PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

# FUN-PROSE: A deep learning approach to predict condition-specific gene expression in fungi

**Ananthan Nambiar**[1,2]*, **Veronika Dubinkina**[1,2,3], **Simon Liu**[2,4], **Sergei Maslov**[1,2,5,6]*

**1** Department of Bioengineering, University of Illinois Urbana-Champaign, Urbana, Illinois, United States of America, **2** Carl R. Woese Institute for Genomic Biology, Urbana, Illinois, United States of America, **3** The Gladstone Institute of Data Science and Biotechnology, San Francisco, California, United States of America, **4** Department of Computer Science, University of Illinois Urbana-Champaign, Urbana, Illinois, United States of America, **5** Department of Physics, University of Illinois Urbana-Champaign, Urbana, Illinois, United States of America, **6** Computing, Environment and Life Sciences, Argonne National Laboratory, Lemont, Illinois, United States of America

* nambiar4@illinois.edu (AN); maslov@illinois.edu (SM)





**Citation:** Nambiar A, Dubinkina V, Liu S, Maslov S (2023) FUN-PROSE: A deep learning approach to predict condition-specific gene expression in fungi. PLoS Comput Biol 19(11): e1011563. https://doi.org/10.1371/journal.pcbi.1011563

**Editor:** Marc Robinson-Rechavi, Universite de Lausanne Faculte de biologie et medecine, SWITZERLAND

**Received:** February 19, 2023

**Accepted:** September 30, 2023

**Published:** November 16, 2023



**Data Availability Statement:** All code and processed data for simulations used in this manuscript can be found at https://github.com/maslov-group/FUN-PROSE.

**Funding:** This work was funded by the DOE Center for Advanced Bioenergy and Bioproducts Innovation (U.S. Department of Energy, Office of Science, Office of Biological and Environmental Research under Award Number DE-SC0018420). Any opinions, findings, and conclusions or recommendations expressed in this publication are

## Abstract

mRNA levels of all genes in a genome is a critical piece of information defining the overall state of the cell in a given environmental condition. Being able to reconstruct such condition-specific expression in fungal genomes is particularly important to metabolically engineer these organisms to produce desired chemicals in industrially scalable conditions. Most previous deep learning approaches focused on predicting the average expression levels of a gene based on its promoter sequence, ignoring its variation across different conditions. Here we present FUN-PROSE—a deep learning model trained to predict differential expression of individual genes across various conditions using their promoter sequences and expression levels of all transcription factors. We train and test our model on three fungal species and get the correlation between predicted and observed condition-specific gene expression as high as 0.85. We then interpret our model to extract promoter sequence motifs responsible for variable expression of individual genes. We also carried out input feature importance analysis to connect individual transcription factors to their gene targets. A sizeable fraction of both sequence motifs and TF-gene interactions learned by our model agree with previously known biological information, while the rest corresponds to either novel biological facts or indirect correlations.

## Author summary

In this paper we develop a deep learning method to predict condition specific gene expression in various kinds of fungi ranging from baker's yeast to red bread mold. Predicting condition-specific gene expression is useful because it measures the response of organisms to different environmental conditions. Among other uses, our model would allow us to predict the effects of TF knockout experiments and discover novel genes that play an important role in a given environment. In addition, our framework allows us to predict

those of the authors and do not necessarily reflect the views of the U.S. Department of Energy. This funder partially paid the salaries of A.N. and V.D. Part of this work was performed under the auspices of the U.S. Department of Energy by Argonne National Laboratory under Contract DE-AC02-06-CH11357. This funder partially paid the salary of S.M. This work utilizes resources supported by the National Science Foundation's Major Research Instrumentation program, grant #1725729, as well as the University of Illinois at Urbana-Champaign. S.L. has been supported by the James Scholar Honors Program and the Illinois Scholars Undergraduate Research Program. V.D. has been supported by the San Simeon Fund. The funder partially paid the salary of V.D. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**Competing interests:** The authors have declared that no competing interests exist.

how an organism will express genes in a novel environment. Being able to make these predictions is an important part of the metabolic engineering of fungi to optimize their use in the production of desired chemicals in industrially scalable conditions.

## Introduction

Transcriptional regulation of gene expression is one of the key mechanisms used by biological organisms in general and fungi in particular to modify their phenotype in response to changes in the environment. Protein abundances directly responsible for the phenotypic state of the cell are known to be strongly correlated with mRNA levels of the corresponding genes (for fungi, see e.g. [1–4]). Therefore, investigating how genes are differentially expressed in specific conditions are important in understanding how an organism regulates its response to different conditions. Hence, the ability to predict condition-specific mRNA expression of relevant genes is a crucial step for developing industrial applications of fungal species [5–7].

Deciphering the complicated process of gene regulation was one of the key research objectives during the past several decades [8–10]. Due to the synchronized work of multiple systems controlling mRNA synthesis and decay [11, 12], the average steady state levels of individual mRNAs may vary from less than one copy per cell to several hundreds per cell across different environmental conditions [13]. The majority of the information shaping this response is encoded in each gene's two cis-regulatory sequence regions [14]. One of them, referred to as the promoter, is located upstream of the protein coding sequence. It contains sequence motifs recognized by DNA-binding transcription factors (TFs) that enhance or repress mRNA gene expression. Another one is the 3' UTR sequence located downstream of the protein coding sequence and containing motifs for RNA-binding proteins responsible for mRNA stability and decay [11]. For instance, in the model fungal organism *Saccharomyces cerevisiae*, sequence properties of individual cis-regulatory regions can explain up to half of the variation in mRNA levels across conditions [14]. However, the functional relationship between expression levels of multiple TFs and their gene targets is highly non-linear, and its mechanistic details remain poorly understood, especially in eukaryotic genomes.

Deep neural networks (DNNs) have been extremely successful in learning such complex non-linear relationships in biological data. In particular, convolutional networks (CNNs) are specifically suitable to learn hierarchical patterns in sequence data such as promoters and 3' UTR regions of individual genes. CNNs were previously applied to extract TF-binding motifs and their higher-order organizational context from ChIP–seq [15, 16], ChIP-exo [17], and artificial sequence experiments [18]. Other DNN architectures, e.g. fully connected perceptrons, were used to work with other biological data and prediction tasks, e.g., learning the internal state of a cell from gene expression counts [19]. DNNs were also previously used to predict the *average* mRNA level of a gene across many conditions based only on its cis-regulatory sequences [20, 21]. Alternatively, there has been work done that incorporates data from transcription factor-DNA binding assays [22], which is not readily available for many fungal species, to predict expression. However, these studies did not address the question of predicting condition-specific gene expression using sequence information, which is the main subject of our study. Predicting condition-specific gene expression is useful because it measures the response of organisms to different conditions. Among other uses, it could allow us to predict the effects of TF knockout experiments and discover novel genes that play an important role in a given environment.

Here we present a broadly applicable DNN model called FUN-PROSE (FUNgal PRomoter to cOndition-Specific Expression) which was trained to predict the relative expression level of a gene in a specific condition based on the gene's promoter sequence and the expression levels of all TFs of a given fungal species. We tested our model on existing gene expression datasets for three different fungal species and demonstrated its practical applicability not only for model organisms such as *Saccharomyces cerevisiae* but also for less studied fungal species such as *Neurospora crassa* and *Issatchenkia orientalis*, where patterns of gene regulation remain virtually unexplored.

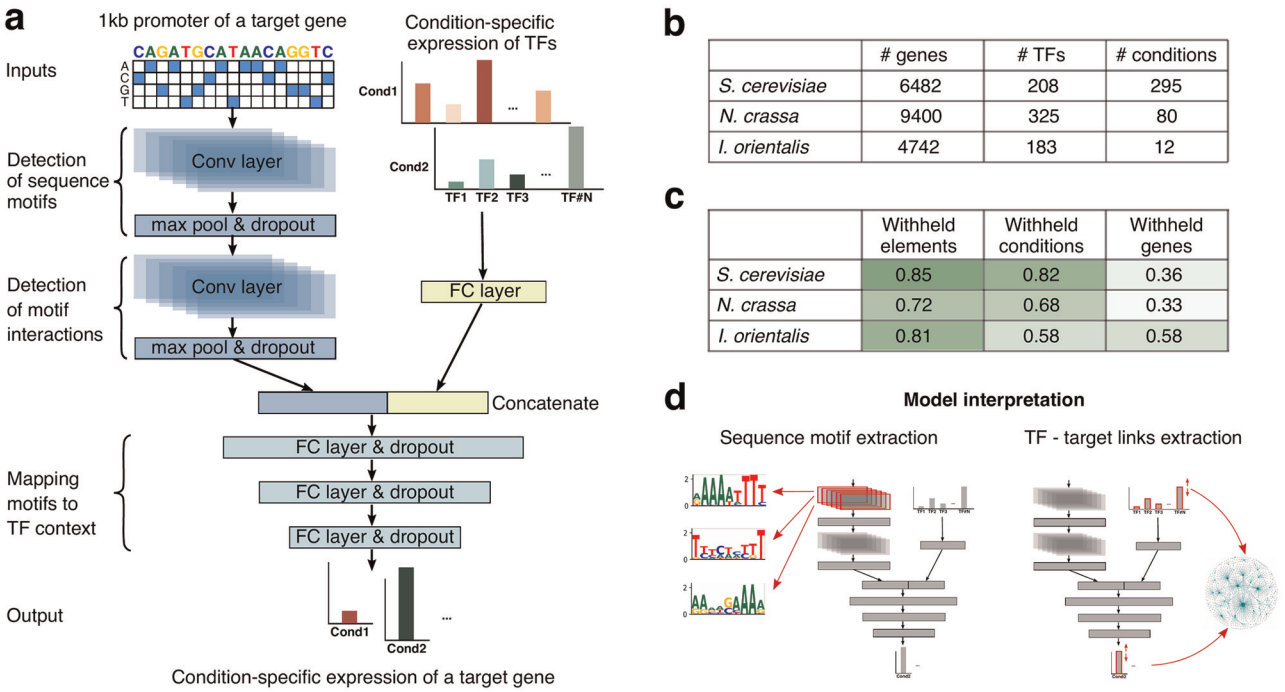
One of the challenges when using DNN models lies in mechanistic interpretation of their results and extraction of new biological knowledge from them [17, 23]. To address this challenge, we interrogate our model to identify biologically relevant information of two types. One type is composed of recurrent sequence motifs relevant for regulation of gene expression, e.g., TF-binding motifs. The other type of biological information is the Gene Regulatory Network (GRN) of a species, linking each of the TFs to their gene targets. To learn GRNs in each of our three fungal species, we used input a feature attribution technique to assign tentative TF regulators to individual genes. In *S. cerevisiae*, many sequence motifs and regulatory interactions discovered by our model agree with previously known biological information, while the rest correspond to either novel biological facts or indirect correlations.

While there has been a study that explored predicting condition specific gene expression in *S. cerevisiae* [24], our studies differ in several aspects. In particular, we perform extensive hyperparameter tuning of our algorithm and test its performance to separately predict the expression of novel genes and novel conditions. We also take a more systematic approach to model interpretation by providing the top TFs that had an effect on predictions as a whole, as well as extracting and analyzing entire gene regulatory networks to identify the TFs that were important for predicting the expression of specific genes. We also show that including 3'-UTR sequences improved the accuracy of our predictions. Finally, our work demonstrates the generalizability of these methods to non-model yeasts and multicellular fungi.

In conclusion, our model can be used to both extract new biological knowledge and to tackle a practically important task of manipulating the expression level of a given gene by either changing its promoter sequence or modifying the TF levels.

## Results

To predict the relative expression level of a gene in a particular environmental condition, we reasoned that most of the necessary information should be contained in two sets of data: the promoter sequence of this gene, which contains cis-regulatory sequences recognized by TFs, and individual expression levels of all TFs in this condition. With that in mind, we designed a deep neural network with the following architecture (see Fig 1A). The first type of inputs (i.e. promoter sequences) is processed through two convolutional layers. The first layer is designed to capture simple sequence motifs in promoters, while the second one should be able to learn combinations of these motifs to account for complex combinatorial effects (e.g. TF-TF interactions, helper proteins, etc.). Convolutional layers are used here to take advantage of their translational invariance. The second type of inputs (i.e. expression levels of all TFs) is processed through a fully-connected layer. The resulting latent representations are concatenated together and passed through several fully-connected layers that establish a connection between any of the TFs and corresponding motifs. The final layer then predicts the condition-specific gene expression level. It is important to note that the condition-specific gene expression here is defined as the Z-score of the log-transformed gene expression calculated across all conditions in our data (see Methods). That is, the expression level of each gene is standardized to have the




| | # genes | # TFs | # conditions |
|---|---|---|---|
| *S. cerevisiae* | 6482 | 208 | 295 |
| *N. crassa* | 9400 | 325 | 80 |
| *I. orientalis* | 4742 | 183 | 12 |

| | Withheld elements | Withheld conditions | Withheld genes |
|---|---|---|---|
| *S. cerevisiae* | 0.85 | 0.82 | 0.36 |
| *N. crassa* | 0.72 | 0.68 | 0.33 |
| *I. orientalis* | 0.81 | 0.58 | 0.58 |

**Fig 1. FUN-PROSE model predicts condition-specific gene expression in fungi and allows extracting transcription factor binding motifs and edges of gene regulatory networks. (a)** Schematic of the FUN-PROSE architecture. The model uses the 1000bp promoter sequence of a gene and condition-specific expression levels of all TFs in the genome as inputs and predicts the expression of this gene in a given condition. FC denotes fully connected layers; Conv denotes convolutional layers. For specific layer parameters (sizes, stride, kernels, etc.), see Table 1 and Methods. **(b)** The statistics of fungal datasets used in this study. **(c)** FUN-PROSE performance on different datasets for three different train/test set splits (see text for details). Background color represents accuracy on a test set. **(d)** Schematic of model interpretation procedures to extract sequence motifs (by analysis of the first layer of convolutional filters) and TF-gene regulatory interactions (by Integrated Gradient technique).

https://doi.org/10.1371/journal.pcbi.1011563.g001

mean of 0 and the standard deviation of 1. Thus the goal of FUN-PROSE is to predict the deviation of gene expression in a particular condition from the average expression level of that gene.

To evaluate the performance of FUN-PROSE, we collected several previously published RNA-seq datasets for different fungal species (see Fig 1B). In particular, we gathered datasets for *S. cerevisiae*, *N. crassa* and *I. orientalis* as described in Methods. The compiled *S. cerevisiae* dataset included 6482 genes, 208 TFs and 295 different stress conditions (see S1 Fig). The *N. crassa* dataset was made up of 9400 genes, 325 TFs and 80 different combinations of growth on different carbon sources and strains with gene knockouts (see S2 Fig). Finally, the *I. orientalis* dataset was our smallest with 4742 genes, 183 TFs and only 12 conditions of growth on different carbon sources (see S3 Fig).

The performance of our optimized model architecture for all three species are shown in Fig 1C. We also took steps to interrogate the network for biologically meaningful information: first is to extract sequence motifs that our DNN model learned during training, we expect some of them to correspond to transcription factor binding motifs; second is to extract TF-gene target links, i.e., edges in GRN using Integrated Gradients [25] (Fig 1D).

## Hyperparameter optimization

In order to make sure that we obtain the best possible performance out of the neural network, we tuned the hyperparameters that define our architecture (see Table 1) to maximize the Pearson correlation coefficient of the predicted and true gene expression levels. In our procedure,

**Table 1. Configuration search space for hyperparameter optimization and best hyperparameters identified in the space.** *Uniform* and *LogUniform* indicate that values are uniformly sampled from the domain and log domain of the provided range, respectively. *RandInt* indicates that values are uniformly sampled from integers in the provided range. Bracketed range indicates that values are sampled from the discrete set.

| Hyperparameter Optimiztaion Space and Results | | |
|---|---|---|
| | **Search Space** | **Best in Search** |
| Promoter Seq. Length | 100*[2, 3, 4, 5, 6, 7, 8, 9, 10] | 1000 |
| Batch Size | [64, 128, 256, 512, 1024] | 256 |
| Peak LR | *LogUniform*[1e-5, 1e-2) | $1.02*10^{-4}$ |
| Weight Decay | *LogUniform*[1e-3, 1e-1) | $1.64*10^{-3}$ |
| Conv. Layer 1 Kernel Size | [9, 11, 13, 15] | 9 |
| Conv. Layer 1 Kernel # | [16, 32, 64, 128, 256] | 256 |
| Pool. Layer 1 Kernel Size | *RandInt*[5, 20) | 19 |
| Conv. Layer 2 Kernel Size | [9, 11, 13, 15] | 13 |
| Conv. Layer 2 Kernel # | [16, 32, 64, 128, 256] | 64 |
| Pool. Layer 2 Kernel Size | *RandInt*[5, 10) | 8 |
| Conv. Activation Function | [ReLU, GeLU, ELU, SeLU] | ReLU |
| Conv. Dropout | *Uniform*[0, 0.50) | 0.398 |
| TF Hidden Layer Size | [32, 64, 128, 256, 512, 1024] | 1024 |
| TF Hidden Layer Dropout | *Uniform*[0, 0.50) | 0.189 |
| F.C. Activation Function | [ReLU, GeLU, ELU, SeLU] | ELU |
| F.C. Layers Dropout | *Uniform*[0, 0.50) | 0.0166 |

https://doi.org/10.1371/journal.pcbi.1011563.t001

we used Bayesian optimization in combination with the Async Successive Halving Algorithm (ASHA) scheduler [26]. We performed 250 trials and selected the configuration yielding the highest correlation on the withheld genes data for *N. crassa*.

Our hyperparameter search was performed using this dataset because it was our weakest performing result. We split the gene-condition data into train, validation and test sets by randomly withholding 10% of the elements for the validation set and 20% for the test set. The optimal hyperparameters found through this search were then used for all the other species and train/test splits without further hyperparameter tuning due to computational limitations. However, we found that the hyperparameters found by tuning on *N. crassa* worked well for the other two species as well, alluding to the applicability of FUN-PROSE across different fungal species.

After tuning, our final model's promoter module was composed of two convolutional layers to process the entire 1kb sequence, where the first layer has 256 filters of 9-bp length to capture relevant sequence motifs, followed by the second layer with 64 filters of length 13 to capture more complex sequence patterns. It is interesting to note that as shown in Table 1 and S4 Fig, the optimal kernel size for the first convolutional filter was on the smaller side of the search space, while the pool kernel size for the first layer was on the larger side. This indicates that the first layer of the neural network looks for multiple short motifs and connects them over a longer range. At the same time, S4 Fig and Table 1 show that using the whole 1kb sequence appears to be the best option. This indicates that the model looks for short motifs that appear throughout the entire promoter sequence.

For the TF-processing fully-connected layer, we found a hidden size of 1024 to be optimal prior to concatenation with the convolution output. We also discovered that applying dropout to the convolutional and fully-connected layers improved our network's performance.

Although some of the optimized hyperparameter values are at the high end of the search range, we believe that these ranges should not be increased. For example, exceeding the 1000

bp promoter length is undesirable given that the average promoter sequence length in *S. cerevisiae* and *I. orientalis* is shorter than 1000 bp. In addition, we did not want to decrease the first layer's convolutional kernel size below 9 because the shorter length of discovered motifs would complicate model interpretation.

## Predicting condition-specific gene expression in fungi

For each species, we first split the gene-condition data into train, validation and test sets by randomly withholding 10% of the elements for the validation set and 20% for the test set. That is, when the neural networks are being tested, they will not be receiving the exact combinations of genes and conditions used in training and validation. With this set-up, for *S. cerevisiae*, our neural network achieved a Pearson correlation coefficient of 0.85 between predicted and observed gene expression values. The *N. crassa* and *I. orientalis* models had correlations of 0.72 and 0.81 respectively. This shows that the FUN-PROSE framework can be generalized to different fungal species with varying sizes of training datasets.

To further understand the performance of our model, we created scatter plots and confusion matrices for the predictions made on the test set (see Fig 2). We generated the confusion matrices by trinarizing expression levels on each axis into three sections labelled "Low" (below one standard deviation from the mean), "Mean" (within one standard deviation on either side of the mean) or "High" (above one standard deviation from the mean). The scatter plots and

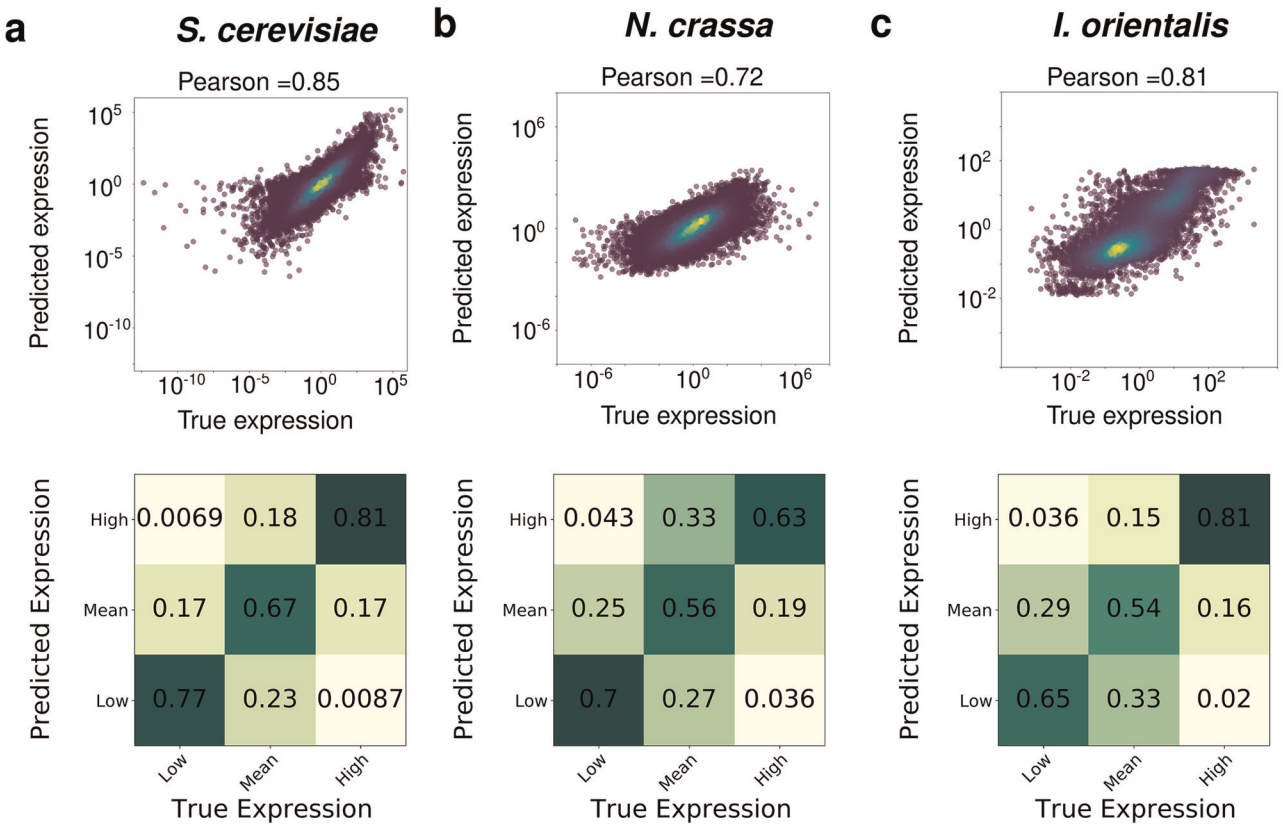



**Fig 2. FUN-PROSE model accurately predicts condition-specific gene expression for three different fungal species.** The results of FUN-PROSE predicting condition-specific gene expression of *N. crassa* **(a)**, *S. cerevisiae* **(b)**, and *I. orientalis* **(c)**. The top panel shows scatter plots of predicted (y-axis) and experimentally measured (x-axis) expression levels, with the color representing density of points. The bottom panels show confusion matrices of expression levels discretized into three categories (Low, Medium, and High) (see text for details).

https://doi.org/10.1371/journal.pcbi.1011563.g002

confusion matrices in Fig 2A–2C show that a gene-condition pair predicted to have low expression level is rarely measured to have a high expression level, and vice versa for all three species. Instead, most of the errors seem to arise from genes with either low or high expression levels being predicted to have a mean expression level. In addition, we plotted scatter plots individually for each of the top five and bottom five performing genes (see S9 Fig). These plots show that all of the 5 worst performing genes are genes that often have expression levels fall below the level of detection. Moreover, we attempted to quantify whether certain genes or conditions were predicted to be consistently lower or higher than the true values. In particular, we calculated the mean of the residuals, divided by the standard deviation of residuals for each condition and each gene. The histograms for these values are shown on S10 Fig and show that there are no significant biases for neither genes nor conditions. To also provide a baseline of what a very good model would do, we plotted a scatterplot comparing replicates of the *N. crassa* data and found the correlation between replicates to be 0.79 (shown on S11 Fig). This is comparable to our withheld elements result of 0.72, indicating that FUN-PROSE's performance is close to the limit of what is possible in predicting condition-specific gene expression.

We then designed two other more stringent splits to evaluate model performance. One with 10%/20% of conditions withheld for test and validation, i.e, our model is never shown a particular condition at all during training but tries to predict it. This setup allows one to evaluate how well our model will fare in a practical scenario when we use it to predict gene expression in a new condition which has not been experimentally tested as well as the scenario where TF expression levels have been manipulated. In this scenario, the FUN-PROSE model's performance stayed the same for *S. cerevisiae* and slightly dropped to 0.68 and 0.58 Pearson correlation between predicted and observed gene expression for *N. crassa* and *I. orientalis* respectively. We expect this drop in performance to depend on how different the new, unseen condition is from all other conditions used in the training set.

To further test the ability of FUN-PROSE to predict the expression of unseen conditions, we created an additional split to show that a model can make accurate predictions on conditions that do not resemble any of its training conditions. To do this, we first clustered the conditions, as shown in S1 and S2 Figs. Then, we built our test split ensuring that no condition in the test set is in the same cluster as any condition in the train or validation sets. As shown in S5 Fig, although this clustered split showed a slight decrease in accuracy, as expected, FUN-PROSE still kept most of its accuracy with correlation coefficients of 0.70 and 0.56 for *S. cerevisiae* and *N. crassa* respectively. We did not run this particular experiment on *I. orientalis* as there was not enough data to properly cluster the conditions. The performance of FUN-PROSE is particularly impressive when compared to that of a Nearest Neighbour Regression model which achieved correlation coefficients of 0.39 (*S. cerevisiae*) and 0.31 (*N. crassa*).

Another split is to completely withhold some gene promoters during training. This allows to assess how well FUN-PROSE will work for the task of predicting expression of a novel gene in the given condition set. This situation could be experimentally realized, e.g., if our model is used in predicting the expression of genes with synthetic or mutated promoter sequences. This setup is the least accurate with 0.36, 0.33 and 0.58 Pearson correlation between predicted and observed gene expression for *S. cerevisiae*, *N. crassa*, and *I. orientalis* respectively. *I. orientalis* may have better performance than the other two species due to a bimodal distribution of condition-specific gene expression (see S3 Fig). This bimodal distribution, in turn, may be an artifact of a very small number of conditions in our training set for this species.

To explore another potential experimental use-case of FUN-PROSE, we focused our predictions on TF-knockout data that was available in our dataset for *N.crassa*. As shown in S5 Fig, FUN-PROSE predicts the effects of TF-knockouts almost as well as it predicts the effects of novel conditions in general.

We also verified that expression levels of individual TFs are indeed used by the FUN-PROSE model to make predictions. To do this, we trained three new models for each of our fungal species, where instead of sending the expression levels of individual TFs as inputs, we send only one given by the mean expression level across all TFs. This singular value is then concatenated to the outputs of the convolutional layers before being fed to the fully connected layers to predict condition-specific gene expression. When this network was evaluated on the withheld elements datasets, we obtained correlation coefficients of 0.36, 0.34 and 0.55 for *S. cerevisiae, N. crassa* and *I. orientalis*, respectively. These are much lower than the correlations of 0.85, 0.72 and 0.81 obtained by our original FUN-PROSE model. This alleviates a potential concern that only the overall levels of TF expression for a cell are necessary to make predictions. Instead, this shows that the expression levels of individual TFs play an important role in accurately predicting condition-specific gene expression.

## Sequence motifs and their interactions can be extracted from the convolutional filters

We then set to explore the information learned by the FUN-PROSE model trained on each fungal dataset and extract sequence features that were most predictive of gene expression. To do this we used the following procedure: for all genes we extracted all feature maps from the first convolutional layer; then for each of 256 kernels we calculated statistics for base pair frequency in the 9-bp sequence windows around the top-0.5% activations. These sequence motifs quantified by base pair frequency profiles (see Methods) are analogous to Transcription Factor Binding Motifs (TFBMs) traditionally used to quantify sequence patterns recognized by individual TFs. We also calculated positional activation profiles (see Methods) for every extracted sequence motif across all promoters of a given fungal species to look for non-random positional preferences along the promoter sequence.

We hypothesized that sequence motifs extracted from CNN kernels should sometimes match TF-binding motifs. To test this hypothesis, we compared sequence motifs extracted from our model to the known *S. cerevisiae* TFBMs from the YEASTRACT database [27]. We were able to tentatively match 77, 87, and 68 of our 256 sequence motifs to at least one known TFBM in *S. cerevisiae*, *N. crassa*, and *I. orientalis* genomes respectively (see Tables A-C in S1 File and Methods for details). Fig 3A–3C shows several examples of motifs extracted from FUN-PROSE model for different species, along with their best match to a known transcription factor binding motif. In the right panel of Fig 3A–3C and 3D and S6 Fig, we show the positional activation profile of these motifs across all promoter sequences. We found that most sequence motifs extracted from our model exhibit non-random positional preferences indicative of biological function. Indeed, transcriptional regulation typically requires a TF to bind a promoter sequence not too far from the transcription start site [28, 29]. Moreover, we compared the activation profiles from Fig 3A–3C to the known binding locations of the corresponding TFs, as reported by the Yeast Epigenome Project in S14 Fig [30].

Interestingly, for all three species, we independently discovered sequence motifs similar to the one recognized by Azf1 in *S. cerevisiae*. Azf1 is known to be a transcription activator of genes involved in carbon metabolism and energy production [31] and is expected to be actively working in the set of conditions we used for our model training for all three species.

## Regulatory interactions between transcription factors and target genes can be inferred from the neural network

To understand the role of different TFs in making predictions, we generated a list of TF-target gene interactions for the model trained on the *S. cerevisiae* dataset using Integrated Gradients

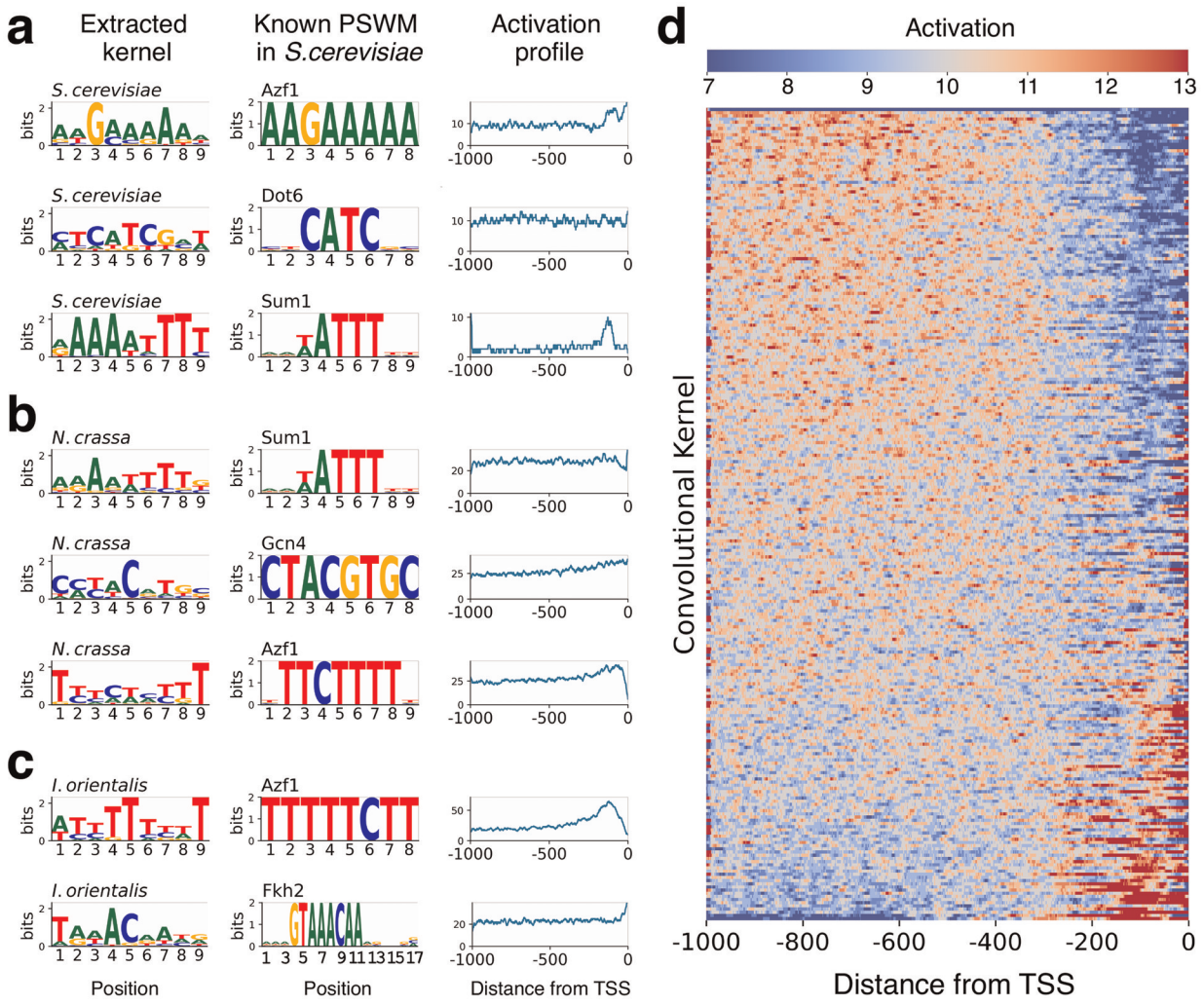
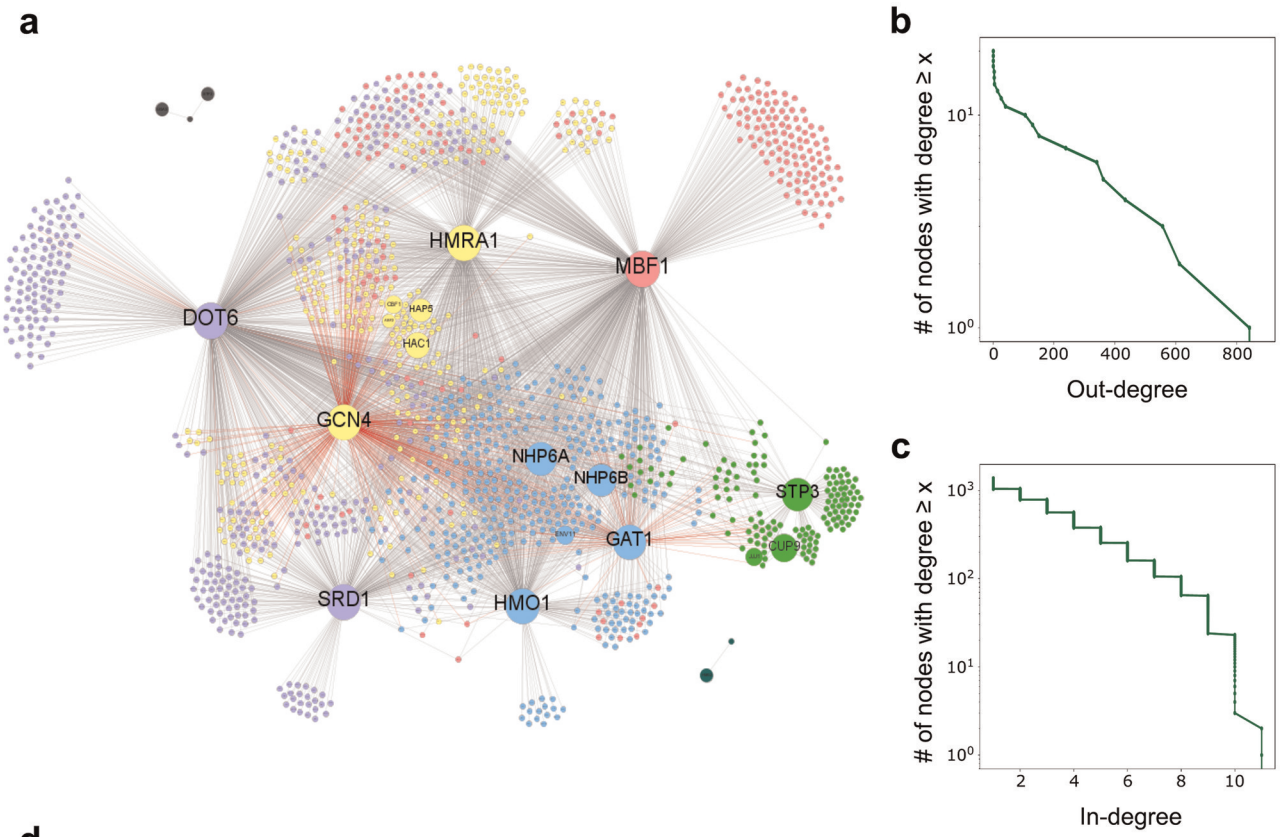



**Fig 3. Discovery of sequence motifs by the FUN-PROSE model. (a)-(c)** Examples of sequence motifs extracted from the convolutional kernels of the FUN-PROSE model trained on the respective species: *S. cerevisiae* **(a)**, *N crassa* **(b)**, *I. orientalis* **(c)**. The best matching *S. cerevisiae* motif in the YEASTRACT database is shown in the middle column and motif's positional activation profile—in the right column. **(d)** The heatmap showing the positional distribution of the top 0.5% of activations for each motif, i.e. kernel in the first convolutional layer, over all *S. cerevisiae* promoter sequences. The rows of this heatmap are sorted by the average activation level within 300bp from the transcription start site. Note that most motifs exhibit non-random positional preferences indicative of biological function.

https://doi.org/10.1371/journal.pcbi.1011563.g003

[25], which quantifies how much each of the input variables contributed to the final prediction of a given output (see Methods). This data was then binarized using the threshold of 3 standard deviations from the mean absolute TF-gene score to create a network of TF-target gene interactions made up of 20 TFs, 1343 target genes and 4144 edges as shown in Fig 4A. The size of the nodes and their labels represent the out-degree of a given TF. As seen in this network diagram, a handful hub TFs that make up the most of TF-gene interactions, a trend that is also seen in the other species (see S7 and S8 Figs). The shape of the cumulative histogram of out-degrees Fig 4B shows a sharp transition between around 10 TF hubs and the rest of TFs. In fact, these 10 hubs cover 95.7% of edges in the network. These results indicate that, when considering stress response in *S. cerevisiae*, where our training data came from, a few TFs are sufficiently predictive of condition-specific gene expression. The reasons behind the predictive power of these specific TFs is better understood by taking a closer at their biological function.




d

| TF | Out-degree | Description |
|---|---|---|
| DOT6 | 882 | Protein involved in rRna and ribosome biogenesis. Acts in response to DNA replication stress |
| MBF1 | 842 | Bridges the DNA-binding region of Gcnp and TATA-binding protein Spt15p. Acts in response to DNA replication stress |
| HMRA1 | 557 | Silenced copy of a1 at HMR; homeobox corepressor that interacts with Alpha2p to repress haploid-specific gene transcription in diploid cells |
| GCN4 | 434 | bZIP transcriptional activator of amino acid biosynthetic genes; activator responds tp amino acid starvation |

**Fig 4. *S. cerevisiae* Gene Regulatory Network learned by the FUN-PROSE model. (a)** The network of TF-target gene interactions obtained by applying a 3 standard deviation threshold to the TF-target gene Integrated Gradients scores for the *S. cerevisiae* dataset. Red edges mark experimentally confirmed interactions from the YEASTRACT database. Nodes are colored by clusters obtained by modularity optimization. **(b)** Cumulative histogram (number of nodes with degree $>= x$) of out-degrees of TFs and **(c)** in-degrees of target genes. **(d)** A table of properties of the top TFs with the highest degree-centrality, including their out-degree and the biological function according to the Saccharomyces Genome Database (SGD).

https://doi.org/10.1371/journal.pcbi.1011563.g004

The top four TFs by out-degree are DOT6, MBF1, HMRA1, and GCN4 included in Fig 4D. DOT6 has been previously shown to be a master regulator of the stress response that can encode the nature of many environmental stresses [32]. This makes sense as the variety of conditions in the dataset includes rich media, synthetic media, heat shock, hyperosmotic shock, glucose depletion, endoplasmic reticulum stress, oxidative stress, proteotoxic stress and antifungal drug exposure. Given the wide range of conditions the model was trained on, the model

would need to to depend heavily on TFs that respond to diverse stresses. Going down the list, MBF1 and GCN4 TFs were also shown to be potentially interacting master regulators in response to stress in yeast, especially nutritional stress [33, 34]. In addition, while HMRA1 is primarily known as a mating type protein, studies have linked it to both pH and oxygenation stress [35, 36].

Fig 4C shows that the in-degree of gene targets (the number of TFs regulating a given gene) approximately follows an exponential distribution, This is consistent with previous results obtained for experimentally observed regulatory interactions in *S. cerevisiae* [37].

While only a small proportion of links in Fig 4A have experimental validation (red links), it is important to note that the links we discover are not necessarily direct causal interactions. This is because machine learning methods routinely pick up indirect interactions, such as when a TF may regulate the expression of another (hidden) TF, which in-turn regulates the expression of a target gene. However, an interesting aspect about our inferred network is that the two TFs with the highest proportion of experimentally confirmed links are GCN4 and GAT1. Both of them are well known stress response regulators, with GCN4 mentioned above and GAT1 being linked to salt stress response.

Next, the network was analyzed by clustering the nodes to maximize the modularity of the network (shown as node color on Fig 4A) [38]. As expected, the clusters agree with the targets of individual hubs or tightly interconnected groups of hubs. Then, we ran Gene Ontology (GO) term enrichment analysis on genes in each cluster [39]. Through this, we saw that often, the GO terms associated with a module could be explained by the TFs in the cluster. For example, most of the GO terms associated with the Green cluster were related to RNA processing and ribosome biogenesis. The TFs in the cluster include JJJ1 which has been shown to be involved in 60S ribosomal subunit biogenesis [40]. Another TF in the Green cluster is STP3, a protein similar to STP1. While STP3 has not been widely studied, STP1 is known to be involved in tRNA splicing [41]. Another example of how the GO term of a cluster matches what is known about the TFs that regulate the cluster is the Blue cluster, which is mostly assigned GO terms relating to DNA repair. As it turns out, NHP6A/B loss leads to increased genomic instability, hypersensitivity to DNA-damaging agents [42].

## Discussion

The focus of this study is on prediction of condition-specific gene expression in fungi. This prediction task is unlike most previous attempts at predicting gene expression [20, 21] because we aim to predict the variation of expression of each gene across different conditions (Z-score), instead of predicting its absolute expression level (mRNAs/cell) averaged over all conditions. To do so, we use as inputs, promoter sequences (to encode the information about genes) and the expression levels of all TFs (to encode the information about conditions). These inputs are processed by the FUN-PROSE network, which is made up of a convolutional neural network to extract features from the promoter sequences, a fully-connected feed-forward network to learn features from the TF expression levels.

We found that it is possible to predict condition-specific gene expression with high accuracy for three different fungal species: *S. cerevisiae*, *N. crassa* and *I. orientalis*, indicating that FUN-PROSE accuracy generalizes well. We also found that FUN-PROSE model can also be applied to predict gene expression in previously unseen conditions and, separately, on previously unseen promoter sequences, although the prediction accuracy on new promoter sequences was considerably lower than that for previously unseen conditions. These results indicate that our model could have practical applications in predicting how the transcriptome of a fungal species will react to a new condition that has not been tested yet, predicting the

consequences of TF knockout/overexpression as well as predicting the effect of promoter modification.

Next, we showed that FUN-PROSE can be successfully interpreted to extract the biological information it used to make predictions. We did this in two parts: first, we interpreted the convolutional neural network module of the FUN-PROSE model to extract the sequence motifs. Second, we interpreted the fully-connected module to extract interactions between TFs and their gene targets. This was done using Integrated Gradients, allowing one to predict which inputs significantly contributed to prediction of a given output variable (gene expression). The results of both exercises were compared to the existing biological knowledge, which is especially significant for *S. cerevisiae*. The fact that a sizeable fraction of our predictions overlapped with known TFBS or regulatory interactions convinced us that the FUN-PROSE model uses biologically relevant information to make its predictions. This also alludes to how we might be able to use FUN-PROSE to generate novel biological hypotheses for less studied species.

In addition, we attempted the ambitious task of trying to connect the TF-gene network shown in Fig 4A to the motifs extracted from convolutional filters in the previous section. For each combination of TF and convolutional filter (only TFs with $\geq 20$ targets in the network were considered), we ran a hypergeometric test to evaluate if the filter activation was over-represented among the TF's targets. Filter activation was binarized using a threshold of 1 standard deviation from the mean. The results of this analysis are shown in S12 Fig. Over half of the tested TFs were matched to a convolutional filter. The resulting network of TF-gene links supported by such motif presence is shown in S13 Fig. This analysis can also be used to narrow down TF-gene links of particularly well connected hubs.

In order to understand if additional information regarding cis-regulatory sequences of genes would help our model, we trained a variant of our model with an additional input of 1000bp from 3' UTR located downstream from the protein coding region of the gene. Similar to the promoter sequence, this information was processed by a two-layer CNN. The latent representations of the 3' UTR is then concatenated with that of the promoter and the TF expression levels before being fed to the fully connected layers. With this additional piece of input, for the withheld elements split of the data, the model was able to achieve correlation coefficients of 0.87, 0.77 and 0.83 in *S. cerevisiae*, *N. crassa* and *I. orientalis*, respectively. The performance of this model is somewhat higher than that of the original FUN-PROSE model: 0.85, 0.72 and 0.81. We expected that adding 3' UTR region controlling mRNA degradation would lead to an improvement in accuracy of our predictions. The fact that the magnitude of this improvement was relatively small could be tentatively attributed to the fact that we did not explicitly include condition-specific expression levels of proteins controlling mRNA degradation. So, the model had to use indirect relationships due to co-expression of these proteins with transcription factors. In other words, TF expression levels impact the expression levels of RNA binding proteins, which in turn affect the mRNA decay. Future work could extend our model by identifying families of proteins responsible for mRNA degradation and including their expression levels as inputs alongside TFs.

Many sequence motifs we discovered for different species are similar to each other. In order to quantify the overall level of conservation of these motifs, we performed the following experiment: we took the weights of the convolutional layers from the *S.cerevisiae* model, froze them and retrained the rest of the model for *I. orientalis*. The final model performance for withheld genes test/train split was Pearson correlation equal to 0.45, which is still significant but somewhat smaller than Pearson correlation equal to 0.58 in the FUN-PROSE model in which CNN weights were independently trained.

We focused our attention on making predictions for fungi for two reasons: (i) the mechanisms for gene regulation in fungi are simpler than that of higher eukaryotes and (ii) being

able to make these predictions is an important part of the metabolic engineering of fungi to optimize their use in the production of desired chemicals in industrially scalable conditions. While we tested our framework on fungi, we believe that the success of FUN-PROSE across different species in the fungal kingdom suggests that it might be generalizable to other kingdoms, e.g. plants and animals. In doing so, one of the potential modifications of the model is changing the length of the promoter sequence that the model takes. The expression of some genes in metazoan species is known to be regulated by distal transcriptional enhancers [43]. Directly incorporating these distal enhancers in our modeling framework may not be computationally tractable. One possible solution to this problem is to extend our model by taking into account the 3D chromosome structure connecting enhancers to their gene targets. One can also incorporate into our model additional input features that might be especially important for more complex organisms. One example of this is the information about epigenetic modifications in the neighborhood of a given gene. We are currently working on this problem for human tissues, where we include more regulatory mechanisms in our model.

Another possible direction for a future study is to predict condition-specific gene expression using single cell transcriptomics data. Machine learning models may potentially perform better in this setting than for spatially averaged expression data due to the lack of averaging over distinct subpopulations.

## Materials and methods

### Data sources and preprocessing

We used previously published RNA-seq data on *N. crassa* (wild type and gene-deletion mutants) growing on different carbon sources [44], recent *S. cerevisiae* RNA-seq data for 28 analog sensitive kinase alleles across 12 different conditions (stresses and different media) (GEO: GSE115556) [45] and *I. orientalis* RNA-seq data for growth in different media conditions (YPD+glucose and lignocellulosic extracts). The goal of the *N. crassa dataset* was to gain to a more complete understanding of cross talk between transcription factors and their target genes, which are involved in regulating nutrient sensing [44]. On the other hand, the goal of the *S. cerevisiae* dataset was to explore the role of various protein kinases in the stress response. Finally, the *I. orientalis* data can be used to understand the effects of different types of media on gene expression.

*Processing the expression data.* The data was processed and raw/FPKM counts were obtained by the authors of the respective studies. To standardize the data, for each media condition we first renormalized raw counts to FPKM (if not already) and averaged data for multiple replicates of the same condition. We then filter out genes that have mean expression below 0.05 and genes that have coefficient of variation below 0.3. S1–S3 Figs (top-left) shows the distribution of counts after applying the two filters. Finally, we log-transformed counts and performed z-score normalization for each gene. The result of these transformations on the distribution of counts is shown in S1–S3 Figs (top-right).

*Processing the sequence data.* For *N. crassa* we used reference genome Neurospora crassa OR74A v2.0 [46] obtained from MycoCosm; for *S. cerevisiae* we used S. cerevisiae S288C R64–3-1 [47] obtained from SGD; and for *I. orientalis* we used Pichia kudriavzevii CBS573 [48] obtained from MycoCosm. For each gene we defined promoter sequence as 1kb upstream of the start codon (such that the 5'UTR region is included in promoter) and extracted them from the corresponding reference genomes. We chose 1kb as a number typically used as fungal promoter length in the literature [29, 49, 50]. We filtered out genes that had promoters shorter than 1kb as it sometimes happens at the ends of chromosome. In the end, we worked with

9725, 6645, and 4925 genes for which we had all 3 types of data for *N. crassa*, *S. cerevisiae*, and *I. orientalis* respectively.

*Predicting transcription factors.* We used InterProScan v5.52–86.0 to annotate reference genomes. To obtain all putative TFs for each species, we used this annotation to extract genes that correspond to the list of TF-specific pfams from DNA-binding domain database (DBD) [51] v2.03 and TF-specific Interpro terms from Fungal Transcription Factor Database (FTFD) [52] v1.2. Overall we identified 325, 208, and 183 putative TFs in *N. crassa*, *S. cerevisiae*, and *I. orientalis* genomes respectively.

## The model

As shown in Fig 1, our final neural network is composed of three modules: a convolutional neural network that takes the promoter sequence as input, a feed-forward layer that takes the transcription factor expression levels as input, and a multi-layer feed-forward neural network that takes the concatenation of the outputs of the previous two modules as input.

The convolutional neural network module is composed of two convolutional layers. The first convolutional layer has 256 kernels of length 9 and stride of one. This is followed by max pooling with kernel size and stride of 19, a ReLU activation function, dropout (elements of the hidden layer are randomly set to 0 with probability, $p$), and batch normalization (the hidden layer is normalized to follow a standard normal). The second convolutional layer, on the other hand, has 64 kernels. Each of these kernels has a length of 13 and a stride of one. This layer is followed by max pooling with kernel size and stride of 8, a ReLU activation function, dropout, and batch normalization.

The feed-forward layer that processes the TF expression levels is a fully-connected layer with a ELU activation function.

Finally, the outputs of the convolutional module and the feed-forward module are concatenated and sent through a final module with three fully-connected layers with ELU activation, dropout, batch normalization, and then a final fully-connected layer to predict the expression of a particular gene.

## Hyperparameter optimization and model training

During training, we defined the loss function as the mean squared error of the gene expression level predictions. The AdamW optimizer was used to minimize this loss. To select the optimal set of hyperparameters, we looked at the configuration yielding the highest validation Pearson correlation.

During hyperparameter optimization, we allowed a maximum of 20 epochs for each trial, with a minimum of 5 epochs before stopping. Trials were also stopped early if they reached a plateau, as defined by the standard deviation of the validation correlation coefficient not exceeding 0.01 in the final 5 epochs. The ASHA scheduler was configured with a reduction factor of 3 and 1 bracket. We used the following software for hyperparameter optimization: Ray v1.8.0, PyTorch v1.9.1, and CUDA v11.5 [53–55]. Models were trained on an NVIDIA V100 GPU with 16GB of RAM using automatic mixed-precision training.

In model training, we allowed for a maximum of 60 epochs and training was stopped early if the validation correlation coefficient did not improve for 5 epochs in a row. We ran our model training on a NVIDIA GeForce GTX 1080 Ti GPU.

## Interpreting the convolutional kernels

*Inferring promoter motifs from convolutional kernels.* We inferred the promoter motifs learned by each model by examination of the 256 kernels in the first convolutional layer, which have

been shown to capture such information [23]. For each kernel, denoted by $Conv1d_x$ for $x \in [0...255]$, we processed all unique one-hot-encoded 1000-bp promoter sequences ($P_1, \ldots, P_N$, each of shape $5 \times 1000$) to generate a feature map $F_x$ of shape $N \times 1000$. Notation: the $x$th feature map $F_x$ is indexed as $F_x^{i,k}$, where $i$ identifies the promoter and $k$ identifies the sequence position. Each promoter $P_i$ is indexed as $P_i^{j,k}$ where $j$ identifies the nucleotide base (A, C, G, T, and N to represent an unknown base) and $k$ identifies the sequence position. All indexes are zero-based unless otherwise noted.

$$F_x^{i,*} = ReLU(Conv1d_x(P_i))$$

We then constructed a motif representation $T_x$ with shape $5 \times 9$ as a weighted aggregate of the 9-bp sequence windows corresponding to the top 0.5% activations in $F_x$ (denoted by the paired promoter and sequence position index lists $i_x^\dagger, k_x^\dagger$). Notation: $T_x$ is indexed as $T_x^{j,k}$ where $j$ identifies the nucleotide base and $k$ identifies the window position. $M_x$ is indexed similarly.

$$T_x^{j,k} = \Sigma_{(\hat{i},\hat{k}) \in (i_x^\dagger, k_x^\dagger)} F_x^{\hat{i},\hat{k}+k} P_{\hat{i}}^{j,\hat{k}+k}$$

$$M_x^{j,k} = T_x^{j,k} / \sum_{\bar{j} \in [0...4]} (T_x^{\bar{j},k})$$

The final result $M_x$ is a position-specific weight matrix (PSWM) representing the sequence motif learned by the $x$th kernel.

*Positional activation profiles for convolutional kernels.* We also constructed positional activation profiles for each kernel based on the locations of the top 0.5% non-zero activations in the promoter sequences. Values were smoothed using a moving average over the sequence with a window size of 15 bps.

*Comparing against previously reported motifs.* We compared these motifs against reported TF-binding *S. cerevisiae* motifs in the YEASTRACT database (version 20130918) [27] using *Tomtom* [56]. For motif comparison, we used the Pearson Correlation Coefficient function and complete scoring with a statistical significance threshold of E-value $<0.5$ (corresponding to a p-value of 0.0007).

**Interpreting TF-gene relationships.** *Generating a table of TF-gene interactions.* To investigate the importance of TFs for prediction, we used Integrated Gradient (IG) scores [25]. IG scores are obtained by creating a linear interpolation between a baseline input and an actual input and calculating gradients at small steps on this interpolation to determine which features have a strong impact on the model's prediction. For each prediction made in the withheld elements test set, we calculated IG scores using Captum v0.5.0 with zero-tensors as the baselines and 20 approximation steps. We only used the IG scores for TF expression levels and averaged them for each target gene. This gives us a table of TF-gene interactions in the model. To find the overall importance of each TF, we then averaged the values of this table for each TF across all target genes.

*Creating and analyzing a network of TF-target gene interactions.* The table of TF-target gene relationships was then thresholded (3 standard deviations from the mean for *S. cerevisiae*) to obtain a network of TF-gene relationships. This network was vizualized using Gephi [57].

To perform our GO enrichment analysis, we first clustered our network by maximizing modularity, a measure of the density of links inside communities as compared to links between communities [38]. Once again, we did this using a built in function in Gephi [57]. Using the clusters obtained, we then performed Gene Ontology (GO) term enrichment analysis to identify terms that are significantly overrepresented in each cluster. This was conducted with the

BiNGO tool using a hypergeometric statistical test and a Benjamini-Hochberg FDR corrected significance level of $\alpha = 0.05$ [39].

## Supporting information

**S1 Fig. Vizualizing the *Saccharomyces cerevisiae* dataset. (Top-left)** Histogram of gene expression with CV filter. **(Top-right)** Histogram of Z-scored expressions. **(Bottom)** The different conditions in the dataset clustered by gene expression using agglomerative clustering.
(PDF)

**S2 Fig. Vizualizing the *Neurospora crassa* dataset. (Top-left)** Histogram of gene expression with CV filter. **(Top-right)** Histogram of Z-scored expressions. **(Bottom)** The different conditions in the dataset clustered by gene expression using agglomerative clustering.
(PDF)

**S3 Fig. Vizualizing the *Issatchenkia orientalis* dataset. (Top-left)** Histogram of gene expression with CV filter. **(Top-right)** Histogram of Z-scored expressions. **(Bottom)** The different conditions in the dataset clustered by gene expression using agglomerative clustering.
(PDF)

**S4 Fig. The factor-response plot for our hyperparameter optimization runs.** The y-axis of the plot shows the correlation between predicted and measured expression for the validation set. The trials on the plot are sorted in ascending correlations. On the x-axis, we show the various hyperparameters that we are optimizing. The size of the black markers represent the value of the hyperparameter for that trial. For example, this plot can be interpreted to show that a smaller learning rate leads to better performance.
(PDF)

**S5 Fig. A deeper investigation into the performance of FUN-PROSE on withheld conditions.** The performance of FUN-PROSE on the withheld conditions split (in blue) is compared to a more rigorous split (clustered split; in orange). To generate the clustered split, the conditions are clustered using hierachical clustering, as shown on S1 and S2 Figs. Then, the train and test splits are generated such that no condition in found in the train split is found in the test split, and vice-versa. The performance of FUN-PROSE is then compared to a simpler Nearest Neighbor Regression model (in green). Finally, FUN-PROSE is used to predict the effects of TF knockout (shown in red).
(PDF)

**S6 Fig. The heatmaps showing the positional distribution of the top 0.5% of activations for each motif, i.e. kernel in the first convolutional layer, over all (left)*N. crassa* and (right) *I. orientalis* promoter sequences.** The rows of this heatmap are sorted by the average activation level within 300bp from the transcription start site. Note that most motifs exhibit non-random positional preferences indicative of biological function.
(PNG)

**S7 Fig. *N. crassa* Gene Regulatory Network learned by the FUN-PROSE model. (a)** The network of TF-target gene interactions obtained by applying a 2.5 standard deviation threshold to the TF-target gene Integrated Gradients scores for the *N. crassa* dataset. Nodes are colored by clusters obtained by modularity analysis. Node sizes are proportional to their out-degree. **(b)** Cumulative histogram (number of nodes with degree $> = x$) of out-degrees of TFs and **(c)** in-degrees of target genes.
(PDF)

**S8 Fig. *I. orientalis* Gene Regulatory Network learned by the FUN-PROSE model. (a)** The network of TF-target gene interactions obtained by applying a 2 standard deviation threshold to the TF-target gene Integrated Gradients scores for the *I. orientalis* dataset. Nodes are colored by clusters obtained by modularity analysis. Node sizes are proportional to their out-degree. **(b)** Cumulative histogram (number of nodes with degree $> = x$) of out-degrees of TFs and **(c)** in-degrees of target genes.
(PDF)

**S9 Fig. Scatter plots of predicted (y-axis) and experimentally measured (x-axis) expression levels for each of the top 5 performing genes (top row) and the bottom 5 performing genes (bottom row) based on correlation between measured expression and the expression level predicted by FUN-PROSE for *S. cerevisiae*.**
(PDF)

**S10 Fig. The distribution of residuals.** The distribution of the mean residuals divided by standard deviation of residuals for each gene **(left)** and for each condition **(right)** in the *S. cerevisiae* test data.
(PNG)

**S11 Fig. Comparing replicates in the experimental data for *N. crassa*.** This is done using a scatter plot of replicates 1 and 2 of the data which have a Pearson correlation coefficient of 0.79.
(PNG)

**S12 Fig. Matching of *S. cerevisiae* TFs to filters in FUN-PROSE with a q-value cutoff of 0.2.** The matching was made by first taking the network shown in Fig 4A. Then, for each combination of TF (only TFs with $\geq$ 20 targets were considered) and filter, we ran a hypergeometric test to evaluate if the filter activation was over-represented in the TF's targets. Filter activation was binarized using a threshold of 1 standard deviation from the mean.
(PDF)

**S13 Fig. The *S. cerevisiae* Gene Regulatory Network with only edges supported by convolutional filters matched to each TF.** The network from Fig 4A was filtered so that only edges So that each TF-gene link was only kept if the gene was activated by a convolutional filter associated with the TF as listed in S12 Fig.
(PNG)

**S14 Fig. Comparing the some of the activation profiles to known binding locations of corresponding TFs in *S. cerevisiae*.** The activation profiles of motifs from Fig 4 are compared to binding locations presented by the Yeast Epigenome Project.
(PDF)

**S1 File. Supplementary data for model interpretation results.** Tables A-C are the full list of TOMTOM matches for the motifs discovered by FUN-PROSE for each of the three species. Supp. Table D Gives the edge list that makes up the network in Fig 4a. Supp. Tables E-H give the GO term enrichment results for the different modules detected in Fig 4a.
(XLSX)

## Acknowledgments

We thank Peter Koo, Saurabh Sinha and Angelo Miskalis for insightful discussions.

## Author Contributions

**Conceptualization:** Ananthan Nambiar, Veronika Dubinkina, Sergei Maslov.

**Data curation:** Veronika Dubinkina.

**Funding acquisition:** Sergei Maslov.

**Investigation:** Ananthan Nambiar, Veronika Dubinkina, Simon Liu, Sergei Maslov.

**Methodology:** Ananthan Nambiar, Veronika Dubinkina, Simon Liu, Sergei Maslov.

**Project administration:** Sergei Maslov.

**Software:** Ananthan Nambiar, Simon Liu.

**Supervision:** Sergei Maslov.

**Visualization:** Ananthan Nambiar, Veronika Dubinkina, Simon Liu.

**Writing – original draft:** Ananthan Nambiar, Veronika Dubinkina, Simon Liu, Sergei Maslov.

**Writing – review & editing:** Ananthan Nambiar, Veronika Dubinkina, Simon Liu, Sergei Maslov.

## References

1. Gygi SP, Rochon Y, Franza BR, Aebersold R. Correlation between protein and mRNA abundance in yeast. Molecular and cellular biology. 1999; 19(3):1720–1730. https://doi.org/10.1128/mcb.19.3.1720 PMID: 10022859
2. Greenbaum D, Colangelo C, Williams K, Gerstein M. Comparing protein abundance and mRNA expression levels on a genomic scale. Genome biology. 2003; 4(9):1–8. https://doi.org/10.1186/gb-2003-4-9-117 PMID: 12952525
3. Griffin TJ, Gygi SP, Ideker T, Rist B, Eng J, Hood L, et al. Complementary Profiling of Gene Expression at the Transcriptome and Proteome Levels in Saccharomyces cerevisiae* S. Molecular & Cellular Proteomics. 2002; 1(4):323–333. https://doi.org/10.1074/mcp.M200001-MCP200
4. Liu Y, Beyer A, Aebersold R. On the dependency of cellular protein levels on mRNA abundance. Cell. 2016; 165(3):535–550. https://doi.org/10.1016/j.cell.2016.03.014 PMID: 27104977
5. Cameron DE, Bashor CJ, Collins JJ. A brief history of synthetic biology. Nature Reviews Microbiology. 2014; 12(5):381–390. https://doi.org/10.1038/nrmicro3239 PMID: 24686414
6. Michael DG, Maier EJ, Brown H, Gish SR, Fiore C, Brown RH, et al. Model-based transcriptome engineering promotes a fermentative transcriptional state in yeast. Proceedings of the National Academy of Sciences. 2016; 113(47):E7428–E7437. https://doi.org/10.1073/pnas.1603577113 PMID: 27810962
7. Nielsen J, Keasling JD. Engineering cellular metabolism. Cell. 2016; 164(6):1185–1197. https://doi.org/10.1016/j.cell.2016.02.004 PMID: 26967285
8. Kemmeren P, Sameith K, Van De Pasch LA, Benschop JJ, Lenstra TL, Margaritis T, et al. Large-scale genetic perturbations reveal regulatory networks and an abundance of gene-specific repressors. Cell. 2014; 157(3):740–752. https://doi.org/10.1016/j.cell.2014.02.054 PMID: 24766815
9. Hackett SR, Baltz EA, Coram M, Wranik BJ, Kim G, Baker A, et al. Learning causal networks using inducible transcription factors and transcriptome-wide time series. Molecular systems biology. 2020; 16(3):e9174. https://doi.org/10.15252/msb.20199174 PMID: 32181581
10. Kang Y, Patel NR, Shively C, Recio PS, Chen X, Wranik BJ, et al. Dual threshold optimization and network inference reveal convergent evidence from TF binding locations and TF perturbation responses. Genome research. 2020; 30(3):459–471. https://doi.org/10.1101/gr.259655.119 PMID: 32060051
11. Foat BC, Houshmandi SS, Olivas WM, Bussemaker HJ. Profiling condition-specific, genome-wide regulation of mRNA stability in yeast. Proceedings of the National Academy of Sciences. 2005; 102(49):17675–17680. https://doi.org/10.1073/pnas.0503803102 PMID: 16317069
12. Inukai S, Kock KH, Bulyk ML. Transcription factor–DNA binding: beyond binding site motifs. Current opinion in genetics & development. 2017; 43:110–119. https://doi.org/10.1016/j.gde.2017.02.007 PMID: 28359978

13. Wodicka L, Dong H, Mittmann M, Ho MH, Lockhart DJ. Genome-wide expression monitoring in Saccharomyces cerevisiae. Nature biotechnology. 1997; 15(13):1359–1367. https://doi.org/10.1038/nbt1297-1359 PMID: 9415887
14. Cheng J, Maier KC, Avsec Ž, Rus P, Gagneur J. Cis-regulatory elements explain most of the mRNA stability variation across genes in yeast. Rna. 2017; 23(11):1648–1659. https://doi.org/10.1261/rna.062224.117 PMID: 28802259
15. Yang J, Ma A, Hoppe AD, Wang C, Li Y, Zhang C, et al. Prediction of regulatory motifs from human Chip-sequencing data using a deep learning framework. Nucleic acids research. 2019; 47(15):7809–7824. https://doi.org/10.1093/nar/gkz672 PMID: 31372637
16. Quang D, Xie X. FactorNet: a deep learning framework for predicting cell type specific transcription factor binding from nucleotide-resolution sequential data. Methods. 2019; 166:40–47. https://doi.org/10.1016/j.ymeth.2019.03.020 PMID: 30922998
17. Avsec Ž, Weilert M, Shrikumar A, Krueger S, Alexandari A, Dalal K, et al. Base-resolution models of transcription-factor binding reveal soft motif syntax. Nature Genetics. 2021; 53(3):354–366. https://doi.org/10.1038/s41588-021-00782-6 PMID: 33603233
18. Vaishnav ED, de Boer CG, Molinet J, Yassour M, Fan L, Adiconis X, et al. The evolution, evolvability and engineering of gene regulatory DNA. Nature. 2022; 603(7901):455–463. https://doi.org/10.1038/s41586-022-04506-6 PMID: 35264797
19. Culley C, Vijayakumar S, Zampieri G, Angione C. A mechanism-aware and multiomic machine-learning pipeline characterizes yeast cell growth. Proceedings of the National Academy of Sciences. 2020; 117(31):18869–18879. https://doi.org/10.1073/pnas.2002959117 PMID: 32675233
20. Agarwal V, Shendure J. Predicting mRNA abundance directly from genomic sequence using deep convolutional neural networks. Cell reports. 2020; 31(7):107663. https://doi.org/10.1016/j.celrep.2020.107663 PMID: 32433972
21. Zrimec J, Börlin CS, Buric F, Muhammad AS, Chen R, Siewers V, et al. Deep learning suggests that gene expression is encoded in all parts of a co-evolving interacting gene regulatory structure. Nature communications. 2020; 11(1):1–16. https://doi.org/10.1038/s41467-020-19921-4 PMID: 33262328
22. Song Q, Lee J, Akter S, Rogers M, Grene R, Li S. Prediction of condition-specific regulatory genes using machine learning. Nucleic Acids Research. 2020; 48:e62–e62. https://doi.org/10.1093/nar/gkaa264 PMID: 32329779
23. Koo PK, Ploenzke M. Improving representations of genomic sequence motifs in convolutional networks with exponential activations. Nature Machine Intelligence. 2021; 3(3):258–266. https://doi.org/10.1038/s42256-020-00291-x PMID: 34322657
24. Liu B, Hussami N, Shrikumar A, Shimko T, Bhate S, Longwell S, et al. A multi-modal neural network for learning cis and trans regulation of stress response in yeast. arXiv preprint arXiv:190809426. 2019;.
25. Sundararajan M, Taly A, Yan Q. Axiomatic Attribution for Deep Networks. In: Proceedings of the 34th International Conference on Machine Learning—Volume 70. ICML'17. JMLR.org; 2017. p. 3319–3328.
26. Li L, Jamieson K, Rostamizadeh A, Gonina E, Ben-tzur J, Hardt M, et al. A System for Massively Parallel Hyperparameter Tuning. In: Dhillon I, Papailiopoulos D, Sze V, editors. Proceedings of Machine Learning and Systems. vol. 2; 2020. p. 230–246. Available from: https://proceedings.mlsys.org/paper/2020/file/f4b9ec30ad9f68f89b29639786cb62ef-Paper.pdf.
27. Teixeira MC, Monteiro PT, Guerreiro JF, Gonçalves JP, Mira NP, dos Santos SC, et al. The YEASTRACT database: an upgraded information system for the analysis of gene and genomic transcription regulation in Saccharomyces cerevisiae. Nucleic Acids Research. 2013; 42(D1):D161–D166. https://doi.org/10.1093/nar/gkt1015 PMID: 24170807
28. Erb I, Van Nimwegen E. Transcription factor binding site positioning in yeast: proximal promoter motifs characterize TATA-less promoters. PLoS One. 2011; 6(9):e24279. https://doi.org/10.1371/journal.pone.0024279 PMID: 21931670
29. Rossi MJ, Kuntala PK, Lai WK, Yamada N, Badjatia N, Mittal C, et al. A high-resolution protein architecture of the budding yeast genome. Nature. 2021; 592(7853):309–314. https://doi.org/10.1038/s41586-021-03314-8 PMID: 33692541
30. Rossi MJ, Kuntala PK, Lai WK, Yamada N, Badjatia N, Mittal C, et al. A high-resolution protein architecture of the budding yeast genome. Nature. 2021; 592(7853):309–314. https://doi.org/10.1038/s41586-021-03314-8 PMID: 33692541
31. Slattery MG, Liko D, Heideman W. The function and properties of the Azf1 transcriptional regulator change with growth conditions in Saccharomyces cerevisiae. Eukaryotic cell. 2006; 5(2):313–320. https://doi.org/10.1128/EC.5.2.313-320.2006 PMID: 16467472

32. Granados AA, Pietsch JMJ, Cepeda-Humerez SA, Farquhar IL, Tkačik G, Swain PS. Distributed and dynamic intracellular organization of extracellular information. Proceedings of the National Academy of Sciences. 2018; 115(23):6088–6093. https://doi.org/10.1073/pnas.1716659115 PMID: 29784812

33. Amorim-Vaz S, Coste AT, Tran VDT, Pagni M, Sanglard D. Function Analysis of MBF1, a Factor Involved in the Response to Amino Acid Starvation and Virulence in Candida albicans. Frontiers in Fungal Biology. 2021; 2. https://doi.org/10.3389/ffunb.2021.658899 PMID: 37744106

34. Hinnebusch AG, Natarajan K. Gcn4p, a Master Regulator of Gene Expression, Is Controlled at Multiple Levels by Diverse Signals of Starvation and Stress. Eukaryotic Cell. 2002; 1(1):22–32. https://doi.org/10.1128/EC.01.1.22-32.2002 PMID: 12455968

35. Serrano R, Ruiz A, Bernal D, Chambers JR, Ariño J. The transcriptional response to alkaline pH in Saccharomyces cerevisiae: evidence for calcium-mediated signalling. Molecular microbiology. 2002; 46 (5):1319–1333. https://doi.org/10.1046/j.1365-2958.2002.03246.x PMID: 12453218

36. Baumann K, Dato L, Graf AB, Frascotti G, Dragosits M, Porro D, et al. The impact of oxygen on the transcriptome of recombinant S. cerevisiae and P. pastoris-a comparative analysis. BMC genomics. 2011; 12:1–16.

37. Maslov S, Sneppen K. Computational architecture of the yeast regulatory network. Physical biology. 2005; 2(4):S94. https://doi.org/10.1088/1478-3975/2/4/S03 PMID: 16280626

38. Blondel VD, Guillaume JL, Lambiotte R, Lefebvre E. Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment. 2008; 2008(10):P10008. https://doi.org/10.1088/1742-5468/2008/10/P10008

39. Maere S, Heymans K, Kuiper M. BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. Bioinformatics. 2005; 21(16):3448–3449. https://doi.org/10.1093/bioinformatics/bti551 PMID: 15972284

40. Meyer AE, Hung NJ, Yang P, Johnson AW, Craig EA. The specialized cytosolic J-protein, Jjj1, functions in 60S ribosomal subunit biogenesis. Proceedings of the National Academy of Sciences. 2007; 104 (5):1558–1563. https://doi.org/10.1073/pnas.0610704104 PMID: 17242366

41. Jørgensen MU, Gjermansen C, Andersen HA, Kielland-Brandt MC. STP1, a gene involved in pre-tRNA processing in yeast, is important for amino-acid uptake and transcription of the permease gene BAP2. Current Genetics. 1997; 31(3):241–247. https://doi.org/10.1007/s002940050201 PMID: 9065387

42. Giavara S, Kosmidou E, Hande MP, Bianchi ME, Morgan A, d'Adda di Fagagna F, et al. Yeast Nhp6A/B and Mammalian Hmgb1 Facilitate the Maintenance of Genome Stability. Current Biology. 2005; 15 (1):68–72. https://doi.org/10.1016/j.cub.2004.12.065 PMID: 15649368

43. Bulger M, Groudine M. Functional and mechanistic diversity of distal transcription enhancers. Cell. 2011; 144(3):327–339. https://doi.org/10.1016/j.cell.2011.01.024 PMID: 21295696

44. Wu VW, Thieme N, Huberman LB, Dietschmann A, Kowbel DJ, Lee J, et al. The regulatory and transcriptional landscape associated with carbon utilization in a filamentous fungus. Proceedings of the National Academy of Sciences. 2020; 117(11):6003–6013. https://doi.org/10.1073/pnas.1915611117

45. Mace K, Krakowiak J, El-Samad H, Pincus D. Multi-kinase control of environmental stress responsive transcription. PloS one. 2020; 15(3):e0230246. https://doi.org/10.1371/journal.pone.0230246 PMID: 32160258

46. Galagan JE, Calvo SE, Borkovich KA, Selker EU, Read ND, Jaffe D, et al. The genome sequence of the filamentous fungus Neurospora crassa. Nature. 2003; 422(6934):859–868. https://doi.org/10.1038/nature01554 PMID: 12712197

47. Engel SR, Dietrich FS, Fisk DG, Binkley G, Balakrishnan R, Costanzo MC, et al. The Reference Genome Sequence of Saccharomyces cerevisiae: Then and Now. G3 Genes—Genomes—Genetics. 2014; 4(3):389–398. https://doi.org/10.1534/g3.113.008995 PMID: 24374639

48. Douglass AP, Offei B, Braun-Galleani S, Coughlan AY, Martos AA, Ortiz-Merino RA, et al. Population genomics shows no distinction between pathogenic Candida krusei and environmental Pichia kudriavzevii: one species, four names. PLoS pathogens. 2018; 14(7):e1007138. https://doi.org/10.1371/journal.ppat.1007138 PMID: 30024981

49. Habib N, Wapinski I, Margalit H, Regev A, Friedman N. A functional selection model explains evolutionary robustness despite plasticity in regulatory networks. Molecular systems biology. 2012; 8(1):619. https://doi.org/10.1038/msb.2012.50 PMID: 23089682

50. Bergenholm D, Liu G, Holland P, Nielsen J. Reconstruction of a global transcriptional regulatory network for control of lipid metabolism in yeast by using chromatin immunoprecipitation with lambda exonuclease digestion. Msystems. 2018; 3(4):e00215–17. https://doi.org/10.1128/mSystems.00215-17 PMID: 30073202

51. Wilson D, Charoensawan V, Kummerfeld SK, Teichmann SA. DBD—-taxonomically broad transcription factor predictions: new content and functionality. Nucleic acids research. 2008; 36(suppl_1):D88–D92. https://doi.org/10.1093/nar/gkm964 PMID: 18073188

52. Park J, Park J, Jang S, Kim S, Kong S, Choi J, et al. FTFD: an informatics pipeline supporting phylogenomic analysis of fungal transcription factors. Bioinformatics. 2008; 24(7):1024–1025. https://doi.org/10.1093/bioinformatics/btn058 PMID: 18304934

53. Paszke A, Gross S, Chintala S, Chanan G, Yang E, DeVito Z, et al. Automatic differentiation in PyTorch. 2017;.

54. Liaw R, Liang E, Nishihara R, Moritz P, Gonzalez JE, Stoica I. Tune: A research platform for distributed model selection and training. arXiv preprint arXiv:180705118. 2018;.

55. NVIDIA, Vingelmann P, Fitzek FHP. CUDA, release: 11.5; 2020. Available from: https://developer.nvidia.com/cuda-toolkit.

56. Gupta S, Stamatoyannopoulos JA, Bailey TL, Noble WS. Quantifying similarity between motifs. Genome Biology. 2007; 8(2):R24. https://doi.org/10.1186/gb-2007-8-2-r24 PMID: 17324271

57. Bastian M, Heymann S, Jacomy M. Gephi: An Open Source Software for Exploring and Manipulating Networks; 2009. Available from: http://www.aaai.org/ocs/index.php/ICWSM/09/paper/view/154.