

Title:

Directed mutational scanning reveals a balance between acidic and hydrophobic residues in strong human activation domains

Authors:

Max V. Staller^{1,2,*}, Eddie Ramirez¹, Sanjana R. Kotha², Alex S. Holehouse^{3,4}, Rohit V. Pappu^{4,5}, Barak A. Cohen^{1,6,*}

Affiliations:

¹ Edison Family Center for Genome Sciences and Systems Biology & Department of Genetics, Washington University in St. Louis School of Medicine, Saint Louis, MO

² Center for Computational Biology, University of California, Berkeley, Berkeley, CA, USA

³ Department of Biochemistry and Molecular Biophysics, Washington University in St. Louis School of Medicine, Saint Louis, MO

⁴ Center for Science and Engineering of Living Systems Washington University in St. Louis, St. Louis, MO, USA

⁵ Department of Biomedical Engineering, Washington University in St. Louis, St. Louis, MO, USA

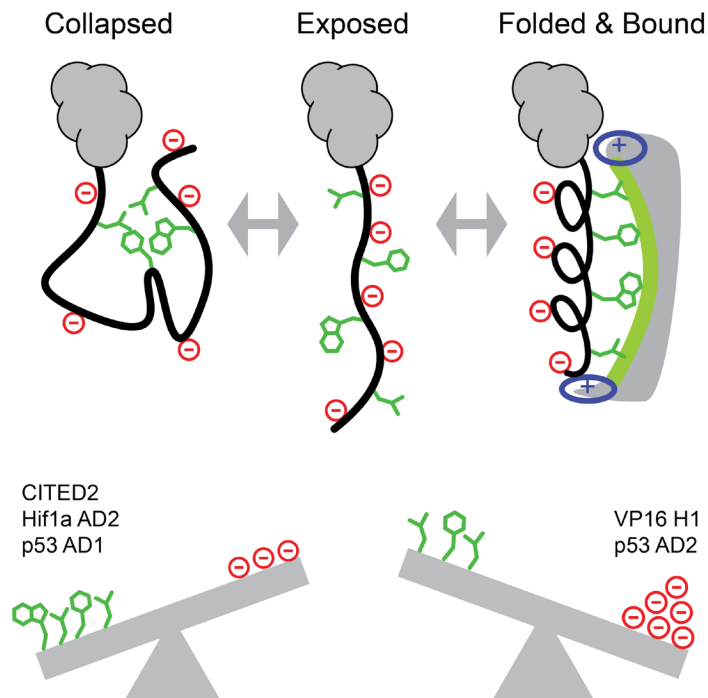
* Correspondence: mstaller@berkeley.edu, cohen@wustl.edu

⁶ Lead contact

Abstract

Acidic activation domains are intrinsically disordered regions of transcription factors that bind coactivators. The intrinsic disorder and low evolutionary conservation of activation domains have made it difficult to identify the sequence features controlling activity. To address this problem, we designed thousands of variants in seven acidic activation domains and measured their activities with a high-throughput assay in human cell culture. We found that strong activation domain activity required a balance between the number of acidic residues and aromatic and leucine residues. These findings motivated a predictor of acidic activation domains that scans the human proteome for clusters of aromatic and leucine residues embedded in regions of high acidity. This predictor identifies known activation domains and accurately predicts previously unidentified ones. Our results support a flexible, Acidic Exposure Model of activation domains in which acidic residues solubilize hydrophobic motifs so that they can interact with coactivators.

Key words: Transcription factor, activation domain, intrinsically disordered protein



Graphical Abstract:

- A high-throughput assay quantifies the activities of activation domain variants in human cells
- Strong activation domains require both acidic and hydrophobic residues
- The combination of acidic and hydrophobic residues predicts new activation domains

Introduction

Transcription factors (TFs) activate gene expression using DNA binding domains (DBDs) and activation domains (ADs). DBDs are structured, evolutionarily conserved and bind related DNA sequences (Latchman, 2008). ADs are intrinsically disordered, poorly conserved, and bind structurally diverse coactivator subunits (Dyson and Wright, 2016). Bioinformatics tools can predict DBDs from protein sequence, but there are few tools for predicting ADs (El-Gebali et al., 2019; Finn et al., 2016). When a new genome is sequenced, scanning for DBDs can predict candidate TFs, but it is not possible to predict which candidate TFs contain ADs.

Predicting ADs from amino acid sequence has been difficult for five reasons: 1) ADs have diverse primary sequences (Latchman, 2008), 2) ADs have poor sequence conservation that hinders comparative genomics, 3) ADs are intrinsically disordered and have diverse modes of binding coactivators (Dyson and Wright, 2016), 4) until recently, measuring AD activity has been low throughput, and 5) the key sequence properties that control AD activity remain unresolved. Many ADs are acidic (have a net negative charge), but site-directed mutagenesis has shown that clusters of hydrophobic residues, called motifs, make the largest contributions to activity (Cress and Triezenberg, 1991; Dyson and Wright, 2016; Warfield et al., 2014). Here, we test the hypothesis that ADs are composed of hydrophobic motifs surrounded by an acidic context.

Based on our work in yeast (Staller et al., 2018), we developed an *Acidic Exposure Model* for AD function: acidity and intrinsic disorder keep hydrophobic motifs exposed to solvent where they are available to bind coactivators (Figure 1A). Hydrophobic residues tend to interact with each other and drive intramolecular chain collapse, suppressing interactions with coactivators. Surrounding the hydrophobic residues with acidic residues that repel one another exposes the motifs to solvent, promoting interactions with coactivators. For example, in the VP16 AD, the critical F442 is highly exposed to solvent in solution, but exposure decreases

upon coactivator binding (Shen et al., 1996a, 1996b). Three recent papers in yeast (Erijman et al., 2020; Ravarani et al., 2018; Sanborn et al., 2021) also found that strong ADs contain both acidic and hydrophobic residues, which supports the Acidic Exposure Model. However, whether this model can explain the properties of human ADs remains unknown.

Here, we show that the Acidic Exposure Model extends from yeast to human cells. We introduce a high-throughput reporter system to test more than 3500 variants in seven ADs. We designed these variants to interrogate two aspects of the Acidic Exposure Model: acidic residues and aromatic residues. We found that strong ADs balance the number of acidic residues against the number of aromatic and leucine residues. Based on these results, we found that scanning the proteome for clusters of eight amino acids (acidic, basic, aromatic and leucine residues) was sufficient to accurately predict new and known ADs. Taken together, our results suggest that the Acidic Exposure Model may be a general explanation for the function of eukaryotic ADs from yeast to humans and provide a framework for unifying the roles of acidity, hydrophobicity, and intrinsic disorder in acidic ADs.

Results

To test the Acidic Exposure Model, we developed a high-throughput method to assay AD variants in parallel in human cell culture (Figure 1B). We engineered a cell culture system with a synthetic TF that binds and activates a genome-integrated GFP reporter. Each cell receives one AD variant marked by a unique DNA barcode integrated into the same genomic “landing pad” with CRE recombinase and asymmetric loxP sites. The landing pad equalizes the effects of genomic position on expression (Maricque et al., 2018). The synthetic TF contains an mScarlet red fluorescent protein for measuring abundance, but after trying four different red fluorescent proteins, each with low signal, we did not normalize for protein abundance in this

study. To avoid cell toxicity, the synthetic TF contains an engineered DBD (Park et al., 2019) and an estrogen response domain for inducible nuclear localization (McIsaac et al., 2013). AD variants that drive different levels of GFP expression are separated by Fluorescent Activated Cell Sorting (FACS), and the barcodes in each sorted pool are counted by deep sequencing (Kinney et al., 2010; Sharon et al., 2012; Staller et al., 2018). We used the barcode counts to compute a probability mass function for each AD across the four pools and the GFP signal of each pool to compute a weighted average GFP signal. The assay is reproducible (average Pearson correlation between replicates of 0.69) and recapitulates the activity of known mutations in human ADs (Figure S1). A synthetic TF without an AD (No AD control) was used to define baseline activity in our assay. In the library, the ADs are cloned into the N terminus of the synthetic TF, between the ATG start codon (M) and a GSGS linker. In the No AD control plasmid, nothing is between the initial M and the GSGS linker. We combined biological replicates by normalizing the activity of the No AD control to 2000 (arbitrary fluorescence units, AU, methods) and averaging together the fluorescence values.

In our first experiment, we performed Deep Mutational Scans (DMS, where every position is mutated to all 19 other residues) and rational mutagenesis on the two ADs of the tumor suppressor TF, p53. This library contained 2991 variants, each paired with 5 barcodes. After extensive analysis, we determined that most point substitutions had small effects on activity and that five barcodes was not sufficient to resolve these small changes in activity. DMS has been very informative for structured proteins (Gray et al., 2017) but not for intrinsically disordered regions, where most point substitutions do not cause measurable changes in activity (Giacomelli et al., 2018; Majithia et al., 2016). In the rational mutagenesis, we introduced multiple substitutions to test the roles of acidic residues, aromatic residues and intrinsic disorder. These perturbations had large effects on activity that could be resolved with five barcodes (Figure S1, S2).

In our second experiment, we examined 525 rationally designed variants of five ADs, each tagged with twenty-eight barcodes. Using more barcodes allowed us to resolve smaller changes in activity and assaying mutations with larger effect sizes increased measurement reproducibility (Figure S1). We focus the main text on these high quality data and use the noisier p53 data to corroborate trends.

Using this assay, we investigated three key features of acidic ADs: acidic residues, hydrophobic motifs and disorder-to-order transitions (Figure 1C). We designed sequence variants that systematically added and subtracted acidic residues or aromatic residues in seven ADs: VP16 (H1 region, 415-453), Hif1 α (AD2, 781-896), CITED2 (220-258), Stat3 (719-764), p65 (AD2, 521-551), p53 AD1 (1-40) and p53 AD2 (40-60) (Berlow et al., 2017; Brady et al., 2011; Lecoq et al., 2017; Raj and Attardi, 2017; Regier et al., 1993; Vogel et al., 2015; Wojciak et al., 2009). Most variants had five or fewer substitutions (Figure 1D). For each AD, we hand designed 6-10 'supercharge' variants that added aromatic residues next to existing acidic residues and added acidic residues next to existing aromatic residues. For each disordered region that folds into an alpha helix upon coactivator binding, we introduced proline or glycine residues to break these helices (Figure S3). The complete list of substitutions and activities are located in Dataset 1 (VP16, Hif1 α , CITED2, Stat3, p65. 525 variants, 28 barcodes per variant, and 4 replicates) and Dataset 2 (p53 AD1, p53 AD2. 2991 variants, 5 barcodes per variant and 3 replicates). Note that the activity values in the two experiments are not comparable because they were collected on different cell sorters and normalized differently.

Compared to the No AD control, all ADs activated the GFP reporter (Figure 2, S2). For each AD, we identified variants that significantly changed activity after correcting for multiple hypotheses (two sided t-test and 5% FDR, Dataset 1). For VP16, CITED2 and Hif1 α we recovered variants that increase or decrease their activities (Dataset 1, Figure 2A). p65 and Stat3 are weakly active in the assay, reducing our sensitivity, and none of these variants significantly changed activity after correcting for multiple hypotheses (Dataset 1). Either our

assay is not sensitive enough to interrogate these two ADs or the residues we mutated made small contributions to activity.

Hydrophobic motifs are necessary for AD activity

We confirmed that hydrophobic motifs make large contributions to AD activity. We included published motifs (LPEL in CITED2, LPQL and LLxxL in Hif1 α , and LxxF $\times$ L in VP16 (Berlow et al., 2017; Regier et al., 1993)) and predicted additional motifs by looking for clusters of W,F,Y,L,M residues. Substituting all the residues that comprise a motif with alanine residues decreased activity (Figure 2B, S2). In CITED2 and VP16, every cluster of aromatic and leucine residues tested contributed to activity.

Acidic residues are necessary for AD activity

We systematically increased and decreased the net negative charge of each AD and plotted the resulting activities (Figure 3A, S2). For CITED2, Hif1 α , VP15, p53 AD1 and p53 AD2, acidic residues were necessary for full activity and regressing activity against net charge had significant negative slopes (Figure 3A, S2C, S4A. For CITED2, Hif1 α , and VP16, slopes were significant when using the charge variants or all variants). Removing negatively charged residues (D,E) had similar effects to adding positively charged (K,R) residues and vice versa (Figure S5), suggesting that net charge and not residue identity is the key parameter.

For ADs with moderate acidity (CITED2, Hif1 α and p53 AD1) adding acidic residues increased activity in the majority of variants (Figure 3A). For p53 AD1, this effect mirrors how phosphorylation increases activity (Raj and Attardi, 2017). For the more acidic p53 AD2, adding acidic residues rarely increased activity; for the most acidic AD, VP16, adding acidic residues never increased activity. Thus, the starting net charge of the wild type AD determined whether it was possible to increase activity by adding acidic residues.

For CITED2 and Hif1 α , adding acidic residues could either increase or decrease activity (Figure 3A: red vs blue) depending on the location of the substitution (Figure 3B). For CITED2, variants with increased activity (Figure 3A, red) frequently added acidic residues in the flanks, near the hydrophobic motifs (Figure 3B, 3C, red), while variants with decreased activity frequently removed the positive residues in the center of the AD (Figure 3B, 3C blue). For Hif1 α , variants with increased activity were more likely to add acidic residues in the C-terminus near L812, L813 or L819 or to remove R820. These data suggest that the location of added acidic residues can determine how they modulate activity. This result agrees with our work in yeast and two random peptide screens which found that [DE][WFY] dipeptides make large contributions to AD activity (Erijman et al., 2020; Ravarani et al., 2018; Staller et al., 2018). To further test this idea, we used the Omega statistic to quantify how the mixture of aromatic and leucine (W,F,Y,L) residues with acidic residues (D,E) related to activity (Martin et al., 2016). We found a modest correlation: variants with more evenly mixed (W,F,Y,L) and (D,E) (i.e. low Omega values) had higher activities (Figure S4D). Together with the literature, our data support the idea that acidic residues near key hydrophobic motifs boost activity. For VP16, we could not increase activity by adding acidic residues, perhaps because 5/7 residues in the motifs are already adjacent to acidic residues. VP16 is the most acidic AD we examined and appears to be saturated for the effect of negative charge on activity.

Context dependent effects of adding acidic or aromatic residues

When we systematically added and removed aromatic residues we saw expected and unexpected changes in AD activity. Based on experiments in human cell culture with the VP16, p53 and ETV ADs, we expected aromatic residues to be critical for activity, and that adding aromatic residues would increase activity (Currie et al., 2017; Raj and Attardi, 2017; Regier et al., 1993). The VP16 variants generally matched this expectation: any substitution in F442 (A,L,W,Y) decreased activity. Furthermore, adding up to four aromatic residues increased

activity in the majority of cases (Figure 2B, 4A, S6C). Similarly, for both p53 ADs removing aromatics decreased activity and adding aromatic residues increased activity (Figure S2D).

In CITED2, adding and removing aromatic residues did not yield the expected results. Mutating the aromatic residues to alanine led to small decreases in activity, and mutating aromatic residues to leucine residues caused small increases in activity (Figure 4A, S6A). Adding aromatic residues decreased activity for all but one variant. If instead, we plot activity against the number of W,F,Y,L residues, CITED2 activity peaks at the WT number, 10 (Figure 4B, S4C).

In Hif1 α , adding or removing aromatic residues generally did not change activity. Mutating the lone Y to L caused a small, and not significant, increase in activity. Adding aromatic residues to Hif1 α caused small, and frequently not significant, decreases in activity (Figure S4B, S6B).

We saw two responses to adding acidic residues and two responses to adding aromatic residues. For the moderately acidic ADs CITED2 and Hif1 α , we could increase activity by adding acidic residues and, for CITED2, decrease activity by adding aromatic residues. For the more acidic ADs VP16 and p53 AD2, we could increase activity by adding aromatic residues but not by adding acidic residues. Even for VP16, adding more than 4 aromatic residues always decreases activity, suggesting that for all ADs there is a regime where adding aromatic residues will eventually decrease activity.

The Acidic Exposure Model can explain why the two responses to adding acidic residues mirror the opposite responses to adding aromatic residues. The model predicts that adding acidic residues will increase AD activity only when there are hydrophobic motifs that can be further exposed. Once the hydrophobic motifs are maximally exposed, adding more acidic residues will not increase activity. In contrast, adding more aromatic residues can increase activity only when there is excess acidity to keep these added residues exposed. Adding too many aromatic residues eventually reduces activity because they overwhelm the acidic residues

and drive collapse. Thus, a prediction of the Acidic Exposure Model is that acidic residues promote expanded AD conformations while aromatic residues promote chain collapse. We tested this prediction with all-atom Monte Carlo simulations of the VP16 and CITED2 variants (Methods) (Staller et al., 2018; Vitalis and Pappu, 2009) and calculated the radius of gyration, which captures the size of the conformational ensemble. Although the dispersion in the predicted radius of gyration is large for any given net charge, we found that adding acidic residues increased the radius of gyration, consistent with expansion (Figure S7A) and adding aromatic residues decreased the radius of gyration, consistent with partial collapse (Figure S7B). These trends hold for the supercharge variants that add both aromatic and acidic residues (Figure S7C).

Leucine residues are critical for AD activity

We found that leucine residues made large contributions to activity. In yeast ADs, aromatic residues contribute more to activity than smaller hydrophobic residues like leucine and methionine (Erijman et al., 2020; Jackson et al., 1996; Ravarani et al., 2018; Staller et al., 2018). In human cells, VP16 and both p53 ADs fit this pattern: substituting aromatic residues decreased activity (Figure 2B, 4A, S2, S4) (Cress and Triezenberg, 1991; Lin et al., 1994). However, in VP16, 'motif 2' contains only leucine residues and is necessary for full AD activity (Figure 2C). In CITED2, summarizing the activities of all substitutions at each position reveals that leucine residues make the largest contributions to activity, followed by the acidic residues (Figure 5A). Similarly, for VP16 aggregating the data by position shows that the key positions are F442, the leucine residues and the acidic residues (Figure S8A). Acidic and leucine residues make large contributions to activity in these ADs.

The mechanism by which leucine residues make large contributions to activity is exemplified by the structure of the CITED2 interaction with TAZ1. TAZ1 has a canyon with a hydrophobic floor and basic rim that tightly embraces the compact alpha helix of CITED2

(Figure 7B). The leucine residues on CITED2 interact with the hydrophobic canyon floor and the acidic residues interact with the basic canyon rim. This tight structural constraint explains the activities of many variants. The positions where mutations cause large decreases in activity in Figure 5A point towards the coactivator surface in the NMR structure of CITED2 bound to the TAZ1 domain of CBP/p300 (Figure 5B) Hif1 α showed a similar pattern (Figure 7B, 7C). Replacing leucines with aromatics reduces activity because the larger side chains do not fit in the canyon (Figure 5B,E). Disrupting the helix folding by adding two proline residues reduces activity, because a helix is very compact and the unfolded peptide likely does not fit (Figure 5B, S9). Adding two glycine residues does not disrupt activity because they do not disrupt helix formation and they are very small (Figure S9). Finally, the D244E substitution reduces activity because, the D224 acidic side chain (negative) sits between the narrowest point of the basic canyon rim, sandwiched between the basic (positive) side chains R439 and K365 of TAZ1, and replacing D244 with the larger glutamic acid residue impairs this fit (Figure 5C, S10). Overall, mutations that increase the size of side chains decrease activity because they impede the helix from fitting into the narrow canyon on TAZ1.

Strong ADs balance hydrophobic and acidic residues

We found that AD activity requires a combination of aromatic and leucine (W,F,Y,L) residues and acidic residues. Plotting the number of W,F,Y,L residues against net charge separates high and low activity variants (Figure 6A, S11A). This separation is less apparent when we count only aromatic residues (Figure S11B), and somewhat visible when we use calculated Kyte Doolittle hydropathy (Figure S11C). Many points on this grid contain both strong and weak variants (Figure S11A), indicating that composition is not the sole determinant of activity and that the arrangement of residues also matters. We found that composition-based machine learning classifiers could separate active and inactive variants (Figure S11). When we removed individual parameters from the model we found that net charge and leucine residues

made the largest contributions to model performance (Table S1). Our results suggest that the balance between W,F,Y,L and acidic residues is critical for AD activity.

Predicting ADs

We examined whether the balance of acidic and W,F,Y,L residues could predict ADs in human TFs. For a third of human TFs, the only annotated domain is the DNA binding domain (Lambert et al., 2018) and only 8% of TFs have an AD annotated in Uniprot (Methods). *In silico*, we broke the protein sequences of 1608 TFs (Lambert et al., 2018) into 39 residue tiling windows (“tiles”) spaced every one amino acid. For each tile we calculated the net charge and counted W,F,Y,L residues. We plotted the joint distribution of these properties as a heatmap (Figure 6B, blue) and found that VP16 and CITED2 are on the periphery. Tiles that have both the net charge and hydrophobicity of these ADs are rare: only 0.02% and 0.03% of tiles were as extreme or more extreme than VP16 or CITED2, respectively. Interpolating between these ADs yields 0.13% of tiles ($n = 1139$, Figure 6B, red), which combine to predict 144 ADs from 136 TFs (Dataset 3). These predicted regions overlap with 17 Uniprot ADs – far more than expected by chance ($p < 1e-5$ in permutation tests). In addition, 11 predicted regions overlap 10 published ADs that are not in Uniprot ($p < 1e-5$ in permutation tests), including the N-terminal AD of c-Myc (Andresen et al., 2012) and the Zn473 KRAB domain (Tycko et al., 2020). The predictor requires the combination of net negative charge and W,F,Y,L residues because neither property alone is sufficient for specific predictions: using only net charge (≤ -9 from CITED2) yielded 18086 tiles that combine to 856 predictions, 30 of which overlap Uniprot ADs; using only the W,F,Y,L count (≥ 7 from VP16) yielded 302161 tiles that combine to 3411 predictions, 99 of which overlap Uniprot ADs. The high degree of overlap between our predicted regions and literature-validated ADs motivated us to test the predictions experimentally.

Testing predicted ADs

We tested the predicted regions and found that our composition model accurately predicted ADs in the human proteome (Figure 6C). In a new experiment, we designed a library with 150 predicted regions (we split long regions to meet synthesis limits), 150 length-matched random regions and 94 published ADs (methods). We did not allow the random regions to overlap predicted regions or Uniprot ADs. We recovered 149 predicted regions, 146 random regions and 78 published ADs (Dataset 4). In this dataset, we normalized the three replicates with the No AD control (set to 200 GFP AU). Note, the activity values are not directly comparable to the 5 AD experiment above because the FACS was performed on a different day. Using the No AD TF as threshold for activity, 108/149 (72%) of predicted ADs are active, 75/89 (84%) of published ADs are active, and 55/149 (38%) of random regions are active. As a threshold for strong AD activity, we chose the 95th percentile of the random regions (221 AU). At this high threshold, 58/149 (39%) of our predicted regions are strong ADs and 52/89 (58%) of published ADs are highly active. Although we do not expect all published ADs will work in our assay, because some ADs have promoter-specific activities (Goodrich and Tjian, 2010), this analysis demonstrates that our predictor identifies known ADs and accurately predicts previously unidentified ADs.

Discussion

The critical feature of strong activation domains is a balance between hydrophobic and acidic residues. Both types of residues are necessary, neither is sufficient, and too much of either decreases activity. Hydrophobic residues make the critical contacts with coactivators (Dyson and Wright, 2016). Acidic residues can have long range interactions with basic residues on coactivators (Ferreira et al., 2005; Hermann et al., 2001; Jonker et al., 2005), but these interactions cannot explain the balance requirement. We argue that balance between hydrophobic and acidic residues can be explained by the Acidic Exposure Model.

In the Acidic Exposure Model, acidic residues and intrinsic disorder keep hydrophobic motifs exposed to solvent where they are available to bind coactivators. Acidic residues prevent local chain compaction through electrostatic repulsion and favorable free energies of solvation. This expansion exposes leucine and aromatic residues to the solvent, so they are available to interact with cofactors. Intrinsic disorder reduces the entropic cost of organizing water around solvent exposed hydrophobic residues because fluctuating between solvent exposed and solvent protected conformations lowers the average cost compared to constant exposure. The Acidic Exposure Model explains why activation domains are both negatively charged and intrinsically disordered: the acidic residues and intrinsic disorder combine to keep aromatic and leucine-rich motifs exposed and available to bind coactivators.

Promoting the exposure of hydrophobic motifs is compatible with the other known functions of acidity and disorder in ADs. In specific cases, intramolecular electrostatic interactions between negatively charged ADs and positively charged DBDs can increase DNA binding specificity (Krois et al., 2018; Liu et al., 2008). We speculate that acidic residues may also reduce non-specific DNA binding by repelling the negatively charged DNA backbone. Intrinsic disorder gives ADs the flexibility to fold into different conformations when bound to different coactivators (Dyson and Wright, 2016). Intrinsic disorder and acidic residues together

can also increase the fraction of molecular collisions that lead to productive coactivator binding by enabling multiple folding trajectories (Kim and Chung, 2020). The Acidic Exposure Model is compatible with these biophysical properties of AD-coactivator interactions.

A key property of ADs is the balance between the strength of their hydrophobic binding motifs and their capacity to keep those motifs exposed to solvent. Adding more hydrophobic residues to ADs will increase their activity so long as their intrinsic disorder and acidic residues can keep the excess hydrophobicity from collapsing the amino acid chain into an inactive conformation. ADs have high hydrophobicity and high acidity and their activity requires a balance between these two physical properties. We exploited this observation to create an AD predictor that scans for a high, but balanced composition of acidity and hydrophobicity. This predictor can be used to prioritize candidate acidic ADs on poorly characterized TFs in any metazoan genome.

Not all hydrophobic residues make an equal contribution to AD activity. In yeast aromatic residues make the largest contributions to activity (Erijman et al., 2020; Ravarani et al., 2018; Sanborn et al., 2021; Staller et al., 2018) and, here, in human cells we find that leucine residues make large contributions to activity. In human cells and yeast, valine and isoleucine (V,I), do not make large contributions to activity, which explains why the choice of hydrophobicity table determines if AD activity is correlated with hydrophobicity: when we used the Kyte-Doolittle hydropathy table, we found no correlation between activity and hydrophobicity, because on this table, V and I have large values while W has and Y have small values (Kyte and Doolittle, 1982; Staller et al., 2018). In contrast, Sanborn et al. chose the Wimley-White hydrophobicity table which perfectly matches the order of residue contributions to AD activity, leading to a correlation between activity and hydrophobicity (Sanborn et al., 2021; Wimley and White, 1996). These are two among more than 27 published tables, each of which remains an approximation (Colwell et al., 2010). These results further emphasize that I-rich ADs in *Drosophila* are a distinct functional class (Attardi and Tjian, 1993).

There is accumulating evidence that hydrophobic residues in ADs do not always need to be organized into motifs with strict arrangements (sequence grammar). When we designed our mutations, the dominant model was that ADs had motifs surrounded by an acidic context. For example, some ADs contain Φ xx $\Phi\Phi$ motifs (where Φ is a hydrophobic residue) in an amphipathic alpha helix that presents the hydrophobic residues as a continuous surface to the coactivator, but this is one solution among many. In yeast, screens of random peptide and extant TFs have failed to find enriched motifs longer than two residues and have shown that some ADs behave as 'bags of amino acids' that can be scrambled with minimal loss of activity (Erijman et al., 2020; Ravarani et al., 2018; Sanborn et al., 2021). For the ADs identified by our predictor, we did not see signatures of grammar. The success of our composition-based predictor, which has no grammar requirement, is evidence for very flexible grammar. We speculate that hydrophobic residues in ADs may simply need to be clustered and not arranged in motifs with specific spacing grammar.

At the same time, we found that within some ADs, there are very strong constraints that reflect the structural constraints of the AD-coactivator interface. Variants that shuffle AD sequences abolish activity, evident for some grammar. Within a hydrophobic motif, the presence of aromatic or leucine residues reflects the structural constraints in AD-coactivator interaction surfaces, which looks like a strong grammar (Figure 5). Contrasting the CITED2-TAZ1 interaction with the Gcn4-Med15 interaction explains why aromatic residues make large contributions to activity in while leucines predominate in CITED2 Gcn4 (Berlow et al., 2017; Brzovic et al., 2011). Both ADs fold into alpha helices and both coactivators contain a binding canyon with a hydrophobic floor and basic rim (Figure 7A). On TAZ1, the canyon is large and the CITED2 alpha helix is engulfed (Figure 7B). Leucines fit this structure better than aromatics because they are smaller and promote folding into a compact helix (Pace and Scholtz, 1998). On Med15, the canyon is shallow and Gcn4 only inserts side chains. A recent structure of the Gal4-Med15 binding interaction shows a similar fuzzy interaction centered on aromatic and

leucine residues (Pacheco et al., 2018; Tuttle et al., 2021). Aromatics fit the Med15 binding surface better than leucine residues because they more easily reach the hydrophobic canyon floor. The increased importance of leucine residues in human ADs such as CITED2 likely reflects the structural constraints imposed by an expanded repertoire of coactivators. Going forward, new approaches to high-throughput mutagenesis will be efficient methods for exploring the structural constraints of protein-protein interaction surfaces (Diss and Lehner, 2018; Rollins et al., 2019; Schmiedel and Lehner, 2019).

The Acidic Exposure Model can explain several results in the literature. Screens of random peptides found enrichment of [DE][WIFY] “mini motifs,” which support our model (Erijman et al., 2020; Ravarani et al., 2018). Sanborn et al, 2021 examined synthetic 9-mer peptides that mixed aromatic and D residues and observed an increase and decrease in activity as aromatics are added (Sanborn et al., 2021). Peak activity occurs when the D’s and F’s are well mixed, or when the F’s are on the C-terminus of the TF, both of which promote F exposure and activity. Sanborn et al. screened the ability of diverse sequences to modulate the activity of the Pdr1 AD and found that hydrophobic residues decreased activity and acidic residues boosted activity. This modulation is consistent with hydrophobic residues promoting collapse, and acidic residues promoting exposure. Balanced sequences are the most active.

We synthesize our findings in three conclusions: strong acidic activation domains balance hydrophobic motifs and acidic residues; clusters of W,F,Y,L residues surrounded by acidic residues are sufficient to predict new ADs; and the choice between aromatic and leucine residues in an acidic AD is constrained by the structure of the coactivator interaction surface. These rules apply to a subset of traditional acidic ADs, and our work implies there are multiple subclasses of acidic ADs. These insights will help refine computational models for predicting ADs, guide engineering of ADs, and inform models that predict the impact of genetic variation on AD function.

Acknowledgments

We thank Minhee Park and Ahmed Khalil for sharing the synthetic DBD and promoter sequence ahead of publication; Kiersten Ruff, Avi Ramu and Nicole Rockweiler for bioinformatics help; Brittany Pioso for help with the cartoons; Jessica Hoisington-Lopez and MariaLynn Crosby for DNA sequencing. We thank members of the Cohen Lab and Thomas Graham for helpful discussions and comments on the manuscript.

Author contributions: MVS and BAC designed the project and wrote the manuscript. MVS and ER collected data. MVS, ER, SRK and ASH analyzed data. RVP and BAC interpreted data. All authors edited the manuscript. The authors declare no competing interests.

Funding: MVS was supported by Burroughs Wellcome Fund Postdoctoral Enrichment Program, American Cancer Society Postdoctoral Fellowship, and NIGMS K99131022. ER was supported by the McDonnell Genome Institute Opportunities in Genomics Research Program under grant NIH-R25HG006687. This work was supported by grants from the National Institutes of Health, NINDS 5R01NS056114 to RVP and NIGMS R01GM092910 to BAC, and the from the Children's Discovery Institute CDI-LI-2018-765 to BAC.

Competing interests: Authors declare no competing interests.

STAR Methods

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to Barak Cohen, cohen@wustl.edu

Materials availability

The plasmids generated in this study have been deposited at AddGene.
The cell lines generated in this study are available upon request.

Data and code availability

The sequencing data have been deposited at NIH GEO and are publicly available as of the date of publication. Accession numbers are listed in the key resources table.

The analysis code has been deposited in Github and is publicly available as of the date of publication. See Key Resources Table.

The raw simulation data is available upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Cell line construction

To engineer the K562 cell line we began with LP3 from (Maricque et al., 2018). These female cells were authenticated in Maricque et al., 2018 just before we started using them. This landing pad is located on chromosome 11. First, we introduced a frameshift mutation to the GFP in the landing pad using CaS10 (from Shondra Miller, Washington University School of Medicine GeiC) and a gRNA against GFP (AddGene 41819). Second, we integrated our reporter at the AAVS1 locus using CaS10, the SM58 SSBD2 T2 gRNA (from Shondra Miller) and the pMVS184 reporter plasmid. Starting on Day 2, we selected for integrations with 1 ug/ml puromycin for three days. We tested candidate reporter clones with transfections of a synthetic TF (pMVS 223) carrying p53 AD1, choosing the clone with the largest dynamic range between baseline GFP and the brightest transfected cells. Our internal name for this clone is T7.1E3.

Cells were grown in Iscove's Modified Dulbecco's Medium (IMDM) medium +10% FBS +1% Non Essential Amino Acids +1% PennStrep (Gibco). All transfections used the Invitrogen Neon electroporation machine using a 100 ul tip, 1.2 M cells and 5 ug of DNA.

METHOD DETAILS

Rational mutagenesis

The sequences of all 525 VP16, Hif1 α , CITED2, Stat3 and p65 variants are listed in Dataset 1. The systematic mutagenesis added and removed charged residues or aromatic residues. Net charge of ADs was changed in two ways: subsets of charged residues were changed to each of the four charged residues and alanine, or subsets of polar residues were changed to charged residues. Aromatic residues were changed to alanine, leucine or other aromatic residues, and aromatic residues were added by replacing leucine, isoleucine, alanine, methionine, and valine residues.

The "Hand Designed" p53 AD variants contained the same systematic mutations and more hand designed variants listed in Dataset 2. The p53 mutagenesis also included a deep mutational scan, a double alanine scan and sequences from orthologous TFs. Activity values of each dataset are normalized separately and should not be compared.

Plasmid library construction

The plasmid sequences for the GFP reporter (pMVS184, Addgene 176294) and synthetic TF chassis (pMVS223, Addgene 176293) are in Supplemental Dataset 8.

In the 5 AD library, we designed the AD variants as protein sequences and reverse translated using optimal human codons. We attached each variant to 28 unique 12 bp FREE barcodes (Hawkins et al., 2018). WT ADs had 84 barcodes each. Between the AD and the barcode are BamHI, SacI and NheI restriction sites. For ADs that were less than 46AA, we added random filler DNA between the BamHI and SacI sites. We added PCR primers at the start (CCCAGCTTAAGCCACCATG) and end (CTCGAGATAACTTCGTATAATGTATGCTAT). Note there is an XhoI site after the barcode, included in the downstream primer. We ordered 14968 unique 217 bp ssDNA oligos from Agilent.

We cloned the AD variant library by HiFi assembly. We added plasmid homology to the ssDNA oligos by PCR, yielding a 232 bp product, with 4 cycles, Q5 polymerase, 0.5 pmol template and 8 reactions. The cloning primers were: TCACCGACCTCTCTCCCCAGCTTAAGCCACCATG and ATAGCATACATTATACGAAGTTATCTCGAG. We digested the pMVS223 backbone with AflII, XhoI and KpnI-HF and gel purified it. Each assembly had 100 ng of backbone and 5x molar ratio of insert. We electroporated bacteria and collected ~20 million colonies. We checked the library with paired end Illumina sequencing. We recovered 98.7% of our barcodes and all AD variants. For the second step of library cloning, we digested the library and pMVS223 with BamHI-HF and NheI-HF, and inserted the synthetic TF by T4 ligation. We electroporated bacteria and collected 400K colonies. We recovered 93% of designed barcodes and all ADs. In the final plasmid library the ADs are on the N terminus of the protein, between the initial methionine (ATG start) and a GSGS linker. In the No AD control there is nothing between the starting methionine and the GSGS linker. The synthetic TF is followed by a P2A cleavage sequence and an in frame Neomycin resistance gene. As a result, 1 and 2 bp deletions, the most common oligo synthesis errors, lead to frameshifts and are selected against after the library is integrated into the genome.

The p53 library was constructed in the same way with 5 barcodes per variant, 30 for WT AD1 and 25 for WT AD2. We collected 4 million colonies after step one and 26 million after step two. We recovered 14355 of 14998 designed barcodes and 2990 of 2991 designed ADs. In this work we used data from both WT ADs and 171 hand designed variants.

All restriction enzymes, HiFi mix, and competent bacteria were purchased from NEB. Library Maxipreps were performed using the ZymoPURE II Plasmid Maxiprep Kit (Zymo).

Plasmid library integration and measurement

In each transfection, we used 1.2 M cells, 2 ug of CMV-CRE (Maricque et al., 2018) and 3 ug of Plasmid Library. We transfected 102 M cells in 86 transfections split into 22 flasks. The next day, we began selection with 400 ng/ml G418 for 10 days. On Day 11 we performed magnetic enrichment of live cells (MACS by Miltenyi Biotec). We combined flasks 1-5 into biological replicate 1, flasks 6-10 into biological replicate 2, flasks 11-15 into biological replicate 3, and flasks 16-22 into biological replicate 4. On Day 12, we added β -estradiol to a final concentration of 1 μ M.

On Day 15 we sorted cells on a Sony HAPS 2 at the Siteman Cancer Center Flow Cytometry Core. We set an ON/OFF threshold for GFP as the 90th percentile of the uninduced population. The lowest bin was the bottom 50% of the OFF population. The ON region was split into 3 bins with equal populations. For each replicate, we collected 750K cells in each of the four bins. We noted the median fluorescence of each bin and used that number to calculate AD activity (see below). The dynamic range of the measurement is determined by the fluorescence values of the dimmest and brightest bin. After collecting the 4 independent replicates, we

combined all the cells and sorted them into 8 bins each with ~12% of the population—this sample has a larger dynamic range. These values are included in DataSet 1.

Barcode amplicon sequencing libraries

Genomic DNA was collected using the Qiamp DNA Mini kit (Qiagen). We performed 8 PCR on each sample. The sequencing libraries were prepared in 2 batches: Batch 1 contained biological replicates 1-3 and Batch 2 contained biological replicate 4 and the 8 bin sort. We did 25 cycles with NEB Q5 polymerase using CP36.P10 and LP_019 primers. We pooled the PCR, cleaned up the DNA (NEB Monarch), quantified it, digested the entire sample with NheI and EcoRI-HF (NEB) for 90 minutes and then ligated sequencing adaptors with T4 ligase (NEB) for 30 min. These adaptors contained sample barcodes in Read1 and Index1. We used 4 ng of this ligation for a 20 cycle enrichment PCR with Q5 and the EPCR_P1_short and EPCR_PE2_short primers. We sequenced each Batch on a NextSeq 500 1x75 High Output run.

Each biological p53 replicate was sorted on a different day, so each sequencing library was sequenced separately with a NextSeq 500 1x75 High Output run.

CP36.P10	ctcccgaattcgacgcgcac
LP_019	GCAGCGTATCCACATAGCGTAAAAG
EPCR_P1_short	AATGATACGGCGACCAACGAG
EPCR_PE2_short	CAAGCAGAAGACGGCATACGAGAT

To assess the number of integrations in each experiment, we saved 1 ml of culture (0.5-1M cells) from each flask (4 transfections) before the magnetic enrichment for live cells (Day 11). We extracted gDNA, amplified barcodes, and sequenced. We identified 96,000 unique integrations, an underestimate. In the sorted samples we recovered 14015 barcodes (93.6% of designed) total, 7164 in all four replicates and 10798 in three or more replicates. All ADs were present in all replicates.

QUANTIFICATION AND STATISTICAL ANALYSIS

Data processing

We demultiplexed samples using a combination of Index1 reads and Read1 inline barcodes using the 'fastqconvert_Xbal.py.' We identified barcodes (grep), sorted the barcodes (sort) and counted them (unique -c) with the 'processMSS18_Sort4_5_LigAdaptors.sh' scripts. Demultiplexed fastq files have been deposited in GEO.

Using perfect matches, we counted the abundance of each FREE barcode in each sample using the 'Preprocessing_MSS18_MergeBCs_NextSeq_p53_2021_forpublication.ipynb' script. We normalized the read counts first by the total reads in each sample and then renormalized each barcode across bins to create a probability mass function. We used the probability mass function and the median GFP fluorescence of each bin (Table S2) to calculate the activity of each barcode. To remove outlier barcodes, we found all barcodes for an AD, computed the activity of each barcode, computed the mean and variance of the set of barcodes and then removed any barcodes whose activity was more than two standard deviations away from the mean. We then took all the reads from all remaining barcodes, pooled them and recomputed activity. This approach led to one activity measurement for each biological replicate.

To combine data across replicates we used the 'No AD control.' We thank an anonymous reviewer for inspiring this analysis. In our library cloning, the parent plasmid (pMVS223), which does not carry an AD, is present at a low background, and this plasmid carries a unique 9 bp barcode. We computed the activity of this No AD control in each replicate. Next we adjusted the raw activity values (by addition or subtraction) so that the No AD control had an activity of 2000 in each replicate. We chose 2000 so that no activity values would be negative.

We used the average and standard deviations of these adjusted activity values for further analysis.

For the p53 data, each replicate was collected on a different day using 2 sorters (Replicate 1 and 3: Sony HAPS 2; Replicate 2: a highly modified Beckman Coulter MoFlow). Replicates 1 and 2 were sorted into 4 bins; Replicate 3 was sorted into 8 bins). To combine these data we converted activities into Z scores and computed the mean and standard error of the mean (SEM).

In the next step of preprocessing, we added physical property calculations and AD sequence names to the activity data using the 'MSS18 Step 2 of Preprocessing (fix names and add columns) For publication.ipynb'. This script also manually corrected errors in sequence names.

Analysis

All analysis was performed in Jupyter Notebooks with python 2.7 and Matplotlib, seaborn, pandas, localcider, biopython, logomaker (Tareen and Kinney, 2020), scipy, statsmodels, sklearn, and itertools. Colors are from Colorbrewer (<https://colorbrewer2.org/>). AD sequence properties were calculated with localcider (Holehouse et al., 2017). To identify AD variants that were statistically significantly different from each WT, we used a two-sided t test and 5% FDR correction. We computed the regressions with statsmodels.api.OLS.

Structures were downloaded from the RSCB PDB (www.rcsb.org) and visualized with VMD (Humphrey et al., 1996). We normalized activity values to [0-1], mapped the values to the Beta column of the pdb file and visualized positions with normalized activity < 0.2 (Figure 4E, S3C).

To summarize the effects of substitutions at each position (Figure 5A, S8C), we identified all variants that changed each position, collected the activity measurements from all biological replicates and created a boxplot. We excluded the shuffle variants.

Sequence properties were calculated with the localcider package or by counting amino acids. The Omega parameter was computed with localcider using the `get_kappa_X(['W','F','Y'], ['D','E'])` command.

Figure panels were generated with the 'MSS18 PaperFiguresRevision.ipynb' and 'MSS19_predictedADs_forpublication_v2.ipynb' jupyter notebooks in the Github repository.

Machine Learning

The machine learning analysis was carried out in python with the sklearn package. We started with all variants of VP16, Hif1 α and CITED2 and then excluded the shuffle variants. The High Activity set (N = 121) had variants with a mean activity above 3400. The Low Activity set (N = 134) had variants with a mean activity below 2900 (Figure S11F). We normalized all parameters to be between [0,1]. We performed 5-fold cross validation and assessed model performance with the Area Under the Curve (AUC) of the Receiver Operator Characteristic (ROC). We compared Support Vector Machines, Logistic Regression and Random Forest classifiers.

All atom simulations

We ran all-atom, Monte Carlo simulations in the CAMPARI simulation engine (23campari.sourceforge.net) using the ABSINTH implicit solvent paradigm (Vitalis and Pappu, 2009). This simulation framework is a well established approach to study the conformational ensembles of intrinsically disordered regions (Martin et al., 2016; Metskas and Rhoades, 2015; Vitalis and Pappu, 2009) and we have previously used it to study the Central Acidic AD of the yeast TF, Gcn4 (Staller et al., 2018). We simulated all VP16 and CITED2 variants. For Hif1 α , we simulated all hand designed variants and the WT sequence.

For each variant, we ran ten simulations starting in a helix and ten starting in a random coil. For the WT sequences, we ran 30 simulations from each start. In total we ran 4300 simulations. Each simulation had a pre-equilibration run of 2M steps. Then we began the real simulation with 10M steps of equilibration and the main simulation of 50M steps, extracting the conformation every 10K steps, yielding 5000 conformations per simulation. Simulation analysis was performed with the CAMPARItraj (ctrj.com) software suite. This software suite calculated helicity with the DSSP algorithm (Kabsch and Sander, 1983) and radius of gyration as the distribution of atoms in each confirmation without weighting by mass (Holehouse et al., 2017). The accessibility was calculated by rolling a 1.5 nm spherical marble around each confirmation and summing the solvent accessible surface area of the W,F,Y,L residues (Staller et al., 2018). To speed up this analysis accessibility was assessed every 20 confirmations.

The summary statistics for all the simulations is in Supplemental Dataset 9.

Predicting ADs in human TFs

We downloaded protein sequences from Uniprot for 1608 TFs (Lambert et al., 2018). For each TF, we created 39 AA tiling windows, spaced every 1 AA, yielding 881,344 tiles. For each tile, we computed the net charge (counting D,E,K&R) and counted W,F,Y,L residues.

We identified tiles that were as extreme or more extreme than VP16 and CITED2. We used a diagonal line to extrapolate between these ADs. The tiles predicted to cover ADs (Figure 6B, red pixels), fulfill 3 criteria:

$(\text{Charge} < -9) \text{ AND } (\text{WFYL} > 7) \text{ AND } (((\text{Charge}+9)-(\text{WFYL}-10)) \leq 0)$

This algorithm identified 1139 tiles, 0.129% of the total. We aggregated overlapping tiles to predict 144 ADs on 136 TFs. To test these predictions, we used ADs annotated in Uniprot. We downloaded .gff files for the 1608 TFs from Uniprot. We used 4 regular expressions to search the “regions” column of the .gff files for “activation”, “TAD”, “Required for transcriptional activation” and “Required for transcriptional activation.” These searches yielded 110 unique ADs, including 7 proline rich ADs (>20% proline) and 3 glutamine rich ADs (>20% glutamine).

We used permutation tests to determine if our predictor was better than random. We randomly selected 136 TFs, randomly selected 144 length matched regions and determined how many overlapped the 110 known ADs. For the 4 TFs with 2 predicted ADs, we preserved the coupling between these lengths. In 100K permutations, we never observed more than 11 overlaps. 17 of our predicted ADs overlapped the 110 Uniprot ADs.

Testing predicted ADs

We built a third plasmid library to test the predicted ADs. Due to DNA synthesis limits, we split long predicted regions and tested 150 regions of 39-76 residues. To create an empirical distribution for the prevalence of ADs on TFs, we included 150 length-matched regions randomly drawn from TF sequences (Lambert et al., 2018). We required that these random regions did not overlap our predicted ADs or Uniprot ADs. The 92 positive control ADs were drawn from: 36 hand-curated ADs (RegionType=Hand_Curated_ADs), 35 ADs from a published list (RegionType=Choi_2000_Pmid_10821850) (Choi et al., 2000), 19 Uniprot domains that overlapped our predictions (RegionType=Uniprot), and 2 published synthetic DW or DF runs (RegionType=Controls) (Ravarani et al., 2018). We also included 3 KRAB domains from Uniprot, 22 mutant ADs and 26 regions tiling the human TF, Crx. Due to human error, we did not test the correct predicted region of AEBP1 (Q8IUX7) and tested 2 other regions instead. The full list of sequences and activities is included in Dataset 4. The ‘Known ADs’ in Figure 6C are flagged in the ‘Positive Controls’ column. The ‘Negative Controls’ column indicates mutant ADs.

The plasmid library was cloned in a similar manner as above. The oligos were ordered as a oPool from IDT. Oligo length varied. For each AD, we included one 9 bp ‘AD barcode’ (Hawkins et al., 2018). During the second step of cloning, we added 6 Ns downstream of the synthetic TF by PCR, which became the ‘integration barcode.’ In principle, a different integration

barcode marks each plasmid integration event, analogous to a Unique Molecular Identifier in single cell RNA-seq protocols. The resulting 'composite barcode' contained a 6 bp 'integration barcode', the NheI restriction site and the 9 bp 'AD barcode.' 76 transfections were split into 3 biological replicates. G418 selection began on Day 1, magnetic separation was performed on Day 11, β -estradiol induction began on Day 11 and cell sorting on Day 14. For each biological replicate, we sorted into 4 bins. During the sequencing library preparation, we performed 24 PCRs for each gDNA sample. We added Index1 and Index2 barcodes by PCR.

After integrating the plasmid library into cells, we deeply sequenced the unsorted pool to build a 'composite barcode' table (Check_Complexity_MSS19_nextSeq.ipynb). This table contained 44077 composite barcodes with at least 10 reads in one of the 3 biological replicates. For all subsequent analysis, we matched reads to this table. We used perfect matches to designed 'AD barcodes' and combined reads for all 'integration barcodes' attached to each 'AD barcode' as described above: we first removed outliers and then combined read counts (MSS19_preprocessing.ipynb). In this experiment, we found that the No AD Control TF had (x,x,x) integration barcodes. We combined replicates by first setting the activity of the No AD Control to 200 and then computing the mean and standard deviation. The biological replicates contained 20850, 19758 and 21656 uniquely identifiable integrations. We designed 443 ADs, detected 434 in the plasmid library, and detected 431 integrated into cells. We required 5 or more unique integration barcodes in at least one replicate, yielding 428 ADs for downstream analysis. In Figure 6C, the threshold for AD activity was 200, and the threshold for strong AD activity was 223, the 95th percentile of the random regions.

Data and code availability

- The AD activity data are included in Supplementary Datasets 1,2 and 4.
- The Illumina sequencing data have been deposited at GEO and are publicly available as of the date of publication. Accession numbers are listed in the key resources table. The plasmids have been deposited in AddGene.
- All simulation data and flow cytometry data reported in this paper will be shared by the lead contact upon request.
- The analysis code has been deposited in a public Github repository with a Zenodo DOI listed in the key resources table.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.
- **Data availability:** All processed activity data is available in the attached Datasets. Simulation data is freely available upon request. The raw sequencing data will be available in the NIH GEO database. Our code is available on Github.

Supplemental item titles (Datasets)

Supplemental Dataset 1: All variants and activity measurements for the 5 AD library (VP16, CITED2, Hif1 α , Stat3 and p65).

Supplemental Dataset 2: All variants and activity measurements for the p53 ADs.

Supplemental Dataset 3: Predicted acidic ADs on human TFs.

Supplemental Dataset 4: Sequences and activity measurements from testing predicted acidic ADs

Supplemental Dataset 5: AD sequences, DNA barcodes, and barcode counts for the 5 AD library.

Supplemental Dataset 6: AD sequences, DNA barcodes, and barcode counts for the p53 AD library.

Supplemental Dataset 7: AD sequences and DNA barcodes for the predicted AD library.

Supplemental Dataset 8: Plasmid sequences for pMVS184 and pMVS223.

739 Supplemental Dataset 9: Table with features extracted from the all-atom simulations

740 References

- 741 Andresen, C., Helander, S., Lemak, A., Farès, C., Csizmok, V., Carlsson, J., Penn, L.Z., Forman-Kay, J.D.,
742 Arrowsmith, C.H., Lundström, P., et al. (2012). Transient structure and dynamics in the disordered c-Myc
743 transactivation domain affect Bin1 binding. *Nucleic Acids Res.* **40**, 6353–6366.
- 744 Attardi, L.D., and Tjian, R. (1993). *Drosophila* tissue-specific transcription factor NTF-1 contains a novel isoleucine-
745 rich activation motif. *Genes & Development* **7**, 1341–1353.
- 746 Berlow, R.B., Dyson, H.J., and Wright, P.E. (2017). Hypersensitive termination of the hypoxic response by a
747 disordered protein switch. *Nature* **543**, 447–451.
- 748 Brady, C.A., Jiang, D., Mello, S.S., Johnson, T.M., Jarvis, L.A., Kozak, M.M., Kenzelmann Broz, D., Basak, S., Park,
749 E.J., McLaughlin, M.E., et al. (2011). Distinct p53 transcriptional programs dictate acute DNA-damage responses and
750 tumor suppression. *Cell* **145**, 571–583.
- 751 Brzovic, P.S., Heikaus, C.C., Kisselev, L., Vernon, R., Herbig, E., Pacheco, D., Warfield, L., Littlefield, P., Baker, D.,
752 Klevit, R.E., et al. (2011). The acidic transcription activator Gcn4 binds the mediator subunit Gal11/Med15 using a
753 simple protein interface forming a fuzzy complex. *Mol. Cell* **44**, 942–953.
- 754 Chang, J., Kim, D.H., Lee, S.W., Choi, K.Y., and Sung, Y.C. (1995). Transactivation ability of p53 transcriptional
755 activation domain is directly related to the binding affinity to TATA-binding protein. *J. Biol. Chem.* **270**, 25014–25019.
- 756 Choi, Y., Asada, S., and Uesugi, M. (2000). Divergent hTAFII31-binding motifs hidden in activation domains. *J. Biol.*
757 *Chem.* **275**, 15912–15916.
- 758 Colwell, L.J., Brenner, M.P., and Ribbeck, K. (2010). Charge as a selection criterion for translocation through the
759 nuclear pore complex. *PLoS Comput. Biol.* **6**, e1000747.
- 760 Cress, W.D., and Triezenberg, S.J. (1991). Critical structural elements of the VP16 transcriptional activation domain.
761 *Science* **251**, 87–90.
- 762 Currie, S.L., Doane, J.J., Evans, K.S., Bhachech, N., Madison, B.J., Lau, D.K.W., McIntosh, L.P., Skalicky, J.J.,
763 Clark, K.A., and Graves, B.J. (2017). ETV4 and AP1 Transcription Factors Form Multivalent Interactions with three
764 Sites on the MED25 Activator-Interacting Domain. *J. Mol. Biol.* **429**, 2975–2995.
- 765 Diss, G., and Lehner, B. (2018). The genetic landscape of a physical interaction. *Elife* **7**.
- 766 Dyson, H.J., and Wright, P.E. (2016). Role of Intrinsic Protein Disorder in the Function and Interactions of the
767 Transcriptional Coactivators CREB-binding Protein (CBP) and p300. *J. Biol. Chem.* **291**, 6714–6722.
- 768 El-Gebali, S., Mistry, J., Bateman, A., Eddy, S.R., Luciani, A., Potter, S.C., Qureshi, M., Richardson, L.J., Salazar,
769 G.A., Smart, A., et al. (2019). The Pfam protein families database in 2019. *Nucleic Acids Res.* **47**, D427–D432.
- 770 Erijman, A., Kozłowski, L., Sohrabi-Jahromi, S., Fishburn, J., Warfield, L., Schreiber, J., Noble, W.S., Söding, J., and
771 Hahn, S. (2020). A High-Throughput Screen for Transcription Activation Domains Reveals Their Sequence Features
772 and Permits Prediction by Deep Learning. *Mol. Cell* **78**, 890–902.e6.
- 773 Ferreira, M.E., Hermann, S., Prochasson, P., Workman, J.L., Berndt, K.D., and Wright, A.P.H. (2005). Mechanism of
774 transcription factor recruitment by acidic activators. *J. Biol. Chem.* **280**, 21779–21784.
- 775 Finn, R.D., Coghill, P., Eberhardt, R.Y., Eddy, S.R., Mistry, J., Mitchell, A.L., Potter, S.C., Punta, M., Qureshi, M.,
776 Sangrador-Vegas, A., et al. (2016). The Pfam protein families database: towards a more sustainable future. *Nucleic*
777 *Acids Res.* **44**, D279–D285.
- 778 Freedman, S.J., Sun, Z.-Y.J., Kung, A.L., France, D.S., Wagner, G., and Eck, M.J. (2003). Structural basis for
779 negative regulation of hypoxia-inducible factor-1 α by CITED2. *Nat. Struct. Mol. Biol.* **10**, 504–512.
- 780 Giacomelli, A.O., Yang, X., Lintner, R.E., McFarland, J.M., Duby, M., Kim, J., Howard, T.P., Takeda, D.Y., Ly, S.H.,
781 Kim, E., et al. (2018). Mutational processes shape the landscape of TP53 mutations in human cancer. *Nat. Genet.*
782 **50**, 1381–1387.
- 783 Goodrich, J.A., and Tjian, R. (2010). Unexpected roles for core promoter recognition factors in cell-type-specific
784 transcription and gene regulation. *Nat. Rev. Genet.* **11**, 549–558.

785 Gray, V.E., Hause, R.J., and Fowler, D.M. (2017). Analysis of Large-Scale Mutagenesis Data To Assess the Impact
786 of Single Amino Acid Substitutions. *Genetics* 207, 53–61.

787 Hawkins, J.A., Jones, S.K., Jr, Finkelstein, I.J., and Press, W.H. (2018). Indel-correcting DNA barcodes for high-
788 throughput sequencing. *Proc. Natl. Acad. Sci. U. S. A.* 115, E6217–E6226.

789 Hermann, S., Berndt, K.D., and Wright, A.P. (2001). How transcriptional activators bind target proteins. *J. Biol. Chem.*
790 276, 40127–40132.

791 Holehouse, A.S., Das, R.K., Ahad, J.N., Richardson, M.O.G., and Pappu, R.V. (2017). CIDER: Resources to Analyze
792 Sequence-Ensemble Relationships of Intrinsically Disordered Proteins. *Biophys. J.* 112, 16–21.

793 Humphrey, W., Dalke, A., and Schulten, K. (1996). VMD: visual molecular dynamics. *J. Mol. Graph.* 14, 33–38, 27–
794 28.

795 Jackson, B.M., Drysdale, C.M., Natarajan, K., and Hinnebusch, A.G. (1996). Identification of seven hydrophobic
796 clusters in GCN4 making redundant contributions to transcriptional activation. *Mol. Cell. Biol.* 16, 5557–5571.

797 Jonker, H.R.A., Wechselberger, R.W., Boelens, R., Folkers, G.E., and Kaptein, R. (2005). Structural properties of the
798 promiscuous VP16 activation domain. *Biochemistry* 44, 827–839.

799 Kabsch, W., and Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-
800 bonded and geometrical features. *Biopolymers* 22, 2577–2637.

801 Kim, J.-Y., and Chung, H.S. (2020). Disordered proteins follow diverse transition paths as they fold and bind to a
802 partner. *Science* 368, 1253–1257.

803 Kim, J.-Y., Meng, F., Yoo, J., and Chung, H.S. (2018). Diffusion-limited association of disordered protein by non-
804 native electrostatic interactions. *Nat. Commun.* 9, 4707.

805 Kinney, J.B., Murugan, A., Callan, C.G., Jr., and Cox, E.C. (2010). Using deep sequencing to characterize the
806 biophysical mechanism of a transcriptional regulatory sequence. *Proc. Natl. Acad. Sci. U. S. A.* 107, 9158–9163.

807 Krois, A.S., Dyson, H.J., and Wright, P.E. (2018). Long-range regulation of p53 DNA binding by its intrinsically
808 disordered N-terminal transactivation domain. *Proc. Natl. Acad. Sci. U. S. A.* 115, E11302–E11310.

809 Kyte, J., and Doolittle, R.F. (1982). A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.*
810 157, 105–132.

811 Lambert, S.A., Jolma, A., Campitelli, L.F., Das, P.K., Yin, Y., Albu, M., Chen, X., Taipale, J., Hughes, T.R., and
812 Weirauch, M.T. (2018). The Human Transcription Factors. *Cell* 175, 598–599.

813 Latchman, D.S. (2008). *Eukaryotic Transcription Factors* (Elsevier Science).

814 Lecoq, L., Raiola, L., Chabot, P.R., Cyr, N., Arseneault, G., Legault, P., and Omichinski, J.G. (2017). Structural
815 characterization of interactions between transactivation domain 1 of the p65 subunit of NF- κ B and transcription
816 regulatory factors. *Nucleic Acids Res.* 45, 5564–5576.

817 Lin, J., Chen, J., Elenbaas, B., and Levine, A.J. (1994). Several hydrophobic amino acids in the p53 amino-terminal
818 domain are required for transcriptional activation, binding to mdm-2 and the adenovirus 5 E1B 55-kD protein. *Genes*
819 *Dev.* 8, 1235–1246.

820 Liu, Y., Matthews, K.S., and Bondos, S.E. (2008). Multiple intrinsically disordered sequences alter DNA binding by
821 the homeodomain of the *Drosophila* hox protein ultrabithorax. *J. Biol. Chem.* 283, 20874–20887.

822 Majithia, A.R., Tsuda, B., Agostini, M., Gnanapradeepan, K., Rice, R., Peloso, G., Patel, K.A., Zhang, X., Broekema,
823 M.F., Patterson, N., et al. (2016). Prospective functional classification of all possible missense variants in PPARG.
824 *Nat. Genet.* 48, 1570–1575.

825 Maricque, B.B., Chaudhari, H.G., and Cohen, B.A. (2018). A massively parallel reporter assay dissects the influence
826 of chromatin structure on cis-regulatory activity. *Nat. Biotechnol.*

827 Martin, E.W., Holehouse, A.S., Grace, C.R., Hughes, A., Pappu, R.V., and Mittag, T. (2016). Sequence Determinants
828 of the Conformational Properties of an Intrinsically Disordered Protein Prior to and upon Multisite Phosphorylation. *J.*
829 *Am. Chem. Soc.* 138, 15323–15335.

830 McIsaac, R.S., Oakes, B.L., Wang, X., Dummit, K.A., Botstein, D., and Noyes, M.B. (2013). Synthetic gene
831 expression perturbation systems with rapid, tunable, single-gene specificity in yeast. *Nucleic Acids Res.* **41**, e57.

832 Metskas, L.A., and Rhoades, E. (2015). Conformation and Dynamics of the Troponin I C-Terminal Domain:
833 Combining Single-Molecule and Computational Approaches for a Disordered Protein Region. *J. Am. Chem. Soc.* **137**,
834 11962–11969.

835 Pace, C.N., and Scholtz, J.M. (1998). A helix propensity scale based on experimental studies of peptides and
836 proteins. *Biophys. J.* **75**, 422–427.

837 Pacheco, D., Warfield, L., Brajcich, M., Robbins, H., Luo, J., Ranish, J., and Hahn, S. (2018). Transcription activation
838 domains of the yeast factors Met4 and Ino2: Tandem activation domains with properties similar to the yeast Gcn4
839 activator. *Mol. Cell. Biol.* **38**.

840 Park, M., Patel, N., Keung, A.J., and Khalil, A.S. (2019). Engineering Epigenetic Regulation Using Synthetic Read-
841 Write Modules. *Cell* **176**, 227–238.e20.

842 Raj, N., and Attardi, L.D. (2017). The Transactivation Domains of the p53 Protein. *Cold Spring Harb. Perspect. Med.*
843 **7**.

844 Ravarani, C.N., Erkina, T.Y., De Baets, G., Dudman, D.C., Erkin, A.M., and Babu, M.M. (2018). High-throughput
845 discovery of functional disordered regions: investigation of transactivation domains. *Mol. Syst. Biol.* **14**, e8190.

846 Regier, J.L., Shen, F., and Triezenberg, S.J. (1993). Pattern of aromatic and hydrophobic amino acids critical for one
847 of two subdomains of the VP16 transcriptional activator. *Proc. Natl. Acad. Sci. U. S. A.* **90**, 883–887.

848 Rollins, N.J., Brock, K.P., Poelwijk, F.J., Stiffler, M.A., Gauthier, N.P., Sander, C., and Marks, D.S. (2019). Inferring
849 protein 3D structure from deep mutation scans. *Nat. Genet.* **51**, 1170–1176.

850 Sanborn, A.L., Yeh, B.T., Feigerle, J.T., Hao, C.V., Townshend, R.J., Lieberman Aiden, E., Dror, R.O., and Kornberg,
851 R.D. (2021). Simple biochemical features underlie transcriptional activation domain diversity and dynamic, fuzzy
852 binding to Mediator. *Elife* **10**.

853 Schmiedel, J.M., and Lehner, B. (2019). Determining protein structures using deep mutagenesis. *Nat. Genet.* **51**,
854 1177–1186.

855 Sharon, E., Kalma, Y., Sharp, A., Raveh-Sadka, T., Levo, M., Zeevi, D., Keren, L., Yakhini, Z., Weinberger, A., and
856 Segal, E. (2012). Inferring gene regulatory logic from high-throughput measurements of thousands of systematically
857 designed promoters. *Nat. Biotechnol.* **30**, 521–530.

858 Shen, F., Triezenberg, S.J., Hensley, P., Porter, D., and Knutson, J.R. (1996a). Transcriptional activation domain of
859 the herpesvirus protein VP16 becomes conformationally constrained upon interaction with basal transcription factors.
860 *J. Biol. Chem.* **271**, 4827–4837.

861 Shen, F., Triezenberg, S.J., Hensley, P., Porter, D., and Knutson, J.R. (1996b). Critical amino acids in the
862 transcriptional activation domain of the herpesvirus protein VP16 are solvent-exposed in highly mobile protein
863 segments. An intrinsic fluorescence study. *J. Biol. Chem.* **271**, 4819–4826.

864 Staller, M.V., Holehouse, A.S., Swain-Lenz, D., Das, R.K., Pappu, R.V., and Cohen, B.A. (2018). A High-Throughput
865 Mutational Scan of an Intrinsically Disordered Acidic Transcriptional Activation Domain. *Cell Syst* **6**, 444–455.e6.

866 Tareen, A., and Kinney, J.B. (2020). Logomaker: beautiful sequence logos in Python. *Bioinformatics* **36**, 2272–2274.

867 Tuttle, L.M., Pacheco, D., Warfield, L., Wilburn, D.B., Hahn, S., and Klevit, R.E. (2021). Mediator subunit Med15
868 dictates the conserved “fuzzy” binding mechanism of yeast transcription activators Gal4 and Gcn4. *Nat. Commun.* **12**,
869 1–11.

870 Tycko, J., DelRosso, N., Hess, G.T., Aradhana, Banerjee, A., Mukund, A., Van, M.V., Ego, B.K., Yao, D., Spees, K.,
871 et al. (2020). High-Throughput Discovery and Characterization of Human Transcriptional Effectors. *Cell* **183**, 2020–
872 2035.e16.

873 Vitalis, A., and Pappu, R.V. (2009). ABSINTH: a new continuum solvation model for simulations of polypeptides in
874 aqueous solutions. *J. Comput. Chem.* **30**, 673–699.

875 Vogel, T.P., Milner, J.D., and Cooper, M.A. (2015). The Ying and Yang of STAT3 in Human Disease. *J. Clin.*

876 Immunol. 35, 615–623.

877 Warfield, L., Tuttle, L.M., Pacheco, D., Klevit, R.E., and Hahn, S. (2014). A sequence-specific transcription activator
878 motif and powerful synthetic variants that bind Mediator using a fuzzy protein interface. *Proc. Natl. Acad. Sci. U. S. A.*
879 111, E3506–E3513.

880 Wimley, W.C., and White, S.H. (1996). Experimentally determined hydrophobicity scale for proteins at membrane
881 interfaces. *Nat. Struct. Biol.* 3, 842–848.

882 Wojciak, J.M., Martinez-Yamout, M.A., Dyson, H.J., and Wright, P.E. (2009). Structural basis for recruitment of
883 CBP/p300 coactivators by STAT1 and STAT2 transactivation domains. *EMBO J.* 28, 948–958.

Figure legends

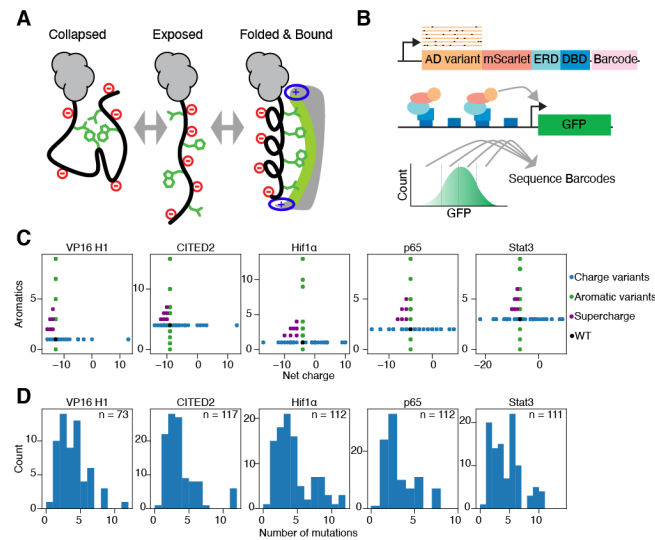


Figure 1. A high throughput assay for measuring the activities of AD variants in parallel.
(A) In the Acidic Exposure Model, ADs fluctuate between collapsed and exposed states. Exposed ADs can bind coactivators and partially fold.
(B) The high-throughput AD assay uses a synthetic DNA binding domain (DBD), an estrogen response domain (ERD), a GFP reporter, FACS and barcode sequencing. The reporter is integrated at the AAVS1 locus.
(C) We designed mutations that varied net charge or the number of aromatic residues. We designed a small set of supercharge variants to vary both properties.
(D) Histograms of the number of mutations in each variant. Most variants had 5 or fewer substitutions.

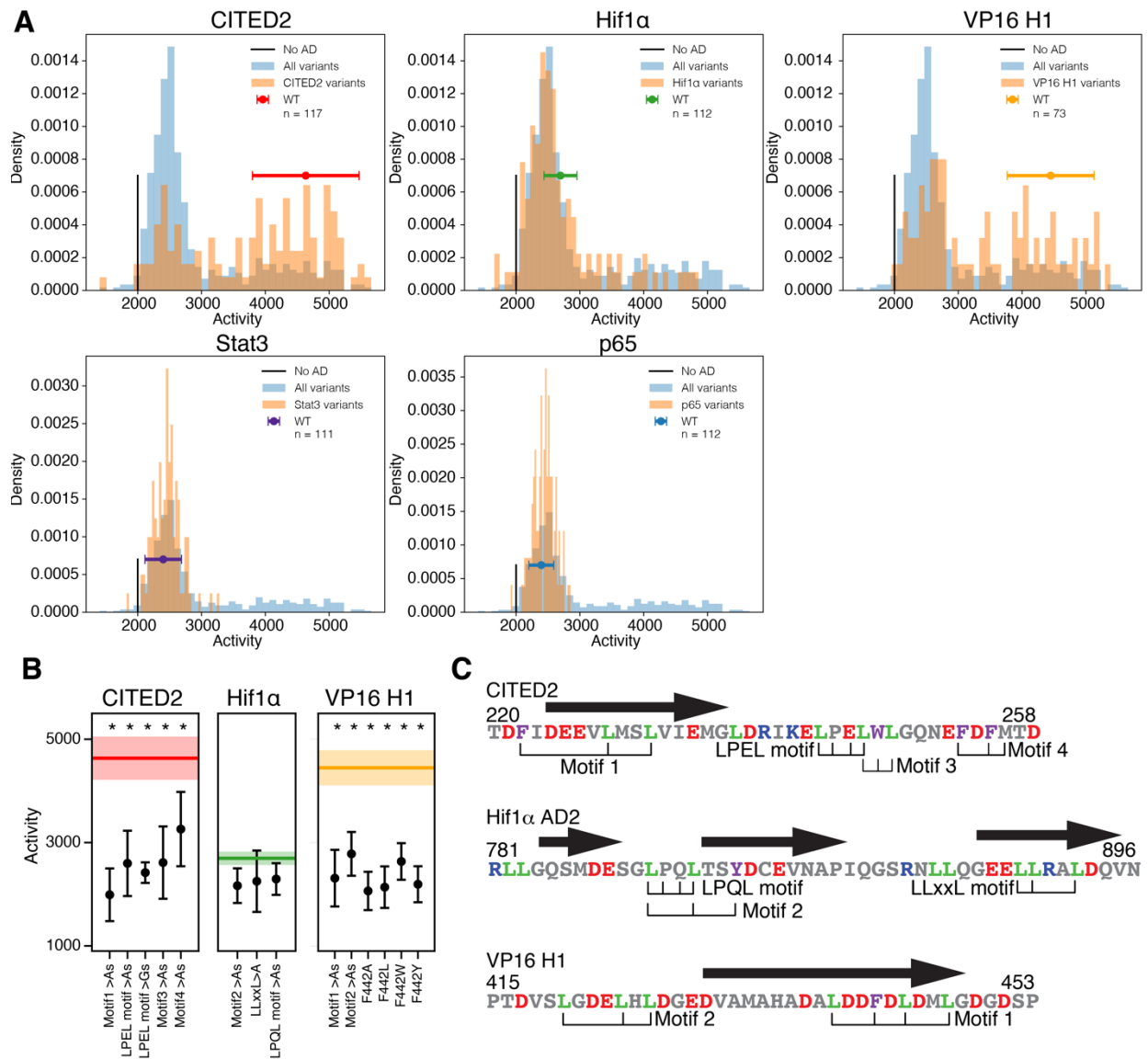


Figure 2. Rationally designed variants increase and decrease activation domain activity

(A) Histograms of the activities of all variants (n = 525, blue) and variants of each AD (orange). On the X-axis, Activity is the calculated GFP fluorescence (Arbitrary Units). On the Y-axis, density is the normalized counts of variants in each bin of activity. Biological replicates were normalized so that the No AD control had an activity of 2000 and then averaged together. The vertical black line indicates the activity of the No AD control. For each WT AD, the mean and standard deviation across the four replicates are shown.

(B) The effects of mutating hydrophobic motifs to alanine or glycine residues. For VP16, the effects of substituting aromatic residues with alanine or leucine. For each panel, the thick line is the mean activity of the WT AD and the shaded box is the standard deviation. Colors match panel A. *, p<0.05, 2-sided t-test with 5% FDR correction.

(C) Motif locations and alpha helices (arrows).

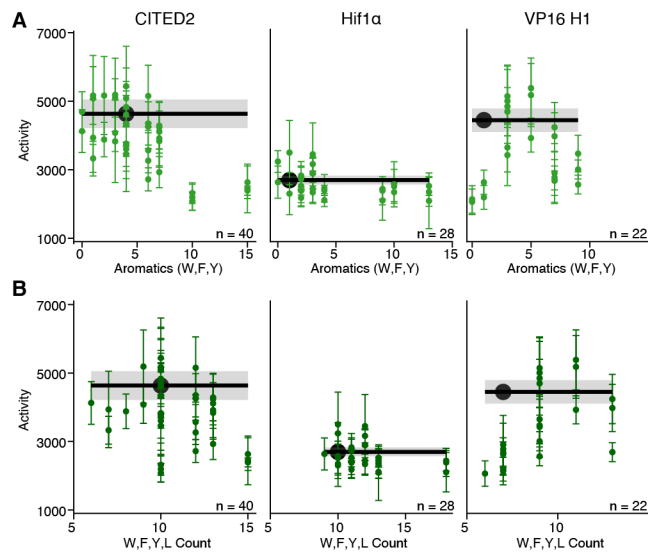


Figure 4: Adding aromatics has context dependent effects on AD activity

(A) For variants that add or remove aromatic residues, the mean and standard deviation are plotted. Activity is plotted against the number of aromatic (W,F,Y) residues. WT activity, black dot and line; WT standard deviation, gray box.

(B) Variants that add or remove aromatic residues, with activity plotted against the number of W,F,Y,L residues.

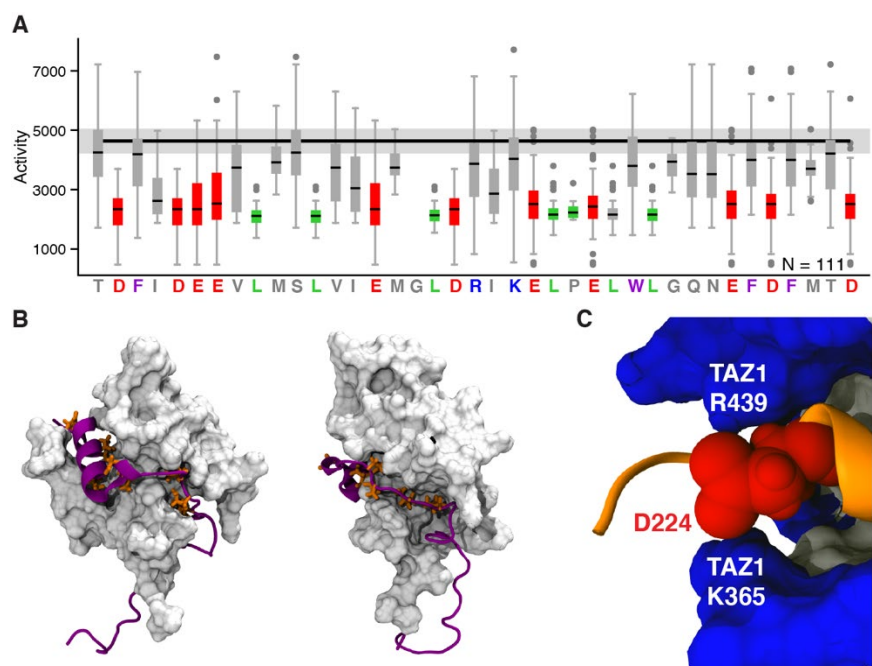


Figure 5. Rational mutagenesis reveals the structural constraints of an AD-coactivator interaction interface

(A) For each position in CITED2, all variants that change that residue are summarized as a boxplot (Activity, AU). Note that each position has different substitutions and variants with multiple substitutions are included in multiple columns. Acidic (red) and leucine (green) residues. Medians, black lines. Whiskers are 1.5 times the interquartile range. Outliers, gray dots. WT mean and standard, black line and gray box. This analysis excludes the shuffle variants.

(B) For all the positions in panel A with a median less than 3000 AU, we visualized these residues (orange) on the NMR structure of CITED2 bound to the TAZ1 coactivator (1R8U). CITED2 backbone, purple; visualized residues, orange; TAZ1, white. The residues that have large decreases in activity when mutated point towards the coactivator surface.

(C) D224 (red) of CITED2 is sandwiched between the narrowest point of the basic rim (blue) of the binding canyon of TAZ1. See Figure S10 for snapshots of all 20 structures in 1R8U.

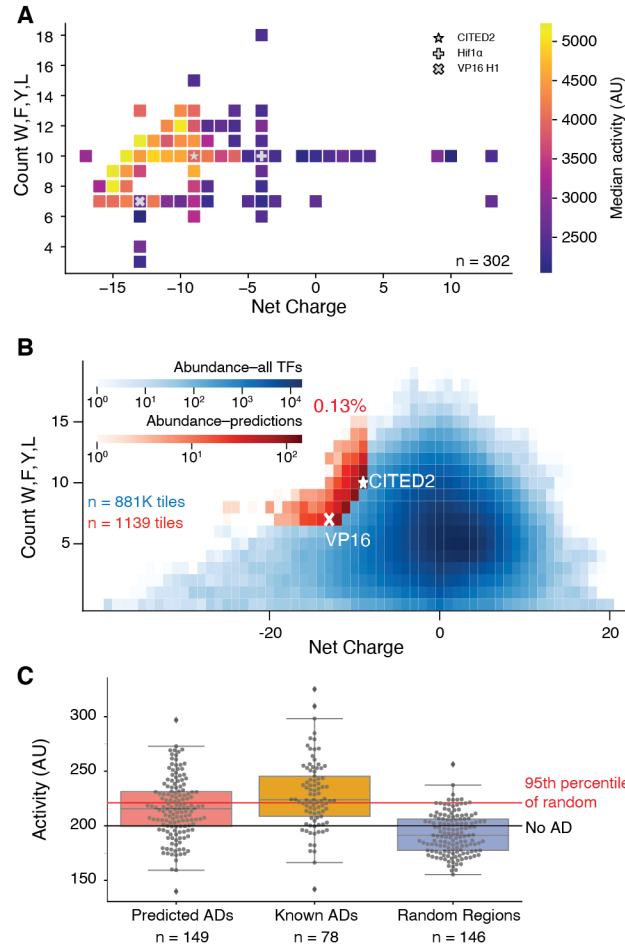


Figure 6. Strong activation domains are acidic and contain many W,F,Y,L residues.

(A) For each point, the x position indicates the net charge and y position the number of W,F,Y,L residues. Color (AU) indicates the median activity of all variants with each combination (See Figure S11 for all individual variants). All variants of VP16, CITED2 and Hif1a are included. The No AD control is 2000 AU. N = 302.

(B) A heatmap of all 39AA tiles from human TFs (log scale). The pixel location indicates the net charge and W,F,Y,L count, and the blue intensity indicates the number of tiles with that combination. Only 0.13% of tiles (red, rescaled heatmap) are as extreme or more extreme than VP16 (x) and CITED2 (*). The red tiles peak at CITED2.

(C) TF regions spanned by the red tiles (red, n = 149) are more likely to have AD activity than random regions (blue, n=146). Most, but not all, published ADs (orange, n = 78) have high activity in this assay. In this experiment the No AD control was normalized to 200 AU. The boxplot shows the quartiles and whiskers are drawn at 1.5 times the interquartile range.

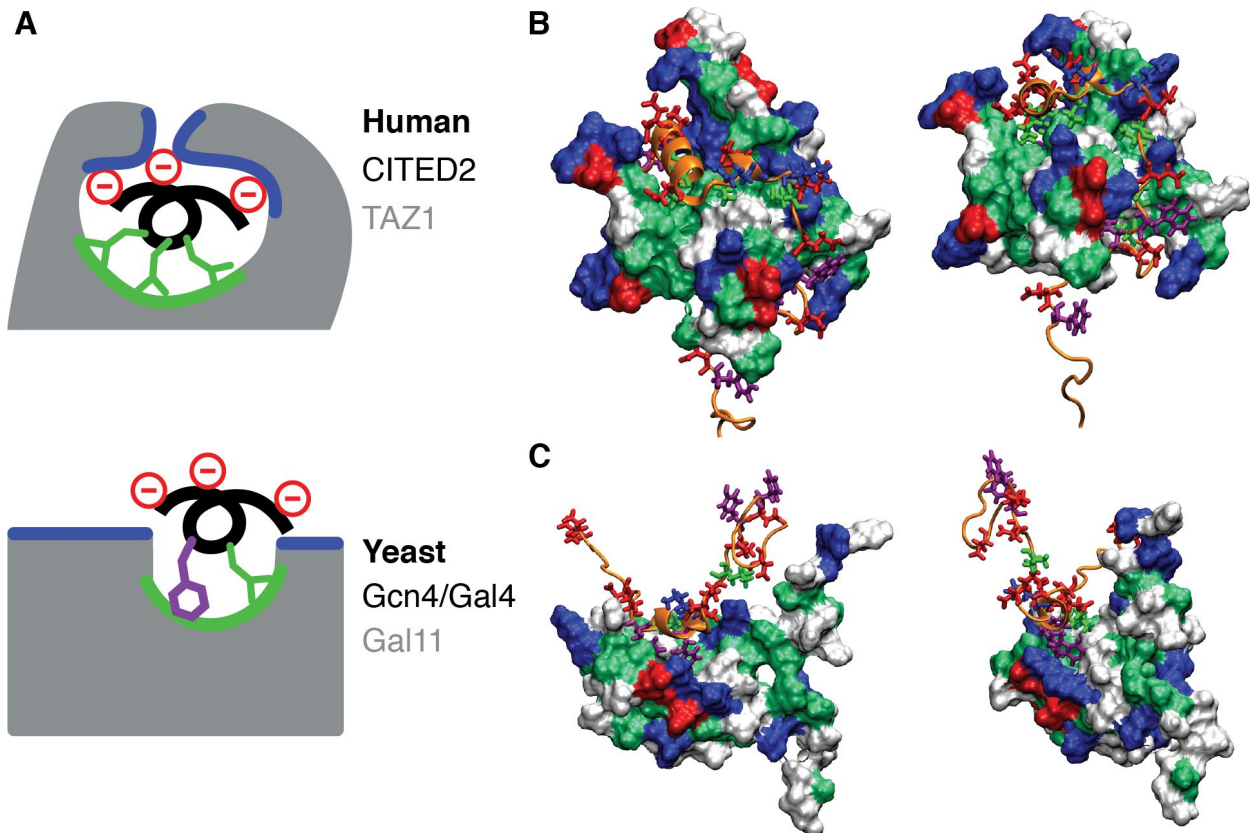


Figure 7: The structure of the coactivator AD-binding canyon constrains AD sequence.

(A) The CITED2 AD is inside the Taz1 canyon, a structural constraint that favors leucine residues. The yeast Gcn4 AD is outside the Med15/Gal11 canyon, enabling a fuzzy interaction that favors aromatic residues.

(B) The deep canyon of Taz1 embraces CITED2 (orange, 1R8U).

(C) The binding canyon of Gal11 (Med15) is shallow and the Gcn4 central acidic AD inserts aromatic side chains (2LPB). Colors in B and C are: red, acidic (negative); blue, basic (positive); green, hydrophobic; purple, aromatic; other, white.