PNAS RESEARCH ARTICLE | APPLIED PHYSICAL SCIENCES

Hyperedge prediction and the statistical mechanisms of higher-order and lower-order interactions in complex networks

Marta Sales-Pardo[a,1], Aleix Mariné-Tena[b], and Roger Guimerà[a,c]

Edited by Luís A. Amaral, Northwestern University, Evanston, IL; received March 8, 2023; accepted November 2, 2023 by Editorial Board Member Simon A. Levin

Complex networked systems often exhibit higher-order interactions, beyond dyadic interactions, which can dramatically alter their observed behavior. Consequently, understanding hypergraphs from a structural perspective has become increasingly important. Statistical, group-based inference approaches are well suited for unveiling the underlying community structure and predicting unobserved interactions. However, these approaches often rely on two key assumptions: that the same groups can explain hyperedges of any order and that interactions are assortative, meaning that edges are formed by nodes with the same group memberships. To test these assumptions, we propose a group-based generative model for hypergraphs that does not impose an assortative mechanism to explain observed higher-order interactions, unlike current approaches. Our model allows us to explore the validity of the assumptions. Our results indicate that the first assumption appears to hold true for real networks. However, the second assumption is not necessarily accurate; we find that a combination of general statistical mechanisms can explain observed hyperedges. Finally, with our approach, we are also able to determine the importance of lower and high-order interactions for predicting unobserved interactions. Our research challenges the conventional assumptions of group-based inference methodologies and broadens our understanding of the underlying structure of hypergraphs.

complex networks | higher-order interactions | stochastic block models | probabilistic inference | link prediction

Significance

Complex networked systems are influenced by interactions between multiple elements, known as hyperedges. Researchers have used inference methodologies to understand hypergraph structure, assuming that the statistical mechanism behind hyperedges is assortative, and relying on a specific set of groups to explain interactions. However, our results show that these assumptions are not always accurate. The same groups of nodes can explain hyperedges of different sizes, but interactions between nodes are not always assortative. Nonassortative patterns can also explain some observed hyperedges. This research challenges existing inference methodologies and broadens understanding of complex networked systems, encouraging the development of new approaches to studying hypergraphs.

In a networked system, interactions are often represented as relationships between pairs of units within the system. However, this representation is in many cases inaccurate and fails to fully represent the complexity of the interactions. In fact, units in the system often participate in so-called higher-order or polyadic interactions that involve more than two units. This is the case of, for instance, collaboration networks in which a team of researchers coauthor papers together, groups of enzymes that form protein complexes to perform a function within the cell, or substances combined into drugs approved for medical use (1, 2). The fact that higher-order interactions are commonplace in natural networked systems makes us question how to reassess previous findings on dyadic networks. For example, higher-order interactions have been suggested to be at the root of the high heterogeneity observed in the density of many real-world networks (3). Moreover, recent studies also show that considering higher-order interactions can have quite dramatic effects in dynamical processes occurring on networks such as the spread of epidemics and misinformation, or synchronization (4–7).

Over the last decade, network scientists have been very successful at developing a set of inference methodologies with which to model the large-scale organization of complex networks (the so-called community detection problem) (8); these methodologies have been proven useful to predict unobserved interactions and detect errors (9–11) and for network reconstruction (9, 12, 13). However, the focus so far has been on dyadic interactions, and only recently the corresponding inference framework has started to emerge for higher-order interactions. In particular, recent works have looked into how inference can help reconstruct higher-order networks based on the heterogeneities observed in dyadic networks (13); how to obtain communities of nodes using generative models for higher-order interactions (14–16); and how inference models can help in the recovery of unobserved hyperedges (17). Importantly, current approaches focus on the limit of almost complete knowledge of the hypergraph, an assumption that is not valid in empirical biological sciences in which experiments are costly and therefore a small percentage of interactions can actually be assessed (18, 19).

Here, we want to further investigate the problem of predicting unobserved interactions, a problem that is closely related to the problem of network reconstruction when we have

Author affiliations: [a]Department of Chemical Engineering, Universitat Rovira i Virgili, Tarragona E-43007, Spain; [b]Catalan Institute of Chemical Research, Tarragona E-43007, Spain; and [c]Institució Catalana de Recerca i Estudis Avançats, Barcelona E-08010, Spain

Author contributions: M.S.-P. and R.G. designed research; M.S.-P. and A.M.-T. performed research; M.S.-P. and A.M.-T. analyzed data; and M.S.-P. and R.G. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission. L.A.A. is a guest editor invited by the Editorial Board.

[1]To whom correspondence may be addressed. Email: marta.sales@urv.cat.

Published December 7, 2023.

a partial observation of a network with higher-order interactions (9, 12). Understanding how inference approaches can help in prediction tasks is a necessary step toward gaining confidence about the validity of the assumptions behind group-based generative models that we can use for clustering and hyperedge prediction (such as in refs. 14, 15, 17, and 20) and also to understand the limitations of these models (21).

We address the problem of predicting unobserved interactions between tuples of m nodes (m-hyperedges) given a set of observed m-hyperedges and assess up to which point observed interactions between tuples of $n(\neq m)$ nodes are useful to predict m-hyperedges. Specifically, as it is commonplace (14, 15), we assume that there is an underlying group structure that explains observed hyperedges; our focus is then to predict interactions between tuples of nodes of fixed size m by considering extra information that might be useful in the predictive task (that is, interactions between tuples of a different size n) in a similar way to when we include node attributes in the inference process (22, 23). To this end, we introduce a mixed-membership stochastic block model for hypergraphs and its corresponding inference equations. This simple model allows us to explore in a straightforward manner the extent up to which hyperedges of different sizes inform one another and help make predictions of unobserved data, and in turn, whether a unique set of group memberships can be used to explain hyperedges of any size.

We find that only in the cases in which interactions have been generated with similar underlying group structures, hyperedges of size m are informative about the existence of unobserved hyperedges of size n, and that this seems to be the case for real data as well. Remarkably, lower-order interactions typically carry more information than higher-order interactions, which highlights the importance of reliably measuring low-order interactions for properly reconstructing hyperedges involving a large number of nodes. Additionally, because our model does not assume group assortativity in the interactions, we can also explore what types of patterns better explain the interactions we observe. Our results for real data show that a group-assortative mechanism is not the only pattern of interaction that explains hyperedges.

Mixed-Membership Stochastic Block Models for Higher-Order Interactions

We start by considering a Bernoulli mixed-membership stochastic block model (SBM) (24, 25) for unipartite, undirected networks with binary interactions between pairs of nodes. Each element in the dyadic adjacency matrix $A^{(2)}$ can take two possible values $A^{(2)}_{ij} \in \{0, 1\}$. Here and in what follows, the superindex of the adjacency matrix indicates the size of the hyperedges we consider.

A mixed-membership SBM assumes that there are K underlying groups of nodes, and that each node i has probability $\theta_{i\alpha}$ of belonging to node group $\alpha = 1, \ldots, K$ with $\sum_\alpha \theta_{i\alpha} = 1$. The SBM further assumes that edges are conditionally independent and that there is a matrix $\boldsymbol{q}^{(2)}$ that expresses the probability that pairs of nodes interact given their group memberships. The probability that $A^{(2)}_{ij} = 1$ is then expressed as

$$p\left(A^{(2)}_{ij} = 1|\boldsymbol{\theta}, \boldsymbol{q}^{(2)}\right) = \sum_{\alpha,\beta} \theta_{i\alpha}\, \theta_{j\beta}\, q^{(2)}_{\alpha\beta}, \qquad [1]$$

where $q^{(2)}_{\alpha\beta}$ is the probability that a node that belongs to group α interacts with a node that belongs to group β, and $p(A^{(2)}_{ij} = 0|\boldsymbol{\theta}, \boldsymbol{q}^{(2)}) = 1 - p(A^{(2)}_{ij} = 1|\boldsymbol{\theta}, \boldsymbol{q}^{(2)})$. Note that if the order of the nodes matters (such as in directed networks), then we would consider tuples instead of sets of nodes. In order to generalize our model for tuples, one would have to consider different group membership vectors depending on the position of the node in the tuple. For instance, for tuples of size two (or directed edges), this would be equivalent to having different membership vectors for incoming and outgoing edges.

If we have higher-order interactions such as those involving sets of three nodes (3-node interactions), then we can make the exact same assumptions, namely, that there are K underlying groups for nodes, and that the probability that node i participates in group α is $\theta_{i\alpha}$. We can then use a tensorial SBM (11, 26) to express the probability that nodes (i, j, k) interact (that is, that $A^{(3)}_{ijk} = 1$) as

$$p\left(A^{(3)}_{ijk} = 1|\boldsymbol{\theta}, \boldsymbol{q}^{(3)}\right) = \sum_{\alpha,\beta,\gamma} \theta_{i\alpha}\, \theta_{j\beta}\, \theta_{k\gamma}\, q^{(3)}_{\alpha\beta\gamma}, \qquad [2]$$

where $q^{(3)}_{\alpha\beta\gamma}$ is the probability that three nodes that belong to groups α, β and γ, respectively, interact. As before, $p(A^{(3)}_{ijk} = 0|\boldsymbol{\theta}, \boldsymbol{q}^{(3)}) = 1 - p(A^{(3)}_{ijk} = 1|\boldsymbol{\theta}, \boldsymbol{q}^{(3)})$.

Note that, using this approach, it is straightforward to model hyperedges of any size, so that for interactions involving sets of n nodes ($A^{(n)}_{i_1\ldots i_n}$), we would have to consider an n-dimensional connection probability tensor $q^{(n)}_{\alpha_1\ldots\alpha_n}$ (Fig. 1).

In the same spirit as other generative models for hypergraph and tensorial models (11, 14, 15, 26), our model assumes that the probability that a set of nodes interact depends only on the group memberships of the nodes. Because in an SBM the probabilities that hyperedges of size n exist are conditionally independent once $\boldsymbol{\theta}$ and $\boldsymbol{q}^{(n)}$ are fixed, the likelihood of all observed hyperedges of size n, $\left(\mathbf{A^0}\right)^{(n)}$, is the product of probabilities for each of the observed interactions

$$L\left(\left(\mathbf{A^O}\right)^{(n)}|\boldsymbol{\theta}, \boldsymbol{q}^{(n)}\right) = \prod_{(i_1\ldots i_n)\in\left(\mathbf{A}^O\right)^{(n)}} p\left(\left(A^O\right)^{(n)}_{ij}|\boldsymbol{\theta}, \boldsymbol{q}^{(n)}\right). \qquad [3]$$

Importantly, this approach allows us to model simultaneously hyperedges of different sizes considering group-based mechanisms that can vary for every hyperedge size. For instance, we could have an assortative mechanism for 2-node interactions and a disassortative mechanism for 3-node interactions. If we assume that the underlying K groups of nodes can explain the observed hyperedges of any size, the likelihood of the observed interactions $\mathbf{A^O} := \{\left(\mathbf{A^O}\right)^{(2)}, \ldots, \left(\mathbf{A^O}\right)^{(n)}\}$ given the model parameters $\{\boldsymbol{\theta}, \boldsymbol{q}^{(2)}, \ldots, \boldsymbol{q}^{(n)}\} \equiv \{\boldsymbol{\theta}, \boldsymbol{q}\}$ is then the product of the likelihoods of the observed set of hyperedges of all sizes is

$$L\left(\mathbf{A^O}|\boldsymbol{\theta}, \boldsymbol{q}\right) = \prod_n L\left(\left(\mathbf{A^0}\right)^{(n)}|\boldsymbol{\theta}, \boldsymbol{q}^{(n)}\right), \qquad [4]$$

where the product is over all observed hyperedge sizes.

If we assume flat priors over model parameters, then the log-posterior probability of model parameters $\{\boldsymbol{\theta}, \boldsymbol{q}\}$ given the observed data is (up to a normalizing constant)

$$\log P(\{\boldsymbol{\theta}, \boldsymbol{q}\}|\mathbf{A^O}) = \sum_n \log L\left(\left(\mathbf{A^0}\right)^{(n)}|\boldsymbol{\theta}, \boldsymbol{q}^{(n)}\right), \qquad [5]$$

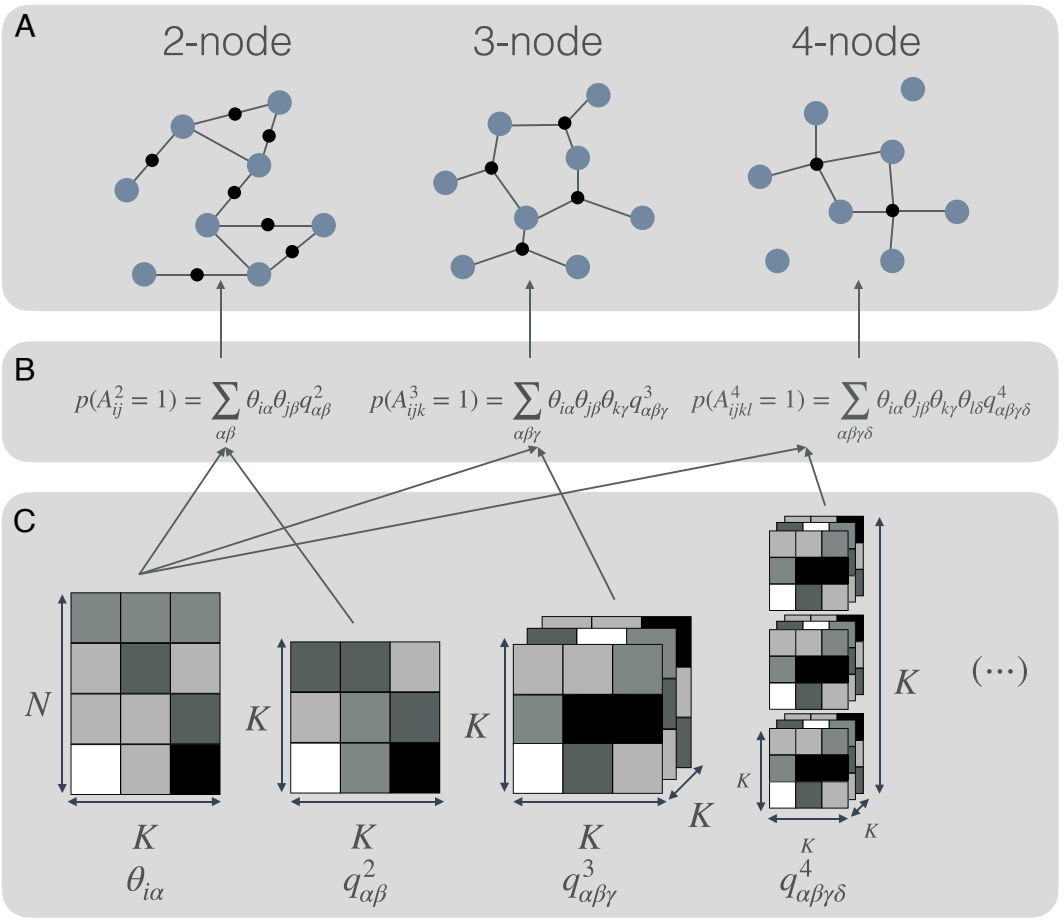

Fig. 1. A mixed-membership stochastic block model (SBM) for higher-order interactions. Given N nodes, we have hyperedges involving a different number of nodes. (*A*) Hyperedges (black circles) involving pairs, triplets, and quadruplets of nodes (blue circles) (*B*) In a mixed-membership SBM for hypergraphs, the probability of observing each n-hyperedge (or n-node interaction) depends on: i) the group memberships vectors of the nodes $\boldsymbol{\theta}$ illustrated for $K = 3$; and ii) a n-tensor $\boldsymbol{q}^{(n)}$ of connection probabilities. Each tensor element $q^{(n)}_{\alpha_1\ldots\alpha_n}$ describes the probability that n nodes that belong to groups $\{\alpha_1 \ldots \alpha_n\}$ interact. We assume that the group memberships of nodes are the same for the different interaction orders, which allows to use observed hyperedge of one type to predict hyperedges of another type. (*C*) Membership vectors and tensors for $K = 3$ node groups.

which is amenable to using an expectation-maximization algorithm to find local maxima of the posterior landscape. We can then average over maxima in the posterior landscape to make predictions, as shown in refs. 11, 22, and 25 (see *Data and Methods* for details).

Limitations for Large Hyperedge Sizes. For a fixed number of groups K, the number of elements in the connection probability tensor grows with hyperedge size n as K^n. This poses a problem to the usability of our approach, since even for small values of K and n, it would not be possible to estimate reliably all tensor elements given finite observed data. However, the majority of hypergraph data are not directed, which means that tensors must be symmetric under index permutations (for instance for hyperedges of size $n = 2$ we have that: $q^{(2)}_{ij} = q^{(2)}_{ji}$). This limits the number of unique tensor elements, which for large n become a vanishingly small fraction of the total number of tensor elements. For instance, for $K = 6$, the number of unique tensor elements for hyperedge size $n = 7$ is 270,036, but the number of unique index combinations is 720 (0.26%). The limitation for very large hyperedge sizes still stands, but for practical analyses in which hyperedges of large sizes are a vanishingly small fraction of the hyperedge space, our approach can still be used on the majority of the data with a substantial advantage over assortative approaches as our analysis for real data shows.

Related Work

Hypergraph Stochastic Block Models for Clustering and Community Detection. There is some literature addressing the problem of community detection in hypergraphs (see for instance ref. 2). Most relevant here are works that use hypergraph stochastic block models (HSMB) to obtain communities or clusters of nodes in hypergraphs (14, 15, 20, 27–29). These approaches consider that nodes can belong to only one group of nodes. Additionally, they focus on situations in which we have block-diagonal SBMs, that is, on models in which the underlying block structure is assortative (or, if we exchange edges and nonedges, fully disassortative), so that it is more probable to have hyperedges among nodes in the same group. Recent work shows that assortativity can be formalised in different manners in HSBMs, resulting in different groupings of nodes (15).

By contrast, in our mixed-membership approach, nodes have a finite probability of belonging to each one of the groups, which makes the model more expressive. Additionally, our model can account for different group mechanisms generating hyperedges of different sizes, as ref. 16 shows for the case of weighted

hypergraphs. This means we do not impose specific constraints on the probability tensor entries, and therefore we do not need to assume full-assortativity (or full disassortativity). Therefore, our formulation could capture situations in which it is not necessary for nodes to belong to the same groups (or have similar membership vector profiles) to interact, or situations in which this happens for some types of hyperedges but not others. Arguably, the approach used in ref. 15 allows, with a smaller number of parameters, to model hypergraphs with interactions of any size. However, for hypergraphs with a large number of interactions of smaller size, a more flexible approach like the one we propose could provide a better description of the data.

Finally, we note that, while in terms of model selection single group memberships are easier to interpret, our goal here is to make predictions of unobserved data, which requires averaging over models (9, 10). Previous work shows that mixed-membership models are, in general, more expressive than single-membership ones and make better predictions of unobserved edges (25).

Prediction of Higher-Order Interactions. Approaches to hyperedge prediction can be divided in roughly two classes: structural and mechanistic approaches. Structural approaches use topological information to establish how likely it is that an unobserved hyperedge exists in the hypergraph. Local structural approaches assume that there is an underlying assortative mechanism that explains hyperedge formation. Then, these approaches use the topology of the hypergraph to establish similarities between nodes and use this similarity to estimate the probability that a set of nodes participate in the same hyperedge. To make predictions of unobserved hyperedges, these approaches use methods that range from adapted common neighbors or Katz centrality metrics to assess similarity between nodes (29), to more sophisticated metrics such as resource allocation (30), or even consider a wide array of topological features of hyperedges to train a binary classifier (31). Global structural approaches focus on overall properties of the adjacency tensor to predict unobserved hyperedges; these approaches include matrix factorization-based inferential approaches (29) and spectral approaches (32). Finally, very recent work uses an inferential approach based on a Poisson formulation of stochastic block models to predict hyperedges of any size (17).

Mechanistic approaches focus on the creation of higher-order hyperedges (or depending on the work, simplices or motifs) from lower-order hyperedges assuming specific closure mechanisms (33, 34). These approaches are better suited, in general, to describe the time evolution of a hypergraph. However, they are not ideal when we have a single, incomplete observation of a hypergraph that we want to reconstruct, because the proposed mechanisms have a temporal dimension.

Structural approaches are, in general, better suited for this task as the goal is to find the topological or statistical regularities that help predict unobserved interactions. Nonetheless, it remains to be explored whether regularities are also useful to make predictions of the future evolution of networks as in the case of dyadic interactions (35–38).

Our approach is a structural global approach in which we use an inference approach to uncover the statistical regularities of the whole adjacency tensor. Therefore, it is more closely related to the tensor-factorization inferential approach presented in ref. 29, and to the inferential approach in ref. 17. Indeed, the latter also considers a mixed membership approach albeit in a Poisson formulation which is better suited the weighted graphs they study as in ref. 16; however, in contrast to ours, the tensor that determines connection rates between groups is considered to be diagonal, that is, it assumes an assortative mechanism as in ref. 15 or, in terms of mixed-membership models, follows a matrix-factorization-like approach as in ref. 29.

In the context of recommendation systems, we have shown that mixed-membership SBMs formally are more general models than those implicit in matrix factorization; in mixed-membership SBMs, nodes with similar latent representation are not forced to be connected and, conversely, dissimilar nodes can be connected (25). As a result, mixed-membership SBMs are more flexible to explore different patterns of interaction and more predictive in a number of contexts (25, 39).

Our primary focus here is on showing up to what extent lower-order interactions are informative of higher-order interactions and vice versa, and to which extent standard assumptions in the modeling of higher-order networks are valid. In doing so, however, we are also able to show that because of its expressibility, and despite its scalability limitations for very large hyperedge sizes, our approach outperforms assortative approaches for hyperedge prediction when observations are limited.

Results

Synthetic Data. First, we want to assess how informative are n-node interactions for the prediction of m-node interactions. To that end, we generate synthetic hypergraphs with interactions involving pairs and triplets of nodes. To generate these graphs, we use a mixed-membership stochastic block model in which the membership vectors of nodes are the same for all types of interactions, as described above (see *Materials and Methods* for details). We then look at the performance of our approach at predicting unobserved hyperedges when we consider only one type of interaction and when we consider both.

Fig. 2 shows that, as expected, as we increase the number of observed interactions of the same type, the model makes better predictions of unobserved hyperedges and reaches optimal predictive power.*

We find differences between the impact that 2-node interactions have in the prediction of 3-node interactions, and the impact that 3-node interactions have in the prediction of 2-node interactions. While adding 2/3-node interactions to small training sets of 3/2-node interactions boosts predictive accuracy, in this case, 2-node interactions carry more information and result in a larger increase in predictive accuracy, due to the larger density of positive interactions (Fig. 2 *A* and *B*). Our results show that, for the task of predicting out-of-sample data, including any order of interactions in the training set helps in the inference process and to predict unobserved data and reconstruct the network.

We can look at this result from the perspective of an issue that has raised a lot of interest recently, namely, assessing the importance of additional information (for example, node attributes or metadata) for elucidating the underlying community structure of networks (23, 40) or for link prediction (22). In link prediction, additional information can be detrimental if node attributes are not correlated with observed network structure. Here, we face a similar problem. In general, generative models to predict higher-order interactions rely on the fact that the same underlying groups of nodes and interaction mechanisms are responsible for all the interactions we observe between nodes. Indeed, this seems a reasonable assumption, but one could

*Note that in our synthetic hypergraphs, the average density for the model parameters we use is of 50% and 30% for 2- and 3-node interactions, respectively. Therefore, neither the optimal predictive performance for each prediction task nor the size of the training set necessary to achieve it are the same in both predictive tasks.

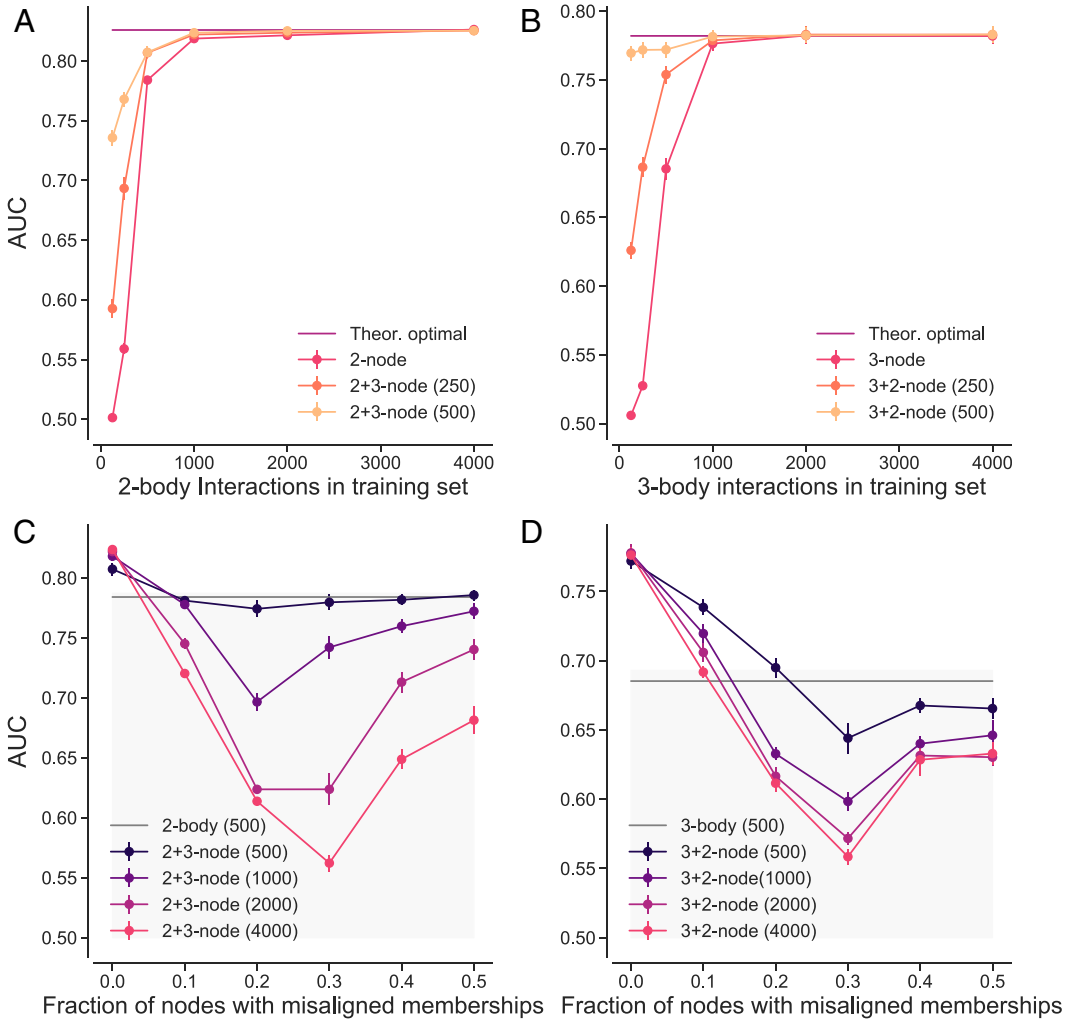

Fig. 2. Effect of 3-node interactions in the prediction of 2-node interactions and vice-versa. (*A* and *B*) We generate synthetic hypergraphs using a mixed-membership stochastic block model with $K = 2$ (see Text and *Materials and Methods* for details) with 2-node and 3-node interactions. Node membership vectors are the same for 2-node and 3-node interactions. (*A*) We compare the performance at predicting unobserved 2-node interactions when we observe only 2-node interactions, and when we observe 2-node interactions and additional 3-node interactions (250, 500). We use the AUC as a metric, that is, the probability that if we pick a noninteraction and an interaction at random, the model is able to properly classify them. (*B*) We compare the performance at predicting unobserved 3-node hyperedges when we observe only 3-node interactions, and when we observe additional 2-node interactions (250, 500). (*C* and *D*) We generate synthetic hypergraphs with 2-node and 3-node interactions using a mixed-membership stochastic block model with $K = 2$ in which nodes can have two possible membership vectors (see *Materials and Methods* for details). With probability f, we consider that a node has different group memberships to generate 2-node and 3-node interactions, so that there is a fraction f of nodes with misaligned group memberships in the hypergraphs we generate. (*C*) We show the AUC to compare the ability of the model of predicting unobserved 2-node interactions when we observe 500 2-node interactions and 500, 1,000, 2,000, and 4,000 3-node interactions with different fractions of nodes with misaligned group membership vectors. The gray line corresponds to the average AUC obtained for 500 observations of 2-node interactions in the training set. The gray shaded area shows AUC values that are within a SD or lower than the reference value. (*D*) Same as *C*, exchanging 2- and 3-node interactions. In all plots, each point corresponds to an average over 10 different hypergraphs and 10 different training-test set combinations. For each one of these, we perform 25 expectation maximization runs with $K = 2$ to make predictions (*Materials and Methods*). Error bars correspond to the SE of the mean. The test size is always 1,000 2-node or 3-node hyperedges.

think of situations in which different group memberships or mechanisms are responsible for different types of hyperedges. For instance, if we investigate the effect of simultaneous gene mutations on a specific phenotype, it could be that higher-order effects are the result of the interaction of different pathways that do not take place in a small number of genes. Because genes can have an effect on multiple pathways it could well be that the grouping that explains both kinds of interactions are not necessarily the same. The question is, then, what is the effect on the prediction of unobserved interactions (and therefore in network reconstruction) of the assumption that a unique group structure can explain all of the interactions among sets of n-nodes.

To address this question, we again generate synthetic networks with 2-node and 3-node interactions using a mixed membership stochastic block model (*Materials and Methods*). We then consider the case in which, for a fraction of the nodes, group memberships generating 2-body and 3-body interactions are different, and therefore have misaligned memberships (Fig. 2 *C* and *D*; *Materials and Methods*). We find that, when we observe a small number of (2-body or 3-body) interactions, including information of interactions generated by a different set of group memberships decreases predictability in both cases (Fig. 2 *C* and *D*). Indeed, even when the fraction of nodes with misaligned memberships is as low as 20%, predictability falls even when only a few 2-body/3-body interactions are observed. We also find that the effect is not symmetric. In our synthetic data, with 500 2-node observations, we get an accuracy that is close to the theoretical maximum, so that the improvement introduced by considering

3-node interactions is modest. Any misalignment in the 3-node interactions is likely to erase that contribution and lower accuracy (gray area in Fig. 2*C*). By contrast, for the prediction of 3-node interactions we see a larger difference between observing only 3-node interactions or 3-node and 2-node interactions. Therefore, more misinformation is necessary for the same negative effect (larger fraction of misalignment or a larger number of observed 2-node interactions).

Summarizing, our approach allows one to leverage higher-order interactions to predict lower-order interactions and vice-versa. At the same time, it shows that if the mechanisms leading to 2- and 3-node interactions are different, then mixing them can have negative effects on link prediction. Therefore, by measuring the effect of such combinations of 2- and 3-node interactions in real data, it is possible to elucidate whether interactions of different orders in a given network are likely to have been generated with the same underlying group structure or not.

Real Data. To illustrate the use of our approach with real data in the case in which the amount of observations is limited, we consider two different datasets. First, we consider a drug substance dataset collected in ref. 33, from which we use the information about drugs that have at most five substances in their composition (see *Data* for details). We represent drugs that have n different substances in their composition as n-substance interactions in the substance network.

Second, we consider a dataset on the effect of gene knockouts on cell growth for the yeast *Saccharomyces cerevisiae*. In particular, the dataset contains knockouts of 410,399 pairs and 91,111 triplets of genes (18) (see *Data* for details). From these data, we represent double mutants that present significant alterations of the expected growth as 2-gene (digenic) interactions (9,363 in total) and triple mutants that result in significantly negative differential growth as 3-gene (trigenic) interactions (3,196 in total; see *Data* and ref. 18 for details).

In Figs. 3 *A* and *B* and 4, we show that, in both cases, our approach is able to make predictions for all hyperedge sizes we investigate. Our results also show that 3-node interactions hold relevant information that helps predict 2-node interactions and, conversely, that 2-node interactions hold relevant information that helps predict 3-node interactions. Thus, these results indicate that for these two case studies, the underlying groups for both types of hyperedge are the same, since otherwise we would expect to see a decline in predictive performance, rather than an increase. Results for higher-order interactions in Fig. 3 *C* and *D* confirm this result and also confirm that our algorithm can also be applied to predict larger-order interactions even when the number of observed hyperedges is low.

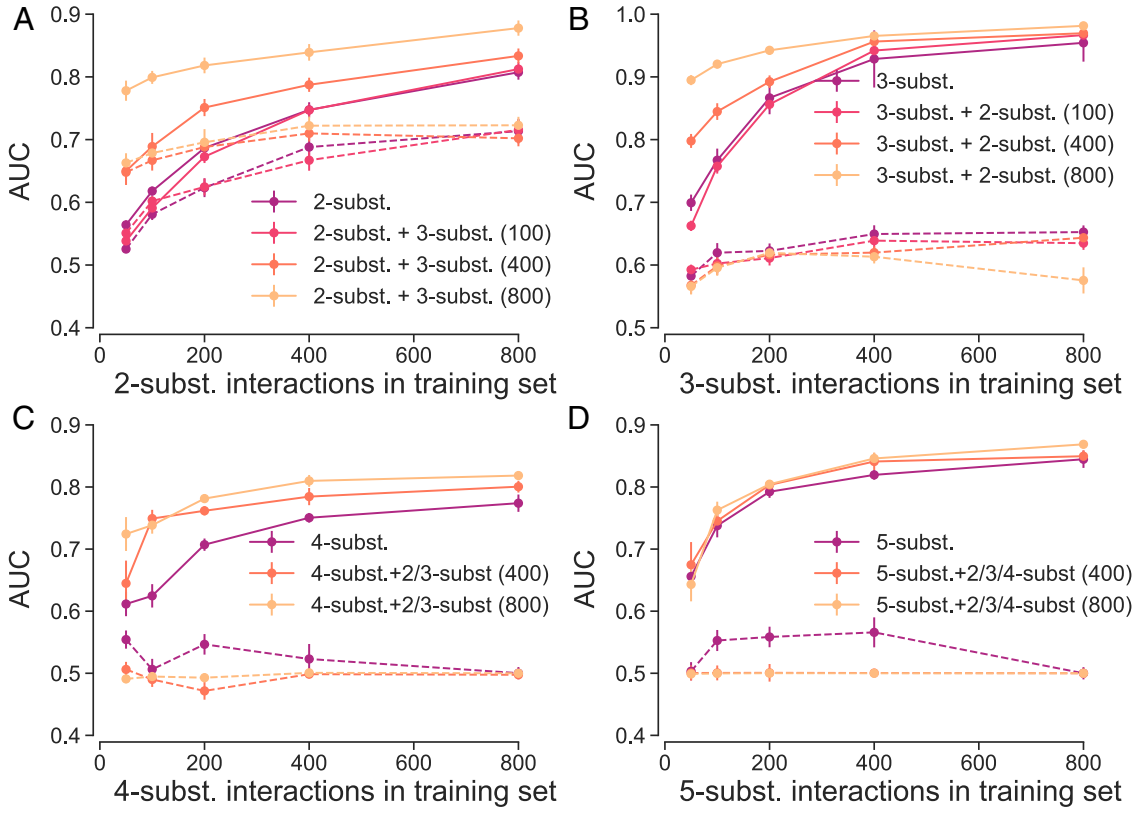

Fig. 3. Prediction of the co-occurrence of substances in drugs. We represent each drug as a hyperedge that connects the substances within that drug. We perform different prediction experiments to predict unobserved 2-substance (*A*), 3-substance (*B*), 4-substance (*C*), and 5-substance (*D*) drugs (see *Materials and Methods* for data description). In each plot, we consider predictions when the training set comprises only one type of hyperedge, and by adding an equal number of lower- or higher-order hyperedges to the training set as indicated by the legend in each plot (solid lines). The # of interactions in the training set comprises the total number of hyperedges (0 and 1) that are observed. The rest of hyperedges are not observed and therefore are not taken into account in the likelihood. Note that as we increase the training set size, the prediction ability increases. Including information of hyperedges of a different order also increases accuracy. Discontinuous lines correspond to predictions using a mixed-membership Bernoulli model in which connection probability tensors are diagonal. Lines are colored according to the legend in each panel. See *Materials and Methods* for details on how we make predictions. Results for our approach are for $K = 6$; results for the assortative model are for $K = 8$ (see *Materials and Methods* for details). Error bars (some smaller than the symbols) show the cross-validation SEM.

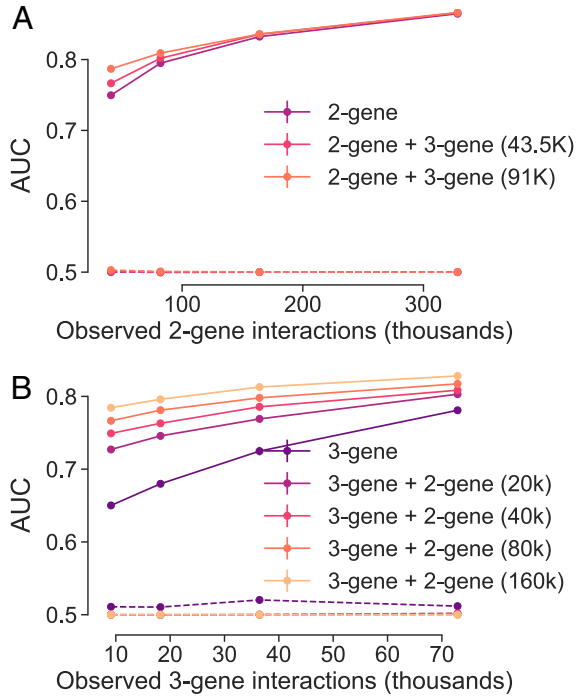

Fig. 4. Prediction of gene interactions. (*A*) Prediction of interactions between pairs of genes. Out of the 410,399 digenic interactions available, we perform cross-validation experiments in which we consider five test sets with 20% of available interactions and training sets with a percentage $p \in [10\%, 80\%]$ of the remaining 2-gene interactions. We show the AUC of the held-out 2-gene interactions versus the number of 2-gene interactions in the training set. We also show the AUC for the case in which we add trigenic interactions to each training set (50% and 100% of the 91,111 available trigenic interactions in the dataset). Note that for the same fraction of 2-gene interactions in the training set, adding 3-gene interactions improves prediction accuracy, especially, for small training set sizes. (*B*) Prediction of interactions between triplets of genes. We perform the same experiments as in *A* exchanging 2-gene and 3-gene interactions. We consider training sets with a percentage $p \in [10\%, 80\%]$ of 3-gene interactions and additional 2-gene interactions (between 20K and 160K). Note that for the same fraction of 2-gene interactions in the training set, adding 3-gene interactions improves prediction accuracy, especially for small training set sizes. Discontinuous lines correspond to predictions using a mixed-membership Bernoulli model in which connection probability tensors are diagonal. Results for 2-gene and 2,3-gene interactions combined are for $K = 6$; results for 3-gene interactions are for $K = 2$ (see *Materials and Methods* for details); results for the assortative model are for $K = 8$. Error bars (some smaller than the symbols) show the cross-validation SEM.

To further, understand the role of assortativity, we also compare against the performance of a mixed-membership Bernoulli SBM model in which we assume that connection probability tensors $q^{(n)}$ have a diagonal structure.[†] In all the cases we consider, the performance of the assortative model is clearly inferior to our approach, which imposes no constraints on the elements of the $q^{(n)}$ tensors. These results showcase the advantage of using more expressible models.

Figs. 3 and 4 also highlight that the information that different types of interactions carry is not symmetric when it comes to making predictions. For the drug composition dataset, we find that, similar to observations for synthetic data, having information of 2-substance drugs is more informative to predict whether 3 substances are present in the same drug or not than the other way around. Again this is to be somehow expected since the overall fraction of 2-substance drugs (out of possible pairs) is larger than that of 3-substance drugs. In fact, our results suggest that, overall, predicting 2-substance drugs is a harder problem (consistent with what ref. 33 suggests). For higher-order interactions, our results in Fig. 3 also show that, the higher the order of the interaction, the more lower-order interactions needed to make a difference for predicting unobserved hyperedges.

For interactions between genes, our results again show that 3-gene interactions are useful to predict 2-gene interactions when the number of observed 2-gene interactions is comparable to the number of trigenic interactions in the training set. Instead, digenic interactions are very helpful to improve the prediction of trigenic interactions. Indeed, given the small size of the trigenic dataset (0.3% of possible interactions with respect to 59% of possible digenic interactions), our analysis provides an excellent example of how our approach can be leveraged to obtain unobserved higher-order interactions.

An advantage of the expressiveness of our model is that we can use model parameters to further explore the statistical mechanisms that lead to good predictions. Fig. 5 shows $\boldsymbol{q}^{(2)}$ and $\boldsymbol{q}^{(3)}$, the 2-node and 3-node connectivity matrices for the models with the largest posterior we sample, for the drug composition and gene interaction datasets. In both cases, we can see that $\boldsymbol{q}^{(2)}$ and $\boldsymbol{q}^{(3)}$ matrices do not have an assortative, block-diagonal structure. For 2-node and 3-node interactions, we see that there are a few groups of nodes that have a tendency to interact with nodes in other groups, which are reminiscent of core-periphery interactions. At the same time, we also find that in some cases, pairs of groups with nonzero connection probability are also involved in three-group interactions, which is compatible with a certain degree of assortativity. However, 3-node interactions in both cases have richer patterns of connection than 2-node interactions, showing that while 2-node interactions are informative, we need specific 3-node interaction mechanisms different from the 2-node interaction mechanisms. Note also that, while 3-gene and 2-gene interactions have different patterns of interaction, they are related, and 3-gene patterns of interactions are only discernible once we include 2-gene and 3-gene interactions in the training of the model (Fig. 6), thus again highlighting the importance of lower-order interactions to properly model higher-order ones.

Furthermore, our results also capture interaction patterns that would not be compatible with a simplicial closure mechanism (33); in the two cases we analyze, we find in $\boldsymbol{q}^{(3)}$ nonzero probability of interaction between groups for which $\boldsymbol{q}^{(2)}$ is close to zero, which means that for a 3-node hyperedge to exist, a lower-order interaction between all the pairs of nodes is not necessary. Our analysis thus shows that our approach is useful to understand statistical patterns behind higher-order interactions that go beyond assortative mechanisms and also beyond simplicial closure.

Discussion and Conclusions

Our manuscript shows how we can use inferential approaches based on mixed-membership stochastic block models to predict unobserved interactions in hypergraphs. The flexibility of our model allows us to investigate a number of fundamental questions about higher-order interactions.

First, our approach enables us to assess the role that observed hyperedges of different order play in the prediction of unobserved hyperedges. In particular, we have shown that we can leverage the information available about hyperedges of a given order to make

[†]We also tried to use the approach in ref. 17 but it resulted in a lower performance that our implementation of an assortative HMMSBM. The are two main reasons that can explain this: i) because the Poisson model is a worse model for hypergraphs with binary hyperedges; or ii) because the assumption that non observed hyperedges are zeroes in the adjacency matrix is bad in the limit of sparse observations we consider.

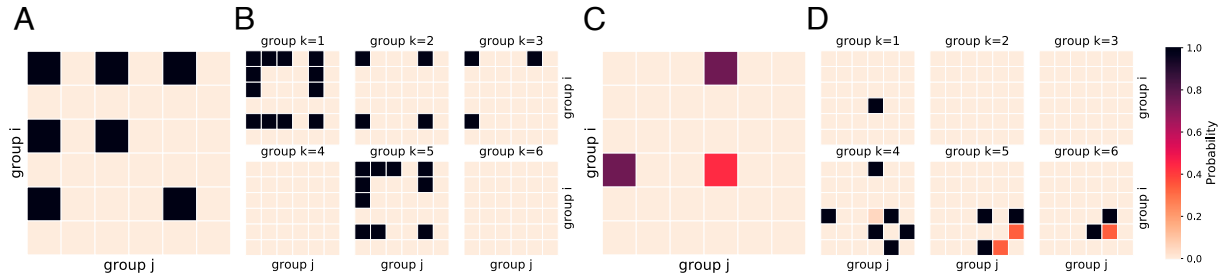

Fig. 5. Connexion probability tensors. (*A* and *B*) Drug substance dataset. (*A*) $\boldsymbol{q}^{(2)}$. Each element shows the probability that a substance in group *i* and a substance in group *j* are part of the same drug. Elements are colored according to the colorbar on the right-hand side (light colors are values very close to zero). (*B*) $\boldsymbol{q}^{(3)}$; each $K \times K$ matrix represent a slice $q^3_{..k}$ of the matrix for a fixed value of index *k*. Each element shows the probability that three substances in groups *i*, *J*, and *k* are in the same drug. (*C* and *D*) Gene interaction dataset. (*C*) $\boldsymbol{q}^{(2)}$. (*D*) $\boldsymbol{q}^{(3)}$; same representation as in panel *B*. These tensors were obtained by obtaining model parameters for all the available data using $K = 6$ for the drug substance dataset and gene interaction dataset (see *Materials and Methods* for the selection of the values of *K*). We show only the matrix for the best posterior of all the maxima we obtained after applying the EM algorithm to different random initializations of model parameters. Other matrices display similar features.

predictions about hyperedges of different orders. We find that, typically, when the number of observed interactions is low, 2-node interactions are more informative about 3-node interactions than vice-versa. This is due to the fact that, while hypergraphs are generally sparse, there is typically a larger observed fraction of 2-node interactions than 3-node interactions. Our analysis shows that the same reasoning can be extended to higher-order interactions as well.

Second, our approach enables us to explore the validity of the assumptions typically made in generative models for hypergraphs, namely that a unique set of group memberships is valid to describe hyperedges of any size and that statistical interaction mechanisms driving the formation of higher-order interactions are assortative. Our analysis for real data shows that while the former hypothesis seems to hold true, this is not the case for the latter necessarily. The comparison of our approach with the equivalent assortative model shows that, despite the apparent limitations of our model in terms of scalability, it clearly outperforms the assortative approach in the task of edge prediction, thus showing that the advantages of model expressiveness outweigh algorithm scalability in this case. Furthermore, the inspection of connection probability tensors reveals that there are connection mechanisms that go beyond both assortativity and simplicial closure.

Our results thus highlight the need for good reconstructions of low-order interactions in order to have reliable reconstructions of full hypergraphs, put forward the importance of considering different statistical mechanisms in the formation of higher-order interactions, and shed light on previous algorithms that consider low-order simplex closure as a temporal mechanism hyperedge formation (33). Our work opens the door to investigating in more depth statistical mechanisms of hyperedge formation and understanding where current inference approaches for predicting higher-order interactions might fail.

Materials and Methods

Synthetic Data. We generate synthetic networks using a mixed-membership stochastic block model with the following parameters:

1) We consider two underlying groups ($K = 2$). Therefore, we consider membership θ vectors of length two, $\boldsymbol{q}^2$ is 2×2 matrix, and $\boldsymbol{q}^3$ is $2\times 2\times 2$ tensor.
2) Nodes can have two possible membership vectors $\theta_1 = [0.05, 0.95]$ and $\theta_2 = [0.95, 0.05]$ nodes are split into two even groups.
3) We use the following connectivity tensors:

$$\boldsymbol{q}^{(2)} = \begin{pmatrix} 0.9 & 0.1 \\ 0.1 & 0.9 \end{pmatrix}, \quad [6]$$

$$\boldsymbol{q}^{(3)}_{..0} = \begin{pmatrix} 0.9 & 0.1 \\ 0.1 & 0.1 \end{pmatrix}, \quad [7]$$

$$\boldsymbol{q}^{(3)}_{..1} = \begin{pmatrix} 0.1 & 0.1 \\ 0.1 & 0.9 \end{pmatrix}. \quad [8]$$

The resulting model has a larger density of 2-body and 3-body interactions between nodes that have the same node memberships. In our analysis, we consider 100 nodes. The average density of interactions is 50% for two body interactions and 30% for three body interactions. A density which is typically higher than in real networks (and therefore an "easy" problem) but that it is useful to describe the leading phenomenological traits.

Real Data.

Drug substances. We consider the NDC-substances dataset provided in ref. 34. From this dataset, we consider two different datasets: i) drugs which comprise 2 and 3 substances. We limit 3-substance drugs to those that have substances for which at least one 2-substance interaction has been reported. Overall this means that we have 302 different substances for which we have 568 substances for which we have 850 2-substance drugs and 408 3-substance drugs (which involve 302 substances)–Fig. 3 *A* and *B*; ii) drugs which comprise, 2, 3, 4, or 5 substances. We only consider 4 and 5-substance drugs for which 2 or 3-substance drugs have been reported. Overall, this means we consider 415 different substances for which have been reported: 402 2-substance drugs, 340 3-substance drugs, 322 4 substance-drugs and 225 5-substance drugs–Fig. 3 *C* and *D*.

In both cases, to construct observed sets of binary interactions, we assume that drug combinations that do not appear in the database correspond to noninteracting combinations of substances. We add as many noninteractions as needed to have a 10% of interactions in the observation. The choice of a 10% is a choice we make in order to ensure that even for small test/training sets there is a measurable amount of substance combination that correspond to existing hyperedges in the dataset.

Gene interactions in S. cerevisiae. In ref. 18, Kuzmin et al. investigated the effect of gene-knockouts involving a total of 1,182 different genes on the growth of *S. cerevisiae*. We consider the set of 410,339 digenic and 91,111 trigenic interactions reported in ref. 18 (Data_S1.csv in their supplementary material). We followed their analysis and considered as "significant" digenic interactions (i.e., edge weight equal to 1) those interactions such that the absolute value of differential growth ϵ_{ij} of the double mutant (f_{ij}) with respect to that single mutant (f_i) is $|\epsilon ij| > 0.08$ and has a *P*-value < 0.05, which amounts to 9.363 significant digenic interactions. We considered as significant (negative) trigenic interactions those in which the differential growth of the triple mutant

$\tau_{ijk} = f_{ijk} - f_i f_j f_k - \epsilon ij f_k - \epsilon_{jk} f_i - \epsilon_{ki} f_j$ fullfils the condition $\tau_{ijk} < -0.08$ and has a P-value < 0.05. In total, there are 3,196 significant negative trigenic interactions.

Inference Equations. Expressing the log posterior in Eq. **5** in terms of the mixed-membership SBM model for the probability of observing an interaction between a set of nodes, we obtain that

$$\begin{aligned}
\log P(\{\boldsymbol{\theta}, \boldsymbol{q}\}|\mathbf{A^0}) &= \sum_n \log L\left(\left(\mathbf{A^0}\right)^{(n)}|\boldsymbol{\theta}, \boldsymbol{q}^{(n)}\right) \\
&= \sum_n \log \prod_{\{i_j\}\in\left(\mathbf{A}^O\right)^n} p\left(\left(A^O\right)^n_{\{i_j\}}|\boldsymbol{\theta}, \boldsymbol{q}^{(n)}\right) \\
&= \sum_n \sum_{\{i_j\}\in\left(\mathbf{A}^O\right)^{(n)}} \log\left(p\left(\left(A^O\right)^n_{\{i_j\}}|\boldsymbol{\theta}, \boldsymbol{q}^{(n)}\right)\right) \\
&= \sum_n \left(\sum_{\{i_j\}/\left(A^O\right)^{(n)}_{\{i_j\}}=1} \log \sum_{\{\alpha_j\}} \prod_{j=1}^n \theta_{i_j\alpha_j} q^{(n)}_{\{\alpha_i\}} \right. \\
&\quad \left. + \sum_{\{i_j\}/\left(A^O\right)^{(n)}_{\{i_j\}}=0} \log \sum_{\{\alpha_j\}} \prod_{j=1}^n \theta_{i_j\alpha_j}(1 - q^{(n)}_{\{\alpha_j\}}) \right), \quad \textbf{[9]}
\end{aligned}$$

where $\{i_j\} := (i_1 \dots i_n)$ and $\{\alpha_j\} := (\alpha_1 \dots \alpha_n)$. To obtain the parameters $\boldsymbol{\theta}^\star$ and $\boldsymbol{q}^\star$ that maximize the log posterior, we follow an Expectation Maximization approach.

By introducing an auxiliary function $\omega^x_{\{i_j\}}(\{\alpha_j\})$; $x = 0, 1$ with $\sum_{\{\alpha_j\}} \omega^x_{\{i_j\}}(\{\alpha_j\}) = 1$ for each term in the logarithm and using Jensen's inequality for each term in the sum:

$$\begin{aligned}
&\log\left(\sum_{\{\alpha_j\}} \prod_{j=1}^n \theta_{i_j\alpha_j} q^{(n)}_{\{\alpha_j\}}\right) \\
&= \log\left(\sum_{\{\alpha_j\}} \frac{\prod_{j=1}^n \theta_{i_j\alpha_j} q^{(n)}_{\{\alpha_j\}}}{\omega^1_{\{i_j\}}(\{\alpha_j\})} \omega^1_{\{i_j\}}(\{\alpha_j\})\right) \\
&\geq \sum_{\{\alpha_j\}} \omega^1_{\{i_j\}}(\{\alpha_j\}) \log\left(\frac{\prod_{j=1}^n \theta_{i_j\alpha_j} q^{(n)}_{\{\alpha_j\}}}{\omega^1_{\{i_j\}}(\{\alpha_j\})}\right).
\end{aligned}$$

$$\begin{aligned}
&\log\left(\sum_{\{\alpha_j\}} \prod_{j=1}^n \theta_{i_j\alpha_j}(1 - q^{(n)}_{\{\alpha_j\}})\right) \\
&= \log\left(\sum_{\{\alpha_j\}} \frac{\prod_{j=1}^n \theta_{i_j\alpha_j} q^{(n)}_{\{\alpha_j\}}}{\omega^0_{\{i_j\}}(\{\alpha_j\})} \omega^0_{\{i_j\}}(\{\alpha_j\})\right) \\
&\geq \sum_{\{\alpha_j\}} \omega^0_{\{i_j\}}(\{\alpha_j\}) \log\left(\frac{\prod_{j=1}^n \theta_{i_j\alpha_j} q^{(n)}_{\{\alpha_j\}}}{\omega^0_{\{i_j\}}(\{\alpha_j\})}\right).
\end{aligned}$$

Note that the index x is unnecessary because from the indices $\{i_j\}$ we already know whether $A_{\{i_j\}}$ is 1 or 0. However, we keep it for clarity.

Putting everything together, we obtain the following inequality for the log posterior:

$$\begin{aligned}
\log P\left(\{\boldsymbol{\theta}, \boldsymbol{q}\}|\mathbf{A^0}\right) \geq & \sum_n \sum_{\{i_j\}/\left(A^0\right)^n_{\{i_j\}}=1} \sum_{\{\alpha_j\}} \omega^1_{\{i_j\}}(\{\alpha_j\}) \\
& \times \log\left(\frac{\prod_{j=1}^n \theta_{i_j\alpha_j} q^{(n)}_{\{\alpha_j\}}}{\omega^1_{\{i_j\}}(\{\alpha_j\})}\right) \\
& + \sum_n \sum_{\{i_j\}/\left(A^0\right)^n_{\{i_j\}}=0} \sum_{\{\alpha_j\}} \omega^0_{\{i_j\}}(\{\alpha_j\}) \\
& \times \log\left(\frac{\prod_{j=1}^n \theta_{i_j\alpha_j}(1 - q^{(n)}_{\{\alpha_j\}})}{\omega^0_{\{i_j\}}(\{\alpha_j\})}\right). \quad \textbf{[10]}
\end{aligned}$$

In the above expression, the equality is met when

$$\omega^1_{\{i_j\}}(\{\alpha_j\}) = \frac{\prod_{j=1}^n \theta_{i_j\alpha_j} q^{(n)}_{\{\alpha_j\}}}{\sum_{\{\alpha'_j\}} \prod_{j=1}^n \theta_{i_j\alpha'_j} q^{(n)}_{\{\alpha'_j\}}}, \text{ and} \quad \textbf{[11]}$$

$$\omega^0_{\{i_j\}}(\{\alpha_j\}) = \frac{\prod_{j=1}^n \theta_{i_j\alpha_j}\left(1 - q^{(n)}_{\{\alpha_j\}}\right)}{\sum_{\{\alpha'_j\}} \prod_{j=1}^n \theta_{i_j\alpha'_j}\left(1 - q^{(n)}_{\{\alpha'_j\}}\right)}. \quad \textbf{[12]}$$

The evaluation of these equations is called the expectation step.

Then, maximizing the right hand side of the above equation subject to the constraints $\sum_\alpha \theta_{i\alpha} = 1$ yields the following (maximization) equations for the $\boldsymbol{\theta}$ and $\boldsymbol{q}$:

$$\theta_{i_k\alpha_k} = \frac{\sum_n \sum_{\{i_j\}_k;\{\alpha_j\}_k} \omega_{\{i_j\}_k;i_k}(\{\alpha_j\}_k;\alpha_k)}{\sum_n d_{i_k,n}}, \quad \textbf{[13]}$$

$$q^n(\{\alpha_j\}) = \frac{\sum_{\left\{\{i_j\}/\left(A^0\right)^n_{\{i_j\}}=1\right\}} \omega^1_{\{i_j\}}(\{\alpha_j\})}{\sum_{\{i_j\}} \omega_{\{i_j\}}(\{\alpha_j\})}, \quad \textbf{[14]}$$

where we have dropped the superindices in ω for simplicity; $\{i_j\}_k$ represents the set of node indices except i_k, $\{\alpha_j\}_k$ represents the set of groups indices except α_k and $d_{i_k,n}$ is the number of interactions involving n nodes in which node i_k participates.

The EM algorithm then works as follows:

1. Generate random initial conditions for $\boldsymbol{\theta}$ and $\boldsymbol{q}$.
2. Expectation Step: compute auxiliary functions ω using Eqs. **11** and **12**.
3. Maximization Step: compute new values of $\boldsymbol{\theta}$ and $\boldsymbol{q}$ using Eqs. **13** and **14**.
4. Iterate Steps 2 and 3 until convergence.

Assortative mixed-membership SBM. In the assortative version of our model, connection probability tensors have elements equal to zero except for the diagonals so that the probability of an edge existing is then:

$$p(A^{(n)}_{i_1,\dots,i_n} = 1|\boldsymbol{\theta}, \boldsymbol{q}^{(2)}) = \sum_\alpha q^{(n)}_\alpha \prod_{k=1}^n \theta_{i_k\alpha}. \quad \textbf{[15]}$$

The inference equations are then derived in the same way as for the full model.

Making predictions. The EM maximization method will find a local optimum in the posterior landscape. However, in order to make predictions the best estimate for $p(A_{\{i_1\cdots_M\}} = 1|\mathbf{A^0})$ comes from integrating over all possible parameters $\{\boldsymbol{\theta}, \boldsymbol{q}\}$, that is, by computing

$$p(A_{\{i_k\}}|\mathbf{A^0}) = \int d\boldsymbol{\theta} \int d\boldsymbol{q}\, p(A_{\{i_k\}}|\{\boldsymbol{\theta}, \boldsymbol{q}\})\, .p(\{\boldsymbol{\theta}, \boldsymbol{q}\}|\mathbf{A^0}). \quad \textbf{[16]}$$

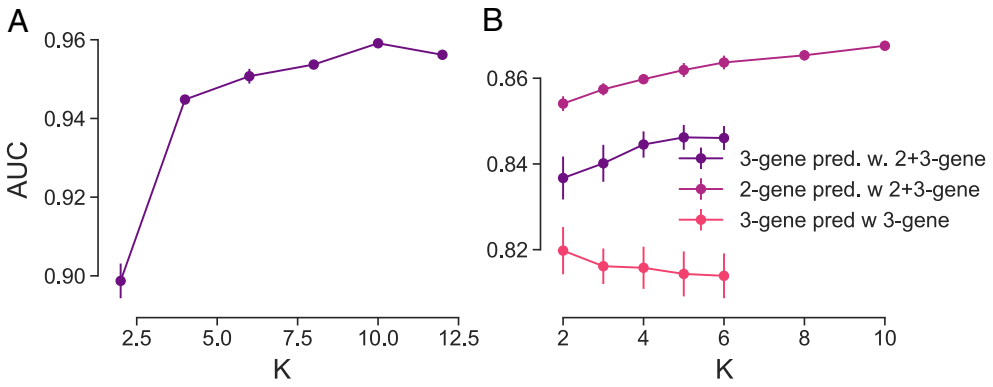

Fig. 6. Selection of the value of K for real datasets. (*A*) Drug substances. Performance at predicting 2 and 3-substance drug compositions using 2 and 3 substance drug compositions in the dataset versus the number of latent groups K. (*B*) Gene interactions. Performance at predicting 2 and 3-gene interactions using 2-gene and 2 and 3-gene interactions in the training set, 3-gene and 2 and 3-gene interactions in the training set, respectively.

However, because this is an infeasible task, we estimate this probability by averaging over local maxima $\mathcal{M}$ in the posterior $p(\{\boldsymbol{\theta}, \boldsymbol{q}\}|\mathbf{A^0})$ landscape so that

$$p(A_{\{i_k\}}|\mathbf{A^0}) = \sum_{\{\boldsymbol{\theta},\boldsymbol{q}\}\in\mathcal{M}} p(A_{\{i_k\}}|\{\boldsymbol{\theta}, \boldsymbol{q}\}). \quad [17]$$

Selection of K. To select the best value of K for rel datasets, we follow a cross-validation procedure varying the value of K. We select the smallest value of K for which the cross-validation prediction performance starts to saturate. For the drug substance dataset, we use a value of $K = 6$ (Fig. 6*A*). For the gene interaction dataset, we use a value of $K = 6$ (for digenic only, and digenic and trigenic interactions) and a value of $K = 2$ for trigenic interactions only since due to the small number of trigenic interactions available, larger values of K already overfit in this case (Fig. 6*B*).

Data, Materials, and Software Availability. The code for the implementation of the full model and the assortative model is available at https://github.com/seeslab/HyGMMSBM. Previously published data were used for this work (18, 33).

ACKNOWLEDGMENTS. This research was funded by projects PID2019-106811GB-C31 and PID2022-142600NB-I00 from MCIN/AEI/10.13039/501100011033, and by the Government of Catalonia (2021SGR-633).

1. G. Bianconi, "Higher-order networks" in *Elements in Structure and Dynamics of Complex Networks* (Cambridge University Press, 2021).
2. C. Bick, E. Gross, H. A. Harrington, M. T. Schaub, What are higher-order networks? *SIAM Review* **65**, 686–731 (2023).
3. F. Battiston *et al*., Networks beyond pairwise interactions: Structure and dynamics. *Phys. Rep.* **874**, 1–92 (2020).
4. F. Battiston *et al*., The physics of higher-order interactions in complex systems. *Nat. Phys.* **17**, 1093–1098 (2021).
5. P. Skardal, A. Arenas, Higher order interactions in complex networks of phase oscillators promote abrupt synchronization switching. *Commun. Phys.* **3**, 218 (2021).
6. G. St-Onge *et al*., Influential groups for seeding and sustaining nonlinear contagion in heterogeneous hypergraphs. *Commun. Phys.* **5**, 25 (2022).
7. R. Ghorbanchian, J. G. Restrepo, J. J. Torres, G. Bianconi, Higher-order simplicial synchronization of coupled topological signals. *Commun. Phys.* **14**, 120 (2021).
8. T. P. Peixoto, "Descriptive vs. inferential community detection in networks: Pitfalls, myths and half-truths" in *Elements in the Structure and Dynamics of Complex Networks* (Cambridge University Press, 2023).
9. R. Guimerà, M. Sales-Pardo, Missing and spurious interactions and the reconstruction of complex net works. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 22073–22078 (2009).
10. T. Vallès-Català, T. P. Peixoto, M. Sales-Pardo, R. Guimerà, Consistencies and inconsistencies between model selection and link prediction in networks. *Phys. Rev. E* **97**, 062316 (2018).
11. M. Tarres-Deulofeu, A. Godoy Lorite, R. Guimerà, M. Sales-Pardo, Tensorial and bipartite block models for link prediction in layered networks and temporal networks. *Phys. Rev. E* **99**, 032307 (2018).
12. T. P. Peixoto, Network reconstruction and community detection from dynamics. *Phys. Rev. Lett.* **123**, 128301 (2019).
13. J. G. Young, G. T. Cantwell, M. E. J. Newman, Bayesian inference of network structure from unreliable data. *J. Complex Networks* **8**, cnaa046 (2021).
14. C. Kim, A. S. Bandeira, M. X. Goemans, Stochastic block model for hypergraphs: Statistical limits and a semidefinite programming approach. arXiv [Preprint] (2018). https://doi.org/10.48550/arXiv.1807.02884 (Accessed 15 November 2022).
15. P. S. Chodrow, N. Veldt, A. R. Benson, Generative hypergraph clustering: From blockmodels to modularity. *Sci. Adv.* **7**, eabh1303 (2021).
16. N. Ruggeri, M. Contisciani, F. Battiston, C. D. Bacco, Community detection in large hypergraphs. *Sci. Adv.* **9**, eadg9159 (2023).
17. M. Contisciani, F. Battiston, C. De Bacco, Inference of hyperedges and overlapping communities in hypergraphs. *Nat. Commun.* **13**, 7229 (2022).
18. E. Kuzmin *et al*., Systematic analysis of complex genetic interactions. *Science* **360**, eaao1729 (2018).
19. M. Menden *et al*., Community assessment to advance computational prediction of cancer drug combinations in a pharmacogenomic screen. *Nat. Commun.* **10**, 2674 (2019).
20. A. Vazquez, Finding hypergraph communities: A Bayesian approach and variational solution. *J. Stat. Mech.* **2009**, P07006 (2009).
21. R. Guimerà, One model to rule them all in network science? *Proc. Natl. Acad. Sci. U.S.A.* **117**, 25195–25197 (2020).
22. O. Fajardo-Fontiveros, R. Guimerà, M. Sales-Pardo, Node metadata can produce predictability crossovers in network inference problems. *Phys. Rev. X* **12**, 011010 (2022).
23. D. Hric, T. P. Peixoto, S. Fortunato, Network structure, metadata, and the prediction of missing nodes and annotations. *Phys. Rev. X* **6**, 031038 (2016).
24. E. M. Airoldi, D. M. Blei, S. E. Fienberg, E. P. Xing, Mixed membership stochastic blockmodels. *J. Mach. Learn. Res.* **9**, 1981–2014 (2008).
25. A. Godoy-Lorite, R. Guimerà, C. Moore, M. Sales-Pardo, Accurate and scalable social recommendation using mixed-membership stochastic block models. *Proc. Natl. Acad. Sci. U.S.A.* **113**, 14207–14212 (2016).
26. C. De Bacco, E. A. Power, D. B. Larremore, C. Moore, Community detection, link prediction, and layer interdependence in multilayer networks. *Phys. Rev. E* **95**, 042317 (2017).
27. D. Ghoshdastidar, A. Dukkipati, "Consistency of spectral partitioning of uniform hypergraphs under plan ted partition model" in *Advances in Neural Information Processing Systems*, Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, K. W. Einberger, Eds. (Curran Associates, Inc., 2014), vol. 27.
28. I. Chien, C. Y. Lin, I. H. Wang, "Community detection in hypergraphs: Optimal statistical limit and efficient algorithms" in *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics and Proceedings of Machine Learning Research*, A. Storkey, F. Perez-Cruz, Eds. (PMLR, 2018), vol. 84, pp. 871–879.
29. M. Zhang, Z. Cui, S. Jiang, Y. Chen, "Beyond link prediction: Predicting hyperlinks in adjacency space" in *Proceedings of the 32nd AAAI Conference on Artificial Intelligence* (2018).
30. T. Kumar, K. Darwin, S. Parthasarathy, B. Ravindran, "HPRA: Hyperedge prediction using resource allocation" in *12th ACM Conference on Web Science, WebSci '20* (Association for Computing Machinery, New York, NY, 2020), pp. 135–143.
31. G. Abuoda *et al*., "Link prediction via higher-order motif features" in *Machine Learning and Knowledge Discovery in Databases*, U. Brefeld, Ed. (Springer International Publishing, Cham, Switzerland, 2020), pp. 412–429.
32. D. Maurya, B. Ravindran, Hyperedge prediction using tensor eigenvalue decomposition. arXiv [Preprint] (2021). https://doi.org/10.48550/arXiv.2102.04986 (Accessed 15 November 2022).
33. A. R. Benson, R. Abebe, M. T. Schaub, A. Jadbabaie, J. Kleinberg, Simplicial closure and higher-order link prediction. *Proc. Natl. Acad. Sci. U.S.A.* **115**, E11221–E11230 (2018).
34. R. A. Rossi, A. Rao, S. Kim, E. Koh, N. Ahmed, "From closing triangles to closing higher-order motifs" in *Companion Proceedings of the Web Conference 2020, WWW '20* (Association for Computing Machinery, New York, NY, 2020), pp. 42–43.
35. R. Guimerà, M. Sales-Pardo, A network inference method for large-scale unsupervised identification of novel drug–drug interactions. *PLoS. Comput. Biol.* **9**, e1003374 (2013).
36. T. P. Peixoto, Inferring the mesoscale structure of layered, edge-valued, and time-varying networks. *Phys. Rev. E* **92**, 042807 (2015).
37. T. Peixoto, M. Rosvall, Modelling sequences and temporal networks with dynamic community structures. *Nat. Commun.* **8**, 582 (2017).
38. A. Divakaran, A. Mohan, Temporal link prediction: A survey. *New Gener. Comput.* **38**, 213–258 (2020).
39. S. Cobo-López, A. Godoy-Lorite, J. Duch, M. Sales-Pardo, R. Guimerà, Optimal prediction of decisions and model selection in social dilemmas using block models. *EPJ Data Sci.* **7**, 48 (2018).
40. M. E. J. Newman, A. Clauset, Structure and inference in annotated networks. *Nat. Commun.* **7**, 11863 (2016).