Conservation of function without conservation of amino acid sequence in intrinsically disordered transcriptional activation domains

Claire LeBlanc[1,2], Jordan Stefani[1,2], Melvin Soriano[1,2], Angelica Lam[1,2,#], Marissa A. Zintel[1], Sanjana R. Kotha[1,2], Emily Chase[1,2], Giovani Pimentel-Solorio[1,2,##], Aditya Vunnum[1], Katherine Flug[1], Aaron Fultineer[3], Niklas Hummel[4], Max V. Staller[1,2,5,*]

Affiliations:

[1] Department of Molecular and Cell Biology, University of California Berkeley, Berkeley, 94720

[2] Center for Computational Biology, University of California Berkeley, Berkeley, 94720

[3] Department of Physics, University of California Berkeley, Berkeley, 94720

[4] Department of Biology, Technische Universität Darmstadt, Darmstadt, Germany

[5] Chan Zuckerberg Biohub–San Francisco, San Francisco, CA 94158

[#] Present address: University of California San Francisco, San Francisco, CA 94158

[##] Present address: University of California Davis, Davis, CA

*Corresponding author: 16 Barker Hall, Berkeley, CA 94720, USA. mstaller@berkeley.edu

Abstract:

Protein function is canonically believed to be more conserved than amino acid sequence, but this idea is only well supported in folded domains, where highly diverged sequences can fold into equivalent 3D structures. In contrast, intrinsically disordered protein regions (IDRs) do not fold into a stable 3D structure, thus it remains unknown when and how function is conserved for IDRs that experience rapid amino acid sequence divergence. As a model system for studying the evolution of IDRs, we examined transcriptional activation domains, the regions of transcription factors that bind to coactivator complexes. We systematically identified activation domains on 502 orthologs of the transcriptional activator Gcn4 spanning 600 MY of fungal evolution. We find that the central activation domain shows strong conservation of function without conservation of sequence. This conservation of function without conservation of sequence is facilitated by evolutionary turnover (gain and loss) of key acidic and aromatic residues, the positions most important for function. This high sequence flexibility of functional orthologs mirrors the physical flexibility of the activation domain coactivator interaction interface, suggesting that physical flexibility enables evolutionary plasticity. We propose that turnover of short functional elements, sometimes individual amino acids, is a general mechanism for conservation of function without conservation of sequence during IDR evolution.

Key words

Intrinsically disordered proteins; transcription; transcription factor; activation domains; evolution; evolutionary turnover; high-throughput assays

Introduction:

The evolution of eukaryotic transcription factor (TF) function contains a paradox: TF protein sequences diverge quickly but maintain function over long evolutionary distances. For example, the master regulator of eye development in mice, Pax6, induces ectopic eyes in fly, and fly Pax6 (*eyeless*) creates ectopic eye structures in frogs and mice[1–3]. While the DNA-binding domains (DBD) are 96% identical, eye induction requires the intrinsically disordered regions (IDRs), which are only 35.5% identical. These IDRs must share a conserved function despite substantial sequence divergence. In contrast, small sequence changes in TFs can lead to large functional changes that drive the evolution of new traits[4,5]. Some TFs maintain function despite low conservation of sequence[6], while other TFs drive evolutionary innovations with limited sequence changes.

For folded domains, function is more conserved than sequence because highly diverged sequences can fold into the same 3D structure and maintain function[7–9]. Here, we seek an analogous framework for understanding the evolution of and functional constraint on IDRs. Small-scale studies have found examples of diverged IDRs that conserve function[10–12] and diverged IDRs that do not conserve function[13,14]. Transcriptional activation domains provide an excellent model system for studying IDR evolution because they are one of the oldest classes of functional IDRs[15], they are required for TF function, and their activity can be measured in high throughput[16]. Our goal is to identify molecular mechanisms by which TF IDR function can be conserved in the face of rapid sequence divergence.

We hypothesized that TF IDRs can maintain function despite sequence divergence through evolutionary turnover of functional elements. Evolutionary turnover is repeated gain and loss of functional elements. Mutations create new functional elements and negative selection maintains a minimum number of elements, allowing ancestral elements to be lost. As a result, on long timescales, neutral drift will give the appearance of functional elements moving around the sequence. For TFs, it is unclear if the functional elements will be entire activation domains, short linear interaction motifs (SLiMs)[17], or individual amino acids. Here, we aim to identify the functional units and test the hypothesis that evolutionary turnover can explain conservation of function without conservation of sequence.

Evolutionary studies of acidic activation domains in yeast benefit from high-throughput data that define sequence features controlling their function[16,18–23]. These data have trained neural network models for predicting activation domains from protein sequence[18,21,23–26]. Our acidic exposure model further provides a biophysical mechanism for the observed features: aromatic and leucine residues make key contacts with hydrophobic surfaces of coactivator complexes, but these residues can also interact with each other and drive collapse into an inactive state[16,27–30]. The acidic residues repel each other, expand the activation domain, and promote exposure of the hydrophobic residues. In many cases, the aromatic and leucine residues are arranged into short linear motifs. Large-scale mutagenesis showed the acidic exposure model applies to hundreds of human activation domains[31].

We investigated the molecular mechanisms by which full-length TFs can maintain activator function over long evolutionary distances despite divergence of their amino acid sequences. As a model system, we used 502 diverse orthologs of Gcn4, a nutrient stress TF, and screened for activation domains with a high-throughput functional assay in *Saccharomyces cerevisiae*[16]. All orthologs contain at least one 40 AA region that functions as an activation domain, and we see widespread conservation of function without

conservation of sequence. We demonstrate evolutionary turnover of entire activation domains and turnover of key residues within an activation domain. The N-terminal activation domains are repeatedly gained and lost. In contrast, the central activation domain is functionally conserved because of turnover of key acidic and hydrophobic residues. This work illustrates how functional screening can unravel the complex evolution of activation domains and IDRs.

Results:

Characterization of a tiling-library of Gcn4 orthologs

To study the evolutionary dynamics of activator function, we sought to experimentally map activation domains across a diverse collection of orthologous TFs. We and others have shown that protein fusion libraries, designed to tile across protein sequences with short, 30-60 amino acid peptides, can faithfully measure activation domain activity[16,18,19,21,22,32]. Furthermore, because activation domain function in yeast is a reliable measure of endogenous function in humans[33], viruses[34], *Drosophila*[35,36], plants[23,37,38], and other yeast species[39], we reasoned that the activity of fungal orthologues in our assay would serve as a reliable measure of activity in their native context. In all subsequent analysis, we assume that tile activity measured in *S. cerevisiae* is a good proxy for TF function in their native species.

As a null hypothesis, we assumed the TF function is conserved and that the observed diversity of sequence is the result of neutral drift. Absent strong evidence to the contrary, neutral drift is a strong null hypothesis[40]. Mutation processes introduce changes, and selection acts at the level of the full protein. Purifying (negative) selection will tolerate all changes that do not reduce function below a minimum level. The neutral space for IDRs is potentially much larger than that of folded proteins because there are no structural constraints. Supporting this assumption, we found evidence for weak negative selection on the full-length TF using a high-quality set of thirty-six true Gcn4 homologs from the yeast gene order browser (**Figure S1E**)[41]. It follows that most of the sequence differences we see in extant species are neutral. We aim to find the (potentially rare or diffuse) sequence features that are functional and conserved.

We chose a diverse set of orthologous Gcn4 protein sequences for functional characterization in *S. cerevisiae*. We found 502 unique Gcn4 ortholog sequences from 129 genomes that span the Ascomycota, the largest phylum of Fungi, representing >600 million years of evolution[42] (**Figure S1, S2**). While the Gcn4 orthologs vary in length (**Figure 1A**), 500 have the DBD at the C-terminus, and the distance between the WxxLF motif and the DBD is very consistent (**Figure 1B**).

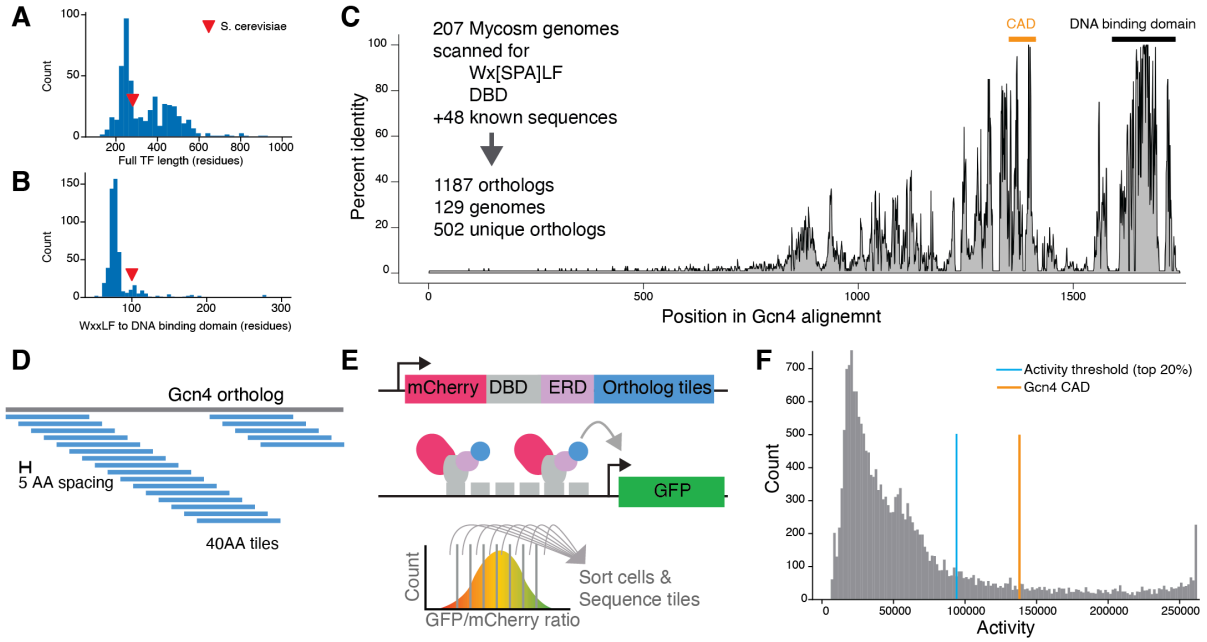


Figure 1: Screening fragments of Gcn4 orthologs for activation domain activity in *S. cerevisiae*.

A) Gcn4 ortholog lengths. Red arrow, *S. cerevisiae*. B) The distance between the WxxLF motif and the start of the DBD is conserved. C) The MSA of 500 orthologs shows the DBD binding domain is highly conserved, and the Central Activation Domain around the WxxLF motif is moderately conserved. D) The tiling strategy for oligo design and the high-throughput activation domain assay. E) The high-throughput assay for measuring activation domain function uses a synthetic TF with mCherry for quantification of abundance, the Zif268 DNA binding domain (DBD), an estrogen response domain (ERD) for inducible activation, and a C-terminally fused tile. Tile activity was calculated based on barcode abundance in eight equally sized bins of a FACS sorting experiment. Bins were set based on GFP/mCherry ratios. F) The distribution of measured tile activities with our activity threshold (top 20%). *S. cerevisiae* Gcn4 CAD activity is shown in orange.

The Gcn4 multiple sequence alignment (MSA) typifies eukaryotic TF evolution, with a highly conserved DBD and lower conservation in the rest of the protein (**Figure 1C**). The central activation domain (CAD) shows intermediate levels of conservation, driven in part by the WxxLF motif (**Figure 2B, S3**). Sequence divergence is driven by insertions: 54% of columns in the MSA contain fewer than 1% of sequences (**Figure 1C, S4**). Distant pairs of sequences do not align outside of the DBD.

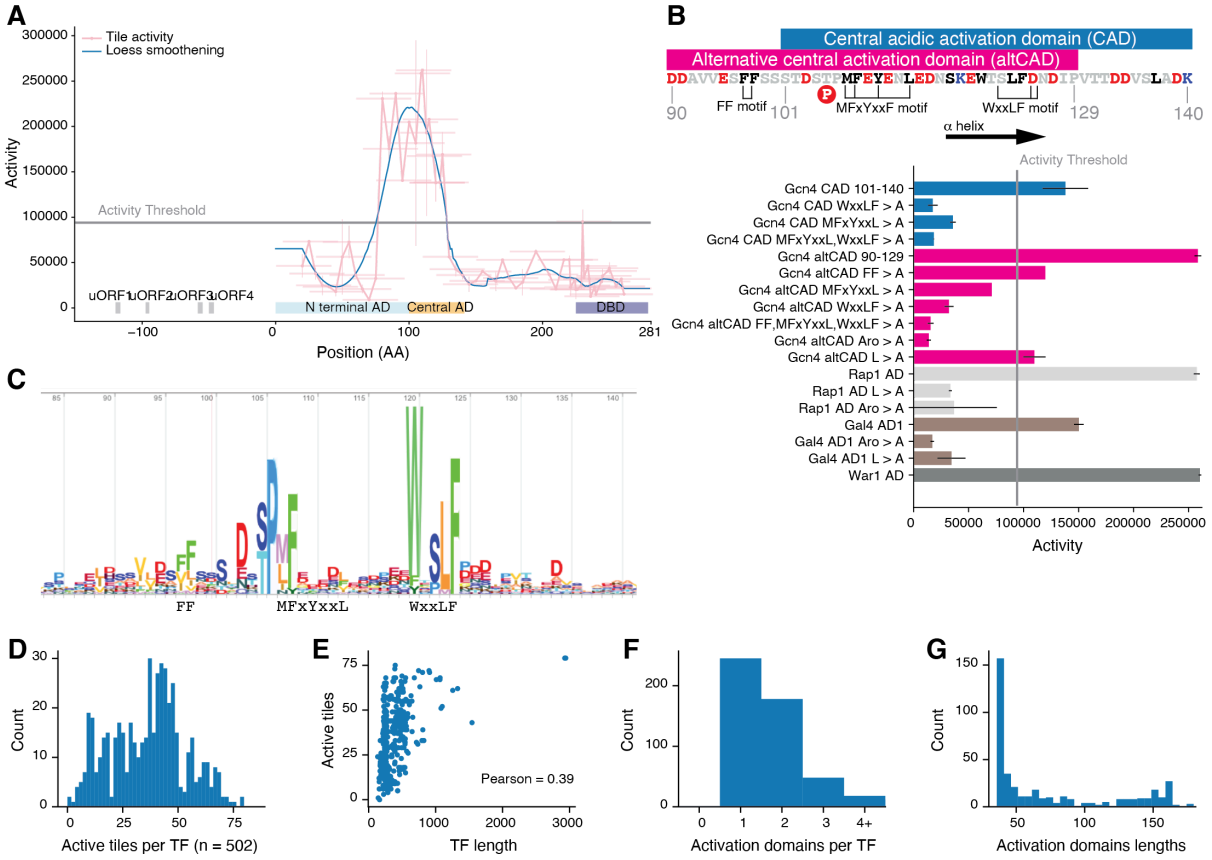


Figure 2: In the *S. cerevisiae* central activation domain, residues that are critical for activity are poorly conserved.

A) Schematic of *S. cerevisiae* Gcn4 with the upstream open reading frames (uORFs) that regulate translation, the NAD and the CAD. Individual measured tiles are indicated as pink lines with a pink point at the center, and the standard deviation of the two replicates is shown vertically. We imputed activity at each position with a Loess smoothing (blue). **B**) Schematic of the CAD and altCAD (most active tile) with key motifs, α-helix, and phosphosites indicated. Mutating motifs, aromatic residues, or leucine residues reduced activity in all cases. **C**) The sequence logo from the 4th iteration of a search for Gcn4 orthologs in fungal genomes with HMMER. This independent analysis confirmed the WxxLF motif is more conserved than the FF and MFxYxxL motifs. **D**) The number of active tiles found on each full-length TF (tiles that map to multiple orthologs can count multiple times in this analysis). **E**) There is a weak correlation between TF length and the number of active tiles. **F-G**) Combining overlapping active tiles shows that most TFs have 2 or more activation domains with a wide distribution of lengths.

High-throughput measurement of orthologs for activation domain function

To study the evolution of TF function, we measured the activation domain activity of all the orthologs. For each of the 502 Gcn4 orthologs, we tiled across the full-length protein with 40 AA tiles spaced every 5 AA, and measured activities of all tiles in *S. cerevisiae* using our established high-throughput assay[16] (**Figure 1D, 1E**). We recovered 18947 of 20731 designed tiles (91.4%), and these data were of high quality (Methods, **Figure S5, S6**). The tiles had a range of activities (**Figure 1F**), and mutations in control activation domains behaved as expected (**Figure 2A, 2B, S7**). As a threshold for highly-active tiles, we used the

top 20% of sequences, but other thresholds led to similar results (Methods, **Figure S8**). Many more tiles are active than datasets that naively tile all TFs in a proteome, as we would expect if most Gcn4 orthologs are activators. Due to the divergence of the orthologs, the sequences of the active tiles are very diverse, allowing us to study sequence-to-function relationships controlling activation domain function. To our knowledge, this dataset is the largest functional study of TF evolution to date.

Activator function is conserved across the Gcn4 orthologs

All the Gcn4 orthologs had at least one tile that functioned as an activation domain in our assay, indicating that activator function is conserved across 600 million years of evolution (**Figure 2D,** Supplemental note 1). *A priori*, it was not a given that all the Gcn4 orthologs would be activators, because on long evolutionary timescales, a family of TFs that share a conserved DBD will include both activators and repressors [23,31,32,38]. Gcn4 activator function is highly conserved despite divergence of the sequence.

The central acidic activation domain shows strong functional conservation.

Our finding that all the orthologs are activators combined with the sequence divergence in the MSA indicates there is conservation of function without conservation of primary amino acid sequence. We examined three hypotheses for this conservation of function without conservation of sequence: 1) turnover of entire activation domains, 2) turnover of motifs within activation domains, and 3) turnover of key residues within activation domains. We found turnover of entire N-terminal activation domains and turnover of key residues within the central activation domain.

The central activation domain is functionally conserved across the orthologs. An advantage of our tiling strategy is the ability to infer the activity of each position in each full-length protein (**Figure 3,** Methods). We found that all orthologs had high activity in the central region (Supplemental note 1). The peak of activity is ten AA residues upstream of the WxxLF motif (**Figure 3, inset**). Aligning on the WxxLF motif or the DBD led to similar results (**Figure S9-S12)**. Projecting the activity heatmap onto the local species tree or gene tree illustrates how the central activation domain can drift side-to-side but stays near the WxxLF motif(**Figure S13, S14**). Intriguingly, the integral of activity across each ortholog was highly consistent, suggesting conservation of total activity (**Figure S15C**).

The second major result is that N-terminal activation domains come and go, providing evidence for turnover of entire activation domains (**Figure 3**). After combining overlapping active tiles, the majority of orthologs have more than one activation domain (**Figure 2F**). Projecting activity onto the MSA or sorting the heatmap by activity at the WxxLF motif emphasizes how the N-terminal activation domains come and go (**Figure S11, S12**). Using our stringent threshold for activity (top 20%), thirteen orthologs lost activity at the WxxLF motif, but all of these have gained additional upstream activation domains. The N-terminal activation domains show intermediate conservation in the MSA (**Figure S15**) and their sequences are very diverse, ruling out the possibility that one ancestral activation domain is recurrently lost (**Figure S16**). Together, these data demonstrate turnover of entire activation domains.

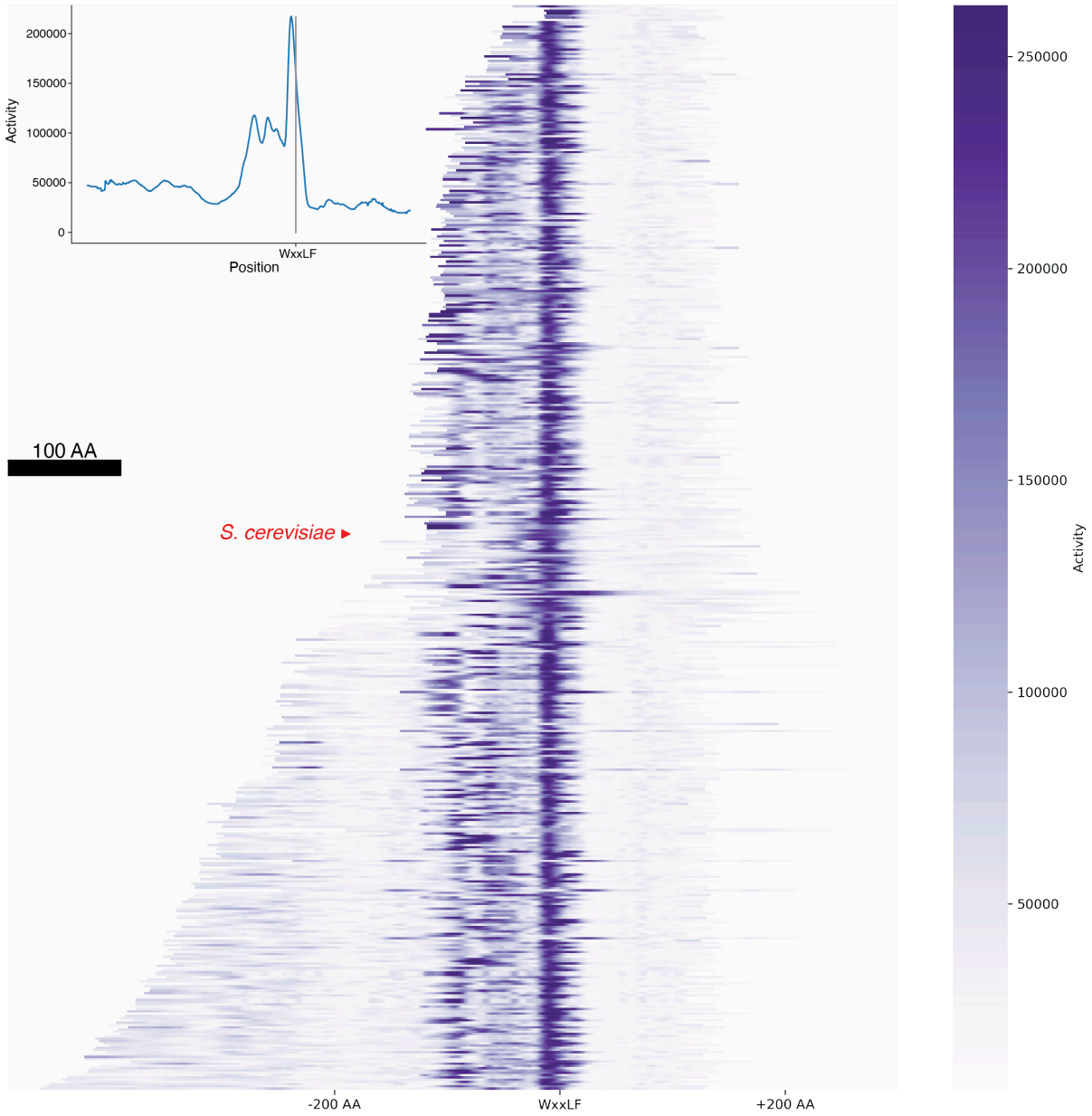


Figure 3: The central acidic activation domain of Gcn4 is functionally conserved. We used the tile activity data to impute the activity of each residue in all the orthologs. These activities are visualized as a heatmap, with color representing imputed activity. The 476 shortest orthologs are sorted by length and aligned on the WxxLF motif. Inset, vertically averaging the heatmap. Activity is consistently high around the WxxLF motif, indicating deep functional conservation. Upstream, N-terminal activity is more salt and pepper, indicating recurrent gain and loss of activation domains. Aligning on the DBD or including the longer sequences yields similar results (**Figure S9-11**). Red arrow, *S. cerevisiae*. Black scale bar, 100 AA.

Conservation of function without conservation of sequence in the Central Acidic Activation Domain of Gcn4

The central activation domain region with high-functional conservation shows intermediate conservation in the multiple-sequence alignment (**Figure 1C, S3A**). We conclude that there is conservation of activation domain function without conservation of the

sequence.To understand the sequence features underlying this conservation of function without conservation of sequence, we first describe the amino acid sequence features controlling activity of individual tiles and then apply these lessons to the orthologs.

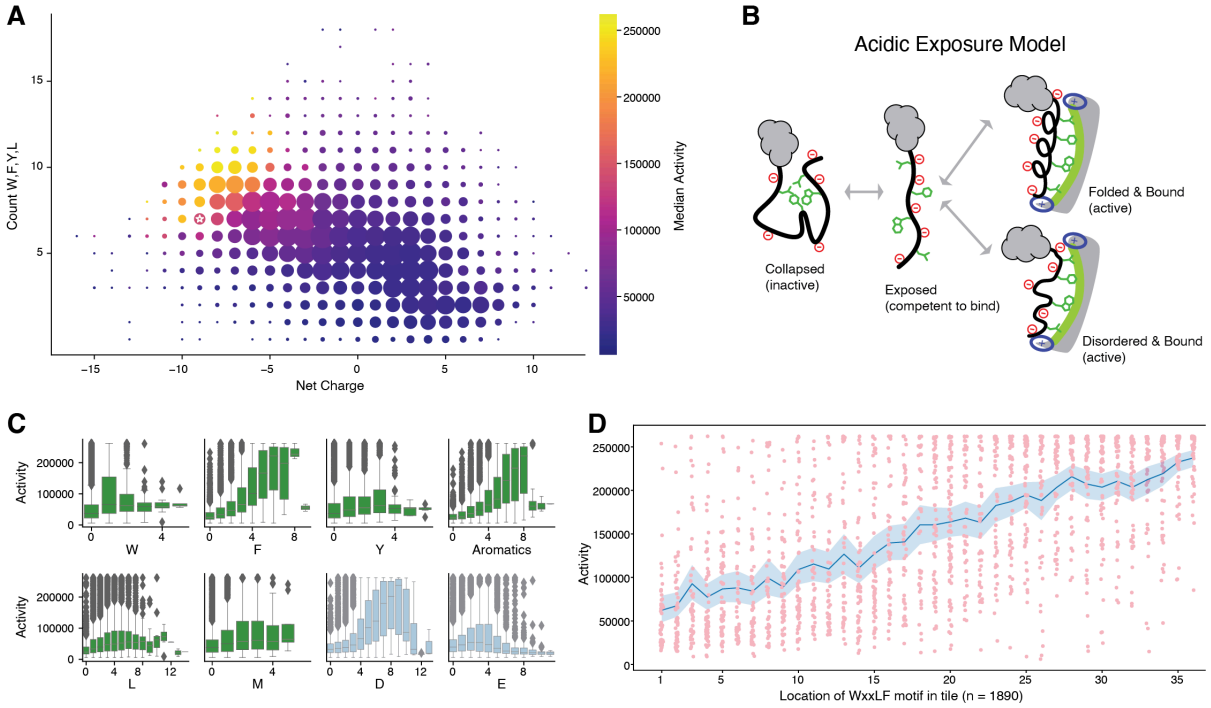


Figure 4: Highly active tiles contain many acidic, aromatic, and leucine residues, supporting the acid exposure model of acidic activation domain function.
A) For each tile, we compute net charge and count the number of WFYL residues. The size of the point indicates the number of tiles with the combination of properties. The color is the median activity of tiles with each combination. White star, *S. cerevisiae* Gcn4. **B**) The acidic exposure model of acidic activation domain function. **C**) Boxplots for the residues that make the largest contributions to activity. **D**) For each tile with the WxxLF motif, activity is plotted against the location of the W. Blue, mean and 95% confidence interval. The location of the motif is correlated with activity.

The sequence features of active tiles support the acidic-exposure model

The Gcn4 ortholog dataset contains all previously observed relationships between sequence and function, but many relationships are stronger and more visible than previously reported (Supplemental Note 1). As predicted by the acidic exposure model, many active tiles contain both acidic residues and WFYL residues (**Figure 4A**, **4B**). These key residues make quantitatively different contributions to activity (**Figure 4C, S17, S18**). Aspartic acid (D) makes stronger contributions to activity than glutamic acid (E), likely because the charge is slower to the backbone and better promotes exposure[43] (**Figure 4C, S18**). In the control activation domains, all published motifs of aromatic and leucine residues made large contributions to activity, but no individual motif was sufficient for full activity (**Figure S7**). These sequence features of active tiles with or without the WxxLF motif are highly similar, suggesting the N-terminal activation domains function similarly to the central activation

domain, as has been shown in *S. cerevisiae*[44] (**Figure 19**). Tiling orthologs reveals sequence rules more efficiently than tiling genomes (**Figure S20**).

Amino acid composition strongly contributes to activation domain function. Ordinary least squares (OLS) regression on single amino acids explains 49.9% of variance in activity (**Table 1**, AUC = 0.9346, PRC = 0.7620, **Table S9**). Regression on dipeptides[21] led to 69 significant parameters that explain 60.2% of the variance in activity (**Table 1**, AUC = 0.9472, PRC = 0.8190). More complex sequence motifs did not improve the regression models: published motifs explained 33.1%, and 40 *de novo* motifs explained 50.5% of the variance in activity (**Table 1**). Combining the *de novo* motifs with single amino acids performed similarly to dipeptides. This result implies that complex motifs capture very little additional information beyond adjacent pairwise amino acid relationships in dipeptides.

Model	Number parameters	Number of statistically significant parameters	Adjusted R^2
Single AAs	20	16	.498
Single AAs - reduced	16		.498
Dipeptides	400	69	.651
Dipeptides - reduced	69		.608
Published Motifs	7	5	.334
de novo motifs	40	27	.502
de novo motifs - reduced	27		.500
de novo motifs + single AAs	60	37	.606
de novo motifs + single AAs	37		.604

Table 1: Ordinary Least Squares regression on tile composition explains a large fraction of the variance in measured activation domain activity

The WxxLF motif requires acidic context and supporting hydrophobic residues.

The absence of clear motifs raises the question of how the arrangement of amino acids, the sequence grammar, controls activation domain function. As an anchor point, we used the WxxLF motif, which makes large contributions to activity in the CAD but not all tiles with this motif are active (**Figure 2B, S8, S20**). We compared tiles with the WxxLF motif that had high or low activity: highly active tiles were more acidic and had more WFYLM residues (**Figure S21C,D**). The first grammar signal we found is that tiles with more evenly intermixed acidic and W,F,Y,L residues are more active, supporting the acidic exposure model (**Figure S21E**). The strongest grammar signal is that tiles with the WxxLF motif near the C-terminus are active (**Figure 4D, S22)**. The additional negative charge of the C-terminus may increase exposure of the motif. Weak C-terminal effects have been seen for aromatic residues[20,37]. This result emphasizes how even a conserved short linear motif requires an acidic context and supporting hydrophobic residues to create an activation domain. Together, our analysis indicates that yeast activation domains are nucleated by a cluster of aromatic residues surrounded by acidic residues and supported by leucine and methionine residues.

The alpha helix from *S. cerevisiae* is dispensable for full activity.

The sequence diversity of strongly active tiles with the WxxLF motif strongly suggests that coupled folding and binding is not necessary for activity. In *S. cerevisiae*, the disordered CAD folds into a short alpha helix upon binding the Gal11/Med15 coactivator[45,46]. Inserting a proline into this helix has little effect on activity[16,45]. The immediate vicinity of the helix in *S.*

cerevisiae has 115 unique sequences: 23 contain 3 prolines (20%), and 3 contain 4 prolines, e.g. G**P**SD**P**WY**P**LF**P**SDTA. Using a 70 residue region, we predicted alpha helix propensity, but only 38/138 (28%) are predicted to form a helix (**Figure S23**, methods [47]). Amphipathic helices are enriched in activation domains[18,32] because they are a convenient way to present hydrophobic residues to a partner, but they are not the only way to create a strong activation domain. CAD function is more conserved than alpha helix formation. This analysis rules out the alpha helix as the relevant functional unit for evolutionary turnover.

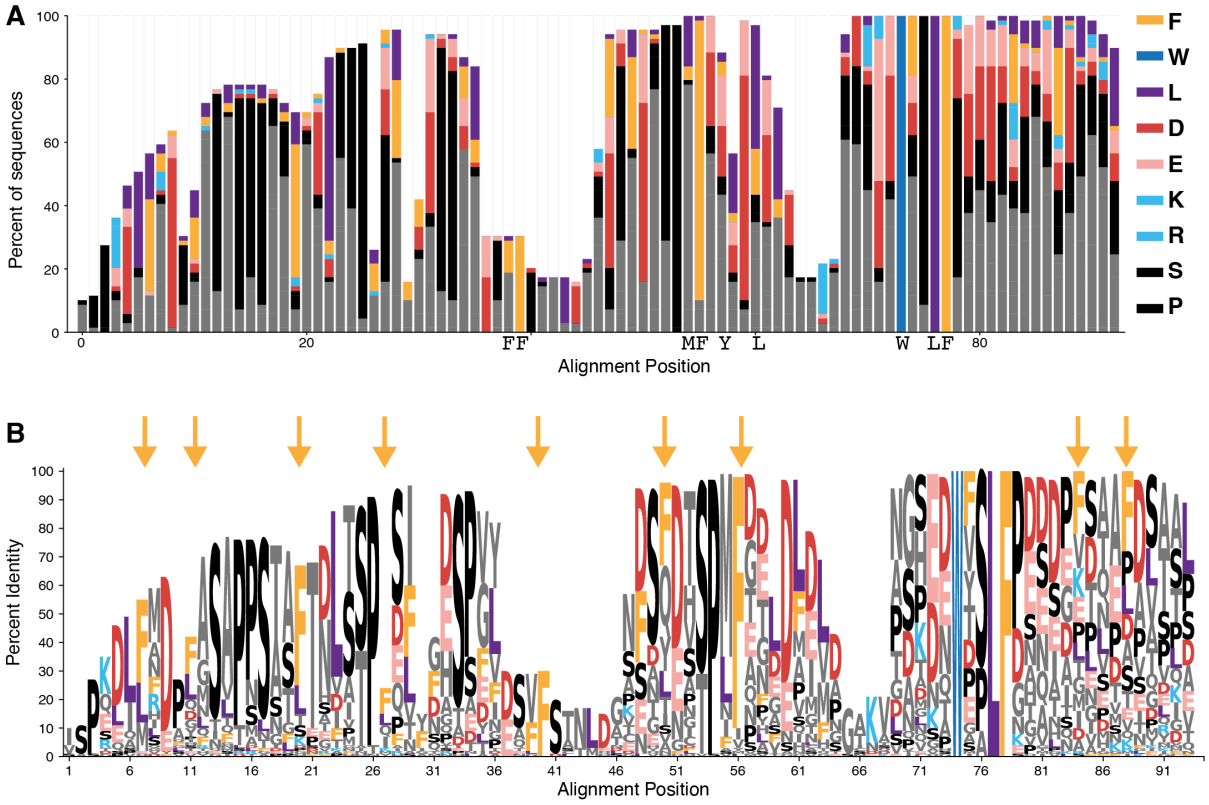


Figure 5: Evolutionary turnover of aromatic and acidic residues explains the conservation of function without conservation of sequence in the central activation domain of Gcn4. **A**) For the 69 most active unique regions around the WxxLF motif, a bar plot showing the relative amino acid frequencies from the MSA. MSA positions with >90% gaps have been removed. The acidic residues, D and E, interchange. **B**) A sequence logo for the MSA. Arrows indicate the 9 positions where F is the most abundant residue. There is some interchange between F and L. Black, SP motifs. See **Figure S25** for the MSA.

Conservation of function without conservation of sequence in the Central Acidic Activation Domain of Gcn4.

The central activation domain region of the Gcn4 orthologs showed strong conservation of function without conservation of sequence. Our two hypotheses for this phenomenon were evolutionary turnover of motifs or evolutionary turnover of key residues.

We found no evidence for turnover of motifs. Each of the published motifs contributed to activity (**Figure 2B**) and was enriched in active tiles (**Figure S21A**), but only the WxxLF

motif was conserved (**Figure 2C, S3**). We did not detect the emergence of new instances of these motifs, so we can reject the motif turnover hypothesis.

The most conserved sequence feature of the CAD region besides the WxxLF motif is an SP motif, which is not typically associated with activation domain function. The full-length orthologs contain up to 4 SP motifs upstream of the WxxLF motif. In *S. cerevisiae*, this SP is a TP (T105), which is phosphorylated to create a phosphodegron that shuts down the Gcn4 program during the recovery from starvation[48,49]. The majority of tiles (10752) contain an SP motif, so it makes little contribution to activity on its own. We believe these motifs are conserved due to regulated degradation. However, it remains possible that multisite phosphorylation can increase activation domain activity[50,51]. For 32 tiles we performed a followup mutagenesis of the SP motifs to test the hypothesis that phosphorylation can control activity (**Figure S24**). These data support the possibility that some of the orthologs utilize phosphorylation to modulate activation domain function.

We observe evolutionary turnover of acidic and F residues within the central activation domain. We focused on a 70 residue region around the WxxLF motif (W-50 : W+19) that contained many active tiles and the peak of inferred activity (**Figure S25**). The top half of sequences contain many acidic residues, but individual acidic positions (D and E) are not well conserved because they interconvert (**Figure 5A,B, S26**). When pooled together, D+E conservation matches or exceeds the conservation level of the aromatic residues. In addition, the F residues that are critical for high activity exhibit evolutionary turnover. There are only 2 positions where an F is present in the majority of sequences, but an additional 7 positions where F is the most common residue. The critical F residues experience evolutionary turnover, giving the appearance of moving around the activation domain.

To this point, all of our analysis has used only the MSA, so we next leveraged the additional information present in the species tree. We tested the hypothesis that the gains of F residues precede the loss of F residues. In most cases, there is too much evolutionary distance between the species to answer this question. However, in the high quality YGOB alignment[41], we see the gain of an F precedes the loss of an ancestral F (**Figure S27**). This example of gains preceding loss bolster then evidence for evolutionary turnover of key residues.

The turnover and conservation patterns we observed in the Gcn4 orthologs generalized to other systems. We reanalyzed a set of orthologs of Pdr1 (**Figure S28**)[18]. For four TFs, we searched for orthologs in the Y1000+ collection and made alignments of their activation domains (**Figure S28**). In these MSAs, aromatic residues were highly conserved and acidic residues interchange at many positions. Some positions also showed interchange between aromatic residues. Other regions showed turnover of aromatic and leucine residues. We propose that evolutionary turnover of key aromatic, leucine, and acidic residues is a general feature of eukaryotic acidic activation domains.

Machine learning insights into activation domain function

The Gcn4 orthologs provide a large, unique dataset to evaluate deep learning models that predict activation domains from amino acid sequence. We compared two first-generation neural networks[18,21] with a second-generation model that we developed[23,24]. All the models can approximate the locations of activation domains in full-length TFs, but the new model, TADA, is substantially more accurate at predicting the activities of individual tiles and identifying activation domain boundaries (**Figure S29**). TADA was intentionally built to ignore sequence grammar by blurring the raw sequence with sliding windows and its high

performance supporting the idea that there is very weak or very little grammar in these orthologs. The machine learning models cannot detect 'missing' grammar, supporting the weak grammar hypothesis. The high accuracy of these models suggests they may be ready to enable evolutionary studies.

We used TADA to predict the contributions of F residues to activity in the central activation domain. The model predicts that all the F residues contribute to activity (**Figure S30**). The contributions of the most conserved F positions are indistinguishable from recently evolved F positions (**Figure S30D**). This analysis further supports evolutionary turnover of key F residues.

Discussion:

By functionally screening protein fragments from a family of orthologous sequences, we demonstrate how activation domains show strong conservation of function without conservation of sequence through turnover of critical acidic and phenylalanine residues. Conservation of function without conservation of sequence was established for full-length TFs, but here we demonstrate how this phenomenon emerges from turnover of entire activation domains and turnover of key residues within activation domains. Our results emphasize how IDR function can be highly conserved and constrained yet invisible in traditional comparative genomics.

The observed turnover of critical residues supports our acidic exposure model for activation domain function and explains why it is so difficult to identify motifs in activation domains. Multiple screens for activation domains have found only one recurrent motif, LxxLL motif with an acidic context, which can be important for binding the Kix domain[17,18,21,31,52,53]. These screens have also shown that the 9aaTAD is not enriched in active sequences[54]. We argue that activation domains are nucleated by Clusters of W and F residues surrounded by acidic residues and boosted by Y, L, and M residues. Under this weak molecular grammar, individual residues are easily replaced, facilitating turnover. The WxxLF motif is one solution among many. When only a few sequences are examined, clusters look like motifs. Each TF family has a different conserved cluster of hydrophobic residues that represents a very good solution to binding the preferred coactivator. Each TF family will appear to have a conserved, essential motif, but convergent evolution of motifs is rare (Supplemental note 1).

We propose that the physical flexibility of the protein interaction interface between Gcn4 and Med15 allows for evolutionary plasticity. The Gcn4 CAD undergoes coupled folding and binding with the Med15 activation domain binding domains, but this interaction is a physically flexible, fuzzy interaction[44,45,55]. The short helix presents the WxxLF motif in many orientations to a shallow hydrophobic canyon on Med15. Molecular dynamics simulations suggest that these orientations interconvert[46]. This binding interaction imposes few structural constraints on the Gcn4 CAD.

The turnover of hydrophobic residues is possible because of this physical flexibility of the Gcn4-Med15 protein-protein interaction. The weak structural constraint of this interaction enables evolutionary plasticity. Binding one sequence in multiple orientations is a step towards binding diverse orthologs, which in turn is a step towards binding to many activation domains[18,56,57]. This flexibility likely requires at least one disordered partner[58]. Coactivators that impose weak structural constraints on activation domains can become engines for evolutionary diversification of activation domains through neutral drift, creating an enormous sequence reservoir for later selection. Although we favor the hypothesis that the observed sequence divergence in Gcn4 orthologs is neutral, stabilizing selection, it remains possible

that there is selection to diversify. Acidic activation domains are highly evolutionarily successful, representing more than half of all known examples[27]. Our observation that acidic activation domains can easily diversify without compromising function suggests they are highly evolvable. This evolvability creates a diverse sequence reservoir that allows for rapid selection on standing variation. We speculate this evolvability allowed for acidic activation domains to bind new coactivators as they emerged with multicellularity[59].

Activation domain evolution exemplifies how protein-protein interactions mediated by IDRs can drive evolutionary plasticity and sequence diversity. Another example of an IDR engaged in flexible PPIs enabling evolutionary plasticity is the human TRIM5 antiviral caging system, wherein short disordered loops make multivalent contacts with the viral capsid[60]. Physically flexible binding and avidity provide the emergent specificity to keep up in evolutionary arms races with fast-evolving viruses[61].

Our results fit well with findings that at long evolutionary distances, transcriptional regulatory networks rewire, substituting individual TFs but maintaining circuit logic[39,62,63]. Here, we examined longer evolutionary distances and found that all the Gcn4 orthologs are activators. This consistency of TF function shows that the sign of TF connections in regulatory networks are more conserved than individual connections. Changes in TF function are pleiotropic, affecting many targets. Slow or rare changes in TF function likely make it easier to substitute TFs at individual regulatory elements.

Our deep dive into the evolution of one IDR family complements other studies of IDR evolution. Using small numbers of sequences, conservation of IDR function across orthologs has been observed, but often the essential residues are unknown[10]. In other systems, there is functional conservation of diverged IDRs, but the key residues are conserved[12] or motifs are conserved[64]. In other cases, functional conservation results from the composition, but not the arrangement, of residues through emergent properties like net charge[11,65–69]. The closest parallel to our turnover of key residues is *de novo* evolution of phosphorylation motifs[70]. TF IDRs are not always functionally conserved, for example in Abf1[13] and the Msn2/4 IDRs have two overlapping functions, only one of which is conserved[14]. Sox family members from Chianoflagelites can substitute for Sox2 in mouse iPSC reprogramming[6]. Cases where function emerges from physical properties may allow for even more turnover than we observe in Gcn4. There remains a need for better IDR-alignment algorithms or alignment-free methods to group functionally related IDRs.

The turnover of key hydrophobic residues in activation domain evolution bears strong parallels to the turnover of TF binding sites in enhancer evolution. In metazoans, enhancers are regulatory DNA that contain clusters of TF binding sites (TFBS). The DNA sequence of enhancers diverges rapidly as individual TFBS are gained and lost while maintaining function[71–73]. Orthologous enhancers can be impossible to detect in sequence alignments but are readily identified by searching for clusters of TFBS[74,75]. Two mechanistic insights led to this predictive power: 1) understanding that the key functional subunit is the TFBS and 2) understanding that individual TFBS can turnover. This conservation of total binding site content enables complex of regulatory DNA to identify conserved enhancers[74,76]. We find strong parallels in the evolution of TF protein sequence. TF protein sequence changes rapidly and is hard to align, but activation domain function is conserved. Analogous to the TFBS in enhancers, the functional units of activation domains are individual aromatic residues. In both cases, the grammar is extremely flexible[77]. Given that TFs function by binding to enhancers, it is striking that both the protein and the DNA are evolving in the same way. Turnover of TF binding sites endows enhancers with robustness to genetic variation, robustness to environmental stress, and evolutionary plasticity. Turnover of key

residues in activation domains may similarly endow TFs with plasticity and robustness. If TFs and enhancers are evolving in the same way, it increases the potential for compensatory mutations, expanding the neutral space and creating diverse sequence reservoirs that can be selected in new environments.

The primary limitation of this work is that we measured the activities of short fragments in one species. Measuring short uniform fragments makes the experiments possible but can miss longer 'emergent' activation domains[55,78]. If, in some species, an activation domain and cognate coactivator together experience many compensatory mutations, the assay will miss these sequences. Our analysis of Med15 coactivator conservation shows that the four activation domain binding domains are conserved (**Figure S31**). Activity of our reporter is well correlated with Med15 binding affinity *in vitro*[18]. The most active tiles are computationally predicted to bind Med15[79] (**Figure S32**). In the future, limited screening in additional species or screening tiles of multiple tile lengths would enrich this work. A secondary limitation is that we measured activity in just one condition. A future direction is to explore activity in other conditions and on other promoters.

Materials and Methods

Identification of ortholog sequences

We computationally screened for Gcn4 orthologs of *S. cerevisiae*. We started with a hand-collected set of 49 orthologs, 48 of which contained the WxxLF motif [16,55]. To find new orthologs, we used two criteria: the bZIP DNA binding domain (IPR004827) and the regular expression Wx[SPA]LF for the WxxLF motif. These criteria distinguished Gcn4 orthologs from other leucine zipper DNA binding domain TFs. We scanned 207 diverse and representative proteomes from the MycoCosm database (mycocosm.jgi.doe.gov). This computational screen yielded 1188 gene models from 129 genomes. These 1188 gene models combine to yield 502 unique proteins (**Table S1**, **Figure S1, S2**).). Of these, >99% were reciprocal Blast best hits with *S. cerevisiae* Gcn4. This initial analysis was performed in 2020 by Sumanth Mutte of MyGen Informatics. 84 of the genomes were from MycoCosm, while the original ortholog collection contributed 45 species. Genomes contained 1-32 gene models and 1-11 unique protein sequences (**Figure S1**). These sequences span nearly all the Ascomycota, the largest phylum of Fungi, representing >600 million years of evolution [42]. The 502 unique orthologs have variable lengths (**Figure 1A**), but the DBD is at the C-terminus in 500 orthologs, and the distance between the WxxLF motif and the DBD is very consistent (**Figure 1B**).

All species were from the Ascomycota except for five entries with three unique sequences from Blastocladiomycota (**Figure S1**). The Blastocladiomycota orthologs are the only proteins where the WxxLF motif does not align in the MSA. The sequence context of their WxxLF motif is H-rich instead of acidic:

e.g. AAAQ**H**VPAADGQ**W**LA**LF**P**H**PSSIDFDFNSF**H**QSFSSPPP**H**

The Blastocladiomycota tiles with the WxxLF motif have high activity in the assay. The regions of Blastocladiomycota orthologs that align to the WxxLF motif in the MSA have low activity in the assay. We suspect the N-terminal WxxLF in the Blastocladiomycota may have been gained by convergent evolution (Supplemental note 1).

The Yeast Gene Order Browser has reconstructed the local synteny of the Gcn4 locus for 37 genomes yielding a high-quality set of true homologs [41]. 36/37 species and the inferred ancestor contain one Gcn4 gene. *Kazachstania saulgeensis CLIB1764T* is missing a Gcn4 homolog. All of the post whole genome duplication species in this set contain only one Gcn4 homolog, suggesting there is no advantage of retaining two copies. All but one of the 36 the orthologs, *Zygosaccharomyces bailiii* ZYBA0L03268g, contain the WxxLF motif. Instead, *Z. bailiii* has an insertion in the WxxLF motif yielding W**PSL**EPLF. This sequence was not included in our experiment but was previously measured in a 44 AA tile, LDQAVVDEFFVNDDAPMFELDDGASGAWPSLEPLFGEDEERVAV, and had high activity in Replicate 2 of our previous paper [16]. This example further supports the observed conservation of function without conservation of sequence.

Despite substantial sequence divergence, all homologs show negative selection at the level of the full protein in the precomputed YGOB analysis. We downloaded a list of 36 pairwise Ka, Ks, and omega coefficients calculated from the yn00 output of Phylogenetic Analysis by Maximum Likelihood (PAML) (**Table S14,** November 2024).

We confirmed that the WxxLF motif is well conserved in fungal TFs with HMMER. We ran the web server for HMMER with default parameters, using *S.cerevisiae* Gcn4 as the seed sequence and restricting our search to Fungi. In the second, third, and fourth iterations of this search, the WxxLF motif was the most prominent feature of the profile HMM in the central region of the TF and always much more prominent than all other published motifs [21,78]. **Figure 2C** shows the pHMM from the fourth iteration.

For the full-length orthologs, MSAs were performed in Genious with the MAFFT algorithm (**Table S2**). We removed the two longest orthologs that had the DBD near the center. In the MSA, 54% of positions had less than 1% identity and 88% had less than 5% identity.

Short alignments were created with MUSCLE online (https://www.ebi.ac.uk/Tools/msa/muscle/) or with or with MAFFT v7.526 and visualized with weblogo.berkeley.edu or the LogoMaker Python package.

Design of the Gcn4 oligo library

We took the 502 unique protein sequences and computationally chopped them into 40 AA tiles spaced every 5 AA (e.g. 1-40, 6-45, 11-50 etc.). As a result, if two closely related sequences contain identical regions, insertions or alternatives (start sites) that change the phasing, a single tile can map to multiple full-length orthologs. We removed duplicate tile sequences, yielding 20679 unique tiles. We added 52 control sequences (controls were included twice in the oligo pool to increase the probability they were recovered in the plasmid pool during cloning). The controls included hand-designed mutants in control activation domains and a handful of sequences from our previous study [16] (**Table S3**, Control sequences). The final design file contained 20783 entries.

We reverse-translated tile sequences using *S. cerevisiae* preferred codons. We added primer sequences for PCR amplification and HiFi cloning ('ArrayDNA' column in **Table S5**). We also added four Stop codons in three reading frames to ensure translational termination, even if there were one or two bp deletions, the most common synthesis errors. We used synonymous mutations to remove instances where the same base occurred four or more times in a row to reduce DNA synthesis errors. The resulting oligo pool was ordered from Agilent Technologies. The final oligos were of the form (see primer sequences in **Table S4**):

```
FullDNAseq = primer1 + ActivationDomainDNAseq + stopCodons + primer2
```

Plasmid Library construction

The oligos were resuspended in 100 uL of water, yielding a 1 pM solution. The oligos were amplified with eight reactions of Q5 polymerase (NEB) using 1 ul of template, five cycles, Tm =72C and the LC3.P1 and LC3.P2 primers. The eight reactions were combined into a single PCR clean-up column (NEB Monarch).

The backbone was prepared by digesting 16 ug of pMVS219 with NheI-HF, PacI and AscI in eight reactions. We digested for seventeen hours at 37C and heat-inactivated for one hour at 80C. The desired 7025 bp fragment was run on a 0.8% gel, visualized with SYBR Safe (Invitrogen), and gel purified (NEB Monarch Kit). Note pMVS219 and pMVS142 have the same sequence, but the pMVS142 stock developed heteroplasmy, so we repurified it as pMVS219 and submitted the corrected stock to AddGene. Both pMVS219 and pMVS142 correspond to AddGene #99049.

We used NEB HiFi 2x mastermix to perform Gibson Isothermal Assembly to create the plasmid library. The 4x reaction volume had 328 ng of backbone and excess molar insert. We incubated at 50C for 15 min and assembled a backbone-only control in parallel. The assemblies were electroporated three times each into ElectroMax 10b E.coli (Invitrogen 18290-015) following the manufacturer's protocol. A dilution series was plated and the bulk of the cells grown overnight in 140mL LB+Amp. These cultures overgrew, so they were spundown and frozen. The cultures were regrown with 105 mL LB+Amp and a MaxiPrep was performed (Zymo). An estimated 4.2 million colonies were collected, covering the library 200-fold.

To assess the quality of the plasmid library, we prepared an amplicon sequencing library (see below). Three independent amplicon libraries were prepared, and sequences present in all three were considered to be present in the plasmid pool with high confidence. GREP for the flanking NheI and AscI sites was used to pull out the designed fragments. Only perfect matches were used in this analysis. 20717 of 20731 designed sequences were detected (99.9%). The vast majority sequence abundances were within 4-fold of each other, indicating minimal skew in library member abundance.

Yeast transformation

The plasmid library was integrated into the DHY213 BY superhost strain, MATa his1Δ1 leu2Δ0 ura3Δ0 met15Δ0 MKT1(30G) RMEI(INS-308A) TAO3(1493Q), CAT5(91M), MIP(661T) SAL1+ HAP1+, a generous gift from Angela Chu and Joe Horecka. Requests for the parent strain are best directed to them. We integrated our library into the URA3 locus with a three-piece PCR [80]. The upstream homology between URA3 and the ACT1 promoter was created by PCR amplifying the pMVS295 (Strader 6161) with the primers YP18 and CP19.P6. The downstream homology between the TEF terminator of KANMX and URA3 was amplified from pMVS196 (Strader 6768) with the primers YP7 and YP19. These template plasmids were a generous gift from Nick Morffy and Lucia Strader. To avoid PCR, the plasmid library was digested with Sal I-HF and EcoRI-HF (NEB) overnight, but not cleaned up. The homology arms were in 3:1 molar excess. 1.25 ug of total DNA was used (225 ng of upstream homology 626 bp, 225 ng of downstream homology 665 bp, and 800 ng of digested plasmid 4583 bp). Cells were streaked out from the -80C on YP+Glycerol. Four transformation cells were grown overnight in YPD, diluted into YPD, and allowed to grow for at least two doublings. We performed a Lithium Acetate transformation with 30 minutes at 30 C and 60 minutes at 42 C followed by a two hour recovery in synthetic dextrose minimal media without a nitrogen source, as recommended by Sasha Levy. We integrated plasmids in seven transformation batches, which were plated overnight on YPD and replica-plated onto YPD+G418 (200 ug/ml). Plates were stored at 4 C and then scraped with water, pooled, frozen into glycerol stocks, and mated. We collected an estimated 100,000 colonies, approximately five-fold coverage of the tiles. For 6/7 pools we sequenced tiles before and after mating, finding that 67-97% of tiles were detected both before and after mating, indicating that the mating sometimes reduced library complexity.

Yeast Mating

We mated each of the seven transformations independently to MY435 (FY5, MATalpha, YBR032w::P3 GFP ClonNat-R (pMVS102)). Downstream sequencing revealed that transformations with modest numbers of colonies (e.g. 4500) experienced no significant loss of complexity during mating, but transformations with more colonies (e.g. >20,000) experienced loss of complexity, up to 40% in one case. Subsequent matings were performed in larger volumes to avoid creating a bottleneck. Mated diploids were selected in liquid culture with YPD with 200 ug/ml G418 and 100 ug/ml ClonNat. After overnight selection, matings were concentrated and frozen as glycerol stocks.

Cell Sorting

The day before sorting, a glycerol stock of mated cells (~100 ul) was thawed into 5 mL SC+Glucose with 200 ug/ml G418 and 100 ug/ml ClonNat and grown overnight, shaking at 30 C. The morning, the culture was diluted 1:5 into SC+Glucose with G418, ClonNat, and 10 uM ß-estradiol (Sigma). The culture was grown for 3.5-4 hours before sorting.

Cells were sorted on a BD Aria Fusion equipped with four Lasers (488 blue, 405 Violet, 561 Yellow-green and 640 Red) and eleven fluorescent detectors. We used two physical characteristics gates, first to enrich for live cells (FSC vs SSC) and second to enrich for single cells (FSC-Height vs FSC-Area). Cells were sorted by the GFP signal, the mCherry signal, or the ratio of GFP:mCherry signal. The ratio is a synthetic parameter that is very easy to saturate on the eighteen-bit scale available in the BD software. Great care was taken to change PMT voltage and the ratio scaling factor (5-10% depending on the day) to make the value of the top and bottom bins as different as possible. The dynamic range of our final estimate for activation domain activity is set by the value of the top and bottom bins. The maximum activation domain strength is 100% in the top bin, and assumes the value of the top bin. The minimum activation domain strength is 100% in the bottom bin and assumes the value of the bottom bin.

We performed our sorting experiment twice. In the first run, we pooled all of the transformants into one sample and sorted it by GFP/mCherry ratio, GFP-only, mCherry-only.

We sorted one million cells per bin. For the ratio sort, we split the ratio histogram in eight approximately equal bins [16].

In the second round of sorting, we split the transformants into two pools, labeled A and B, so we could assess measurement reproducibility for independent transformants. Pool A and Pool B are true biological replicates. We sorted each pool by GFP/mCherry ratio, GFP-only, mCherry-only. We used the comparison of the A and B pool measurements to assess measurement reproducibility of true biological replicates. We have never previously measured this biological reproducibility. On this day, we sorted 250000 cells per bin.

Sorted cells were grown overnight in SC-glucose. The next morning, gDNA was extracted with the Zymo YeaSTAR D2002 kit, using Protocol I with chloroform according to the manufacturer instructions. We have previously shown that growing cells overnight makes the gDNA extraction easier but does not change the computed activation domain activity [16].

Amplicon Sequencing Library preparation

Amplicon sequencing libraries were prepared from genomic DNA in three steps. First, the general vicinity of the tile sequence was amplified with CP21.P14 and CP17.P12 using 100 ng of gDNA as template and yielding a 604 bp product that was cleaned up (Monarch PCR cleanup). In the second PCR, we added 1-4 bp of phasing on each end and the Illumina sequencing primer in 7-10 cycles with SL5.F[1-4] and SL5.R[1-3]. These seven phased primers were pooled and added to all samples. Four nanograms of the first PCR were used as template for the second PCR. Two microliters of the second PCR served as template for the third PCR. The third PCR added unique Index1 and Index2 sequences to each sample with an additional 7-10 cycles. These final products were cleaned up with PCR columns or magnetic beads (MacroLab at UC Berkeley) and submitted for sequencing. We performed 2x150 bp paired end sequencing in a shared Nova-Seq lane at the Washington University School of Medicine Genome Technology Access Center (GTAC). GTAC provided demultiplexed fastq files. We sequenced additional samples in shared Nova-seq lanes with MedGenome.

Sequencing Analysis

After demultiplexing samples and pairing reads with PEAR, we kept only the reads where the tile DNA sequence contained a perfect match to a designed tile. For each eight bin sort, we performed two normalizations. We first normalized the reads by the total number of reads in each bin. Then, we normalized across the eight bins to calculate a relative abundance. We then converted relative abundances to an activity score for each tile by taking the dot product of the relative abundance with the median fluorescence value of each bin (**Table S8**). This weighted average is the measured activation domain activity. Tiles with fewer than forty-one reads were not included in the final dataset. These analysis scripts are available at github.com/staller-lab/labtools/tree/main/src/labtools/adtools. This preprocessing computed an activity for each tile in each experiment. Activity is uncorrelated with total reads (**Figure S5E**). The pooled ratio sort (BSY2) had 115.6 M reads. The Replicate A ratio sort had 934.5 M reads, and the Replicate B ratio sort had 697 M reads. Replicate A GFP had 33.1 M reads, Replicate B GFP had 31.6 M reads, Replicate A mCherry had 32.8 M reads, and Replicate B mCherry had 30.3 M reads.

Measurement Reproducibility

We used the two measurements of independent transformants to assess the reproducibility of our measurements of true biological replicates (R = .870; **Figure S5A-D**). Reproducibility is higher (R= .919) for highly abundant tiles (>1000 reads).

We combined data from the two biological replicates. For tiles present in both populations (n= 11797), we averaged the two measurements and used the standard deviation as the error bar. For tiles present in only one population, we used that measurement and did not report error bars. These combined data agree very well with the pooled sort (R= .919; **Figure S5C**). Activity was saturated for forty-nine tiles, but most of

these were measured with low fidelity because they had low read depth, and forty-seven were present in only one biological replicate. We identified forty-one tiles that were very highly active in both replicates and had high read depth in both replicates (**Table S11**). These we recommend for CRISPR Activation studies in yeast.

We assessed whether the mating introduced biological variability. We remated seven pools of the integrated library to the same reporter line, selected for diploids, pooled them, and resorted cells. This time we sorted 500,000 cells per bin. This measurement agreed with the initial experiments (R = 0.920; **Figure S5D**).

Inferred activity was not correlated with read count, which, as previously shown, is another indicator of high-quality data (**Figure S5E**).

We compared activity measurements to our previously published results [16]. Previously, we used forty-four AA regions, and here we used forty AA tiles. We considered any forty-four AA tile that contained one or our forty AA tiles to be corresponding pairs. The extra four AA can modify activity, so the correspondence of these measurements will not be perfect. The observed Pearson correlation of 0.786 and Spearman correlation of 0.731 indicate the new data are of high quality and consistent with previous measurements (**Figure S5F**).

The technical reproducibility of our measurements at UC Berkeley are lower than the published reproducibility from sorting at Washington University in St. Louis [16]. In both cases, we sorted the same cell population twice and created independent sequencing libraries. In 2018, the technical reproducibility was high, Pearson R = 0.988. The 2018 work had a smaller library (<5000 unique sequences) and sorted more cells (1-2 million cells per bin). Sorting more cells per library member increases the technical reproducibility of the measurement. The sorter operator in the 2018 work was more experienced than the sorter operator in this work (MVS), and the machine was maintained to a higher standard of operation, so the sorted populations were purer.

The eight bin ratio activity measurements are primarily driven by the GFP signal. Activity (ratio) is largely separable from abundance assessed by the mCherry sort (**Figure S5G-I**) and well-correlated with the GFP sort (**Figure S5J-L**).

Determining a threshold for active tiles

The full distribution of tile activities has a peak at low activity, which, based on control sequences, is clearly inactive, with a heavy right shoulder and a heavy right tail (**Figure 1F**). The tail contains the control sequences with known high activity (**Figure S7**). We set out to fit the inactive sequences to a Gaussian distribution and use this distribution to create a threshold for active sequences. We first bin all tiles according to their activity score such that there are ~ 200 tiles per bin and plot a histogram. We hypothesized tile density is highest around inactive tiles and thus refer to all tiles to the left of the resulting histogram's peak as inactive tiles. We fit a one-sided Gaussian to these inactive tiles (**Figure S8A**) and call the two-sided extension of this Gaussian the inactive tile distribution (**Figure S8B**). Treating this Gaussian inactive tile distribution as our null hypothesis, we calculate p-values for each tile (not including tiles earlier used as inactive, **Figure S8D**). We then correct for multiple comparisons using FDR [81] and Bonferroni [82] corrections. The 1% FDR threshold was 33821 (60.6% of tiles active). The 1% FWER threshold was 45373 (46.6% of tiles active). As a conservative threshold to call active sequences, we used the 1% FWER threshold of 45,373. All of our designed inactive control sequences are below this threshold.

After trying many thresholds (**Figure S8)**, we ultimately chose the top 20% (94,031) as a threshold for high activity. The choice of threshold had very little effect on our results. In particular, a wide range of threshold has almost no effect on the number of orthologs with an active tile.

Protein sequence parameters

We computed protein sequence parameters (Net charge, local net charge, Kyte Doolittle Hydrophobicity, Wimley White hydrophobicity, Kappa [83]) with localCIDER [84]. The

OmegaWFYL_DE mixture parameter computes the mixture statistic between W,F,Y,L residues and D,E residues using the seq.get_kappa_X(['D','E'],['W','F','Y','L']) function in localCIDER [85]. We predicted intrinsic disorder with MetaPredict2 [86]. We counted motifs with regular expressions in Python with the "re" package.

The MAFFT algorithm aligns the WxxLF motif for all but three orthologs. For three orthologs, in the Full_length_ortholog_dataframe, we corrected the "WxxLF motif location" parameter using the coordinates from the MSA. These species are the only ones outside the Ascomycota that have the motif. We suspect the WxxLF motif convergently evolved in these distance orthologs because the context is very different and H rich. Blastocladiomycota_jgi|Catan2|1097078|CE97078_6759, Blastocladiomycota_jgi|Catan2|1466814|fgenesh1_pg.199_#_9, and Blastocladiomycota_jgi|Catan2|1506241|gm1.11555_g.

To predict helical propensity of ortholog sequences, we used the Sparrow package in Python [https://github.com/idptools/sparrow]. A region was called helical if it contained five adjacent residues with over 50% chance of being helical. A large proportion of sequences have no residues with a >50% probability of being helical in this region. We consider this predictor to capture the propensity to form a helix in some context. To count proline residues in the region homologous to the known helix, we used the five AA upstream and five AA downstream of the WxxLF motif. From the 500 orthologs in the MSA, there are 115 unique 15 AA regions around the WxxLF motif; twenty-three contain three prolines (20%) and three contain four prolines (2.6%).

Imputing activity in the full-length orthologs

We used the tile data to impute the activity of each position in each of the full-length orthologs. The 19099 recovered tiles mapped to 68577 locations on the orthologs (each tile matched to 3.6 orthologs on average). We used a second order Loess smoothing (20 nearest points with the loess.loess_1d.loess_1d() function) across tiles to impute the activities of all positions in the 502 unique orthologs. This quadratic smoothing can cause artifacts on the extreme ends of the protein, such as predicting negative activity. To remove this artifact, we constrained the imputed activity to be no more than the maximum measured and no less than the minimum measured in that ortholog.

To validate the Loess smoothing, we averaged together all activities for all tiles that overlapped a position, equally weighing all tiles. These averages were more jagged because of the stepwise nature of the tiles. This simple average also created artifacts at the ends of the protein where only one tile is present. The Loess and average smoothing methods agreed well (97% had Pearson R > 0.80) (**Figure S33**).

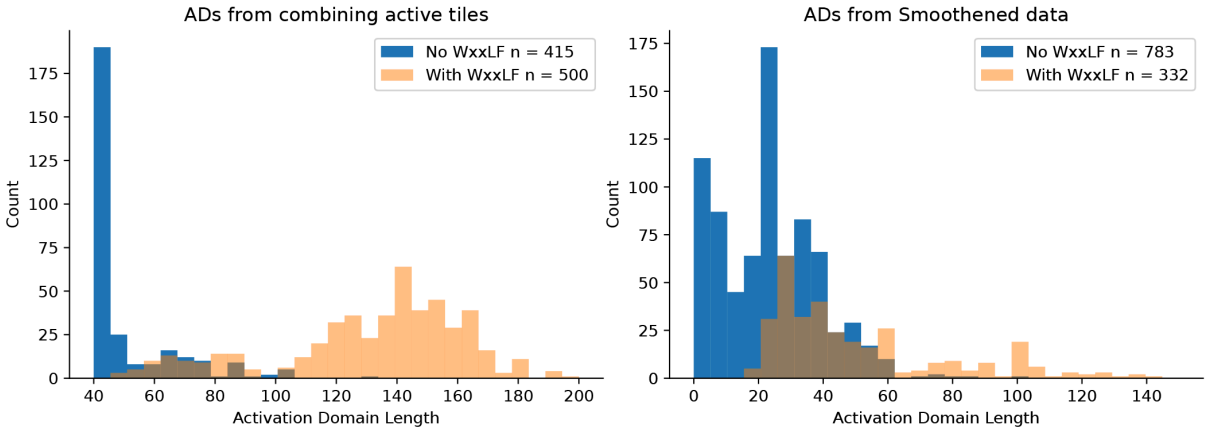
We used the imputed activities to create the heatmaps to visualize activity across the orthologs. We tried many variations of these heatmaps but ultimately found that aligning the sequences on the start of the DBD or on the WxxLF motif was most informative. In the main text, we removed the twenty-seven longest sequences to make the visualization easier to display but added most of them back in Figure S9.

We tested the hypothesis that insertions are enriched for active tiles by projecting activity onto the MSA. We defined insertions as the positions in the MSA with residues (non-gaps) in less than 1% of sequences (n < 5), which yielded 880/2690 (32.7%) of positions. In a two-sided t-test of the imputed activities of the insertion positions compared to all other positions, insertions were less active (p < 1e-52). We concluded that insertions are depleted for sequences with activation domain activity in *S. cerevisiae*.

To estimate the activity at the WxxLF motif, we used the integral of the imputed activity from -10 to +10 around the W of the WxxLF motif. When this integral was below our activity threshold, we called sequences inactive in this region. Using this integral, ninety-two unique sequences had high activity (>150000) and thirteen unique sequences had low activity, less than our activity threshold. Thirty-three had intermediate activity.

For motif enrichment, we performed a Welch's t-test assuming unequal variances stats.ttest_ind(Sequences_WITH_Motif,Sequences_WITHOUT_Motif, equal_var=False).

To count activation domains on each TF, we combined active overlapping tiles, taking the union. With this method, we found 500 ADs with the WxxLF motif and 415 ADs without the WxxLF motif. We required more than forty residues between activation domains before they were called as two separate domains. Calling activation domains from the imputed activity map gives slightly different results because some very close double peaks are split. With the smoothed data, there are 332 ADs with the WxxLF motif and 783 ADs without the WxxLF motif.



ANOVA

We used ordinary least squares regression (OLS) to create a baseline model for how composition controls activation domain function. We used ANOVA, OLS, and adjusted R-squared to compare models. See the Composition_ANOVA jupyter notebook for the full analysis. Briefly, we used the ols(formula, ANOVA_DF).fit() function from the statsmodels package to fit the model, find coefficients, and compute adjusted R-squared values. We used the anova_lm(model, typ=2) function to find the sum of squares explained by each parameter. We used a Bonferroni multiple hypothesis correction to remove non-significant parameters and refit the model. In most cases, one iteration was sufficient to get a model where all parameters were significant. For the dipeptides, we used two interaction terms. All ANOVA parameters are in **Table S9**.

OLS regression on single amino acids explains 49.9% of variance in activity (**Table 1**, AUC = 0.9346, PRC = 0.7620, **Table S9**). Iteratively removing non-significant parameters led to sixteen residues which explain 49.9% of variance. We repeated the regression with 400 dipeptides and found 69 significant parameters that explain 60.2% of the variance in activity (**Table 1**, AUC = 0.9472, PRC = 0.8190). Half the variance in activity could be explained by composition alone and dipeptides offered ~10% improvement.

We predicted *de novo* motifs using the DREAM suite and then repeated the OLS ANOVA analysis using the motifs. We performed *de novo* motif searching on multiple slices of the data, but highly active (n=3524) vs. inactive (n=15575) were the most interpretable and gave the clearest signal in the ANOVA analysis. First, we ran the package STREME from the MEME suite to discover motifs that are enriched in a list of sequences relative to a user-provided control list.

For the OLS on *de novo* motifs, we used the motif counts provided by the DREAM motif prediction software (**Table S10**). For simplicity, in the parameter table, we refer to each motif as a string, but we used the PWM for actually finding motifs in each sequence with FIMO.

Machine learning

We predicted activities on full length orthologs using publicly available models, TADA, ADpred, and PADDLE[18,21,23,24]. All models were run on the SAVIO high performance computing cluster at UC Berkeley. TADA uses 40 AA windows, ADpred, 30 AA windows, and PADDLE 53 AA windows. For each TF, we tiled at 1 AA increments, spanning the full

proteins (e.g. 1-40, 2-41 etc). For full length TF analysis, we corrected the inferred activity at each position (Loess smoothing) with the predictions at each position. The smoothed data averages out some measurement noise so all the model performance is improved on smoothed data. For individual tile analysis, we used the center aligned score. We also tried maximum scores, average scores, and other variations, but chose center aligned. ROC and PRC analyses were performed with the sklearn python package.

Predicting the impact of mutating F residues in the central activation domains. We tile the 138 unique 70AA central regions into 40AA tiles spaced every 1 amino acid. For each tile, we computationally mutated each F individually, all pairs, all triplets, and all sets of four or more. For each mutant, we predicted activity. The mutants are predicted to have less activity. For each mutant, we also computed the change in activity. Finally, we grouped the changes in activity based on the conservation of each F residue.

Pax6 alignments

BLAST alignment of mouse Pax6 (P63015) and *D. melanogaster* Eyeless (O18381) was performed with the Uniprot canonical sequences. We calculated the DBD percent identity using the longest aligned region that encompassed the annotated DBD (5-135 and 157-187, respectively). We realigned the regions C-terminal to the end of this DBD alignment and found three regions with modest-to-high scores: (79+16+7)/287 = 35.5% residues identical and (88+28+11)/287 = 44.3% residues similar in the three regions. We summed the number of identical or similar residues to compute similarity. We used the shorter mouse IDR length as the denominator, overstating conservation. Alignments are in **Figure S34**. Using the more permissive BLOSSUM90 matrix yielded a fourth small aligned region that increased the similarities: (79+16+14+7)/287 40.4% residues identical and (88+26+18+11)/287 = 50% residues similar.

Datafiles

All the raw sequencing data has been deposited at NIH SRA Accession #PRJNA1186961:
http://www.ncbi.nlm.nih.gov/bioproject/1186961
All the analysis scripts are deposited on github and Zenodo:
10.5281/zenodo.14201918
https://github.com/staller-lab/Gcn4-evolution
github.com/staller-lab/labtools/tree/main/src/labtools/adtools
https://github.com/staller-lab/Gcn4-evolution
All the processed data is attached in supplemental tables (**Tables S5 - S7**).
Processed sequencing read counts are in **Table S13**.

The ‘masterDF’ dataframe contains each designed tile (**Table S5**). Tiles that were not measured have activity recorded as nan or 0. The ‘orthorlogDF’ dataframe contains all tiles associated with each original full-length ortholog (**Table S6**). As a result, tiles occur multiple times because they map to multiple orthologs. The ‘NativeLocation’ is the position of the tile relative to the first amino acid. The ‘NormLocation’ is the position of the tile relative to the WxxLF motif. Finally, the ‘FullOrthoDF’ dataframe contains one entry for each full-length ortholog, and each column contains an array with values for each position (**Table S7**), such as imputed activity at each position and local charge from localCIDER. The location of the bZIP DNA-binding domain was identified with the InterPro signature (IPR004827).

Acknowledgments

We would like to thank Nick Ingolia, Zeba Wunderlich, Rachel Brem, Alex Holehouse, Shahar Sukenik, and Ashley Wolf for helpful comments on the manuscript. Sumanth Mutte for finding the initial orthologs. We thank Lucia Strader, Nicholas Morffy, Ross Sozzani, Lisa Van den Broeck, Hunter Nisonoff, and Jennifer Listgarten for helpful discussions, and Nick Morffy and Lucia Strader for the yeast genome targeting plasmids. Igor Grigoriev identified the deprecated *Tortispora caseinolytica* gene models. Weijing Tang performed exploratory analyses not included in the final manuscript. The Regents of the University of California have filed a patent based on the findings of this study. The DHY213 BY super host strain used for library construction was a generous gift from Angela Chu and Joe Horecka, and requests for this strain should be directed to them.

Funding

CJL training grant T32HG4725. AL UC Berkeley URAP. MAZ T32GM148378. MS and SRK UC Berkeley SEED Scholars Program. SRK UC Berkeley SURF. AF biophysics training grant T32GM146614**.** GPS UC Berkeley BSP scholar, McNair Scholar, and UC Berkeley SURF. This work was supported by the Burroughs Wellcome Fund PEDP, Simons Foundation grant 1018719 to MVS, NSF grant 2112057 to MVS, and NIH grant R35GM150813 to MVS. MVS is a Chan Zuckerberg Biohub – San Francisco Investigator.

Supplementary Note: Additional analysis of the orthologs

Selection of the Gcn4 orthologs

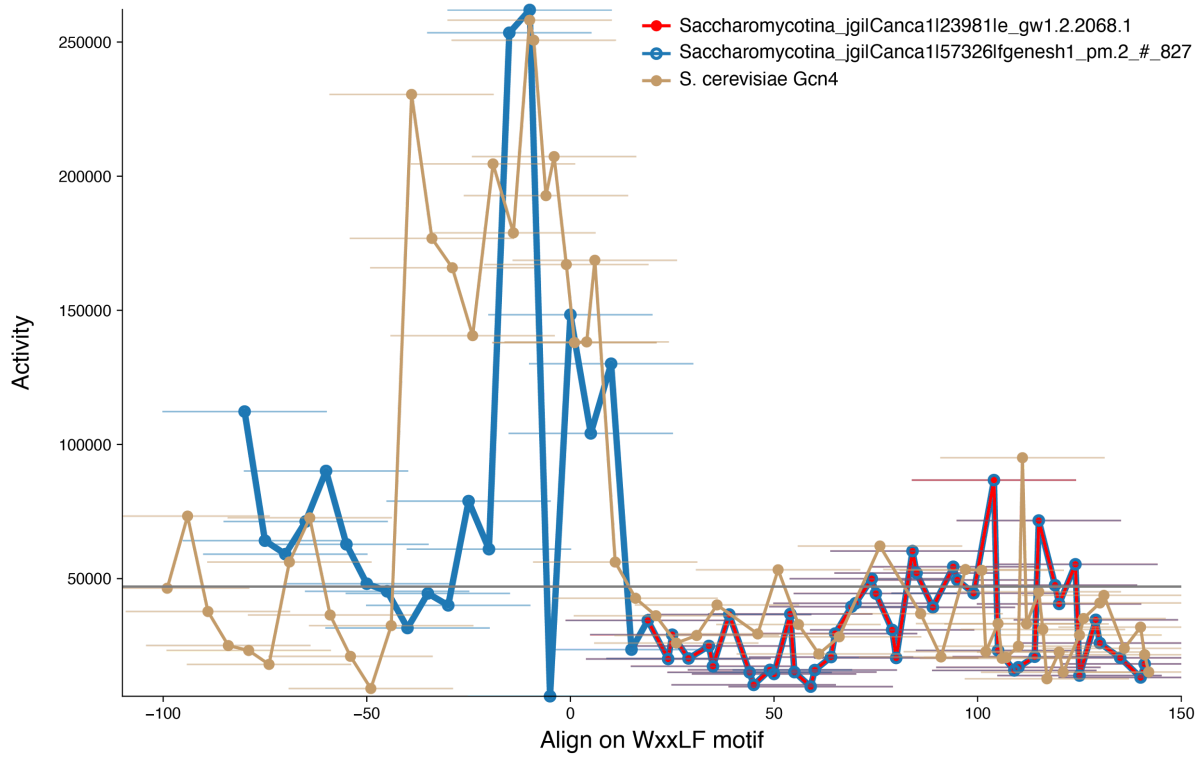
We chose a diverse set of orthologous Gcn4 protein sequences for functional characterization in *S. cerevisiae*. We started with a set of forty-nine previously identified orthologs [16,41,55]. In these, 48/49 contain an WxxLF motif. Next, we scanned 207 representative proteomes from the MycoCosm database, sampling the diversity of fungal genomes (**Figure S1, S2**). To distinguish Gcn4 orthologs from other basic-leucine zipper (bZIP) domain TFs, we required the presence of both a bZIP DNA-binding domain (IPR004827) and the WxxLF motif. This computational screen yielded 1188 hits in 129 genomes. There are 502 unique Gcn4 ortholog sequences that we used for all our experiments and analyses (**Figure S1**). These sequences span nearly all the Ascomycota, the largest phylum of Fungi, representing >600 million years of evolution [42]. The 502 unique orthologs have variable lengths (**Figure 1A**), but the DBD is at the C-terminus in 500, and the distance between the WxxLF motif and the DBD is very consistent (**Figure 1B**).

The Gcn4 MSA typifies eukaryotic TF evolution, with a highly conserved DBD and lower conservation in the rest of the protein (**Figure 1C**). Sequence divergence is driven by insertions: 88% of columns in the MSA contain fewer than 5% of sequences (n<25) and 54% of columns contain <1% of sequences (n<5) (**Figure S4**). Without user input, the MAFT algorithm aligned the WxxLF motif in nearly all sequences (Methods). We suspect that MAFT aligned nearly all WxxLF motifs because the distance between this motif and the DBD is highly consistent. Distant pairs of sequences do not align outside of the DBD, but we have

enough sequences to bridge the full diversity of the collection. The central activation domain shows intermediate levels of conservation largely driven by the WxxLF motif. Since we required all the orthologs to contain a WxxLF motif, the conservation of this motif is overstated in **Figure 1C**, but we independently verified that this motif is the most conserved sequence outside the DNA-binding domain using a HMMER search of fungal TFs (**Figure 2C**).

All orthologs are activators

To show that all the orthologs contain at least one active tile, we used multiple thresholds. As an unbiased threshold for modest activity, we fit a Gaussian distribution to the inactive sequences. Using this highly permissive threshold, all orthologs have at least one tile that is active. As a stringent threshold for activity we doubled this threshold, or used the top 20% of sequences, which yielded very similar values. At the stringent threshold, there is only one ortholog with no active tiles, Canca1_23981 from *Tortispora caseinolytica.* This ortholog is an alternative gene model for the Canca1_57326 protein, which contains an additional 99 N-terminal residues with twenty-three overlapping active tiles that comprise two activation domains, the second of which overlaps the WxxLF motif. The short form of the protein starts at the WxxLF motif. Based on improved, transcript-based gene models, the short version, Canca1_23981, is likely a computational annotation error. There is more support for the long version, Canca1_57326. Given the relatively weak evidence supporting the one potential exception, we conclude all of the Gcn4 orthologs are activators.

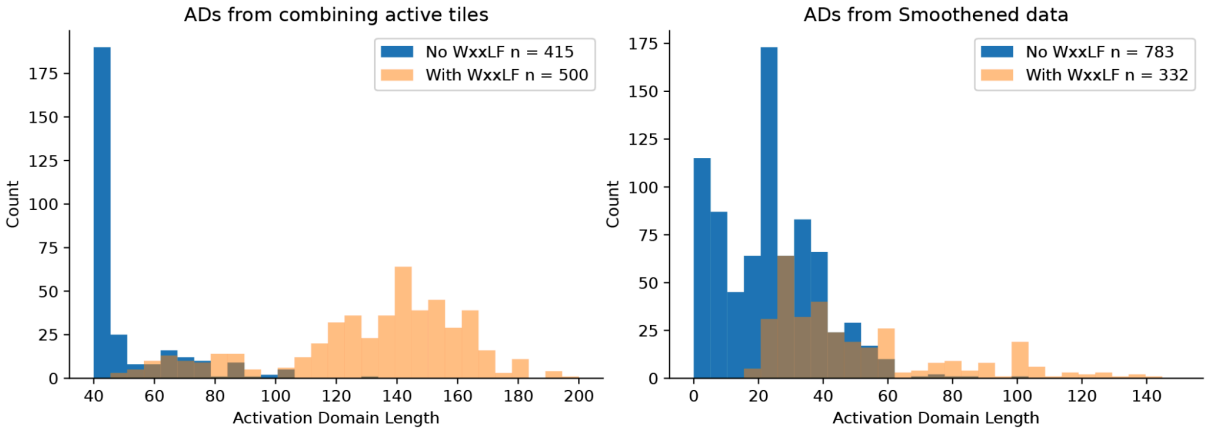


Alternative gene models from *Tortispora caseinolytica.* Canca1_23981 (red) and Canca1_57326 (dark blue) are alternative gene models for the same locus. Importantly, they are identical, so the red overlaps the dark blue.

Activation domains per ortholog

Longer TFs often have more active tiles (**Figure 2E**). When we merged overlapping active tiles, most orthologs had more than one activation domain (**Figure 2F**). The lengths of the merged activation domains are bimodal, but they are generally <200 AA (**Figure 2G**).

We used two methods to count activation domains on each ortholog. First, we aggregated overlapping active tiles. This method biases towards fewer longer activation domains because there must be more than forty AA between active regions for them to be called as separate activation domains. With this method, 245 orthologs (48.8%) have only one activation domain, and for all these orthologs, the AD overlaps the WxxLF motif. In total 500 activation domains contained the WxxLF motif, and these were longer than N-terminal activation domains. There were also many single-tile activation domains. Second, we used the smoothed data to find activation domains. This method averages out some experimental noise and shortens active regions. In this approach, there are only 332 orthologs with an activation domain that contains the WxxLF motif, consistent with the peak of activity being upstream of this motif. There are more N-terminal activation domains, and they are shorter than activation domains with the WxxLF motif. In both methods, the sequences of the N-terminal activation domains are diverse.



Clusters of aromatic and leucine residues make large contributions to function

In the control activation domains, all published motifs of aromatic and leucine residues made large contributions to activity, but no individual motif was sufficient for full activity. Historically, *S. cerevisiae* Gcn4 is annotated with two activation domains: the CAD is residues 101-140, while the N terminal activation domain (NAD) is residues 1-100 (**Figure 2A**)[78,87,88]. There are six published motifs, F9 F16 (FxxxxxxF), F45 F48 (FxxF), F67 F69 (FxF), F97 F98 (FF), M107 Y110 L113 (MxxYxxL or MFxYxxL), and W120 L123 F124 (WxxLF)[78,88]. The CAD has two motifs that make large contributions to activity[78,88] (**Figure 2B**). The strongest tile from Gcn4 was the junction of the NAD and CAD (residues 90-129), which we call the altCAD, a region with three motifs [78] that make large contributions to function (**Figure 2B, S7**). All published motifs are enriched in active tiles (**Figure S20A**), and tiles with multiple motifs are more likely to have high activity (**Figure S20B**). However, in our sequences and an independent set of fungal orthologs, only the WxxLF motif is well conserved (**Figure 1C, 2C, S3**). We do not see reemergence of any published motifs. The hydrophobic motifs essential for function in *S. cerevisiae* are not conserved and do not experience evolutionary turnover.

Sequence features of strongly active tiles

The Gcn4 ortholog tiles efficiently detected known sequence features of strong yeast activation domains. Acidic, aromatic, leucine, and methionine residues make the largest contributions to activity[16,18–23,28,31] (**Figure 4A, C**). Aromatics generally increase activity, but too many aromatic residues reduces activity (**Figure S17F,G**), a non-monotonic trend previously seen only in synthetic peptides[18] and mutant activation domains[28]. This non-monotonicity is a key piece of evidence supporting the acidic exposure model because it shows how too many hydrophobic residues can overwhelm the exposure capacity of the acidic residues[26–28]. Moreover, aspartic acid (D) makes much stronger contributions to activity than glutamic acid (E) (**Figure 4C**), which has only been seen in mutants[18] and weakly in plant activation domains[23]. We suspect this effect occurs because the negative charge is closer to the peptide backbone, leading to a stronger solvation effect and more exposure of nearby hydrophobic residues [43]. This modestly sized dataset gave a much clearer picture of key sequence properties than much larger datasets [18,21,23], indicating that orthologs provide a very efficient set of sequences for learning the sequence features that control function (**Figure S20**).

Evidence for negative (purifying) selection

The Yeast Gene Order Browser (YGOB) contains a high quality set of thirty-six true homologs inferred from chromosomal synteny. All of the species following the whole genome duplication contain only one Gcn4 homolog, suggesting there is no advantage of retaining two copies. This result suggests that most species will have just one true homolog. The YGOB analysis of full-length TFs shows negative selection (**Figure S1E**), implying there is pressure to maintain a functional protein. This weak negative selection and large protein diversity supports the idea that the neutral space is very large and that the Gcn4 sequence can drift.

Enforcing the presence of a strict WxxLF motif left out one true homolog from YGOB, *Zygosaccharomyces bailiii* ZYBA0L03268g, which has an insertion in the WxxLF motif yielding W**PSL**EPLF. This sequence was not included in our current experiment but was measured as highly active in one replicate in Staller et al. 2018, suggesting activation domain function is also conserved in this ortholog. This example reinforces the idea that motifs can be flexible.

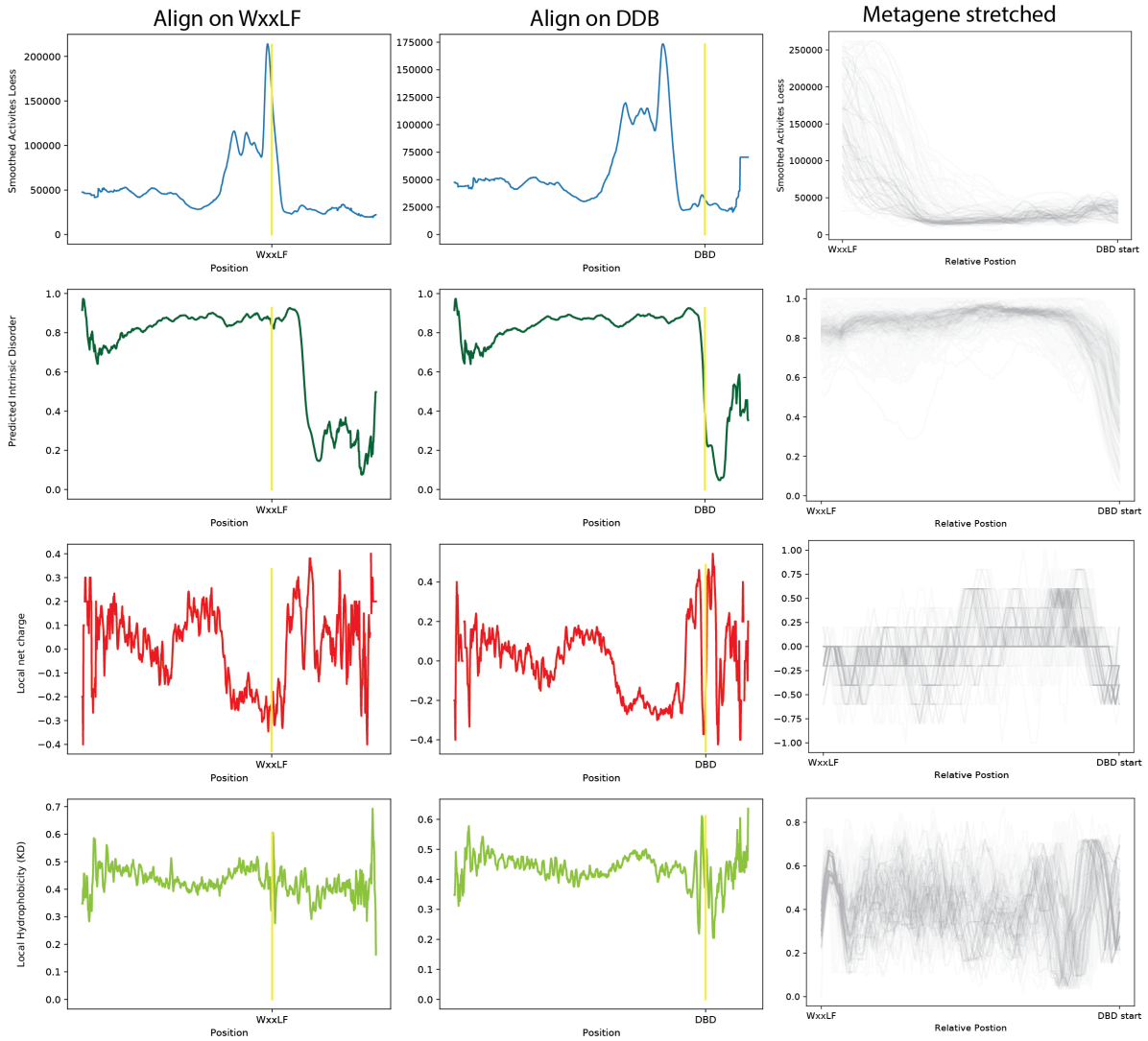
Analysis of the Gal11/Med15 coactivator

The best characterized coactivator of *S. cerevisiae* Gcn4 is Gal11/Med15. Med15 contains four regions that bind to Gcn4, the KIX domain and three activation domain binding domains (ABD1, ABD2, ABD3) [44]. Activity of our P3 promoter is well correlated with *in vitro* binding to Med15 [18], indicating this promoter is a reliable reporter of binding to Med15. We collected a set of 653 Gal11 orthologs from the Y1000+ genomes and created an MSA. The KIX, ABD1, and ABD3 domains are more conserved than the rest of the protein. ABD2 approaches the rest of the protein. The residues of the ABD1 domain that contact Gcn4 [45] are reasonably conserved, but not more conserved than the rest of ABD1 (**Figure S31**). Overall the conservation of Med15 is much higher than Gcn4.

Analysis of the spacer sequence between the CAD and DBD

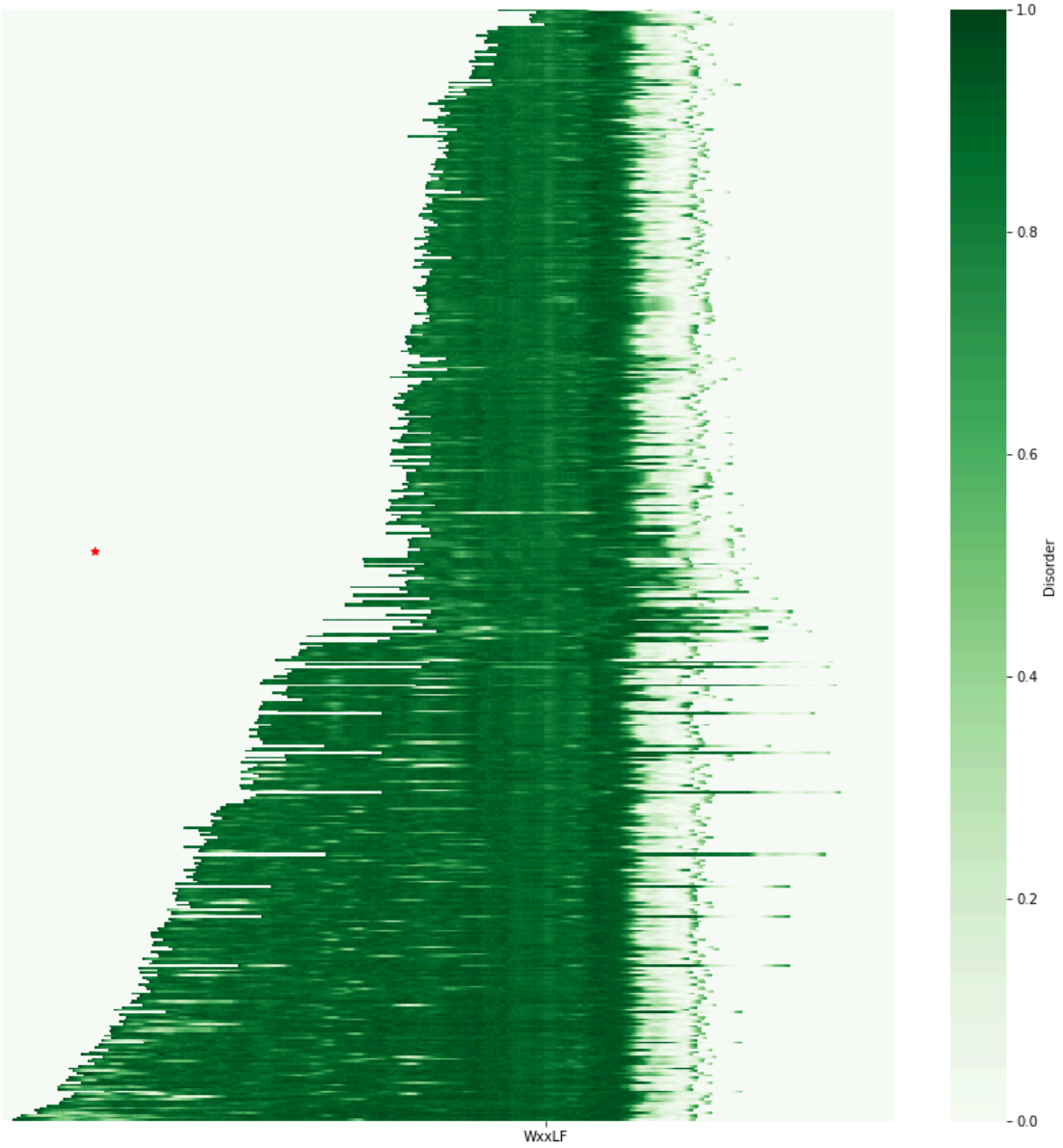
The distance between the WxxLF motif (CAD) and the DBD is highly conserved and may be an entropic spacer. The amino acid sequence of this spacer is very poorly

conserved, but both the undulating charge pattern and the high degree of predicted intrinsic disorder are conserved. NMR data clearly indicates that *S. cerevisiae* Gcn4 is fully disordered in solution and that the DBD folds upon binding DNA and the CAD folds upon binding Med15. Predicting this pattern is difficult, and Gcn4 has become a stringent test for intrinsic disorder prediction algorithms. AlphaFold predicts the DBD correctly. AlphaFold predicts many short, low-confidence helices outside the DBD, but none overlap the CAD NMR helix. To predict intrinsic disorder of the orthologs, we used Metapredict, which carefully examined performance on Gcn4 during algorithm development [86]. Based on this analysis, the most disordered region in all orthologs is the sequence between the CAD and DBD. This region has a positive to negative charge undulation just before the DBD.



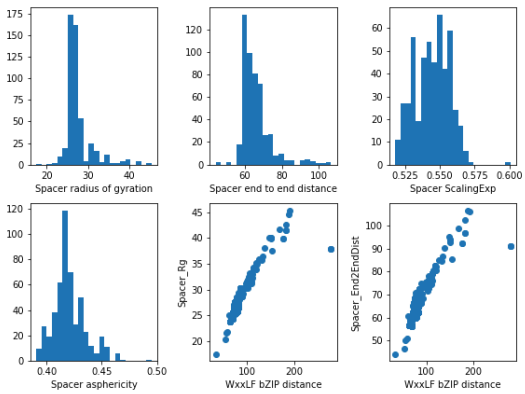
Analysis of the spacer sequence between the WxxLF motif and the DBD

Left panels align position on the WxxLF motif. Middle panels align position on the DBD. The spacer is the sequence between these landmarks. Imputed activity of the spacer is low. Predicted intrinsic disorder of the spacer is high (Metapredict2). Negative change undulates between the landmarks. The region right after the WxxLF is negatively charged, followed by a positively charged region and another net negative region just before the positively charged DBD. Hydrophobicity is high throughout.



Predicted disorder in the spacer sequence peaks between the WxxLF motif and the DBD

We speculate this region is a conserved entropic spacer that keeps the activation domain away from the DBD and exposed to partners. *S. cerevisiae* has uncommonly long spacing between the WxxLF and DBD (**Figure 1B**, red arrow). We tested this idea by predicting biophysical parameters with Albatross [47]. We see that the predicted radius of gyration (estimate of ensemble size) and end-to-end distance distributions are very tight, implying that there might be some selection to maintain a specific 3D spacing distance.



Computationally predicted scaling exponents and biophysical properties of the spacer from Sparrow.
The highly consistent predicted dimensions support they hypothesis that this spacer is keeping the central activation domain away from the DBD.

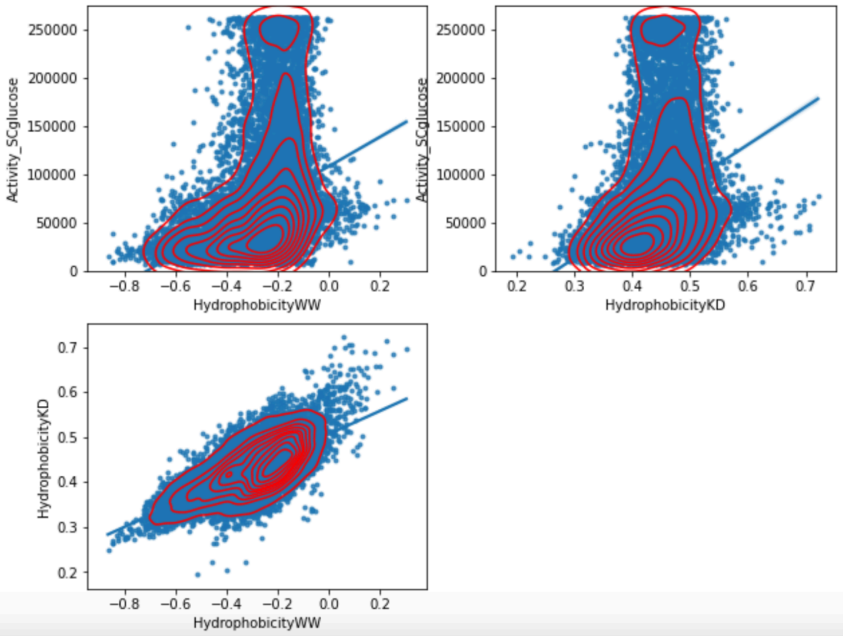
Additional analysis of tile sequence properties

Yeast activation domains are more reliant on aromatic residues than leucine residues. This difference is illustrated by the human CITED2 activation domain. In human cells, the aromatic residues make small contributions to CITED2 function, but in yeast, these residues make large contributions to function. Leucine residues contribute to CITED2 function in both yeast and human cells. The mutant of CITED2 without aromatic residues was the strongest sequence with no aromatic residues (**Figure S7B**). It is mildly surprising that CITED2 works in yeast because its primary coactivator partner, TAZ1, is not present in yeast.

Sanborn et al. argued that the Wimley White hydrophobicity (WW) score was well correlated with AD activity [18]. We had previously used the Kyte Doolittle hydropathy (KD) score and found no correlation in designed mutants [16]. The largest difference between these tables is tryptophan, W, which has a high value on WW and moderate value on KD. Since W makes large contributions to activity, we believe that the number of W's drives the conclusion by Sanborn et al. 2021. In our Gcn4 ortholog tiles, the two hydrophobicity scores are well correlated with each other. Both have similar, low correlations with activity. Some hydrophobicity is required for activity. The combination of acidity and hydrophobicity is more predictive than hydrophobicity alone.

Out[18]:

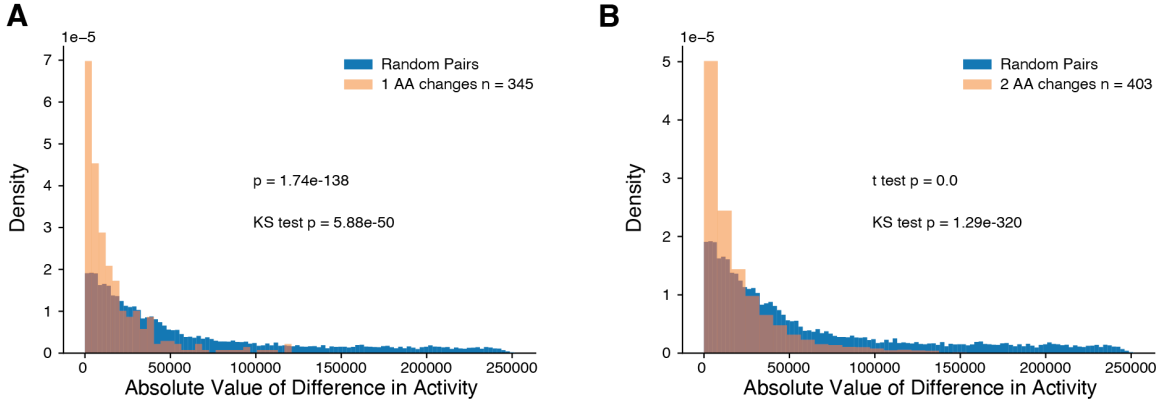
	HydrophobicityKD	HydrophobicityWW	Activity_SCglucose
HydrophobicityKD	1.000000	0.689229	0.341205
HydrophobicityWW	0.689229	1.000000	0.354076
Activity_SCglucose	0.341205	0.354076	1.000000



Naturally occurring changes in sequence generally do not change activity

Most naturally occurring sequence changes do not change activity. Starting with the altCAD as an anchor, we identified related sequences with increasing edit distance. As sequence divergence increased, all the natural sequences maintained high activity. In contrast, designed mutants show that small changes in sequence can cause loss of activity. Large effect changes are absent from the evolutionary record. This result supports a model where neutral drift and weak negative selection maintain activation domain activity.

Next, we compared pairs of sequences that differed by one or two amino acids. As a null model for differences in tile activities, we chose 10000 random pairs of tiles and computed the difference between their activities. The distribution of activity differences between tiles that differ at 1-2 amino acids is much smaller.



In most cases, there was little-to-no change in activity. We imposed a strong threshold for change in activity: either one member of the pair was active and the other inactive, or both were active but differed in activity by more than 50%. In the majority of cases that change activity, the sequence change was interpretable by our acidic exposure model: the stronger tile had additional acidic or hydrophobic residues. Of the 345 pairs of tiles that differ at a single position, 15 pairs (2.5%) had different activities and 9 supported the acidic exposure model. In four cases, an L or M was added that increased activity. In one case, an E>D change increased activity. In three cases, adding an S or G, which promotes disorder and expansion, increased activity. Of the 403 pairs of tiles that differ at two positions, 27 changed activity (7%). Two of these were designed mutants in the altCAD, FF>AA and LL>AA, both of which caused large decreases in activity (**Figure S7**). 17/27 cases (or 15/25 natural cases) supported the acidic exposure model. These data further support the mounting evidence that activation domains are robust enough to maintain because most single and double AA changes do change activity.

References

1. Onuma, Y., Takahashi, S., Asashima, M., Kurata, S. & Gehring, W. J. Conservation of Pax 6 function and upstream activation by Notch signaling in eye development of frogs and flies. *Proceedings of the National Academy of Sciences* **99**, 2020–2025 (2002).
2. Lynch, V. J. & Wagner, G. P. Revisiting a classic example of transcription factor functional equivalence: are Eyeless and Pax6 functionally equivalent or divergent? *J. Exp. Zool. B Mol. Dev. Evol.* **316B**, 93–98 (2011).
3. Halder, G., Callaerts, P. & Gehring, W. J. Induction of Ectopic Eyes by Targeted Expression of the eyeless Gene in Drosophila. *Science* **267**, 1788–1792 (1995).
4. Andersson, L. S. *et al.* Mutations in DMRT3 affect locomotion in horses and spinal circuit function in mice. *Nature* **488**, 642–646 (2012).
5. Lynch, V. J., May, G. & Wagner, G. P. Regulatory evolution through divergence of a phosphoswitch in the transcription factor CEBPB. *Nature* **480**, 383–386 (2011).
6. Gao, Y. *et al.* The emergence of Sox and POU transcription factors predates the origins of animal stem cells. *Nat. Commun.* **15**, 1–16 (2024).
7. Chothia, C. & Finkelstein, A. V. The classification and origins of protein folding patterns. *Annu. Rev. Biochem.* **59**, 1007–1039 (1990).
8. Lim, W. A. & Sauer, R. T. Alternative packing arrangements in the hydrophobic core of lambda repressor. *Nature* **339**, 31– 36 (05 1989).
9. Metcalf, P., Blum, M., Freymann, D., Turner, M. & Wiley, D. C. Two variant surface glycoproteins of Trypanosoma Brucei of different sequence classes have similar 6 Å resolution X-ray structures. *Nature* **325**, 84–86 (1987).
10. Chin, A. F., Zheng, Y. & Hilser, V. J. Phylogenetic convergence of phase separation and mitotic function in the disordered protein BuGZ. *Protein Sci.* **31**, 822–834 (2022).
11. Beh, L. Y., Colwell, L. J. & Francis, N. J. A core subunit of Polycomb repressive complex 1 is broadly conserved in function but not primary sequence. *Proceedings of the National Academy of Sciences* **109**, E1063–71 (05 2012).
12. Schmidt, H. B., Barreau, A. & Rohatgi, R. Phase separation-deficient TDP43 remains functional in splicing. *Nat. Commun.* **10**, 4890 (2019).
13. Langstein-Skora, I. *et al.* Sequence- and chemical specificity define the functional landscape of intrinsically disordered regions. *bioRxiv* 2022.02.10.480018 (2022) doi:10.1101/2022.02.10.480018.
14. Mindel, V. *et al.* Intrinsically disordered regions of the Msn2 transcription factor encode multiple functions using interwoven sequence grammars. *Nucleic Acids Res.* **52**, 2260–2272 (2024).
15. Sigler, P. B. Transcriptional activation. Acid blobs and negative noodles. *Nature* **333**, 210–212 (05 1988).
16. Staller, M. V. *et al.* A high-throughput mutational scan of an intrinsically disordered acidic transcriptional activation domain. *Cell Syst.* **6**, 444–455.e6 (2018).

17. Kumar, M. *et al.* ELM-the Eukaryotic Linear Motif resource-2024 update. *Nucleic Acids Res.* **52**, D442–D455 (2024).
18. Sanborn, A. L. *et al.* Simple biochemical features underlie transcriptional activation domain diversity and dynamic, fuzzy binding to Mediator. *Elife* **10**, e68068 (2021).
19. Ravarani, C. N. *et al.* High-throughput discovery of functional disordered regions: investigation of transactivation domains. *Mol. Syst. Biol.* **14**, e8190 (2018).
20. Broyles, B. K. *et al.* Activation of gene expression by detergent-like protein domains. *iScience* **24**, 103017 (2021).
21. Erijman, A. *et al.* A High-Throughput Screen for Transcription Activation Domains Reveals Their Sequence Features and Permits Prediction by Deep Learning. *Mol. Cell* **78**, 890–902.e6 (2020).
22. Arnold, C. D. *et al.* A high-throughput method to identify trans-activation domains within transcription factor sequences. *EMBO J.* **37**, e98896 (2018).
23. Morffy, N. *et al.* Identification of plant transcriptional activation domains. *Nature* **632**, 166–173 (2024).
24. Mahatma, S. *et al.* Prediction and functional characterization of transcriptional activation domains. in *2023 57th Annual Conference on Information Sciences and Systems (CISS)* 1–6 (2023).
25. Erkina, T. Y. & Erkine, A. M. Nucleosome distortion as a possible mechanism of transcription activation domain function. *Epigenetics Chromatin* **9**, 40 (2016).
26. Kotha, S. R. & Staller, M. V. Clusters of acidic and hydrophobic residues can predict acidic transcriptional activation domains from protein sequence. *Genetics* **225**, (2023).
27. Udupa, A., Kotha, S. R. & Staller, M. V. Commonly asked questions about transcriptional activation domains. *Curr. Opin. Struct. Biol.* **84**, 102732 (2024).
28. Staller, M. V. *et al.* Directed mutational scanning reveals a balance between acidic and hydrophobic residues in strong human activation domains. *Cell Systems* **13**, 334–345.e5 (2022).
29. Cress, W. D. & Triezenberg, S. J. Critical structural elements of the VP16 transcriptional activation domain. *Science* **251**, 87–90 (01 1991).
30. Shen, F., Triezenberg, S. J., Hensley, P., Porter, D. & Knutson, J. R. Critical amino acids in the transcriptional activation domain of the herpesvirus protein VP16 are solvent-exposed in highly mobile protein segments. An intrinsic fluorescence study. *J. Biol. Chem.* **271**, 4819–4826 (03 1996).
31. DelRosso, N. *et al.* Large-scale mapping and mutagenesis of human transcriptional effector domains. *Nature* (2023) doi:10.1038/s41586-023-05906-y.
32. Alerasool, N., Leng, H., Lin, Z.-Y., Gingras, A.-C. & Taipale, M. Identification and functional characterization of transcriptional activators in human cells. *Mol. Cell* **82**, 677–695.e7 (2022).
33. Kato, S. *et al.* Understanding the function-structure and function-mutation relationships of p53 tumor suppressor protein by high-resolution missense mutation analysis. *Proc. Natl. Acad. Sci. U. S. A.* **100**, 8424– 8429 (07 2003).
34. Sadowski, I., Ma, J., Triezenberg, S. & Ptashne, M. GAL4-VP16 is an unusually potent transcriptional activator. *Nature* **335**, 563–564 (10 1988).
35. Burz, D. S. & Hanes, S. D. Isolation of Mutations that Disrupt Cooperative DNA Binding by the Drosophila Bicoid Protein☆. *J. Mol. Biol.* **305**, 219–230 (2001).
36. Lebrecht, D. *et al.* Bicoid cooperative DNA binding is critical for embryonic patterning in Drosophila. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 13176– 13181 (09 2005).
37. Hummel, N. F. C., Markel, K., Stefani, J., Staller, M. V. & Shih, P. M. Systematic identification of transcriptional activation domains from non-transcription factor proteins in plants and yeast. *Cell Syst* (2024) doi:10.1016/j.cels.2024.05.007.
38. Hummel, N. F. C. *et al.* The trans-regulatory landscape of gene networks in plants. *Cell Syst* **14**, 501–511.e4 (2023).
39. Tsong, A. E., Tuch, B. B., Li, H. & Johnson, A. D. Evolution of alternative transcriptional circuits with identical logic. *Nature* **443**, 415–420 (2006).
40. Lynch, M. The evolution of genetic networks by non-adaptive processes. *Nat. Rev. Genet.* **8**, 803– 813 (2007).
41. Byrne, K. P. & Wolfe, K. H. The Yeast Gene Order Browser: combining curated homology and syntenic context reveals gene fate in polyploid species. *Genome Res.* **15**, 1456–1461 (2005).
42. Bennett, R. J. & Turgeon, B. G. Fungal Sex: The Ascomycota. *Microbiol Spectr* **4**, (2016).
43. Roesgaard, M. A. *et al.* Deciphering the Alphabet of Disorder-Glu and Asp Act Differently on Local but Not Global Properties. *Biomolecules* **12**, (2022).
44. Tuttle, L. M. *et al.* Gcn4-Mediator Specificity Is Mediated by a Large and Dynamic Fuzzy Protein-Protein Complex. *CellReports* **22**, 3251– 3264 (03 2018).

45. Brzovic, P. S. *et al.* The acidic transcription activator Gcn4 binds the mediator subunit Gal11/Med15 using a simple protein interface forming a fuzzy complex. *Mol. Cell* **44**, 942– 953 (12 2011).
46. Scholes, N. S. & Weinzierl, R. O. J. Molecular Dynamics of 'Fuzzy' Transcriptional Activator-Coactivator Interactions. *PLoS Comput. Biol.* **12**, e1004935 (2016).
47. Lotthammer, J. M., Ginell, G. M., Griffith, D., Emenecker, R. J. & Holehouse, A. S. Direct prediction of intrinsically disordered protein conformational properties from sequence. *Nat. Methods* **21**, 465–476 (2024).
48. Shemer, R., Meimoun, A., Holtzman, T. & Kornitzer, D. Regulation of the transcription factor Gcn4 by Pho85 cyclin PCL5. *Mol. Cell. Biol.* **22**, 5395– 5404 (2002).
49. Chi, Y. *et al.* Negative regulation of Gcn4 and Msn2 transcription factors by Srb10 cyclin-dependent kinase. *Genes Dev.* **15**, 1078– 1092 (05 2001).
50. Conti, M. M. *et al.* Phosphosite Scanning reveals a complex phosphorylation code underlying CDK-dependent activation of Hcm1. *Nat. Commun.* **14**, 310 (2023).
51. Raj, N. & Attardi, L. D. The Transactivation Domains of the p53 Protein. *Cold Spring Harb. Perspect. Med.* **7**, a026047–19 (2017).
52. Dyson, H. J. & Wright, P. E. Role of Intrinsic Protein Disorder in the Function and Interactions of the Transcriptional Coactivators CREB-binding Protein (CBP) and p300. *J. Biol. Chem.* **291**, 6714–6722 (2016).
53. Ludwig, C. H. *et al.* High-throughput discovery and characterization of viral transcriptional effectors in human cells. *Cell Syst* **14**, 482–500.e8 (2023).
54. Piskacek, M., Vasku, A., Hajek, R. & Knight, A. Shared structural features of the 9aaTAD family in complex with CBP. *Mol. Biosyst.* **11**, 844– 851 (2015).
55. Warfield, L., Tuttle, L. M., Pacheco, D., Klevit, R. E. & Hahn, S. A sequence-specific transcription activator motif and powerful synthetic variants that bind Mediator using a fuzzy protein interface. *Proceedings of the National Academy of Sciences* **111**, E3506– E3513 (08 2014).
56. Pacheco, D. *et al.* Transcription activation domains of the yeast factors Met4 and Ino2: tandem activation domains with properties similar to the yeast Gcn4 activator. *Mol. Cell. Biol.* MCB.00038–18 – 39 (03 2018).
57. Tuttle, L. M. *et al.* Mediator subunit Med15 dictates the conserved 'fuzzy' binding mechanism of yeast transcription activators Gal4 and Gcn4. *Nat. Commun.* **12**, 1–11 (2021).
58. Schuler, B. *et al.* Binding without folding - the biomolecular function of disordered polyelectrolyte complexes. *Curr. Opin. Struct. Biol.* **60**, 66–76 (2020).
59. Dunker, A. K., Bondos, S. E., Huang, F. & Oldfield, C. J. Intrinsically disordered proteins and multicellular organisms. *Semin. Cell Dev. Biol.* **37**, 44–55 (2015).
60. Tenthorey, J. L., Young, C., Sodeinde, A., Emerman, M. & Malik, H. S. Mutational resilience of antiviral restriction favors primate TRIM5α in host-virus evolutionary arms races. *Elife* **9**, (2020).
61. Koonin, E. V. & Dolja, V. V. A virocentric perspective on the evolution of life. *Curr. Opin. Virol.* **3**, 546–557 (2013).
62. Dalal, C. K. & Johnson, A. D. How transcription circuits explore alternative architectures while maintaining overall circuit output. *Genes Dev.* **31**, 1397–1405 (2017).
63. Fowler, K. R., Leon, F. & Johnson, A. D. Ancient transcriptional regulators can easily evolve new pair-wise cooperativity. *Proc. Natl. Acad. Sci. U. S. A.* **120**, e2302445120 (2023).
64. Liu, Y. *et al.* Evolution of the activation domain in a Hox transcription factor. *Int. J. Dev. Biol.* **62**, 745– 753 (2018).
65. Zarin, T. *et al.* Proteome-wide signatures of function in highly diverged intrinsically disordered regions. *eLife* **xx**, xxx–45 (03 2019).
66. Zarin, T. *et al.* Identifying molecular features that are associated with biological function of intrinsically disordered protein regions. *Elife* **10**, e60220 (2021).
67. Zarin, T., Tsai, C. N., Ba, A. N. N. & Moses, A. M. Selection maintains signaling function of a highly diverged intrinsically disordered region. *Proc. Natl. Acad. Sci. U. S. A.* **114**, E1450–E1459 (2017).
68. Parker, M. W. *et al.* A new class of disordered elements controls DNA replication through initiator self-assembly. *Elife* **8**, e48562 (2019).
69. Parker, M. W., Kao, J. A., Huang, A., Berger, J. M. & Botchan, M. R. Molecular determinants of phase separation for Drosophila DNA replication licensing factors. *Elife* **10**, (2021).
70. Davey, N. E., Cyert, M. S. & Moses, A. M. Short linear motifs - ex nihilo evolution of protein regulation. *Cell Commun. Signal.* **13**, 43 (2015).
71. Wong, E. S. *et al.* Deep conservation of the enhancer regulatory code in animals. *Science* **370**, (2020).

72. Ludwig, M. Z., Bergman, C., Patel, N. H. & Kreitman, M. Evidence for stabilizing selection in a eukaryotic enhancer element. *Nature* **403**, 564–567 (2000).
73. Ludwig, M. Z. *et al.* Functional evolution of a cis-regulatory module. *PLoS Biol.* **3**, e93 (2005).
74. Hare, E. E., Peterson, B. K., Iyer, V. N., Meier, R. & Eisen, M. B. Sepsid even-skipped Enhancers Are Functionally Conserved in Drosophila Despite Lack of Sequence Conservation. *PLoS Genet.* **4**, e1000106 (2008).
75. Peterson, B. K. *et al.* Big genomes facilitate the comparative identification of regulatory elements. *PLoS One* **4**, e4688 (2009).
76. Kaplow, I. M. *et al.* Relating enhancer genetic variation across mammals to complex phenotypes using machine learning. *Science* **380**, eabm7993 (2023).
77. Arnosti, D. N. & Kulkarni, M. M. Transcriptional enhancers: Intelligent enhanceosomes or flexible billboards? *J. Cell. Biochem.* **94**, 890– 898 (2005).
78. Jackson, B. M., Drysdale, C. M., Natarajan, K. & Hinnebusch, A. G. Identification of seven hydrophobic clusters in GCN4 making redundant contributions to transcriptional activation. *Mol. Cell. Biol.* **16**, 5557–5571 (1996).
79. Ginell, G. M., Emenecker, R. J., Lotthammer, J. M., Usher, E. T. & Holehouse, A. S. Direct prediction of intermolecular interactions driven by disordered regions. *bioRxivorg* 2024.06.03.597104 (2024).
80. Amberg, D. C., Burke, D. & Strathern, J. N. *Methods in Yeast Genetics: A Cold Spring Harbor Laboratory Course Manual*. (CSHL Press, 2005).
81. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc. Series B Stat. Methodol.* **57**, 289–300 (1995).
82. Dunn, O. J. Multiple Comparisons among Means. *J. Am. Stat. Assoc.* **56**, 52–64 (1961).
83. Das, R. K. & Pappu, R. V. Conformations of intrinsically disordered proteins are influenced by linear sequence distributions of oppositely charged residues. *Proceedings of the National Academy of Sciences* **110**, 13392–13397 (08 2013).
84. Ginell, G. M. & Holehouse, A. S. Intrinsically Disordered Proteins, Methods and Protocols. *Methods Mol. Biol.* **2141**, 103–126 (2020).
85. Martin, E. W. *et al.* Sequence determinants of the conformational properties of an intrinsically disordered protein prior to and upon multisite phosphorylation. *J. Am. Chem. Soc.* **138**, 15323–15335 (2016).
86. Emenecker, R. J., Griffith, D. & Holehouse, A. S. Metapredict V2: An update to metapredict, a fast, accurate, and easy-to-use predictor of consensus disorder and structure. *bioRxiv* 2022.06.06.494887 (2022) doi:10.1101/2022.06.06.494887.
87. Hope, I. A., Mahadevan, S. & Struhl, K. Structural and functional characterization of the short acidic transcriptional activation region of yeast GCN4 protein. **333**, 635– 640 (06 1988).
88. Drysdale, C. M. *et al.* The transcriptional activator GCN4 contains multiple activation domains that are critically dependent on hydrophobic amino acids. *Mol. Cell. Biol.* **15**, 1220–1233 (1995).