BMC Bioinformatics

RESEARCH

Open Access

The maximum entropy principle for compositional data

Corey Weistuch[1], Jiening Zhu[2], Joseph O. Deasy[1] and Allen R. Tannenbaum[2,3*]

*Correspondence: allen.tannenbaum@stonybrook.edu

[1] Department of Medical Physics, Memorial Sloan Kettering Cancer Center, New York, USA
[2] Department of Applied Mathematics & Statistics, Stony Brook University, Stony Brook, USA
[3] Departments of Computer Science and Applied Mathematics & Statistics, Stony Brook University, Stony Brook, USA

Abstract

Background: Compositional systems, represented as parts of some whole, are ubiquitous. They encompass the abundances of proteins in a cell, the distribution of organisms in nature, and the stoichiometry of the most basic chemical reactions. Thus, a central goal is to understand how such processes emerge from the behaviors of their components and their pairwise interactions. Such a study, however, is challenging for two key reasons. Firstly, such systems are complex and depend, often stochastically, on their constituent parts. Secondly, the data lie on a simplex which influences their correlations.

Results: To resolve both of these issues, we provide a general and data-driven modeling tool for compositional systems called Compositional Maximum Entropy (CME). By integrating the prior geometric structure of compositions with sample-specific information, CME infers the underlying multivariate relationships between the constituent components. We provide two proofs of principle. First, we measure the relative abundances of different bacteria and infer how they interact. Second, we show that our method outperforms a common alternative for the extraction of gene-gene interactions in triple-negative breast cancer.

Conclusions: CME provides novel and biologically-intuitive insights and is promising as a comprehensive quantitative framework for compositional data.

Keywords: Networks, Compositional data, Inference, Maximum entropy

Background

Describing the compositions of physical systems, such as in mixtures of industrial chemical reactions, across bacteria in the microbiome, or relative influences in cancer networks is of significant practical importance. In the present work, these systems are modeled as networks of components (or nodes) and their unknown node-node interactions. However, the challenge of inferring these interactions lies in incorporating the defining feature of such compositions: the total proportion across components must always stay fixed.

Much recent interest has been devoted to improving the statistical analysis of compositional data [1–5]. The typical strategies that have been employed broadly fall into three categories. First, many apply traditional statistics (such as correlational analyses).

Applied to compositional data, however, such tools are known to generate spurious results [6–8]. A second approach considers analyses that are unaffected by data rescaling ("scale invariance") and the addition of new components ("subcompositional coherence") [1, 2, 9]. However, such methods cannot natively handle zeros in the data and require transformations (e.g. log ratios) that may introduce unwarranted biases into downstream analyses [10, 11]. A third approach considers more general models of the simplicial geometry, or the set of coordinates that sum to a fixed quantity, inherent to compositional data [12–14]. What is needed, however, is an approach for modeling compositional data that is both general and principled.

In contrast to previous approaches, we aim to infer the structure of our model from the data. The natural method for this is the principle of maximum entropy or Max Ent [15–18]. Here, one provides constraints, such as means, variances, and even the geometry of the data itself, and Max Ent provides the model. The advantage of this approach is twofold. First, as opposed to other modeling approaches, Max Ent makes minimal assumptions that are not warranted by the data itself; we simply require our principle to provide a unique, coordinate-independent answer that preserves independence of subcomponents [19]. Second, Max Ent is a widely and successfully utilized modeling framework for complex biological systems [20–25]. We provide theory and practical demonstrations of our new approach in the present work.

Results

The model

Suppose one is given several stochastic observations of the relative abundances of N different components. Each of these observations may be represented as a vector $\Gamma = \{s_1, s_2, \dots s_N\}$. Our goal is to infer the most likely and least-biased inter-component relationships that give rise to these observations (see Fig. 1). The unique model with this property is provided by the *principle of maximum entropy*, which selects the model P that both maximizes the entropy $S = -\sum_\Gamma P_\Gamma \log P_\Gamma$ and satisfies known constraints from the data. Here, the standard constraints are the estimated first and second moments, $M_i = \langle s_i \rangle$ and $\chi_{ij} = \langle s_i s_j \rangle$ [26], as well as the special compositional constraint, $\sum_i s_i = 1$ (or 100%). The resulting solution P^*, obtained through the method of Lagrange multipliers, is given by:

$$\begin{aligned} P^*_\Gamma &= Z^{-1} \exp\left[\sum_i \left(h_i + \frac{1}{2}\sum_{j\neq i} K_{ij} s_j\right) s_i\right], \\ Z &= \int_{\sum_i^N s_i = 1} \exp\left[\sum_i \left(h_i + \frac{1}{2}\sum_{j\neq i} K_{ij} s_j\right) s_i\right] d\vec{s} \end{aligned} \tag{1}$$

Here h_i and K_{ij} enforce, respectively, the means M_i and the covariances $\chi_{ij} - M_i M_j$. The normalizing constant Z is defined by an intractable integral over the simplex. Thus, the model parameters are found using an adapted pseudolikelihood approximation (see Methods: The simplex pseudolikelihood method). Finally, as $\sum_i s_i = 1$, several constraints are redundant. Thus, we set $h_N = 0$ and K_{ii} $(i = 1, 2, \dots N)$ to 0 (see Methods: Refining the maximum entropy parameters).

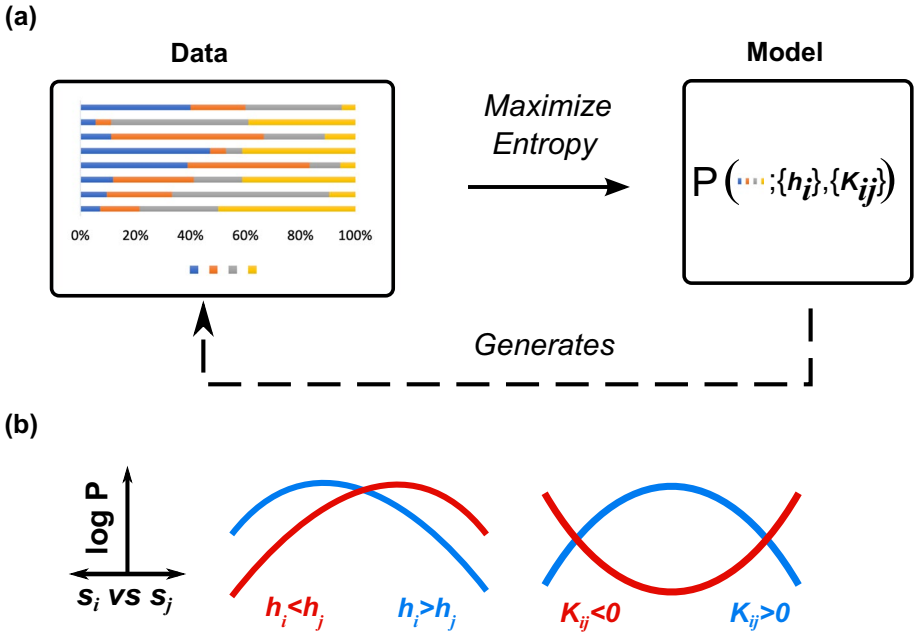

Fig. 1 The Compositional Maximum Entropy (CME) approach. **a** Through maximum entropy, CME infers the unknown generative model of the observed component abundances. **b**. h_i embodies the influence of each (i) component. Components with large h_i tend to have higher abundances than those with small h_i. K_{ij} embodies the interaction between pairs of components. Pairs with $K_{ij} > 0$ tend to coexist, while pairs with $K_{ij} < 0$ tend to be mutually exclusive

In summary, Eq 1 provides the Compositional Maximum Entropy model (CME) subject to known means and covariances. The CME method provides interpretable influence weights h_i for each component node i as well as the interaction strengths K_{ij} between each pair of components (i and j). Below, we provide two proofs of principle of the method: in a model of the abundances of co-evolving species and the analysis of gene expression data in cancer.

Quantifying competition among co-evolving species

The quantification of competition among bacteria in the gut, market forces in the economy (or even among scientists) is of course of great interest. A simple and widely-used mechanism is provided by the competitive Lotka-Volterra model (cLV), which describes the population dynamics (i.e., the abundances) of different species vying for a shared resource [27–29]. The population ($\tilde{s}_i$) of each species i depends on its growth rate r_i and interaction α_{ij} with each other species j. Furthermore, the population of each type stops growing as it nears its carrying capacity κ_i, representing the complete exhaustion of resources.

$$\frac{d\tilde{s}_i}{dt} = r_i\tilde{s}_i \times \left(1 - \frac{\sum_j \alpha_{ij}\tilde{s}_j}{\kappa_i}\right) \tag{2}$$

While cLV remains a powerful model for predicting population dynamics, several challenges remain in calibrating it to experimental data. First, we are often only provided with relative (normalized) species abundances. Tools handling both this information loss

and the resulting compositional data remain problematic [7, 30]. In addition, we rarely have access to the full time series [31]. Bacterial abundances, for example, are typically measured sparsely but across many different conditions and environments [7].

Here we show that CME can provide accurate quantitative estimates of inter-species interactions, as predicted by cLV, using only available experimental information. The simulated cLV abundances $\tilde{s}_i$ are first normalized to resemble experimental data:

$$s_i = \frac{\tilde{s}_i}{\sum_j \tilde{s}_j}, \quad i = 1, 2, \ldots N \tag{3}$$

The time-evolving relative abundances $s_i(t)$ are then randomly sampled to apply CME. Compared to the cLV model, our proposed approach requires fewer parameters that are thus more resolvable from the limited available data [31].

cLV models exhibit three broad classes of stable inter-species behaviors: mutualism (they coexist), neutralism (they ignore each other), and competition (only one type can exist at a time) [30]. To illustrate these behaviors, we consider a cLV model of three different species with equal interactions $\alpha_{ij} = \alpha$. Figure 2 shows the dynamics and abundance distributions for each of three different regimes: $\alpha = 0.6$ (mutualism, Fig. 2a), $\alpha = 1.2$ (neutralism, Fig. 2b), and $\alpha = 4.0$ (competition, Fig. 2c). For simplicity, r_i, κ_i, and the self-interactions α_{ii} are fixed at 1. Gaussian noise was then added to the simulated dynamics to introduce additional inter-sample variability.

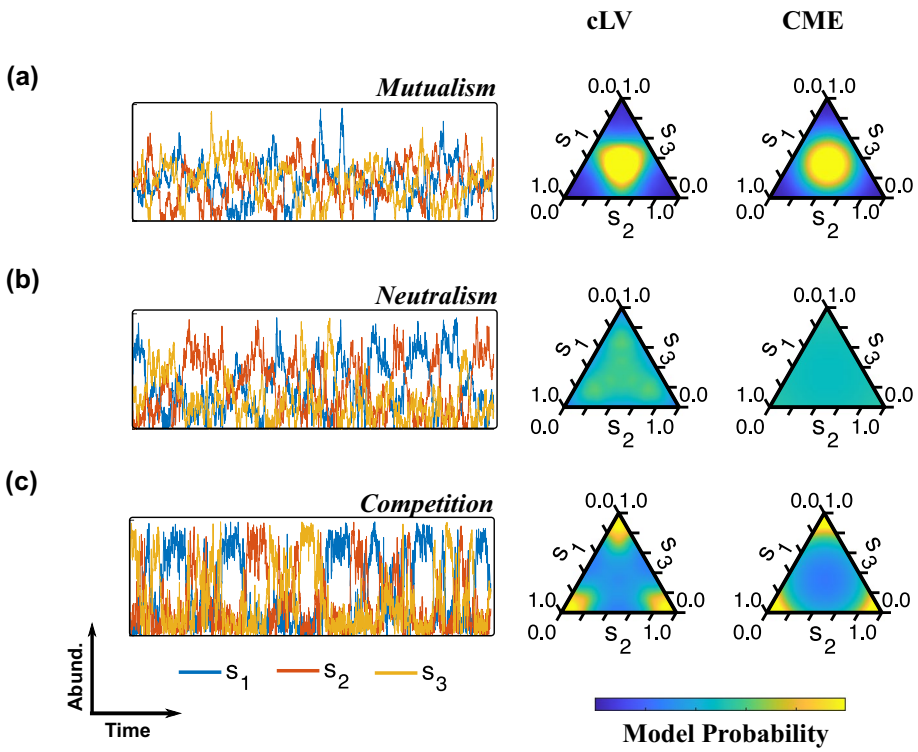

Fig. 2 Simulated abundances of three co-evolving species under mutualism (**a**), neutralism (**b**), and competition (**c**). Left, the cLV simulated abundances of each of the three interacting species over time. Center, the corresponding abundance distribution (cLV). Right, the best fit maximum entropy distribution (CME)

The cLV model exhibits a sharp qualitative transformation in its abundance distribution, from a unimodal (Fig. 2a) to a trimodal (Fig. 2c) behavior [32], as α is increased above a critical value ($\alpha \approx 1.2$, Fig. 2b). Despite requiring fewer parameters ($h = 0$ and K, compared to the original four of cLV), CME (right) captures the cLV model behavior (center) across this transformation; $K > 0$ describes mutualism while $K < 0$ describes competition.

In summary, our model provides a simple, data-driven framework for modeling interspecies relationships from limited experimental information. We next consider the more complex case involving heterogeneous interactions from gene expression data in cancer.

Revealing driving interactions in cancer networks

Cancer is a heterogeneous disease involving complex molecular interactions between many genes. Despite the wealth of information provided by modern experimental tools, the application of such molecular data, including gene expression, to identify effective drug targets continues to face two significant obstacles. First, the accuracy of experimental expression profiles differs between genes [33]. Thus influences from biologically critical but more poorly resolved genes may be overlooked. Second, genes of typical interest often interact, and their effects overlap [34].

Novel network analysis techniques have been developed to refine the genetic signatures of critical genes in cancer. These approaches have been utilized to discover feedback structures in gene interaction networks, identify hubs and bridges, and define measures of robustness and fragility [35–37]. The Wasserstein distance from optimal transport lies as the basis for such methodologies, and in addition to the above references has been directly applied to the stationary (normalized) measures of the networks in question to derive biological information, e.g. showing that pediatric sarcoma data forms a unique cluster [38]. We will now show that CME may provide an important tool for such problems and help point to potential driver genes and their most important interactions.

To test our method, we analyze whole-genome expression data of triple-negative breast tumors, a highly aggressive and complex type of cancer. While many genes are known to be dysregulated in this disease, the relative influence of individual genes is far from established [39]. The data consist of expression profiles from 299 disease samples in METABRIC (Methods: The METABRIC dataset) [40]. We obtained normalized weights for each of $N = 3147$ genes using the Human Protein Reference Database (HPRD) for each sample (Methods: Network identification) [41]. As most of these genes provide no signal in the data, we renormalized these weights after considering only the top 17 highest variabiliy genes with known relevance to cancer (according to OncoKB, see [42] and Methods: Data preparation). Figure 3 illustrates the known connectivity of these genes, but with node size and color proportional to their inferred maximum entropy node weights (h_i). We immediately notice two key details. First, our genes of interest all form a tightly connected network. Second, despite being highly correlated with each other (as the topology would suggest), these genes have unequal influences on the data. The highest-ranked genes, SRC and TP53, are also known master regulators of cancer [43, 44].

A major strength of maximum entropy methods is identifying key node-node interactions underlying the more complex covariances measured from data. This is

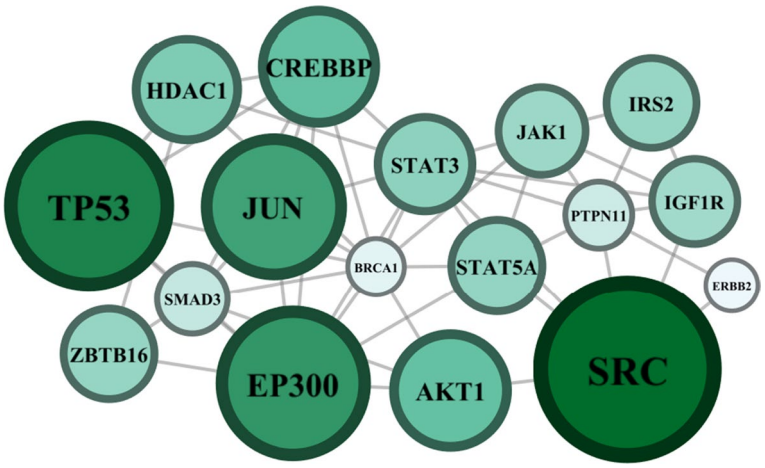

Fig. 3 Maximum entropy ranking of key genes in triple-negative breast cancer. Edges correspond to protein-protein interactions obtained from HPRD. Node color and size correspond to their influence (h_i)

illustrated in Fig. 4, which compares the maximum entropy pairwise interactions K_{ij} to those inferred from a widely-used alternative statistical model, the logit-normal distribution (Methods: Implementation of the logit-normal distribution [1]); there is an identifiable mapping between the strongest magnitude maximum entropy interactions (Fig. 4a), in contrast to these obtained from the logit-normal (Fig. 4b), and their corresponding gene-gene covariances (Fig. 4c).

We also note that the two top maximum entropy interactions alone (SRC/TP53 and BRCA1/PTPN11) provide an intuitive explanation for some of the key features of the data. SRC and TP53 maintain the critical balance between growth (SRC) and damage repair (TP53): enhanced SRC (or repressed TP53) promotes cell survival, growth, and metastasis, while the reverse leads to accelerated aging [43–45]. This known and critical negative interaction between SRC and TP53 separates most of the 17 genes into two distinct (and negatively covarying) clusters. Thus, since BRCA1 and PTPN11 belong to opposing clusters, their corrected interaction, as revealed by both maximum entropy and logit-normal modeling, is much larger than expected from their weak, positive covariance. Interestingly, both BRCA1 and PTPN11, along with SRC and TP53, are involved in the JAK-STAT pathway [46, 47]. Thus, these genes may have a general and synergistic role in cancer that remains to be explored.

Yet, while the logit-normal model does appear to resolve some features (such as the subtle covariance between AKT1 and EP300) that CME neglects, the interactions predicted by this method generally appear difficult to interpret in the context of the original covariance matrix: it predicts many interactions between uncorrelated genes and fails to resolve, among others, the clear negative covariance between SRC and TP53. Overall, the CME method provides a parsimonious biological mechanism, involving known cancer drivers and only a few of their interactions, for the genetic variability in this poorly understood disease.

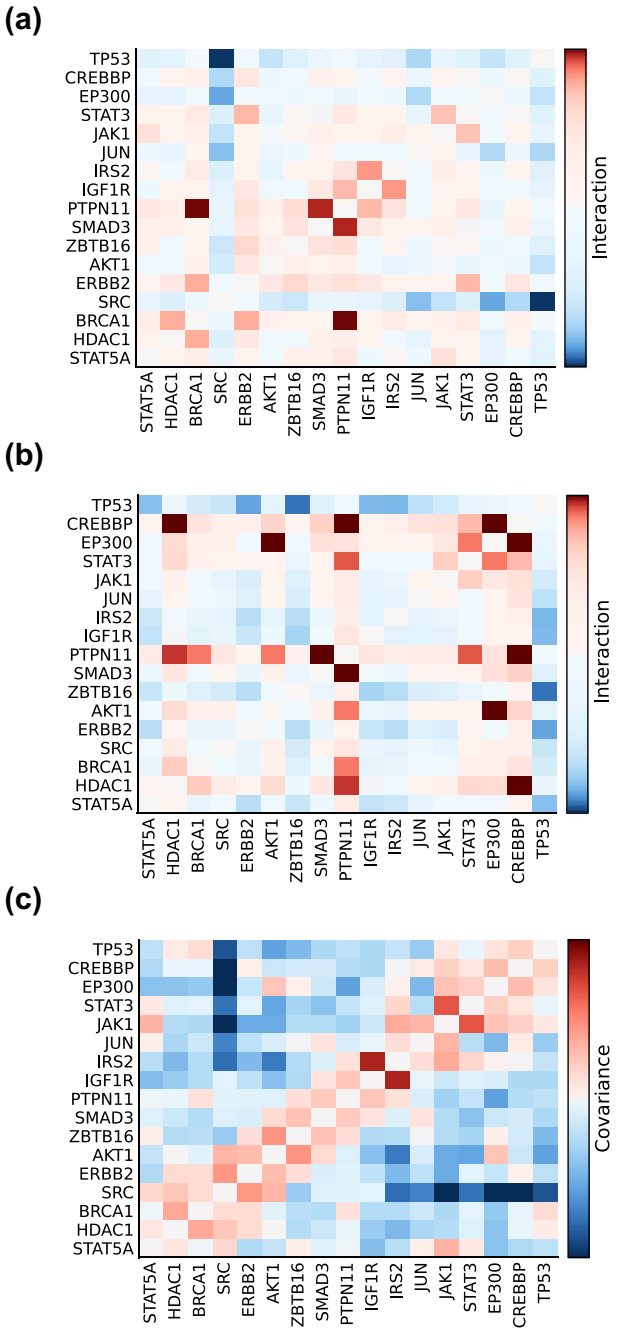

Fig. 4 Comparison between three breast cancer network analyses: CME (**a**), logit-normal (**b**), and the data covariances (**c**). Maximum entropy and logit-normal results are shown on a log-scale to reveal the most influential positive (red) and negative (blue) interactions

Discussion

We have provided CME, a probabilistic framework for inferring the behaviors of compositional systems from data. Typically, models are deduced bottom-up, starting from mathematical relationships between individual components and combining them often in a complex, nonlinear way. However, as we have described for the Lotka-Volterra

model, these interactions can rarely be resolved from the available experimental data. CME, instead, takes a top-down approach – starting from the data and learning the most parsimonious model for it. As evidenced by our breast cancer analysis, CME may also provide more interpretable insights into the organization of compositional systems.

Similar to partial correlational analysis [48–50], maximum entropy computes direct pairwise interactions by controlling for the confounding indirect effects of the other nodes. Despite being widely used in data analysis and machine learning, partial correlations are only appropriate for linear associations or Gaussian-like data [48]. Maximum entropy methods, such as our application to compositional data [49], are, by contrast, much more general.

For simplicity, we have considered only small networks; however, our method can be easily extended to much larger networks. First, the pseudolikelihood approach at the core of our method has been successfully applied, with the proper regularization, to networks consisting of thousands of nodes [51]. Second, the implementation of our algorithm uses a scalable L-BFGS algorithm and is fully parallelized across multiple CPU cores.

The principle of maximum entropy deduces the simplex-truncated normal distribution from the given first and second moment constraints. While such models have been previously studied in compositional data analysis [13], our approach provides two key advantages. First, maximum entropy can naturally incorporate more general model constraints including higher-order moments [26], more complex geometries [52], additional types of data [53], and domain-specific assumptions [1, 2]. Second, our simplex pseudolikelihood method provides consistent [54] and asymptotically efficient [55] parameter estimates and is asymptotically equivalent to maximum likelihood estimation [54]. Furthermore, a recent study demonstrates that score matching approaches can be viewed as approximations of pseudolikelihood [56], suggesting a relationship between our approach and [13] that could be explored in a future work.

Conclusion

We proposed CME, a data-driven framework for modeling compositions in multi-species networks. We utilize maximum entropy, a first-principles modeling approach, to learn influential nodes and their network connections using only the available experimental information. Our method requires minimal assumptions and no modifications of the experimental data. Furthermore, the method can be easily generalized to incorporate new types of constraints and data that may emerge.

Methods

The simplex pseudolikelihood method

Fitting maximum entropy models to data is generally computationally intractable. Thus, to fit CME, we will adapt the widely-used *pseudolikelihood* approximation [51]. This method requires two pieces of information. First, we need a formula to compute the conditional distribution $P(s_i|s_{\sim i})$, where $s_{\sim i}$ represents all of the variables of interest s_j $(j = 1, 2, \ldots N-1)$ excluding s_i and $s_N = 1 - \sum_{i=1}^{N-1} s_i$. For the simplex model, we have:

$$P(s_i|s_{\sim i}) = \tilde{Z}_i(s_{\sim i})^{-1} \exp\left[\left(\tilde{h}_i + \frac{1}{2}\tilde{K}_{ii}s_i + \sum_{j\neq i}\tilde{K}_{ij}s_j\right)s_i\right] \tag{4}$$

$$\tilde{Z}_i(s_{\sim i}) = \int_0^{1-\sum_{j\neq i}^{N-1} s_j} \exp\left[\left(\tilde{h}_i + \frac{1}{2}\tilde{K}_{ii}s_i + \sum_{j\neq i}\tilde{K}_{ij}s_j\right)s_i\right] ds_i \tag{5}$$

$$\tilde{h}_i = h_i + \frac{1}{2}(K_{iN} + K_{Ni}), \quad \tilde{K}_{ij} = K_{ij} - K_{iN} - K_{Nj} \tag{6}$$

Unlike that of Eq 1, $\tilde{Z}_i$ is a tractable Gaussian-like integral. However, its value is sample dependent. Thus, the second required piece of information is the actual samples of the agent proportions s_i^d $(d = 1, 2, \ldots D)$ rather than simply the summary means and covariances. Together these enable the maximization of the *pseudo*likelihood functions ℓ_{PL}^i (see Methods: Model implementation):

$$\ell_{PL}^i = \tilde{h}_i M_i + \frac{1}{2}\tilde{K}_{ii}\chi_{ii} + \sum_{j\neq i}^{N-1}\tilde{K}_{ij}\chi_{ij} - D^{-1}\sum_{d=1}^{D}\log\tilde{Z}_i(s_{\sim i}^d) \tag{7}$$

Refining the maximum entropy parameters

One challenge in modeling compositional data is handling the parameter redundancies induced by the compositional constraint $\sum_i s_i = 1$. Specifically, $M_N = 1 - \sum_{i\neq N} M_i$ and $\chi_{iN} = \chi_{Ni} = M_i - \sum_{j\neq N}\chi_{ij}$ are entirely determined from the other data constraints. We could set the associated Lagrange multipliers to 0, but this would hide information about node N (as all of its connections would be forced to 0).

Instead, we recover interpretable model parameters with the following transformations:

$$K_{ij} = \frac{1}{2}(\tilde{K}_{ij} + \tilde{K}_{ji} - \tilde{K}_{ii} - \tilde{K}_{jj}), \quad h_i = \tilde{h}_i - K_{iN} \tag{8}$$

By forcing K_{ii} to be 0 in Eq 1, we can resolve the interaction strengths between all pairs of nodes in the data. For simplicity, we have defined $h_N = 0$. However, we can increase or decrease all h_i by any constant and still have an equally good fit. Thus we introduce another transformation to facilitate intra-model comparison of these node weights:

$$Q_i = \frac{e^{h_i}}{\sum_{i=1}^{N} e^{h_i}} \tag{9}$$

Conceptually, the quotient Q_i compares the relative probability of observing a network configuration with influence dominated ($P^*(s_i = 1)$) by node i. We posit this as a useful comparison metric for future studies of compositional systems modeled under different conditions.

Model implementation

To provide a high-accuracy, low overhead approximate maximum of the CME log pseudolikelihood functions, we performed convex optimization using L-BFGS [57] augmented by automatic differentiation. To validate our method, we also designed a custom Monte-Carlo scheme to simulate from CME models. This scheme considers the fitted h_i and K_{ij} parameters and numerically estimates the corresponding means M_i and covariances $\Sigma_{ij} = \chi_{ij} - M_i M_j$. In contrast to CME, such simulation is prohibitively expensive for even moderately-sized, strongly-interacting networks. However, it enabled us to confirm the high accuracy of our model on our Lotka-Volterra simulations (see Fig. 5).

The METABRIC dataset

Microarray gene expression data for METABRIC were downloaded from the cBioPortal database [58, 59]. The METABRIC dataset, containing 1904 samples, is one of the most extensive publicly-available breast cancer studies [40]. We utilized microarray gene expression data containing 24368 genes from the 299 triple-negative samples.

Network identification

To quantify the (normalized) influence of genes relevant to triple-negative breast cancer, we utilized the method of network Markov chains [35–37]. The Human Protein Reference Database (HPRD) provides a curated interaction network of most human proteins [41]. Thus, to perform our analysis, we utilized the largest connected component, consisting of 3147 genes, obtained from the intersection of HPRD with the METABRIC gene list. We then performed network analysis as in [37] using the subset of 288 genes annotated in OncoKB, a curated database of prominent cancer genes [42].

Data preparation

For each sample, we obtain a measure of the relative influence of each of 288 genes. To identify potential drivers of the variability of these influences across the data, we computed their inter-sample Pearson correlations. We identified two distinct clusters of highly correlated genes: one containing a small number of immune-adjacent genes and the other, a

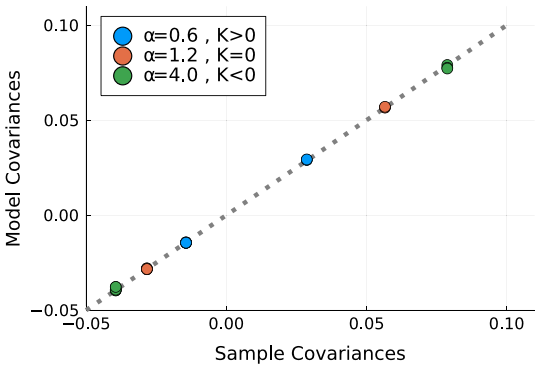

Fig. 5 Comparison of CME model covariances to the sample covariances of the cLV model. We observe complete agreement between our model and the data (see Figure 2), confirming the correctness of our maximum entropy fitting algorithm. The means, not shown, were all equal to 1/3 as expected

much larger component, containing prominent breast cancer genes such as TP53 and BRCA1. Thus, we utilized only this second component for our analysis.

Our primary goal is to identify genes and their interactions that potentially drive the variability in treatment responses observed in triple-negative breast cancer [39]. Likely genes include only those with large influence and inter-subject variability. Upon computing the variance in the influence of each gene, we found 17 candidates with markedly higher variance than the remaining bulk. We thus renormalized node influence across these 17 prime candidates before performing our maximum entropy analysis.

Implementation of the logit-normal distribution

An alternative to CME, the logit-normal distribution is given by [1]:

$$P_\Gamma = Z_{LN}^{-1} \frac{1}{\prod_{i=1}^{N} s_i} e^{-\frac{1}{2}\left\{ \log\left(\frac{s_{\tilde{N}}}{s_N}\right) - M_{LN} \right\}^\top \Sigma_{LN}^{-1} \left\{ \log\left(\frac{s_{\tilde{N}}}{s_N}\right) - M_{LN} \right\}} \tag{10}$$

where M_{LN} and Σ_{LN} are the means and covariances of the transformed data: $y = \left[\log(\frac{s_1}{s_N}), \ldots, \log(\frac{s_{N-1}}{s_N}) \right]$. Here, the feature of interest is the precision matrix $K_{LN}^* = -\Sigma_{LN}^{-1}$ which, under fairly general circumstances, has been shown to approximate maximum entropy interactions [20]. As with CME, we then utilized Eq 8 to define symmetric interactions between all pairs of nodes rather than simply the first $N-1$.

Acknowledgements
We acknowledge our use of Julia software and MATLAB software. The results are in part based upon data derived from TCGA database. We appreciate the platforms and the authors who uploaded their data. CW also acknowledges the Marie-Josée Kravis Fellowship in Quantitative Biology.

Author contributions
CW: Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Writing - original draft, Writing - review & editing. JZ: Formal analysis, Investigation, Software, Writing - original draft, Writing - review & editing. JD: Funding acquisition, Project administration, Resources, Supervision, Writing - review & editing. AT: Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Supervision, Writing - review & editing. All authors read and approved the final manuscript.

Funding
C.W. is funded by the Marie-Josée Kravis Fellowship in Quantitative Biology. A.R.T. is funded by AFOSR grants FA9550-17-1-0435, FA9550-20-1-0029, NIH grant R01AT011419, ARO grant W911NF2210292, and a grant from the Cure Alzheimer's Foundation. J.O.D. is funded by NIH/NCI Cancer Center Support grant P30 CA008748. J.O.D. and A.R.T. are funded by Breast Cancer Research Foundation Grant BCRF-17-193. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Data availibility
The dataset and code supporting the conclusions of this article are available at Github Git repository: https://github.com/Corey651/Compositional_Maximum_Entropy.

Declarations

Ethical approval and consent to participate
Not applicable.

Consent to Publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Received: 18 August 2022 Accepted: 26 October 2022
Published online: 29 October 2022

References

1. Aitchison J. The statistical analysis of compositional data. J Roy Stat Soc Ser B (Methodol). 1982;44(2):139–60.
2. Greenacre M. Compositional data analysis. Annual Rev Stat Appl. 2021;8:271–99.
3. Barceló-Vidal C, Martín-Fernández J-A. The mathematics of compositional analysis. Austrian J Stat. 2016;45(4):57–71.
4. Billheimer D, Guttorp P, Fagan WF. Statistical interpretation of species composition. J Am Stat Assoc. 2001;96(456):1205–14.
5. Pawlowsky-Glahn V, Egozcue JJ. Geometric approach to statistical analysis on the simplex. Stoch Env Res Risk Assess. 2001;15(5):384–98.
6. Pearson K. Mathematical contributions to the theory of evolution.-on a form of spurious correlation which may arise when indices are used in the measurement of organs. In: Proceedings of the royal society of london 1897;60(359-367):489–498.
7. Carr A, Diener C, Baliga NS, Gibbons SM. Use and abuse of correlation analyses in microbial ecology. ISME J. 2019;13(11):2647–55.
8. Gloor GB, Macklaim JM, Pawlowsky-Glahn V, Egozcue JJ. Microbiome datasets are compositional: and this is not optional. Front Microbiol. 2017;8:2224.
9. Greenacre M, Lewi P. Distributional equivalence and subcompositional coherence in the analysis of compositional data, contingency tables and ratio-scale measurements. J Classif. 2009;26(1):29–54.
10. Weiss S, Xu ZZ, Peddada S, Amir A, Bittinger K, Gonzalez A, Lozupone C, Zaneveld JR, Vázquez-Baeza Y, Birmingham A, et al. Normalization and microbial differential abundance strategies depend upon data characteristics. Microbiome. 2017;5(1):1–18.
11. Calle ML. Statistical analysis of metagenomics data. Genom Inf. 2019;17(1).
12. Pawlowsky-Glahn V, Buccianti A. Compos Data Anal. New York: Wiley Online Library; 2011.
13. Scealy JL, Wood AT. Score matching for compositional distributions. J Am Stat Assoc. 2022. https://doi.org/10.1080/01621459.2021.2016422.
14. Ongaro A, Migliorati S, Ascari R. A new mixture model on the simplex. Stat Comput. 2020;30(4):749–70.
15. Jaynes E, et al. The maximum entropy formalism. In: Levine RD, Tribus M, editors. Where do we stand 1979.
16. Jaynes ET. On the rationale of maximum-entropy methods. Proc IEEE. 1982;70(9):939–52.
17. Pressé S, Ghosh K, Lee J, Dill KA. Principles of maximum entropy and maximum caliber in statistical physics. Rev Mod Phys. 2013;85(3):1115.
18. Dixit PD, Wagoner J, Weistuch C, Pressé S, Ghosh K, Dill KA. Perspective: maximum caliber is a general variational principle for dynamical systems. J Chem Phys. 2018;148(1):010901.
19. Shore J, Johnson R. Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy. IEEE Trans Inf Theory. 1980;26(1):26–37.
20. Weistuch C, Agozzino L, Mujica-Parodi LR, Dill KA. Inferring a network from dynamical signals at its nodes. PLoS Comput Biol. 2020;16(11):1008435.
21. Weistuch C, Mujica-Parodi LR, Razban RM, Antal B, van Nieuwenhuizen H, Amgalan A, Dill KA. Metabolism modulates network synchrony in the aging brain. In: Proceedings of the national academy of sciences. 2021;118(40).
22. Weistuch C, Mujica-Parodi LR, Dill K. The refractory period matters: unifying mechanisms of macroscopic brain waves. Neural Comput. 2021;33(5):1145–63.
23. Schneidman E, Berry MJ, Segev R, Bialek W. Weak pairwise correlations imply strongly correlated network states in a neural population. Nature. 2006;440(7087):1007–12.
24. Morcos F, Pagnani A, Lunt B, Bertolino A, Marks DS, Sander C, Zecchina R, Onuchic JN, Hwa T, Weigt M. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. Proc Natl Acad Sci. 2011;108(49):1293–301.
25. Ghosh K, Dixit PD, Agozzino L, Dill KA. The maximum caliber variational principle for nonequilibria. Annu Rev Phys Chem. 2020;71:213–38.
26. Merchan L, Nemenman I. On the sufficiency of pairwise interactions in maximum entropy models of networks. J Stat Phys. 2016;162(5):1294–308.
27. Berry D, Widder S. Deciphering microbial interactions and detecting keystone species with co-occurrence networks. Front Microbiol. 2014;5:219.
28. Marasco A, Picucci A, Romano A. Market share dynamics using lotka-volterra models. Technol Forecast Soc Chang. 2016;105:49–62.
29. Stein RR, Bucci V, Toussaint NC, Buffie CG, Rätsch G, Pamer EG, Sander C, Xavier JB. Ecological modeling from time-series inference: insight into dynamics and stability of intestinal microbiota. PLoS Comput Biol. 2013;9(12):1003388.
30. Faust K, Raes J. Microbial interactions: from networks to models. Nat Rev Microbiol. 2012;10(8):538–50.
31. Egozcue JJ, Jarauta-Bragulat E. Differential models for evolutionary compositions. Math Geosci. 2014;46(4):381–410.
32. Mateu-Figueras G, Pawlowsky-Glahn V, Egozcue JJ. The normal distribution in some constrained sample spaces. arXiv preprint. 2008. arXiv:0802.2643.
33. Consortium S, et al. A comprehensive assessment of rna-seq accuracy, reproducibility and information content by the sequencing quality control consortium. Nat Biotechnol. 2014;32(9):903.
34. Paull EO, Aytes A, Jones SJ, Subramaniam PS, Giorgi FM, Douglass EF, Tagore S, Chu B, Vasciaveo A, Zheng S, et al. A modular master regulator landscape controls cancer transcriptional identity. Cell. 2021;184(2):334–51.
35. Sandhu R, Georgiou T, Reznik E, Zhu L, Kolesov I, Senbabaoglu Y, Tannenbaum A. Graph curvature for differentiating cancer networks. Sci Rep. 2015;5(1):1–13.
36. West J, Bianconi G, Severini S, Teschendorff AE. Differential network entropy reveals cancer system hallmarks. Sci Rep. 2012;2(1):1–8.
37. Zhu J, Oh JH, Deasy JO, Tannenbaum AR. vwcluster: vector-valued optimal transport for network based clustering using multi-omics data in breast cancer. PLoS One. 2022;17(3):0265150.
38. Chen Y, Cruz FD, Sandhu R, Kung AL, Mundi P, Deasy JO, Tannenbaum A. Pediatric sarcoma data forms a unique cluster measured via the earth mover's distance. Sci Rep. 2017;7(1):1–9.

39. Shah SP, Roth A, Goya R, Oloumi A, Ha G, Zhao Y, Turashvili G, Ding J, Tse K, Haffari G, et al. The clonal and mutational evolution spectrum of primary triple-negative breast cancers. Nature. 2012;486(7403):395–9.
40. Curtis C, Shah SP, Chin S-F, Turashvili G, Rueda OM, Dunning MJ, Speed D, Lynch AG, Samarajiwa S, Yuan Y, et al. The genomic and transcriptomic architecture of 2000 breast tumours reveals novel subgroups. Nature. 2012;486(7403):346–52.
41. Peri S, Navarro JD, Amanchy R, Kristiansen TZ, Jonnalagadda CK, Surendranath V, Niranjan V, Muthusamy B, Gandhi T, Gronborg M, et al. Development of human protein reference database as an initial platform for approaching systems biology in humans. Genome Res. 2003;13(10):2363–71.
42. Chakravarty D, Gao J, Phillips S, Kundra R, Zhang H, Wang J, Rudolph JE, Yaeger R, Soumerai T, Nissan MH, et al. Oncokb: a precision oncology knowledge base. JCO Precis Oncol. 2017;1:1–16.
43. Irby RB, Yeatman TJ. Role of src expression and activation in human cancer. Oncogene. 2000;19(49):5636–42.
44. Finlay CA, Hinds PW, Levine AJ. The p53 proto-oncogene can act as a suppressor of transformation. Cell. 1989;57(7):1083–93.
45. Niu G, Wright KL, Ma Y, Wright GM, Huang M, Irby R, Briggs J, Karras J, Cress WD, Pardoll D, et al. Role of stat3 in regulating p53 expression and function. Mol Cell Biol. 2005;25(17):7432–40.
46. Gao B, Shen X, Kunos G, Meng Q, Goldberg ID, Rosen EM, Fan S. Constitutive activation of jak-stat3 signaling by brca1 in human prostate cancer cells. FEBS Lett. 2001;488(3):179–84.
47. Liu X, Qu C-K. Protein tyrosine phosphatase shp-2 (ptpn11) in hematopoiesis and leukemogenesis. J Signal Transduct 2011;2011.
48. Baba K, Shibata R, Sibuya M. Partial correlation and conditional correlation as measures of conditional independence. Australian N Z J Stat. 2004;46(4):657–64.
49. Erb I. Partial correlations in compositional data analysis. Appl Comput Geosci. 2020;6:100026.
50. Williams DR, Rast P. Back to the basics: Rethinking partial correlation network methodology. Br J Math Stat Psychol. 2020;73(2):187–212.
51. Stein RR, Marks DS, Sander C. Inferring pairwise interactions from biological data using maximum-entropy probability models. PLoS Comput Biol. 2015;11(7):1004182.
52. Arnold BC, Sarabia JM. Conditional specification of statistical models: classical models, new developments and challenges. J Multivar Anal. 2022;188:104801.
53. Mo Q, Wang S, Seshan VE, Olshen AB, Schultz N, Sander C, Powers RS, Ladanyi M, Shen R. Pattern discovery and cancer gene identification in integrated cancer genomic data. Proc Natl Acad Sci. 2013;110(11):4245–50.
54. Mozeika A, Dikmen O, Piili J. Consistent inference of a general model using the pseudolikelihood method. Phys Rev E. 2014;90(1):010101.
55. Janžura M, Boček P. Relative asymptotic efficiency of the maximum pseudolikelihood estimate for gauss-markov random fields. Stat Infer Stoch Process. 2002;5(2):179–97.
56. Hyvarinen A. Connections between score matching, contrastive divergence, and pseudolikelihood for continuous-valued variables. IEEE Trans Neural Netw. 2007;18(5):1529–31.
57. Malouf R. A comparison of algorithms for maximum entropy parameter estimation. In: COLING-02: the 6th conference on natural language learning 2002 (CoNLL-2002) 2002.
58. Cerami E, Gao J, Dogrusoz U, Gross BE, Sumer SO, Aksoy BA, Jacobsen A, Byrne CJ, Heuer ML, Larsson E. et al. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. AACR 2012.
59. Gao J, Aksoy BA, Dogrusoz U, Dresdner G, Gross B, Sumer SO, Sun Y, Jacobsen A, Sinha R, Larsson E, et al. Integrative analysis of complex cancer genomics and clinical profiles using the cbioportal. Sci Signal. 2013;6(269):1–1.

Publisher's Note