Collective Gene Expression Fluctuations Encode the Regulatory State of Cells

Fabrizio Olmeda[1,2,3,*], Yiteng Dang[1,4,6], Fabian Rost[1,4,5,6], Vincenzo Maria Schimmenti[1], Ivan Di Terlizzi[1,2], and Steffen Rulands[1,2,†]

[1]Max Planck Institute for the Physics of Complex Systems, Noethnitzer Str. 38, 01187 Dresden, Germany
[2]Ludwigs-Maximilians-Universität München, Arnold Sommerfeld Center for Theoretical Physics, Theresienstr. 37, 80333 München, Germany
[3]Institute of Science and Technology Austria, Klosterneuburg, Austria
[4]Max Planck Institute for Cell Biology and Genetics, Pfotenhauer Str., Dresden, Germany
[5]Dresden-concept Genome Center (DcGC), Center for Molecular and Cellular Bioengineering (CMCB), Technische Universität Dresden, 01062 Dresden, Germany
[6]Center for Systems Biology, Pfotenhauer Str., Dresden, Germany
[*]fabrizio.olmeda@ista.ac.at
[†]rulands@lmu.de

Abstract

Gene expression is inherently stochastic, but how fluctuations propagate at the macroscopic scale of gene networks and whether they play a functional role remains unclear. Here, we develop a theoretical and statistical learning framework that shows how collective gene expression fluctuations encode the regulatory state of cells. Specifically, we map fluctuations in gene expression to a disordered, non-equilibrium system and derive the macroscopic phase behaviour of these fluctuations. Applying our theory to single-cell RNA sequencing data, we show that cells operate close to a dynamical transition separating strongly correlated fluctuations, reminiscent of glassy dynamics, from weakly correlated, disordered regimes. Drawing on these experiments, we adapt Restricted Boltzmann Machines to infer cell states by the structure of their collective gene expression fluctuations. This shows that strongly correlated fluctuations characterize cells with the ability to differentiate into more specialised cell types. Our results point to the emergence of collective fluctuations as a layer of regulation in cells and as a hallmark of high regulatory flexibility. They also demonstrate how physical principles can yield functional insights into cell states from widely used single-cell data.

The self-organisation of cells into complex tissues relies on a precise definition of cell states, such that the right kinds of cells are produced at a given position and time [1, 2]. In most cases, cell states progress through a hierarchy of distinct states with decreasing levels of potency, which is the capacity to differentiate into multiple specialized cell types, ranging from stem and progenitor cells to differentiated cells. On the molecular scale, these cells states are controlled by complex biochemical networks. As a precursor of protein production, the expression of genes and their interaction in gene regulatory networks (GRNs) is commonly seen as the primary determinant of cell states.

The expression of genes is intrinsically highly stochastic. Previous studies have highlighted the role of noise in single genes or minimal circuits unvealing transcriptional bursts and switching

dynamics [3, 4, 5, 6, 7, 8]. By contrast, large-scale models of gene regulatory networks often neglect stochasticity, relying on Boolean or coarse-grained representations [9, 10]. Recent advances in single-cell RNA sequencing (scRNA-seq) enable measurements of the expression of thousands of genes in single cells [11, 12], and thereby variability of gene expression across genes and cells. Despite their widespread use, we still lack a clear understanding of how fluctuations propagate across large-scale gene regulatory networks and how such data can inform physical readouts of cell states, though some progress has been made in this direction [13]. The reason for this is that noise in these experiments is heavily influenced by technical factors, which require robust theoretical frameworks for their interpretation.

Here, we show that the macroscopic structure of fluctuations in gene networks encodes functional information of cell states. Specifically, drawing on methods from the physics of disordered systems, we develop a theoretical framework that describes how fluctuations propagate through GRNs at the macroscopic scale (Fig.1a). This theory predicts the existence of phase transitions between a glassy regime, where fluctuations are strongly correlated, and a paramagnetic phase, where fluctuations are largely independent between genes. By using large-scale single-cell RNA sequencing experiments, we show that human cells are poised in the vicinity of these phase transitions, with a tendency of stem and progenitor cells to have more strongly correlated fluctuations than differentiated cells. The architecture of a restricted Boltzmann machine (RBM) emerges naturally from our theoretical framework. We train RBMs on single-cell gene expression data from different human cell types and predict how macroscopic quantities are associated with the relative position of cells in the phase diagram.

Results

Understanding whether macroscopic fluctuations in gene expression reflect underlying regulatory states is key to linking cellular variability with function. To explore this connection, we began by examining covariance matrices from single-cell RNA sequencing data [14], which quantify how gene expression levels co-fluctuate across cells (Supplementary section VII.1). These matrices reveal strong correlations in stem cells, which weaken along differentiation (Figure 1b). Motivated by these observations, we developed a theoretical framework that connects fluctuation patterns to a disordered, non-equilibrium model of gene regulation, enabling a quantitative interpretation of macroscopic variability in terms of collective dynamics.

Collective behaviour of gene expression fluctuations

As a first step, we define a stochastic description of how fluctuations propagate through gene networks. To this end, we consider a set of N genes, each of which is associated with m_i mRNA molecules and n_i proteins. Both proteins and mRNAs degrade with rates γ_i and d_i, respectively. mRNA molecules are translated to proteins at a rate g_i. Genes interact via the expression of transcription factors, which are proteins that, often cooperatively, regulate the expression of other genes. Hence, the transcription rate of a gene i depends on the protein abundance of other genes j through a sigmoidal function $f_{ji}(n_j)$ [15], Fig. 2a. The time evolution of the probability $P(\mathbf{n}, \mathbf{m}, t)$ of finding protein and mRNA abundances $\mathbf{n} = (n_1, \ldots, m_N)$ and $\mathbf{m} = (m_1, \ldots m_N)$, respectively,

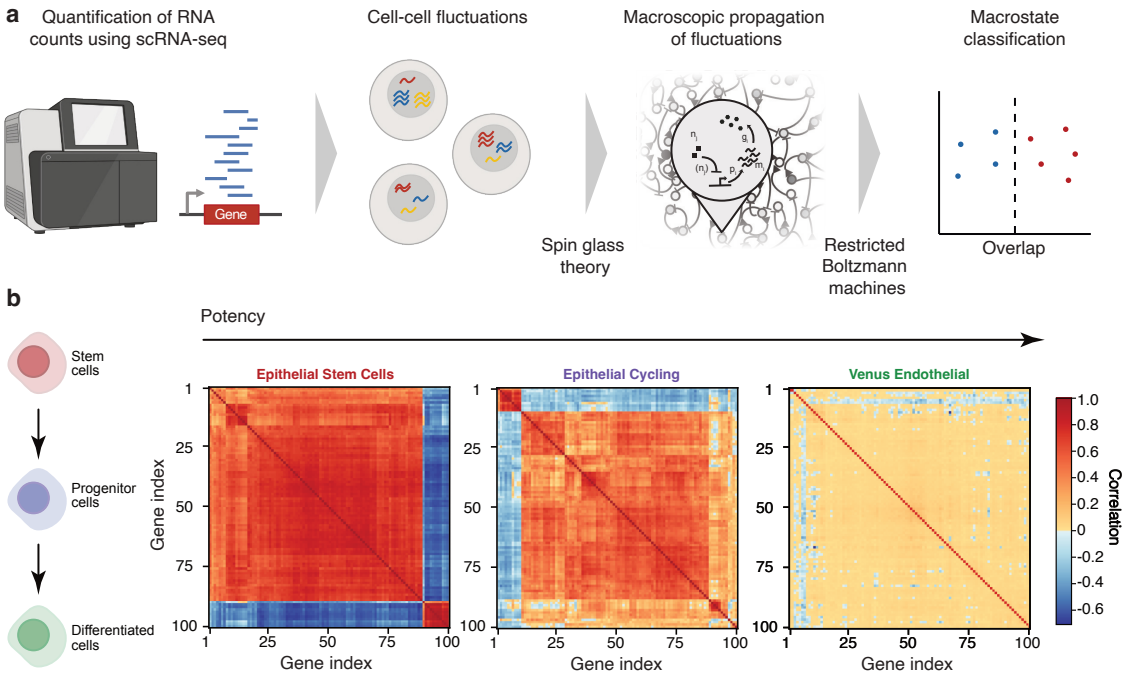

Figure 1: (**a**) Graphical outline of the approach used in this manuscript. mRNA at single-cell resolution is measured in RNA sequencing experiments from which we extract gene expression fluctuations. A theory for macroscopic fluctuations of gene expression allows us, using restricted Boltzmann machines, to classify cells based on the macroscopic properties of these fluctuations. (**b**) Left: Simplified schematic of a lineage hierarchy. Right: Typical covariance matrices of gene expression fluctuations obtained from a sequencing experiment of representative cell types of different stages of the lineage hierarchy. The structure of the covariance matrices reveals different structures and patterns which emerges with increasing cell potency (from left to right: stem cells, progenitor cells, and differentiated cells from the gut [14]). Details are given in Supplementary section VII.1.

is then given by a master equation,

$$\frac{\partial P}{\partial t} = \sum_{i=1}^{N} \left[d_i \left(\mathcal{E}_i^1 - 1\right) n_i + \left(\mathcal{F}_i^1 - 1\right) \gamma_i m_i + \Omega \left(\mathcal{E}_i^{-1} - 1\right) g_i m_i + \Omega \left(\mathcal{F}_i^{-1} - 1\right) \sum_{j \in e(i)} f_{ji}^e \left(\frac{n_j}{\Omega}\right) \right] P \tag{1}$$

where $j \in e(i)$ signifies that the sum goes over all genes j that interact with gene i and we used the notation of step operators $\mathcal{E}_i^{\pm} G(\mathbf{n}, \mathbf{m}) = G(\{\mathbf{n}, n_i \pm 1\}, \mathbf{m})$, with G an arbitrary function of protein and mRNA, and similarly for $\mathcal{F}_i^{\pm}$ which acts on the mRNA dependent part.

While the absolute values of mRNA and protein abundance are highly specific to a given type of cell, we reasoned that how *fluctuations* around these absolute values behave macroscopically might give a robust read-out of the cell state. To derive a description in terms of fluctuations alone, we expanded Eq. (1) around steady-state values of the average mRNA and protein abundances (Supplementary section I), capturing the dynamics of collective fluctuations of gene products (Fig. 2b). This expansion yields a Fokker–Planck equation for the joint probability distribution of fluctuations, whose stationary solution resembles a Boltzmann distribution with an effective "Hamiltonian",

$$\mathcal{H} = \sum_{i=1}^{N} V^{\xi}(\xi_i) + \sum_{i=1}^{N} V^{\eta}(\eta_i) + \sum_{i=1}^{N} K_i^{\xi\eta} \xi_i \eta_i + \sum_{ij} J_{ij} \xi_i \eta_j \,, \tag{2}$$

Here, the first two terms describe intrinsic mRNA and protein fluctuations, and they are proportional to ξ_i^2 and η_i^2, respectively. The third term gives the contribution of mRNA translation, and

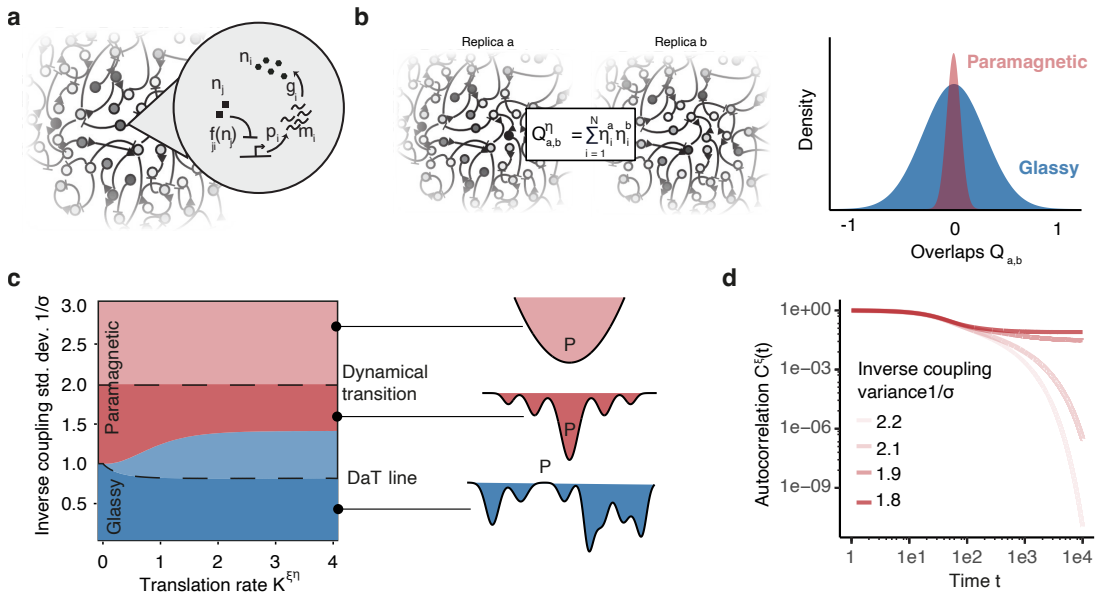

Figure 2: (**a**) Graphical representation of the different regulatory interactions in Eq. (1). (**b**) Left: Two different replicas for GRNs and possible distributions of the overlaps between pairs of cells sharing the same type $Q_{a,b}$. Right: The width of probability density functions of overlaps reflects the phase behaviour of fluctuations. (**c**) Phase boundary between a paramagnetic like phase (P) and a sping-glass phase (SG) obtained theoretically in Supplementary section III and V.The upper black dashed line denotes the dynamical phase transition below which time autocorrelation functions of protein and mRNA fluctuations have a non-vanishing plateau, Eq. (3). The lower black dashed line denote the De Almeida Thouless (DAT) line of the instability of the replica symmetric solution for the spherical spin glass. In the gray region the replica symmetric solution is stable. The light-blue region denotes the DAT line for the Sherrington Kirkpatrick model where the fluctuations are binary rather than spherically constrained. On the right hand side we show how the free-energy landscapes changes while increasing the σ. The paramagnetic state passes from being the unique minima of the free energy to a maxima in the SG-phase after new metastable states are formed. (**d**) Autocorrelation functions of mRNA fluctuations above and below the dynamical transition. A plateau emerges as the system approaches the critical value $\sigma = 1/2$, reflecting long-term correlations in time.

the fourth term quantifies fluctuations arising from interactions between genes. The parameters in Eq. (2) are explicitly defined in Supplementary Section. II.

In the following, we take the couplings J_{ij} to be Gaussian distributed with vanishing mean and standard deviation $\sigma/\sqrt{N}$, in accordance with previous studies [16]. The "Hamiltonian" in (2) is identical to the one of a bipartite, soft spin-glass with asymmetric couplings of the form J_{ij} [17, 18]. While models of this kind have already been explored in the literature [19, 20], we extend their treatment to analyse the fluctuation phases of GRNs. To this end, we first define a parameter that quantifies the degree to which gene expression fluctuations are predetermined by the interactions between genes. This parameter is given by the average correlation of fluctuations between different realisations (replicas) a and b, $Q^{\xi}_{ab} = N^{-1}\sum_{i=1}^{N}\langle \xi_i^a \xi_i^b \rangle$ and $Q^{\eta}_{ab} = N^{-1}\sum_{i=1}^{N}\langle \eta_i^a \eta_i^b \rangle$, termed overlaps. Biologically, the overlaps measure how consistent gene expression fluctuations are between cells that rest in the same steady state. With this, we calculated an effective free energy that only depends on the overlaps (Supplementary Section. II).

Minimization of the free energy reveals the GRN's fluctuation phases. When the overlaps vanish, fluctuations are uncorrelated (paramagnetic regime). The emergence of gene-gene correlation instead indicates a glass-like phase with multiple metastable states and complex dynamical properties, like long-lived time correlation functions. As σ increases, we find a transition from the paramagnetic to the glassy phase, marked by the emergence of non-trivial overlap distributions and slow collective dynamics. This framework thus enables us to directly link gene expression variability to the underlying dynamical phases of the GRN. Figure 2c shows the phase diagram derived from Eq. (2) (Supplementary sections II, III).

Because these phase transitions are dynamical in nature, our theory not only captures the structure of GRN-level fluctuations, summarized by the overlap distribution and accessible through single-cell sequencing data, but also enables analytical predictions of dynamical properties such as autocorrelation functions. These provide a complementary information to the static data available in scRNA-seq experiments. To explore these dynamical aspects, we calculated the autocorrelation functions of mRNA and protein fluctuations [21, 22] (Supplementary sections IV, V). In the paramagnetic regime, autocorrelation functions decay exponentially, with a correlation time governed by the kinetic parameters of the system. However, the system undergoes a dynamical transition, marked by the emergence of a persistent plateau in the autocorrelation functions, once the scaled standard deviation of the effective interactions exceeds a critical threshold given by

$$\sqrt{N\,\mathbb{E}\left[\left(J_{ij}-\mathbb{E}\left[J_{ij}\right]\right)^2\right]} \equiv \sigma > \frac{1}{2}\,. \tag{3}$$

This plateau signals the emergence of long-lived fluctuations and metastable states, even in regimes that are paramagnetic. Since this behavior does not appear in the spherical 2-spin glass with a unipartite interaction structure [23], it can be attributed to the underlying bipartite architecture. Therefore, we predict that even in a region of the paramagnetic phase, the multi-step process of protein production with mRNA as an intermediary product may lead to long-lived correlations in fluctuations. In Fig. 2d we perform numerical simulation of the dynamical equation for autocorrelation functions, (Eq. (S.69) in Supplementary section VI) and show their scaling for different values of σ as well as the emergence of a plateau near the dynamical transition. Together, these results indicate that gene–gene interactions play a central role in sustaining memory of expression states beyond the intrinsic timescales set by mRNA and protein degradation.

Empirical validation of correlated fluctuations

To test our theoretical predictions empirically, we used single-cell RNA sequencing data. These experiments provide genome-wide measurements of mRNA abundance in thousands of individual cells, providing a high-dimensional view of gene expression variability. However, testing theoretical predictions against such data is challenging due to substantial technical noise and biases [24, 25]. To address this, we employed two complementary strategies designed to be robust to technical artifacts. First, we tested for long-range correlations in gene expression fluctuations by quantifying the scaling of the standard deviation of the overlap distribution with the number of genes for several cell types. Secondly, we exploited the formal similarity between the Hamiltonian in Eq. (2) and the energy function of a restricted Boltzmann machine (RBM) [26] to infer the covariance structure of gene expression fluctuations.

To begin, we reasoned that the variance of the overlap distribution of mRNA fluctuations, $P(Q^{\eta}_{ab})$, is larger in the glassy phase compared to the paramagnetic phase. Because the scale of the overlap distribution is influenced by technical parameters, it cannot unambiguously inform about the phase behaviour of gene expression fluctuations. To test our predictions, we therefore defined statistical observables that are independent of such technical factors. In general, because glassy behaviour is defined by correlations between replicas, glassy fluctuations in gene expression networks are reflected in a broad distribution of single-gene overlaps, $q^i_{ab} = \xi^a_i \xi^b_i$. If mRNA and protein fluctuations are uncorrelated with similar mean and variance, the central limit theorem implies that the standard deviation of the distribution of the overlaps, $Q^{\xi}_{ab} = N^{-1}\sum_{i=1}^{N} q^i_{ab}$ scales with the number of genes as $N^{-1/2}$. However, if either of the above conditions is violated, the central limit theorem no longer holds, and we expect the standard deviation of Q^{ξ}_{ab} to deviate significantly from the $N^{-1/2}$ scaling form (Fig. 3a). Violations of the central limit theorem for the

overlap distribution do not necessarily imply glassy properties, but they reveal strong correlations between gene-product fluctuations. Within the framework of our theory, these correlations can only be due to the structure of the interaction matrix and therefore imply glassy behaviour. To test for strongly correlated mRNA fluctuations, we analysed a single-cell RNA sequencing data set of cells from the developing and adult mouse brain [27]. For cell states that are longer lived than typical mRNA and protein lifetimes, we identify different steady states of the GRN by using clustering algorithms (Fig. 3b) and computed the fluctuations of each gene in a cell by taking the difference between the expression level and the average in the respective cluster (Supplementary section VI).

We found that the standard deviation of the overlap distribution for several cell types decreases slower than predicted by the central limit theorem. This is particularly true for progenitor cell types, which reside at the apex of the lineage hierarchy (Fig. 3c and Fig. S1). To rule out that this observation arises from confounding factors that do not directly reflect the collective interactions between genes, we performed two orthogonal analyses. Firstly, we took successively smaller random subsets of cells by defining local neighbourhoods in gene expression space as steady states and found our results to be robust with respect to the number of cells in the subset (Fig. 3d, Fig. S2 and Supplementary section VI). Hence, the deviation from the $N^{-1/2}$ scaling is not a consequence of a too broad definition of the steady states, nor do they merely reflect differences in cluster sizes. To exclude that our findings reflect generic features of technical noise in such experiments, we simulated random, uncorrelated gene expression in a way that realistically mimics both biological and technical noise of gene expression, Supplementary section VI. We found that these simulations strictly follow the scaling predicted by the central limit theorem, which implies that the observation of strong correlations in gene expression fluctuation does not reflect technical noise in single-cell RNA sequencing data (Fig. S3). Hence, differences between stem, progenitor, and differentiated cells lie not only in their absolute gene expression levels, but also in how fluctuations are collectively organized and correlated between genes.

Restricted Boltzmann machines capture macroscopic fluctuations

To further investigate the macroscopic structure of gene expression fluctuations and to quantitatively estimate values of σ across different cell types, we adopted a physics-informed machine learning approach. Specifically, we leveraged the formal analogy between our theoretical framework and restricted Boltzmann machines (RBMs) [26]. RBMs have deep connections to the theory of disordered systems [28] and have also been applied in biological contexts [29, 30]. Formally, a restricted Boltzmann machine (RBM) is an energy-based probabilistic model that defines a joint distribution over observed variables, $\boldsymbol{v}$, and latent variables, $\boldsymbol{h}$, through a bipartite interaction graph (Fig. 4a). The probability distribution takes the Boltzmann form $P(\boldsymbol{v},\boldsymbol{h}) = Z^{-1}\exp(-E(\boldsymbol{v},\boldsymbol{h}))$, where the energy is given by $E(\boldsymbol{v},\boldsymbol{h}) = \sum_i V^i(v_i) + \sum_j U^j(h_j) + \sum_{ij} W_{ij} v_i h_j$, with V and U representing self-energies and W_{ij} the weights connecting visible and hidden units.

If the variables are continuous, the self-energies are quadratic and the dimensionalities of the visible and hidden layers are identical, the RBM energy function is formally equivalent to the Hamiltonian in Eq. (2). The observed variables correspond to mRNA fluctuations, $\boldsymbol{v} = \boldsymbol{\eta}$, and the latent variables correspond to protein fluctuations, $\boldsymbol{h} = \boldsymbol{\xi}$. In this equivalence, the RBM weights W_{ij} correspondto the interaction matrix J_{ij} in Eq.(S.5), up to diagonal terms, which scale with N and have negligible impact on the inferred statistics compared to the N^2 off-diagonal couplings. This equivalence enables us to estimate the standard deviation of gene–gene interactions σ from Eq. (3) by computing $\widetilde{\sigma} = \sqrt{N\,\mathbb{E}\left[(W_{ij} - \mathbb{E}\left[W_{ij}\right])^2\right]}$, where N is the number of genes considered and W_{ij} are the inferred RBM couplings. This allows us to place different cell types within the

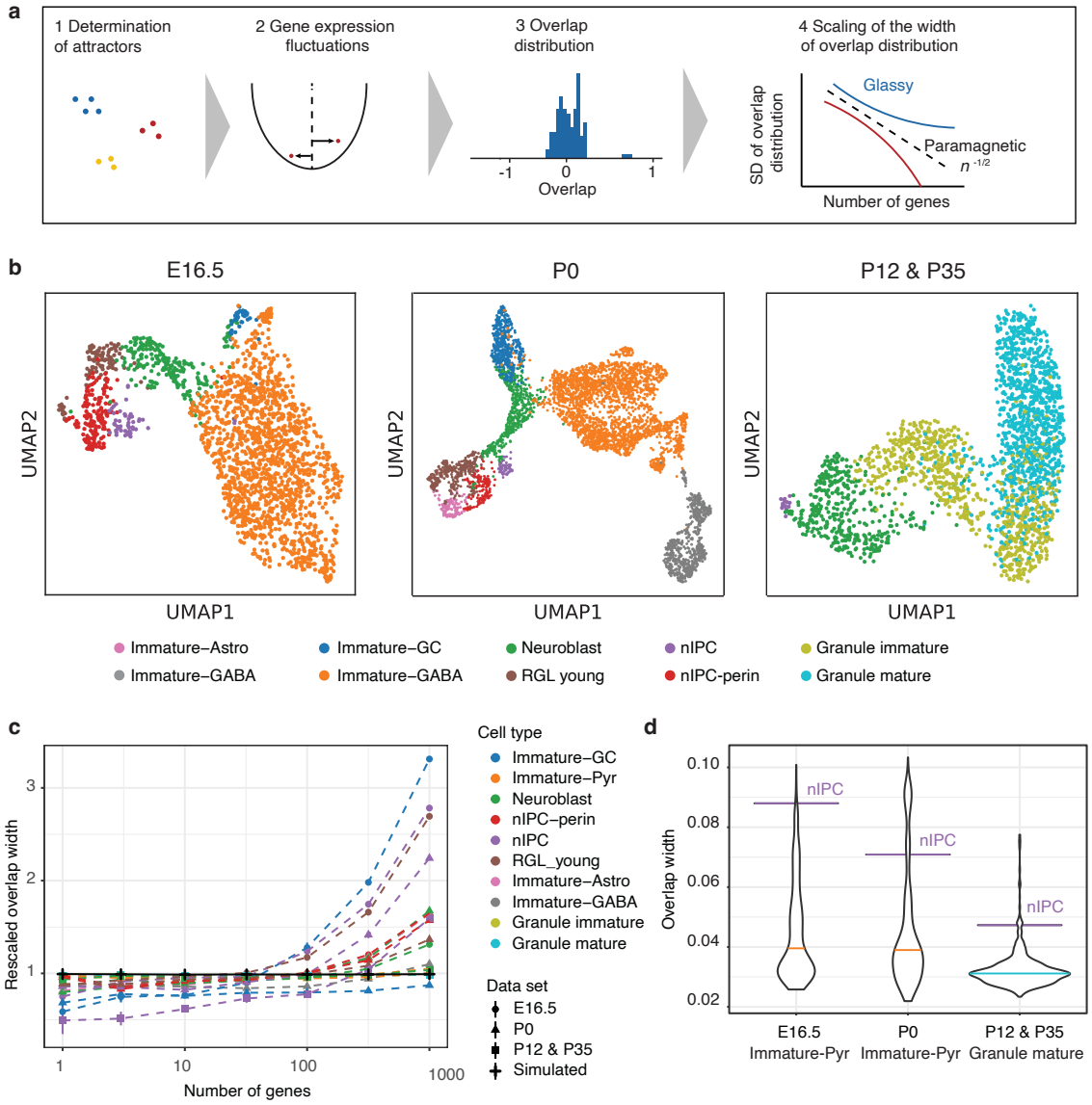

Figure 3: (**a**) Schematic of the computational pipeline used to detect spin-glass behaviour in single-cell RNA-seq data. We start with clustered and annotated data from sequenced single cells from the mouse brain [27]. After filtering out irrelevant clusters, we computed the gene fluctuations for each cell and each gene (VI). From this we computed the overlap distribution for each cluster. Finally, we subsampled genes to obtain how the overlap width scales with the number of genes. The black line shows the mean overlap width over different subsamples (blue circles). The dashed blue line shows the $1/\sqrt{N}$ scaling obtain from the central limit theorem. (**b**) Uniform Manifold Approximation and Projection (UMAP) plot of the data at three different developmental times, with colours indicating different cell clusters. Each dot represents a cell. (**c**) The rescaled overlap width, obtained by dividing the width of the overlap distr1ibution by $1/\sqrt{N}$, as a function of the number of genes N. The simulated uncorrelated data set (black line) shows a flat line, indicating that the overlap width scales as $1/\sqrt{N}$ as expected. The different clusters of the single-cell data shown in (b) show different deviations from this $1/\sqrt{N}$ scaling. (**d**) To test whether the difference in the scaling of the overlap width stems from the different number of cells contained in different clusters, we compared the overlap width of the nIPC cluster (which shows large deviations from $1/\sqrt{N}$ scaling, as seen in (**c**)) to subsamples of cells from the largest clusters. For each of the three data sets, we took one of the largest clusters for which the overlap width is close to $1/\sqrt{N}$ scaling. We then subsampled cells from the large clusters to match the number of cells of the nIPC clusters (E16.5: 51 cells, P0: 60 cells, P12 & P35: 16 cells). The violin plots show the distributions of overlap widths obtained from 1000 independent subsamples of the large clusters. Horizontal line indicates the median.

phase diagram predicted by the theory.

The physical interpretation of sequencing experiments is complicated by the fact that they provide non-stoichiometric readouts of mRNA counts. To ensure that our inference is robust to the

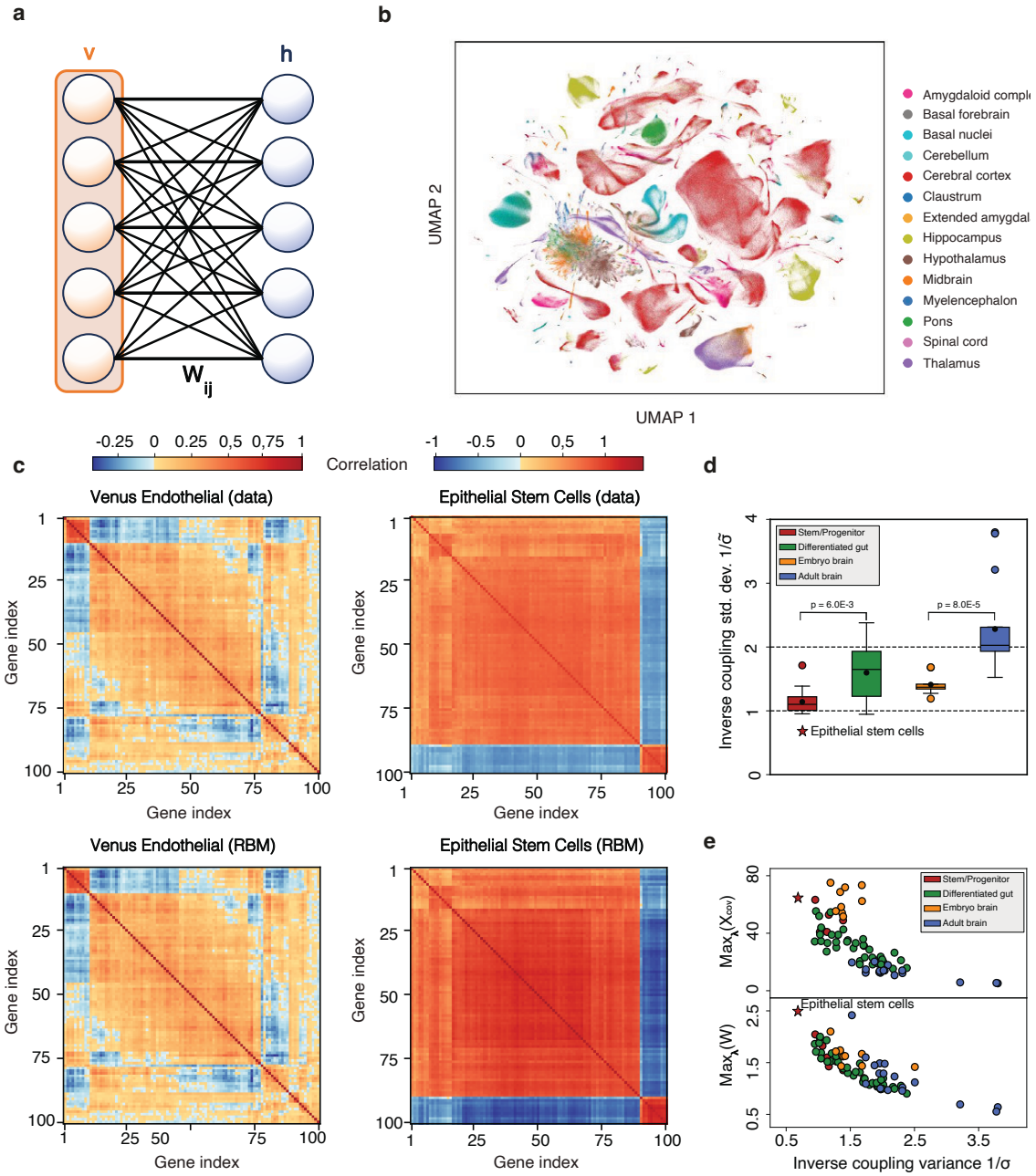

Figure 4: (**a**) Graphical representation of our Boltzmann machine, where the hidden layer encodes the proteome, while the visible layer are mRNA expression values for individual genes.(**b**) Clustering (colors) shown on UMAP dimensional reduction of neural cells from Ref. [31]. Each dot represents one cell. (**c**) Covariance matrices obtained from the data compared to the one fitted from the Rectricted Boltzmann machine. (**d**) Inverse coupling standard deviation inferred from the Restricted Boltzmann machine for several stem and differentiated cells clusters in adult and embrionic human tissues. P-values were calculated using the Mann-Whitney U test. (**e**) Maximum eigenvalue from the eigenvalue distribution of the covariance and the interaction matrices obtained from the RBM for stem and differentiated cells clusters as in (**d**). Interaction matrices, covariance matrices and eigenvalue distribution for other cell types are shown in Figs. S4-S7.

arbitrary scaling inherent in scRNA-seq experiments, we preprocessed all data by standardizing gene expression, such that each gene has unit variance across cells. While this normalization removes information about the absolute scale of gene expression, it enables meaningful comparisons across samples. Although this normalization differs slightly from the spin constraint in the spherical spin

model, Eq. (2), we expect the estimated numerical values of $\widetilde{\sigma}$ to be approximately correct.

To estimate values of $\widetilde{\sigma}$ for different cell states, we trained the RBM model on three distinct single-cell RNA-seq datasets, encompassing approximately 10^6 cells after filtering. The first dataset, the Pan-GI Cell Atlas [14], includes stem, progenitor, and differentiated cell types profiled from the gastro-intestinal tract in human. This enables a controlled comparison of gene-gene co-fluctuation patterns within the same experimental context. We also analysed two further datasets, one from human embryonic development [32] and from adult brain tissue [31]. A dimensionality-reduced visualization of the latter dataset is shown in Fig. 4b. We obtained putative steady states of the GRN by using clustering and calculated fluctuations by subtracting gene expression values by the mean in each cluster. For each cluster, we trained a Gaussian-Gaussian RBM on a selected set of 100 genes, as described in Supplementary Section VII. The model accurately reproduces the empirical gene-gene covariance structure, capturing the degree of correlation across cells in the same state(Fig. 4c and Fig. S4-S7). Stem and progenitor cells display strongly structured off-diagonal covariances in gene expression fluctuations, while more differentiated states exhibit weaker and more localized correlations.

The estimated values of $\widetilde{\sigma}$ are of order one for all clusters,suggesting that the analysed cell types operate in a regime close to the phase transitions predicted by the theory (Fig. 2d). This raises the question of whether these phase transitions play a biological role in the regulation of cell identity. To test whether there is a tendency for $\widetilde{\sigma}$ to be higher in cells with lower potency compared to more potent cell types, as suggested by the statistical analysis in Fig. 3, we grouped cell clusters into stem, progenitor, and differentiated categories according to the annotations in the original publications. To minimize technical confounders, we first compared cell types within the Pan-GI Cell Atlas. Fig. 4d shows that $1/\widetilde{\sigma}$ is significantly lower in stem and progenitor cells compared to differentiated ones (Mann-Whitney U test, $p = 6 \cdot 10^{-3}$). We then extended the analysis to embryonic and adult brain datasets, again observing a significant difference ($p = 8 \cdot 10^{-5}$), consistent with a reduction in correlated fluctuations as cells progress towards a differentiated cell state.

Finally, we investigated the link between the value of $\widetilde{\sigma}$ and the structure of correlations in gene expression fluctuations by analysing the largest eigenvalues of the covariance matrix of gene expression fluctuations. In random matrices, eigenvalues follow probability distributions. Eigenvalues which are isolated from the bulk of the distribution indicate correlations in the random matrix [33, 34], which is also used in the training and interpretation of RBMs [35]. As shown in Fig.4e, the leading eigenvalue of the covariance matrix of gene expression fluctuations decreases with $1/\widetilde{\sigma}$, reflecting stronger and more structured fluctuations in highly potent cells. At the same time, the spectrum of W_{ij} follows a Wigner semicircle distribution, consistent with the hypothesis of uncorrelated couplings, and its edge scales with $\widetilde{\sigma}$ (Fig.4e), as expected. Together, these findings establish a link between variability of interactions between genes, long-range correlations of gene expression fluctuations, and cellular potency.

Discussion

We combined methods from the statistical physics of disordered systems with tools from statistical learning and bioinformatics to show that cells exhibit macroscopic fluctuations in gene expression, which are associated with their potency. By mapping these fluctuations onto disordered systems governed by asymmetric, random interactions, we found that GRNs display a dynamical phase transition, controlled by the standard deviation of the underlying interaction matrix, separating regimes of weakly and strongly correlated fluctuations. We demonstrated that this transition leaves detectable signatures in the distribution of overlaps even when inferred from static measurements

such as scRNA-seq. We tested this prediction using published scRNA-seq data and found that stem and progenitor cells exhibit fluctuation patterns reminiscent of glassy dynamics, with strong temporal correlations. In contrast, differentiated cells display more rapidly decorrelating gene expression fluctuations over time.

To identify correlated fluctuations in cells, we implemented RBMs trained directly on scRNA-seq data. The RBM bipartite probabilistic model mirrors the structure of our theoretical Hamiltonian and provides a physically motivated framework for learning patterns of co-fluctuation in gene expression [20]. We found that cell types associated with higher potency, such as stem or progenitor cells, tend to exhibit broader coupling distributions and more structured covariance patterns in gene expression fluctuations. These signatures are consistent with a regime of stronger correlations, as predicted by our theory. In contrast, differentiated cell types show narrower coupling distributions and weaker co-fluctuation patterns.

Our results suggest that the identity of cells is not only defined by absolute concentrations of biochemical species, but also by the how fluctuatation propagate on the system-level. Because these fluctuations can be long-lived, they may potentially be used by cells to encode biological information. Whether strongly correlated gene expression are indeed exploited by cells to perform biological functions is an interesting question for future research. Our work also indicates a way for the biophysical interpretation of widely used single-cell sequencing experiments: although absolute gene expression values are usually strongly confounded by technical factors and highly specific to a given cell type, the way fluctuations propagate on the GRN scale gives a robust and interpretable readout of cell states.

Taken together, our results suggest a new interpretation of the Waddington landscape: not as a static potential, but as a high-dimensional, fluctuating system governed by dynamical regimes [36]. Our findings challenge the dichotomous view of stem versus differentiated states and point to a more complex organization of cell identity.

Acknowledgments

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 101034413 (F.O.). This project has received funding from the European Research Council (ERC, grant agreement no. 950349, S.R.). Figure 1a has been created with BioRender.com.

Authors contributions

F.O. designed, planned, and conducted the research; F.O. and S.R. conceptualized and supervised the work; F.O. performed the theoretical and numerical work; Y.D. and F.R performed bionformatics and data analysis; I.D.T. and V.S. performed bionformatics and machine learning; S.R. acquired funding.

References

[1] C. H. Waddington. *The strategy of the genes: A discussion of some aspects of theoretical biology*. London: Allen & Unwin, 1957.

[2] Lewis Wolpert. "Positional information and the spatial pattern of cellular differentiation". In: *Journal of theoretical biology* (1969).

[3] Masaki Sasai and Peter G. Wolynes. "Stochastic gene expression as a many-body problem". In: *Proceedings of the National Academy of Sciences* (2003).

[4] Aleksandra M Walczak, Masaki Sasai, and Peter G Wolynes. "Self-consistent proteomic field theory of stochastic gene switches". In: *Biophysical journal* (2005).

[5] Michael Strasser, Fabian J Theis, and Carsten Marr. "Stability and multiattractor dynamics of a toggle switch based on a two-stage model of stochastic gene expression". In: *Biophysical journal* (2012).

[6] Jonathan W Young et al. "Measuring single-cell gene expression dynamics in bacteria using fluorescence time-lapse microscopy". In: *Nature protocols* (2012).

[7] Christoph Zechner et al. "Moment-based inference predicts bimodality in transient gene expression". In: *Proceedings of the National Academy of Sciences* (2012).

[8] Niraj Kumar, Abhyudai Singh, and Rahul V Kulkarni. "Transcriptional bursting in gene expression: analytical results for general stochastic models". In: *PLoS computational biology* (2015).

[9] Graham J Hickman and T Charlie Hodgman. "Inference of gene regulatory networks using boolean-network inference methods". In: *Journal of bioinformatics and computational biology* (2009).

[10] Gabriel Torregrosa and Jordi Garcia-Ojalvo. "Mechanistic models of cell-fate transitions from single-cell data". In: *Current Opinion in Systems Biology* (2021).

[11] Rory Stark, Marta Grzelak, and James Hadfield. "RNA sequencing: the teenage years". In: *Nature Reviews Genetics* (2019).

[12] Sagar and Dominic Grün. "Deciphering cell fate decision by integrated single-cell sequencing analysis". In: *Annual review of biomedical data science* (2020).

[13] Anika Gupta et al. "Inferring gene regulation from stochastic transcriptional variation across single cells at steady state". In: *Proceedings of the National Academy of Sciences* (2022).

[14] Amanda J Oliver et al. "Single-cell integration reveals metaplasia in inflammatory gut diseases". In: *Nature* (2024).

[15] Donald Voet and Judith G Voet. *Biochemistry*. 3rd ed. Includes CD-ROM. Hoboken, New Jersey: John Wiley, 2004.

[16] Mor Nitzan and Michael P Brenner. "Revealing lineage-related signals in single-cell gene expression using random matrix theory". In: *Proceedings of the National Academy of Sciences* (2021).

[17] Gavin S Hartnett, Edward Parker, and Edward Geist. "Replica symmetry breaking in bipartite spin glasses and neural networks". In: *Physical Review E* (2018).

[18] Romeo Brunetti, Giorgio Parisi, and F Ritort. "Asymmetric Little spin-glass model". In: *Physical Review B* (1992).

[19] Gavin S Hartnett, Edward Parker, and Edward Geist. "Replica symmetry breaking in bipartite spin glasses and neural networks". In: *Physical Review E* (2018).

[20] Anthony Szedlak, Giovanni Paternostro, and Carlo Piermarocchi. "Control of asymmetric Hopfield networks and application to cancer attractors". In: *PloS one* (2014).

[21] L. F. Cugliandolo. "Course 7: Dynamics of Glassy Systems". In: *Slow Relaxations and nonequilibrium dynamics in condensed matter*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2003.

[22] A. Crisanti and H.J. Sommers. "The spherical p-spin interaction spin glass model: the statics". In: *Zeitschrift für Physik B Condensed Matter* (1992).

[23] Alain Barrat. “The p-spin spherical spin glass model”. In: *arXiv: Disordered Systems and Neural Networks* (1997).

[24] Byungjin Hwang, Ji Hyun Lee, and Duhee Bang. “Single-cell RNA sequencing technologies and bioinformatics pipelines”. In: *Experimental & molecular medicine* (2018).

[25] Jong Kyoung Kim et al. “Characterizing noise structure in single-cell RNA-seq distinguishes genuine from technical stochastic allelic expression”. In: *Nature communications* (2015).

[26] Geoffrey E. Hinton. “Training Products of Experts by Minimizing Contrastive Divergence”. In: *Neural Computation* (2002).

[27] Hannah Hochgerner et al. “Conserved properties of dentate gyrus neurogenesis across postnatal development revealed by single-cell RNA sequencing”. In: *Nature Neuroscience* (2018).

[28] Aurélien Decelle, Cyril Furtlehner, and Beatriz Seoane. “Equilibrium and non-equilibrium regimes in the learning of restricted Boltzmann machines”. In: *Advances in Neural Information Processing Systems* (2021).

[29] Jérôme Tubiana, Simona Cocco, and Rémi Monasson. “Learning protein constitutive motifs from sequence data”. In: *Elife* (2019).

[30] Martin Treppner et al. “Synthetic single cell RNA sequencing data from small pilot studies using deep generative models”. In: *Scientific reports* (2021).

[31] Kimberly Siletti et al. “Transcriptomic diversity of cell types across the adult human brain”. In: *Science* (2023).

[32] Yichi Xu et al. “A single-cell transcriptome atlas profiles early organogenesis in human embryos”. In: *Nature cell biology* (2023).

[33] Daniel J Amit, Hanoch Gutfreund, and Haim Sompolinsky. “Storing infinite numbers of patterns in a spin-glass model of neural networks”. In: *Physical Review Letters* (1985).

[34] Jinho Baik, Gŕrared bem Aroud, and Sandrine Péché. “Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices”. In: *The Annals of Probability* (2005).

[35] Aurélien Decelle and Cyril Furtlehner. “Restricted Boltzmann machine: Recent advances and mean-field theory”. In: *Chinese Physics B* (2021).

[36] Jin Wang et al. “Quantifying the Waddington landscape and biological paths for development and differentiation”. In: *Proceedings of the National Academy of Sciences* (2011).