Beyond model organisms: robust prediction of functional properties across protein evolution

Lucas Waldburger[1,2] **Hunter Nisonoff**[1] **Marissa Zintel**[1] **Liam D. Kirkpatrick**[1,2]
Angelica Lam[1] **Nathan Lanclos**[1,2] **Jay D. Keasling**[1,2,3] **Max V. Staller**[1,4] **Patrick M. Shih**[1,2]

Abstract

Biological discovery and design are increasingly being guided by surrogate models trained on data from high-throughput assays in place of costly experimentation. However, existing datasets are often biased due to an overrepresentation from model organisms, leading to failures when performing evolutionary studies in non-model species. We present a hybrid framework that leverages high-throughput molecular assays and active learning to quantify biological properties across evolutionary space. We focus on transcriptional activators, which contain activation domains (ADs) that promote gene expression. ADs are intrinsically disordered and poorly conserved, which limits their study using alignment-based algorithms. Here, we develop ADhunter, a high-capacity regression model that outperforms state-of-the-art algorithms in identifying and quantifying the strength of ADs. Predictive uncertainty was used to guide evolutionary sampling across 7,842,516 proteins from 2,400 fungal genomes. We functionally characterized 9,836 ADs from 1,071 fungal genomes, providing a 15.5-fold expansion in genome representation compared to existing datasets. Comprehensive sampling improved model generalizability and provides the first functional annotation for 3,416 proteins in non-model fungi, highlighting the importance of sampling from non-model genomes to build evolutionarily robust models for predicting biological properties.

1 Introduction

Advances in DNA synthesis and sequencing have enabled high-throughput assays that are transforming biological research [1], [2]. These methodologies have shifted experimentation from small-scale characterization of a few hypotheses to large-scale assays capable of evaluating hundreds to thousands of hypotheses in parallel. While small-scale experiments have elucidated gene-level principles, large-scale approaches enable researchers to uncover genome-level features that govern biological functions [3], [4].

In biological discovery and design, high-capacity regression models are increasingly replacing costly and time-consuming functional characterization with cheap and fast inference [5]–[8]. These surrogate models approximate the behavior of a more complex or computationally expensive function, enabling estimation or inference by serving as a proxy for biological properties. High-throughput assays provide functional measurements for training surrogate models, but existing datasets are often biased by overrepresentation of sequences from model organisms, limiting model generalizability in evolutionary studies (Fig. 1A). Decades of sequencing and experimentation have highlighted that model organisms represent only a fraction of biological diversity. While experimental scientists rely on model systems to elucidate biological principles through standardized procedures, high-throughput assays are limited by the constraints of DNA synthesis rather than by organismal origin. Evolutionary space represents a broad distribution of functionally enriched sequences, especially

[1] University of California, Berkeley, [2] Lawrence Berkeley National Laboratory, [3] Technical University of Denmark, [4] Chan Zuckerberg Biohub – San Francisco. Correspondence to: Lucas Waldburger <lwaldburger@berkeley.edu>.

Accepted to the AI for Science workshop (NeurIPS 2025).

from non-model organisms, that challenge high-throughput assays with edge cases, enabling a more rigorous evaluation of the assay's capacity to quantify biological properties (Fig. 1B). Furthermore, most algorithms are known to suffer from pathologies such as overconfident predictions and reduced accuracy when performing inference on regimes far from the training distribution [9], [10]. Therefore, models trained exclusively on sequences from model organisms are unreliable when performing inference on divergent sequences from non-model species. Successful design of predictive models for evolutionary studies thus requires comprehensive training datasets to capture underrepresented features relevant to a property of interest.

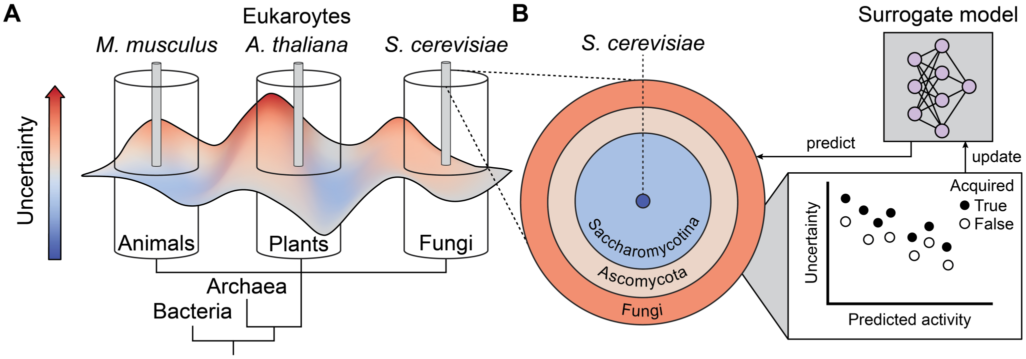

Figure 1: Surrogate models for evolutionary studies of biological properties. (A) Model organisms (e.g., *S. cerevisiae*, *A. thaliana*, *M. musculus*) are useful experimental systems, however, their sequence space represents a fraction of biological diversity for their respective branches of life. (B) Surrogate models used to predict biological properties trained exclusively on sequences from model organisms perform poorly in evolutionary studies. These models can be made into generalizable predictors of biological properties by acquiring labeled data with high uncertainty across protein evolution and updating the initial model.

One such property is sequence-to-function prediction of transcriptional activators, which promote gene expression [11]. Activation domains (ADs) undergo dynamic interactions with components of transcriptional machinery to enhance expression of target genes. Predicting the extent to which ADs promote gene expression remains difficult due to their intrinsic disorder, multiple modes of binding, and poor sequence conservation, limiting comparative genomics approaches [12]. Recently developed high-throughput assays have enabled quantification of thousands of ADs in parallel [11], [13], providing functional measurements for training surrogate models. While existing models can identify key properties of AD sequences, most datasets are biased toward the model fungus, *Saccharomyces cerevisiae* [14], [15], and the model plant, *Arabidopsis thaliana* [16], [17]. Surrogate models trained on these datasets can identify a functionally-conserved class of ADs known as acidic ADs [18]–[20], but fail to detect less-characterized AD classes, limiting their generalizability in non-model genomes. Since ADs lack structural constraints, they can explore a much larger sequence space compared to structured proteins, making model generalizability imperative for evolutionary studies of gene regulation.

In this study, we leverage the growing volume of sequencing data from non-model fungal genomes and active learning to discover transcriptional activators across fungal evolutionary space. Fungi are largely understudied, yet biosynthesize natural products that have transformed modern medicine [21]. Functional characterization has primarily focused on sequences from *Saccharomycotina*, while the broader protein diversity across *Ascomycota* and other fungal divisions remains largely uncharted. We developed ADhunter, a high-capacity regression model that outperforms the state-of-the-art model, TADA [22], at identifying and quantifying the strength of transcriptional activators. Machine-based uncertainty was used to predict the activity for 7,842,516 proteins across 2,400 fungal genomes. The predicted activity and associated uncertainty guided the acquisition and downstream functional characterization of 9,836 proteins from 1,071 fungal genomes. We demonstrate how performing active learning on non-model genomes significantly enhanced ADhunter's ability to quantify the activity of diverse transcriptional activators, especially non-acidic sequences that are leucine- and phenylalanine-enriched. Our results demonstrate how integrating high-throughput assays with active learning from non-model genomes enables scaling of genome-level characterization towards evolutionary-level functional genomics.

2 Results

2.1 Sequence-to-function modeling for precise quantification of transcriptional activators

Algorithms that are robust across evolutionary space require comprehensive training datasets to predict a biological property of interest. We sought to create a regression model to quantitatively predict AD activity. A quantitative model would enable accurate identification of AD boundaries and peak activity in natural sequences to study intrinsically disordered protein evolution, such as in non-model organisms that have evolved genetic regulation to biosynthesize natural products and adapt to ecological niches. Furthermore, a quantitative model can be used in protein engineering contexts to design transcription factors (TFs) for fine-tuned gene expression in synthetic biology.

We systematically evaluated protein sequence representations to enhance sample efficiency, using performance on a held-out test dataset as the benchmark. Functional characterization of AD sequences is costly and time-consuming, making it essential for surrogate models to extract the most information from each datapoint. Binary encodings of protein sequences are frequently used due to their simplicity; however, this approach fails to capture secondary structure and biochemical properties of amino acids [23], [24]. Alternatively, features from human-selected sequence descriptors may introduce biases. Combining both approaches can lead to unequal weighting and increased complexity, resulting in reduced performance (Table 3). For instance, PADDLE [15] performs well at identifying acidic ADs, but had poor quantitative performance on the initial dataset (Pearson r = 0.261; RMSE = 0.336). Excluding secondary structure predictions improved performance (Pearson r = 0.338; RMSE = 0.329). Continuous neural encodings from pretrained protein language models provide general, task-agnostic features that capture evolutionary signals [25]–[27].

Our initial dataset consists of 17,609 53 AA tiles from fungal and plant proteins previously characterized for AD activity with a high-throughput assay [16]. We trained a convolutional neural network (CNN) and found that neural encodings from an evolutionary-scale protein language model (ESM) [28], [29] outperformed one-hot and all other representations on a held-out test dataset (Pearson r = 0.744; RMSE = 0.664). These results indicate that neural encodings from pretrained models outperform simple encodings, and that pretrained protein language models learn representations that are informative for modeling intrinsically disordered and poorly conserved regions like ADs.

After determining the protein sequence representation, we evaluated model architectures to balance optimal performance while minimizing model complexity. Neural encodings combined with lightweight regressors have been shown to outperform complex architectures trained on simpler features [29]. A CNN with residual connections (ResNet) had the highest performance on the held-out test dataset. This model with ESM encodings was named ADhunter. The state-of-the-art AD predictor model, TADA [17], has been shown to outperform first-generation models ADpred [14] and PADDLE [15]. When trained on the initial dataset, ADhunter achieved better quantitative performance than TADA (Pearson r = 0.538; RMSE = 0.995) (Table 4). An ablation study of TADA's architecture reveals that the CNN layer alone is sufficient to achieve optimal predictive performance (Table 5). These results demonstrate how a ResNet trained with neural encodings can outperform a more complex architecture trained on human-selected features at quantitatively predicting AD activity.

2.2 Deep ensembling enables uncertainty estimation and improves model generalizability

Quantifying uncertainty enables researchers to distinguish between confident predictions and those that are unreliable, guiding prioritization of candidates for experimental validation and exploration of uncharacterized sequence space. Existing predictive models of AD activity have no notion of uncertainty, leading to biased, overconfident, or misleading predictions. While neural networks do not inherently represent uncertainty, deep ensembles provide an approach for estimating prediction uncertainty [30]. We used a deep ensemble for scalable and robust estimates of epistemic uncertainty in high-dimensional, non-linear settings. This approach avoids the restrictive kernel assumptions and computational bottlenecks of sparse Gaussian processes, as well as the optimization and calibration challenges of Bayesian neural networks. Machine-based uncertainty can inform the optimization of predictive models in evolutionary studies by guiding sampling from non-model genomes to build a more comprehensive training dataset.

Uncertainty can originate from the assay, known as aleatoric uncertainty, or from the predictive model, known as epistemic uncertainty. We incorporated machine-based uncertainty into ADhunter to quantify epistemic uncertainty (See Appendix 5.1). Furthermore, model ensembling improves predictive performance by combining the strengths of multiple regressors, which reduces individual model biases. As expected, the model ensemble outperformed a single model on a held-out test dataset. In particular, performance improved as a function of ensemble size and we selected an ensemble with 20 models (Pearson r = 0.775; RMSE = 0.632) for computational tractability of downstream tasks. These results indicate that the model ensemble improved ADhunter performance and provides a machine-based estimation of uncertainty to select diverse sequences that will improve model generalizability for evolutionary studies.

We simulated out-of-distribution inference by performing spectral clustering of the training dataset. When evaluated on a held-out test cluster, TADA (Pearson r = 0.381; RMSE = 1.021) underperformed relative to ADhunter (Pearson r = 0.512; RMSE = 0.826) (Table 6). A single instance of ADhunter with neural encodings outperforms one-hot encodings, and ensembling further improved out-of-distribution performance. These results indicate that ADhunter has superior generalizability performance relative to the state-of-the-art AD prediction model.

2.3 Evolutionary sampling of transcriptional activators using machine-based uncertainty

We sought to further improve ADhunter for use in evolutionary studies with active learning from sequences in non-model genomes. Active learning involves performing inference on a design space, functional characterization of selected sequences defined by an acquisition function, and updating the model on the active learning dataset [31], [32]. This framework has successfully been used in protein engineering applications, such as in directed evolution, where the objective is to optimize a property of interest [33]–[35]. In comparison to structured proteins, we have limited knowledge of the activity landscape for disordered proteins, such as transcriptional activators. For studying gene regulation, our objective is to minimize prediction error. In place of mutagenesis libraries for protein engineering, we created an active learning dataset from the sequence space sampled by evolution.

To enhance ADhunter's robustness across evolutionary space, we performed *in silico* discovery of ADs. From the MycoCosm collection [36] we obtained 2,400 fungal genomes totaling 87,542,943 proteins. Sequences were deduplicated, resulting in 9,395,825 unique proteins (Fig. 2A). We clustered sequences with 90% identity to retain maximum diversity of protein space. The 7,842,516 representative sequences from each cluster were sliced into 53 AA tiles with 10 AA stride, totaling 72,057,539 tiles. ADhunter predicted the activity and associated uncertainty for each tile. In contrast to previously characterized sequences from the model fungus, the evaluated sequences encompass both unicellular and multicellular fungi with diverse morphologies and ecological niches.

Using quantile-balanced uncertainty sampling as our acquisition function, we selected 8,935 sequences with maximum uncertainty across the range of predicted activity. The acquired sequences are represented in 1,050 of 2,400 fungal genomes (43.8%) compared to 21 fungal genomes (0.00875%) from the initial dataset and 69 fungal genomes (0.0288%) from the *S. cerevisiae* TF dataset [15] (Fig. 2B). Once harmonized, a process that maps the new dataset onto the same distribution of empirical activity as the initial dataset, the updated dataset will represent the diversity from 1,071 fungal genomes (44.6%). Interestingly, acquired sequences with high uncertainty originate from all clades except *Saccharomycotina*, which is sampled by the initial dataset and contain the model fungus *S. cerevisiae*. Our active learning dataset provides comprehensive exploration of evolutionary space for functional characterization of diverse ADs from non-model fungi, including deeper sampling within *Ascomycota* and the first high-throughput characterization across fungal divisions.

Completing the active learning cycle requires functional characterization of the new dataset, harmonizing the initial and new datasets, then retraining ADhunter on the harmonized dataset. We used a previously developed high-throughput assay to quantify AD activity [11]. Overall, the test tiles from non-model fungi had lower activity relative to the initial dataset. Interestingly, there was poor correlation between the empirical and predicted activity for test tiles from non-model fungi (Pearson r = 0.178; RMSE = 1.15). These results indicate that there are novel sequence-to-function relationships within proteins from non-model fungi in the active learning dataset that were not in the initial dataset.

The distribution of net charge versus fraction of non-polar residues in the initial dataset and the *S. cerevisiae* TF dataset show that prediction error originates outside the sampled distribution (Fig. 2C).

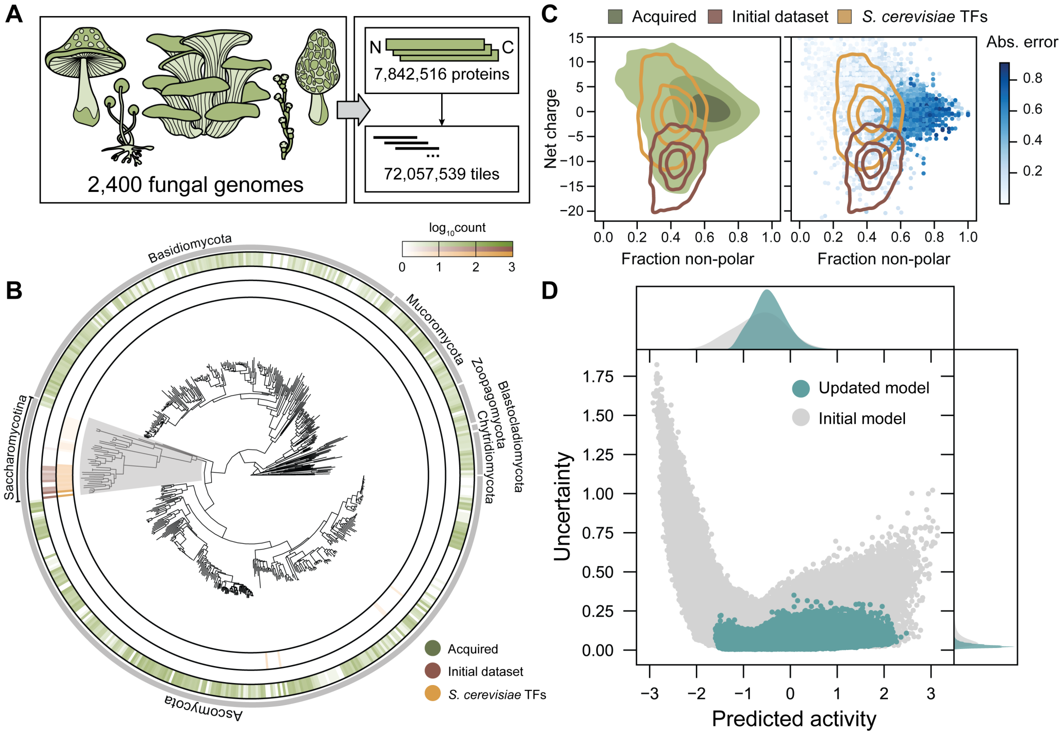

Figure 2: Machine-guided exploration of fungal protein evolution. (A) The MycoCosm collection consists of 9,395,825 unique proteins from 2,400 fungi. Clustering by 90% sequence identity resulted in 7,842,516 proteins that were sliced into 72,057,539 tiles to perform inference with ADhunter. (B) Test tiles from evolutionary space (green) represent the diversity of 1,050 non-model fungal genomes compared to 21 fungal genomes from the initial dataset (brown) and 69 fungal genomes from the *S. cerevisiae* TF dataset (yellow). Uncertainty sampling avoided sequences from the *Saccharomycotina* subdivision (grey), which contains the model fungus *S. cerevisiae*. (C) Tiles acquired from non-model fungal genomes exhibit more diverse sequence properties, particularly in net charge and fraction of non-polar residues, compared to prior datasets. The initial model fails to accurately predict the activity of tiles with sequence features that lie outside the range observed in previously characterized datasets. (D) Performing inference using the updated model across fungal protein evolution reduces uncertainty, enabling ADhunter to better capture sequence-to-function relationships for use in evolutionary studies of gene regulation.

The active learning dataset acquired sequences that span both distributions with a higher overall fraction of non-polar residues. In particular, the tiles with the strongest activity from non-model fungi, as well as those with the largest prediction error, contain a higher fraction of leucine and phenylalanine residues compared to previously characterized datasets. Tiles above the median absolute prediction error are found in 7,236 proteins across 849 fungal genomes. Therefore, adding these sequences with high uncertainty to the training dataset should enable ADhunter to identify ADs from a larger sequence space relative to the initial dataset.

2.4 Active learning enables quantification of protein codes from non-model genomes

After harmonizing the datasets, we evaluated the improvement in performance from active learning on a new held-out test dataset. This dataset included sequences from both the initial and newly acquired dataset, representing the diversity of 406 fungal genomes. ADhunter trained on the initial dataset was able to perform well on the test sequences from yeast, but there was a clear subset of sequences with high prediction error and uncertainty from non-model fungi (Pearson r = 0.541; RMSE = 0.963). In particular, prediction error and uncertainty were highest outside sequences from *Saccharomycotina*. ADhunter trained on the harmonized dataset was able to identify these patterns and achieved much lower prediction error and uncertainty across the held-out test dataset (Pearson r = 0.824; RMSE =

0.570). These results demonstrate that active learning reduced uncertainty in ADhunter and identified novel sequence-to-function relationships of ADs from a diverse sampling of fungal evolutionary space. Therefore, active learning improved ADhunter's ability to generalize across fungal proteins, enabling deeper insights for evolutionary studies of gene regulation.

We further evaluated ADhunter with respect to TADA by partitioning performance contributions attributed to the dataset composition versus the prediction task (Table 1). Comparison across models is complicated by differences in AD sequence lengths. To assess the role of AD sequence length, we retrained ADhunter and TADA separately on the 40 AA from Morffy et al. or the 53 AA harmonized dataset. ADhunter achieved superior performance relative to TADA when evaluated on both held-out test datasets. Comparison across models is further complicated by differences in model optimization objectives. ADhunter minimizes a regression objective (i.e., mean squared error loss) whereas TADA minimizes a classification objective (i.e., focal loss). To assess classification performance, we binarized the predicted activity by ADhunter as described in Morffy et al. and found that ADhunter outperformed TADA. To assess regression performance, we used the TADA score as the predicted activity and found that ADhunter provided more quantitative predictions. Overall, these findings further demonstrate that ADhunter achieves state-of-the-art performance across datasets and prediction tasks.

Prediction Task	**Model**	**Harmonized Dataset**	**Morffy et al. Dataset**

Regression	**TADA**	Pearson correlation = 0.621;
		RMSE = 0.935
			Pearson correlation = 0.635;
			RMSE = 0.961

	ADhunter	Pearson correlation = 0.818;
		RMSE = 0.556
			Pearson correlation = 0.682;
			RMSE = 0.740

Classification	**TADA**	Accuracy = 0.861; F1 score = 0.866	Accuracy = 0.924; F1 score = 0.922
	ADhunter	Accuracy = 0.905; F1 score = 0.910	Accuracy = 0.932; F1 score = 0.935

Table 1: ADhunter achieves state-of-the-art performance as a surrogate model for AD activity. ADhunter was compared to the state-of-the-art AD predictor, TADA, when trained and tested on either the harmonized dataset or the Morffy et al. dataset. ADhunter outperformed TADA on both datasets as well as regression and classification prediction tasks. Metrics are averaged across three random seeds.

Revisiting fungal protein evolution using ADhunter trained on the harmonized dataset revealed an overall reduction in uncertainty and a narrower range of predicted activity compared to ADhunter trained on the initial dataset (Fig. 2D). Since only a small subset of proteins activate gene expression, this may account for the greater variance in uncertainty among sequences with low predicted activity compared to those with high predicted activity in the initial model. Given that most proteins lack an AD, this likely explains the low median predicted activity across fungal protein evolution. The reduced uncertainty in the characterized tiles from non-model fungi and across fungal protein evolution suggest that active learning improved model generalizability. ADhunter enables accurate quantification of ADs across fungal protein evolution, and our framework can be extended to study biological properties across underexplored branches of life.

3 Discussion

Biological discovery and design are increasingly being guided by surrogate models trained on data from high-throughput assays. While large-scale experimentation is often performed in model species, acquisition of ground-truth labels should not be limited to sequences from these organisms. In this study, we demonstrate how a deep ensemble and active learning can be used to comprehensively explore fungal protein evolution, thereby enhancing model generalizability for predicting biological properties. We use machine-guided exploration to traverse the sequence landscape of transcriptional activators in fungi, a largely uncharacterized branch of life that represent 26.9% of all eukaryotic reference genomes. Predicting properties of intrinsically disordered proteins remains a significant challenge, even with the most advanced computational models. We present a framework that fine-tunes pretrained neural encodings to accurately predict transcriptional activation of disordered proteins across evolutionary space. By integrating high-throughput assays with active learning, this framework extends beyond transcriptional activators to enable robust functional genomics models that quantify biological properties at an evolutionary scale.

ADhunter outperformed the state-of-the-art model in classification and quantitation of ADs for use in evolutionary studies. We optimized model performance by replacing binary protein representations with continuous neural encodings and deep ensembling. These features also improved model generalizability, as demonstrated by spectral clustering analysis. While design spaces in protein engineering often center on mutagenesis libraries, our approach leverages the rich diversity of sequencing data to explore protein codes in non-model genomes. Evolved proteins are highly diverse and enriched for functional sequences, offering a more expansive foundation for surrogate models to quantify biological properties.

A key feature of ADhunter is its integration of machine-based uncertainty. By prioritizing sequences with high uncertainty across the range of predicted activity, we focused experimental efforts on maximally informative samples. We performed functional characterization for a library of diverse sequences, which significantly improved ADhunter's ability to generalize outside *Saccharomycotina* and across fungal divisions. This approach enabled quantification of underrepresented sequence-to-function relationships from non-model organisms compared to existing models that tend to identify overrepresented patterns. As surrogate models trained on high-throughput datasets continue to guide biological discovery and design, addressing prediction biases from the training dataset composition is crucial. Our work highlights the importance of expanding functional characterization beyond model organisms to include sequences from non-model genomes. By leveraging advances in scalable molecular technologies and machine learning, we can accelerate the study of underexplored branches of life to identify universal principles of living systems and how machine-guided design can reprogram organisms for novel purposes.

4 References

[1] C. H. Ludwig, A. R. Thurm, D. W. Morgens, *et al.*, "High-throughput discovery and characterization of viral transcriptional effectors in human cells," *Cell Systems*, vol. 14, no. 6, 482–500.e8, Jun. 2023. DOI: 10.1016/j.cels.2023.05.008.

[2] N. DelRosso, J. Tycko, P. Suzuki, *et al.*, "Large-scale mapping and mutagenesis of human transcriptional effector domains," en, *Nature*, vol. 616, no. 7956, pp. 365–372, Apr. 2023. DOI: 10.1038/s41586-023-05906-y.

[3] M. L. Bileschi, D. Belanger, D. H. Bryant, *et al.*, "Using deep learning to annotate the protein universe," en, *Nature Biotechnology*, pp. 1–6, Feb. 2022. DOI: 10.1038/s41587-021-01179-w.

[4] T. Sanderson, M. L. Bileschi, D. Belanger, and L. J. Colwell, "ProteInfer, deep neural networks for protein functional inference," en, *eLife*, vol. 12, e80942, Feb. 2023. DOI: 10.7554/eLife.80942.

[5] K. K. Yang, Z. Wu, and F. H. Arnold, "Machine-learning-guided directed evolution for protein engineering," en, *Nature Methods*, vol. 16, no. 8, pp. 687–694, Aug. 2019. DOI: 10.1038/s41592-019-0496-6.

[6] Z. Wu, S. B. J. Kan, R. D. Lewis, B. J. Wittmann, and F. H. Arnold, "Machine learning-assisted directed protein evolution with combinatorial libraries," en, *Proceedings of the National Academy of Sciences*, vol. 116, no. 18, pp. 8852–8858, Apr. 2019. DOI: 10.1073/pnas.1901979116.

[7] C. N. Bedbrook, K. K. Yang, J. E. Robinson, E. D. Mackey, V. Gradinaru, and F. H. Arnold, "Machine learning-guided channelrhodopsin engineering enables minimally invasive optogenetics," en, *Nature Methods*, vol. 16, no. 11, pp. 1176–1184, Nov. 2019. DOI: 10.1038/s41592-019-0583-8.

[8] S. Biswas, G. Khimulya, E. C. Alley, K. M. Esvelt, and G. M. Church, "Low-N protein engineering with data-efficient deep learning," en, *Nature Methods*, vol. 18, no. 4, pp. 389–396, Apr. 2021. DOI: 10.1038/s41592-021-01100-y.

[9] I. Y. Chen, F. D. Johansson, and D. Sontag, “Why is my classifier discriminatory?” In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, Montréal, Canada: Curran Associates Inc., 2018, pp. 3543–3554.

[10] D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané, *Concrete Problems in AI Safety*, Jul. 2016. DOI: 10.48550/arXiv.1606.06565.

[11] M. V. Staller, A. S. Holehouse, D. Swain-Lenz, R. K. Das, R. V. Pappu, and B. A. Cohen, “A High-Throughput Mutational Scan of an Intrinsically Disordered Acidic Transcriptional Activation Domain,” *Cell Systems*, vol. 6, no. 4, 444–455.e6, Apr. 2018. DOI: 10.1016/j.cels.2018.01.015.

[12] A. Udupa, S. R. Kotha, and M. V. Staller, “Commonly asked questions about transcriptional activation domains,” *Current Opinion in Structural Biology*, vol. 84, p. 102 732, Feb. 2024. DOI: 10.1016/j.sbi.2023.102732.

[13] C. D. Arnold, F. Nemčko, A. R. Woodfin, *et al.*, “A high-throughput method to identify trans-activation domains within transcription factor sequences,” eng, *The EMBO journal*, vol. 37, no. 16, e98896, Aug. 2018. DOI: 10.15252/embj.201798896.

[14] A. Erijman, L. Kozlowski, S. Sohrabi-Jahromi, *et al.*, “A High-Throughput Screen for Transcription Activation Domains Reveals Their Sequence Features and Permits Prediction by Deep Learning,” *Molecular Cell*, vol. 78, no. 5, 890–902.e6, 2020. DOI: https://doi.org/10.1016/j.molcel.2020.04.020.

[15] A. L. Sanborn, B. T. Yeh, J. T. Feigerle, *et al.*, “Simple biochemical features underlie transcriptional activation domain diversity and dynamic, fuzzy binding to Mediator,” *eLife*, vol. 10, e68068, Apr. 2021. DOI: 10.7554/eLife.68068.

[16] N. F. C. Hummel, K. Markel, J. Stefani, M. V. Staller, and P. M. Shih, “Systematic identification of transcriptional activation domains from non-transcription factor proteins in plants and yeast,” English, *Cell Systems*, vol. 15, no. 7, 662–672.e4, Jul. 2024. DOI: 10.1016/j.cels.2024.05.007.

[17] N. Morffy, L. Van den Broeck, C. Miller, *et al.*, “Identification of plant transcriptional activation domains,” en, *Nature*, pp. 1–8, Jul. 2024. DOI: 10.1038/s41586-024-07707-3.

[18] J. Ma and M. Ptashne, “A new class of yeast transcriptional activators,” eng, *Cell*, vol. 51, no. 1, pp. 113–119, Oct. 1987. DOI: 10.1016/0092-8674(87)90015-8.

[19] M. V. Staller, E. Ramirez, S. R. Kotha, A. S. Holehouse, R. V. Pappu, and B. A. Cohen, “Directed mutational scanning reveals a balance between acidic and hydrophobic residues in strong human activation domains,” en, *Cell Systems*, vol. 13, no. 4, 334–345.e5, Apr. 2022. DOI: 10.1016/j.cels.2022.01.002.

[20] S. R. Kotha and M. V. Staller, “Clusters of acidic and hydrophobic residues can predict acidic transcriptional activation domains from protein sequence,” *Genetics*, vol. 225, no. 2, iyad131, Oct. 2023. DOI: 10.1093/genetics/iyad131.

[21] A. H. Aly, A. Debbab, and P. Proksch, “Fifty years of drug discovery from fungi,” en, *Fungal Diversity*, vol. 50, no. 1, pp. 3–19, Sep. 2011. DOI: 10.1007/s13225-011-0116-y.

[22] S. Mahatma, L. Van den Broeck, N. Morffy, M. V. Staller, L. C. Strader, and R. Sozzani, “Prediction and functional characterization of transcriptional activation domains,” in *2023 57th Annual Conference on Information Sciences and Systems (CISS)*, Mar. 2023, pp. 1–6. DOI: 10.1109/CISS56502.2023.10089768.

[23] V. Frappier and A. E. Keating, “Data-driven computational protein design,” *Current Opinion in Structural Biology*, vol. 69, pp. 63–69, Aug. 2021. DOI: 10.1016/j.sbi.2021.03.009.

[24] D. Erhan, A. Courville, Y. Bengio, and P. Vincent, “Why does unsupervised pre-training help deep learning?” In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, Y. W. Teh and M. Titterington, Eds., vol. 9, Chia Laguna Resort, Sardinia, Italy: PMLR, 13–15 May 2010, pp. 201–208.

[25] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” en, *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015. DOI: 10.1038/nature14539.

[26] R. Rao, J. Meier, T. Sercu, S. Ovchinnikov, and A. Rives, *Transformer protein language models are unsupervised structure learners*, en, Dec. 2020. DOI: 10.1101/2020.12.15.422761.

[27] A. Rives, J. Meier, T. Sercu, *et al.*, "Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences," *Proceedings of the National Academy of Sciences*, vol. 118, no. 15, e2016239118, 2021. DOI: 10.1073/pnas.2016239118. eprint: https://www.pnas.org/doi/pdf/10.1073/pnas.2016239118.

[28] Z. Lin, H. Akin, R. Rao, *et al.*, "Evolutionary-scale prediction of atomic level protein structure with a language model," en, Synthetic Biology, preprint, Jul. 2022. DOI: 10.1101/2022.07.20.500902.

[29] B. Hie, B. D. Bryson, and B. Berger, "Leveraging uncertainty in machine learning accelerates biological discovery and design," *Cell systems*, vol. 11, no. 5, pp. 461–477, 2020.

[30] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles," in *Advances in Neural Information Processing Systems*, vol. 30, Curran Associates, Inc., 2017.

[31] B. L. Hie and K. K. Yang, "Adaptive machine learning for protein engineering," *Current opinion in structural biology*, vol. 72, pp. 145–152, 2022.

[32] M. Huot, D. Wang, J. Liu, and E. I. Shakhnovich, "Predicting high-fitness viral protein variants with bayesian active learning and biophysics," *Proceedings of the National Academy of Sciences*, vol. 122, no. 24, e2503742122, 2025. DOI: 10.1073/pnas.2503742122. eprint: https://www.pnas.org/doi/pdf/10.1073/pnas.2503742122.

[33] J. Yang, R. G. Lal, J. C. Bowden, *et al.*, "Active learning-assisted directed evolution," en, *Nature Communications*, vol. 16, no. 1, p. 714, Jan. 2025. DOI: 10.1038/s41467-025-55987-8.

[34] D. H. Bryant, A. Bashir, S. Sinai, *et al.*, "Deep diversification of an AAV capsid protein by machine learning," en, *Nature Biotechnology*, vol. 39, no. 6, pp. 691–696, Jun. 2021. DOI: 10.1038/s41587-020-00793-4.

[35] K. Jiang, Z. Yan, M. D. Bernardo, *et al.*, "Rapid in silico directed evolution by a protein language model with evolvepro," *Science*, vol. 387, no. 6732, eadr6006, 2025. DOI: 10.1126/science.adr6006. eprint: https://www.science.org/doi/pdf/10.1126/science.adr6006.

[36] S. R. Ahrendt, S. J. Mondo, S. Haridas, and I. V. Grigoriev, "MycoCosm, the JGI's Fungal Genome Portal for Comparative Genomic and Multiomics Data Analyses," en, in *Microbial Environmental Genomics (MEG)*, F. Martin and S. Uroz, Eds., New York, NY: Springer US, 2023, pp. 271–291, ISBN: 978-1-07-162871-3. DOI: 10.1007/978-1-0716-2871-3_14.

[37] W. Li and A. Godzik, "Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences," *Bioinformatics*, vol. 22, no. 13, pp. 1658–1659, Jul. 2006. DOI: 10.1093/bioinformatics/btl158.

[38] A. M. Bolger, M. Lohse, and B. Usadel, "Trimmomatic: A flexible trimmer for Illumina sequence data," *Bioinformatics*, vol. 30, no. 15, pp. 2114–2120, Aug. 2014. DOI: 10.1093/bioinformatics/btu170.

[39] A. P. Masella, A. K. Bartram, J. M. Truszkowski, D. G. Brown, and J. D. Neufeld, "PANDAseq: Paired-end assembler for illumina sequences," *BMC Bioinformatics*, vol. 13, no. 1, p. 31, Feb. 2012. DOI: 10.1186/1471-2105-13-31.

5 Appendix

5.1 Uncertainty quantification

We approximate the predictive distribution by averaging over M neural networks, each of which outputs a mean $\mu_m(x)$ and assuming heteroscedastic variance $\sigma_m^2(x)$:

$$p(y \mid x, \mathcal{D}) \approx \frac{1}{M} \sum_{m=1}^{M} \mathcal{N}\big(y;\, \mu_m(x),\, \sigma_m^2(x)\big), \tag{1}$$

$$\hat{\mu}(x) := \frac{1}{M} \sum_{m=1}^{M} \mu_m(x). \tag{2}$$

Aleatoric uncertainty captures the irreducible noise in the data, such as measurement error or inherent biological variability, and corresponds to the average predicted variance. Epistemic uncertainty captures uncertainty due to limited knowledge or data scarcity, reflecting the disagreement between ensemble members and can be reduced with more informative training data. Together they form the total predictive variance:

$$\widehat{\mathrm{Var}}[Y \mid x, \mathcal{D}] = \underbrace{\frac{1}{M} \sum_{m=1}^{M} \sigma_m^2(x)}_{\text{aleatoric uncertainty}} + \underbrace{\frac{1}{M} \sum_{m=1}^{M} \big(\mu_m(x) - \hat{\mu}(x)\big)^2}_{\text{epistemic uncertainty}}. \tag{3}$$

$$\widehat{\mathrm{Var}}[Y \mid x, \mathcal{D}] = \mathbb{E}_m\big[\sigma_m^2(x)\big] + \mathrm{Var}_m[\mu_m(x)]. \tag{4}$$

5.2 Data, code, and sequencing availability

All code related to this study is publicly available on GitHub (github.com/shih-lab/ADhunter). The base model of ADhunter using binary encodings is publicly available on Github (github.com/staller-lab/adhunter). Raw sequencing reads are publicly available through the NCBI SRA Database under BioProject accession PRJNA1183837.

5.3 Model implementation

ADhunter consists of a convolutional layer, a series of residual blocks, a pooling layer, and a fully-connected output layer. Each residual block contains two convolutional layers with batch normalization and ReLU activation. ADhunter optimizes mean squared error loss using Adam and outputs evaluation metrics for the root mean squared error, Pearson correlation, and Spearman correlation. To prevent overfitting, we added early stopping if the validation loss did not improve after 5 epochs. Labeled datasets were preprocessed by removing duplicate entries and z-score normalizing the activity values where 80% was used for training, 10% for validation, and 10% for testing. Since AD activity is continuous, the value was binarized with respect to the median activity to enable stratified shuffling of the data. Protein sequences and activation measurements from Hummel et al. were used for the initial training dataset. ESM embeddings were obtained as described on the ESM GitHub repository (github.com/facebookresearch/esm).

5.4 Sequencing analysis

Fungal genomes were obtained from MycoCosm via the Joint Genome Institute. Protein sequences were deduplicated then clustered by sequence identity using CD-HIT [37]. The test tiles from the MycoCosm collection were codon optimized for expression in *S. cerevisiae* whereas the original codons were used from dataset harmonization and control tiles.

Raw sequencing reads were demultiplexed using bcl2fastq v2.19.0, trimmed using trimmomatic [38] v0.39, and assembled using PANDAseq [39] v2.11. Only reads with a perfect match to a tile in the library were retained. Tile and barcode sequences were extracted using a custom regular expression matching conserved regions. For each set of eight sorted samples, reads were normalized by the total number of reads in each bin. For each tile, counts were normalized across the eight sets to calculate a relative abundance. Activity scores were calculated by taking the inner product between relative abundances and the median fluorescent value of each bin, resulting in a weighted average. Tiles with less than 50 reads were discarded. We used labtools v0.0.3 to quantify activity for each AD tile (github.com/staller-lab/labtools). The activity score for a given tile was aggregated by taking the mean activity across all DNA barcodes.

5.5 Active learning dataset and harmonization

In addition to the tiles from non-model fungi, we added control tiles and tiles that harmonize the initial and active learning datasets. For dataset harmonization, 437 tiles from the held-out test dataset were selected with the smallest difference across the range of experimental and predicted AD activity. As controls, the same acquisition function as the evolutionary sampling was used to select a subset of 464 tiles with the highest uncertainty that span the range of empirical activity from the yeast TF dataset. For each tested tile there was a median of 12 DNA barcodes. We recovered 379 of 464 control tiles (81.7%), 404 of 437 harmonization tiles (92.4%) and 7,681 of 8,935 non-model fungal tiles (86.0%). As expected, the control tiles spanned the range of empirical activity with high correlation to the activity reported in Sanborn et al. (Pearson r = 0.695; RMSE = 0.841). Harmonization tiles in the active learning dataset and the initial dataset had high correlation (Pearson r = 0.934; RMSE = 0.274).

The FACS saturates at high and low signal. We determined the linear range of the assay by maximizing the Pearson correlation between harmonization tiles in the initial and new dataset then validating the thresholds on the Pearson correlation of the control tiles. To harmonize the two datasets, we used a linear fit to map the activity values onto the same distribution. When evaluating the performance improvement of active learning, ADhunter was retrained on the harmonized dataset, and a random held-out test dataset was used for evaluation. ADhunter was retrained on the entire harmonized dataset when performing inference on the MycoCosm collection.

5.6 Supplementary tables

Secondary Structure Included	Pearson r	RMSE
True	0.261	0.336
False	0.338	0.329

Table 2: Evaluation of PADDLE performance. PADDLE was evaluated on the initial dataset and achieved poor prediction performance. The model combines one-hot encodings and secondary structure predictions. However, the secondary structure predictions result in worse performance likely due to increased complexity of input features.

Encoding	Pearson r	RMSE
BLOSUM	0.730	0.692
NLF	0.698	0.722
One-hot	0.742	0.672
ESM1_650M_1	0.740	0.669
ESM1_650M_2	0.732	0.678
ESM1_650M_3	0.734	0.678
ESM1_650M_4	0.739	0.674
ESM1_650M_5	0.728	0.685
ESM2_8M	0.744	0.664
ESM2_35M	0.744	0.664
ESM2_150M	0.725	0.686
ESM2_650M	0.706	0.708
ESM2_3B	0.744	0.668

Table 3: Evaluation of protein sequence encodings. One-hot encodings outperform other simple protein representations on a held-out test dataset. Neural encodings from pretrained protein language models slightly outperform one-hot encodings on a held-out test dataset. Metrics were averaged across 10 random seeds.

Model	Pearson r	RMSE
TADA	0.538	0.995
ADhunter	0.775	0.631

Table 4: Evaluation of the state-of-the-art model relative to ADhunter on the initial dataset. ADhunter outperforms the state-of-the-art AD predictor, TADA, when trained on the initial dataset and evaluated on a held-out test dataset. PADDLE was not included since the untrained model is not available for fair comparison.

Model variant	Accuracy	F1 score
Conv1D-Dropout-Conv1D-Dropout-Attention-BiLSTM-BiLSTM	0.845	0.845
Conv1D-Dropout-Conv1D-Dropout-Attention-BiLSTM	0.845	0.845
Conv1D-Dropout-Conv1D-Dropout-Attention-MaxPooling1D	0.839	0.839
Conv1D-Dropout-Conv1D-Dropout-MaxPooling1D	0.852	0.852
Conv1D-Dropout-MaxPooling1D	0.849	0.849
Conv1D-Dropout-Conv1D-Dropout-BiLSTM-BiLSTM	0.850	0.850
Conv1D-Conv1D-Attention-BiLSTM-BiLSTM	0.846	0.846
Conv1D-Conv1D-BiLSTM-BiLSTM	0.852	0.852
Conv1D-Conv1D-MaxPooling1D	0.866	0.866
Conv1D-MaxPooling1D	0.860	0.860
Attention-MaxPooling1D	0.839	0.838

Table 5: TADA ablation study. We evaluated the contribution of each architectural component of TADA by incrementally removing or replacing key elements then training on the initial dataset and evaluating on a held-out test dataset. Metrics were averaged across three random seeds.

Model	Pearson r	RMSE
TADA	0.381	1.021
ADhunter_simple	0.416	0.884
ADhunter_neural	0.476	0.856
ADhunter	0.512	0.826

Table 6: Evaluation of model generalizability. Using spectral clustering analysis, the ensemble model of ADhunter with neural encodings outperforms single models of ADhunter with simple (one-hot) or neural (ESM) encoding at generalizing on the initial dataset. Metrics were averaged across three random seeds.