and compared the accuracy of the D-SPIN model to 3 different network reconstruction methods (PIDC, GENIE3, and GRNBoost2) that were the top performers in a previous benchmarking study [41] (Fig. 2D,E). We applied D-SPIN to generate a single network model encoded in the gene interaction matrix $\boldsymbol{J}$ and generated perturbation vectors $\boldsymbol{h}^{(n)}$ for each perturbation condition. We applied the algorithms to individual genes, rather than gene programs, given the relatively compact size (11 nodes) of the network used in the BEELINE benchmarking evaluation. Each method generates a set of inferred edges and provides confidence estimates for each edge. We compared the methods by considering the top 10, top 15 and top 20 edges identified by each method.

On the simulated network reconstruction task, D-SPIN consistently achieved higher performance than comparable methods, primarily due to the natural way in which D-SPIN can accommodate perturbation data. For D-SPIN, on average 0.96 of the top ten edges found by the model were correct across inference runs, as compared to 0.6, 0.77, and 0.7 for PIDC, GRNBOOST2, and GENIE3, respectively (Fig. 2E). For D-SPIN, 0.75 of the top 20 edges were correct, compared with 0.625, 0.645, and 0.665 for PIDC, GRNBOOST2, and GENIE3, respectively. As an example where D-SPIN outperforms existing methods, the networks inferred by PIDC, GENIE3, and GRNBoost2 all identified a spurious activating interaction inside the Mega.-Ery. gene module, e.g., Scl-Fog1 or Scl-Eklf, while D-SPIN pruned these false positive edges and revealed the central regulatory role of Gata1. D-SPIN also pruned the spurious inhibiting interactions between the two modules, e.g., Pu1-Eklf or Pu1-Fli1, revealing the actual regulatory target of Pu1. When spurious edges occur they were similar across the different inference frameworks: all 4 methods identified spurious edges in strongly connected regions of the HSC network, struggling to separate direct versus indirect connections [43]. However, for D-SPIN uniquely, spurious edges can be further reduced through sampling of higher-order perturbations (SI Text).

D-SPIN predicts the impact of simulated perturbations on differentiation

D-SPIN is a probabilistic or generative model that can predict the distribution of transcriptional states in a cell population subjected to a perturbation, unlike the other models, which use either regression or information-theoretic measures to find gene linkages. D-SPIN infers a quantitative model of the cell population generated by the HSC network that can be sampled to generate quantitative predictions regarding the distribution of cell-types induced by each gene knockdown or activation (Fig 2F). Unlike existing methods, D-SPIN allowed us to ask how specific gene knockdown or over-expression strategies could be applied to modulate the differentiation trajectory of HSC cells.

We applied D-SPIN to compute the population structure of a simulated HSC cell population when different transcription factors are perturbed. The model distributions quantitatively agree with data, and capture important biological impacts of the perturbations (Fig. 2F). Analysis of the model allowed us to gain insight into how disruption of different underlying network components might impact the differentiation trajectory of the cells. For example, when Pu1 is knocked down, it results in a complete loss of Monocyte cell-states and an increase in Granulocyte cell-states. The change in population structure occurs because Pu1 directly activates EgrNab and cJun, and the three factors together induce the Monocyte cell-state while EgrNab represses Gfi1, a key regulator of the Granulocyte cell-state. Knocking down Pu1, turns off the Pu1-cJun-EgrNab group leading to loss of the Monocyte cell-state as well as activation of Gfi1. Gfi1 activation leads to activation of Cebpa and the Granulocyte differentiation program. The Mega-Ery module (Eklf,Fog1,Gata1, and Scl) is affected by both the reduced inhibition from Pu1 and increased inhibition from Cebpa. Consequently, the cell states Mega and Ery remain present, albeit in reduced proportions.

Both Pu1 activation and the knockdown of Cebpa and Gfi1 lead to an increased prevalence of Monocyte cell-states (Fig. 2G). Pu1 directly activates the monocyte regulatory network that

contains EgrNab and cJun. Gfi1 represses monocyte differentiation through repression of EgrNab, and knock-down of Gfi leads to activation of EgrNab which in turn activates Pu1 and cJun, resulting in increased Monocyte states. The knockdown of Cebpa turns off Gfi1 because of the positive interaction between the two, leading to the activation of Monocyte states in a similar mechanism as Gfi1 knockdown.

Thus, D-SPIN, using simulated data, can infer an accurate and generative regulatory network model of a single-cell population by integrating information across perturbation conditions. D-SPIN prunes the ambiguity of alternative models, and builds a coherent integrated network upon which perturbations act. The inferred regulatory network reveals functional interactions between regulatory genes and provides a map of how the distribution of cell-states in the underlying HSC cell population are generated through interactions between internal regulators.

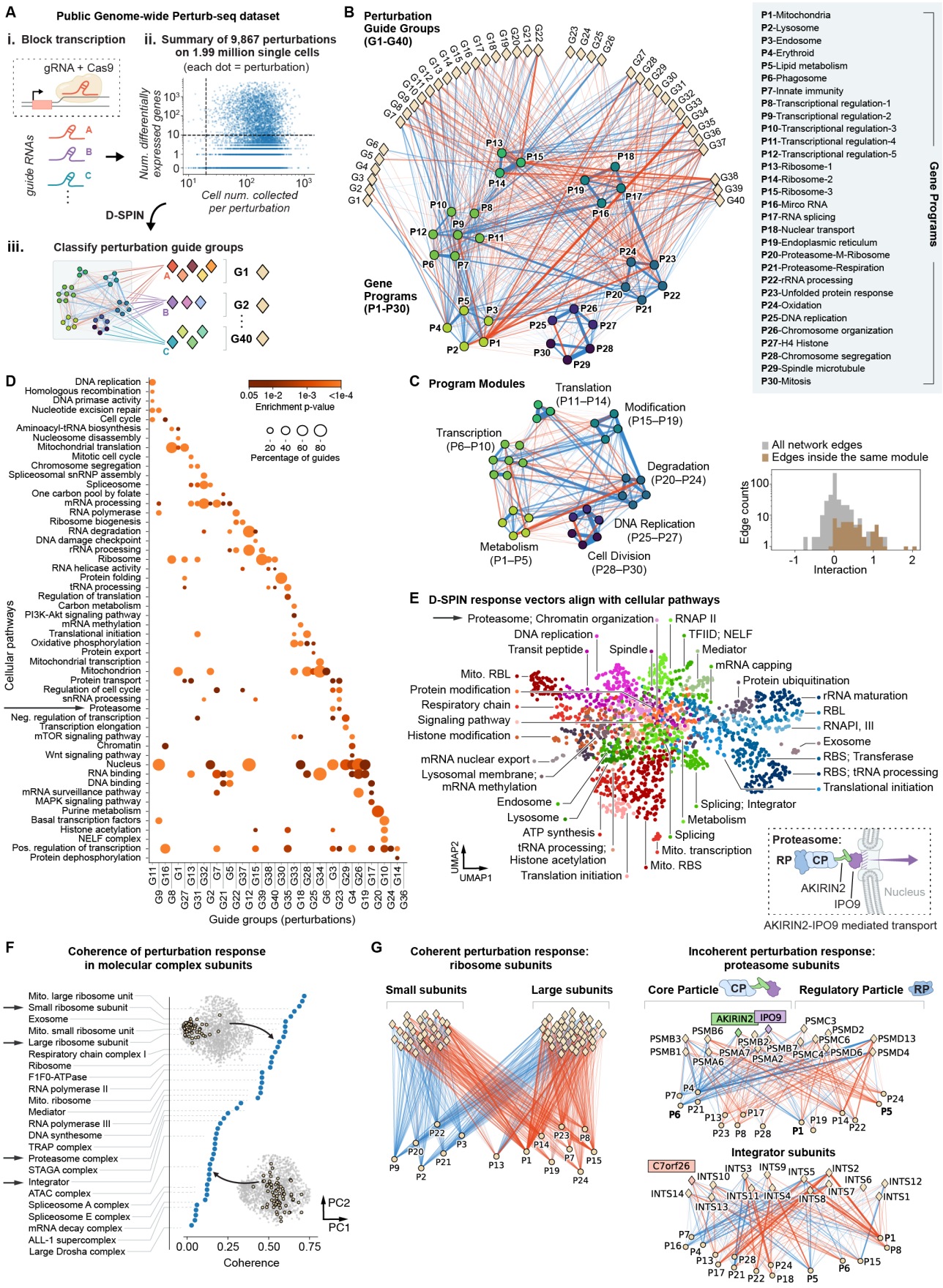


Figure 3: Caption on following page

Figure 3 **D-SPIN generates a regulatory network model from genome-wide Perturb-seq data that reveals the architecture of cellular pathways and sub-functions for pathway components.** (A) Genome-wide Perturb-seq experiment [28] delivered 9,867 gene knockdown perturbations individually to ∼2M human K562 cells using guide RNAs. (ii) D-SPIN model construction was applied to knockdown perturbations that induced more than 10 differentially expressed genes and more than 20 cells passed quality filtering. (iii) D-SPIN model was applied to generate a regulatory network model, inferring interactions between gene-expression programs and impacts of perturbations that classify guides into groups. (B) Inferred D-SPIN model contains a core gene regulatory network composed of 30 gene-expression programs and response of each gene knock down. Gene programs are annotated by a combination of function annotation tool and manual look up. D-SPIN classified gene knockdowns into 40 different guide classes with Leiden clustering. The interactions between gene-expression programs and perturbations are rendered as positive (blue) or negative (red) edges, and edge thickness scales with the absolute value of edge weight. (C) D-SPIN network model exhibits modular architecture with 7 core network modules that can be assigned physiological functions: metabolism, transcription, translation, modification, protein degradation, DNA replication, and cell division. Interactions between programs in the same modules are mostly activation and contain the majority of strong activation interactions. (D) Guide RNA classes are each enriched for physiological functions and cellular pathways in gene ontology and KEGG pathway databases. Bubble size represents percentage of enriched guides in the class. (E) UMAP representation of response vector of each guide RNA agrees with global organization of major cellular pathways including transcription, translation, respiration and cell cycle. (F) Regulatory coherence of macromolecular complex subunits, quantified by the average pairwise cosine similarity between the perturbation response vector for knockdown of complex sub-components. High-coherence complexes have subunits with similar impact on the D-SPIN gene regulatory network. In low-coherence complexes, knocking down different subunits leads to different impact, suggesting distinct functions for complex subunits. (G) Perturbation vectors for subunits of the ribosome, proteasome, and integrator complexes. The ribosome has a coherent signature where components of the complex have similar inferred response vectors, while the proteasome and integrator each have signatures of sub-functionalization. In the example of proteasome, low-coherence is observed as only a subset of subunits have impact on a gene program (P5, P6), or even display opposite effects on a gene program (P1). D-SPIN also identified recently annotated pathway components IPO9 and AKIRIN2 (proteasome) and C7orf26 (integrator).

Constructing a regulatory network model from genome-wide Perturb-seq data

Peturb-seq (or its derivatives) involves the knockdown of individual genes, using Cas9 and a targeting guide RNA. Single-cell mRNA-seq is used to assess the transcriptional state of each cell and to identify the guide RNA delivered to the cell. The technique can be applied at scale to interrogate the impact of thousands of distinct gene knock-downs on cellular gene expression. We used D-SPIN to construct a gene regulatory network model using data from a genome-wide perturb-seq experiment on the erythrocyte leukemia K562 cell line [28]. K562 is an erythrocyte precursor cell line with a BCR-ABL mutation, derived from a chronic myelogenous leukemia (CML) patient [44, 45]. In the genome-wide Perturb-seq experiment, $9,867$ genes were knocked down individually across 2 million single cells using CRISPRi. In the genome-wide perturb-seq experiment, 9,867 genes were knockdowned individually across 2 million single cells using CRISPRi. Prior to D-SPIN analysis, we pre-processed the data to identify guide RNAs that have an impact on the K562-cell transcriptional state. Guide RNA perturbation impact varies (Fig. 3A): 70.5% of perturbations had less than 10 differentially expressed genes (DEGs), 14.7% had 10-100 DEGs, and 14.8% had more than 100 DEGs. There is also great variation in the number of collected cells: 20.3% of perturbations had less than 100 cells collected. We focused our analysis on 3,138 perturbations with more than 10 DEGs and 20 collected cells, and D-SPIN was able to construct the network model in 8 hours with parallelization on 500 CPU cores.

We constructed a D-SPIN model using a set of 30 gene programs that D-SPIN extracted with its automated oNMF procedure (Fig. 3B). The extracted gene programs reflected both general cell biology as well as lineage-specific gene programs that reflect the specific biology of K562 cells. We annotated the gene programs with a combination of bioinformatic database including DAVID, Enrichr and manual lookup [46, 47]. Among the programs, D-SPIN extracted a series of programs that are enriched for genes involved in specific core cell processes including transcription, translation, DNA replication, mitosis, and RNA processing. D-SPIN also identified lineage-specific programs including an erythroid-specific program with hemoglobin

(HBG1, HBG2, HGZ) and glycophorin (GYPA, GYPB, GYPE) genes, as well as two myeloid-associated programs with phagosome/actin-organization (ACTB, ACTG1, ARPC3) and immune-response (LAPTM5, VASP, RAC2) genes respectively, which agrees with the multi-lineage potential of K562 cells (Fig. 3B).

Given the set of 30 gene programs, D-SPIN generated a gene regulatory network model that provides a wiring diagram of K562 internal regulation and the response of the K562 regulatory network to gene knockdown perturbations (Fig. 3A). The D-SPIN network contains gene programs, as nodes, and interactions between gene programs as edges that represent inferred interactions. D-SPIN identified 84 activating and 21 inhibiting interactions between gene programs, corresponding respectively to 19.3% and 4.8% of all gene-program pairs. The D-SPIN regulatory network reconstruction contained a set of sub-networks or modules of gene programs that interact with one another through dense positive (blue) interactions (Figs. 3B,C). We automatically decomposed the D-SPIN network into seven modules through the Leiden community detection algorithm [48]. Each identified module reflects a core cellular function, including transcription, translation, protein-degradation, and cell-cycle-related functions.

The network also contains a set of inhibitory interactions between gene programs that are expressed in distinct cellular states. The strongest inhibitory interaction was between gene programs expressed at different stages of mitosis, for example P29 Spindle microtubule inhibits both P25 DNA replication and P27 Histone. We also found strong inhibiting interactions between distinct stress-response programs, P23 Unfolded protein response and P2 Lysosome. The Unfolded protein response program consists of heat-shock protein (HSP90AB1, HSPD1, HSP90AA1) and other molecular chaperone (CCT8, CCT3, CCT5) genes, while the Lysosome program contains genes involved in lysosome genesis (PSAP, ASAH1, PLD3). We also observed inhibition between P4 Erythroid and P6 Phagosome, where the phagosome is an innate immune response mounted by macrophages and other myeloid cell types [49]. This indicates the presence of two, mutually exclusive differentiation paths that lead to erythroid and myeloid cell fates.

D-SPIN reconstructs the architecture of core cellular pathways

In addition to inferring the structure of the core K562 gene regulatory network, D-SPIN also inferred interactions between each gene program and each gene knockdown perturbation. The interactions between perturbations and gene programs are represented by perturbation response vectors $\boldsymbol{h}^{(n)}$. The entries of $\boldsymbol{h}^{(n)}$ represent positive or negative interaction between perturbations and individual gene programs. The response vectors can be represented as directed edges that connect perturbations to the core regulatory network (Fig. 3B).

Similarly to the grouping of the K562 gene regulatory network into interacting gene modules, we grouped gene knockdown perturbations into a set of 40 perturbation classes that we refer to as G1-G40 (Guide RNA group 1 through 40) through unsupervised Leiden clustering of the $\boldsymbol{h}^{(n)}$ vectors (Figs. 3B-E). Identified clusters reflect the pathway-level organization in the K562 cell, revealing both well-known cell biology and more novel or cryptic organization of pathways (Figs. 3D,E). Guide clusters were enriched for specific cell-biological functions, including DNA replication, MAPK signaling pathway, and RNA degradation (Fig. 3D). The guide clusters also provide detailed information regarding the structure of cellular pathways. As a specific example, the core components of the proteasome fell into a single perturbation cluster. The cluster also contains AKIRIN2 and IPO9, two protein components recently associated with the proteasome. Similarly, D-SPIN identified C7orf26 as associated with the integrator subunits INTS10, INTS13 and INTS14, a key result of the original primary Perturb-seq study [28].

Predicting macro-molecular complex component function

Macromolecular complexes like the ribosome, the proteasome, and the integrator complex play a critical role in cell physiology and gene regulation. We used D-SPIN perturbation vectors to analyze the functions of individual macromolecular-complex components. D-SPIN generates a perturbation vector $\boldsymbol{h}^{(n)}$ for each subunit of each macromolecular complex that was knocked down in the Perturb-seq experiment. We applied D-SPIN to gain insight into sub-functions of complex components by comparing cosine similarity between $\boldsymbol{h}^{(n)}$ for different components of known complexes in the CORUM database [50]. We selected complexes that have more than two thirds of subunits in the dataset and picked representative ones from subunits with more than half overlapping subunits, resulting in a list of of 55 core cellular complexes. For each complex in the list, we defined a coherence score by computing the average cosine similarity between $\boldsymbol{h}^{(n)}$ for all pairs of complex components. When subunits of a complex interact to achieve a common function, disruption of

individual subunits will disrupt that common function, leading to high cosine similarity between subunits and an average cosine similarity approaching 1. When complexes can achieve distinct functions through combinations of distinct protein components, then such complexes have average cosine similarity closer to 0 which is the random expectation.

Across the 55 complexes in CORUM, our analysis revealed a wide range of coherence scores across different complexes, ranging from 0.01 (large Drosha complex) to 0.73 (mitochondrial large ribosome) (Fig. 3F). Notably, complexes that play critical roles in specific biological functions including translation and metabolism, such as the mitochondrial ribosome, ribosome, exosome, DNA repair, and respiratory chain, exhibited the highest coherence scores. In contrast, complexes with known regulatory roles including the large Drosha complex, DGCR8 complex, ALL-1 supercomplex (involved in chromatin regulation), the ATAC complex (histone acetyltransferase) and spliceosome all had lower coherence scores $< .25$. The low-scoring complexes are involved in versatile regulatory roles in epigenetic regulation and microRNA and mRNA processing. D-SPIN provides a list of the candidate cellular sub-processes (gene programs) that are controlled by individual sub-units of each macromolecular regulatory complex (SI Table) providing a specific framework for generating experimental hypothesis to determine how modulation of subunits composition of these macromolecular regulatory complexes might enable the complex to achieve a range of regulatory functions.

The ribosome behaves as a coherent complex where knockdown of individual genes in the large or small ribosome subunit had similar impact on D-SPIN gene programs as illustrated by the monochromatic edge connections between subunit perturbations and gene programs in the interaction diagram (Fig. 3G). Perturbations to ribosome subunits activate protein digestion-related gene programs including P2 Lysosome, P3 Endosome, and P20 P21 Proteasome. The perturbations inhibit ribosome programs, P1 Mitochondria, P19 Endoplasmic reticulum, P23 Unfolded protein response, and P24 Oxidation. Thus, perturbing any ribosome subunits induced the cell into a stressed state with upregulated protein digestion and inhibition of protein production and secretion processes.

The proteasome and integrator behave as incoherent complexes, and have response vectors that split subunits into smaller sub-groups. The integrator complex subunits were partitioned into 4 sub-groups that agree with the original Perturb-seq study (Fig. 3J): an INTS10-13-14-C7orf26 module, a cleavage module (INTS3, INTS4, INTS8, INTS11), and two shoulder and backbone modules (INTS2, INTS5, INTS7, INTS8; INTS1, INTS6, INTS12). We found knockdown of the cleavage module activated P21 Proteasome and inhibited P22 rRNA processing, P18 Nuclear transport, P24 Oxidation, and P5 Lipid metabolism, while knockdown of the shoulder and backbone modules activated P4 Erythroid and P7 Innate immunity and inhibited P21 Proteasome, P28 Chromosome segregation, P17 RNA splicing, and P13 Ribosome programs.

Similarly, the proteasome complex subunits are split into two groups of core particles (20S proteasome) (PSMA6, PSMB1, PSMB3, PSMB6; PSMA2, PSMA7, PSMB2, PSMB7) and two regulatory particles (PSMC3, PSMC3, PSMC3, PSMD2, PSMD6; PSMD4, PSMD13). D-SPIN found that knockdown of proteasome subunit PSMC6 (also known as TBP7 and Rpt4) has a negative impact on respiration, and this result is consistent with the finding that PSMC6 interacts with mitochondrial proteins including TRAP1 in the endoplasmic reticulum to assist folding of mitochondrial proteins [51]. Further, AKIRIN2 and IPO9 were grouped together with the 20S proteasome subunits by the response vectors. These results demonstrated that the response vectors reveal impacts of genetic perturbations, and provide a compact signature for identifying functional associations between genes [52].

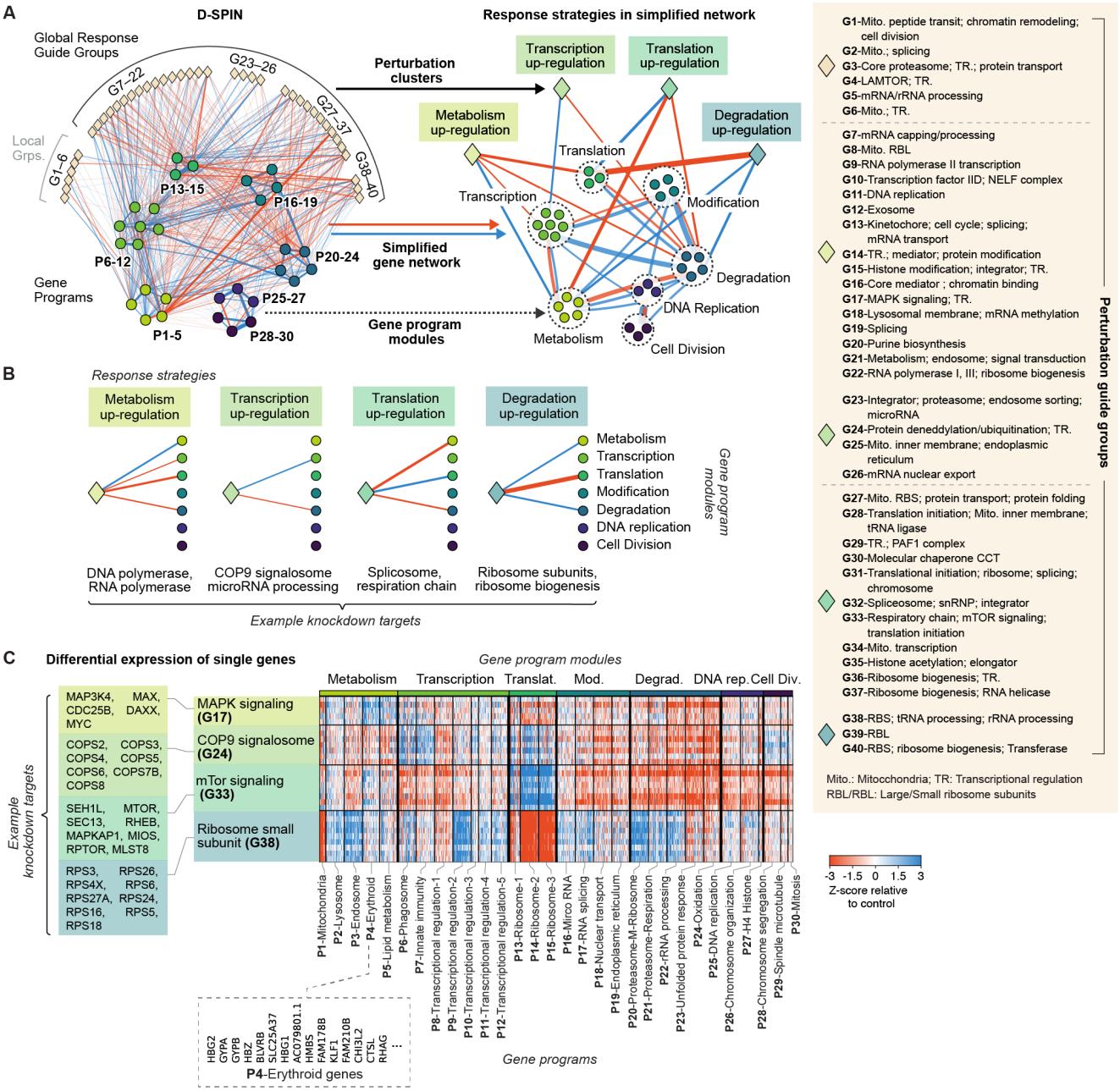


Figure 4: **D-SPIN network model identifies global regulatory response strategies mounted by K562 cells to distinct classes of gene perturbation.** (A) A coarse-grained network model is generated by grouping gene programs into 7 identified gene modules and grouping guide RNA gene knockdown classes based upon network impact. The resulting coarse-grained model enables global analysis of cellular regulatory responses. (B) The minimal model reveals that K562 cells mount 4 distinct classes of global regulatory responses in response to gene knockdown perturbations. For example, when RNA polymerase is knocked down, cells generate the metabolism upregulation response, where the cell up-regulates metabolism and down-regulates transcription, translation and degradation. As a second example, when ribosome subunits are knocked down, cells up-regulate protein degradation and metabolism while down-regulating translation. (C) Gene-expression heatmap showing the upregulation (blue) or downregulation (red) of genes in gene programs to perturbation of individual genes in a series of cellular pathways. Heatmap indicates the change in gene expression for perturbed cell population compared to the control where a non-targeting guide RNA is introduced to cells.

D-SPIN coarse-graining reveals global regulatory strategies

While the D-SPIN model in Fig. 3 is compact, we found that the model could be simplified further through an automated coarse-graining strategy (diagrammed in Fig. 4A) that reduces the D-SPIN network by grouping gene programs into sets that we call *modules* and single gene perturbations into perturbation clusters. We call this strategy *network coarse-graining* because of its similarity with coarse-graining strategies in physical models that group strongly correlated degrees of freedom into a single variable to aid computation and interpretation. The coarse-graining strategy produced a minimal network that we analyzed to reveal 4 classes of damage-response strategies, patterns of gene activation and inhibition generated by the cell when specific cellular processes are disrupted.

The D-SPIN-identified regulatory strategies suggest non-trivial forms of homeostatic regulation. We named the strategies by the most prominent upregulation response to the perturbations, because in the context of knockdown perturbations, gene activation represents the active regulatory feedback of cell to encounter the perturbed or stressed condition. Therefore the four global response strategies are metabolism upregulation, transcription upregulation, translation upregulation and degradation upregulation. For example, perturbation to components impacting translation initiation (G28) leads to upregulation of translation and downregulation of protein degradation, suggesting that the cell compensates for decreased translational flux through a regulatory response that both up-regulates translation while also down-regulating processes that degrade proteins. Interestingly, perturbations to ribosome subunits and ribosomal rRNA (G38, G39, G40) production actually leads to a distinct response where the cell down-regulates translation while up-regulating protein degradation, suggesting a cellular attempt to contain the impact of unassembaled ribosome subunits and damaged proteins.

The regulatory strategies also provide insight into the flow of information in the cell. knockdown of genes often impacted processes that suggested long-range regulation or coupling between seemingly distinct processes. For example, a large proportion of metabolism upregulation response are related to transcription processes, including RNA polymerase (G9, G22), Transcription Factor IID, NELF complex (G10) and mediator (G16). The knockdown of these genes led to upregulation of metabolism, and downregulation of transcription, translation, and degradation. It has been found that transcription stress would induce an elevated ATP pool and rewired metabolism states [53]. As another example, the COP9 signalosome (CSN) is a protein complex involved in deubiquitination and deneddylation of proteins. Knockdown of components including COPS2, COPS4, and COPS6 led to upregulation of the phagosome and downregulation of protein-degradation machinery (Fig. 4C). Knockdown of mTOR signaling components led to upregulation of ribosomal programs in a translation upregulation response. Finally, knockdown of small ribosomal subunits led to downregulation of translation and upregulation of transcription and protein degradation.

These results demonstrate that D-SPIN-generated regulatory networks provide insights into control strategies used by cells to regulate a distributed set of processes to achieve homeostasis in response to various forms of internal damage. D-SPIN models can provide global insights into classes of regulatory strategies, suggest mechanisms of compensation and feedback, and also identify information flow between cellular processes that might not be intuitively connected, thereby providing insights into principles of information processing and homeostatic control.

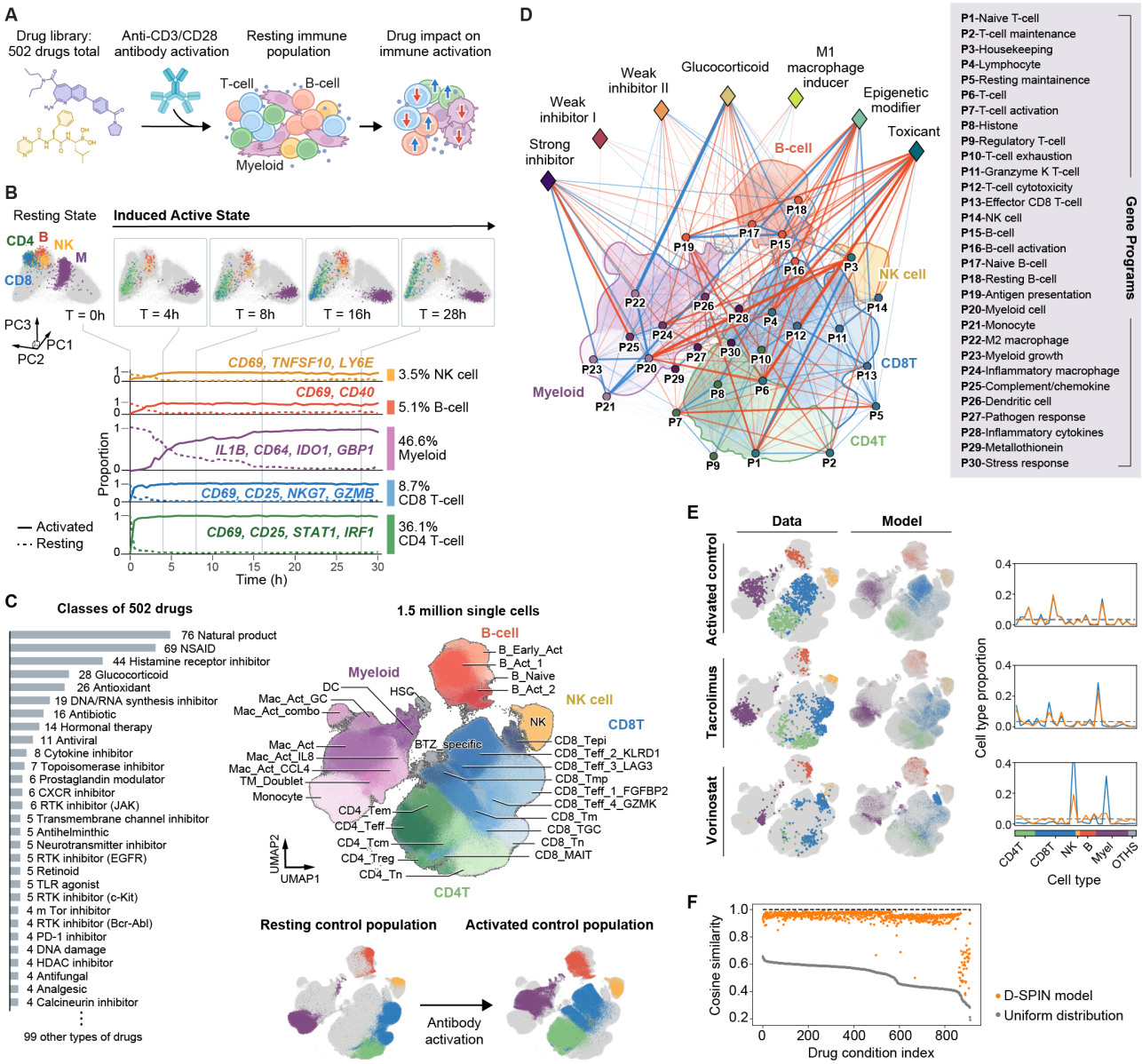


Figure 5: Caption on following page

Figure 5 **D-SPIN derives drug-response network model from human immunomodulatory drug-response single-cell mRNA-seq profiling experiments.** (A) Illustration of drug-response experiment design. Peripheral blood mononuclear cells (PBMCs) were harvested from a human donor. The cell population was treated by anti-CD3/CD28 antibody that specifically activate T-cell and immunomodulatory drugs drawn from a library. The cell population was profiled after 24 hours of drug and antibody treatment. (B) PCA projections derived from single-cell mRNA-seq of T-cell mediated immune activation at selected timepoints within a 30-hr time-course experiment. Time-courses of the proportion of activated and resting cells in each cell type with activation gene markers. T-cells reach activated states first in 2 hrs and myeloid cell activation lasts until 16 hr. (C) Histogram of classes of profiled drugs. The selected drugs have a variety of biochemical properties and target pathways. In the profiled cell population We identified 32 cell states in the major cell types of T, B, NK, and myeloid cell and each cell state is curated by marker genes and gene differential expression. The UMAP embedding is primarily divided into two distinct regions of resting and activated cell population. (D) The regulatory network extracted by D-SPIN contains 30 gene programs that are identified from data with oNMF and annotated with cellular function by a combination of bioinformatic database and manual lookup. Drugs were classified into seven drug classes based on response vectors identified by D-SPIN. D-SPIN network is rendered on top of a UMAP schematic of the cell population: the UMAP shows the underlying cell states, and D-SPIN nodes are positioned on the UMAP based upon the cell state in which each program highly express. (E) Examples of model-generated cell state distribution compared with data. The model and data agree well in the control and Tacrolimus-treated samples. The model fits less well in the Vorinostat-treated sample but still captures qualitative features of the distribution. The model UMAP are rendered by directly sample from the model and map each state to nearest data. Dashed line in cell state distribution are uniform distribution as reference. (F) Model generated cell-state distribution are highly similar to experimental data in cosine similarity measure. The cosine similarity is higher than 0.9 for the majority of conditions. For comparison, the null model is a uniform distribution over cell subtypes. Drug conditions are ordered by cosine similarity with uniform distribution for visualization.

Modeling immunomodulatory drug responses in primary human immune cells

D-SPIN models can contribute to the design of therapeutic strategies by providing global insights into how different cell types respond to therapeutic interventions. Given the role of the human immune system in many diseases, we applied D-SPIN to dissect the response of human immune cells to a library of small molecule immunomodulatory drugs. We developed an experimental data set in which we applied single-cell mRNA-seq to profile the response of human immune cells to 502 different immunomodularity drugs. Our experimental system was designed to probe the modulation of a cell population undergoing T-cell driven hyper-activation as is observed in auto-immune and hyperinflamatory states [54, 55]. We culture a heterogeneous population of donor derived human immune cells (peripheral blood mononucleated cells, PBMC) that contains T-cells, B-cells, macrophages/monocytes, and NK-cells. T-cells are activated using a pair of T-cell receptor antibodies (Fig. 5A), leading to activation of macrophage, B-cell, and NK-cell cell-types (Fig 5B). We applied the system to analyze the action of different immunomodulatory drugs on the activating immune cell population (Fig 5C).

Single-cell mRNA-seq was used to measure the effect of 502 immunomodulatory small molecules on T-cell driven-immune activation (Fig. 5C). The drug library contained a diverse set of small molecules targeting pathways including mTor, MAPK, glucocorticoids, JAK/STAT, and histone deacetylases (HDAC) (Fig. 5C, SI Table). The library contains drugs used for treating auto-immune disease (e.g., Tacrolimus, Budesonide, Tofacitinib) as well as FDA-approved anti-cancer drugs (e.g., Bosutinib, Crizotinib). We added each compound at 1μM at the same time as anti-CD3/CD28 antibody, and profiled the cell population using single-cell mRNA-seq at 24 hours [56]. For additional context, we also profiled drug responses in resting cell populations without CD3/CD28 activation for a subset of drugs. In total, we profiled 1.5 million single cells in resting and activated conditions, with over 1,200 total conditions and 31 different immune cell states including 4 CD4 T-cell states, 10 CD8 T-cell states, NK-cell state, 4 B-cell states, and 8 myeloid cell states (Fig. 5C).

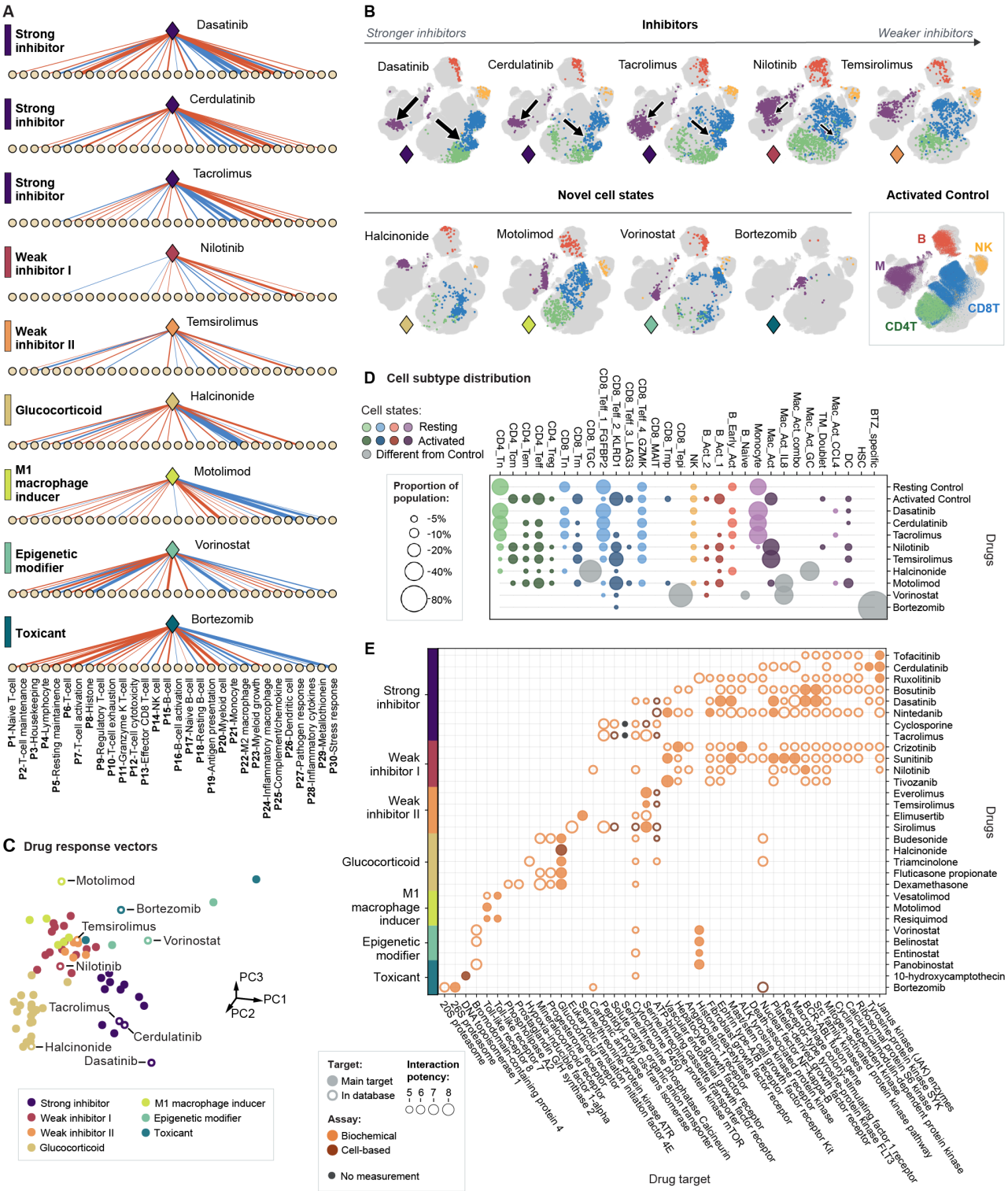
Classification of small molecules using D-SPIN

Using the single-cell mRNA-seq data, we applied D-SPIN to construct a regulatory network that models the interaction of individual small molecules in the drug library with the immune cell population, which contains various cell-types (T-cells, B-cells, macrophages, and NK-cells) as well as states of activation and rest. We applied D-SPIN in an unsupervised manner, using the algorithm to construct a single, integrated regulatory network model that can capture all cell-types and cell states by modeling interactions between gene-expression programs and interactions between gene-expression programs and applied drugs. D-SPIN extracted a set of 30 gene programs by oNMF of the transcription data with seeding to include drug-specific programs (SI Text) and the programs were sufficient to describe the transcriptional state of all cells in the population. The decomposition yielded known T-cell and macrophage gene-expression programs including T-cell resting (P5) and activation programs (P7), macrophage anti-inflammatory (P22 M2 macrophage) and pathogen-response programs (P27), an antigen-presentation program (P19), and global cell-state programs for T cells (P6), B cells (P15), NK cells (P14), and myeloid cells (P20). The function of each program was annotated by a combination of informatics database [46] and manually annotation, and validated on a single-cell atlas of human immune cells [57].

D-SPIN inferred a single interaction network $\boldsymbol{J}$ and a set of 502 drug-specific biasing vectors $\boldsymbol{h}^{(n)}$. The D-SPIN model contains a total 465 network parameters and 30 parameters for each drug condition but was able to describe all profiling experiments of 502 drug with low quantitative and qualitative error (SI) (Fig. 5E,F). Quantitatively, we computed the cosine similarity between model and data distribution, and in 92.4% of samples the model is >90% similar to data (Fig. 5F). The drugs with lower cosine similarity tended to be drugs that drove the cell population into a small number of cell states, like the proteasome-inhibitor Bortezomib which produced a strong signature of cell stress in all cells. For these drugs, the D-SPIN model produced a cell-population distribution that had qualitative agreement with the data but differed quantitatively, likely to do the extremely focused cell-type distribution for these drugs.

Like the K562 cell network, the immunomodulatory drug-response network had a distinct modular structure with a set of core network modules composed of tightly interacting gene programs (Fig. 5D). The modules in the drug data contained groups of gene programs that are expressed together in individual cell types. For example, a module corresponding to anti-inflammatory myeloid cells contains P21 Monocyte, P22 M2 macrophage and P23 Myeloid growth and a network hub P20 Myeloid cell. Similarly in T-cells, another module containing P11 Granzyme K T-cell, P12 T-cell cytotoxicity, P13 Effector CD8 T-cell and P14 NK cell corresponds to the population of CD8 T cells and NK cells. In the D-SPIN model, negative interactions primarily occur between gene programs expressed in different cell types, e.g., P2 T-cell maintenance and P19 Antigen presentation, as antigen presentation happens in B-cells and myeloid cells but not T-cells.

In addition the regulatory network model, D-SPIN also inferred an interaction vector $\boldsymbol{h}^{(n)}$ that quantifies interactions between each drug and the regulatory network. By comparing the inferred interaction vector with control experiments, we identified 71 out of 502 drugs that had statistically significant interactions with the core gene network. Moreover, as in the case of the Perturb-seq data, we applied unsupervised clustering to group drugs into a set of 7 classes including strong inhibitor, weak inhibitor I, weak inhibitor II, glucocorticoid, M1 macrophage inducer, epigenetic regulator, and toxicant based on their interactions with the underlying regulatory network (Fig. 5C, SI Fig). We named the classes by analyzing the structure of the cell population induced by drugs within the class. While we defined the drug classes based only on D-SPIN's inferences regarding how each drug interacts with the core network, the classes contained drugs from specific chemical families and mechanisms of action.


A
Strong inhibitor
Dasatinib
Strong inhibitor
Cerdulatinib
Strong inhibitor
Tacrolimus
Weak inhibitor I
Nilotinib
Weak inhibitor II
Temsirolimus
Glucocorticoid
Halcinonide
M1 macrophage inducer
Motolimod
Epigenetic modifier
Vorinostat
Toxicant
Bortezomib
P1-Naive T-cell
P2-T-cell maintenance
P3-Housekeeping
P4-Lymphocyte
P5-Resting maintainence
P6-T-cell
P7-T-cell activation
P8-Histone
P9-Regulatory T-cell
P10-T-cell exhaustion
P11-Granzyme K T-cell
P12-T-cell cytotoxicity
P13-Effector CD8 T-cell
P14-NK cell
P15-B-cell
P16-B-cell activation
P17-Naive B-cell
P18-Resting B-cell
P19-Antigen presentation
P20-Myeloid cell
P21-Monocyte
P22-M2 macrophage
P23-Myeloid growth
P24-Inflammatory macrophage
P25-Complement/chemokine
P26-Dendritic cell
P27-Pathogen response
P28-Inflammatory cytokines
P29-Metallothionein
P30-Stress response
B
Inhibitors
Stronger inhibitors
Weaker inhibitors
Dasatinib
Cerdulatinib
Tacrolimus
Nilotinib
Temsirolimus
Novel cell states
Activated Control
Halcinonide
Motolimod
Vorinostat
Bortezomib
B
NK
M
CD8T
CD4T
C Drug response vectors
Motolimod
Bortezomib
Temsirolimus
Vorinostat
Nilotinib
Tacrolimus
Cerdulatinib
Halcinonide
Dasatinib
PC3
PC1
PC2
Strong inhibitor
Weak inhibitor I
Weak inhibitor II
Glucocorticoid
M1 macrophage inducer
Epigenetic modifier
Toxicant
D Cell subtype distribution
Cell states:
Resting
Activated
Different from Control
Proportion of population:
-5%
-10%
-20%
-40%
-80%
CD4_Tn
CD4_Tcm
CD4_Tem
CD4_Teff
CD4_Treg
CD8_Tn
CD8_Tm
CD8_TGC
CD8_Teff_1_FGFBP2
CD8_Teff_2_KLRD1
CD8_Teff_3_LAG3
CD8_Teff_4_GZMK
CD8_MAIT
CD8_Tmp
CD8_Tepi
NK
B_Act_2
B_Act_1
B_Early_Act
B_Naive
Monocyte
Mac_Act
Mac_Act_IL8
Mac_Act_combo
Mac_Act_GC
TM_Doublet
Mac_Act_CCL4
DC
HSC
BTZ_specific
Resting Control
Activated Control
Dasatinib
Cerdulatinib
Tacrolimus
Nilotinib
Temsirolimus
Halcinonide
Motolimod
Vorinostat
Bortezomib
Drugs
E
Strong inhibitor
Weak inhibitor I
Weak inhibitor II
Glucocorticoid
M1 macrophage inducer
Epigenetic modifier
Toxicant
Tofacitinib
Cerdulatinib
Ruxolitinib
Bosutinib
Dasatinib
Nintedanib
Cyclosporine
Tacrolimus
Crizotinib
Sunitinib
Nilotinib
Tivozanib
Everolimus
Temsirolimus
Elimusertib
Sirolimus
Budesonide
Halcinonide
Triamcinolone
Fluticasone propionate
Dexamethasone
Vesatolimod
Motolimod
Resiquimod
Vorinostat
Belinostat
Entinostat
Panobinostat
10-hydroxycamptothecin
Bortezomib
Drugs
20S proteasome
26S proteasome
DNA topoisomerase I
Toll-like receptor 8
Phosphatase 1
Phosphatase 2A
Hydroxysteroid 11-beta dehydrogenase 1
Mineralocorticoid receptor
Glucocorticoid receptor
Eukaryotic translation initiation factor 4A
Cytochrome P450, family 51, subfamily A, polypeptide 1
Peptidyl-prolyl cis-trans isomerase
Solute carrier organic anion transporter
Sterol O-acyltransferase 1
Cytochrome P450, family 19, subfamily A, polypeptide 1
Sterol 14-demethylase
Lanosterol 14-alpha demethylase
Glucocorticoid receptor
Cytochrome P450, family 11, subfamily A, polypeptide 1
Cytochrome P450, family 17, subfamily A, polypeptide 1
Cytochrome P450, family 11, subfamily B, polypeptide 2
Cytochrome P450, family 11, subfamily B, polypeptide 1
Cytochrome P450, family 11, subfamily A, polypeptide 1
Cytochrome P450, family 21, subfamily A, polypeptide 2
Cytochrome P450, family 21, subfamily A, polypeptide 1
Cytochrome P450, family 17, subfamily A, polypeptide 1
Cytochrome P450, family 19, subfamily A, polypeptide 1
Cytochrome P450, family 51, subfamily A, polypeptide 1
Cytochrome P450, family 11, subfamily A

Three of the seven classes include drugs that inhibit activation of the immune-cell population (Fig. 5D, SI), including drugs we classified as strong inhibitor, weak inhibitor I, and weak inhibitor II. Across these drug classes, analysis of cell-population fractions and UMAP visualization indicated that these drugs block the transition of T , B , and myeloid cell types to the activated state, shifting the balance of the cell population to resting cell states but acting with a spectrum of inhibition strengths (Fig. 6A-D). Very strong inhibitors, including the immunosupressive drug Tacrolimus and the cancer drug Dasatinib, completely block the activation of T-cells and myeloid cells, inducing a cell population with a similar cell-state profile to unstimulated PBMC cells (Fig. 6D). Weak inhibitors, such as Temsirolimus, only slightly bias the cell population towards resting, compared with the control population.

The individual drugs that fall into the inhibitor drug classes act on a variety of molecular targets and signaling pathways [58] (Fig. 6E). The strong-inhibitor class targets include JAK (Tofacitinib, Cerdulatinib, Ruxolitinib), BCR-ABL and Src family kinases (Bosutinib, Dasatinib), and calcineurin (Cyclosporine, Tacrolimus), which are all pathways downstream of T-cell receptor signaling. The drugs in the weak-inhibitor I class target another group of receptors including ALK (Crizotinib), FLT3 (Sunitinib), and various growth-factor receptors (Crizotinib, Sunitinib, Nilotinib, Tivozanib). The drugs in the weak-inhibitor II class target protein kinases including mTOR (Everolimus, Temsirolimus, Sirolimus) and ATR (Elimusertib). Broadly, the analysis suggests that inhibition can be achieved via a range of distinct biochemical pathways and mechanisms. However, the strongest inhibitors in our data have high affinity for signaling molecules immediately downstream of T-cell receptor activation, including the Src family kinase Lck. Further, while many of the drugs in our data set have been previously reported as immunosuppressive including well-known drugs such as Tacrolimus, D-SPIN organized the drugs into a quantitative series and provides a single integrated framework for comparing drug action at the transcriptome scale across a large number of inhibitors with different chemical mechanisms of action.

Detailed comparison between the effects of drugs with the same primary target demonstrates that the single-cell assay and D-SPIN model can identify phenotypic patterns that capture potential underlying differences in the biochemical mechanism of drug action (Fig. 6E). A specific set of BCR-ABL kinase inhibitors (Bosutinib, Dasatinib) developed for the treatment of leukemia generated strong inhibition, while another set of BCR-ABL inhibitors, including Nilotinib, fall into the weak-inhibitor class (Fig. 6E). Biochemically, the BCR-ABL strong inhibitors can be distinguished from the BCR-ABL weak inhibitors by their increased Lck/Src kinase specificity (Fig. 6E). Lck kinase is a central signaling node directly downstream of T-cell receptor activation. Similarly, the JAK-kinase inhibitor Cerdulatinib generated a stronger inhibitory response than the FDA-approved anti-inflammatory JAK-kinase inhibitor Tofacitinib. While both drugs have JAK-family kinases as their primary target, Cerdulatinib also targets the kinase SYK. Therefore, the response vectors generated by D-SPIN allowed us to identify potential signaling nodes that differentiate responses from related drugs.

Beyond the strong and weak inhibitors, D-SPIN identified a class of drugs that contains glucocorticoids (GCs), steroid-derived small molecules that activate the glucocorticoid receptor (Fig. 5D, 6A-D). The glucocorticoid class of drugs included well-known immunosuppressive drugs including Halcinonide, Budesonide, Triamcinolone and dexamethasone. The GCs generated a population response that was similar to the strong inhibitors like Dasatinib between the classes. However, the two classes of drugs also had important differences in their interaction with the D-SPIN regulatory network. Specifically, D-SPIN inferred that the GCs not only inhibit activation programs in T-cells and myeloid cells, but also activate a specific immunosuppressive gene program in macrophages P22 M2 macrophage named due to the presence of genes associated with immunosuppressive M2 macrophages including CD163 [59, 60]. Therefore, D-SPIN inferred important differences in the immunosuppressive effect of GCs and other inhibitors. GCs had weaker suppression of programs associated with chemokine secretion (P25), but induced the M2 macrophage program that was more weakly activated by drugs in the strong-inhibitor classes.

In addition to the suppressive drug classes, D-SPIN also identified a group of drugs that induced the activation of inflammatory, pathogen-responsive or M1 macrophages. The class included small-molecule activators of Toll-like receptors, receptors in macrophages that sense the presence of pathogens and whose activation induces an innate immune response. The M1 macrophage inducer class contains TLR7 agonists (Vesatolimod, Resiquimod) and TLR8 agonists (Motolimod, Resiquimod), and produced macrophage states related to host defense that highly express P27 Pathogen response and P29 Metallothionein.

Finally, D-SPIN identified two classes of drugs which we call epigenetic modifiers and toxicants that D-SPIN identified to have activated cell-stress programs. Toxicants included the proteasome inhibitor Borte-

zomib and DNA damage agents Panobinostat and DNA topoisomerase inhibitor (10-Hydroxycamptothecin). Toxicants strongly activate P30 Stress response and have mostly inhibitory interactions with other gene programs, especially generic cell type programs such as P6 T cell, P20 Myeloid cell and P4 Lymphocyte. The epigenetic-modifiers class consists of HDAC inhibitors which generated an epigenetically disrupted T-cell state (CD8_Tepi) that had elevated expression of histone-related genes such as histone component genes and DNA topoisomerase TOP2A.

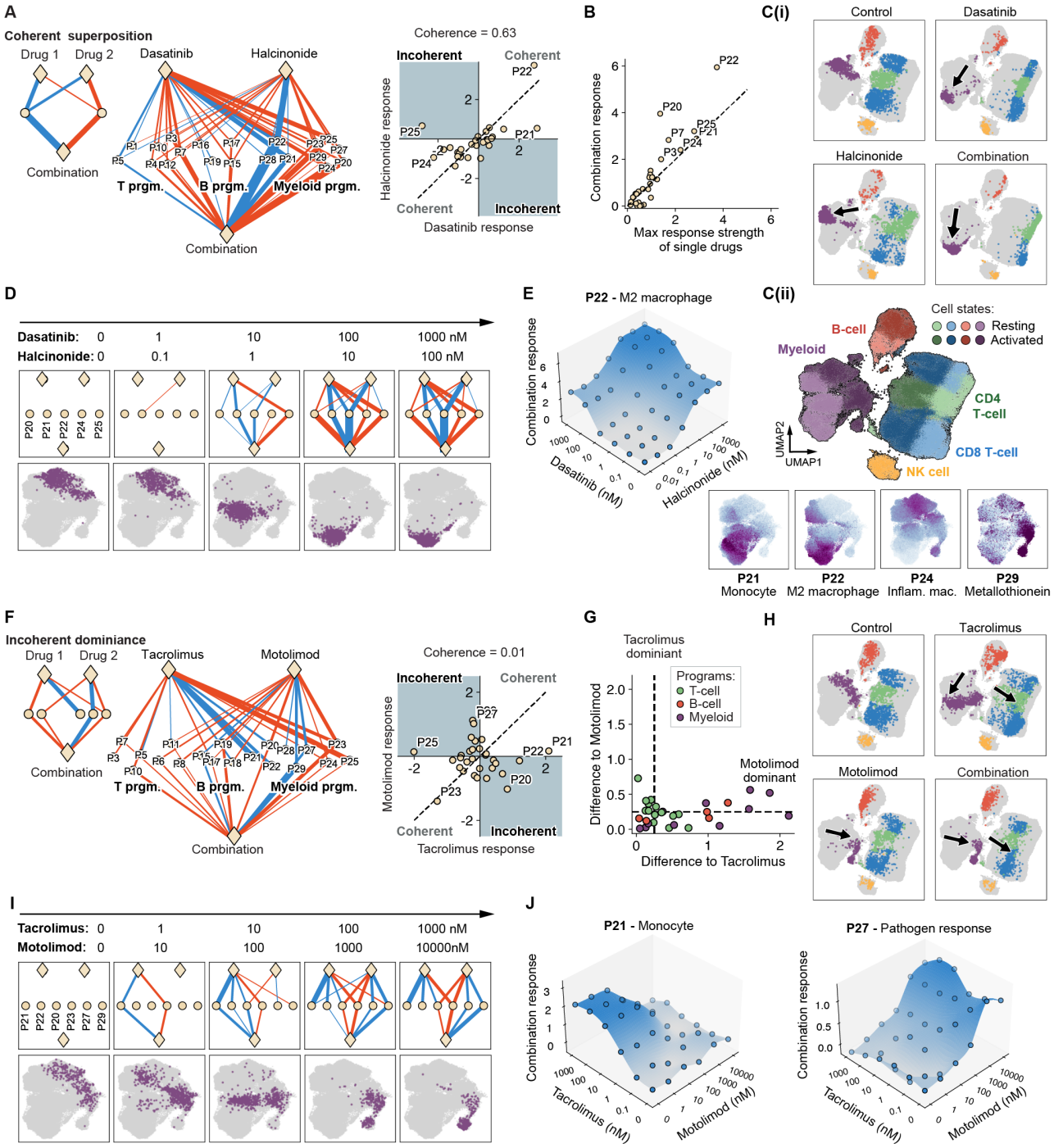


Figure 7: Caption on following page

Figure 7 **D-SPIN reveals network-level mechanisms of combinatorial drug action.** (A) Schematic diagram of coherent superposition and the combination perturbation response of Dasatinib and Halcinonide. The two single drugs and their combination have a coherent effect on gene programs by activating/inhibiting the same set of programs. The two drugs have a high coherence score (0.63), but different-strength interactions with gene programs. (B) The combination responses on gene programs are superposition of single drug responses as they are generally higher than the max strength of the single-drug responses. (C)(i) Comparison of single drugs and combination effect with UMAP rendering. The combinatorial recruitment of gene programs created a novel macrophage state that occupies a distinct region on the UMAP. (ii) UMAP rendering of cell states and gene-program expression of the two dosage-titration experiments. Gene programs are rendered on UMAP based on cell expression intensity. (D) combinatorial response of selected dosage series and most impacted myeloid programs. The mechanism of coherent superposition is consistent across drug dosages combinations. (E) On P22 M2 macrophage program, coherent superposition produced a response function close to the sum of the sigmoid-shaped single-drug dosage responses. (F) Schematic diagram of incoherent dominance and the combination perturbation response of Tacrolimus and Motolimod. The two drugs have an incoherent effect on gene programs and have a low coherence score (0.01). The two drugs act on different gene programs and the combination response is dominated by different drugs (G) On myeloid and B-cells, the response is dominated by Motolimod. On T-cells, some programs are dominated by Tacrolimus and some are dominated by Motolimod. (H) The gene-program dominance of drug effects produced novel cell states, with inhibited T-cell and activated macrophage states induced by Motolimod. (I) Across dosage combinations strengths, Motolimod dominates over Tacrolimus on the key myeloid programs. (J) On the P22 Monocyte program, the activation effect of Tacrolimus is eliminated with increasing Motolimod dosage. On the P27 Pathogen response program, Tacrolimus dosage has no impact on the activation effect of Motolimod.

Drug combinations that generate hyper-suppression

Immunomodulatory drugs are often used in combinations, and drug combinations could in principle be used to tune the transcriptional state of immune cell populations. However, the design and analysis of drug combinations is challenging at the transcriptome scale due to a large number of potential drug-gene interactions. D-SPIN models provide a framework to compare the action of individual drugs on the underlying gene regulatory network in order to identify drug combinations with potentially useful therapeutic applications. To analyze the drug responses classes identified by D-SPIN to identify potentially useful drug combinations, we profiled these combinations experimentally, and applied D-SPIN to interpret the mechanisms of combination action.

We first investigated the network level similarities and differences between glucocorticoids (GCs) and strong inhibitors. In our experiments, GC action on the D-SPIN regulatory network was similar to the action of strong inhibitors, with both classes of drugs suppressing the activation of T-cells, macrophages, B-cells and NK-cells. However, GCs induced weaker suppression, allowing cells to activate inflammatory cytokines, including CXCL10 and CXCL11 in the compliment/chemokine program whereas these cytokine programs are strongly suppressed by strong inhibitors, including the Src family inhibitor Dasatinib. GCs uniquely induce an M2 macrophage gene-expression program, which could be useful in some therapeutic settings, due to the expression of anti-inflammatory signaling molecules. The difference in impact between GCs and strong inhibitors is seen in the comparison of the inferred network impact of the GC Halcinonide with the inhibitor Dasatinb (Fig. 7B): Gene-expression programs P22 (cytokine), P25 (M2 Macrophage), and P21 (resting macrophage/monocyte; Jialong check) are induced or repressed more strongly by one drug then the other.

Therefore, we investigated whether combinations of glucocorticoids and strong inhibitors could induce both the M2 macrophage-activation and cytokine-repression effects, by performing single-cell mRNA-seq on cell populations treated with glucocorticoids in combination with strong inhibitors (including Dasatinib) across a range of doses. We used D-SPIN to analyze the impact of the drug combination on the gene regulatory network and underlying cell population. We found that the drug combination induced both the strong suppression of Dasatinib as well as the induction of macrophage gene-expression programs for specific relative concentrations (Fig. 7BCD). The response vector for the drug combination showed that the impact of Dasatinib and Halcinonide are additive at the gene-program level across combination of different concentrations. The response vector for the drug combination indicated significantly higher M2 macrophage induction than for Dasatinib or Halcinonide alone (Fig. 7E), while also exhibiting increased suppression of

macrophage activation over the single drugs. The interaction vector also maintained strong activation of the resting program P21 Monocyte.

The combination of Dasatinib and Halcinonide, therefore, generated a hyper-repressed cell population with a T-cell population resembling resting T-cells (Fig. 7C) and a macrophage cell population with activation of both resting myeloid program P21 and M2 macrophage program P22. The modulation was monotonic with dose, and manipulation of the drug doses allowed conversion of the cell population from activated to intermediate suppression to hyper-suppression M2 macrophage-like states (Fig. 7D). Conceptually, by combining the two drugs, we were effectively able to fine-tune the transcriptional state of the macrophage population by exploiting a superposition interaction where the drug combination recruited a super-set of gene-expression programs in a dose-dependent fashion that was also observed in previous study on single protein level [61].

Drug combinations that activate the innate immune system while repressing T-cell activation

In a variety of therapeutic settings it can be advantageous to modulate the balance between adaptive (T-cell driven) and innate (macrophage-driven) immune responses. For example, T-cell responses are suppressed in transplant patients because T-cells can react adversely with transplanted organs, however, infection then becomes a significant cause of mortality. Thus, induction of an innate immune response in the background of immuno-suppression provides a strategy for combating infection [62, 63, 64]. To understand whether D-SPIN could be applied to tune the balance between different immune cell states at the population level, we used D-SPIN model-identified M1 macrophage-inducing small molecules, including Motolimod, as well as standard immunosupressive small molecules, including Tacrolimus, which were revealed to be mostly incoherent in their impact on the underlying gene regulatory network (Fig 7F). Specifically, the global cosine similarity between Motolimod and Tacrolimus responses was only 0.01 and the two drugs were biased to impact T-cell (Tacrolimus) or macrophage (Motolimod) gene-expression programs, with Tacrolimus activating resting T-cell programs and Motolimod activating a pathogen-responsive macrophage program, P28 (Fig 7G).

To exploit the incoherence between Tacrolimus and Motolimod, we added the drugs in combination across a series of doses, applied single-cell mRNA-seq to profile the resulting cell population, and analyzed the combinatorial drug response with D-SPIN. We identified a concentration regime (Fig 7H) in which the cell population contains activated macrophages that exhibit strong induction of the pathogen response gene-expression program while maintaining T-cell repression, with higher than 1000nM Motolimod and higher than 100nM Tacrolimus (Fig 7I). D-SPIN inferred that the specific combination of concentrations enables the activation of the pathogen response macrophage program while maintaining activation of resting T-cell gene-expression programs. Conceptually, the Motolomid and Tacrolimus drug combination demonstrates that drugs impacting separate cell-types can be combined to tune the structure of an immune cell population.

Discussion

Here, we introduce D-SPIN, a probabilistic modeling framework, that constructs quantitative, generative models for gene regulatory networks from large scale single-cell mRNA-seq data sets collected from cell populations across a series of perturbation conditions. The D-SPIN framework is general and can be applied to study the response of cells to genetic perturbations, small molecules, and signaling proteins. D-SPIN models can be treated both as network models that are analyzed using the tools of graph theory as well as fully probabilistic models that can predict the structure of a cell population across a series of different conditions. The mathematical structure of D-SPIN allowed us to develop a computationally efficient, parallel inference procedure that can be run on hundreds of CPU cores to perform network inference on data sets with thousands of perturbations and millions of cells. Our work establishes an interpretable model framework for transcriptome-scale perturbation datasets. The D-SPIN framework is a maximum-entropy modeling framework adapted from analysis of systems in condensed-matter physics matter physics [65]. The framework provides a low-dimensional, compact yet predictive framework for analyzing population-level effects in single-cell data through a regulatory- network model.

Single-cell mRNA-seq methods enable large-scale perturbation response studies across cell-types and organisms. D-SPIN provides a computational framework for integrating information from such studies into regulatory network models that can be analyzed and compared to reveal principles of cellular regulation.

We have demonstrated that D-SPIN provides insight into damage response strategies in K562 cells. Stress response strategies play a role in many diseases including cancer, and diseases of ageing and degeneration. Further, cell response strategies change over the life-time of an organisms and also over evolutionary time-scales. Our understanding of gene regulatory network architecture, logic, and evolution have been limited by the limited number of network reconstructions that exist. D-SPIN provides a framework in which to construct, analyze, and compare regulatory networks to understand how they change over time.

Cells are distributed control systems that modulate many internal processes to main homeostasis and execute cell-fate transitions. However, the principles of distributed control at the transcriptome scale are poorly understood. D-SPIN models help reveal the principles of distributed control. We show that K562 cells mount a set of regulatory responses when different internal processes are perturbed through gene knockdown. The response strategies point to a broadly distributed control where the cell, for example, down-regulates. D-SPIN models can identify potential points of feedback control and long-range information flow in cells facilitating the study of control strategies used in cell populations to maintain homeostasis.

D-SPIN can also be used to design interventions that modulate networks and cell states. We show that D-SPIN can be applied to a model of the HSC network to construct strategies for controlling differentiation of a stem cell population (Fig. 2). D-SPIN also reveals how small molecules and small-molecule combinations can generate novel cell-states and population structures (Figs 5-7), suggesting D-SPIN could be used to study a broader set of interventions in stem cells and immune cell populations. In both cases, a major goal is to control the distribution of cell-states in a cell-population, for example, generating a specific set of progeny from stem cells or to modulate the state of the immune system in cancer or auto-immunity. D-SPIN can reveal non-intuitive connections between different cellular processes, revealing how drugs can recruit set of gene programs to generate a cell-state. This principle has been studied on a small scale for a small numbers of drugs [61], and our results suggest that such principles of super-position might hold at the transcriptome scale.

In our analysis of drug responses, we demonstrate that drug combinations can act through coherent and incoherent super-position to generate novel cell-states. In both cases, drugs act to recruit super-sets of gene-expression programs. Super-position refers to the ability of drug combinations to function additively, resulting in the union of the gene-expression programs activated by single drugs. D-SPIN allows us to dissect these combinatorial mechanisms and interpret them through drug response vectors. Drug-combination responses at the gene-program level provide a means for interpreting and predicting the effect of drug combinations. If additivity extends beyond our studied system, our model could facilitate the programming of target states in cell populations. Even though we have shown additivity alone is not able to accurately predict cell-population structure, it could still serve as a guide to narrow down potential drug-combination targets for therapeutic objectives. Additivity could arise from the modularity of gene-regulatory circuits, such that different pathways impact gene expression levels independently. Further work is needed to reveal conditions where additivity holds or breaks down.

D-SPIN has a number of limits. First, the model D-SPIN generates is an equilibrium model and so does not account for network dynamics. In future iterations of D-SPIN, we aim to convert the network into a dynamical systems model that will allow a cell to move between different steady-states. Second, D-SPIN models produce undirected rather than directed graphs, because this allows natural mapping to cells that contain feedback regulation and other forms of cyclic regulation, which are not typically allowed in directed acyclic graph modeling frameworks. Further, undirected graphical models have a powerful inference procedure which we have also extended to consider situations like active learning and perturbation design [43]. Future versions of D-SPIN might relax the undirected framework to consider directed connections between gene programs.

D-SPIN's connection with spin-network models provides some insights into the fundamental nature of cellular regulation. The spin-network model is an equilibrium model used to study physical systems at or near thermal equilibrium. Theoretically, D-SPIN models the cell population as a collection of points moving on an energy landscape that is being titled through perturbation vectors. Theoretically, energy-landscape models represent a highly simplified class of dynamical systems, because the behavior of the system can be captured within a single energy (potential) function. Maximum entropy methods have applied equilibrium spin network models to study a much broader range of systems that are far from thermal equilibrium, including neural networks and bird flocks. However, it remains unclear why equilibrium models can provide such significant predictive power for strongly non-equilibrium systems. The strategy of using an equilibrium model to produce low-error probabilistic reconstructions of the gene expression states of a cell population suggests that cells, in

certain situations, might be effectively modeled as equilibrium systems that are being driven through a series of different configurations by a biasing drive. Such models have been demonstrated for cell-fate regulation and might represent a simplifying principle [10]. The ability to model a cell as an equilibrium system that is being driven through states represents a powerful simplification that has also been studied in other chemical systems and might provide a path to more global theories of gene regulation [66, 67].

Acknowledgements

We would like to acknowledge the NIH, the Heritage Medical Research Institute, the Shurl and Kay Curci Foundation, and the Merkin Institute for Translational Research. The single-cell profiling experiments were performed at the Beckman Institute Single-cell Profiling and Engineering Center (SPEC). Sequencing was performed at the UCSF CAT, supported by UCSF PBBR, RRP IMIA, and NIH 1S10OD028511-01 grants. We acknowledge Dr. Guy Riddihough for editorial assistance with the manuscript. DAS acknowledges the support of an NSERC Discovery Grant and a Tier-II Canada Research Chair. JH acknowledges the support of NIH 5R01GM135337 grants.

References

[1] Tabula Sapiens Consortium*, Robert C Jones, Jim Karkanias, Mark A Krasnow, Angela Oliveira Pisco, Stephen R Quake, Julia Salzman, Nir Yosef, Bryan Bulthaup, Phillip Brown, et al. The tabula sapiens: A multiple-organ, single-cell transcriptomic atlas of humans. *Science*, 376(6594):eabl4896, 2022.

[2] Eric H Davidson. *The regulatory genome: gene regulatory networks in development and evolution*. Elsevier, 2010.

[3] Dennis Bray. Protein molecules as computational elements in living cells. *Nature*, 376(6538):307–312, 1995.

[4] Aviv Regev and Ehud Shapiro. The π-calculus as an abstraction for biomolecular systems. In *Modelling in Molecular Biology*, pages 219–266. Springer, 2004.

[5] Aviv Regev and Ehud Shapiro. Cellular abstractions: Cells as computation. *Nature*, 419(6905):343–343, 2002.

[6] Eric H Davidson. Emerging properties of animal gene regulatory networks. *Nature*, 468(7326):911–920, 2010.

[7] Eric H Davidson and Douglas H Erwin. Gene regulatory networks and the evolution of animal body plans. *Science*, 311(5762):796–800, 2006.

[8] Lacramioara Bintu, Nicolas E Buchler, Hernan G Garcia, Ulrich Gerland, Terence Hwa, Jané Kondev, and Rob Phillips. Transcriptional regulation by the numbers: models. *Current opinion in genetics & development*, 15(2):116–124, 2005.

[9] David A Sivak and Matt Thomson. Environmental statistics and optimal regulation. *PLoS computational biology*, 10(9):e1003826, 2014.

[10] Cameron Sokolik, Yanxia Liu, David Bauer, Jade McPherson, Michael Broeker, Graham Heimberg, Lei S Qi, David A Sivak, and Matt Thomson. Transcription factor competition allows embryonic stem cells to distinguish authentic signals from noise. *Cell systems*, 1(2):117–129, 2015.

[11] Hao Yuan Kueh, Ameya Champhekar, Stephen L Nutt, Michael B Elowitz, and Ellen V Rothenberg. Positive feedback between pu. 1 and the cell cycle controls myeloid differentiation. *Science*, 341(6146):670–673, 2013.

[12] Nir Yosef, Alex K Shalek, Jellert T Gaublomme, Hulin Jin, Youjin Lee, Amit Awasthi, Chuan Wu, Katarzyna Karwacz, Sheng Xiao, Marsela Jorgolli, et al. Dynamic regulatory network controlling th17 cell differentiation. *Nature*, 496(7446):461–468, 2013.

[13] Jacques Monod. The growth of bacterial cultures. *Annual review of microbiology*, 3(1):371–394, 1949.

[14] Arthur B Pardee, François Jacob, and Jacques Monod. The genetic control and cytoplasmic expression of "inducibility" in the synthesis of β-galactosidase by e. coli. *Journal of Molecular Biology*, 1(2):165–178, 1959.

[15] Shai S Shen-Orr, Ron Milo, Shmoolik Mangan, and Uri Alon. Network motifs in the transcriptional regulation network of escherichia coli. *Nature genetics*, 31(1):64–68, 2002.

[16] Ron Milo, Shai Shen-Orr, Shalev Itzkovitz, Nadav Kashtan, Dmitri Chklovskii, and Uri Alon. Network motifs: simple building blocks of complex networks. *Science*, 298(5594):824–827, 2002.

[17] Günter P Wagner, Mihaela Pavlicev, and James M Cheverud. The road to modularity. *Nature Reviews Genetics*, 8(12):921–931, 2007.

[18] Tong Ihn Lee, Nicola J Rinaldi, François Robert, Duncan T Odom, Ziv Bar-Joseph, Georg K Gerber, Nancy M Hannett, Christopher T Harbison, Craig M Thompson, Itamar Simon, et al. Transcriptional regulatory networks in saccharomyces cerevisiae. *science*, 298(5594):799–804, 2002.

[19] Denes Hnisz, Brian J Abraham, Tong Ihn Lee, Ashley Lau, Violaine Saint-André, Alla A Sigova, Heather A Hoke, and Richard A Young. Super-enhancers in the control of cell identity and disease. *Cell*, 155(4):934–947, 2013.

[20] William T Ireland, Suzannah M Beeler, Emanuel Flores-Bautista, Nicholas S McCarty, Tom Röschinger, Nathan M Belliveau, Michael J Sweredoski, Annie Moradian, Justin B Kinney, and Rob Phillips. Deciphering the regulatory genome of escherichia coli, one hundred promoters at a time. *Elife*, 9:e55308, 2020.

[21] Christopher T Harbison, D Benjamin Gordon, Tong Ihn Lee, Nicola J Rinaldi, Kenzie D Macisaac, Timothy W Danford, Nancy M Hannett, Jean-Bosco Tagne, David B Reynolds, Jane Yoo, et al. Transcriptional regulatory code of a eukaryotic genome. *Nature*, 431(7004):99–104, 2004.

[22] Sebastian J Maerkl and Stephen R Quake. A systems approach to measuring the binding energy landscapes of transcription factors. *Science*, 315(5809):233–237, 2007.

[23] Zhanzhi Hu, Patrick J Killion, and Vishwanath R Iyer. Genetic reconstruction of a functional transcriptional regulatory network. *Nature genetics*, 39(5):683–687, 2007.

[24] Linda S Huang and Paul W Sternberg. Genetic dissection of developmental pathways. *Methods in cell biology*, 48:97–122, 1995.

[25] Edwin L Ferguson, Paul W Sternberg, and H Robert Horvitz. A genetic pathway for the specification of the vulval cell lineages of caenorhabditis elegans. *Nature*, 326(6110):259–267, 1987.

[26] Chiou-Hwa Yuh, Hamid Bolouri, and Eric H Davidson. Genomic cis-regulatory logic: experimental and computational analysis of a sea urchin gene. *Science*, 279(5358):1896–1902, 1998.

[27] Daniel Schraivogel, Andreas R Gschwind, Jennifer H Milbank, Daniel R Leonce, Petra Jakob, Lukas Mathur, Jan O Korbel, Christoph A Merten, Lars Velten, and Lars M Steinmetz. Targeted perturb-seq enables genome-scale genetic screens in single cells. *Nature methods*, 17(6):629–635, 2020.

[28] Joseph M Replogle, Reuben A Saunders, Angela N Pogson, Jeffrey A Hussmann, Alexander Lenail, Alina Guna, Lauren Mascibroda, Eric J Wagner, Karen Adelman, Gila Lithwick-Yanai, et al. Mapping information-rich genotype-phenotype landscapes with genome-scale perturb-seq. *Cell*, 2022.

[29] Christopher S McGinnis, David M Patterson, Juliane Winkler, Daniel N Conrad, Marco Y Hein, Vasudha Srivastava, Jennifer L Hu, Lyndsay M Murrow, Jonathan S Weissman, Zena Werb, et al. Multi-seq: sample multiplexing for single-cell rna sequencing using lipid-tagged indices. *Nature methods*, 16(7):619–626, 2019.

[30] Jase Gehring, Jong Hwee Park, Sisi Chen, Matthew Thomson, and Lior Pachter. Highly multiplexed single-cell rna-seq by dna oligonucleotide tagging of cellular proteins. *Nature Biotechnology*, 38(1):35–38, 2020.

[31] Elad Schneidman, Michael J Berry, Ronen Segev, and William Bialek. Weak pairwise correlations imply strongly correlated network states in a neural population. *Nature*, 440(7087):1007–1012, 2006.

[32] William Bialek, Andrea Cavagna, Irene Giardina, Thierry Mora, Edmondo Silvestri, Massimiliano Viale, and Aleksandra M Walczak. Statistical mechanics for natural flocks of birds. *Proceedings of the National Academy of Sciences*, 109(13):4786–4791, 2012.

[33] John J Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the national academy of sciences*, 79(8):2554–2558, 1982.

[34] Assen Roguev, Sourav Bandyopadhyay, Martin Zofall, Ke Zhang, Tamas Fischer, Sean R Collins, Hongjing Qu, Michael Shales, Han-Oh Park, Jacqueline Hayles, et al. Conservation and rewiring of functional modules revealed by an epistasis map in fission yeast. *science*, 322(5900):405–410, 2008.

[35] Eran Segal, Michael Shapira, Aviv Regev, Dana Pe'er, David Botstein, Daphne Koller, and Nir Friedman. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nature genetics*, 34(2):166–176, 2003.

[36] Xiaoqiao Chen, Sisi Chen, and Matt Thomson. Minimal gene set discovery in single-cell mrna-seq datasets with activesvm. *Nature Computational Science*, 2(6):387–398, 2022.

[37] Graham Heimberg, Rajat Bhatnagar, Hana El-Samad, and Matt Thomson. Low dimensionality in gene expression data enables the accurate extraction of transcriptional programs from shallow sequencing. *Cell systems*, 2(4):239–250, 2016.

[38] Sisi Chen, Paul Rivaud, Jong H Park, Tiffany Tsou, Emeric Charles, John R Haliburton, Flavia Pichiorri, and Matt Thomson. Dissecting heterogeneous cell populations across drug and disease conditions with popalign. *Proceedings of the National Academy of Sciences*, 117(46):28784–28794, 2020.

[39] Grace Hui Ting Yeo, Sachit D Saksena, and David K Gifford. Generative modeling of single-cell time series with prescient enables prediction of cell trajectories with interventions. *Nature communications*, 12(1):1–12, 2021.

[40] Caleb Weinreb, Samuel Wolock, Betsabeh K Tusi, Merav Socolovsky, and Allon M Klein. Fundamental limits on dynamic inference from single-cell snapshots. *Proceedings of the National Academy of Sciences*, 115(10):E2467–E2476, 2018.

[41] Aditya Pratapa, Amogh P Jalihal, Jeffrey N Law, Aditya Bharadwaj, and TM Murali. Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. *Nature methods*, 17(2):147–154, 2020.

[42] Jan Krumsiek, Carsten Marr, Timm Schroeder, and Fabian J Theis. Hierarchical differentiation of myeloid progenitors is encoded in the transcription factor network. *PloS one*, 6(8):e22649, 2011.

[43] Jialong Jiang, David A Sivak, and Matt Thomson. Active learning of spin network models. *arXiv preprint arXiv:1903.10474*, 2019.

[44] Nora Heisterkamp, John R Stephenson, John Groffen, Pamela F Hansen, Annelies de Klein, Claus R Bartram, and Gerard Grosveld. Localization of the c-abl oncogene adjacent to a translocation break point in chronic myelocytic leukaemia. *Nature*, 306(5940):239–242, 1983.

[45] CB LOZZIO. Human chronic myelogenous leukemia cell-line with positive philadelphia chromosome. *Blood*, 45:321–334, 1975.

[46] Brad T Sherman, Ming Hao, Ju Qiu, Xiaoli Jiao, Michael W Baseler, H Clifford Lane, Tomozumi Imamichi, and Weizhong Chang. David: a web server for functional enrichment analysis and functional annotation of gene lists (2021 update). *Nucleic Acids Res*, 10, 2022.

[47] Maxim V Kuleshov, Matthew R Jones, Andrew D Rouillard, Nicolas F Fernandez, Qiaonan Duan, Zichen Wang, Simon Koplev, Sherry L Jenkins, Kathleen M Jagodnik, Alexander Lachmann, et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic acids research*, 44(W1):W90–W97, 2016.

[48] Vincent A Traag, Ludo Waltman, and Nees Jan Van Eck. From louvain to leiden: guaranteeing well-connected communities. *Scientific reports*, 9(1):5233, 2019.

[49] Hyo-Ji Lee, Yunseo Woo, Tae-Wook Hahn, Young Mee Jung, and Yu-Jin Jung. Formation and maturation of the phagosome: a key mechanism in innate immunity against intracellular bacterial infection. *Microorganisms*, 8(9):1298, 2020.

[50] Madalina Giurgiu, Julian Reinhard, Barbara Brauner, Irmtraud Dunger-Kaltenbach, Gisela Fobo, Goar Frishman, Corinna Montrone, and Andreas Ruepp. Corum: the comprehensive resource of mammalian protein complexes—2019. *Nucleic acids research*, 47(D1):D559–D563, 2019.

[51] Maria Rosaria Amoroso, Danilo Swann Matassa, Gabriella Laudiero, Anastasia V Egorova, Roman S Polishchuk, Francesca Maddalena, Annamaria Piscazzi, Simona Paladino, Daniela Sarnataro, Corrado Garbi, et al. Trap1 and the proteasome regulatory particle tbp7/rpt3 interact in the endoplasmic reticulum and control cellular ubiquitination of specific mitochondrial proteins. *Cell Death & Differentiation*, 19(4):592–604, 2012.

[52] Melanie de Almeida, Matthias Hinterndorfer, Hanna Brunner, Irina Grishkovskaya, Kashish Singh, Alexander Schleiffer, Julian Jude, Sumit Deswal, Robert Kalis, Milica Vunjak, et al. Akirin2 controls the nuclear import of proteasomes in vertebrates. *Nature*, 599(7885):491–496, 2021.

[53] Chiara Milanese, Cíntia R Bombardieri, Sara Sepe, Sander Barnhoorn, César Payán-Goméz, Donatella Caruso, Matteo Audano, Silvia Pedretti, Wilbert P Vermeij, Renata MC Brandt, et al. Dna damage and transcription stress cause atp-mediated redesign of metabolism and potentiation of anti-oxidant buffering. *Nature communications*, 10(1):4887, 2019.

[54] Yvan Jamilloux, Thomas El Jammal, Lucine Vuitton, Mathieu Gerfaud-Valentin, Sébastien Kerever, and Pascal Sève. Jak inhibitors for the treatment of autoimmune and inflammatory diseases. *Autoimmunity reviews*, 18(11):102390, 2019.

[55] David C Fajgenbaum and Carl H June. Cytokine storm. *New England Journal of Medicine*, 383(23):2255–2273, 2020.

[56] Christopher S McGinnis, David M Patterson, Juliane Winkler, Daniel N Conrad, Marco Y Hein, Vasudha Srivastava, Jennifer L Hu, Lyndsay M Murrow, Jonathan S Weissman, Zena Werb, Eric D Chow, and Zev J Gartner. MULTI-seq: sample multiplexing for single-cell RNA sequencing using lipid-tagged indices. *Nat. Methods*, 16(7):619–626, July 2019.

[57] C Domínguez Conde, C Xu, LB Jarvis, DB Rainbow, SB Wells, T Gomes, SK Howlett, O Suchanek, K Polanski, HW King, et al. Cross-tissue immune cell analysis reveals tissue-specific features in humans. *Science*, 376(6594):eabl5197, 2022.

[58] Ctibor Skuta, Martin Popr, Tomas Muller, Jindrich Jindrich, Michal Kahle, David Sedlak, Daniel Svozil, and Petr Bartunek. Probes & drugs portal: an interactive, open data resource for chemical biology. *Nature methods*, 14(8):759–760, 2017.

[59] Thibaut Desgeorges, Giorgio Caratti, Rémi Mounier, Jan Tuckermann, and Bénédicte Chazaud. Glucocorticoids shape macrophage phenotype for tissue repair. *Frontiers in immunology*, 10:1591, 2019.

[60] Yan-Cun Liu, Xian-Biao Zou, Yan-Fen Chai, and Yong-Ming Yao. Macrophage polarization in inflammatory diseases. *International journal of biological sciences*, 10(5):520, 2014.

[61] Naama Geva-Zatorsky, Erez Dekel, Ariel A Cohen, Tamar Danon, Lydia Cohen, and Uri Alon. Protein dynamics in drug combinations: a linear superposition of individual-drug responses. *Cell*, 140(5):643–651, 2010.

[62] JA Fishman. Infection in organ transplantation. *American Journal of Transplantation*, 17(4):856–879, 2017.

[63] Jay A Fishman and Robert H Rubin. Infection in organ-transplant recipients. *New England Journal of Medicine*, 338(24):1741–1751, 1998.

[64] Marcus R Pereira, Sumit Mohan, David J Cohen, Syed A Husain, Geoffrey K Dube, Lloyd E Ratner, Selim Arcasoy, Meghan M Aversa, Luke J Benvenuto, Darshana M Dadhania, et al. Covid-19 in solid organ transplant recipients: initial report from the us epicenter. *American journal of transplantation*, 20(7):1800–1808, 2020.

[65] Edwin T Jaynes. Information theory and statistical mechanics. *Physical review*, 106(4):620, 1957.

[66] Lacramioara Bintu, Nicolas E Buchler, Hernan G Garcia, Ulrich Gerland, Terence Hwa, Jane Kondev, Thomas Kuhlman, and Rob Phillips. Transcriptional regulation by the numbers: applications. *Current opinion in genetics & development*, 15(2):125–135, 2005.

[67] Jeremy Gunawardena. A linear framework for time-scale separation in nonlinear biochemical systems. *PloS one*, 7(5):e36321, 2012.