DANGO: Predicting higher-order genetic interactions

Ruochi Zhang[1], Jianzhu Ma[2,*], and Jian Ma[1,*]

[1]Computational Biology Department, School of Computer Science,
Carnegie Mellon University, Pittsburgh, PA 15213, USA
[2]Department of Computer Science and Department of Biochemistry
Purdue University, West Lafayette, IN 47907, USA
[*]Correspondence: majianzhu@purdue.edu and jianma@cs.cmu.edu

Abstract

Higher-order genetic interactions, which have profound impact on phenotypic variations, remain poorly characterized. Almost all studies to date have primarily reported pairwise interactions because it is dauntingly difficult to design high-throughput genetic screenings of the large combinatorial search space for higher-order interactions. Here, we develop an algorithm named DANGO, based on a self-attention hypergraph neural network, to effectively predict the higher-order genetic interaction for a group of genes. As a proof-of-concept, we make comprehensive prediction of >400 million trigenic interactions in the yeast *S. cerevisiae*, significantly expanding the quantitative characterization of trigenic interactions. We find that DANGO can accurately predict trigenic interactions that reveal both known and new biological functions related to cell growth. The predicted trigenic interactions can also serve as powerful genetic markers to predict growth response to many distinct conditions. DANGO enables unveiling a more complete map of complex genetic interactions that impinge upon phenotypic diversity.

Introduction

Genetic interactions represent the phenomena where different genes (or more broadly, genetic variants) work synergistically to impact phenotypes (Costanzo et al., 2019). Characterizing the principles and mechanisms of genetic interactions, which have profound implications in understanding human health and disease, has been a central question in systems biology (Boone et al., 2007; Cowen et al., 2017; Domingo et al., 2019). Leveraging workhorse model organism *S. cerevisiae*, both high-throughput experimental approaches and computational methods have been developed to map the landscape of genetic interactions (Costanzo et al., 2016, 2019). Such genetic interactions are often manifested when mutations in the genes involved in the interaction lead to an unexpected growth phenotype that cannot be explained by the additive effects of individual mutations (Costanzo et al., 2016). For instance, positive genetic interactions happen when mutations in multiple genes lead to greater growth fitness than what is expected from individual mutations, while negative genetic interactions, including the most extreme form "synthetic lethality", lead to much worse growth fitness than expected. Together, measuring genetic interactions has shed new light on the functional dependencies between genes, the potential redundancies in the genome, and, most importantly, the complex genotype-to-phenotype relationships.

However, genetic interaction studies so far have been primarily focused on pairwise, digenic interactions (Boone et al., 2007; Costanzo et al., 2016). Growing evidence has suggested that higher-order genetic interactions among three or more genes also play critical roles in controlling phenotypes. A number of studies revealed that higher-order genetic interactions occur across a diverse range of biological contexts (Beh et al., 2001; Suzuki et al., 2011; Bao et al., 2015; Garst et al., 2017; Lian et al., 2017; Zhang et al., 2018, 2019). By mutating tens of thousands of loci of the *E. coli* genome, Garst et al. (2017) identified a higher-order interaction among Fis, Fnr, and FadR related to the viability of cells at high acetate concentrations. Suzuki et al. (2011) developed the "green monster" technology that precisely deleted 16 ATP-binding cassette transporters in a drug-sensitive strain of *S. cerevisiae* and has been applied to delineate higher-order genetic interactions of the 16 transporters in ,353 engineered strains with >85,000 genotype-to-resistance drug-response observations (Celaj et al., 2020). Importantly, in a recent landmark study, Kuzmin et al. (2018) reported the first large-scale profile of trigenic interactions in the yeast *S. cerevisiae* where the impact of different combinations of three-genes deletions on colony growth is ascertained systematically, greatly expanding the reach of genetic interaction analysis. However, even in this most systematic study of trigenic interactions to date (Kuzmin et al., 2018), only ∼100,000 among all possible 36 billion trigenic interactions were tested, leaving the roles of the vast majority of trigenic interactions unclear.

Here, we develop a new machine learning model named DANGO, based on a self-attention hypergraph neural network, to predict higher-order genetic interactions by capturing the complex relationship between higher-order interactions and pairwise interactions extracted from a collection of heterogeneous molecular networks. To evaluate the method, we utilize the trigenic interaction data from Kuzmin et al. (2018). We apply DANGO to multiple pairwise interaction networks that are widely available to make trigenic interaction predictions. We find that DANGO is able to accurately predict trigenic interactions with significant advantages over other state-of-the-art algorithms generalized from methods designed for predicting pairwise interactions. Crucially, we demonstrate that DANGO as a computational framework not only refines the measured trigenic interactions in Kuzmin et al. (2018), but also further predicts more than 400 million interaction scores for potential trigenic interactions that were not measured previously, significantly expanding the repertoire of quantitative predictions of trigenic interactions. We show that the trigenic interactions predicted by DANGO capture known and new functions related to cell growth. In addition, we transfer DANGO to an independent yeast dataset and demonstrate that the predicted trigenic interactions can serve as important features to improve the mapping from genotype to phenotype.

Results

Overall design of DANGO

Fig. 1a illustrates the overall architecture of DANGO for the prediction of trigenic interactions by modeling genes and their trigenic interactions as a hypergraph. DANGO captures the higher-order interaction patterns with four main components: (1) Constructing the hypergraph where genes are represented as nodes and trigenic interactions are represented as 3-way hyperedges (marked as "Labeled Data" in Fig. 1a). The trigenic interaction scores are modeled as the attributes of hyperedges. (2) Pre-training graph neural networks (GNNs) to generate node embeddings based on six pairwise interaction networks from the STRING database (Franceschini et al., 2012). (3) Using meta embedding learning to integrate embeddings from different networks. (4) Training the hypergraph representation learning model Hyper-SAGNN (Zhang et al., 2020) with both labeled data and node features as input (Fig. 1b). Importantly, the Hyper-SAGNN architecture developed in DANGO is trained with a distinct loss function to predict the attributes of hyperedges in a regression manner, different from other applications of Hyper-SAGNN. Note that although we describe DANGO in separate components, the GNNs (after pre-training), the meta embedding learning module, and the Hyper-SAGNN are jointly optimized in an end-to-end fashion. Also, DANGO is generic to extend to higher-order interactions other than trigenic ones. The detailed structures of these components are described in the Methods section. Additional descriptions of the model choice, the training procedure, the computational complexity, and the design of the baseline models used in this work can be found in Supplementary Methods.

Datasets used in DANGO

The six PPI networks used as input to the GNNs are from STRING database v9.1 (Franceschini et al., 2012). Although the latest STRING database v11.0 contains more PPIs, to make a fair comparison with other baseline models, we kept the input the same. These networks are referred to as "Experimental" (experimentally verified PPIs through literature curation), "Database" (imported PPIs from other database), "Neighborhood" (proteins with similar genomic context in different species), "Fusion" (proteins fused in the given genome), "Co-occurrence" (proteins with similar phylogenetic profile), and "Co-expression" (predicted association between genes with co-expression patterns) in the STRING database.

The trigenic dataset studied in this work is from Kuzmin et al. (2018), which contains measured trigenic interaction scores τ for around 91,000 triplets among 1,400 genes. 1,395 out of 1,400 genes also have records in the pairwise PPI networks used as input. The remaining five genes with their related triplets are excluded for the following analysis. All measured trigenic interaction scores are negative, i.e., there are no observed positive trigenic interactions within this dataset. The Pearson correlation between the trigenic interaction scores of two individual replicates is around 0.59, which is much lower than the Pearson correlation between the digenic interaction score of two replicates from the same data source (0.88). Note that these numbers give an approximate upper bound of the performance that a computational framework could possibly achieve, suggesting the difficulties of capturing the higher-order interactions from the dataset.

DANGO can accurately predict trigenic interactions

We systematically evaluated the performance of DANGO in predicting trigenic interaction scores and made comparisons with three baseline models, including two methods based on the state-of-the-art biological network integration method Mashup (Cho et al., 2016) and one graph neural network model that was trained in an end-to-end manner. These baseline methods are referred to as Mashup-RF, Mashup-

GTB, and GCN-Avg-DNN, respectively (see Supplementary Methods A.2 & A.3 for details on our training procedure, computational complexity, and the design of the baseline models). We evaluated the performance of the prediction with several metrics including (1) Pearson and Spearman correlations between the predicted and measured trigenic interactions; (2) AUROC and AUPR for classifying the triplets into "strong" and "weak" interactions with a cutoff 0.05 on the absolute measured interaction score; (3) Pearson and Spearman correlation within the "strong" interactions.

We performed a 5-fold cross-validation on the trigenic interaction dataset from Kuzmin et al. (2018). We used the trained models to make predictions on the test set and collected the predicted values from all five folds. We found that DANGO can make accurate predictions for trigenic interactions (Fig. 2a) and significantly outperforms all baseline models on all evaluation metrics (Fig. 2b). We also observed overall better performance of the GCN-Avg-DNN model than the other two baseline models Mashup-RF and Mashup-GTB, suggesting that training the model in an end-to-end manner allows the node embeddings to capture task-specific information more effectively.

There were two replicates when measuring the trigenic interaction scores from the dataset in Kuzmin et al. (2018). Therefore, a *P*-value indicating the consistency between the two replicates of each trigenic interaction score can be calculated. We tested whether DANGO would achieve higher correlation scores on a smaller but more reliable set of data. As shown in Fig. 2c with Pearson and Spearman correlations, the correlation score goes higher on the test set with more strict *P*-value cut-off, especially when the cut-off changes from 5e-1 to 1e-2. More importantly, we observed that DANGO can achieve equally strong performance on these reliable test sets with a less reliable training data. For around 100,000 trigenic interaction scores measured in Kuzmin et al. (2018), only 4,042 of them pass the *P*-value cut-off 1e-2 (see Fig. S3), making it even harder to obtain reliable trigenic interaction scores for all combinations of genes. DANGO's stable and robust performance strongly suggests its capability to uncover interactions from noisy data source and refine the trigenic interaction scores originally measured by experimental approaches.

Assessing bias in trigenic interaction predictions

We next sought to assess whether the performance improvements were resulted from the model's enhanced ability to capture interaction patterns or simply 'memorizing' potential bias (overfitting to the confounding factors in both training and test sets) (Eid et al., 2020). For instance, if a gene appears many times in the measured trigenic interaction dataset with high interaction scores, the model could achieve good performance by simply assigning all triplets containing this gene with high interaction scores. We specifically assessed if such bias exists in DANGO and the baseline models using two approaches.

The first approach is to randomly shuffle the node embeddings and re-train the model, i.e., each gene is now assigned with an embedding vector of another gene. For Mashup-RF and Mashup-GTB, this can be achieved by shuffling Mashup-embeddings across genes. For GCN-Avg-DNN and DANGO, this can be done by shuffling the adjacency matrix of the six PPI networks used in the GCN. Note that if the model achieves strong performance by overfitting to the training set, random shuffling would not lead to a significant decrease in the performance. On the other hand, if the model indeed learns the interaction patterns, random shuffling would lead to performance drop. Among all methods, we found that DANGO has the most significant performance decrease when the features are shuffled across genes (Fig. 2d). All other baseline models have much smaller decrease in the performance, suggesting the existence of potential bias from the models.

The second method is to use a different training/test split scheme to assess whether the model can generalize to unseen genes. We tested two splitting schemes. (1) We kept around 60 genes unobserved in the training set. All triplets containing at least one of these 60 genes automatically went to the test set

(Fig. 2e). (2) We tested on triplets containing three unobserved genes. We randomly selected 400 genes and made sure that triplets with all three genes from this set are in the test set. Triplets with one or two genes from this set were excluded from both training and test set (Fig. 2f). Under both settings, DANGO achieves best performance among all methods (Fig 2e-f). We also evaluated the performance decrease with randomly shuffled feature space under these two training/test split scheme. Consistent with what we observed under random split scheme (Fig. S2), DANGO again has the largest performance decrease (except the Pearson correlation within strong trigenic interactions).

These evaluations collectively suggest that DANGO is able to make accurate trigenic interaction score predictions and its strong performance on the test set is not due to overfitting to the confounding factors in the dataset. One reason for DANGO to overcome such bias is that the unique self-attention mechanism from Hyper-SAGNN makes the node embeddings to interact with each other while all other baseline methods do not guarantee that.

DANGO discovers novel trigenic interactions with important biological functions

Next, we asked whether DANGO can be used to make *de novo* predictions of trigenic interactions. We used the trained DANGO to make predictions on two sets of combinations of three genes. The first set contains the same triplets measured in Kuzmin et al. (2018). The predicted interaction scores for this set are obtained through a 5-fold cross-validation where the predictions on the test sets in each fold are collected and concatenated. The second set is the combinations of all 1,395 genes in the dataset ($\sim$451 million triplets) except those that have been measured in Kuzmin et al. (2018), i.e., the predictions are obtained by a DANGO model trained with all measured trigenic interactions in Kuzmin et al. (2018).

For both predicted interactions and the originally measured interactions, we selected the top 20% of them for further study (cut-off $|\tau| \sim 0.05$ for the original trigenic interactions). We characterized the functional properties of these selected trigenic interactions with enrichment analysis of Gene Ontology (GO) biological process and cellular component annotations. For genes that have multiple records, we assigned them to the "leaf node" of the GO hierarchy. Each trigenic interaction is classified as "across terms" or "within terms" based on whether all three genes belong to the same GO term. Within each category, we calculated the fold-change of frequencies in the selected trigenic interactions versus the background (all triplets in the two sets) with hypergeometric test to assess the enrichment significance. As shown in Fig. 3a-b, when evaluating on set #1, the enriched GO terms largely overlap with the original trigenic interaction set with fold-change scores correlated well with each other (Pearson correlation 0.56 for within terms and 0.94 for across terms). This again demonstrates that DANGO captures the higher-order genetic interactions within the original datasets. For all GO term patterns (within or across terms) that appear in the original interaction set at least once, its frequency fold-change in set #2 correlates well with that of the original set (Pearson correlation 0.53 for within terms, and 0.72 for across terms). However, the predicted trigenic interactions in set #2 are also enriched with extra GO terms. In particular, we found that most of these extra biological processes are related to the cell growth. For instance, 'cytoplasmic vesicle', 'cell cortex' and 'cytoskeleton' are important cellular components to maintain cell survival. In addition, other biological processes such as 'RNA splicing', 'Golgi vesicle transport', 'Golgi apparatus', and 'microtubule organizing center' are also known to play important roles in regulating growth and development (Parenteau et al., 2008; Feyder et al., 2015; Sawin and Tran, 2006).

Next, we took a closer look at the combinations of different GO terms. Specifically, we calculated the frequency fold-change of a combination of GO terms in the selected trigenic interaction set versus the background. We observed that DANGO can recover a significant amount of combinations of biological processes not enriched in the original set. One example is a novel trigenic interaction hub shown in Fig. 3c, where the network shows all pairs of biological processes that have significant trigenic in-

teractions with genes from the protein targeting process (*P*-value $\leq 10^{-3}$). We found that within the network of set #1, the node that corresponds to the biological process endosomal transport has a very large degree, reflecting a trigenic interaction hub consisting of genes related to protein targeting, genes related to endosomal transport, and genes from various other biological processes.

We then asked if these identified trigenic interactions co-occur frequently in the higher-order interactions among multiple genes captured by orthogonal data. We again calculated the frequency fold-changes of the three genes in the trigenic interaction that co-occur in a specific KEGG pathway (Kanehisa and Goto, 2000) versus the background (Fig. 4a). Within the original trigenic interaction set, there are only two KEGG pathways (MAPK signaling pathway and cell cycle) that significantly overlap with the predicted trigenic interactions (*P*-value $\leq 10^{-3}$, hypergeometric test), and both KEGG pathways received lower *P*-values in set #1. All of the selected trigenic interactions (from set #1 and the original set) that are enriched with MAPK signaling pathway always consist of CLN1 and CLN2 that are paralogs. Compared with the original set, set #1 includes 4 more trigenic interactions, two of which (BEM3, RGA1) with higher predicted interaction scores have recorded physical interactions between CLN1/2 and the third gene based on BioGRID (Stark et al., 2006). Set #1 also reveals an additional significant KEGG pathway, endocytosis. Among the measured combinations of genes, there are 3 triplets with all 3 genes from this pathway, all of which have predicted interaction scores within the top 20% of set #1. Of the triplets not selected in the original set, the one with the highest absolute measured interaction score (-0.005) has a *P*-value 0.46 (within top 10% of all the *P*-values), suggesting that the original measured score may not be reliable. Set #2 has identified all these KEGG pathways with much higher fold-change value and lower *P*-values, but it further reveals 14 additional KEGG pathways that significantly overlap with the predicted trigenic interactions.

Moreover, we tested whether all the genes in the predicted trigenic interactions are likely to form a protein complex by comparing to a manually curated protein complex database (Pu et al., 2009) (Fig. 4b). Consistent with what we observed from the KEGG pathway, both the original trigenic interaction set and set #1 only have 1 complex (COMPASS complex) significantly overlapping with the trigenic interactions while set #2 contains 14 extra protein complexes.

These results strongly suggest that the trigenic interactions predicted by DANGO can uncover both known and new biological functions related to cell growth, suggesting the potential of DANGO to facilitate the discovery of novel biological pathways and protein complexes.

Trigenic interactions predicted by DANGO enhance the prediction of yeast growth

To further demonstrate that the trigenic interactions predicted by DANGO provide critical insights into the wiring mechanism of gene functions and pathways, we challenged it with the task of predicting yeast growth under various conditions that affect physiological and cellular responses of distinct yeast strains. We obtained the data from (Peter et al., 2018), which characterized the single nucleotide polymorphisms (SNPs) of 971 yeast strains with the corresponding phenotypes quantified by the growth fitness under 35 conditions (e.g., adding caffeine into the culture medium or culturing at a high temperature). In our prediction setting, our goal is to use the SNPs in each strain to predict its 35 growth fitness scores. This is a highly challenging prediction problem because there are more than 18,000 SNPs in the dataset (much higher than the number of yeast strains that can be used as the training data). Even after grouping the SNPs based on the genes they reside in, the dimension of features for each strain can still be greater than 6,000. The curse of dimensionality combined with insufficient data points would make pure data-driven approaches such as deep neural networks (DNNs) to converge to less preferred local optima or overfit to the training set.

We used the identified trigenic interactions as prior by enforcing them in the neural network structure

to improve the prediction of the growth fitness scores (see Fig. 5a for illustration of the model structure; Supplementary Methods A.4 for detailed descriptions). Depending on whether we grouped the SNPs according to the genes as the input to the model, we call these models with enforced structures as Sparse (trigenic)-gene and Sparse (trigenic)-SNP, respectively. To make a fair comparisons, we only considered the SNPs in the 1,395 genes included in Kuzmin et al. (2018). As for the baseline methods in our comparisons, we built a two layered neural network with two fully connected layer. The baseline methods are called Dense-gene and Dense-SNP, respectively. The dimension of the hidden layer is tuned between 8 to 1024 for the best performance and is chosen as 256.

As shown in Fig. 5b, each model was trained and evaluated for 10 times. In general, either using the raw SNP frequency as features or aggregating them based on genes in prior, the model with enforced structure based on the identified trigenic interactions (marked as Sparse (trigenic)-SNP and Sparse (trigenic)-gene) outperforms the DNNs without the sparsity constraints (marked as Dense-SNP and Dense-gene). When comparing the Pearson and Spearman correlations under each condition between the model with and without the trigenic interaction guided sparsity constraints, we observed that the performance improvement is consistent across all conditions (Fig. 5c-d). To further assess if the improvement comes from the prior knowledge of the relationships between genotypes to phenotypes in the trigenic interactions or merely the sparsity constraints that reduce overfitting, we constructed another baseline model with the same structure of Sparse (trigenic)-SNPm, with a random structure enforced in the sparse weight matrix. We call this model as Sparse (random)-SNP. We kept the sparsity ratio of this random sparse weight matrix the same as the trigenic interaction guided one. Importantly, we observed that with the same sparsity ratio, the model embedded with the trigenic interaction guided structures outperforms the one with a random structure (Fig. 5b). Also, the Sparse (random)-SNP model performs similarly compared with the Dense-SNP model, suggesting that the two Sparse (trigenic) models indeed benefit from the prior knowledge within the trigenic interactions instead of the sparsity constraint.

These prediction results demonstrate that the trigenic interactions identified by DANGO can serve as an effective prior for the prediction of growth fitness score under various conditions. This also provides strong evidence that DANGO can transfer the model trained under one condition to other settings with the potential to better understand complex genotype-to-phenotype relationships in a wide range of biological contexts.

Discussion

In this work, we developed DANGO based on a self-attention hypergraph neural network to effectively model trigenic interactions. To our knowledge, DANGO is the first algorithm specifically designed for predicting higher-order genetic interactions. DANGO achieved significantly improved performance compared with the models generalized from the state-of-the-art methods for pairwise interactions. We showed that DANGO can accurately predict trigenic interaction scores with high correlation even when the model is being trained on a noisy dataset. We utilized DANGO to make *de novo* predictions of potential trigenic interactions and demonstrated that these newly identified trigenic interactions are enriched with relevant biological processes and protein complexes that are important to cell growth. It is important to note that in addition to the two sets of predicted trigenic interaction scores produced in this project, we also made predictions on a third set of triplets (data not shown but will be made available), where two genes come from the observed 1,395 genes and the third one is an unobserved gene, drastically expanding the repertoire of possible trigenic interactions for future experimental validations. Our results suggest that these identified trigenic interactions from DANGO have the potential to greatly facilitate the discoveries of novel biological pathways and protein complexes. We also demonstrated that these

identified trigenic interactions can be used as a prior for predicting yeast growth under unseen conditions and improve the prediction performance. Additionally, by denoising the measured trigenic interaction scores and making *de novo* predictions of unmeasured interactions, DANGO provides new insights into the mechanisms of the synergistic effects of genes on the phenotypes. Finally, although our main focus in this work is the trigenic interactions in yeast, the DANGO algorithm can be generalized to the genetic interactions in other species (including human).

There are several directions to further improve DANGO to analyze higher-order genetic interactions. First, we could incorporate more modalities into the framework, e.g., adding the sequence features of each gene to the node features. This would allow us to reveal the related sequence patterns, which would be critically important for the understanding of the principles and mechanisms of genetic interactions. Another potential extension of DANGO is to model more phenotypic traits. In this work, we only focused on the growth fitness of yeast strain quantified by the colony size. Our results from the prediction of the phenotypes under unseen conditions demonstrate the potential of using DANGO to predict other phenotype traits such as human disease phenotypes. In addition, the model framework could be extended to specifically investigate duplicated genes as demonstrated very recently in Kuzmin et al. (2020). On the algorithmic side, DANGO can be improved in terms of its computational cost. Compared with wet-lab experiments, enumerating all potential high-order interactions is much more feasible for computational method like DANGO. However, under the effect of combinatorial explosion, it could still be made more efficient to make all predictions ($\sim$8 hours to enumerate all triplets on 1,395 genes in this work). It would be highly useful development if techniques such as buffering frequently-used calculation results and training low-precision models using AMP (Automatic Mixed Precision) could be incorporated into the DANGO implementation.

We expect that DANGO will serve as an effective computational framework for the modeling of high-order genetic interaction data to further the study of the complex genotype-to-phenotype relationships. DANGO has the potential to accelerate the progress of unveiling a more complete map of complex genetic interactions that impinge upon phenotypic diversity in a wide range of biological contexts.

Method

Definitions

Definition 1. (Digenic/Trigenic interactions)
A higher-order genetic interaction is measured by the differences between the phenotypes from the joint deletion of the multiple genes and the expected phenotypes from single gene deletions (Kuzmin et al., 2018). For the yeast data we studied, the phenotype of a strain (cells of the same genotype) refers to the growth fitness where the colony size is commonly used as a proxy. The *digenic interaction score* (i.e., pairwise interaction) ϵ_{ij} between mutants i and j is quantified as: $\epsilon_{ij} = f_{ij} - (f_i f_j)$, where f_{ij}, f_i, f_j represent the measured growth fitness of the strain with double or single mutations (i and/or j). Following this convention, the *trigenic interaction* score τ is defined as (Kuzmin et al., 2018): $\tau_{ijk} = f_{ijk} - f_i f_j f_k - \epsilon_{jk} f_i - \epsilon_{ik} f_j - \epsilon_{ij} f_k$. The trigenic interaction score not only considers the expected fitness based on single mutations, but also accounts for the expected effects from lower-order interactions.

Definition 2. (Hypergraph) A *hypergraph* can be formally defined as $G = (V, E)$, where $V = \{v_1, ..., v_n\}$ represents the set of nodes in the graph, and $E = \{e_i = (v_1^{(i)}, ..., v_k^{(i)})\}$ represents the set of hyperedges (Zhou et al., 2007). A *hyperedge* e can connect more than two nodes ($|e| \geq 2$).

Definition 3. (The hyperedge prediction problem) One of the most common challenges for hypergraph representation learning is the *hyperedge prediction problem* (Zhou et al., 2007; Tu et al., 2018; Zhang et al., 2020). It aims to predict the probability of a group of nodes $(v_1, v_2, ..., v_k)$ forming a hyperedge or the attributes of the hyperedge given the features of these nodes $(E_1, E_2, ..., E_k)$.

GNN pre-training for generating node embeddings

We introduce how the node embeddings are generated based on the pairwise molecular interaction networks. In this work, we use six different types of protein-protein interaction (PPI) networks in *S. cerevisiae* from the STRING database (Franceschini et al., 2012), which are generated based on high-throughput interaction assays, curated PPI databases, and co-expression data. Each of these networks consists of G nodes (i.e., genes) with weighted edges representing a specific type of interactions between the corresponding gene pairs. Since all operations described in this section on each network are the same and work in parallel, we will not specify different networks. For each of these six networks, a 2-layered GNN (Hamilton et al., 2017) is pre-trained to reconstruct the graph structures. Specifically, for a node i, its initial feature vector $H_i^{(0)}$ is obtained through an embedding layer shared across six GNNs. The weight of the embedding layer $H^{(0)}$ has size of $G \times D$, where G is the number of genes and D is the feature dimensions. For each layer in the GNN, to generate the output vector for node i, the input vectors of its neighbors $\boldsymbol{\mathcal{N}}_i$ are aggregated:

$$H_{\boldsymbol{\mathcal{N}}(i)}^{(n)} = \text{Average}\left(\{H_u^{(n-1)}, u \sim \boldsymbol{\mathcal{N}}(i)\}\right) \tag{1}$$

$$H_i^{(n)} = \sigma\left(W_{\text{GNN}}^{(n)} \cdot \text{Concat}(H_i^{(n-1)}, H_{\boldsymbol{\mathcal{N}}(i)}^{(n)}\right) \tag{2}$$

where $H_i^{(n)}$ is the output vector of the node i at the n-th layer of the GNN, $W_{\text{GNN}}^{(n)}$ represents the weight matrix to be optimized at the n-th layer, and σ represents the non-linear activation function. The output of each node from the last layer of the GNN (node embeddings) further passes through a fully connected layer to reconstruct the corresponding rows in the original adjacency matrix. Thus, we can write the

corresponding loss term L_i for the node i as:

$$L_i = \text{weighted MSE}\left(\text{FC}\left(H_i^{(2)}\right), w_{e(i)}\right) \tag{3}$$

$$\text{weighted MSE}(x, z) = \frac{1}{N}\left[\sum_{n=1}^{N}(x_n - z_n)^2 \boldsymbol{\mathcal{I}}(z_n \neq 0) + \lambda \sum_{n=1}^{N}(x_n - z_n)^2 \boldsymbol{\mathcal{I}}(z_n = 0)\right] \tag{4}$$

where FC represents the fully connected layer, $w_{e(i)}$ is the i-th row of the network adjacency matrix. z_n is the n-th element of the vector z of size N. The weighted mean-squared error (MSE) function penalizes the zero entries and non-zero entries in the ground truth differently. We used this approach instead of the standard mean-squared error function as the reconstruction loss because a zero in the molecular interaction networks does not necessarily mean no interaction between a pair of genes. The hyperparameter λ reflects to which degree the zeros in the network are considered no interactions. In this work, we calculated the percentage of decreased zeroes from STRING database v9.1 to v11.0 for each network, ranging from 0.02% (co-occurrence) to 2.42% (co-expression). This metric indicates the percentage of the zeros in the original molecular interaction network that can potentially be validated in the future. Therefore, if the decreased percentage of zeros is greater than 1%, we set $\lambda = 0.1$; otherwise, we set $\lambda = 1.0$.

Meta embedding learning for the integration of multiple embeddings

The six pre-trained GNNs described in the previous section generate node embeddings that capture the neighborhood topology in each PPI network. In DANGO, we also develop a meta embedding learning module to integrate embeddings of the same node across different networks. For a node i, we denote all its node embeddings (six in total) from the GNNs as $\{E_i^{(1)}, E_i^{(2)}, ..., E_i^{(6)}\}$. The meta embedding for that node i is calculated as:

$$E_i = \sum_{k=1}^{6} E_i^{(k)} \frac{\exp\left[\text{MLP}\left(E_i^{(k)}\right)\right]}{\sum_{j=1}^{6} \exp\left[\text{MLP}\left(E_i^{(j)}\right)\right]} \tag{5}$$

where MLP represents the multi-layer perceptron that consists of two fully-connected layers. Compared with simpler methods to integrate embeddings, such as averaging or concatenation, the meta embedding learning methods with trainable weights have the advantage of addressing challenges such as unseen entities or inconsistency of embedding qualities across different sources (Kiela et al., 2018; Xie et al., 2019; Liu et al., 2020). In DANGO, there are genes in certain PPI network that do not have observed interactions with other genes, which, as mentioned, do not necessarily mean no interactions. Both averaging or concatenation would make such embedding as important as the embeddings from other PPI networks with more information of that gene. Moreover, to predict the trigenic interactions, the most relevant properties characterized by the PPI networks from different sources may vary across genes. For instance, some genes involve in trigenic interactions as they are functionally interchangeable and only deleting all three would result in the loss of function, while others could result from the formation of protein complex. Our meta embedding module with trainable weights will allow the model to capture the variable importance of embeddings from different PPI networks.

Hypergraph representation learning to predict trigenic interactions

Here we briefly introduce the structure of Hyper-SAGNN (Fig. 1b; more details in Zhang et al. (2020)) and the modification we made to improve its performance. Each input sample of the model is a tuple

consisting of the meta-embeddings of three genes, i.e., (E_1, E_2, E_3). All these embeddings first go through the same feed-forward neural network with one fully connected layer, respectively, to produce (s_1, s_2, s_3), where $s_i = \sigma\left(\text{FC}\left(E_i\right)\right)$. s_i is called the static embedding of node i since it remains the same for gene i independent to the given triplet. The input also passes through two multi-head self-attention layers (Vaswani et al., 2017), which result in (d_1, d_2, d_3). Each d_i is called the dynamic embedding because it depends on all the node features within this triplet and is thus variable across different triplets. We denote the weight matrices within each self-attention layer as W_Q, W_K, W_V, which represent the linear transformation of features before applying the scaled dot-product attention (Vaswani et al., 2017). The calculation of d_i can be summarized as:

$$\hat{\alpha}_{ij} = \left(W_Q^T E_i\right)^T \left(W_K^T E_j\right), \forall 1 \leq i, j \leq 3, i \neq j \tag{6}$$

$$\hat{d}_i = \sigma\left(\sum_{1 \leq j \leq 3, j \neq i} \frac{\exp(\hat{\alpha}_{ij})}{\sum_{1 \leq l \leq 3, l \neq i} \exp(\hat{\alpha}_{il})} W_V^T E_j\right) \tag{7}$$

where $\hat{\alpha}_{ij}$ represent the attention coefficients before softmax normalization. $\hat{d}_i$ is calculated based on these coefficients as the weighted sum of linear transformed features with activation function.

We improve the self-attention layer with the ReZero technique introduced in Bachlechner et al. (2020). Consistent with the observation in Bachlechner et al. (2020), ReZero accelerates the convergence of the model and allows us to potentially use more stacked self-attention layers than the original Hyper-SAGNN. ReZero can be described as: $d_i = \beta \hat{d}_i + E_i$, where β is also a parameter initialized as 0 as suggested by ReZero and optimized through the training. Next, the Hadamard power (element-wise power) of the difference of the corresponding static/dynamic pair for each node is calculated and then sent through a one-layered neural network with linear activation function to produce a scalar value $\hat{y}_i$. Finally, all $\hat{y}_i$ are averaged to reach the final predicted trigenic interaction score $\hat{y}_{1,2,3}$, i.e.,

$$\hat{y}_{1,2,3} = \frac{1}{3}\sum_{i=1}^{3} \hat{y}_i = \frac{1}{3}\sum_{i=1}^{3} \sigma\left\{\text{FC}\left[\left(d_i - s_i\right)^{\circ 2}\right]\right\} \tag{8}$$

The hypergraph representation learning module is trained with the log cosh loss with all other components of DANGO in an end-to-end manner.:

$$L = \frac{1}{N}\sum_{(i,j,k) \in E} \log\left[\cosh(\hat{y}_{i,j,k} - y_{i,j,k})\right] \tag{9}$$

The log cosh regression loss performs similarly as the mean squared error around 0 and as the mean absolute error for larger target values, making it robust to the outliers while having stable gradients around zeros (Neuneier and Zimmermann, 1998).

Code Availability

Source code of DANGO can be accessed at: https://github.com/ma-compbio/DANGO.

Acknowledgements

This work was supported in part by the National Institutes of Health grant R01HG007352 (J.M.), National Institutes of Health Common Fund 4D Nucleome Program grant UM1HG011593 (J.M.), and National Science Foundation grant 1717205 (J.M.). J.M. is additionally supported by a Guggenheim Fellowship from the John Simon Guggenheim Memorial Foundation.

Author Contributions

Conceptualization, R.Z., J-Z.M., and J.M.; Methodology, R.Z., J-Z.M., and J.M.; Software, R.Z.; Investigation, R.Z., J-Z.M., and J.M.; Writing – Original Draft, R.Z., J-Z.M., and J.M.; Writing – Review & Editing, R.Z., J-Z.M., and J.M.; Funding Acquisition, J.M.

Competing Interests

The authors declare no competing interests.

References

T. Bachlechner, B. P. Majumder, H. H. Mao, G. W. Cottrell, and J. McAuley. Rezero is all you need: Fast convergence at large depth. *arXiv preprint arXiv:2003.04887*, 2020.

Z. Bao, H. Xiao, J. Liang, L. Zhang, X. Xiong, N. Sun, T. Si, and H. Zhao. Homology-integrated CRISPR–Cas (HI-CRISPR) system for one-step multigene disruption in Saccharomyces cerevisiae. *ACS Synthetic Biology*, 4(5):585–594, 2015.

C. T. Beh, L. Cool, J. Phillips, and J. Rine. Overlapping functions of the yeast oxysterol-binding protein homologues. *Genetics*, 157(3):1117–1140, 2001.

C. Boone, H. Bussey, and B. J. Andrews. Exploring genetic interactions and networks with yeast. *Nature Reviews Genetics*, 8(6):437–449, 2007.

A. Celaj, M. Gebbia, L. Musa, A. G. Cote, J. Snider, V. Wong, M. Ko, T. Fong, P. Bansal, J. C. Mellor, et al. Highly combinatorial genetic interaction analysis reveals a multi-drug transporter influence network. *Cell Systems*, 10(1):25–38, 2020.

H. Cho, B. Berger, and J. Peng. Compact integration of multi-network topology for functional analysis of genes. *Cell Systems*, 3(6):540–548, 2016.

M. Costanzo, B. VanderSluis, E. N. Koch, A. Baryshnikova, C. Pons, G. Tan, W. Wang, M. Usaj, J. Hanchard, S. D. Lee, et al. A global genetic interaction network maps a wiring diagram of cellular function. *Science*, 353(6306), 2016.

M. Costanzo, E. Kuzmin, J. van Leeuwen, B. Mair, J. Moffat, C. Boone, and B. Andrews. Global genetic networks and the genotype-to-phenotype relationship. *Cell*, 177(1):85–100, 2019.

L. Cowen, T. Ideker, B. J. Raphael, and R. Sharan. Network propagation: a universal amplifier of genetic associations. *Nature Reviews Genetics*, 18(9):551, 2017.

J. Domingo, P. Baeza-Centurion, and B. Lehner. The causes and consequences of genetic interactions (epistasis). *Annual Review of Genomics and Human Genetics*, 20:433–460, 2019.

F.-E. Eid, H. Elmarakeby, Y. A. Chan, N. F. Martins, M. Elhefnawi, E. Van Allen, L. S. Heath, and K. Lage. Systematic auditing is essential to debiasing machine learning in biology. *bioRxiv*, 2020.

S. Feyder, J.-O. De Craene, S. Bär, D. L. Bertazzi, and S. Friant. Membrane trafficking in the yeast saccharomyces cerevisiae model. *International Journal of Molecular Sciences*, 16(1):1509–1525, 2015.

A. Franceschini, D. Szklarczyk, S. Frankild, M. Kuhn, M. Simonovic, A. Roth, J. Lin, P. Minguez, P. Bork, C. Von Mering, et al. STRING v9. 1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Research*, 41(D1):D808–D815, 2012.

A. D. Garst, M. C. Bassalo, G. Pines, S. A. Lynch, A. L. Halweg-Edwards, R. Liu, L. Liang, Z. Wang, R. Zeitoun, W. G. Alexander, et al. Genome-wide mapping of mutations at single-nucleotide resolution for protein, metabolic and genome engineering. *Nature Biotechnology*, 35(1):48–55, 2017.

W. Hamilton, Z. Ying, and J. Leskovec. Inductive representation learning on large graphs. In *Advances in Neural Information Processing Systems*, pages 1024–1034, 2017.

M. Kanehisa and S. Goto. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 28(1):27–30, 2000.

D. Kiela, C. Wang, and K. Cho. Dynamic meta-embeddings for improved sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Brussels, Belgium, 2018.

E. Kuzmin, B. VanderSluis, W. Wang, G. Tan, R. Deshpande, Y. Chen, M. Usaj, A. Balint, M. M. Usaj, J. Van Leeuwen, et al. Systematic analysis of complex genetic interactions. *Science*, 360(6386), 2018.

E. Kuzmin, B. VanderSluis, A. N. N. Ba, W. Wang, E. N. Koch, M. Usaj, A. Khmelinskii, M. M. Usaj, J. van Leeuwen, O. Kraus, et al. Exploring whole-genome duplicate gene retention with complex

genetic interaction analysis. *Science*, 368(6498), 2020.

J. Lian, M. HamediRad, S. Hu, and H. Zhao. Combinatorial metabolic engineering using an orthogonal tri-functional CRISPR system. *Nature Communications*, 8(1):1–9, 2017.

Q. Liu, J. Lu, G. Zhang, T. Shen, Z. Zhang, and H. Huang. Domain-specific meta-embedding with latent semantic structures. *Information Sciences*, 2020.

R. Neuneier and H. G. Zimmermann. How to train neural networks. In *Neural networks: tricks of the trade*, pages 373–423. Springer, 1998.

J. Parenteau, M. Durand, S. Véronneau, A.-A. Lacombe, G. Morin, V. Guérin, B. Cecez, J. Gervais-Bird, C.-S. Koh, D. Brunelle, et al. Deletion of many yeast introns reveals a minority of genes that require splicing for function. *Molecular Biology of the Cell*, 19(5):1932–1941, 2008.

J. Peter, M. De Chiara, A. Friedrich, J.-X. Yue, D. Pflieger, A. Bergström, A. Sigwalt, B. Barre, K. Freel, A. Llored, et al. Genome evolution across 1,011 Saccharomyces cerevisiae isolates. *Nature*, 556 (7701):339–344, 2018.

S. Pu, J. Wong, B. Turner, E. Cho, and S. J. Wodak. Up-to-date catalogues of yeast protein complexes. *Nucleic Acids Research*, 37(3):825–831, 2009.

K. E. Sawin and P. Tran. Cytoplasmic microtubule organization in fission yeast. *Yeast*, 23(13):1001–1014, 2006.

C. Stark, B.-J. Breitkreutz, T. Reguly, L. Boucher, A. Breitkreutz, and M. Tyers. BioGRID: a general repository for interaction datasets. *Nucleic Acids Research*, 34(suppl_1):D535–D539, 2006.

Y. Suzuki, R. P. St Onge, R. Mani, O. D. King, A. Heilbut, V. M. Labunskyy, W. Chen, L. Pham, L. V. Zhang, A. H. Tong, et al. Knocking out multigene redundancies via cycles of sexual assortment and fluorescence selection. *Nature Methods*, 8(2):159–164, 2011.

K. Tu, P. Cui, X. Wang, F. Wang, and W. Zhu. Structural deep embedding for hyper-networks. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017.

Y. Xie, Y. Hu, L. Xing, and X. Wei. Dynamic task-specific factors for meta-embedding. In *International Conference on Knowledge Science, Engineering and Management*, pages 63–74. Springer, 2019.

H. Zhang, H. Pan, C. Zhou, Y. Wei, W. Ying, S. Li, G. Wang, C. Li, Y. Ren, G. Li, et al. Simultaneous zygotic inactivation of multiple genes in mouse through CRISPR/Cas9-mediated base editing. *Development*, 145(20), 2018.

R. Zhang, Y. Zou, and J. Ma. Hyper-SAGNN: a self-attention based graph neural network for hypergraphs. In *International Conference on Learning Representations (ICLR)*, 2020.

Y. Zhang, J. Wang, Z. Wang, Y. Zhang, S. Shi, J. Nielsen, and Z. Liu. A gRNA-tRNA array for CRISPR-Cas9 based rapid multiplexed genome editing in Saccharomyces cerevisiae. *Nature Communications*, 10(1):1–10, 2019.

D. Zhou, J. Huang, and B. Schölkopf. Learning with hypergraphs: Clustering, classification, and embedding. In *Advances in Neural Information Processing Systems*, pages 1601–1608, 2007.

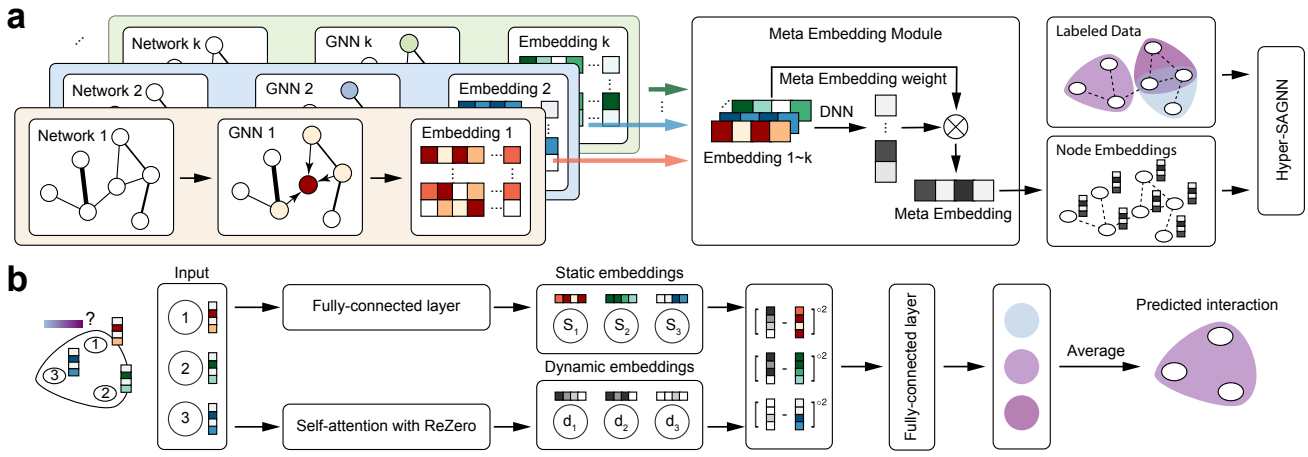

Figure 1: Overview of the DANGO algorithm. **a.** Workflow of DANGO. DANGO takes multiple pairwise molecular interaction networks as input and pre-trains multiple graph neural networks to generate node embeddings. Embeddings for the same node across different networks are integrated through a meta embedding learning scheme. Both the integrated node embeddings and the labeled data from labeled trigenic interaction datasets are used as input to train a hypergraph representation learning framework. **b.** A modified Hyper-SAGNN architecture (Zhang et al., 2020)) for hyperedge regression. The input of the model contains gene triplets with the corresponding node embedding features. The triplets pass through the fully-connected layers and the modified self-attention layers to generate static and dynamic node embeddings, respectively. A pseudo euclidean distance is then calculated for each pair of static/dynamic embeddings that will be averaged to produce the final predicted trigenic interaction score.

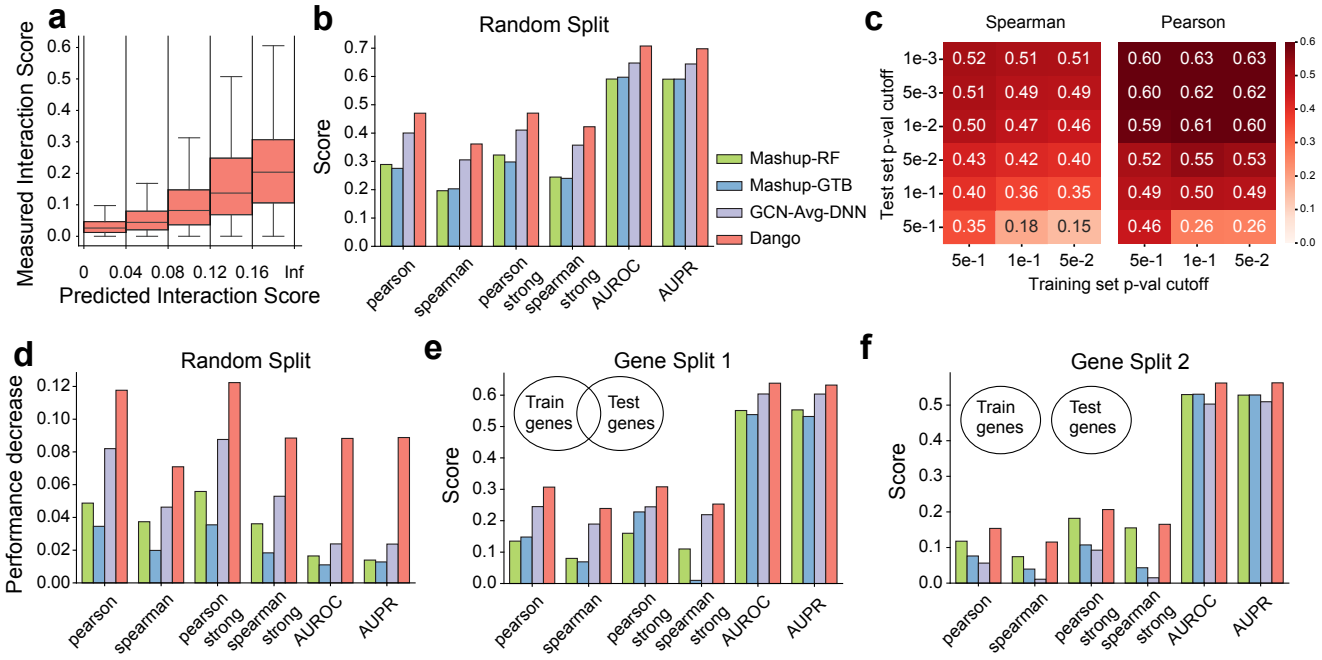

Figure 2: Evaluation of the performance on trigenic interaction score predictions. **a**. Measured versus predicted trigenic interaction scores for each triplet gene disruption genotype. **b**. Evaluation of the predicted trigenic interaction scores with random partition cross validation. The prediction performance is evaluated using Pearson and Spearman correlation scores, Pearson and Spearman correlation scores within strong trigenic interactions ($(|\tau| > 0.05)$), and AUROC/AUPR scores of the predicted interaction value versus whether the measured interaction score is strong or not. **c**. Performance changes when using different *P*-value cutoffs on the training and test datasets. The *P*-values reflect the consistency of trigenic interaction scores calculated on two replicates. **d**. Performance decreases when we randomly shuffle the learned gene embeddings (for GNN based methods, the adjacency matrix of the input network is shuffled) under random cross validations. **e, f**. Evaluation of the predicted trigenic interaction scores with two types of gene based training and testing set split. In Gene Split 1, around 60 genes in the test set are unobserved in the training data. In Gene Split 2, all 400 genes in the test set are unobserved in the training data.

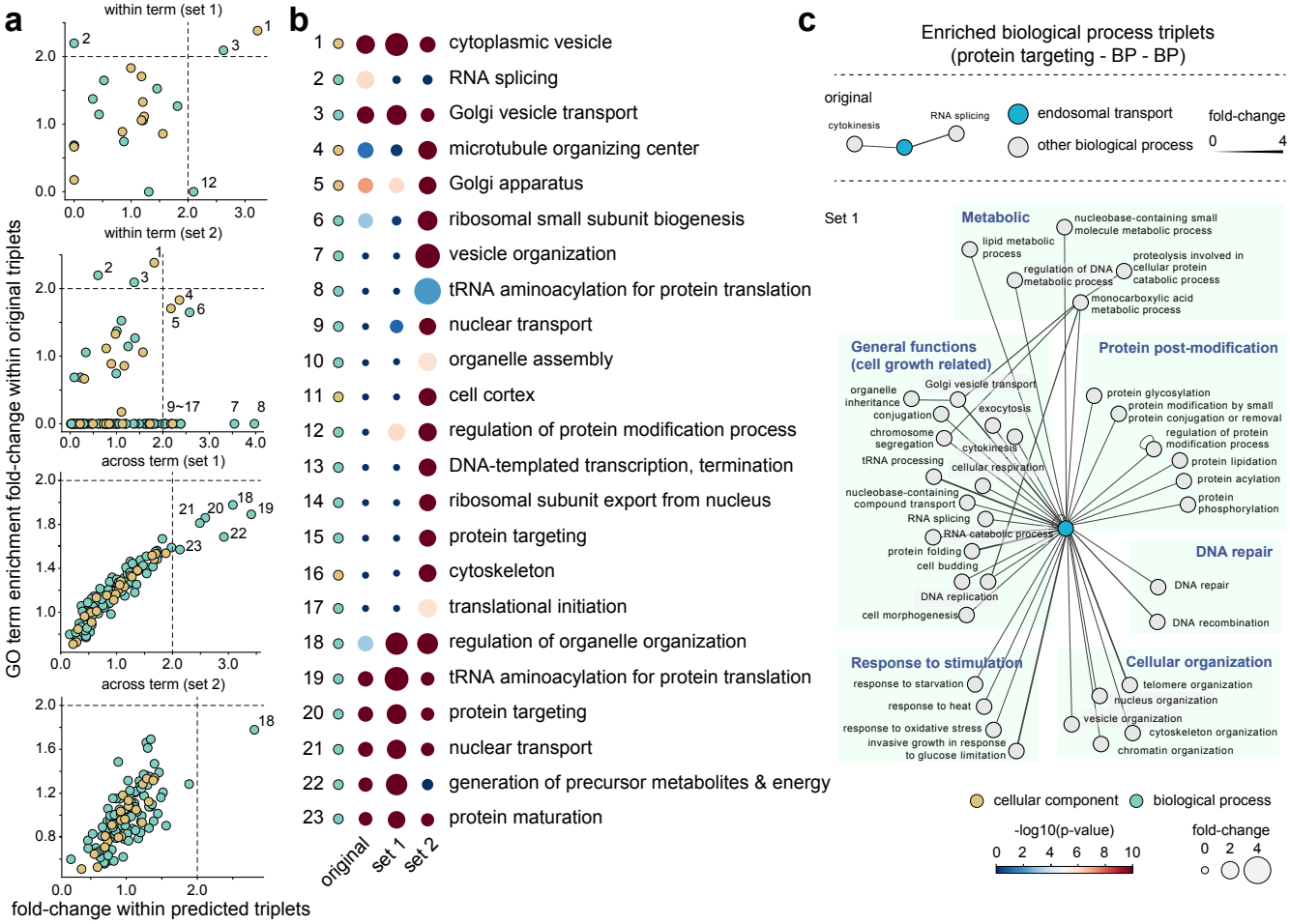

Figure 3: Functional characterization of the predicted trigenic interactions. **a.** Correlation of the frequency fold-change of trigenic interactions within/across GO terms between the predicted and measured trigenic interactions. There are two sets of predicted trigenic interactions. The set #1 corresponds to the same set of triplets measured in Kuzmin et al. (2018). The set #2 corresponds to all possible combinations of three genes that appear in Kuzmin et al. (2018). **b.** Enriched GO terms across different set of trigenic interactions. All terms with frequency fold-change greater than 2.0 in any set of trigenic interactions are visualized with the frequency fold-change and *P*-value (hypergeometric test). **c.** DANGO discovers a functionally important trigenic interaction hub. The networks visualize examples of the enriched combinations of biological process terms. Specifically, the predicted trigenic interactions in set #1 discovers a trigenic interaction hub unobserved in the original dataset, which consists of genes related to protein targeting, endosomal transport, and other biological process (such as protein post-modification, DNA repair, etc.

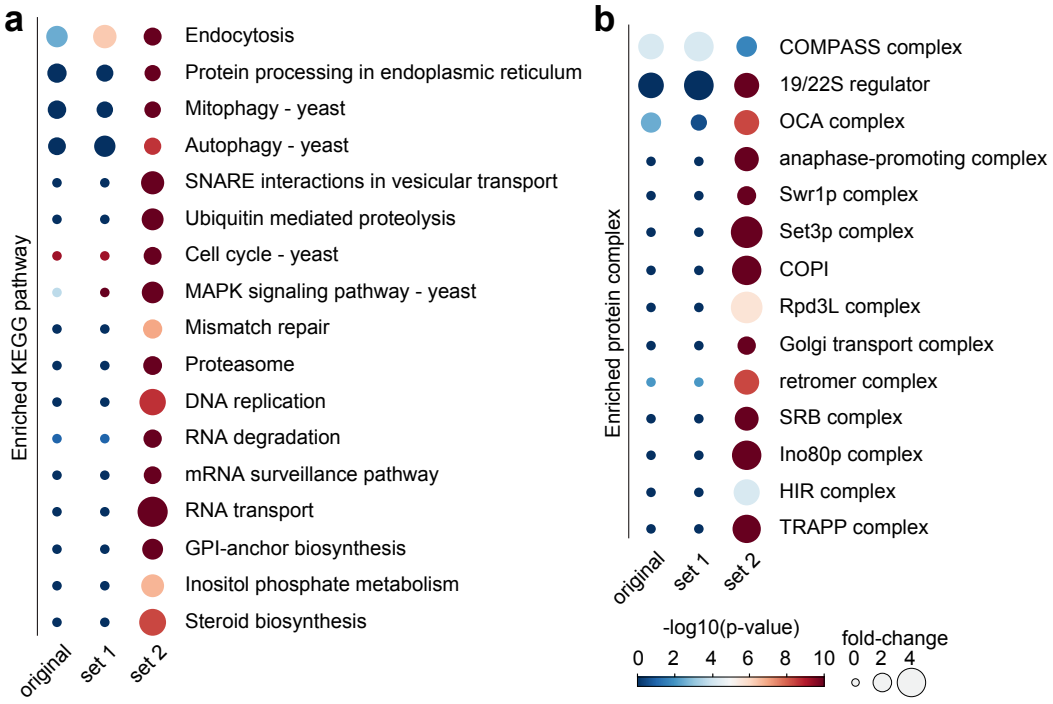

Figure 4: a. Enriched KEGG pathways across different sets of trigenic interactions. **b.** Enriched yeast complex across different sets of trigenic interactions. All pathways and yeast complexes are visualized with the frequency fold-change and *P*-value assessed with the hypergeometric test.

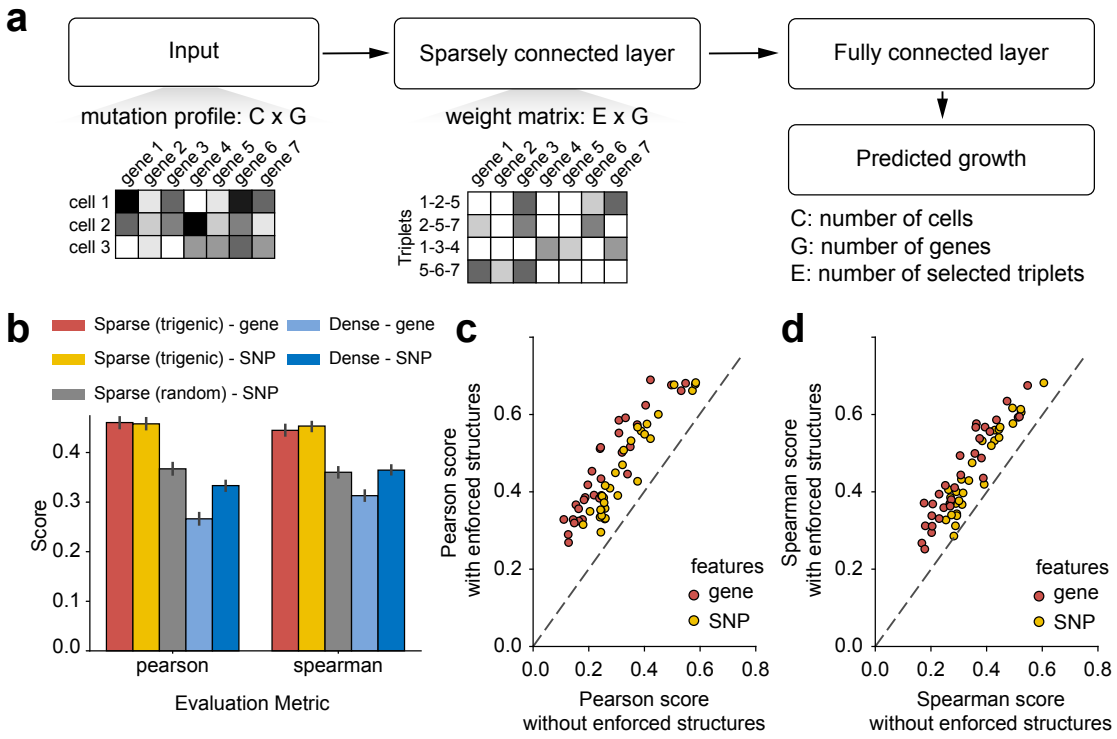

Figure 5: Trigenic interactions identified by DANGO can predict yeast growth under unseen culture conditions. **a**. The illustration of how to embed the selected trigenic interactions into the structure of a deep neural network (DNN). **b**. Performance evaluation of the trigenic interaction guided DNNs and the baseline models for predicting yeast growth. The performances are aggregated across all 35 conditions. **c, d.** Performance evaluation of predicting yeast growth from the trigenic interaction guided DNNs versus the baseline models under each condition.