SiGMoiD: A superstatistical generative model for binary data

Purushottam D. Dixit

Department of Physics, University of Florida, Gainesville, FL and

Genetics Institute, University of Florida, Gainesville, FL

Abbstract

In modern biological physics, there is a great interest in building generative probabilistic models for ensembles of covarying binary variables. A popular approach is to use the maximum entropy principle. Here, one builds generative models that use as constraints lower level statistics estimated from the data. While extremely popular, maximum entropy models have conceptual as well as practical issues; they rely on the modelers' choice of constraints and are computationally expensive to infer when the number of variables is large ($n > 100$). Here, we address both these issues with **S**uperstast**i**stical **G**enerative **Mo**del for b**i**nary **D**ata (SiGMoiD). SiGMoiD is a maximum entropy based framework where we imagine that the data as arising from superstatistical system; individual binary variables are coupled to the same bath whose intensive variables fluctuate from sample to sample. Moreover, instead of choosing the constraints, in SiGMoiD we choose only the number of constraints and let the algorithm infer them from the data itself. Notably, we show that SiGMoiD is orders of magnitude faster than current maximum entropy-based models and allows us to model collections of very large number of binary variables. We also discuss future directions.

Introduction: In recent years, there has been a great interest in understanding the statistics of covarying random binary variables. Significant examples are in neuroscience where neurons are either silent or firing (1), in protein/DNA sequence evolution where amino acid or nucleotide positions are either wild type or mutant (2), or in presence and absence of species in an ecosystem (3), for example, the microbiome (4). Estimating from available samples the frequency of occurance of every possible binary configuration is not possible for any reasonably sized collection; a system with N co-dependent binary variables has 2^N states and the number of samples available is typically orders of magnitude lower than the number of states.

At the same time, given the complexity of interactions, it is not possible to build bottom-up mechanistic models to describe these systems. A popular alternative has been to develop approximate top-down probabilistic models and train those models on the data. Over the past two decades, the maximum entropy (max ent) approach (5) has emerged as a strong candidate for building approximate models. Here, one computes user-specified lower order statistics from the samples and seeks a maximum entropy distribution consistent with these data-drive constraints. These models have been used in a variety of contexts, ranging from

illustrating the critical all or none nature of collective neuron firings (6) as well as in more practical applications to predict physical proximity between amino acids in three dimensional structures of proteins (7).

A key advantage of max ent models is that they are unbiased expect for the constraints imposed by the modeler (5, 8). For binary data, constraints of averages and covariances have become the standard (2, 9). However, learning max ent models using these constraints has practical as well as conceptual issues issues. A system comprising N binary variables has $\sim N^2$ covariances and a max ent model constraining these covariances has $\sim N^2$ Lagrange multipliers (parameters). Inference of these parameters from data require extensive Markov chain Monte Carlo (MCMC) simulations and are computationally expensive. Currently, the inference is limited to $\sim$ 100 binary variables (10). Conceptually, and perhaps more importantly, max ent models require the modelers to *a priori* know which constraints are appropriate for any particular problem. Indeed, different constraints lead to models that fit data with different levels of accuracy and with different sets of predictions (11, 12).

To overcome the conceptual limitations of modeler-prescribed constraints, we recently developed and illustrated using several examples a constraint-agnostic max ent approach, thermodynamic manifold inference (TMI) (13). In TMI, we imagined that we had access to a collection of distributions in the same class (several grayscale images or microbial or mRNA abundances). We assumed that K different constraints ('energies') captured these distributions but did not specify *a priori* the values of those constraints. From the data, we inferred the constraints that were hypothesized to be common across all sample distributions as well as distribution-specific Lagrange multipliers.

However, TMI needed multiple distributions from the same class to learn the constraints. As a result, it was not suitable to be applicable when several samples are given from a single hypothesized distribution (for example, multiple protein sequences or samples of collective neuron firings). Here, we propose SiGMoiD; **S**uperstatist**i**cal **G**enerative**Mo**del for b**i**nary **D**ata; a hierarchical generalization of thermodynamic manifold inference using the superstatistical framework that is applicable broadly to capture covariation in binary data.

In SiGMoiD, we assume that the data is generated in a hiearchical manner. Every binary variable is characterized by K types of energies and is coupled to a bath which can exchange these energies. Invoking the superstatistical approach, we assume that the bath intensive variables vary from sample to sample according to a pre-defined distribution.

Given that all binary variables interact with the same bath, their probabilities become correlated with each other. Importantly, the constraints (energies) are not specified by the user but instead inferred directly from the data itself. Given its constraint-agnostic nature, SiGMoiD is a significant conceptual advance over previous max ent based methods while retaining the intuitive interpretability of statistical physics-based models; the inferred constraints/energies can be used to identify correlations and clusters. Practically speaking, inference of SiGMoiD-based models is orders of magnitude efficient compared to the inference of Lagrange multipliers in max ent models. As a result, SiGMoiD can be applied to very large collections of binary variables ($N \sim 500$) that remain out of the reach of max ent models.

Below, we sketch the theoretical developments of SiGMoiD and then apply it to model three experimental data sets (4, 14, 15). We show the utility of SiGMoiD with systems that are too large to be modeled using max ent models. We also discuss generalizations to other types of data.

The model: Consider N binary variables $\{\sigma_i\}$ that take values 0 or 1. Let us denote by π_i the probability that $\sigma_i = 1$ and by $\boldsymbol{\pi}$ the vector of probabilities $\boldsymbol{\pi} = \{\pi_1, \pi_2, \ldots, \pi_N\}$. We imagine the following physical process: each binary variable is connected to the same bath that can exchange K types of extensive variables (energies). The k^{th} type of extensive variable for each variable in the state when it is active ($\sigma_i = 1$) is E_{ki} and zero when it is inactive ($\sigma_i = 0$) (denoted collectively by $\mathbf{E}$). Under these circumstances, the probability of the i^{th} variable is equal to 1 is given by the Gibbs-Boltzmann distribution:

$$\pi_i = \frac{\exp(-\sum_k \beta_k E_{ki})}{\exp(-\sum_k \beta_k E_{ki}) + 1} \approx \exp\left(-\sum_{k=1}^{K} \beta_k E_{ki}\right). \tag{1}$$

In Eq. 1, $\boldsymbol{\beta} \equiv \{\beta_k\}$ are the bath intensive variables. The second approximation is valid when $\pi_i \ll 1$. While not required in principle, in what follows, we will assume that the approximation holds true. This approximation greatly simplifies our calculations below. For numerical stability, we will assume that $E_{ki} > 0$ to ensure that π_i always remain less than 1. The probabilities in Eq. 1 are also interpreted as the max ent probability distributions when averages of the K types of energies $\langle E_{ki} \rangle (k \in [1, K])$ are specified for each neuron $i, (i \in [1, N])$.

Once $\boldsymbol{\beta}$ and $\mathbf{E}$ are known, samples can be generated using Eq. 1 very easily. However, the propensities $\boldsymbol{\pi}$ in Eq. 1 will be statistically independent of each other. We are however

interested in modeling the correlated binary variables. We model the correlations using a superstatistical framework. We posit that each instantiation of the N variable system (each sample) has its own set of bath intensive variables. We require that the intensive variables are distributed according to a pre-defined joint distribution. Here, for simplicity, we assume that the intensive variables are distributed according to independent exponential distributions:

$$p(\boldsymbol{\beta}) = \prod_k \exp\left(-\beta_k\right). \tag{2}$$

In Eq. 2, we have set the mean values $\langle \beta_k \rangle = 1$ since that E_{ki} and β_k can be multiplied by an arbitrary constant without changing the predictions. Other more complex distributions are possible as well. The superstatistical framework can be thought of as a hierarchical max ent inference (16–19) wherein the distributions in Eq. 1 are identified maximum entropy distributions that reproduce average energies and the distributions in Eq. 2 are identified as max ent distributions that reproduce average intensive variables. In this setup, when the energies $\mathbf{E}$ are known, samples can be generated by first sampling temperatures $\boldsymbol{\beta}$ using Eq. 2, then evaluating the probabilities π_i in Eq. 1, and finally sampling σ_i using those probabilities.

Our goal is to infer the energies $\mathbf{E}$ given samples. In our hierarchical setup, the intensive variables are unobservable (latent) variables. As a result, inference of the energies from data using rigorous likelihood maximization will have to resort to techniques such as expectation maximization (20). Here, we propose a simpler approximate approach. Assuming the approximation in Eq. 1 is valid, we can analytically calculate the average probability that $\sigma_i = 1$ and the pairwise correlations. We have:

$$\langle \pi_i \rangle = \int p(\boldsymbol{\beta}) \pi_i d\boldsymbol{\beta} = \prod_k \frac{1}{1 + E_{ki}} \tag{3}$$

and

$$\langle \pi_i \pi_j \rangle = \int p(\boldsymbol{\beta}) \pi_i \pi_j d\boldsymbol{\beta} = \prod_k \frac{1}{1 + E_{ki} + E_{kj}} \tag{4}$$

We note that the formulae in Eq. 3 and Eq. 4 depend on the specific functional form of the superstatistical distribution.

We can find the energies $\mathbf{E}$ by minimizing the squared error:

$$\mathcal{E} = \sum_i \left(\langle \sigma_i \rangle - \langle \pi_i \rangle\right)^2 + \sum_{i \neq j} \left(\langle \sigma_i \sigma_j \rangle - \langle \pi_i \pi_j \rangle\right)^2. \tag{5}$$

The gradients of this error function with respect to $\mathbf{E}$ can be found analytically (see appendix) and used in a gradient descent set up to estimate the energies. Other error functions, for example, the Kullback-Leibler divergence can also be implemented.

Results: Before we illustrate SiGMoiD using larger data sets, we first show a comparison between SiGMoiD and a max ent model. To that end, we use a previously collected data set measuring collective firing of 160 retinal neurons for duration of a movie that lasted 19 seconds (21). The details of the experiment can be found in the original article (21). We note that inference of a max ent model for the collective firing of all 160 neurons is currently computationally unrealistic. Therefore, we chose 15 most active neurons in the data (15 highest firing propensities). First, we inferred a max ent model from the data that constrained mean firing rates and pairwise correlations. The max ent model describes the probability of any configuration $\boldsymbol{\sigma} \equiv \{\sigma_i\}$ as:

$$p(\boldsymbol{\sigma}) \propto \exp\left(-\sum_{i,j} J_{ij}\sigma_i\sigma_j\right). \tag{6}$$

In Eq. 6, J_{ij} are coupling constants that need to be inferred from the data, typically using gradient descent (11). Given that there are only $2^{15} \sim 3\times10^4$ states for 15 neurons, we could estimate model predictions and therefore the coupling constants by a brute force summation over all possible states without resorting to MCMC simuations. This minimized the errors in max ent inference that arises due to inaccuracies in MCMC-based estimates of average firing rates and neuron-neuron correlations. The model had $15 + \binom{15}{2} = 120$ parameters. Next, we inferred a SiGMoiD-based model using $K = 4$ types of energies (a total of 60 parameters). In Fig. 1, we compare the two models. In panel (a), we show a comparison between the raw probabilities of individual configurations obtained from data (x-axis) to model predicted probabilities (y-axis, red: max ent, blue: SiGMoiD). It is clear that SiGMoiD model has lower error compared to the max ent model (mean absolute error 1.08×10^{-5} vs 1.22×10^{-5}). In panel (b), we plot the probability $p(n)$ that n neurons fire at the same time as observed in the data (black), predicted using SiGMoiD (blue), and using the max ent model (red). Here too, the SiGMoiD model performs better especially when capturing the likelihood that a large number of neurons fire together. In panels (c) and (d), we plot the three-body correlations $\langle\sigma_i\sigma_j\sigma_k\rangle$ as observed in the data (x-axis) and as predicted by the model (y-axis, SiGMoiD, panel (c), max ent, panel (d)). Both models capture the three body correlations with reasonable accuracy; the mean absolute error is 0.042 vs 0.038 for the SiGMoiD and

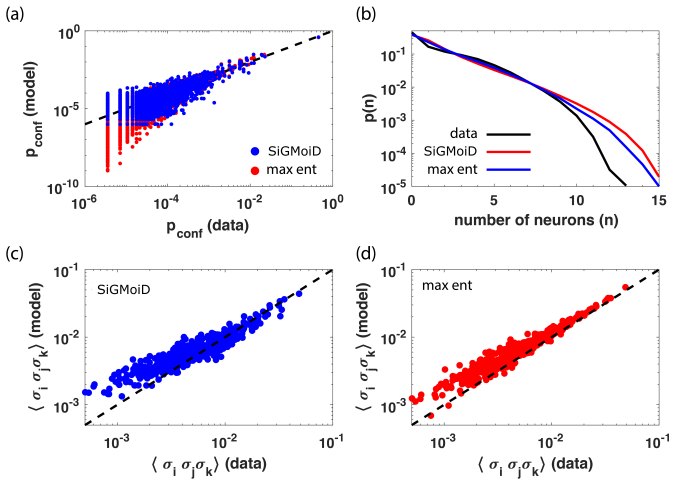

FIG. 1. **Comparison of SiGMoiD with max ent modeling**. (a) the probabilities of individual configurations estimated from the data ($p_{\rm data}(\boldsymbol{\sigma})$, x-axis) and from the two models ($p_{\rm model}(\boldsymbol{\sigma})$, x-axis, red: max ent, blue: SiGMoiD) (b) the probability $p_{\geq}(K)$ that K or more neurons fire in any given configuration as estimated from data (black), the max ent model (red), and SiGMoiD (blue), (c) and (d) comparison between three variable correlations $\langle \sigma_i \sigma_j \sigma_k \rangle$ estimated from data (x-axis) and those using the models (y-axis).

the max ent model respectively. This analysis shows that the SiGMoiD approach is at least as good, if not better, than the max ent based model at capturing the data and predictions.

Next, we analyzed neural recording data from Steinmetz et al. (15). Spiking data was recorded across multiple brain regions of mice. The mice were shown two visual gratings with differing level of contrast and were rewarded to select the grating with a higher contrast. For our analysis, we combined neural data from the root region of the brain in one of the experimental sessions (details in appendix). The data comprised collective firings of 626 neurons. We modeled this data using $K = 2 - 96$ types of energies. We found that while most fits were able to reproduce the average firing rate and two body correlations with similar degree of accuracy (not shown), the collective behavior; the probability that n neurons firing at the same time was well captured only by the most complex model ($K = 96$ energies) (see Fig. A1). We note that inference of max ent models with pair correlation constraints for this large a data set is currently not possible. In contrast, all SiGMoiD calculations were carried out on a personal computer within a matter of minutes and did

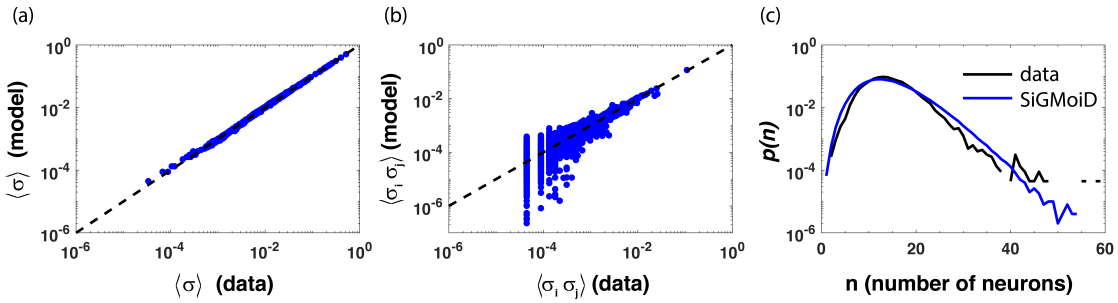

FIG. 2. **SiGMoiD captures the statistics of a large collection of neurons**. We used $K = 96$ energies to model the collective firing of $N = 628$ neurons. (a) The average firing rates $\langle\sigma\rangle$ as calculated in the data (x-axis) and as predicted by SiGMoiD (y-axis). (b) The correlated firing rates of two neurons $\langle\sigma_i\sigma_j\rangle$ as calculated in the data (x-axis) and as predicted by SiGMoiD (y-axis). We have chosen 5000 random pairs from all pairs. (c) The probability $p(n)$ that n neurons fire at the same time (black: data, blue: SiGMoiD).

not require any extensive computational infrastructure. Moreover, the SiGMoiD model had $96 \times 626 \approx 6 \times 10^4$ parameters. In comparison, a pairwise Ising-type model would have a much larger number, $\approx 2 \times 10^5$, of parameters.

Another type of binary data popular in biological physics is presence and absence of species in an ecosystem. Gut microbiomes are an ecosystems whose statistical properties have received significant attention in the last few years (22). Bacteria in the gut live in complex communities where they compete for nutrients and also exchange metabolites with each other. These associations create complex spatial structures of bacterial communities in the gut spanning several length scales. Recently, Sheth et al. (4), devised an experimental method to probe the spatial organization of the gut microbiome at the micron length scale allowing them to identify putative direct interactions between bacteria. In these experiments, Sheth et al. (4) fractured mice guts into particles of a specific size and quantified membership of $\sim$ 300 operational taxonomic units (OTUs) in $\sim$ 1500 particles with median diameter $\sim 30\mu m$. Each particle was characterized by a binary vector representing the OTUs present in that particle.

We used $K = 8$ energies to model the collective behavior of OTUs. In panel (a) of Fig. 3 we show the probability of observing n OTUs in any particle as observed in the data (black circles) and as predicted by SiGMoiD (blue line). It is clear that SiGMoiD

accurately captures this co-occurrence distribution. SiGMoiD also captured the mean occurrence frequency and pairwise correlations accurately (see Fig. A2). Notably, SiGMoiD characterizes each binary variable (here, OTUs) using a K dimensional vector. Therefore, it can be used to identify OTUs who have similar occurrence profiles using hierarchical clustering. There are two types of interactions between bacteria that lead to co-occurrence in any ecosystem (23), especially at the micron length scale (4). First, genetically related bacteria tend to co-occur because they have similar metabolic networks and can compete for the same resources. Second, genetically dissimilar bacteria have different metabolic networks and can cross-feed each; one bacteria utilizing the waste products of another. Given that co-occurrences are transitive (A co-occurs with B, B co-occurs with C $\Rightarrow$ A co-occurs with C), it is not possible to use simple co-occurrence calculations to identify putative pairs of interacting bacteria (24). SiGMoiD-based clustering of OTUs is a more direct clustering that relies on inferred inherent properties of the OTUs (the energies). Panel (b) of Fig. 3 shows a hierarchical clustering plot of all OTUs using Ward's linkage. Instead of the energies E_{ki}, we used the transformed variables $r_{ki} = 1/(1 + E_{ki})$ that are related to the mean occurrence frequencies (see Eq. 3).

Among the several identified clusters, we focus on two biologically interesting ones (details of cluster membership can be found on github, see appendix for the link). The gut microbiome of mice is dominated by OTUs belonging to the family *Lachnospiraceae*; 53% of all the OTUs belonged to this family. However, these OTUs were not equally dispersed across the particles. We found one cluster of OTUs comprising 11 OTUs that was statistically significantly enriched in *Lachnospiraceae* (9 out of 11, single tailed binomial distribution p-value 0.05) and another cluster of OTUs comprising 32 OTUs that was statistically significantly depleted in *Lachnospiraceae* (9 out of 32, single tailed binomial distribution p-value 0.0035). This same cluster was also enriched in the family *Christensenellaceae* (3 out of 32 against a background frequency of 1.7%, single tailed binomial distribution p-value 0.017). In the future, it will be interesting to identify direct metabolic interactions amongst OTUs belonging to these clusters.

Discussion: A deluge of biophysical data in the last decade has necessitated the development of top-down modeling. Here, instead of describing the data from first principles mechanistic models, one constructs probability distributions that represent it. As a result, generative models of collective behavior have become essential to modeling several biophys-

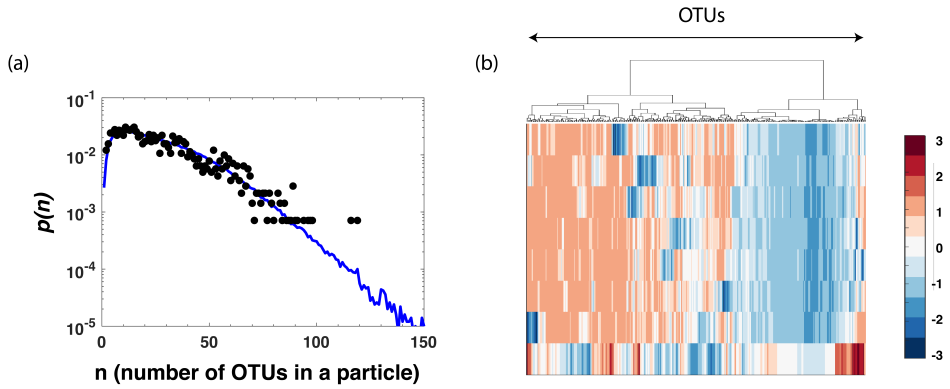

FIG. 3. **SiGMoiD captures the statistics of microbial co-occurrences**. We used $K = 8$ energies to model the co-occurrence of $N = 347$ OTUs. (a) The probability $p(n)$ that n OTUs are observed in any particle (black: data, blue: SiGMoiD). (b) Hierarchical clustering diagram of the OTUs identifies several clusters of OTUs with similar occurrence patterns.

ical systems. The most popular way to generate top-down models is the maximum entropy (max ent) approach wherein one approximates the data using a probabilistic model that reproduces lower order statistics estimated from the data. The max ent approach has a significant conceptual advantage that it represents the simplest model consistent with the imposed constraints (5, 8). However, there are two significant drawbacks. First, the constraints are hand-picked by the modeler and the model therefore depends on these constraints. For binary data, constraints of averages and pair correlations have become popular. Second, the inference of max ent models for large data sets can be computationally expensive and it may be unrealistic to infer models for > 100 binary variables.

To address these issues, we developed SiGMoiD. SiGMoiD takes an agnostic approach about the constraints. In SiGMoiD, instead of specifying the constraints, the user only specifies the total number of constraints. SiGMoiD learns these constraints from the data. Moreover, parameter inference in SiGMoiD is orders of magnitude faster than max ent inference. We showed using three data sets of varying complexity that SiGMoiD not only performs as well as max ent models in terms of accuracy but can also be applied to study very large data sets that are currently out of the reach of max ent inference.

There are several directions in which SiGMoiD can be improved. First, we assumed (1) an approximate expression for the probabilities (see Eq. 1) and (2) exponentially distributed independent intensive variables which greatly simplified the predictions from the

model. A more rigorous approach would be to infer sample specific intensive variables using the Gibbs-Boltzmann definition of probabilities by employing methods such as expectation maximization (20). Second, we currently assumed that the intensive variables are drawn from independent exponential distributions. However, other choices, for example, gamma and inverse gamma distribution, are possible as well. Third, the current approach only applies to binary data, however, in many cases, a more general approach might be required. For example, when modeling variation in DNA sequences or protein sequences (7), we may need to model distributions of four and twenty possible outcomes respectively. SiGMoiD can be easily generalized to model such data. Fourth, the number of energies K used to fit the data remains a free parameter in SiGMoiD inference. However, models with different K are nested therefore the log-likelihood ratio test or any other information theory based criterion can be implemented to determine the optimal number of energies in SiGMoiD. Going forward, we believe that this computationally efficient and conceptually straightforward approach will be immensely valuable in modeling collective behavior of binary and other categorical data.

[1] E. Schneidman, M. J. Berry, R. Segev, and W. Bialek, Nature **440**, 1007 (2006).

[2] S. Cocco, C. Feinauer, M. Figliuzzi, R. Monasson, and M. Weigt, Reports on Progress in Physics **81**, 032601 (2018).

[3] A. E. Noble, T. S. Rosenstock, P. H. Brown, J. Machta, and A. Hastings, Proceedings of the National Academy of Sciences **115**, 1825 (2018).

[4] R. U. Sheth *et al.*, Nature biotechnology **37**, 877 (2019).

[5] S. Pressé, K. Ghosh, J. Lee, and K. A. Dill, Reviews of Modern Physics **85**, 1115 (2013).

[6] G. Tkačik *et al.*, Proceedings of the National Academy of Sciences **112**, 11508 (2015).

[7] F. Morcos *et al.*, Proceedings of the National Academy of Sciences **108**, E1293 (2011).

[8] J. Shore and R. Johnson, IEEE Transactions on information theory **26**, 26 (1980).

[9] C. Savin and G. Tkačik, Current opinion in neurobiology **46**, 120 (2017).

[10] J. P. Barton, E. De Leonardis, A. Coucke, and S. Cocco, Bioinformatics **32**, 3089 (2016).

[11] G. Tkacik, E. Schneidman, M. J. Berry II, and W. Bialek, arXiv preprint arXiv:0912.5409 (2009).

[12] G. Tkačik *et al.*, Journal of Statistical Mechanics: Theory and Experiment **2013**, P03011 (2013).

[13] P. D. Dixit, Physical Review Research **2**, 023201 (2020).

[14] J. Humplik and G. Tkačik, PLoS computational biology **13**, e1005763 (2017).

[15] N. A. Steinmetz, P. Zatka-Haas, M. Carandini, and K. D. Harris, Nature **576**, 266 (2019).

[16] G. E. Crooks, Physical Review E **75**, 041119 (2007).

[17] P. D. Dixit, The Journal of chemical physics **138**, 05B612_1 (2013).

[18] P. D. Dixit, Physical Chemistry Chemical Physics **17**, 13000 (2015).

[19] P. D. Dixit, A. Bansal, W. G. Chapman, and D. Asthagiri, The Journal of Chemical Physics **147**, 164901 (2017).

[20] T. K. Moon, IEEE Signal processing magazine **13**, 47 (1996).

[21] G. Tkačik *et al.*, PLoS Comput Biol **10**, e1003408 (2014).

[22] B. W. Ji, R. U. Sheth, P. D. Dixit, K. Tchourine, and D. Vitkup, Nature Microbiology **5**, 768 (2020).

[23] J. Friedman and J. Gore, Current Opinion in Systems Biology **1**, 114 (2017).

[24] R. Menon, V. Ramanan, and K. S. Korolev, PLoS computational biology **14**, e1005939 (2018).

I. APPENDIX

A. Github link

All raw data and scripts can be found on GitHub: https://github.com/dixitpd/sigmoid

B. Gradient of the error function

We define the error as:

$$\mathcal{E} = \sum_i \left(\langle\sigma_i\rangle - \langle\pi_i\rangle\right)^2 + \sum_{i\neq j} \left(\langle\sigma_i\sigma_j\rangle - \langle\pi_i\pi_j\rangle\right)^2 . \tag{A1}$$

To find the gradient of $\mathcal{E}$ with respect to E_{ki}, we first consider the derivatives with respect to E_{ki} of $\langle\pi_i\rangle$ and $\langle\pi_i\pi_j\rangle$. We have

$$\frac{\partial\langle\pi_i\rangle}{\partial E_{ki}} = -\frac{\langle\pi_i\rangle\lambda_k}{\lambda_k + E_{ki}} \tag{A2}$$

$$\frac{\partial\langle\pi_i\pi_j\rangle}{\partial E_{ki}} = -\frac{\langle\pi_i\pi_j\rangle\lambda_k}{\lambda_k + E_{ki} + E_{kj}} \tag{A3}$$

Therefore, we have

$$\begin{aligned}\frac{\partial\mathcal{E}}{\partial E_{ki}} &= 2\left(\langle\pi_i\rangle - \langle\sigma_i\rangle\right)\frac{\partial\langle\pi_i\rangle}{\partial E_{ki}} \\ &\quad + 2\sum_{j\neq i}\left(\langle\pi_i\pi_j\rangle - \langle\sigma_i\sigma_j\rangle\right)\frac{\partial\langle\pi_i\pi_j\rangle}{\partial E_{ki}}\end{aligned} \tag{A4}$$

We used the gradient descent algorithm to infer the energies from the data. In order to ensure positivity of energies, the gradient descent was performed for logarithms of energies instead of the energies themselves.

C. Details of extraction of data from Steinmetz et al. (15)

Steinmetz et al. (15) collected neural data across several brain regions in 10 mice observed over 39 separate sessions. The goal of the experiment was to study how the mice responded to a visual cue in the form of difference in contrast in gratings on the left and the right side of the mice. From this large amount of data, we selected brain recordings from the root region in the 5^{th} session. We only focussed on those trials where the grating contrast was higher on the right hand side compared to the left hand side. This led to recordings of

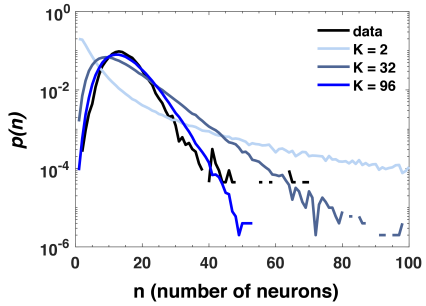

FIG. A1. The probability $p(n)$ that n neurons simultaneously fire in the Steinmetz et al. (15) data (black) and SiGMoiD predictions with $K = 2, 32$, and 96.

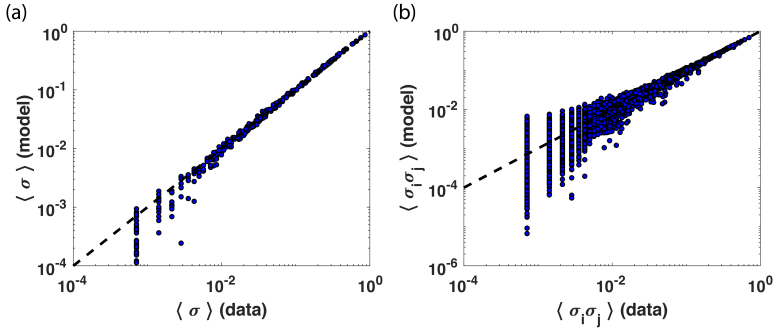

FIG. A2. Panel (a) mean OTU occupancy $\langle\sigma\rangle$ as estimated from the data (x-axis) and as predicted by the model (y-axis). (b) Pair correlations $\sigma_i\sigma_j$ among OTUs as estimated from the data (x-axis) and as predicted from the model (y-axis).

$N = 628$ neurons in 90 trials across 250 time points. The data used here can be found on GitHub.

We analyzed this data using $K = 2, 32$, and 96 types of energies. While these reproduced the average firing rates and pair-correlations to the same degrees of accuracy, only the $K = 96$ model was able to reproduce the probability $p(n)$ that n neurons fired at the same time (Fig. A1). We used $K = 96$ for further analysis.

D. SiGMoiD reproduces mean occupancy and pairwise correlations for microbiome data