scYeast: a Biological-knowledge-guided Foundation Model on Yeast Single-Cell Transcriptomics

Xingcun Fan[1,2,3], Wenbin Liao[1,2,3], Luchi Xiao[1,3], Xuefeng Yan[2*], Hongzhong Lu[1*]

[1]State Key Laboratory of Microbial Metabolism, School of Life Science and Biotechnology, Shanghai Jiao Tong University, Shanghai 200240, P. R. China

[2]Key Laboratory of Smart Manufacturing in Energy Chemical Process, Ministry of Education, East China University of Science and Technology, Shanghai 200237, P. R. China

[3]These authors contributed equally.

*Corresponding authors.

Correspondence: xfyan@ecust.edu.cn and hongzhonglu@sjtu.edu.cn

Abstract

Pre-trained large models have emerged as a pivotal technological approach for foundational cell modeling. However, existing deep learning-based foundational models for cells have predominantly focused on human or murine systems, with a relative scarcity of research on model microorganisms such as Saccharomyces cerevisiae. Furthermore, these models often exhibit limitations in the integration of biological prior knowledge. To address this gap, we introduce scYeast, the first foundational cell model specifically designed for yeast that deeply integrates biological priors. scYeast features an innovative asymmetric parallel architecture that embeds transcriptional regulatory prior information directly into the Transformer's attention mechanism, thereby systematically incorporating and leveraging established biological knowledge during model training. After large-scale pre-training on single-cell transcriptomics data from yeast, scYeast demonstrates robust generalization capabilities and strong biological interpretability. It can perform zero-shot tasks, such as inferring specific regulatory relationships and resolving critical cell states, functional types, and developmental trajectories. Moreover, by constructing fine-tuning networks, scYeast excels at tasks including cell type identification, doubling time prediction, and forecasting responses to gene perturbations. Further leveraging transfer learning, scYeast can be adapted to other multi-omics data, such as proteomics, expanding its application boundaries in systems biology research. scYeast not only provides a novel tool for fundamental research in yeast single-cell biology but also establishes a new paradigm for the organic fusion of foundational models with prior knowledge, laying a solid groundwork for synthetic and systems biology.

Introduction

In recent years, foundation models, with the Transformer architecture at their core, have been reshaping the research paradigms in both natural language processing and life sciences [1]. By leveraging self-supervised learning frameworks and large-scale parameterization, these models can learn highly abstract and transferable biological representations from vast and complex datasets. This provides a unified computational framework for addressing intricate biological questions, such as understanding cellular regulation, integrating multi-omics data, and modeling a panoramic view of life activities. These in silico cell models can be employed to describe the process of transcriptional regulation within cells, and constructing accurate network models is of profound significance for elucidating transcriptional regulatory mechanisms and discovering key regulatory relationships [2]. With the advancement of computational technologies, the forms of virtual cell models have continuously evolved, encompassing Boolean models [3-5], kinetic equation models [6-8], and traditional machine learning models [9-11]. However, the introduction of deep learning has made it possible to construct foundational cell models on an unprecedented scale [12].

The widespread application of single-cell sequencing technologies has led to the accumulation of massive volumes of single-cell transcriptomics data, laying the data foundation for large models based on deep learning frameworks like the Transformer. Currently, most deep learning-based foundational cell models have focused on human or murine cells, significantly enhancing the capabilities for cell state identification and the understanding of physiological mechanisms through the use of Transformer architectures. For instance, scBERT [13] developed a pre-trained deep neural network model, adopting the pre-training and fine-tuning methodology of BERT. It gains a general understanding of gene-gene interactions through pre-training on a vast amount of unlabeled scRNA-seq data, which is then transferred and fine-tuned for supervised cell type annotation tasks on unseen scRNA-seq data. scFoundation [14] developed an asymmetric Transformer-style architecture and pre-training task design, encompassing approximately 20,000 genes and pre-trained on over 50 million human single-cell transcriptome profiles, enabling it to effectively capture the complex contextual relationships between genes across various cell types and states. DeepIMAGER [15] infers cell-specific regulatory networks through deep learning and data integration, employing a supervised learning approach that converts the co-expression patterns of gene pairs into image-like representations and utilizes transcription factor (TF) binding information for model training. Geneformer [16] introduced a modified Transformer encoder with a latent array to eliminate dependency on nucleotide sequences. DeepRIG [17] proposed a graph-based deep learning model that constructs a prior regulatory graph by transforming the gene expression patterns of the data into

co-expression patterns. It then utilizes a graph autoencoder model to embed the global regulatory information contained in the graph into latent gene embeddings, thereby reconstructing the gene regulatory network.

However, despite the significant advancements that deep learning has brought to high-dimensional modeling and the discovery of regulatory mechanisms from single-cell data, existing methods still fall short in the deep integration of biological prior knowledge. The majority of current studies merely embed prior knowledge as a form of multi-modal data, concatenating it with input features to be fed into the model, which fails to directly demonstrate the effective utilization of this knowledge during the model's inference process. For example, GRNPT, a novel Transformer-based framework, leverages textual descriptions of individual genes from the NCBI database, generating embedding vectors via a large language model to capture biological information, which are then used as part of the input data for inferring regulatory relationships [18]. GeneCompass integrates four types of prior biological knowledge—gene connectivity networks, promoter information, gene family annotations, and co-expression relationships—by creating separate embeddings for each category and combining them into a joint input vector for training a deep learning model [19]. Similarly, DNA sequences, protein sequences, and gene descriptions have also been separately embedded to predict inter-genic relationships [20]. It is evident from these approaches that prior knowledge is treated as a multi-modal input, where different forms of information are independently embedded, concatenated into a new vector, and then used for model training and prediction. Analogous methods for integrating multi-modal data are also employed in other research areas[21-23]. Consequently, when faced with complex biological scenarios, the interpretability of existing large models and the efficacy of their biological knowledge utilization remain limited. More importantly, the vast majority of current single-cell deep learning foundation models are focused on human or murine systems, with a notable absence of general-purpose large models for model microorganisms like yeast. As a model organism, Saccharomyces cerevisiae is extensively used in studies of metabolic engineering and transcriptional regulation [24-26], and a wealth of research findings and prior knowledge has been accumulated for it [27-29]. Therefore, constructing a foundational model for the yeast Transcriptional Regulatory Network (TRN) that fully leverages this prior knowledge is of significant value.

Here, we propose scYeast, a foundational model for Saccharomyces cerevisiae single-cells based on the deep fusion of prior knowledge. Inspired by DSGraphformer [30], scYeast introduces an asymmetric parallel transformer architecture that deeply integrates transcriptional regulatory

prior knowledge with the attention computation mechanism. This design ensures that scYeast thoroughly considers the incorporated prior knowledge throughout its training process. The model is first pre-trained in a self-supervised manner on a large corpus of unlabeled single-cell transcriptomics data. Subsequently, it can be adapted to a diverse array of downstream tasks, including the inference of condition-specific transcriptional regulatory relationships, classification of cell states and types, prediction of cell doubling time, and forecasting responses to gene expression perturbations. Furthermore, it supports transfer learning with proteomics data, enabling the prediction of a broader range of biological phenotypes. Overall, scYeast not only provides a novel tool for investigating the functional states and phenotypic variations of yeast cells but also presents a new research paradigm for the deep integration of foundational models with biological knowledge.

Results

Framework of scYeast

scYeast is a transformer-based pre-trained large model designed to construct a foundational model of the yeast Transcriptional Regulatory Network (TRN) that deeply integrates prior knowledge. Drawing inspiration from the static graph concept of DSGraphformer [30], scYeast employs an asymmetric parallel deep learning architecture. This design allows for the introduction of prior transcriptional regulatory relationships, in the form of an adjacency matrix, directly into the attention computation process. Consequently, the model can fully leverage this prior knowledge during computation and training. The overall architecture of scYeast is depicted in Figure 1a. The model begins by embedding single-cell transcriptomics data into high-dimensional vectors through expression token embedding. Concurrently, gene co-expression relationships are inferred from high-quality pan-transcriptome data, using a Spearman correlation coefficient greater than 0.5 as the threshold. Based on these inferred co-expression relationships, a name embedding vector is generated for each gene after an embedding step. Subsequently, the Expression Token Embeddings are randomly masked and then summed with the Gene Name Embeddings before being fed into two parallel encoder modules.

Within the overall framework of scYeast (Figure 1a), two parallel Transformer encoders collaboratively undertake the task of feature extraction, yet they focus on distinct sources of signals. The first branch employs the standard multi-head Query-Key-Value (QKV) attention mechanism, consistent with models like BERT. This "dynamic" encoder relies entirely on the single-cell transcriptomics data itself to capture potential novel gene expression dependencies that may emerge

under varying conditions. In parallel, the "static" encoder explicitly injects mechanistic prior knowledge (a transcription factor–target gene adjacency matrix) at the attention layer. After both encoders complete their respective contextual encoding, their outputs are passed to a learnable gating fusion unit. The fused representation is then processed through a Multilayer Perceptron (MLP) to reconstruct the original, unmasked input vector, thereby accomplishing the self-supervised pre-training (see Methods).

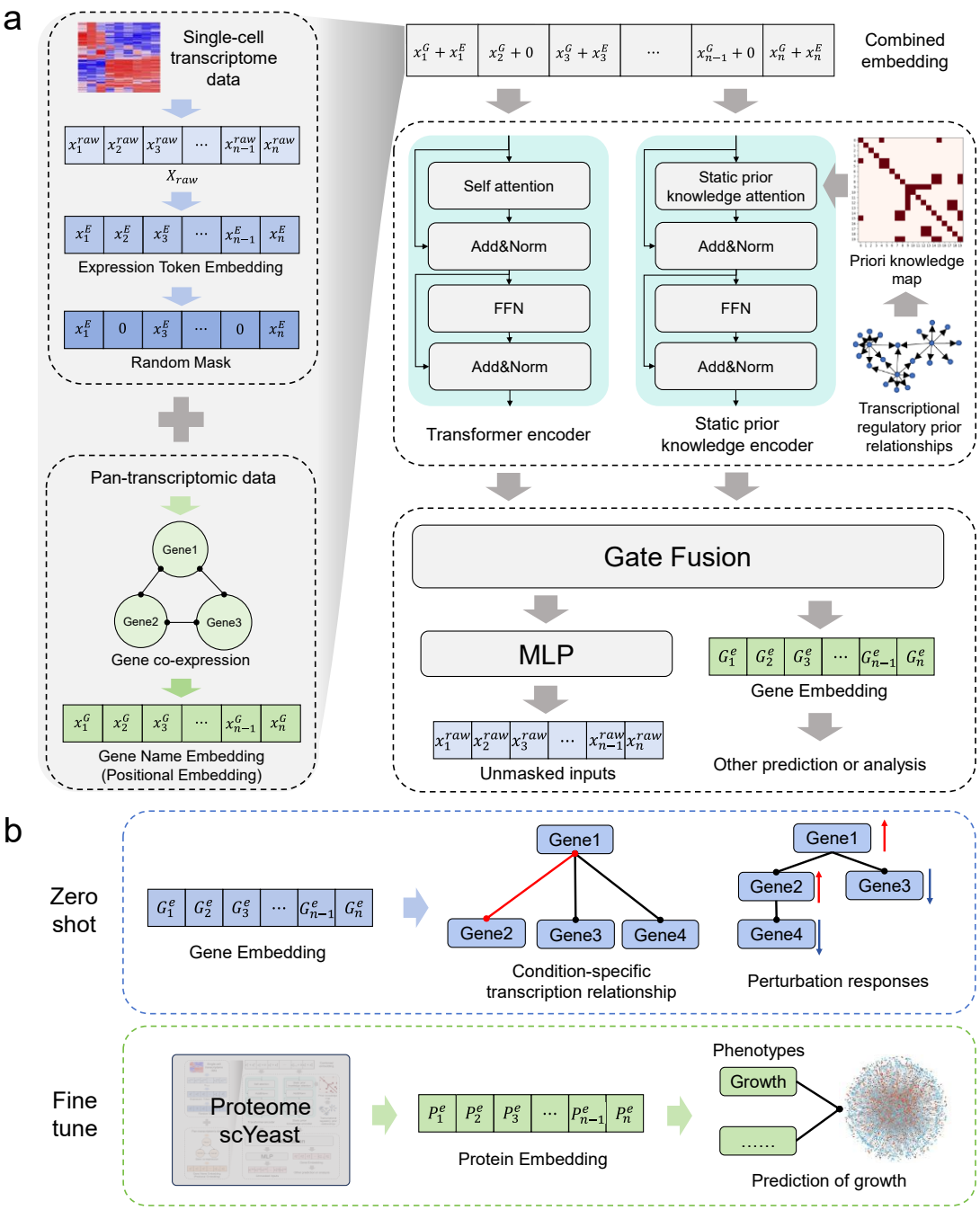

Figure 1. Framework and conceptual diagram of scYeast. (a) The overall framework of scYeast. It employs a parallel architecture to introduce prior knowledge, in the form of an adjacency matrix, into the attention mechanism, achieving a deep fusion of mechanistic knowledge. (b) scYeast can perform various prediction and analysis tasks, such as cell state classification and gene perturbation prediction, through either fixed-parameter or fine-tuning approaches. It can also be transferred to proteomics data to enable additional functions, including cell growth prediction.

Leveraging the features extracted by the pre-trained model and through model fine-tuning, scYeast can be broadly applied to yeast big data analysis, as well as to studies involving the association between omics and phenotypes (Figure 1b). Within the scYeast framework, with all parameters of the pre-trained model frozen, the output of the gate fusion unit for a new transcriptomic input can be regarded as a gene embedding that fully considers both data-driven features and prior knowledge. This embedding can be used to perform other analytical or predictive tasks. Furthermore, all parameters of scYeast can be fine-tuned using proteomics data to develop a proteomics-adapted scYeast model, enabling functions such as phenotype prediction based on protein abundance.

scYeast can analyze condition-specific transcriptional regulatory relationships

The scYeast model effectively leverages prior knowledge combined with information learned from massive pre-training datasets to generate high-dimensional gene embeddings that encapsulate both gene expression levels and regulatory information. This lays a solid foundation for the in-depth investigation of transcriptional regulatory relationships. Using single-cell transcriptomics data from the yeast aging process (young, early-aged, and late-aged) [31], we employed scYeast to explore the changes in intra-cellular gene regulatory relationships across different age stages (Figure 2a).

First, we used scYeast to obtain gene embedding vectors for each sample. Gene pairs with a cosine similarity higher than 0.9 within each sample were defined as highly-correlated gene pairs, suggesting a potentially strong interaction between them. We then took the intersection of these highly-correlated pairs within each of the three age states and compared these sets. The results indicate that the number of highly-correlated gene pairs significantly decreases as cells age. A total of 160,725 highly-correlated pairs were extracted from the young state, whereas only 99,608 pairs remained in the late-aged state. Furthermore, the number of overlapping highly-correlated gene pairs between adjacent age stages was greater (Figure 2b). Additionally, all genes appearing in the highly-correlated pairs were defined as "important genes." We found 297 important genes exclusive to the young stage, while only 38 and 39 were exclusive to the early- and late-aged stages, respectively. This demonstrates that the intracellular transcriptional regulatory landscape in

Saccharomyces cerevisiae undergoes dynamic changes with increasing age. To more intuitively display these age-dependent changes in gene regulation, we selected these important genes and plotted heatmaps for the three states (Figure 2c), which clearly show a significant shift and remodeling of gene-gene regulatory relationships as cells age. To obtain more direct statistical features, we binned and calculated the average cosine similarity for the important genes exclusive to each age stage (see Methods). It is evident that for each state, the average cosine similarity of its corresponding age-specific important genes is the highest. The overall average cosine similarity across all genes also decreases with aging (0.68 to 0.49 to 0.44), possibly reflecting a decline in intracellular gene activity with age.

GO enrichment analysis of the important genes from each stage revealed that among the genes exclusive to the young stage, 14 were significantly enriched (P value $= 0.00715$) in the "external encapsulating structure" under the cellular component annotation (Figure 2d). This reflects the high dependency of young cells on cell wall functions, including growth, division, morphology maintenance, and stress response [32]. As cells age, the function of the cell wall may gradually decline, leading to the disappearance of this significant enrichment. These findings fully demonstrate the rich information contained within scYeast's gene embeddings and underscore its immense potential for inferring gene regulatory networks and their dynamic changes.

Further, we conducted a case study on ASP3 (asparaginase II) to specifically analyze the inferred regulatory relationships. ASP3 is a key enzyme in nitrogen metabolism and stress response, playing a crucial role in various biological processes [33]. As cells age, the number of transcription factors inferred to regulate the ASP3 gene progressively decreases (Figure 2e). In young cells, the expression of ASP3 is synergistically regulated by multiple transcription factors, ensuring a rapid response to environmental changes. However, with advancing age, the regulatory network tends to be simplified, dominated by only a few core transcription factors. This may reflect a shift in the cell's regulatory focus and an optimization of energy expenditure [34]. While this regulatory simplification strategy might offer energy benefits, it could also reduce the adaptability of aging cells to environmental fluctuations and have potential negative impacts on other cellular functions.

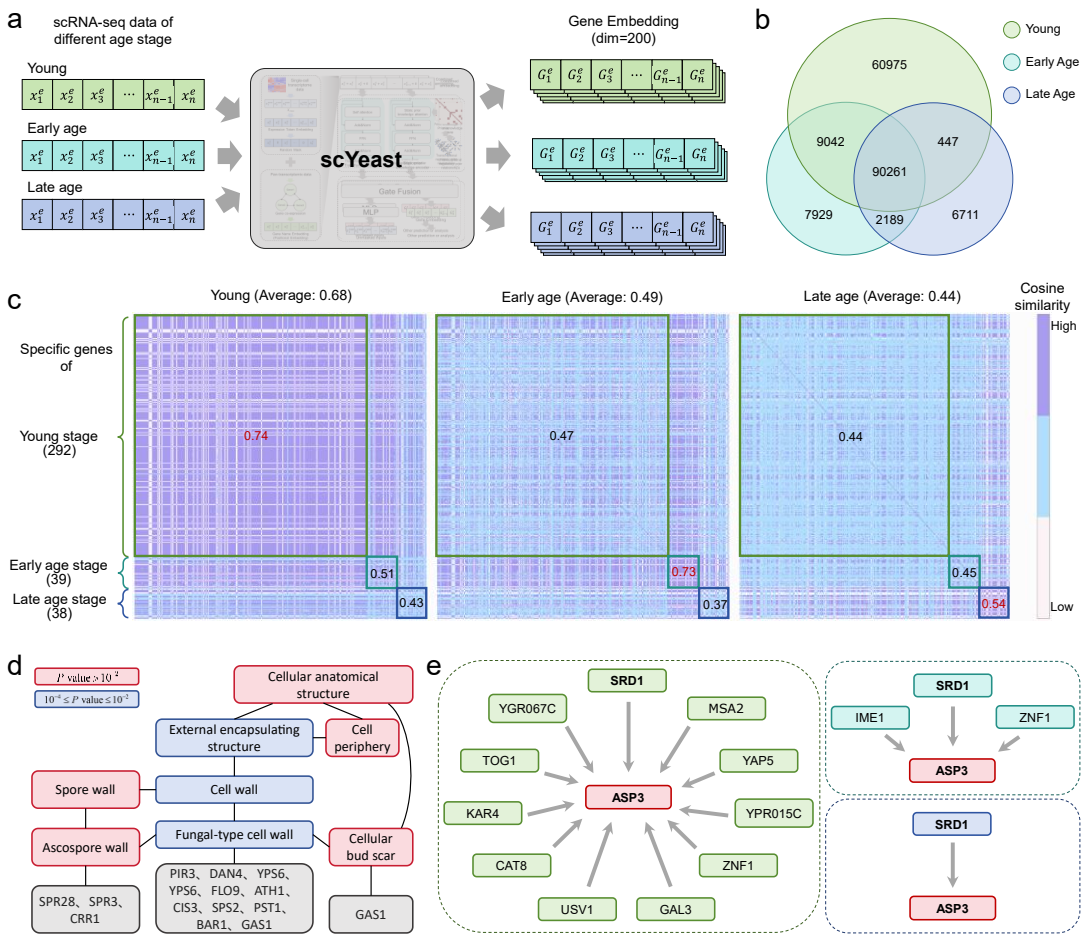

Figure 2. Inference of transcriptional regulatory relationships using scYeast. (a) Gene expression data from Saccharomyces cerevisiae at different age stages (young, early age, late age) are input into the scYeast model. The input data is embedded into 200-dimensional vectors, and the cosine similarity between these gene embeddings is calculated to infer regulatory relationships. (b) A comparative distribution plot of gene pairs with high similarity across the three age stages. (c) Heatmaps of the inferred important gene relationships at each age stage, clearly showing the changes and transitions in regulatory relationships with age. The mean similarity at each stage is calculated to reflect the changes in gene regulatory relationships at different ages. (d) GO enrichment analysis results for cellular components of important genes identified in the young stage. (e) A case study of ASP3, examining the changes in its inferred regulatory network with age, revealing a shift in the network's focus as aging progresses.

scYeast enables phenotype prediction and key gene discovery

scYeast can also achieve precise prediction of cellular phenotypes, including regression prediction of growth doubling time and classification prediction of different cellular states. To validate whether the prior knowledge-infused attention mechanism introduced in scYeast enhances performance in phenotype prediction, we trained an additional model. In this control model, the static, prior knowledge-based encoder was replaced with a standard transformer encoder, and it was used for phenotype prediction in an identical manner.

The network architecture for predicting yeast cell growth doubling time (regression) and different states (classification) is shown in Figure 3a. New data is fed into the pre-trained model to generate gene embeddings, which then serve as the input for a downstream MLP network that performs the classification or regression task. Concurrently, to verify that the proposed pre-training network has effectively learned features from the unlabeled data, we also fed the raw data directly into several traditional machine learning algorithms for a comparative analysis of prediction results (see Methods). The test results for the regression prediction of yeast growth doubling time are presented in Figure 3b. It is evident that the prediction results from models incorporating the pre-trained embeddings outperform those from traditional machine learning algorithms, indicating that the pre-trained model has successfully learned characteristic information from large-scale unlabeled data, which in turn helps to improve the accuracy of other predictive tasks. Furthermore, we observed that the pre-trained model incorporating prior knowledge surpassed the model without it, suggesting that the integration of prior knowledge during pre-training allows the model to further enhance its predictive accuracy. Under identical dataset partitioning, we further compared the effect of including prior knowledge in the fine-tuning network. The results show that the predictive performance of the model with added prior knowledge is superior overall to the one without, with a particularly significant advantage for samples with long doubling times.

Moreover, to extract additional biological insights, we can analyze this predictive model to identify the key genes that have the greatest impact on yeast cell growth doubling time. We calculated the gradient of the output (growth doubling time) with respect to each input variable. The results revealed that the absolute values of the gradients for four genes—SUL1, SNQ2, BNA7, and RPL32—were significantly higher than those of other genes (Figure 3b). Therefore, we hypothesize that these four genes may be the most critical ones affecting yeast growth doubling time. To validate the plausibility of this finding, we searched for the associated phenotypes of these four genes in the Saccharomyces Genome Database (SGD) and found that all of them are linked to the keywords "competitive fitness" and "growth." This provides a reasonable basis for considering these four genes as key determinants of yeast doubling time and further illustrates that our regression model can identify key genes influencing a target phenotype in yeast.

Similarly, the test results for the four-class classification task under different stress conditions are shown in Figure 3e, with several traditional machine learning classification methods included as controls. The conclusions drawn are highly consistent with those from Figure 3b: the classification accuracy of the pre-trained network is higher than that of traditional algorithms, and

the network incorporating prior knowledge outperforms the one without. The ROC curve comparison in Figure 3f between the pre-trained networks with and without prior knowledge further corroborates this conclusion. Subsequently, we calculated the gradients for the glucose starvation and amino acid starvation stress conditions within the four-class classification model to obtain sets of key genes associated with these two stress categories. GO enrichment analysis was performed on these gene sets, and the top ten enriched biological processes ranked by significance are shown in Figure 3g. We found that they are significantly enriched in multiple ribosome-related processes and amino acid metabolic processes. This is consistent with known biology, as yeast adaptation to glucose starvation involves specific changes in ribosomes, which cease to participate in protein synthesis. Likewise, under amino acid starvation stress, protein production is affected, leading to changes in ribosome-related processes. These results confirm the validity of the enrichment of the identified key gene sets in ribosome and amino acid metabolism-related pathways, further demonstrating that scYeast can enhance the predictive performance of downstream tasks and effectively identify key gene sets influencing phenotypic states, thereby facilitating genotype-to-phenotype association studies in yeast.

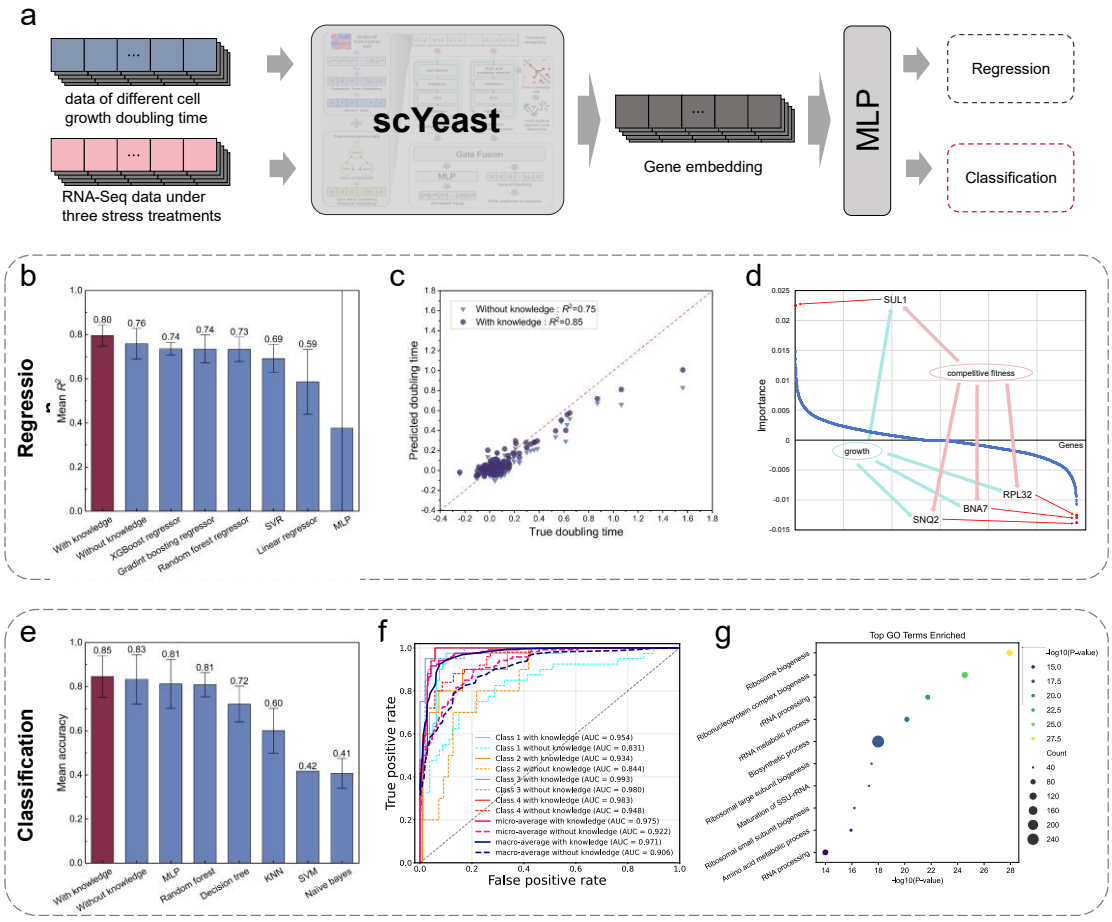

Figure 3. scYeast enables growth prediction and cell state classification tasks. (a) The network architecture for scYeast performing phenotype prediction tasks. The parameters of scYeast are fixed, and it is followed by an MLP for classification or regression prediction. (b) Comparison of results for the regression task of predicting growth doubling time. (c) A scatter plot comparing the prediction results on the test set for fine-tuned networks with and without the inclusion of prior knowledge, under the same dataset split. (d) The gradient distribution of key genes from the yeast growth prediction model, and the four key genes with the most significant gradients. (e) Comparison of results for the classification task of cell stress states, which are divided into four categories: isosmotic, hypoosmotic, glucose starvation, and amino acid starvation. (f) A comparison of the cell stress state classification results between the model with prior knowledge and the model without. The results are displayed using ROC curves, where solid and dashed lines distinguish the two types of models. The plot includes the individual ROC curves for each of the four classes and their average curve. (g) GO enrichment analysis results for the key genes identified by the cell stress state classification model.

scYeast can predict temporal responses to gene perturbations

scYeast can serve as a foundational model to enable the prediction of responses to gene perturbations. Since scYeast has already integrated prior knowledge such as gene co-expression and transcriptional regulatory relationships during its pre-training, we adapted the GEARs [35] framework by using scYeast as the base model for predicting gene perturbation responses (see Methods) (Figure 4a). The network for dynamic gene perturbation consists of three components: scYeast, a Gene Ontology (GO) graph network, and an MLP prediction network. First, the gene expression from the previous time point is used as input to scYeast to obtain a gene embedding representing the transcriptomic state at that moment. Subsequently, the perturbation information is fed into a graph convolutional neural network based on a GO relationship graph to obtain a high-dimensional embedding of the perturbation. Finally, the perturbation embedding and the transcriptomic state embedding are summed and input into an MLP for output prediction (see Methods). Following this approach, we constructed ten independent perturbation networks, each corresponding to a unique gene (Supplementary Table 1). Analysis of the training performance on the validation and test sets showed that each network reached its optimal validation performance within 15 epochs. Testing these ten optimal networks on the test set revealed that their test R^2 values ranged from a high of 0.97 to a low of 0.70 (Figure 4b). This indicates that all networks were able to accurately capture the dynamic changes of these ten genes in response to transcription factor perturbations, successfully achieving the prediction of gene perturbation responses.

To further validate the model's training performance, we selected AQY2 as a case study and retrained a dedicated perturbation prediction network for it. AQY2 is an aquaporin [36] that enhances the permeability of the cell membrane to water. Studying AQY2 helps to provide deeper

insights into how yeast cells regulate water influx and efflux, and its role in osmoregulation and other physiological processes. In the public IDEA dataset, the temporal changes in gene expression following stimulation by a transcription factor (TF) can be broadly categorized into three types: an S-shaped increase or decrease, a unimodal pulse, or remaining essentially stable (indicating the gene is not regulated by the stimulated TF) [37]. If the model truly understands the underlying regulatory logic, it should be able to reproduce these typical dynamic profiles even when the time-series data for a specific perturbation is completely withheld.

Based on this rationale, we entirely masked the time-series data from the intervention experiments of three TFs—GLN3, FZF1, and STE12 (strain numbers SMY117, SMY2066, and SMY2118, respectively)—from the training data, fine-tuning the network only on the remaining intervention conditions. This forces the model to rely on the learned regulatory priors and inter-genic temporal associations during the prediction phase, rather than performing simple regression. After 13 epochs of training, the optimal network achieved an R^2 of 0.7295 on the held-out test set (Figure 4c), indicating that the overall fit remained reliable. Subsequently, we reconstructed the complete temporal response of AQY2 under the three aforementioned interventions, starting from time point zero. The results are shown in Figure 4d: under GLN3 and STE12 interventions, the model-predicted expression of AQY2 exhibited an inverse S-shaped curve, consistent with the experimental sequencing data. In contrast, under the FZF1 intervention, the predicted trajectory remained close to the baseline, accurately reproducing the stable "no significant response" pattern observed experimentally. This is consistent with biological priors obtained from the YEASTRACT database: TFs GLN3 and STE12 are both confirmed to regulate AQY2, participating in the amino acid starvation response and the invasive growth signaling pathway in the cell cycle, respectively. In contrast, there is no known direct regulatory relationship between FZF1 and AQY2.

This experiment validates that scYeast can thoroughly mine the rich information implicit in transcriptomics data. The understanding of transcriptional regulatory relationships extracted by the model is sufficient to support de novo prediction of dynamic expression under unknown conditions. This provides a reliable computational tool for gene perturbation design and temporal control, further expanding the application space of the foundational transcriptional regulation model proposed in this study.

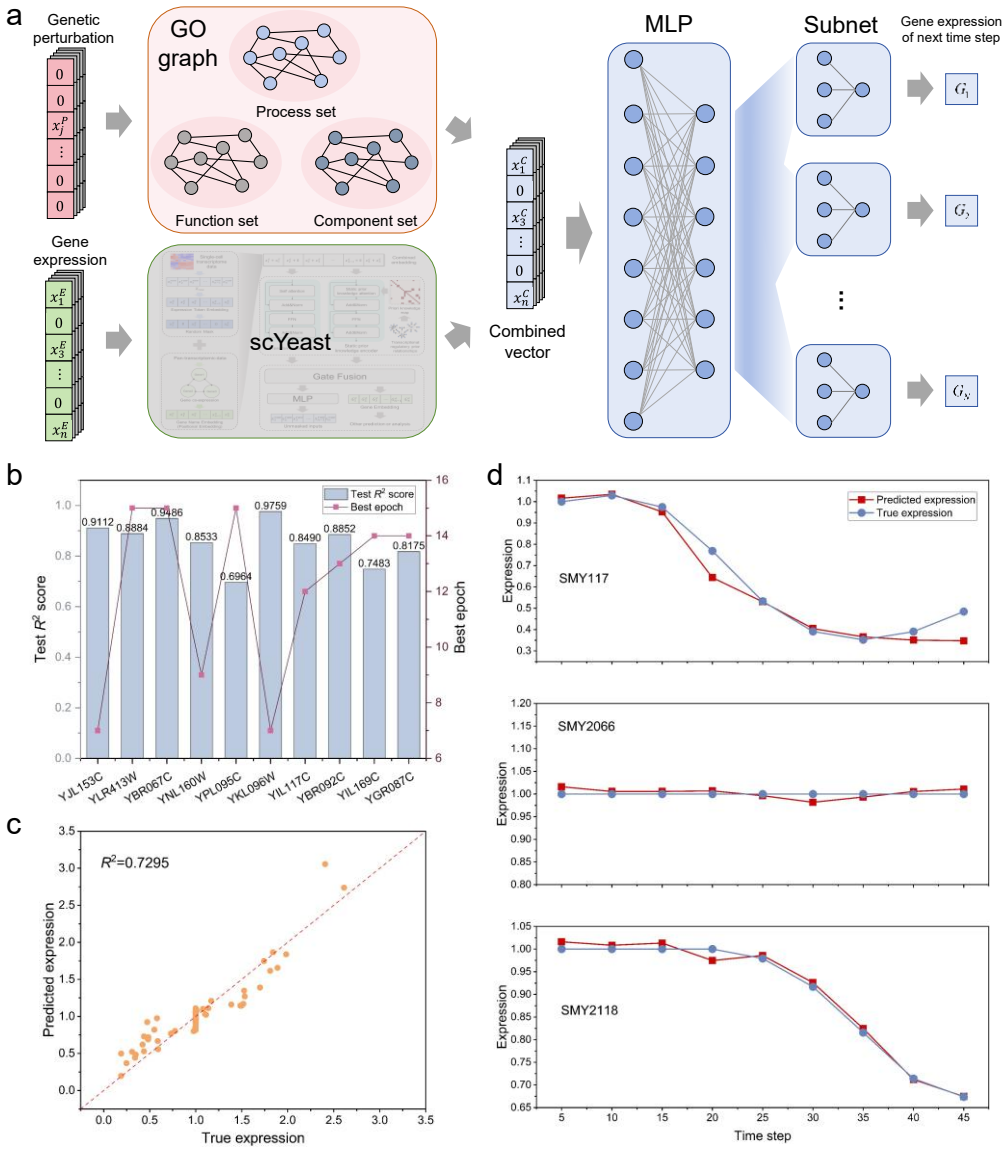

Figure 4. scYeast enables the prediction of gene perturbation responses. (a) The input includes gene expression data at a specific time point and information about the transcription factor perturbation applied to the strain. These inputs are processed separately through the pre-trained scYeast model and a graph neural network to extract embeddings. The outputs from both are then concatenated and passed through multiple linear layers to predict the target gene expression level at the next time point under the given condition. (b) Training results for subnetworks targeting ten key genes. The bar chart represents the R^2 scores of the subnetworks on the test set, while the curve indicates the training epoch at which the model achieved its optimal performance. (c) An R^2 scatter plot on the training set for the subnetwork designed to predict the gene AQY2, after removing three types of perturbations from the training data. (d) A comparison of the predicted full time-series against the actual observed values under three perturbations (SMY117, SMY2066, SMY2118) that were not included in the training. (SMY117, SMY2066, and SMY2118 represent different types of cellular perturbations).

scYeast can be transferred to proteomics data for growth prediction

To validate the multi-omics transferability of the scYeast model, we further explored its

application potential on proteomic data and constructed a transfer learning-based framework for growth rate prediction (Figure 5a). The proteomic data used was derived from a global proteome analysis of 4,699 Saccharomyces cerevisiae single-gene knockout strains[38], in which 1,850 proteins were quantified using micro-flow liquid chromatography-tandem mass spectrometry (micro-flow LC-SWATH-MS), covering approximately 79% of the yeast coding genome. Due to the significantly lower gene dimensionality of proteomic data compared to transcriptomic data, we employed a zero-padding method to expand the proteomic data to the same dimension as the transcriptomic data, thereby fitting the input structure of the pre-trained model. Subsequently, the same random masking auto-regressive training strategy was applied to the proteomic data to generate protein-specific embeddings. Finally, global features of the embeddings were extracted via a pooling operation and input into an Extra Trees ensemble regression model to predict three types of yeast cell phenotype information collected from different literature sources: growth rate[38] , ribosome occupancy of different genes (translation efficiency)[39], and protein half-life [40](see Methods).

To evaluate whether the pre-trained model contributed to its multi-omics transferability, we compared the training processes of a model with loaded pre-trained parameters against one with randomly initialized parameters (Figure 5b). The results showed that the model with loaded pre-trained parameters had a significantly lower initial reconstruction loss at the masked sites compared to the control group, and the loss value after convergence was also significantly lower. In contrast, the model with randomly initialized parameters, which did not load pre-trained parameters, exhibited markedly higher reconstruction loss, both initially and at convergence. This indicates that the pre-trained model effectively enhanced the feature learning efficiency and accuracy for proteomic data by transferring knowledge from the transcriptome.

Finally, we input the pooled features of the protein embeddings obtained from the model into an Extra Trees model for regression prediction of phenotypic data (Figure 5c-e). In all three tasks, the predictions based on protein embeddings demonstrated high R^2 values, indicating that the protein embedding representations obtained by scYeast through cross-omics knowledge transfer can successfully characterize the patterns of protein abundance changes and achieve non-linear prediction of different phenotype types. By selecting relevant genes or gene knockouts based on the model's feature importance parameters and performing GO enrichment analysis, we further revealed the biological pathways influencing growth, protein half-life, and gene translation. GO enrichment related to growth revealed multiple contributing factors, including "pseudohyphal growth" (p-value

= 0.006) as a morphological adaptation mechanism of yeast under nutrient deprivation, "transmembrane transport" (p-value = 0.009) as a key process for cells to acquire necessary nutrients and excrete metabolic waste, and "one-carbon metabolic process" (p-value = 0.039), which provides the foundation for nucleotide and amino acid synthesis, driving cell biomass accumulation and proliferation[41]; GO enrichment for ribosome occupancy was related to "positive regulation of TORC1 signaling" (p-value = 0.005) and "cellular response to amino acid starvation" (p-value = 0.006), which aligns with a recent study mentioning that TORC1, as a master regulator of growth and metabolism, has its activity influenced by amino acid availability and directly determines ribosome activity efficiency and specificity[42]; GO enrichment associated with protein half-life involved processes such as "protein monoubiquitination" (p-value = 0.016), "proteasome assembly" (p-value = 0.030), and "protein processing" (p-value = 0.023), which is consistent with the biological fact that the ubiquitin-proteasome system is a key pathway for protein degradation in eukaryotes, directly affecting protein stability and turnover. These GO enrichment results confirm the biological plausibility captured by the model. Through transfer learning, the model achieved a deep representation of proteomic data and provided complementary mechanistic explanatory perspectives for different phenotypes.

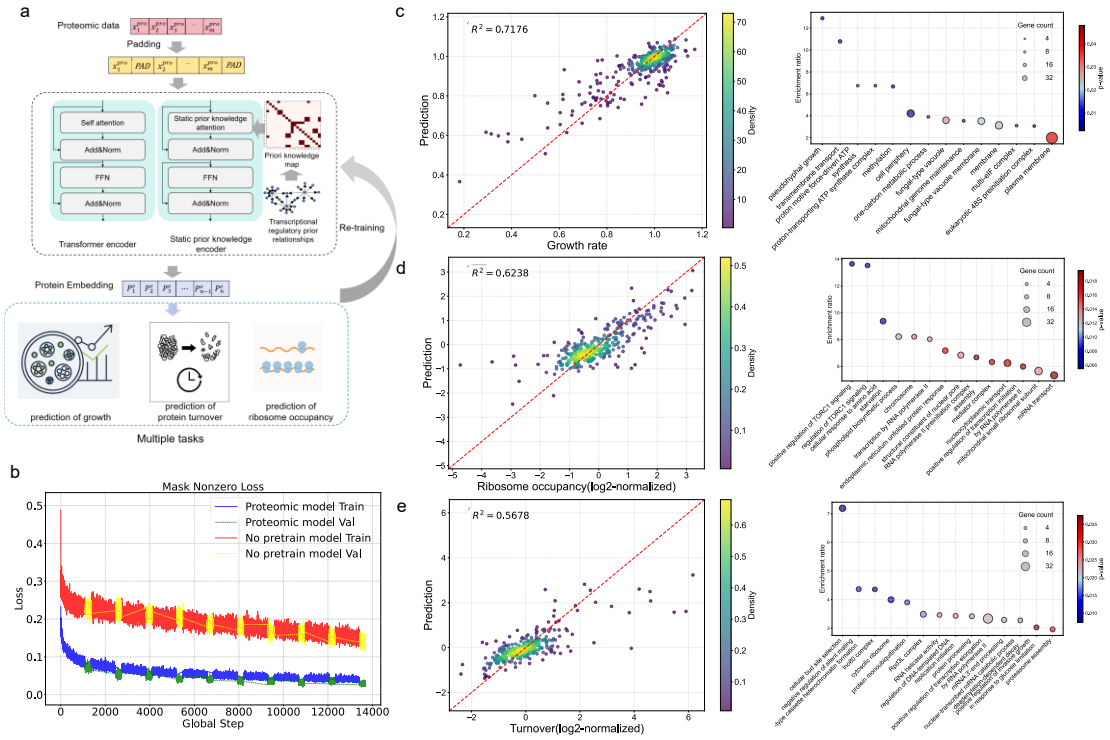

Figure 5. Transfer learning of scYeast to proteomics for phenotype prediction. (a) The network architecture for transfer learning of scYeast on proteomics data. (b) Iteration curves for transfer training with loaded scYeast transcriptome pre-trained parameters versus randomly initialized parameters. It is evident that the reconstruction error of the model with pre-trained parameters is lower than that of the model with randomly initialized parameters, demonstrating the effectiveness of the pre-trained model. (c) The R^2 result for growth phenotype prediction using the proteomic scYeast model, and the corresponding Gene Ontology (GO) enrichment analysis of genes based on the model's feature importance parameters. (d) The R^2 result for ribosome occupancy prediction using the proteomic scYeast model, and the corresponding GO enrichment analysis of genes based on the model's feature importance parameters. (e) The R^2 result for protein half-life (turnover) prediction using the proteomic scYeast model, and the corresponding GO enrichment analysis of selected genes based on the model's feature importance parameters.

Discussion

In this study, we have introduced scYeast—the first foundational single-cell model for yeast that deeply integrates prior regulatory knowledge. scYeast leverages an asymmetric parallel architecture to embed prior transcriptional regulatory information deeply into the Transformer's attention mechanism, yielding gene embeddings for each sample that fully consider both prior knowledge and data-driven features. Based on these gene embeddings, it is possible to infer condition-specific transcriptional regulatory relationships and extract biological insights. Ablation studies demonstrated that the pre-trained model, which integrates regulatory prior knowledge, can effectively learn feature information from large-scale, unlabeled single-cell transcriptomics data. This, in turn, enhances the accuracy of cellular phenotype predictions, such as doubling time and state classification, while also enabling the analysis and identification of key gene sets that critically influence these phenotypes. The potential of scYeast as a base model was also validated, as it exhibited outstanding performance in predicting gene perturbation responses, accurately simulating the dynamic changes of genes under different perturbations. Finally, by applying transfer learning with scYeast to proteomics data, we further confirmed the effectiveness of the pre-trained model, successfully applying it to proteomics-based growth prediction and key gene discovery. All these results indicate that scYeast, with its deep fusion of prior knowledge, has broad application prospects and can accomplish a variety of prediction and analysis tasks.

However, the prior knowledge in scYeast is based on a static relationship matrix derived from regulatory databases. This means that regardless of the conditions under which the input transcriptomics data were obtained, the transcriptional regulatory prior knowledge within the model remains fixed. In reality, studies have shown that yeast transcriptional regulatory relationships are dynamic [43, 44] , varying with different growth conditions and developmental stages. Therefore,

allowing the transcriptional regulatory prior knowledge to adjust dynamically with changes in the input data could further improve the model's accuracy. In the future, integrating more diverse forms of prior knowledge into the model could also be considered. At the same time, the scale of single-cell transcriptomics data for yeast has not yet reached its full potential. As data volume continues to accumulate, it will be feasible to construct pre-trained models with a larger number of parameters, thereby enhancing the precision of simulating transcriptional regulatory relationships, extracting more feature information, and further improving model accuracy. Furthermore, while this paper has highlighted several typical application scenarios, scYeast as a foundational pre-trained model can, in fact, be adapted for many more prediction and analysis tasks, which will be explored in subsequent research.

In summary, as the first pre-trained large model for yeast that deeply integrates prior knowledge, scYeast has demonstrated its universality and utility. It can provide a reliable tool for the study of transcriptional regulation in yeast and serves as a foundational pre-trained model for yeast research, enabling the rapid implementation of other predictive and analytical tasks, the discovery of key gene sets, and the acceleration of research into and elucidation of yeast mechanisms.

Methods

Dataset preparation

The datasets required for the construction and training of scYeast are described below. Each dataset underwent quality control and preprocessing.

Transcriptional Regulatory Prior Knowledge:

The prior knowledge of transcriptional regulation required by scYeast was obtained from the YEASTRACT database [45]. We selected the "documented targets" for each transcription factor, which represent regulatory relationships supported by literature evidence. This approach ensures a high-quality, evidence-based prior knowledge matrix of regulatory interactions, comprising approximately 170,000 regulatory relationships in total.

Pan-transcriptome Data:

The Saccharomyces cerevisiae pan-transcriptome data contains 969 high-quality transcriptomes, covering 4,977 core genes and 1,468 accessory genes [46]. This dataset enabled the calculation of the Spearman correlation coefficient for each gene across different strains, which served as the basis for inferring co-expression relationships in S. cerevisiae.

Single-cell Transcriptomics Data:

The single-cell transcriptomics data refers to the data required for the pre-training of scYeast, primarily consisting of two datasets: GSE125162 [47] and GSE242556 [48]. GSE125162 compiled data from wild-type controls and 11 different genotypes under various environmental conditions, determining their expression states by sequencing 38,285 single cells. GSE242556 involved continuous sampling of S. cerevisiae without metabolic labeling, measuring the expression of 173,361 individual cells. We performed quality control on the obtained single-cell datasets, removing cells with total read counts below 200. Ultimately, a dataset of approximately 210,000 cells was constituted for the pre-training of scYeast.

Yeast Aging Transcriptomics Data:

Transcriptomics data from different age stages of the yeast aging process were sourced from GSE210032[31]. This dataset includes 37, 43, and 45 single cells from the 2-hour (young), 16-hour (early-aged), and 36-hour (late-aged) age groups, respectively. On average, 2,202 genes were detected per cell, accounting for approximately one-third of the total genes in S. cerevisiae. To ensure the completeness of the analysis by scYeast, missing gene expression values were imputed with zeros.

Transcriptomics Data for Growth Prediction:

To predict yeast cell growth phenotypes from transcriptomics data, a dataset comprising transcriptomes of 1,143 single-gene knockout strains was collected [49, 50]. This dataset contains the transcriptomic data for each gene knockout and the corresponding cell doubling times under specific conditions. The cell doubling time was used as the prediction label to simulate and predict yeast growth phenotypes from transcriptomic data.

Stress Condition Data:

The data for classifying different yeast stress states were obtained from GSE201386 [51]. Using a full-length-based single-cell RNA sequencing method, 117 yeast cells were sequenced under four stress conditions (isosmotic, hypoosmotic, glucose starvation, and amino acid starvation). This data was used to perform classification prediction of these four cellular stress states.

Time-Series Perturbation Data:

The yeast time-series perturbation dataset was sourced from IDEA [37]. This dataset was

constructed by independently inducing hundreds of transcription factors (TFs) and measuring the time course of the resulting gene expression responses in budding yeast. IDEA encompasses over 200 TF induction experiments, which are followed by rapid changes in genes directly regulated by these TFs, as well as subsequent changes in indirectly regulated genes. The IDEA dataset provides time-series data on the differential expression of the complete transcriptome of Saccharomyces cerevisiae following perturbation by a single TF, typically spanning eight time points. However, due to the non-uniform distribution of time points in the IDEA dataset, we employed a radial basis function interpolation method with a Gaussian kernel to interpolate the data within a 45-minute window at 5-minute intervals.

GO Relationship Matrix:

For the time-series perturbation prediction, the perturbation relationship graph was derived from the Gene Ontology (GO) database [52, 53]. GO comprises three domains: Biological Process, Molecular Function, and Cellular Component. We retrieved all GO IDs for each gene across all domains and created connections between genes that share the same GO ID. The more GO IDs shared, the greater the connection weight. This process was used to construct an adjacency matrix of inter-genic GO relationships, which served as the perturbation relationship graph for the time-series perturbation prediction.

Proteomics Data for Growth Prediction:

The proteomics data was obtained from a previous global proteome analysis of the Saccharomyces cerevisiae non-essential gene knockout library [38] . This library consists of 4,699 gene knockout strains. Each strain was cultured in Synthetic Minimal (SM) medium, and its proteome was quantified using micro-flow liquid chromatography coupled with data-independent acquisition mass spectrometry (micro-flow LC-SWATH-MS). The experiment quantified a total of 1,850 proteins, covering approximately 79% of the yeast coding genome. The data underwent stringent quality control, including validation of technical replicates (median coefficient of variation, CV = 8.1%) and biological replicates (wild-type strain CV = 11.3%).

The data for the subsequent prediction tasks—growth rate, ribosome occupancy, and protein half-life—were sourced from different studies. The growth rate data, covering 4,550 of the aforementioned knockout strains, were obtained by measuring their growth phenotypes in three types of media (SC, SM, and YPD). The strains were cultured under standardized conditions in 96-well plates, and their growth curves were dynamically monitored by measuring OD600 to calculate

the maximum growth rate. The ribosome occupancy (ribosome protection footprint, RPF) data were derived from a study by McManus et al. [39], which measured translation rates (ribosome occupancy) on a genome-wide scale in S. cerevisiae using ribosomal profiling. The protein half-life data were sourced from a study by Martin-Perez and Villén [40], which systematically measured protein half-lives in S. cerevisiae using a metabolic labeling method.

Static prior knowledge attention

In scYeast, to more deeply integrate prior knowledge within the Transformer framework, we employed a static graph attention method. First, the transcriptional regulatory prior knowledge is stored in the form of an adjacency matrix P. Thus, P is a square matrix where the number of rows and columns equals the number of genes, and a value of 1 in P indicates that a regulatory relationship exists between the genes corresponding to the row and column. Subsequently, considering the attention calculation formula in the Transformer framework:

$$\text{Attention}(Q,K,V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V \tag{1}$$

where Q, K, and V are the query, key, and value matrices, respectively. The term $\text{softmax}(\frac{QK^T}{\sqrt{d_k}})$ can be defined as the attention weight matrix, which is also a square matrix with dimensions equal to the number of variables. When multiplied by the value matrix V, it represents the attention relationships between variables. Therefore, we considered replacing this attention weight matrix with the prior knowledge adjacency matrix P, signifying that the attention relationships between variables can be guided by the provided prior knowledge.

$$\text{Attention}(P,V) = P \cdot V \tag{2}$$

Here, P represents static prior knowledge; thus, the parameters within the P matrix are not updated during the model training process.

Gene name embedding

In scYeast, the gene name embedding is determined by gene co-expression relationships. First, using the pan-transcriptome data, we perform a Spearman correlation analysis on the expression levels of every gene pair. We then filter these correlations, setting a threshold of 0.5. That is, when the Spearman correlation coefficient is greater than 0.5, the two genes are considered to have a significant co-expression relationship, and they are saved as a gene pair. Subsequently, each gene pair is treated as a "sentence" in natural language processing. If a gene has no identified co-expression relationships, it forms a sentence consisting of a single "word." We then use the Word2Vec [54] method to obtain a 200-dimensional embedding vector for each gene. In scYeast,

this gene name embedding replaces the conventional positional embedding, and the position of each gene in the input data must be kept constant.

Expression token embedding

For the gene expression value embedding module, scYeast references the continuous embedding strategy from scFoundation [14] and adapts it to the specific characteristics of the yeast dataset. For the input single-cell transcriptomics data, the model first maps the raw expression value x to a high-dimensional vector representation. Specifically, non-zero expression values generate dynamic attention weights through a two-stage linear transformation. The first MLP layer projects the scalar value into a hidden space, and after LeakyReLU activation, a second MLP layer further learns non-linear associations across dimensions. The formulas are as follows:

$$v_1 = LeakyReLU(W_1 x + b_1),\ W_1 \in \mathbb{R}^{H\times 1} \tag{3}$$

$$v_2 = W_2 v_1 + \alpha \cdot v_1,\ W_2 \in \mathbb{R}^{H\times H} \tag{4}$$

where α (initialized to 1.0) is a learnable scaling factor. The scores are then obtained through Softmax normalization:

$$w = softmax(v_2) \tag{5}$$

Finally, a continuous embedding vector is generated by a weighted summation based on a predefined embedding table T :

$$E_{gene} = \sum_{i=1}^{H} w_i T_i,\ T_i \in \mathbb{R}^{H\times d} \tag{6}$$

where H is the predefined number of bins (default is 10) and d is the preset embedding vector dimension. For special tokens (zero values, masked positions, padding symbols), the model initializes separate embedding layers, E_{zero}, E_{mask} and E_{pad}, respectively. In the implementation, the positions of these special tokens are located by index and replaced with their corresponding embedding vectors. This continuous embedding strategy avoids the information loss associated with traditional binning of continuous data at fixed intervals and eliminates the problem of selecting an appropriate bin size. The gene-specific embedding and the expression value embedding are fused via element-wise addition to form the final input representation.

Pre-training of scYeast

scYeast adopts a pre-training strategy typical of large models. During the pre-training phase, scYeast employs a self-supervised, auto-regressive training process following random masking. Specifically, when the model receives single-cell transcriptomics data as input, the gene name embeddings are not masked. Only 15% of the expression token embeddings are randomly masked.

The model's output is the original single-cell transcriptomics data itself, thereby achieving auto-regressive training.

During pre-training, scYeast utilizes the Huber loss function. Considering the high sparsity of single-cell transcriptomics data, a large number of zero values could potentially impact the loss function and the training process during auto-regressive reconstruction. Furthermore, to emphasize the reconstruction performance on the masked portions, scYeast employs a composite loss function, as shown in the formula:

$$\begin{aligned} loss &= 0.6 \times masked_nonzero_loss + 0.3 \times nonzero_loss \\ &+ 0.4 \times masked_zero_loss + 0.1 \times zero_loss \end{aligned} \tag{7}$$

The training parameters for the model were set as follows: a batch_size of 16, the AdamW optimizer with a weight decay of 0.01, a learning rate set to 5e-5, and a maximum of 10 epochs. The model was constructed and trained within the PyTorch environment, utilizing GPU parallel computing for acceleration.

Predicting gene perturbation response

With scYeast serving as a base model, we predict the gene expression at the next time point by using the gene expression data from the previous time point as input, thus achieving single-step, time-series prediction of gene expression response. To mitigate the impact of the numerous zero values in the IDEA dataset, we applied an exponential transformation to the data. The input data consists of two parts. First, the gene expression data at a specific time point following TF perturbation is fed into scYeast to obtain context-aware gene embeddings. Second, the perturbation vector is input into the composite GO relationship graph to generate a perturbation embedding. These two embeddings are then summed and passed through multiple linear layers for progressive dimensionality reduction to predict the expression of a specific gene at the next time point.

Transfer learning for proteomic data

The knowledge learned by scYeast during its transcriptomic pre-training was transferred to the pre-training of a model for proteomics data. Since the dimensionality of proteomics data is much smaller than that of transcriptomics data, we expanded the dimensionality of the proteomics data to match that of the transcriptome by padding the gene positions that had no protein match. This ensured compatibility with the model's architecture. The new model adopts the same architecture as the transcriptomic scYeast. During its pre-training, two schemes were employed to validate the effectiveness of transferring knowledge from the transcriptomic network to the proteomic network: one involved loading the parameters from the transcriptomic model, while the other used randomly

initialized parameters. This model also utilized a self-supervised, auto-regressive training process following random masking. In the model's forward propagation architecture, the weights of the padded positions in the attention matrix were masked out to prevent the token information from non-protein-matched sites from interfering with the training.

Transfer-learning for proteomic data

The knowledge learned by scYeast during transcriptomic pre-training was transferred to the pre-training of the model on proteomics data. Since the dimensionality of proteomics data is much smaller than that of transcriptomics, the proteomics data dimension was expanded to be consistent with the transcriptome by padding the genes that had no protein match, making it compatible with the model architecture. The new model uses the same architecture as the transcriptomic scYeast. During the pre-training process, two schemes were used to validate the effect of transferring knowledge from the transcriptomic network to the proteomic network: one loading the parameters from the transcriptomic model and the other using randomly initialized parameters. The model also employs a self-supervised, auto-regressive training process following random masking. In the model's forward propagation architecture, the weights of the padded positions in the attention matrix are masked to prevent the token information from non-protein-matched sites from interfering with the training.

Prediction of growth rate, ribosome occupancy, and protein half-life based on proteomics data

The embedding vectors corresponding to the proteomic data, obtained from the pre-trained model, were processed through max-pooling and min-pooling, then concatenated to form feature vectors. These vectors were input into an Extra Trees (n=200) model to predict the growth rate of strains in YPD medium[38], as well as experimentally measured ribosome occupancy[39] and protein half-life[40]. The training was conducted over multiple rounds using a k-fold (k=10) training scheme. For comparison, the raw proteomic abundance information of the strains was also input as feature vectors into the Extra Trees model for prediction using the same scheme. The growth rate prediction was mainly used to analyze the effect of proteins on growth rate. The ribosome occupancy and protein half-life data were used as response variables to analyze the impact of gene knockouts on translation efficiency and protein turnover rate.

Key genes or gene knockouts were identified by screening based on the intrinsic feature parameters of the Extra Trees model. During the prediction process, the Extra Trees (n=200) model, trained on the pooled features of the proteomic embeddings, automatically assesses the predictive

contribution of each input feature to the target variable according to the built-in Gini importance calculation method of the scikit-learn framework. This importance score quantifies the extent to which a feature affects the model's output. These features were ranked by their importance scores, and the top 100 genes or gene knockouts were selected for subsequent GO enrichment analysis using goatools [53].

Acknowledgement

This work was financially supported by the National key research and development program of China (2020YFA0908300), Shanghai Municipal Science and Technology Major Project, and grant 22208211 and 22378263 from the National Natural Science Foundation of China (NSFC).

Code and data availability

The scYeast is available at the repository: https://github.com/hongzhonglu/scYeast .

Conflict of Interests

The authors have no competing interests to declare that are relevant to the content of this article.

Reference

1. Moor M, Banerjee O, Abad ZSH et al. Foundation models for generalist medical artificial intelligence, Nature 2023;616:259-265.
2. Badia-i-Mompel P, Wessels L, Müller-Dott S et al. Gene regulatory network inference in the era of single-cell multi-omics, Nature Reviews Genetics 2023;24:739-754.
3. Li L, Sun L, Chen G et al. LogBTF: gene regulatory network inference using Boolean threshold network model from single-cell gene expression data, Bioinformatics 2023;39.
4. Musilova J, Vafek Z, Puniya BL et al. Augusta: From RNA-Seq to gene regulatory networks and Boolean models, Computational and Structural Biotechnology Journal 2024;23:783-790.
5. Pušnik Ž, Mraz M, Zimic N et al. Review and assessment of Boolean approaches for inference of gene regulatory networks, Heliyon 2022;8.
6. Glass DS, Jin X, Riedel-Kruse IH. Nonlinear delay differential equations and their application to modeling biological network motifs, Nature Communications 2021;12:1788.
7. Ma B, Fang M, Jiao X. Inference of gene regulatory networks based on nonlinear ordinary differential equations, Bioinformatics 2020;36:4885-4893.
8. Xie C, Yang Y, Yu H et al. RNA velocity prediction via neural ordinary differential equation, iScience 2024;27.
9. Josephs-Spaulding J, Rajput A, Hefner Y et al. Reconstructing the transcriptional regulatory network of probiotic L. reuteri is enabled by transcriptomics and machine learning, mSystems 2024;9:e01257-01223.
10. Zito F, Cutello V, Pavone M. A Machine Learning Approach to Simulate Gene Expression and Infer Gene Regulatory Networks, Entropy 2023;25:1214.

11. Alawad DM, Katebi A, Kabir MWU et al. AGRN: accurate gene regulatory network inference using ensemble machine learning methods, Bioinformatics Advances 2023;3.
12. Li Z, Gao E, Zhou J et al. Applications of deep learning in understanding gene regulation, Cell Reports Methods 2023;3.
13. Yang F, Wang W, Wang F et al. scBERT as a large-scale pretrained deep language model for cell type annotation of single-cell RNA-seq data, Nature Machine Intelligence 2022;4:852-866.
14. Hao M, Gong J, Zeng X et al. Large-scale foundation model on single-cell transcriptomics, Nature methods 2024;21:1481-1491.
15. Zhou X, Pan J, Chen L et al. DeepIMAGER: Deeply Analyzing Gene Regulatory Networks from scRNA-seq Data, Biomolecules 2024;14:766.
16. Cui Z, Xu T, Wang J et al. GeneFormer: Learned Gene Compression using Transformer-Based Context Modeling. In: ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2024, p. 8035-8039.
17. Wang J, Chen Y, Zou Q. Inferring gene regulatory network from single-cell transcriptomes with graph autoencoder model, PLOS Genetics 2023;19:e1010942.
18. Weng G, Martin P, Kim H et al. Integrating Prior Knowledge Using Transformer for Gene Regulatory Network Inference, Advanced Science 2025;12:2409990.
19. Yang X, Liu G, Feng G et al. GeneCompass: deciphering universal gene regulatory mechanisms with a knowledge-informed cross-species foundation model, Cell Research 2024;34:830-845.
20. Chen Y, Xu D, Hammer R et al. Predicting Gene Relations with a Graph Transformer Network Integrating DNA, Protein, and Descriptive Data. In: 2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). 2024, p. 3101-3104.
21. Duan J, Xiong J, Li Y et al. Deep learning based multimodal biomedical data fusion: An overview and comparative review, Information Fusion 2024;112:102536.
22. Madan S, Lentzen M, Brandt J et al. Transformer models in biomedicine, BMC Medical Informatics and Decision Making 2024;24:214.
23. Huang D, Yan C, Li Q et al. From Large Language Models to Large Multimodal Models: A Literature Review, Applied Sciences 2024;14:5068.
24. Shi S, Chen Y, Nielsen J. Metabolic Engineering of Yeast, Annual Review of Biophysics 2025;54:101-120.
25. Hu C, Wang M, Zhu C et al. A transcriptional regulation of ERF15 contributes to ABA-mediated cold tolerance in tomato, Plant, Cell & Environment 2024;47:1334-1347.
26. Zheng J, Wen D, Tang C et al. The transcriptional regulation of Arabidopsis ECT8 by ABA-Responsive Element binding transcription factors in response to ABA and abiotic stresses, Physiology and Molecular Biology of Plants 2025;31:343-355.
27. Nadal-Ribelles M, Solé C, de Nadal E et al. The rise of single-cell transcriptomics in yeast, Yeast 2024;41:158-170.
28. Sahana GR, Balasubramanian B, Joseph KS et al. A review on ethanol tolerance mechanisms in yeast: Current knowledge in biotechnological applications and future directions, Process Biochemistry 2024;138:1-13.
29. Nguyen L, Schmelzer B, Wilkinson S et al. From natural to synthetic: Promoter engineering in yeast expression systems, Biotechnology Advances 2024;77:108446.
30. Sun Y, Yan X. Spatial-Temporal relation inference Transformer combined with dynamic relationship and static causality for batch process modeling and the application of erythromycin

fermentation, Engineering Applications of Artificial Intelligence 2025;146:110271.
31. Wang J, Sang Y, Jin S et al. Single-cell RNA-seq reveals early heterogeneity during aging in yeast, Aging Cell 2022;21:e13712.
32. Municio-Diaz C, Muller E, Drevensek S et al. Mechanobiology of the cell wall - insights from tip-growing plant and fungal cells, J Cell Sci 2022;135.
33. League GP, Slot JC, Rokas A. The ASP3 locus in Saccharomyces cerevisiae originated by horizontal gene transfer from Wickerhamomyces, FEMS Yeast Res 2012;12:859-863.
34. Baccolo G, Stamerra G, Coppola DP et al. Mitochondrial Metabolism and Aging in Yeast, Int Rev Cell Mol Biol 2018;340:1-33.
35. Roohani Y, Huang K, Leskovec J. Predicting transcriptional outcomes of novel multigene perturbations with GEARS, Nature Biotechnology 2024;42:927-935.
36. Laizé V, Tacnet F, Ripoche P et al. Polymorphism of Saccharomyces cerevisiae aquaporins, Yeast 2000;16:897-903.
37. Hackett SR, Baltz EA, Coram M et al. Learning causal networks using inducible transcription factors and transcriptome-wide time series, Molecular Systems Biology 2020;16:e9174.
38. Messner CB, Demichev V, Muenzner J et al. The proteomic landscape of genome-wide genetic perturbations, Cell 2023;186:2018-2034. e2021.
39. McManus CJ, May GE, Spealman P et al. Ribosome profiling reveals post-transcriptional buffering of divergent gene expression in yeast, Genome research 2014;24:422-430.
40. Martin-Perez M, Villén J. Determinants and regulation of protein turnover in yeast, Cell Systems 2017;5:283-294. e285.
41. Hammer SE, Polymenis M. One-carbon metabolic enzymes are regulated during cell division and make distinct contributions to the metabolome and cell cycle progression in Saccharomyces cerevisiae, G3: Genes, Genomes, Genetics 2023;13:jkad005.
42. Cecil JH, Padilla CM, Lipinski AA et al. The molecular logic of Gtr1/2-and Pib2-dependent TORC1 regulation in budding yeast, elife 2025;13:RP94628.
43. Shelansky R, Abrahamsson S, Brown CR et al. Single gene analysis in yeast suggests nonequilibrium regulatory dynamics for transcription, Nature Communications 2024;15:6226.
44. Liu L, Liu Y, Min L et al. Most Pleiotropic Effects of Gene Knockouts Are Evolutionarily Transient in Yeasts, Molecular Biology and Evolution 2024;41.
45. Teixeira MC, Viana R, Palma M et al. YEASTRACT+: a portal for the exploitation of global transcription regulation and metabolic model data in yeast biotechnology and pathogenesis, Nucleic Acids Research 2022;51:D785-D791.
46. Caudal É, Loegler V, Dutreux F et al. Pan-transcriptome reveals a large accessory genome contribution to gene expression variation in yeast, Nature Genetics 2024;56:1278-1287.
47. Jackson CA, Castro DM, Saldi GA et al. Gene regulatory network reconstruction using single-cell RNA sequencing of barcoded genotypes in diverse environments, eLife 2020;9.
48. Jackson CA, Beheler-Amass M, Tjärnberg A et al. Simultaneous estimation of gene regulatory network structure and RNA kinetics from single cell gene expression, bioRxiv 2023.
49. Sameith K, Amini S, Groot Koerkamp MJA et al. A high-resolution gene expression atlas of epistasis between gene-specific transcription factors exposes potential mechanisms for genetic interactions, BMC Biology 2015;13:112.
50. Kemmeren P, Sameith K, van de Pasch LA et al. Large-scale genetic perturbations reveal regulatory networks and an abundance of gene-specific repressors, Cell 2014;157:740-752.

51. Su Y, Xu C, Shea J et al. Transcriptomic changes in single yeast cells under various stress conditions, BMC Genomics 2023;24:88.
52. Ashburner M, Ball CA, Blake JA et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium, Nat Genet 2000;25:25-29.
53. Consortium TGO, Aleksander SA, Balhoff J et al. The Gene Ontology knowledgebase in 2023, Genetics 2023;224.
54. Church KW. Word2Vec, Natural Language Engineering 2017;23:155-162.