Title:

Human transcriptional activation domains require hydrophobic and acidic residues

Authors:
Max V. Staller[1], Eddie Ramirez[1], Alex S. Holehouse[2,3], Rohit V. Pappu[3,4], Barak A. Cohen[1]

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] Department of Biochemistry and Molecular Biophysics, Washington University in St. Louis School of Medicine, Saint Louis, MO
[3] Center for Science and Engineering of Living Systems Washington University in St. Louis, St. Louis, MO, USA
[4] Department of Biomedical Engineering, Washington University in St. Louis, St. Louis, MO, USA

Abstract

Transcription factors activate gene expression with separable DNA binding domains and activation domains (Latchman, 2008). High-throughput studies have uncovered rules for how DNA binding domains recognize their cognate DNA motifs, but the design principles of activation domains remain opaque. For over thirty years it has been a mystery why activation domains are acidic and unstructured (Sigler, 1988). Activation domains require hydrophobic motifs to bind coactivators and join transcriptional condensates, but low evolutionary conservation and intrinsic disorder have made it difficult to identify the design principles that govern the sequence to function relationship (Boija et al., 2018; Chong et al., 2018; Cress and Triezenberg, 1991; Dyson and Wright, 2016). Consequently, activation domains cannot be predicted from amino acid sequence (Finn et al., 2016). Here, we resolve the functional roles of acidity and disorder in activation domains and use these insights to build a new predictor. We designed sequence variants in seven acidic activation domains and measured their activities in parallel with a high-throughput assay in human cell culture. Our results support a flexible model in which acidic residues solubilize hydrophobic motifs so that they can interact with coactivators. This model accurately predicts activation domains in the human proteome. We identify three general rules for activation domain function: hydrophobic motifs must be balanced by acidic residues; acidic residues make large contributions to activity when they are adjacent to motifs; and within motifs, the presence of aromatic or leucine residues reflects the structural constraints of coactivator interactions. We anticipate these design principles will aid efforts to predict activations from amino acid sequence and to engineer new domains.

Introduction

Transcription factors (TFs) activate gene expression with a DNA binding Domain (DBD) and an activation domain (AD). DBDs are structured, evolutionarily conserved and bind related DNA sequences. ADs are intrinsically disordered regions (IDRs), poorly conserved, bind structurally diverse coactivator subunits and frequently undergo coupled folding and binding. Bioinformatics tools for predicting DBDs can predict candidate TFs, but we cannot predict which TFs are activators (El-Gebali et al., 2019; Finn et al., 2016). ADs can form clusters in the nucleus, a phenomenon that has been called hub formation, liquid liquid phase separation, or condensate formation. It remains unclear if this clustering is the cause or consequence of AD activity.

Our understanding of ADs has lagged behind the DBDs for two reasons: 1) ADs are hard to study with traditional molecular biology, biochemistry and comparative genomics methods (Dyson and Wright, 2016; Hahn, 1993; Hahn and Young, 2011; Latchman, 2008; Sigler, 1988) and 2) high-throughput methods for studying DNA-protein interactions *in vitro* (Bulyk, 2007; Riley et al., 2014; Stormo et al., 2015; Teytelman et al., 2013) and *in vivo* accelerated our understanding of DBDs (Buenrostro et al., 2013; Noyes et al., 2008; Skene and Henikoff, 2017; Vierstra and Stamatoyannopoulos, 2016). Developing analogous high throughput methods for studying ADs will help close the understanding gap with DBDs. We and others have recently developed high-throughput methods for studying ADs in yeast, fly cells and human cells (Arnold et al., 2018; Erijman et al., 2020; Ravarani et al., 2018; Staller et al., 2018; Tycko et al., 2020). Building upon our work in yeast, here, we present a high-throughput method for studying AD variants in human cell culture. We deployed rational mutagenesis and hypothesis testing to investigate the amino acid sequence features that control activity of strong acidic ADs, with an aim to identify the design principles underlying AD function.

There are thirty years of conflicting data on the sequence to function relationship of acidic ADs. The only shared feature of the first few dozen eukaryotic ADs was an abundance of acidic residues (net negative charge), leading Paul Sigler to propose the "Acid Blob and Negative Noodle" model (Hope et al., 1988; Ma and Ptashne, 1987; Sigler, 1988) wherein ADs interacted with the polymerase electrostatically. This model was challenged when site directed mutagenesis studies showed that hydrophobic residues made larger contributions to activity than acidic residues (Cress and Triezenberg, 1991; Drysdale et al., 1995; Jackson et al., 1996). Acidic residues can collectively contribute to activity (Cress and Triezenberg, 1991) but in some cases all the acidic residues can be removed without loss of activity (Brzovic et al., 2011). Kinetic studies of AD-coactivator interactions have found fast, low-affinity electrostatic interactions followed by slow, high-affinity hydrophobic interactions (Ferreira et al., 2005; Hermann et al., 2001). A more recent explanation for the acidic residues in ADs is the connection to intrinsic disorder. Acidic residues are enriched in IDRs (Oldfield and Dunker, 2014). Virtually all ADs are intrinsically disordered and many AD fold into short alpha helices when bound to coactivators (Dyson and Wright, 2016; Liu et al., 2006). However, this folding is often dispensable because inserting a proline into a helix frequently has no effect on function (Brzovic et al., 2011; Cress and Triezenberg, 1991). It remains unexplained why acidity is a common and conserved feature of diverse ADs.

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). Left to their own devices, hydrophobic residues will interact with each other and drive intramolecular chain collapse, suppressing interactions with coactivators. Surrounding the motifs with acidic residues that repel one another exposes the hydrophobic residues to solvent, promoting interactions with coactivators.

We developed the Acidic Exposure Model in yeast and here, test this model in human cell culture. Although core transcriptional processes are conserved from yeast to humans, metazoans have added additional layers of regulation and new coactivator complexes. We selected one viral and six human ADs and rationally designed mutations to test the contributions of hydrophobic motifs, alpha helices, aromatic residues and acidic residues to activity. To test these AD variants, we developed a new high-throughput method in human cell culture.

We show that strong acidic ADs balance acidic residues against aromatic and leucine residues. All ADs tested required hydrophobic motifs and acidic residues. Adding acidic residues increased the activity of some ADs. For the most acidic ADs, adding aromatics could increase activity, however, in all ADs tested, there came a point where adding more aromatics decreases activity. The added aromatics overwhelmed exposure capacity of the acidity and disorder. To further test the Acidic Exposure Model, we used the balance between acidic, aromatic and leucine residues to predict new activation domains across human TFs. Many of our predictions overlap known ADs. Within motifs, the preference for leucine or aromatic residues reflects the structural constraint of the AD-coactivator interaction interface. Finally, we show that the arrangement of amino acids within ADs is also important for function. We synthesize our findings in three design principles: activation domains balance hydrophobic motifs and acidic residues, acidic residues better promote exposure when adjacent to the motifs; within the motifs, the choice between aromatic and leucine residues is constrained by the structure of the coactivator interaction surface.

Results

We investigated three key features of ADs: acidic residues, hydrophobic motifs and disorder-to-order transitions. We designed sequence variants that systematically added and subtracted acidic residues or aromatic residues in seven ADs: VP16 (H1 region, 415-453), Hif1ɑ (AD2, 781-896), CITED2 (220-258), Stat3 (719-764), p65 (AD2, 521-551), p53 AD1 (1-40) and p53 AD2 (40-60) (Figure 1B). For each disordered region that folds into an alpha helix upon coactivator binding, we introduced proline or glycine residues, which suppress helicity (Figure S1). We measured 525 and 2991 variants in two experiments.

To test these designed variants, we developed a high-throughput method to assay AD variants in parallel in human cell culture (Figure 1C, S2, Methods). 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,” reducing the effects of genomic position on expression (Maricque et al., 2018). AD variants that drive different levels of GFP expression are separated by 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). The assay is reproducible, quantitative and resolves 9-10 levels of activity (Figure 2, S2, S3).

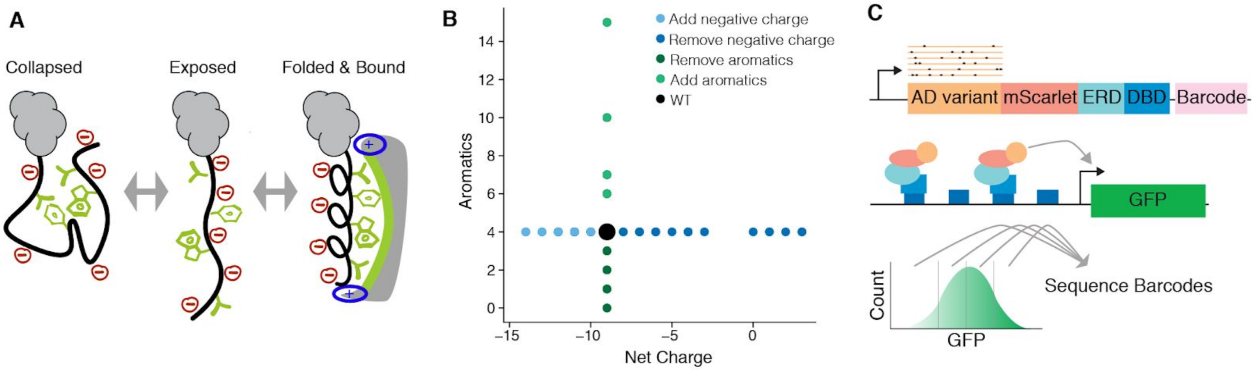


Figure 1. A high throughput assay for measuring the activities of AD variants in parallel. A) In the Acidic Exposure Model, ADs fluctuate between collapsed and exposed states. Exposed ADs can bind coactivators and partially fold. B) Rationally designed mutations add and remove aromatic residues or vary net charge. C) The high-throughput AD assay uses a synthetic DNA binding domain (DBD), and estrogen response domain (ERD), a GFP reporter, FACS and barcode sequencing.

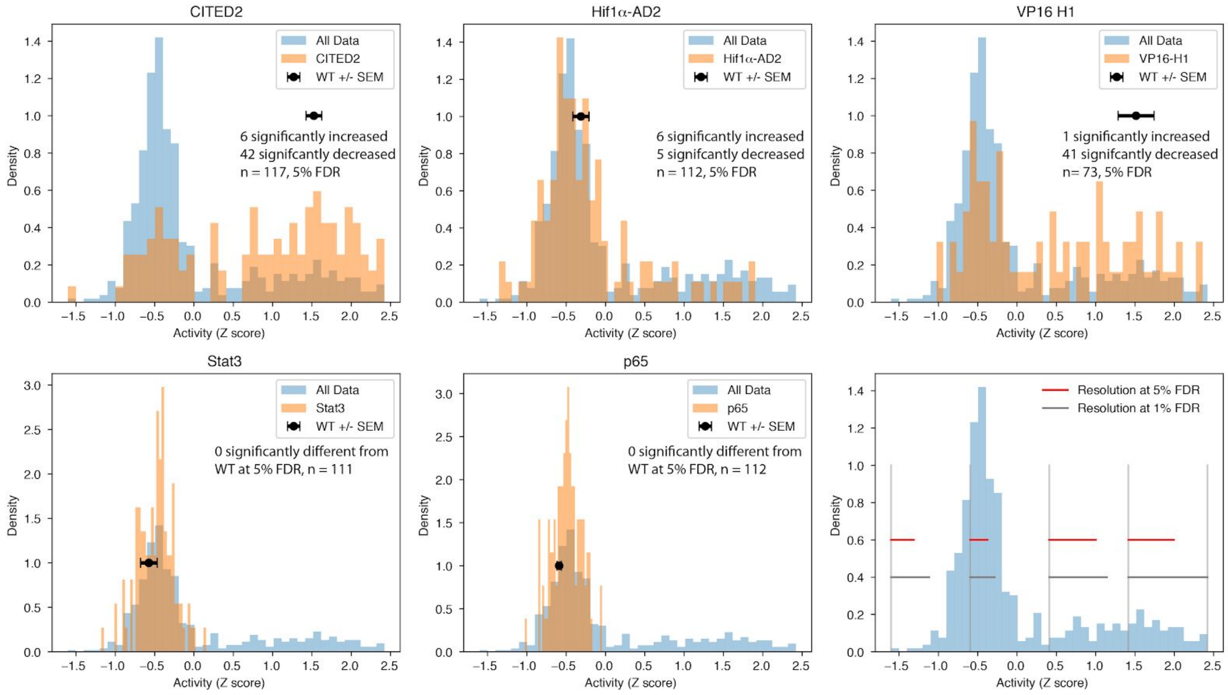


Figure 2. The high throughput assay detected increases and decrease in activation domain activity Histograms of the activities of all variants (blue) and AD variants (orange). The mean and SEM of each WT AD is shown in black. CITED2 and VP16 started with high activity and our designed variants increased and decreased activity. Hif1ɑ started with modest activity and our designed variants increased and decreased activity. The resolution of the method at 5% FDR (red) and 1% FDR (dark gray) in each quartile is shown. For 5% FDR, if we average these four resolutions then we can resolve 9.3 levels of activity. If we apply each resolution to its respective quartile, we can resolve 10.4 levels of activity.

We confirmed that hydrophobic motifs make large contributions to AD activity (Cress and Triezenberg, 1991; Jackson et al., 1996; Lin et al., 1994). Removing aromatic and leucine-rich motifs decreased activity of all ADs (Figure 3A, 3B, S4). Both known motifs (LPEL in CITED2, LPQL and LLxxL in Hif1ɑ, and LxxFxL in VP16 (Berlow et al., 2017; Regier et al., 1993)) and predicted motifs (Figure S1) contributed to activity. Aromatic residues made large contributions to activity in VP16 and both p53 ADs, as expected (Cress and Triezenberg, 1991; Lin et al., 1994), but made smaller contributions to activity in CITED2 and Hif1ɑ (Figure 4, S4-6). Perturbing hydrophobic motifs consistently decreased activity.

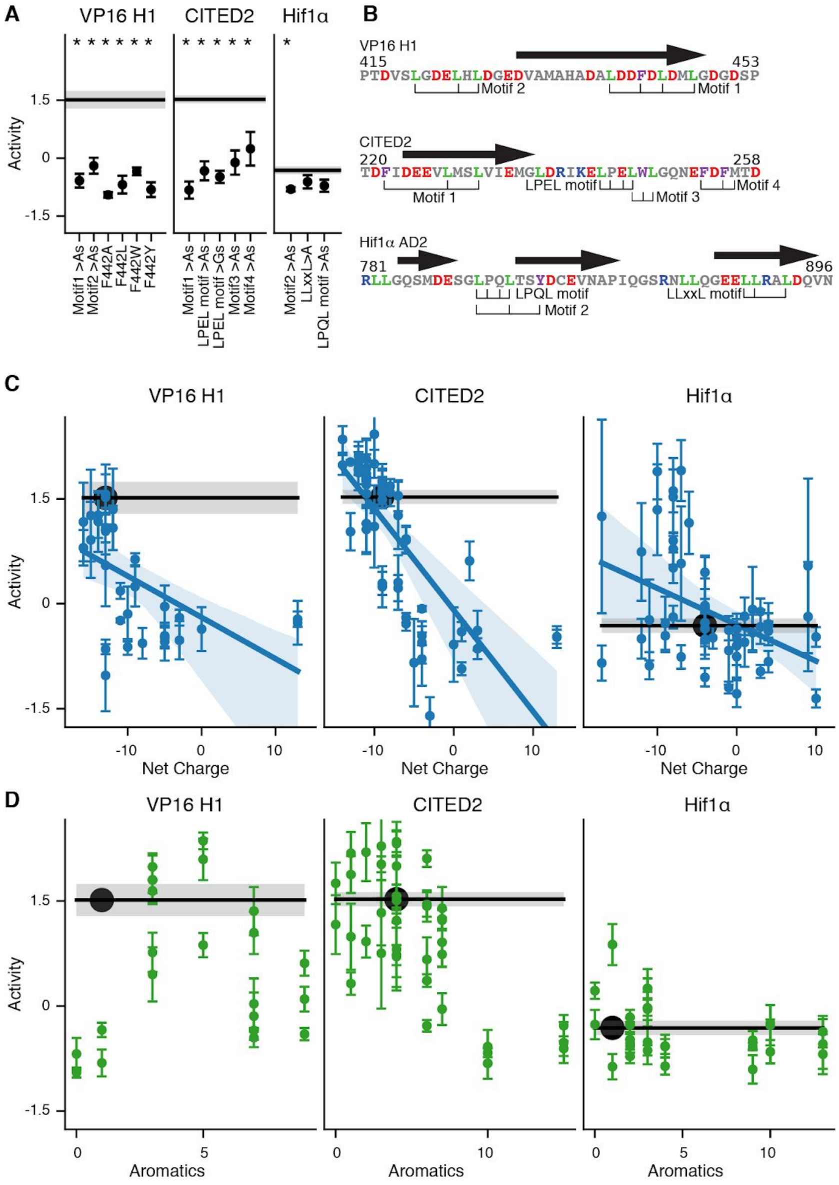


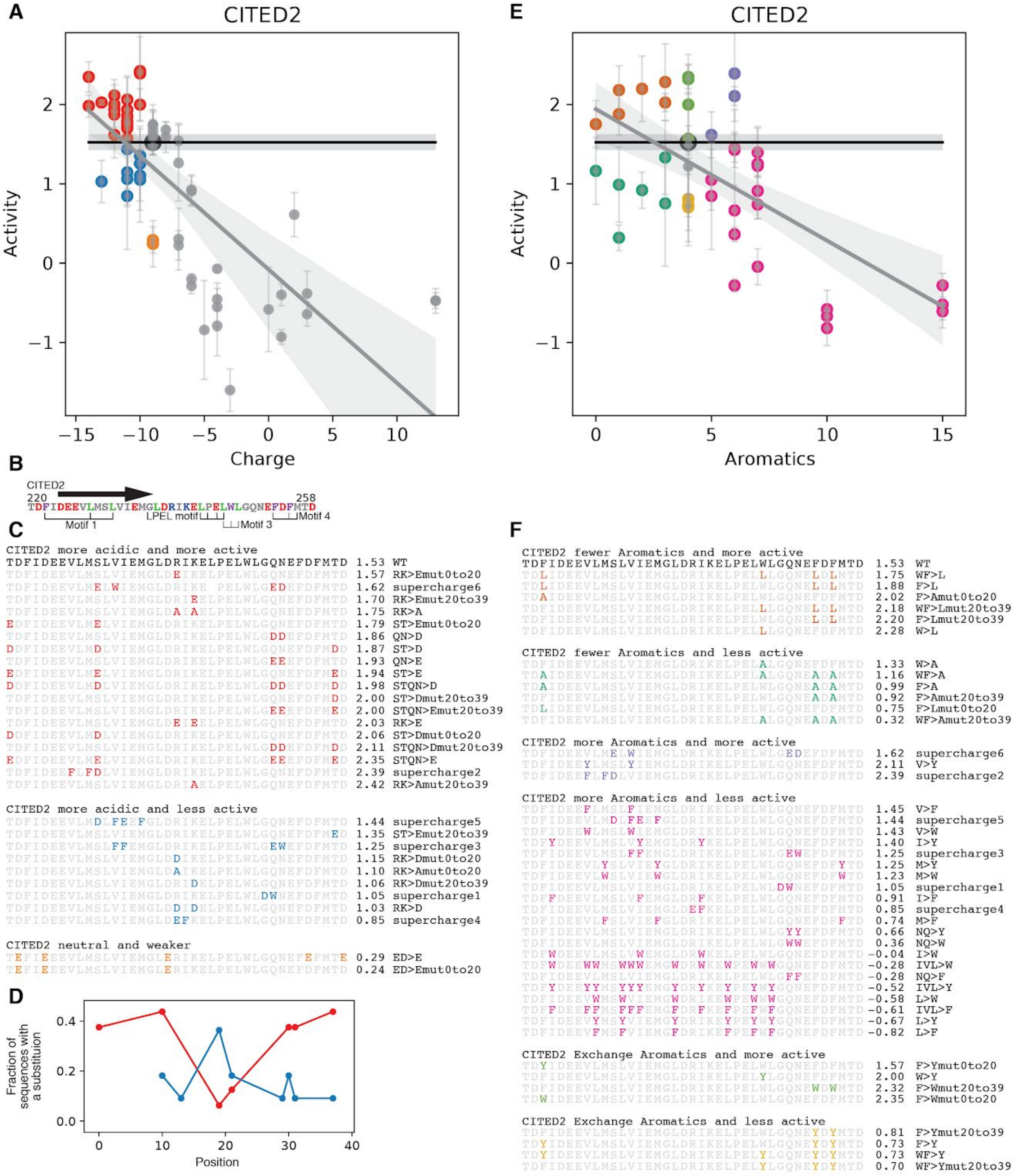
Figure 3. Acidic activation domains require hydrophobic motifs and acidic residues. A) Mutating motifs to alanine decreased activity. Mean and SEM shown. *, p<0.05. B) Motif locations and alpha helices (arrows). C) For variants designed to perturb net charge, the mean and SEM are plotted along with a linear regression and confidence interval. D) Variants that added aromatic residues increased or decreased activity. WT mean, black line and dot. SEM, gray box.

Acidic residues were necessary for the activity of all ADs and adding acidic residues increased the activity of some ADs. For VP16, CITED2, p53 AD1 and p53 AD2, removing acidic residues (moving towards positive net charge) decreased activity (Figure 3C). For Hif1α, four of five variants with significantly reduced activity added basic residues (Figure S7). Adding negatively charged residues frequently increased the activity of CITED2, Hif1α and p53 AD1 (Figure 2C, Figure S4). The location of added acidic residues determined whether activity increased (discussed below, Figure 4,S5, S6). For the most acidic ADs, VP16 and p53 AD2, adding acidic residues rarely increased activity (Figure 3C, S4). Similar trends are weakly visible in the Stat3 variants (Figure S8). Activity was better explained by net charge than by the identities of the added acidic and basic residues: 1) removing negative residues or adding positive residues had similar effects, 2) adding lysine or arginine (both positively charged) had similar effects and 3) adding aspartic acid or glutamic acid (both negatively charged) had similar effects (Figure S9). Adding acidic residues caused two responses: increased activity or near WT activity.

Adding aromatic residues can increase or decrease activity (Figure 3D). For VP16, adding 1-4 aromatic residues increased activity in the majority of variants, but adding more aromatic residues always decreased activity. For both p53 ADs, adding aromatics frequently increased activity (Figure S4). In contrast, adding aromatic residues to CITED2 and Hif1ɑ nearly always decreased activity (Figure 3D). Adding aromatic residues caused two opposite responses: increased activity or decreased activity.

The Acidic Exposure Model can explain why the two responses to adding acidic residues mirror the two opposite responses to adding aromatic residues. In the model, 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, activity is capped, and adding more acidic residues will not increase activity. In contrast, adding more aromatic residues eventually reduces activity, because they overwhelm the acidic residues and drive collapse, reducing total motif exposure. CITED2 has the most aromatic residues and its activity can be increased by adding acidic residues but not by adding aromatic residues. CITED2 has excess hydrophobic residues and activity is limited by exposure capacity. VP16 is the most acidic AD and its activity can be increased by adding aromatic residues but not by adding acidic residues. VP16 has excess exposure capacity and activity is limited by hydrophobic residues. We further tested this hypothesis by running all-atom Monte Carlo simulations of all VP16 and CITED2 variants (Methods, (Staller et al., 2018; Vitalis and Pappu, 2009). Supporting this prediction, introducing aromatic residues leads to more collapsed ensembles in the simulations, while adding acidic residues leads to more expanded ensembles (Figure S10). We also observe the finite exposure capacity of CITED2 (Figure S10). In the Acidic Exposure Model, maximal activity requires a balance between the numbers of acidic residues and hydrophobic motifs.

The Acidic Exposure Model extends from yeast to human cells with one elaboration: a larger role for leucine residues in human cells. In yeast, we focused on the central AD of Gcn4, where aromatic residues made large contributions to activity while leucine and methionine made smaller contributions (Erijman et al., 2020; Jackson et al., 1996; Ravarani et al., 2018; Staller et al., 2018). In human cells, we found three pieces of evidence that leucine residues make large contributions to AD activity. First, some motifs contained only leucine residues (Figure 3B). Second, replacing aromatics with leucines increased activity in Hif1ɑ and CITED2 (Figure 4, S5). Third, replacing leucines with aromatics frequently decreased activity (Figure 4, S5). The increased role of leucine residues likely reflects interactions with the expanded set of coactivators present in human cells.



```
CITED2 more acidic and more active
TDFIDEEVLMSLVIEMGLDRIKELPELWLGQNEFDFMTD 1.53 WT
TDFIDEEVLMSLVIEMGLDEIKELPELWLGQNEFDFMTD 1.57 RK>Emut0to20
TDFIDEEVLMELWIEMGLDRIKE PELWLGEDEFDFMTD 1.62 supercharge6
TDFIDEEVLMSLVIEMGLDRIEELPELWLGQNEFDFMTD 1.70 RK>Emut20to39
TDFIDEEVLMSLVIEMGLDAIAELPELWLGQNEFDFMTD 1.75 RK>A
EDFIDEEVLMELVIEMGLDRIKELPELWLGQNEFDFMTD 1.79 ST>Emut0to20
TDFIDEEVLMSLVIEMGLDRIKELPELWLGDDEFDFMTD 1.86 QN>D
DDFIDEEVLMDLVIEMGLDRIKELPELWLGQNEFDFMDD 1.87 ST>D
TDFIDEEVLMSLVIEMGLDRIKELPELWLGEEEFDFMTD 1.93 QN>E
EDFIDEEVLMELVIEMGLDRIKELPELWLGQNEFDFMED 1.94 ST>E
DDFIDEEVLMDLVIEMGLDRIKELPELWLGDDEFDFMDD 1.98 STQN>D
TDFIDEEVLMSLVIEMGLDRIKELPELWLGQNEFDFMDD 2.00 ST>Dmut20to39
TDFIDEEVLMSLVIEMGLDRIKELPELWLGEEEFDFMED 2.00 STQN>Emut20to39
TDFIDEEVLMSLVIEMGLDEIEELPELWLGQNEFDFMTD 2.03 RK>E
DDFIDEEVLMDLVIEMGLDRIKELPELWLGQNEFDFMTD 2.06 ST>Dmut0to20
TDFIDEEVLMSLVIEMGLDRIKELPELWLGDDEFDFMDD 2.11 STQN>Dmut20to39
EDFIDEEVLMELVIEMGLDRIKELPELWLGEEEFDFMED 2.35 STQN>E
TDFIDEEFLFDLVIEMGLDRIKELPELWLGQNEFDFMTD 2.39 supercharge2
TDFIDEEVLMSLVIEMGLDRIAELPELWLGQNEFDFMTD 2.42 RK>Amut20to39

CITED2 more acidic and less active
TDFIDEEVLMDLFEEFGLDRIKELPELWLGQNEFDFMTD 1.44 supercharge5
TDFIDEEVLMSLVIEMGLDRIKELPELWLGQNEFDFMED 1.35 ST>Emut20to39
TDFIDEEVLMSLFFEMGLDRIKELPELWLGEWEFDFMTD 1.25 supercharge3
TDFIDEEVLMSLVIEMGLDDIKELPELWLGQNEFDFMTD 1.15 RK>Dmut0to20
TDFIDEEVLMSLVIEMGLDAIKELPELWLGQNEFDFMTD 1.10 RK>Amut0to20
TDFIDEEVLMSLVIEMGLDRIDELPELWLGQNEFDFMTD 1.06 RK>Dmut20to39
TDFIDEEVLMSLVIEMGLDRIKELPELWLDWNEFDFMTD 1.05 supercharge1
TDFIDEEVLMSLVIEMGLDDIDELPELWLGQNEFDFMTD 1.03 RK>D
TDFIDEEVLMSLVIEMGLDEFKELPELWLGQNEFDFMTD 0.85 supercharge4

CITED2 neutral and weaker
TEFIEEEVLMSLVIEMGLERIKELPELWLGQNEFEFMTE 0.29 ED>E
TEFIEEEVLMSLVIEMGLERIKELPELWLGQNEFDFMTD 0.24 ED>Emut0to20
```

```
CITED2 fewer Aromatics and more active
TDFIDEEVLMSLVIEMGLDRIKELPELWLGQNEFDFMTD  1.53  WT
TDLIDEEVLMSLVIEMGLDRIKELPELLLGQNELDLMTD  1.75  WF>L
TDLIDEEVLMSLVIEMGLDRIKELPELWLGQNELDLMTD  1.88  F>L
TDAIDEEVLMSLVIEMGLDRIKELPELWLGQNEFDFMTD  2.02  F>Amut0to20
TDFIDEEVLMSLVIEMGLDRIKELPELLLGQNELDLMTD  2.18  WF>Lmut20to39
TDFIDEEVLMSLVIEMGLDRIKELPELWLGQNELDLMTD  2.20  F>Lmut20to39
TDFIDEEVLMSLVIEMGLDRIKELPELLLGQNEFDFMTD  2.28  W>L

CITED2 fewer Aromatics and less active
TDFIDEEVLMSLVIEMGLDRIKELPELALGQNEFDFMTD  1.33  W>A
TDAIDEEVLMSLVIEMGLDRIKELPELALGQNEADAMTD  1.16  WF>A
TDAIDEEVLMSLVIEMGLDRIKELPELWLGQNEADAMTD  0.99  F>A
TDFIDEEVLMSLVIEMGLDRIKELPELWLGQNEADAMTD  0.92  F>Amut20to39
TDLIDEEVLMSLVIEMGLDRIKELPELWLGQNEFDFMTD  0.75  F>Lmut0to20
TDFIDEEVLMSLVIEMGLDRIKELPELALGQNEADAMTD  0.32  WF>Amut20to39

CITED2 more Aromatics and more active
TDFIDEEVLMELWIEMGLDRIKELPELWLGEDEFDFMTD  1.62  supercharge6
TDFIDEEYLMSLYIEMGLDRIKELPELWLGQNEFDFMTD  2.11  V>Y
TDFIDEEFLFDLVIEMGLDRIKELPELWLGQNEFDFMTD  2.39  supercharge2

CITED2 more Aromatics and less active
TDFIDEEFLMSLFIEMGLDRIKELPELWLGQNEFDFMTD  1.45  V>F
TDFIDEEVLMDLFEEFGLDRIKELPELWLGQNEFDFMTD  1.44  supercharge5
TDFIDEEWLMSLWIEMGLDRIKELPELWLGQNEFDFMTD  1.43  V>W
TDFYDEEVLMSLVYEMGLDRYKELPELWLGQNEFDFMTD  1.40  I>Y
TDFIDEEVLMSLFFEMGLDRIKELPELWLGEWEFDFMTD  1.25  supercharge3
TDFIDEEVLYSLVIEYGLDRIKELPELWLGQNEFDFYTD  1.25  M>Y
TDFIDEEVLWSLVIEWGLDRIKELPELWLGQNEFDFWTD  1.23  M>W
TDFIDEEVLMSLVIEMGLDRIKELPELWLDWNEFDFMTD  1.05  supercharge1
TDFFDEEVLMSLVFEMGLDRFKELPELWLGQNEFDFMTD  0.91  I>F
TDFIDEEVLMSLVIEMGLDEFKELPELWLGQNEFDFMTD  0.85  supercharge4
TDFIDEEVLFSLVIEFGLDRIKELPELWLGQNEFDFFTD  0.74  M>F
TDFIDEEVLMSLVIEMGLDRIKELPELWLGYYEFDFMTD  0.66  NQ>Y
TDFIDEEVLMSLVIEMGLDRIKELPELWLGWWEFDFMTD  0.36  NQ>W
TDFWDEEVLMSLVWEMGLDRWKELPELWLGQNEFDFMTD -0.04  I>W
TDFWDEEWWMSWWWEMGWDRWKEWPEWWWGQNEFDFMTD -0.28  IVL>W
TDFIDEEVLMSLVIEMGLDRIKELPELWLGFFEFDFMTD -0.28  NQ>F
TDFYDEEYYMSYYYEMGYDRYKEYPEYWYGQNEFDFMTD -0.52  IVL>Y
TDFIDEEVWMSWVIEMGWDRIKEWPEWWWGQNEFDFMTD -0.58  L>W
TDFFDEEFFMSFFFEMGFDRFKEFPEFWFGQNEFDFMTD -0.61  IVL>F
TDFIDEEVYMSYVIEMGYDRIKEYPEYWYGQNEFDFMTD -0.67  L>Y
TDFIDEEVFMSFVIEMGFDRIKEFPEFWFGQNEFDFMTD -0.82  L>F

CITED2 Exchange Aromatics and more active
TDYIDEEVLMSLVIEMGLDRIKELPELWLGQNEFDFMTD  1.57  F>Ymut0to20
TDFIDEEVLMSLVIEMGLDRIKELPELYLGQNEFDFMTD  2.00  W>Y
TDFIDEEVLMSLVIEMGLDRIKELPELWLGQNEWDWMTD  2.32  F>Wmut20to39
TDWIDEEVLMSLVIEMGLDRIKELPELWLGQNEFDFMTD  2.35  F>Wmut0to20

CITED2 Exchange Aromatics and less active
TDFIDEEVLMSLVIEMGLDRIKELPELWLGQNEYDYMTD  0.81  F>Ymut20to39
TDYIDEEVLMSLVIEMGLDRIKELPELWLGQNEYDYMTD  0.73  F>Y
TDYIDEEVLMSLVIEMGLDRIKELPELYLGQNEYDYMTD  0.73  WF>Y
TDFIDEEVLMSLVIEMGLDRIKELPELYLGQNEYDYMTD  0.70  WF>Ymut20to39
```

Figure 4: In CITED2 the locations of introduced acidic residues impacts whether variants have increased or decreased activity.

A) Variants that added acidity and had increased activity are red. Variants that added acidity and had WT or lower activity are blue. B) The location of the motifs and alpha helix (arrow). C) The sequences of variants from A. Red variants tend to add negative charge in the flanks, near the motifs. Blue variants tend to remove positive charges in the center or be 'supercharge' variants that add acidic and aromatic residues. Most electrostatically neutral substitutions had a small effect on activity, but two variants (orange) that substituted E for D in the N terminal region had reduced activity. In the main text, we argue this effect is mostly likely because D224, which is the most conserved position in the AD (Figure S13), is tightly sandwiched between two positively charged residues of TAZ1 (Figure S15), and replacing the small D with a larger E interferes with the electrostatic interaction. D) Summary of the locations of substitutions in the red and blue sets of variants. The fraction of variants with a substitution at each position is shown. Adding negative charges in the flanks, near the motifs, tends to increase activity while adding acidity in the center tends not to increase activity. E-F): CITED2 does not require aromatic residues for activity. Variants with fewer aromatic residues and higher activity (brown) either replaced aromatics with leucines or were F222A. Variants with fewer aromatic residues and lower activity (green). To our surprise, replacing all aromatics with alanine (WF>A) only mildly decreased activity. Most variants that replaced aromatics with alanines decreased activity. Aromatic residues contribute to activity but are not critical like F442 in VP16. In nearly all cases, adding aromatic residues decreased activity (pink). Two of the three variants that added aromatic residues and increased activity were ‘supercharge’ variants that also added acidic residues. Substituting F’s with W’s can increase activity (lime green). Substituting the F’s with Y’s generally decreased activity (orange). For most substitutions, F222 had the opposite response from F253 and F255. In both panels, WT mean and SEM are shown with a black line and gray horizontal box. There is a linear regression line and CI plotted to guide the eye.

After accounting for the increased contribution of leucine residues, we found that AD activity requires a balance between 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 5A). Neither the W,F,Y&L count nor net negative charge is sufficient for activity; both are necessary. Counting only aromatic residues does not separate high and low activity variants as well as counting W,F,Y&L residues (Figure S11). This separation is also visible when we normalize by AD length or substitute W,F,Y&L count with hydrophobicity (Figure S11). There are points on this grid occupied by both strong and weak variants (Figure S11), indicating that composition is not the sole determinant of activity and that the arrangement of residues matters. When we used machine learning classifiers to separate active and inactive variants, leucines made larger contributions to model performance than any individual aromatic residue (Table S1 and Figure S11). The balance between W,F,Y&L and acidic residues is critical for AD activity.

As a further test of the Acidic Exposure Model, we examined whether the combination of acidic and W,F,Y&L residues could predict known and new 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”), and for each tile we calculated the net charge and counted W,F,Y&L residues. Tiles that are similar to strong 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 (Figure 5B, red), which combine to predict 144 ADs distributed across 136 TFs. These predicted ADs overlap with 17 Uniprot ADs - far more than expected by chance (p<1e-5 in permutation tests). Given the extreme simplicity of the predictor, this predictive power is remarkable. In addition, we accurately predicted three known ADs that are not in Uniprot, including the N terminal AD of c-Myc (Andresen et al., 2012) and the Znf473 KRAB domain which has been recently shown to be an AD (Tycko et al., 2020). This analysis shows how counting eight specific amino acids (acidic, basic, aromatic and leucine residues) is sufficient to find known ADs and predict new ones.

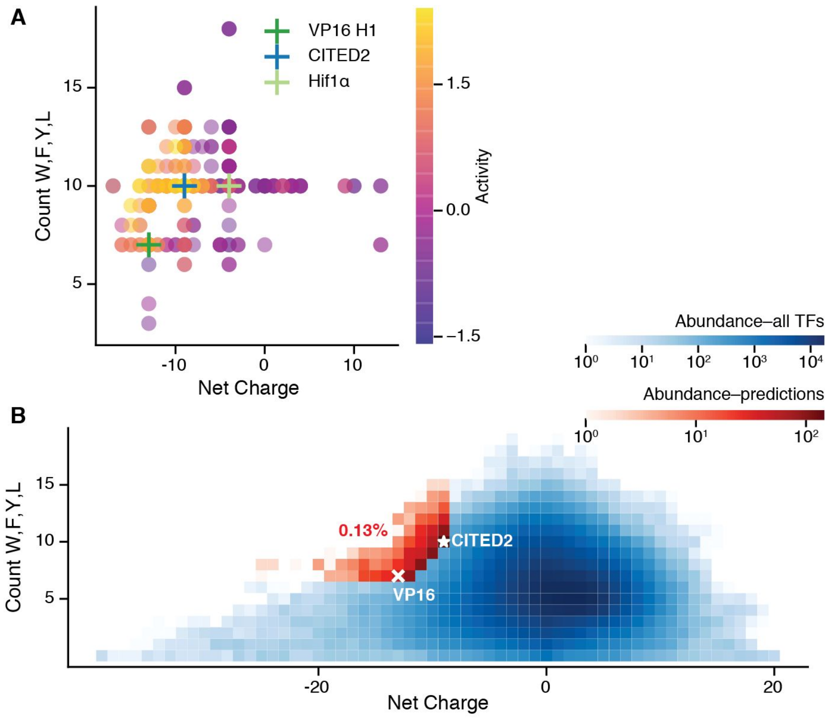


Figure 5. Strong activation domains are acidic and contain many W,F,Y&L residues.
A) The location of each point indicates the net charge and the number of W,F,Y&L residues, while the color indicates activity. All variants of VP16, CITED2 and Hif1ɑ are plotted. B) A heatmap of all 39AA tiles from human TFs. The pixel location indicates the net charge and W,F,Y&L count, while the blue intensity indicates the number of tiles with that combination. Only 0.13% of tiles (red) are as extreme or more extreme than VP16 (x) and CITED2 (*).

The AD predictor required counting leucine residues, prompting us to investigate the molecular role of leucines. Summarizing the activities of all variants with a substitution at each position highlighted how mutating leucines decreased activity (Figure 6A, S12). Moreover, the residues with the lowest mean activities when mutated all point towards the coactivator surface in NMR structures of CITED2 and Hif1ɑ (Figure 6B, S12). These same positions are also the most conserved positions of these ADs (Figure S13). These results support the hypothesis that the structural constraints of the interaction interfaces impose negative selection on the residues that make contact.

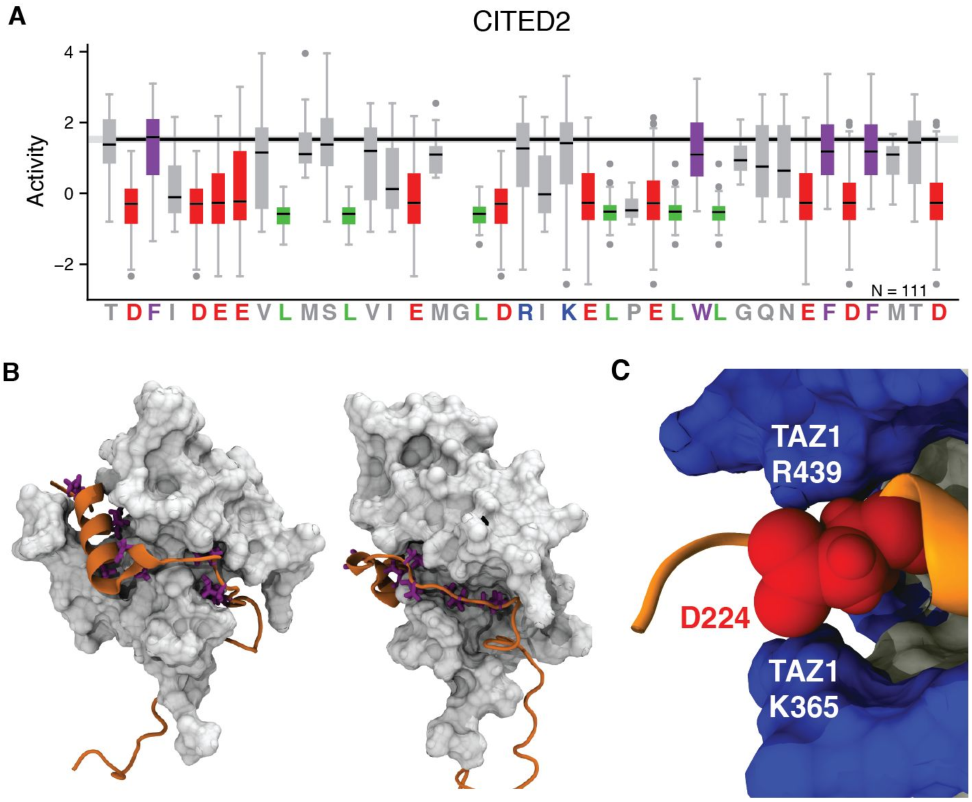


Figure 6. Rational mutagenesis detects the structural constraints of AD-coactivator interactions. A) The distribution of activities for all variants that change each position of CITED2. Acidic (red), aromatic (purple) and leucine (green) residues. Medians, gray bars. Outliers, gray dots. B) The residues of CITED2 (orange) with lowest mean positions in A (pink) point towards the TAZ1 coactivator (white, 1R8U). C) D224 (red) of CITED2 is sandwiched between the narrowest point of the basic rim (blue) of the binding canyon of TAZ1. See Figure S15 for snapshots of all 20 structures in 1R8U.

The mechanism by which leucine residues make large contributions to activity is exemplified by the CITED2 interaction with the TAZ1 domain of CBP/p300. TAZ1 has a canyon with a hydrophobic floor and basic rim that tightly embraces the compact alpha helix of CITED2 (Figure 7). 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. Replacing leucines with aromatics decreases activity because the larger side chains do not fit in the canyon. Disrupting the helix folding, either by adding two prolines or shuffling the sequence (Figure S13), causes expansion, reducing activity. Conversely, the 2xGlycine variant increased activity because it had smaller side chains and retained helicity. This structural constraint also explains why a conservative substitution, replacing aspartic acid residues (D) with larger glutamic acid residues (E), reduced activity: the D224 side chain sits between the narrowest point of the basic canyon rim, sandwiched between R439 and K365 of TAZ1, and the D244E substitution impairs this fit (Figure 4C, S6, S20). If the signature we see in CITED2 carries over to other ADs that bind this TAZ1 canyon, then our results, along with other studies (Diss and Lehner, 2018; Rollins et al., 2019; Schmiedel and Lehner, 2019), could predict AD-coactivator interactions.

Contrasting the CITED2-TAZ1 interaction with the Gcn4-Med15 interaction explains why aromatic residues make large contributions to activity in Gcn4 (Berlow et al., 2017; Brzovic et al., 2011). Both ADs fold into alpha helices and both coactivators contain an AD binding canyon with a hydrophobic floor and basic rim (Figure 7). On TAZ1, the canyon is large and the CITED2 alpha helix is fully engulfed. Leucines fit this structure better than aromatics because they are smaller and promote helix formation (Pace and Scholtz, 1998). On Med15, the canyon is shallow and Gcn4 only inserts side chains. Aromatics fit this structure better than leucines because they are large and better reach the hydrophobic canyon floor. The Gcn4-Med15 interaction is fuzzy with few steric constraints, explaining why Gcn4 is poorly conserved and more tolerant of substitutions in mutagenesis experiments (Brzovic et al., 2011; Erijman et al., 2020; Staller et al., 2018). The increased importance of leucine residues in human cells likely reflects the structural constraints imposed by an expanded repertoire of coactivators.

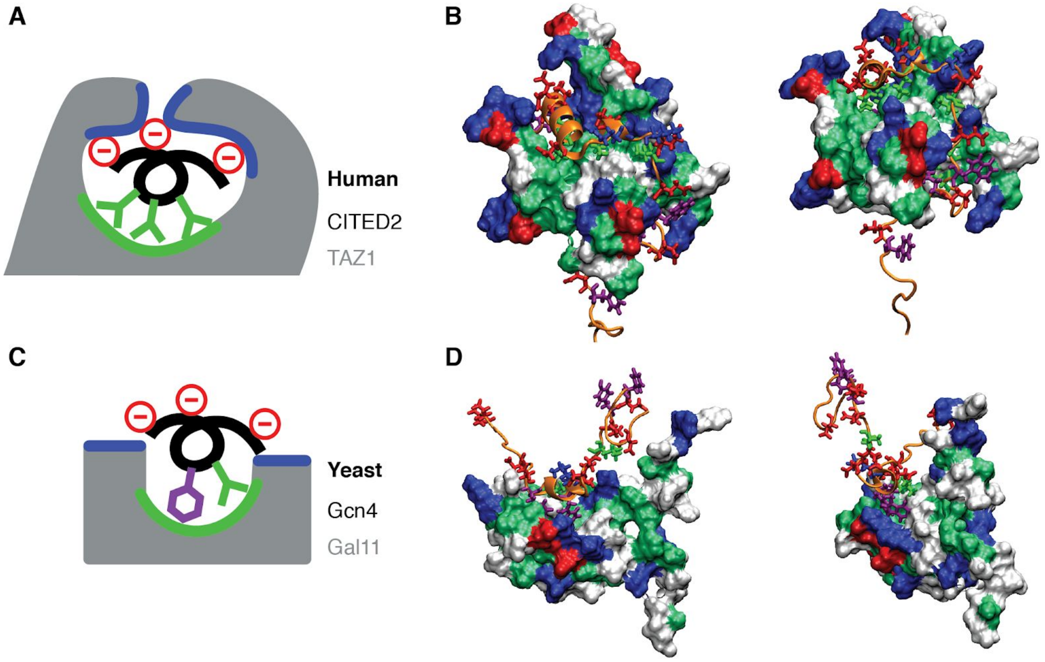


Figure 7: The structure of the coactivator AD-binding canyon constrains AD sequence.
A) The CITED2 AD is inside the Taz1 canyon, a structural constraint that favors leucine residues. The yeast Gcn4 AD is outside the Med15/Gal11 canyon, enabling a fuzzy interaction that favors aromatic residues. B-C) Structures of the CITED2-TAZ1 and Gcn4-Med15 interactions. Both ADs fold into amphipathic alpha helices (orange) that present the hydrophobic residues (Leucine, lime green or aromatic, purple) on one side and the acidic residues (red) on the other side. TAZ1 and Med15 each contain an AD binding canyon with a hydrophobic floor (dark green) and basic rim (blue). B) On TAZ1, the canyon is large and the CITED2 alpha helix is fully engulfed. The right panel gazes down the barrel of the helix, illustrating the TAZ1 embrace. Structure 1 of 1R8U. The yeast Gcn4 AD is outside the Med15/Gal11 canyon, enabling a fuzzy interaction that favors aromatic residues. C) On Med15, the canyon is shallow, the alpha helix of Gcn4 remains outside, and Gcn4 only inserts side chains. The Gcn4-Med15 interaction is fuzzy with few steric constraints, Gcn4 binds with the helix assuming multiple orientations (Brzovic et al., 2011; Scholes and Weinzierl, 2016). Structure 10 of 2LPB.

Although composition is a key determinant of AD activity, we found three ways in which the arrangement of amino acids (i.e. sequence) is also important for function. First, for VP16 and CITED2, 80% of shuffle variants, which maintain composition but rearranged the order of amino acid residues in a region of the AD, disrupted activity (Figure S14, S16). Second, for VP16 and CITED2, forming an alpha helix was necessary for activity because variants that disrupted helicity also reduced activity (Figure S14, S16). Third, adding acidic residues adjacent to W,F,Y&L residues was more likely to increase activity (Figure 4, S5-6). 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). Fourth, adding dipeptides increased the explanatory power of composition based ANOVA models (Figure S17). Interspersing acidic residues between W,F,Y&L residues more efficiently exposes the motifs.

Discussion

Our Acidic Exposure Model explains why many ADs are acidic and unstructured (Sigler, 1988): the acidic residues and intrinsic disorder keep aromatic and leucine residues exposed and available to bind coactivators. In principle, the main role of acidity, exposing hydrophobic residues, could have been achieved by positively charged amino acids, but negative charges repel DNA, preventing non-specific DNA binding and enabling evolution to independently tune the affinity of the DNA binding domain and the strength of the AD. Acidity further enables biphasic binding with positively charged coactivators (Ferreira et al., 2005; Hermann et al., 2001) and intramolecular interactions between acidic ADs and positively charged DBDs that can enhance DNA specificity (Krois et al., 2018). The intrinsic disorder enables ADs to assume different conformations when bound to each partner (Dyson and Wright, 2016) and dynamically exposes the motifs, reducing the entropic cost of exposure compared to constant exposure. Dynamic exposure likely has a larger sequence solution space that is more evolutionarily accessible.

This Acidic Exposure Model resolves three decades of conflicting data. Net acidity and intrinsic disorder are conserved while individual residues are poorly conserved because the effect of acidity and disorder is emergent (Cress and Triezenberg, 1991; Martchenko et al., 2007). Short linear motifs of aromatics and leucines make the largest contributions to activity because they directly contact the coactivators (Dyson and Wright, 2016; Latchman, 2008). Amphipathic alpha helices are an effective way to present a motif as a continuous binding surface, but the helix is not strictly necessary (Brzovic et al., 2011; Cress and Triezenberg, 1991; Giniger and Ptashne, 1987). Leucine residues outside of the core motif contribute to binding by increasing the valence of the interaction, extending the very flexible, “fuzzy” interaction surfaces (Jackson et al., 1996; Tuttle et al., 2019; Warfield et al., 2014). Binding is distributed over many residues in a sometimes redundant manner, explaining why ADs are robust to mutation: missense mutations in ADs are underrepresented in patient samples and ADs have lower signal in deep mutational scans than DBDs (Giacomelli et al., 2018; Johnson and Anthony Weil, 2017; Majithia et al., 2016). On evolutionary time scales, individual residues are easily replaced, explaining the both high degree of evolutionary plasticity and the conservation of function without conservation of primary sequence (Martchenko et al., 2007; Staller et al., 2018).

The Acidic Exposure Model has been further corroborated by two papers that screened random peptides for AD activity in yeast: in both works aromatic and acidic residues are the strongest predictors of activity (Erijman et al., 2020; Ravarani et al., 2018). Both papers found an enrichment of [D/E][W/F/Y] dipeptide motifs in AD-like peptides. In our work, we had limited evidence that placing positive residues adjacent to a hydrophobic motif decreased activity (Staller et al., 2018), and these random peptide screens clearly showed that positive residues decreased AD activity. The random peptide screens also hinted that clusters of aromatics surrounded by acidic residues are sufficient for function suggesting that the motifs do not necessarily require strict grammar or spacing (Erijman et al., 2020).

It is likely that the acidic exposure model only explains one class of acidic ADs and other classes remain (Latchman, 2008). Proline-rich ADs might function in a similar way because proline rich sequences are highly expanded and can keep motifs exposed (Gerber et al., 1994; Martin et al., 2016). Glutamine-rich ADs are likely to be a different functional class entirely, because many have been recategorized as prion-like domains, some of them phase separate, some do not work in yeast and others only activate metazoan promoters (Boija et al., 2018; Boulay et al., 2017; Chong et al., 2017; Kwon et al., 2013; Seipel et al., 1992; Shin and Brangwynne, 2017; Tanaka et al., 1994). Isoleucine-rich ADs from *Drosophila* (Attardi and Tjian, 1993) are likely yet another class because, in our assays, isoleucine residues do not make large contributions to activity (Johnson and Anthony Weil, 2017; Staller et al., 2018).

High throughput methods for assaying ADs promise to close the understanding gap with DBDs (Arnold et al., 2018; Erijman et al., 2020; Ravarani et al., 2018; Staller et al., 2018; Tycko et al., 2020). Our current AD method faces a tradeoff between throughput and resolution. Measuring small changes in activity requires many barcodes per AD while measuring large changes in activity requires fewer barcodes. For a fixed number of barcodes, we can either screen 500-1000 variants at high resolution or more variants at low resolution. In principle, screening for potential ADs for activity can be achieved with very few barcodes per variant. Deep mutational scanning (DMS) of ADs will require many barcodes per variant because most point mutations in disordered regions or ADs have small effects on activity (Giacomelli et al., 2018; Kotler et al., 2018; Majithia et al., 2016). Our method can be adapted to screen for functional effects of clinically identified variants of uncertain significance (VUS), albeit at moderate throughput.

For ADs and other IDRs, rational mutagenesis offers advantages over DMS. Protein sequence space is too large to survey systematically or randomly. Point mutants in ADs and IDRs generally have small effects (Giacomelli et al., 2018; Majithia et al., 2016). In our data, mutating one acidic residue frequently had little or no effect on AD activity, but removing multiple acidic residues significantly reduced activity. Similarly adding one acidic residue sometimes increased activity while adding multiple residues was more likely to increase activity. The function of negative charge is collective and emergent. A common use of DMS is built look up tables for clinical missense VUS, but pathological variants are less common in IDRs compared to globular domains (Mészáros et al., 2020; Starita et al., 2017). Rational mutagenesis is most effective when the designed variants are focused on testing a few specific hypotheses. For uncharacterized protein regions, initial DMS or random mutagenesis may be necessary to generate hypotheses that can then be tested with rational mutagenesis (Bolognesi et al., 2019, 2016; Esposito et al., 2019).

We have found three design principles for mammalian ADs. First, sequence composition: strong ADs require a balance between W,F,Y&L residues and acidic residues. Second, sequence grammar: acidic residues are more potent when adjacent to the W,F,Y&L residues. Third motif spelling: within motifs, the choice between aromatic and leucine residues reflects the structural constraints of AD-coactivator interactions. Going forward these rules will refine computational models for predicting ADs, guide engineering of new ADs, predict

AD-coactivator interactions, and inform models that predict the impact of genetic variation on AD function. These design principles apply to ADs that bind structurally distinct coactivators in humans and yeast. Metazoans have more numerous and more structurally diverse coactivators than yeast, and the larger role for leucines likely reflects this elaboration of targets and new structures. For example, the TAZ1 domain is not present in yeast (El-Gebali et al., 2019). It appears that acidic activation domains have diversified to bind an elaborated ensemble of partners. We speculate that structural flexibility allows evolutionary plasticity.

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 Rottweiler for bioinformatics help; Jessica Hoisington-Lopez and MariaLynn Crosby for DNA sequencing. We thank members of the Cohen Lab for helpful discussions and comments on the manuscript.

Author contributions: MVS: Conceptualization, Methodology, Software, Investigation, Writing - Original Draft, Writing - Review & Editing ER: Investigation, Software ASH: Software RVP: Writing - Review & Editing BAC: Conceptualization, Writing - Review & Editing. **Funding:** MVS was supported by Burroughs Wellcome Fund Postdoctoral Enrichment Program, American Cancer Society Postdoctoral Fellowship, and NIGMS K99131022. ER was supported on 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.

Data availability: All processed activity data is available in the supplementary materials. Simulation data is freely available upon request. The raw data will be available in the NIH GEO database. Our code is available on request and will be eventually available on Github before publication. The plasmids will be deposited in AddGene.

References

Andresen C, Helander S, Lemak A, Farès C, Csizmok V, Carlsson J, Penn LZ, Forman-Kay JD, Arrowsmith CH, Lundström P, Sunnerhagen M. 2012. Transient structure and dynamics in the disordered c-Myc transactivation domain affect Bin1 binding. *Nucleic Acids Res* **40**:6353–6366. doi:10.1093/nar/gks263

Arnold CD, Nemčko F, Woodfin AR, Wienerroither S, Vlasova A, Schleiffer A, Pagani M, Rath M, Stark A. 2018. A high-throughput method to identify trans-activation domains within transcription factor sequences. *EMBO J* **37**:e98896.

Attardi LD, Tjian R. 1993. Drosophila tissue-specific transcription factor NTF-1 contains a novel isoleucine-rich activation motif. *Genes & Development*. doi:10.1101/gad.7.7b.1341

Berlow RB, Dyson HJ, Wright PE. 2017. Hypersensitive termination of the hypoxic response by a disordered protein switch. *Nature* **543**:447–451. doi:10.1038/nature21705

Boija A, Klein IA, Sabari BR, Dall'Agnese A, Coffey EL, Zamudio AV, Li CH, Shrinivas K, Manteiga JC, Hannett NM, Abraham BJ, Afeyan LK, Guo YE, Rimel JK, Fant CB, Schuijers J, Lee TI, Taatjes DJ, Young RA. 2018. Transcription Factors Activate Genes through the Phase-Separation Capacity of Their Activation Domains. *Cell* **175**:1842–1855.e16. doi:10.1016/j.cell.2018.10.042

Bolognesi B, Faure AJ, Seuma M, Schmiedel JM, Tartaglia GG, Lehner B. 2019. The mutational landscape of a prion-like domain. *Nat Commun* **10**:4162. doi:10.1038/s41467-019-12101-z

Bolognesi B, Lorenzo Gotor N, Dhar R, Cirillo D, Baldrighi M, Tartaglia GG, Lehner B. 2016. A Concentration-Dependent Liquid Phase Separation Can Cause Toxicity upon Increased Protein Expression. *Cell Rep* **16**:222–231. doi:10.1016/j.celrep.2016.05.076

Boulay G, Sandoval GJ, Riggi N, Iyer S, Buisson R, Naigles B, Awad ME, Rengarajan S, Volorio A, McBride MJ, Broye LC, Zou L, Stamenkovic I, Kadoch C, Rivera MN. 2017. Cancer-Specific Retargeting of BAF Complexes by a Prion-like Domain. *Cell* **171**:163–178.e19. doi:10.1016/j.cell.2017.07.036

Brzovic PS, Heikaus CC, Kisselev L, Vernon R, Herbig E, Pacheco D, Warfield L, Littlefield P, Baker D, Klevit RE, Hahn S. 2011. 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. doi:10.1016/j.molcel.2011.11.008

Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. 2013. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat Methods* **10**:1213–1218. doi:10.1038/nmeth.2688

Bulyk ML. 2007. Protein binding microarrays for the characterization of DNA-protein interactions. *Adv Biochem Eng Biotechnol* **104**:65–85.

Chang J, Kim DH, Lee SW, Choi KY, Sung YC. 1995. Transactivation ability of p53 transcriptional activation domain is directly related to the binding affinity to TATA-binding protein. *J Biol Chem* **270**:25014–25019. doi:10.1074/jbc.270.42.25014

Chong S, Dugast-Darzacq C, Liu Z, Dong P, Dailey G. 2017. Dynamic and Selective Low-Complexity Domain Interactions Revealed by Live-Cell Single-Molecule Imaging. *bioRxiv*.

Chong S, Dugast-Darzacq C, Liu Z, Dong P, Dailey GM, Cattoglio C, Heckert A, Banala S, Lavis L, Darzacq X, Tjian R. 2018. Imaging dynamic and selective low-complexity domain interactions that control gene transcription. *Science* **361**. doi:10.1126/science.aar2555

Cress WD, Triezenberg SJ. 1991. Critical structural elements of the VP16 transcriptional activation domain. *Science* **251**:87–90.

Diss G, Lehner B. 2018. The genetic landscape of a physical interaction. *Elife* **7**. doi:10.7554/eLife.32472

Drysdale CM, Duenas E, Jackson BM, Reusser U, Braus GH, Hinnebusch AG. 1995. The transcriptional activator GCN4 contains multiple activation domains that are critically dependent on hydrophobic amino acids. *Mol Cell Biol* **15**:1220–1233.

Dyson HJ, Wright PE. 2016. 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. doi:10.1074/jbc.R115.692020

El-Gebali S, Mistry J, Bateman A, Eddy SR, Luciani A, Potter SC, Qureshi M, Richardson LJ, Salazar GA, Smart A, Sonnhammer ELL, Hirsh L, Paladin L, Piovesan D, Tosatto SCE, Finn RD. 2019. The Pfam protein families database in 2019. *Nucleic Acids Res* **47**:D427–D432. doi:10.1093/nar/gky995

Erijman A, Kozlowski L, Sohrabi-Jahromi S, Fishburn J, Warfield L, Schreiber J, Noble WS, Söding J, Hahn S. 2020. A High-Throughput Screen for Transcription Activation Domains Reveals Their Sequence Features and Permits Prediction by Deep Learning. *Mol Cell* **78**:890–902.e6. doi:10.1016/j.molcel.2020.04.020

Esposito D, Weile J, Shendure J, Starita LM, Papenfuss AT, Roth FP, Fowler DM, Rubin AF. 2019. MaveDB: an open-source platform to distribute and interpret data from multiplexed assays of variant effect. *Genome Biol* **20**:223. doi:10.1186/s13059-019-1845-6

Ferreira ME, Hermann S, Prochasson P, Workman JL, Berndt KD, Wright APH. 2005. Mechanism of transcription factor recruitment by acidic activators. *J Biol Chem* **280**:21779–21784. doi:10.1074/jbc.M502627200

Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, Potter SC, Punta M, Qureshi M, Sangrador-Vegas A, Salazar GA, Tate J, Bateman A. 2016. The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res* **44**:D279–85. doi:10.1093/nar/gkv1344

Freedman SJ, Sun Z-YJ, Kung AL, France DS, Wagner G, Eck MJ. 2003. Structural basis for negative regulation of hypoxia-inducible factor-1α by CITED2. *Nat Struct Mol Biol* **10**:504–512. doi:10.1038/nsb936

Gerber HP, Seipel K, Georgiev O, Höfferer M, Hug M, Rusconi S, Schaffner W. 1994. Transcriptional activation modulated by homopolymeric glutamine and proline stretches. *Science* **263**:808–811. doi:10.1126/science.8303297

Giacomelli AO, Yang X, Lintner RE, McFarland JM, Duby M, Kim J, Howard TP, Takeda DY, Ly SH, Kim E, Gannon HS, Hurhula B, Sharpe T, Goodale A, Fritchman B, Steelman S, Vazquez F, Tsherniak A, Aguirre AJ, Doench JG, Piccioni F, Roberts CWM, Meyerson M, Getz G, Johannessen CM, Root DE, Hahn WC. 2018. Mutational processes shape the landscape of TP53 mutations in human cancer. *Nat Genet* **50**:1381–1387. doi:10.1038/s41588-018-0204-y

Giniger E, Ptashne M. 1987. Transcription in yeast activated by a putative amphipathic α helix linked to a DNA binding unit. *Nature*. doi:10.1038/330670a0

Hahn S. 1993. Structure(?) and function of acidic transcription activators. *Cell* **72**:481–483.

Hahn S, Young ET. 2011. Transcriptional regulation in Saccharomyces cerevisiae: transcription factor regulation and function, mechanisms of initiation, and roles of activators and coactivators. *Genetics* **189**:705–736. doi:10.1534/genetics.111.127019

Hawkins JA, Jones SK Jr, Finkelstein IJ, Press WH. 2018. Indel-correcting DNA barcodes for high-throughput sequencing. *Proc Natl Acad Sci U S A* **115**:E6217–E6226. doi:10.1073/pnas.1802640115

Hermann S, Berndt KD, Wright AP. 2001. How transcriptional activators bind target proteins. *J Biol Chem*

276:40127–40132. doi:10.1074/jbc.M103793200

Holehouse AS, Das RK, Ahad JN, Richardson MOG, Pappu RV. 2017. CIDER: Resources to Analyze Sequence-Ensemble Relationships of Intrinsically Disordered Proteins. *Biophys J* **112**:16–21. doi:10.1016/j.bpj.2016.11.3200

Hope IA, Mahadevan S, Struhl K. 1988. Structural and functional characterization of the short acidic transcriptional activation region of yeast GCN4 protein. *Nature* **333**:635–640. doi:10.1038/333635a0

Humphrey W, Dalke A, Schulten K. 1996. VMD: visual molecular dynamics. *J Mol Graph* **14**:33–8, 27–8. doi:10.1016/0263-7855(96)00018-5

Jackson BM, Drysdale CM, Natarajan K, Hinnebusch AG. 1996. Identification of seven hydrophobic clusters in GCN4 making redundant contributions to transcriptional activation. *Mol Cell Biol* **16**:5557–5571.

Johnson AN, Anthony Weil P. 2017. Identification of a transcriptional activation domain in yeast repressor activator protein 1 (Rap1) using an altered DNA-binding specificity variant. *Journal of Biological Chemistry*. doi:10.1074/jbc.m117.779181

Jonker HRA, Wechselberger RW, Boelens R, Folkers GE, Kaptein R. 2005. Structural properties of the promiscuous VP16 activation domain. *Biochemistry* **44**:827–839. doi:10.1021/bi0482912

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

Kinney JB, Murugan A, Callan CG Jr, Cox EC. 2010. Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence. *Proc Natl Acad Sci U S A* **107**:9158–9163. doi:10.1073/pnas.1004290107

Kotler E, Shani O, Goldfeld G, Lotan-Pompan M, Tarcic O, Gershoni A, Hopf TA, Marks DS, Oren M, Segal E. 2018. A Systematic p53 Mutation Library Links Differential Functional Impact to Cancer Mutation Pattern and Evolutionary Conservation. *Mol Cell* **71**:873. doi:10.1016/j.molcel.2018.08.013

Krois AS, Dyson HJ, Wright PE. 2018. Long-range regulation of p53 DNA binding by its intrinsically disordered N-terminal transactivation domain. *Proc Natl Acad Sci U S A* **115**:E11302–E11310. doi:10.1073/pnas.1814051115

Krois AS, Ferreon JC, Martinez-Yamout MA, Dyson HJ, Wright PE. 2016. Recognition of the disordered p53 transactivation domain by the transcriptional adapter zinc finger domains of CREB-binding protein. *Proc Natl Acad Sci U S A* **113**:E1853–62. doi:10.1073/pnas.1602487113

Kwon I, Kato M, Xiang S, Wu L, Theodoropoulos P, Mirzaei H, Han T, Xie S, Corden JL, McKnight SL. 2013. Phosphorylation-regulated binding of RNA polymerase II to fibrous polymers of low-complexity domains. *Cell* **155**:1049–1060. doi:10.1016/j.cell.2013.10.033

Kyte J, Doolittle RF. 1982. A simple method for displaying the hydropathic character of a protein. *J Mol Biol* **157**:105–132.

Lambert SA, Jolma A, Campitelli LF, Das PK, Yin Y, Albu M, Chen X, Taipale J, Hughes TR, Weirauch MT. 2018. The Human Transcription Factors. *Cell* **175**:598–599. doi:10.1016/j.cell.2018.09.045

Lando D, Peet DJ, Gorman JJ, Whelan DA, Whitelaw ML, Bruick RK. 2002. FIH-1 is an asparaginyl hydroxylase enzyme that regulates the transcriptional activity of hypoxia-inducible factor. *Genes Dev* **16**:1466–1471. doi:10.1101/gad.991402

Latchman DS. 2008. Eukaryotic Transcription Factors, 5th Edition. ed. Elsevier Science.

Lecoq L, Raiola L, Chabot PR, Cyr N, Arseneault G, Legault P, Omichinski JG. 2017. Structural characterization of interactions between transactivation domain 1 of the p65 subunit of NF-κB and transcription regulatory factors. *Nucleic Acids Res* **45**:5564–5576. doi:10.1093/nar/gkx146

Lin J, Chen J, Elenbaas B, Levine AJ. 1994. Several hydrophobic amino acids in the p53 amino-terminal domain are required for transcriptional activation, binding to mdm-2 and the adenovirus 5 E1B 55-kD protein. *Genes Dev* **8**:1235–1246. doi:10.1101/gad.8.10.1235

Liu J, Perumal NB, Oldfield CJ, Su EW, Uversky VN, Dunker AK. 2006. Intrinsic disorder in transcription factors. *Biochemistry* **45**:6873–6888. doi:10.1021/bi0602718

Majithia AR, Tsuda B, Agostini M, Gnanapradeepan K, Rice R, Peloso G, Patel KA, Zhang X, Broekema MF, Patterson N, Duby M, Sharpe T, Kalkhoven E, Rosen ED, Barroso I, Ellard S, UK Monogenic Diabetes Consortium, Kathiresan S, Myocardial Infarction Genetics Consortium, O'Rahilly S, UK Congenital Lipodystrophy Consortium, Chatterjee K, Florez JC, Mikkelsen T, Savage DB, Altshuler D. 2016. Prospective functional classification of all possible missense variants in PPARG. *Nat Genet* **48**:1570–1575. doi:10.1038/ng.3700

Ma J, Ptashne M. 1987. A new class of yeast transcriptional activators. *Cell* **51**:113–119.

Maricque BB, Chaudhari HG, Cohen BA. 2018. A massively parallel reporter assay dissects the influence of chromatin structure on cis-regulatory activity. *Nat Biotechnol*. doi:10.1038/nbt.4285

Martchenko M, Levitin A, Whiteway M. 2007. Transcriptional activation domains of the Candida albicans Gcn4p and Gal4p homologs. *Eukaryot Cell* **6**:291–301. doi:10.1128/EC.00183-06

Martin EW, Holehouse AS, Grace CR, Hughes A, Pappu RV, Mittag T. 2016. Sequence Determinants of the Conformational Properties of an Intrinsically Disordered Protein Prior to and upon Multisite Phosphorylation. *J*

Am Chem Soc **138**:15323–15335. doi:10.1021/jacs.6b10272

Mészáros B, Hajdu-Soltész B, Zeke A, Dosztányi Z. 2020. Intrinsically disordered protein mutations can drive cancer and their targeted interference extends therapeutic options. *bioRxiv*.

Metskas LA, Rhoades E. 2015. Conformation and Dynamics of the Troponin I C-Terminal Domain: Combining Single-Molecule and Computational Approaches for a Disordered Protein Region. *J Am Chem Soc* **137**:11962–11969. doi:10.1021/jacs.5b04471

Noyes MB, Meng X, Wakabayashi A, Sinha S, Brodsky MH, Wolfe SA. 2008. A systematic characterization of factors that regulate Drosophila segmentation via a bacterial one-hybrid system. *Nucleic Acids Res* **36**:2547–2560. doi:10.1093/nar/gkn048

Oldfield CJ, Dunker AK. 2014. Intrinsically disordered proteins and intrinsically disordered protein regions. *Annu Rev Biochem* **83**:553–584. doi:10.1146/annurev-biochem-072711-164947

Pace CN, Scholtz JM. 1998. A helix propensity scale based on experimental studies of peptides and proteins. *Biophys J* **75**:422–427. doi:10.1016/s0006-3495(98)77529-0

Park M, Patel N, Keung AJ, Khalil AS. 2019. Engineering Epigenetic Regulation Using Synthetic Read-Write Modules. *Cell*. doi:10.1016/j.cell.2018.11.002

Raj N, Attardi LD. 2017. The Transactivation Domains of the p53 Protein. *Cold Spring Harb Perspect Med* **7**. doi:10.1101/cshperspect.a026047

Ravarani CN, Erkina TY, De Baets G, Dudman DC, Erkine AM, Babu MM. 2018. High-throughput discovery of functional disordered regions: investigation of transactivation domains. *Mol Syst Biol* **14**:e8190. doi:10.15252/msb.20188190

Regier JL, Shen F, Triezenberg SJ. 1993. Pattern of aromatic and hydrophobic amino acids critical for one of two subdomains of the VP16 transcriptional activator. *Proc Natl Acad Sci U S A* **90**:883–887.

Riley TR, Slattery M, Abe N, Rastogi C, Liu D, Mann RS, Bussemaker HJ. 2014. SELEX-seq: a method for characterizing the complete repertoire of binding site preferences for transcription factor complexes. *Methods Mol Biol* **1196**:255–278. doi:10.1007/978-1-4939-1242-1_16

Rollins NJ, Brock KP, Poelwijk FJ, Stiffler MA, Gauthier NP, Sander C, Marks DS. 2019. Inferring protein 3D structure from deep mutation scans. *Nat Genet* **51**:1170–1176. doi:10.1038/s41588-019-0432-9

Schmiedel JM, Lehner B. 2019. Determining protein structures using deep mutagenesis. *Nat Genet* **51**:1177–1186. doi:10.1038/s41588-019-0431-x

Scholes NS, Weinzierl ROJ. 2016. Molecular Dynamics of "Fuzzy" Transcriptional Activator-Coactivator Interactions. *PLoS Comput Biol* **12**:e1004935. doi:10.1371/journal.pcbi.1004935

Seipel K, Georgiev O, Schaffner W. 1992. Different activation domains stimulate transcription from remote ('enhancer') and proximal ('promoter') positions. *EMBO J* **11**:4961–4968.

Sharon E, Kalma Y, Sharp A, Raveh-Sadka T, Levo M, Zeevi D, Keren L, Yakhini Z, Weinberger A, Segal E. 2012. Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters. *Nat Biotechnol* **30**:521–530. doi:10.1038/nbt.2205

Shin Y, Brangwynne CP. 2017. Liquid phase condensation in cell physiology and disease. *Science* **357**. doi:10.1126/science.aaf4382

Sigler PB. 1988. Transcriptional activation. Acid blobs and negative noodles. *Nature* **333**:210–212. doi:10.1038/333210a0

Skene PJ, Henikoff S. 2017. An efficient targeted nuclease strategy for high-resolution mapping of DNA binding sites. *Elife* **6**. doi:10.7554/eLife.21856

Staller MV, Holehouse AS, Swain-Lenz D, Das RK, Pappu RV, Cohen BA. 2018. A High-Throughput Mutational Scan of an Intrinsically Disordered Acidic Transcriptional Activation Domain. *Cell Syst* **6**:444–455.e6. doi:10.1016/j.cels.2018.01.015

Starita LM, Ahituv N, Dunham MJ, Kitzman JO, Roth FP, Seelig G, Shendure J, Fowler DM. 2017. Variant Interpretation: Functional Assays to the Rescue. *Am J Hum Genet* **101**:315–325. doi:10.1016/j.ajhg.2017.07.014

Stormo GD, Zuo Z, Chang YK. 2015. Spec-seq: determining protein–DNA-binding specificity by sequencing. *Brief Funct Genomics* **14**:30–38. doi:10.1093/bfgp/elu043

Tanaka M, Clouston WM, Herr W. 1994. The Oct-2 glutamine-rich and proline-rich activation domains can synergize with each other or duplicates of themselves to activate transcription. *Mol Cell Biol* **14**:6046–6055. doi:10.1128/mcb.14.9.6046

Tareen A, Kinney JB. 2020. Logomaker: beautiful sequence logos in Python. *Bioinformatics* **36**:2272–2274. doi:10.1093/bioinformatics/btz921

Teytelman L, Thurtle DM, Rine J, van Oudenaarden A. 2013. Highly expressed loci are vulnerable to misleading ChIP localization of multiple unrelated proteins. *Proc Natl Acad Sci U S A* **110**:18602–18607. doi:10.1073/pnas.1316064110

Tuttle LM, Pacheco D, Warfield L, Hahn S, Klevit RE. 2019. Mediator subunit Med15 dictates the conserved "fuzzy" binding mechanism of yeast transcription activators Gal4 and Gcn4. *Cold Spring Harbor Laboratory*. doi:10.1101/840348

Tycko J, DelRosso N, Hess GT, Aradhana, Banerjee A, Mukund A, Van MV, Ego BK, Yao D, Spees K, Suzuki P, Marinov GK, Kundaje A, Bassik MC, Bintu L. 2020. High-throughput discovery and characterization of human transcriptional effectors. doi:10.1101/2020.09.09.288324

Vierstra J, Stamatoyannopoulos JA. 2016. Genomic footprinting. *Nat Methods* **13**:213–221. doi:10.1038/nmeth.3768

Vitalis A, Pappu RV. 2009. ABSINTH: a new continuum solvation model for simulations of polypeptides in aqueous solutions. *J Comput Chem* **30**:673–699. doi:10.1002/jcc.21005

Warfield L, Tuttle LM, Pacheco D, Klevit RE, Hahn S. 2014. A sequence-specific transcription activator motif and powerful synthetic variants that bind Mediator using a fuzzy protein interface. *Proc Natl Acad Sci U S A* **111**:E3506–13. doi:10.1073/pnas.1412088111

Wojciak JM, Martinez-Yamout MA, Dyson HJ, Wright PE. 2009. Structural basis for recruitment of CBP/p300 coactivators by STAT1 and STAT2 transactivation domains. *EMBO J* **28**:948–958. doi:10.1038/emboj.2009.30

Yasinska IM, Sumbayev VV. 2003. S-nitrosation of Cys-800 of HIF-1α protein activates its interaction with p300 and stimulates its transcriptional activity. *FEBS Lett* **549**:105–109.

Materials and methods

Cell line construction

To engineer the K562 cell line we began with LP3 from (Maricque et al., 2018). First, we introduced a frameshift mutation to the GFP in the landing pad using Cas9 and a gRNA against GFP (AddGene 41819). Second, we integrated our reporter at the AAVS1 locus using Cas9, the SM58 SSBD2 T2 gRNA (from Shondra Miller, Washington University School of Medicine GEiC) and the pMVS184 reporter plasmid (Figure S18). Starting on Day 2, we selected with 1 ug/ml puromycin for three days. We tested candidate reporter clones with transfections of a synthetic TF (pMVS 223, Figure S19) carrying p53 AD1, choosing the clone with the largest dynamic range between baseline GFP and the brightest transfected cells.

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.

Rational Mutagenesis

The sequences of all 525 VP16, Hif1α, CITED2, Stat3 and p65 variants are listed in Supplemental Dataset 1. “Hand Designed” variants are shown in Figure S1. 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 Supplemental Dataset 2. The p53 mutagenesis also included a deep mutational scan, a double alanine scan and some sequences from orthologous TFs. Activity values of each dataset are normalized separately.

Plasmid Library Construction

The plasmid sequences for the GFP reporter (Figure S23) and synthetic TF (Figure S24) are in Supplemental Dataset 6. Both will be available from Addgene.

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. We added PCR primers at the start and end. 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 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. 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.

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 4M colonies after step one and 26.2M 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). 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 1 uM ß-estradiol.

On Day 14 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 activity below. The dynamic range of the measurement is determined by the fluorescence values of the dimmest and brightest bin.

Barcode amplicon sequencing libraries

Genomic DNA was collected using the Qiamp DNA Mini kit (Qiagen). We performed 8 PCRs 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. We did 25 cycles with NEB Q5 polymerase using CP36.P10 and LP_019 primers. We pooled the PCRs, 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. 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	ctcccgattcgcagcgcatc
LP_019	GCAGCGTATCCACATAGCGTAAAAG
EPCR_P1_short	AATGATACGGCGACCACCGAG
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, 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.

Data processing

We demultiplexed samples using a combination of Index1 reads and Read1 inline barcodes. Using perfect matches, we counted the abundance of each FREE barcode in each sample. 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 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. We converted these activities into Z scores and computed the mean and standard error mean (SEM). The Z score normalization is nearly equivalent to mean centering the replicates (Pearson R^2 =1). The Z score normalization was essential for combining the p53 data, because the fluorescence values differed between days. We used Z scores for both datasets for consistency. Pooling barcodes led to reproducible AD measurements (Figure S3B). Reproducibility improved with more barcodes (Figure S3C) and with more integrations (Figure S3D).

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 ittertools. Colors 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 WT, we used a two sided t test and 5% FDR correction. We plotted the regressions with ‘regplot’ command in seaborn, using the ‘robust’ option for the confidence interval. To estimate the resolution of the assay (Figure 1D and Figure S4), for all pairs of ADs, we compared mean activities with a t-test, and corrected for multiple hypotheses 5% and 1% FDR. For each quartile, we averaged the ten smallest significant differences and plotted this mean difference Figure 1D and S3.

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 4B, S16).

To summarize the effects of substitutions at each position (Figure 4, S16), 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.

To assess the conservation of each AD, we used the HMMER website (hmmer.org) Using the full protein sequences, we ran HMMER for 3 iterations, at which point, VP16 had converged with 47 sequences, CITED2 had converged with 446 sequences, and Hif1α had not yet converged with 126K sequences. We took a screenshot of the logo of the HMM model for the alignment.

We performed ANOVA on composition, regressing activity of all four replicates against all 20 amino acids and a batch term. We iteratively removed parameters that were not significant. Separately, we trained a model with all dipeptides derived from [D,E,W,F,Y,L] and identified significant dipeptides. We added the significant dipeptides to the composition model and iteratively removed parameters that were not significant.

All atom simulations

We ran all-atom, Monte Carlo simulations in the CAMPARI simulation engine (campari.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 Hif1ɑ, we simulated all hand designed variants and the WT sequence. All simulation data is available upon request.

For each variant, we ran ten simulations starting in a helix and ten starting in a random coil. 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 (ctraj.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.

Machine Learning

The machine learning analysis was carried out in python with the sklearn package. We started with all variants of VP16, Hif1ɑ and CITED2 and then excluded the shuffle variants. The High Activity set (N = 122) had variants with a mean Z score above 0.5. The Low Activity set (N = 136) had variants with a mean Z score below 0. 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).

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 4A, red pixels), fulfill 3 criteria:

(Charge < -9) AND (WFYL > 7) AND (((Charge+9)-(WFYL-10)) <= 0)

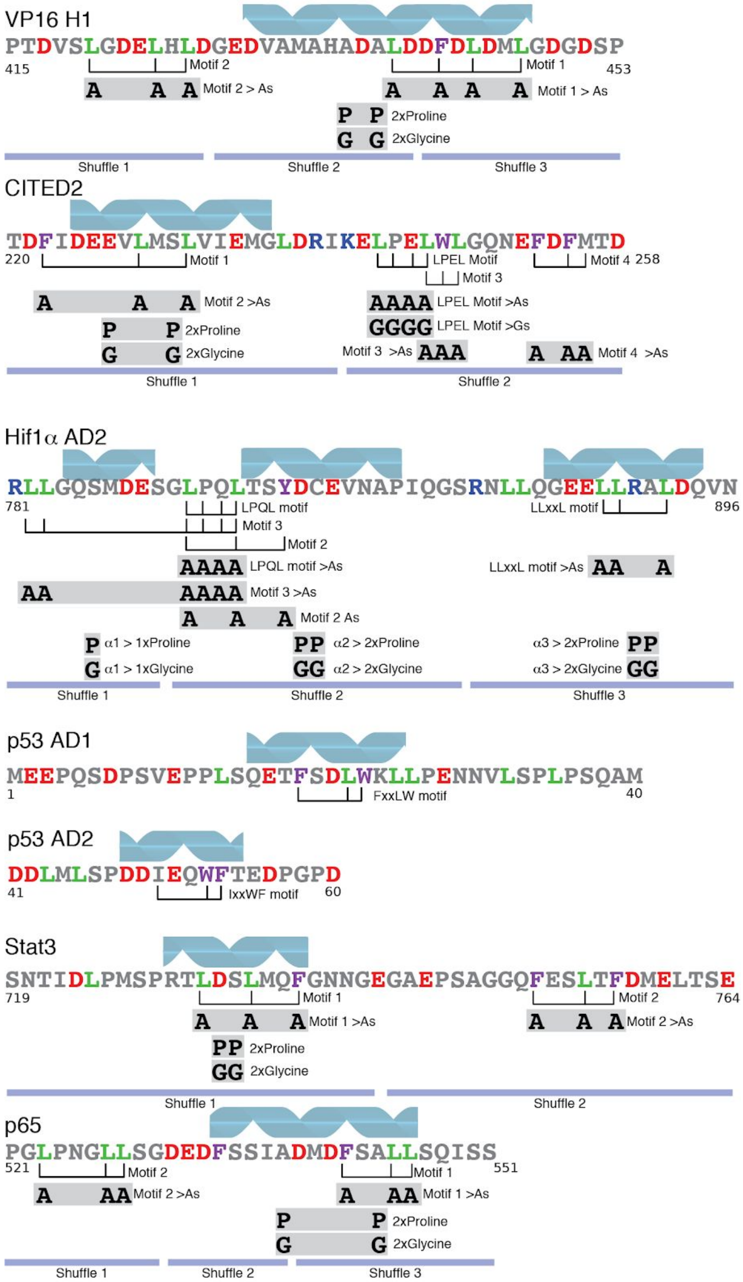
This algorithm identified 1139 tiles, 0.129% of the total. We aggregated overlapping tiles to predict 144 ADs on 136 TFs. To test these predictions, we used ADs annotated in Uniprot. We downloaded .gff files for the 1608 TFs from Uniprot. We used 4 regular expressions to search the “regions” column of 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. We also applied this predictor to our AD variants in Figure S11.

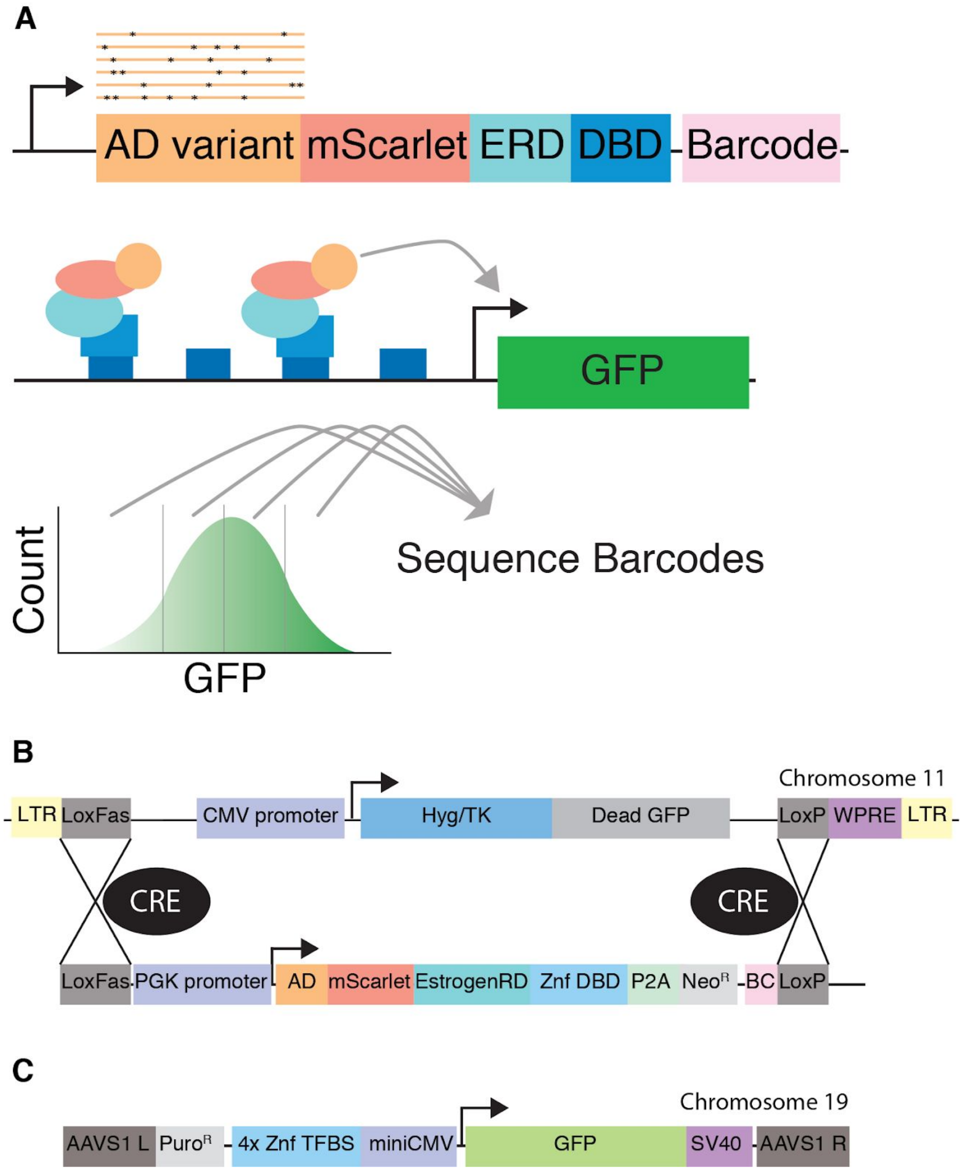
Supplemental Datasets

Dataset 1: All variants and activity measurements for the 5 AD library (VP16, CITED2, Hif1α, Stat3 and p65, N =525).
Dataset 2: All variants and activity measurements for the p53 ADs (N = 2991).
Dataset 3: Predicted acidic ADs on human TFs.
Dataset 4: AD sequences and DNA barcodes for the 5 AD library.
Dataset 5: AD sequences and DNA barcodes for the p53 ADs.
Dataset 6: Plasmid sequences for pMVS184 and pMVS223.

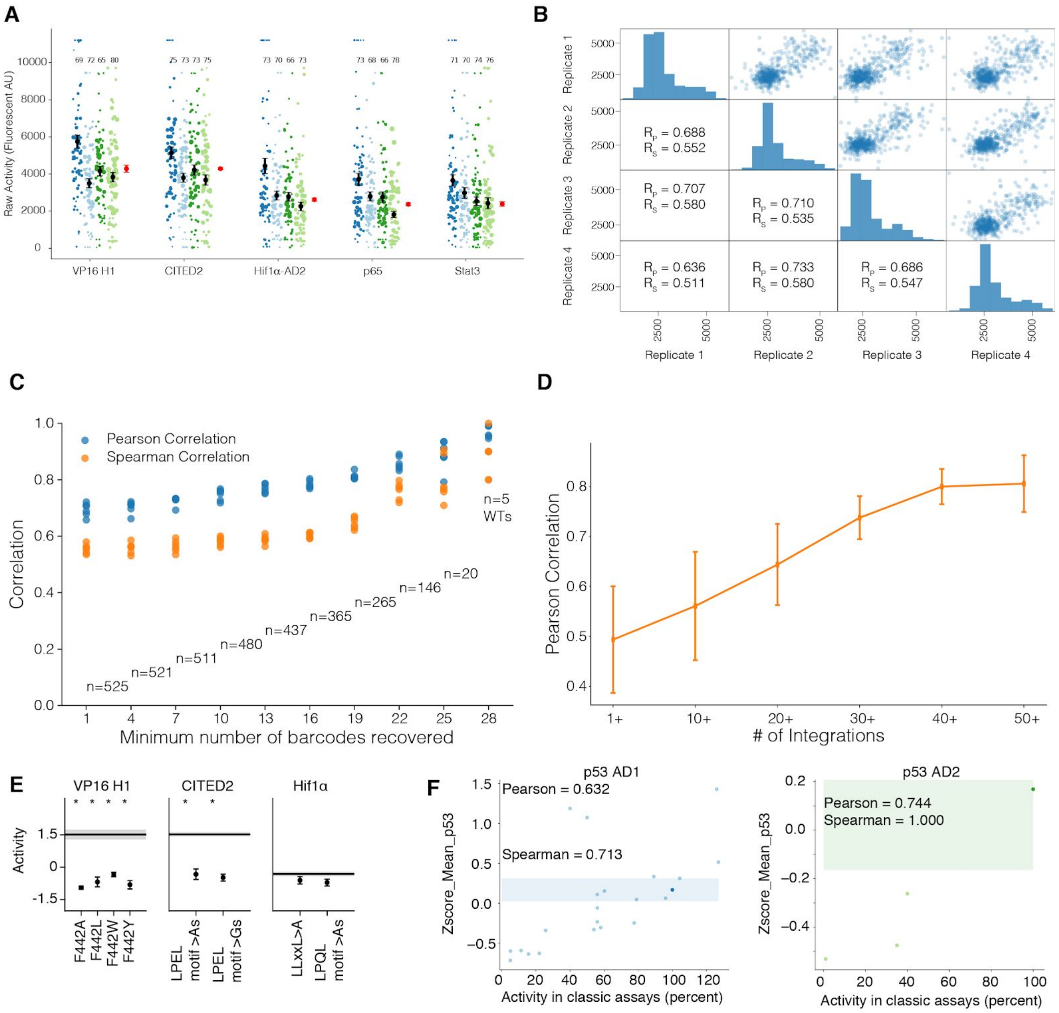
Supplemental Figures and Legends



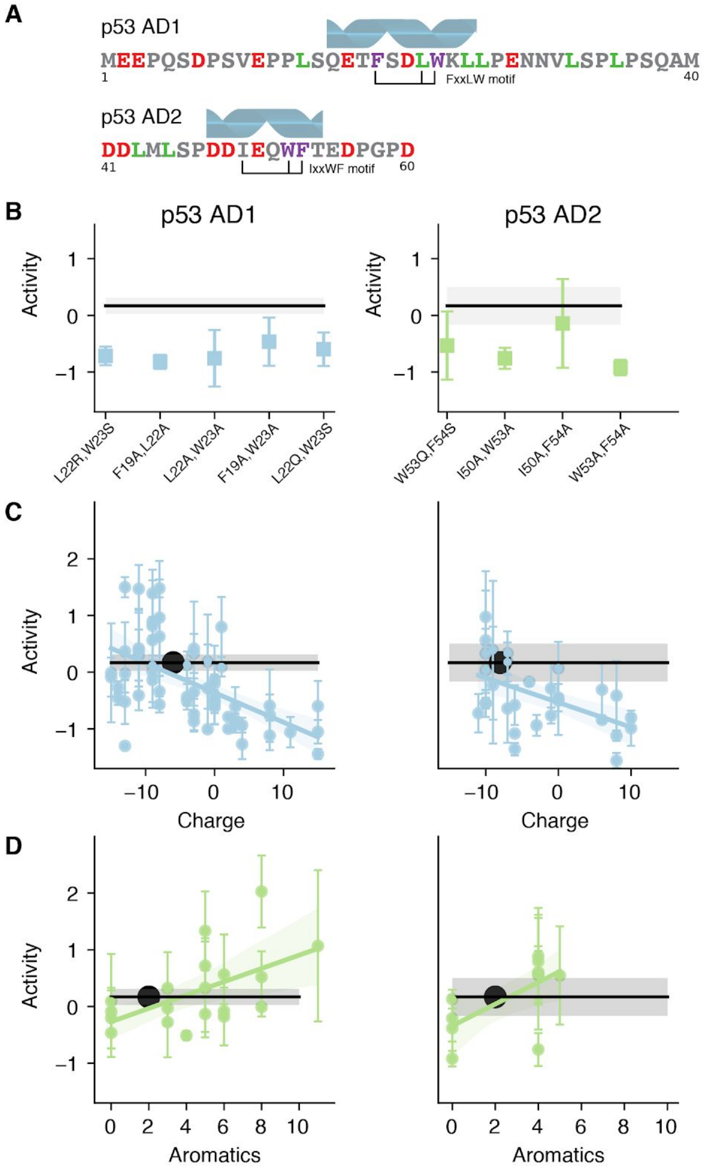
Supplemental Figure S1: Cartoon of hand designed variants in each AD. Known and predicted alpha helices are indicated as blue helices (Berlow et al., 2017; Jonker et al., 2005; Krois et al., 2016; Lecoq et al., 2017; Wojciak et al., 2009). Hydrophobic motifs are indicated with brackets. Published motifs are LPEL in CITED2, LPQL and LLxxL in Hif1ɑ, LxxFxL in VP16, FxxLW in p53 AD1 and IxxWF in p53 AD2. (Berlow et al., 2017; Raj and Attardi, 2017; Regier et al., 1993). Variants that remove motifs are indicated as gray boxes. Variants designed to break the helices include the 2xProline and 2xGlycine substitutions. The regions that were randomly shuffled in each AD are indicated. The shuffle variants disrupt helices, disrupt motif grammar and test if the arrangement of residues contributes to activity.



Supplemental Figure S2: We developed a high throughput method for assaying AD activity in human cell culture. A) We clone a library of designed AD variants into a synthetic TF that contains an mScarlet red fluorescent protein, an estrogen response domain (ERD), a synthetic Zinc Finger DNA binding domain (DBD) and a barcode sequence in the 3’ UTR. When we induce nuclear localization with ß-estradiol, the synthetic TF enters the nucleus and activates a GFP reporter. The reporter has 4 binding sites for the DBD upstream of a miniCMV promoter. We use ‘Sort-seq’ to quantify activity of each variant (Kinney et al., 2010). We used landing pad #3 from (Maricque et al., 2018). B) Detailed cartoons of the Landing Pad and synthetic TF. Hyg/TK, hygromycin resistance and thymidine kinase fusion gene. WPRE, Woodchuck Hepatitis Virus Posttranscriptional Regulatory Element. LTR, lentivirus long terminal repeats. C) Cartoon of the reporter gene integrated at the AAVS1 locus. See Figure S23-S24 for plasmid maps.

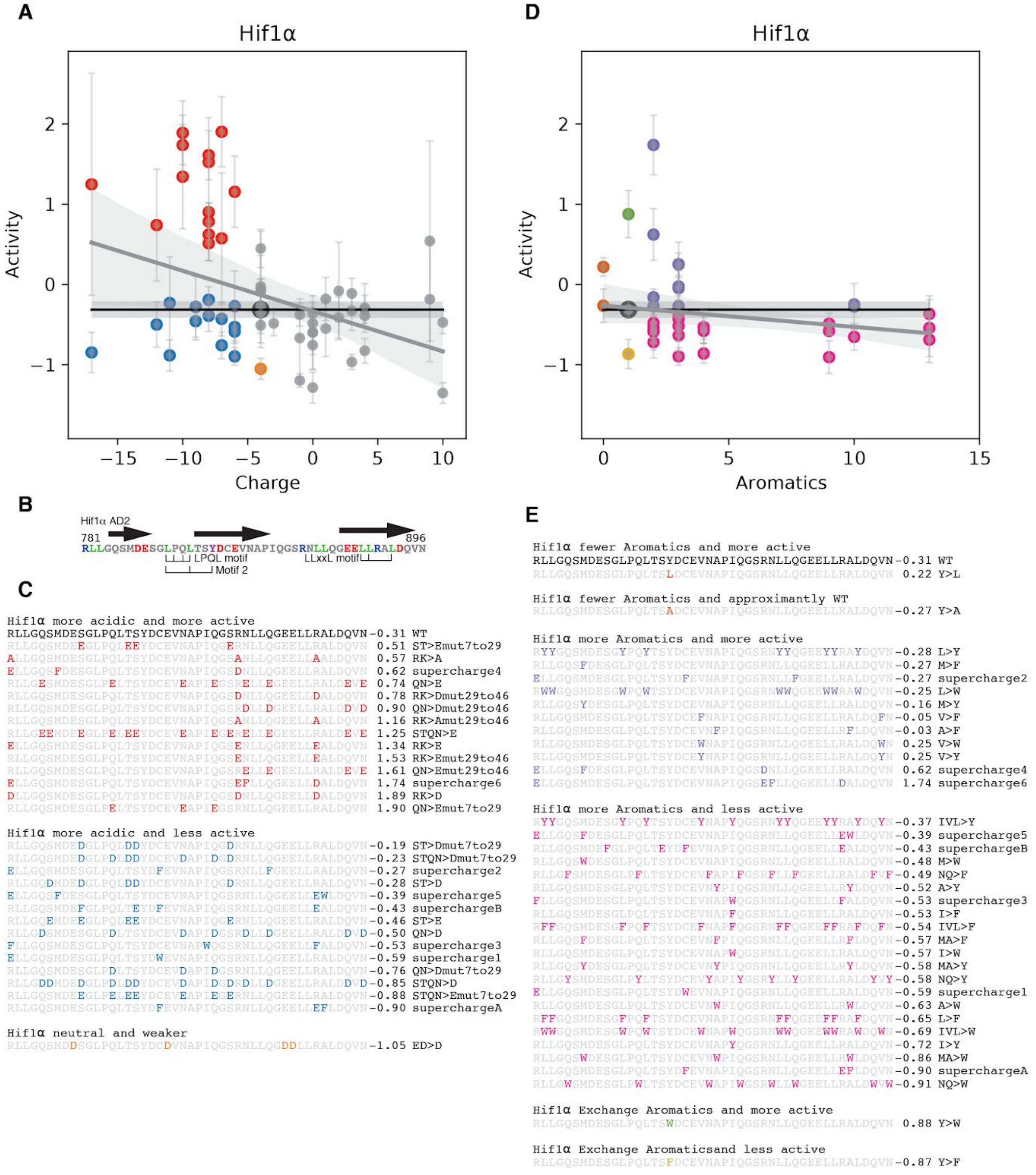


Supplemental Figure S3: Validating the high throughput AD assay.
A) Each WT AD was represented by 84 barcodes. The activity of each recovered barcode is plotted as a dot and the size of the dot is the square root of the number of reads for the barcode. For each AD, four biological replicates are shown and the number of barcodes recovered for each AD in each replicate is indicated. The mean and SEM calculated from combining data across all barcodes (red) agrees with the individual means (black). B) Reproducibility of AD measurements. Diagonal: histograms of AD activity measurements. Upper panels: reproducibility of ADs between biological replicate measurements. Lower panels: Pearson (R_P) and Spearman (R_S) correlation coefficients. C) The correlations between biological measurements improve as more barcodes are recovered. The 6 pairs of replicates are plotted. D) The correlations between biological replicates improve with the number of independent integrations. The mean and standard deviation of six pairs is shown. E) Variants of VP16, CITED2 and Hif1ɑ that have previously been shown to reduce activity reduce activity in our assay (Berlow et al., 2017; Cress and Triezenberg, 1991; Freedman et al., 2003; Regier et al., 1993). *, significant at 5% FDR. F) The activities of published variants of p53 are correlated with the activities we measure (Chang et al., 1995; Lin et al., 1994). WT p53 AD1, dark blue dot; SEM, light blue box. WT p53 AD2, dark green dot; SEM, green box (cutoff at the top).



Supplemental Figure S4: Both activation domains of p53 require acidic residues and hydrophobic motifs for function.

A) The motif and alpha helix locations in both p53 ADs. B) Both ADs of p53 contain aromatic and leucine rich motifs that are important for AD function. C) Removing negative charge and/or adding positive charge decreased the activity of both p53 ADs. In p53 AD1, adding acidity (more negative net charge) increases activity in about half of variants. Note that p53 AD2 is only 20 residues, so its net charge per residues is double that of AD1. WT is indicated with a large black dot. WT means and SEMs are shown in all plots with a black line and gray box. Activity is average Z score. A linear regression with a confidence interval summarizes the general dependence on net charge. D) For both ADs, removing aromatic residues generally decreases activity. Adding aromatic residues can increase activity.

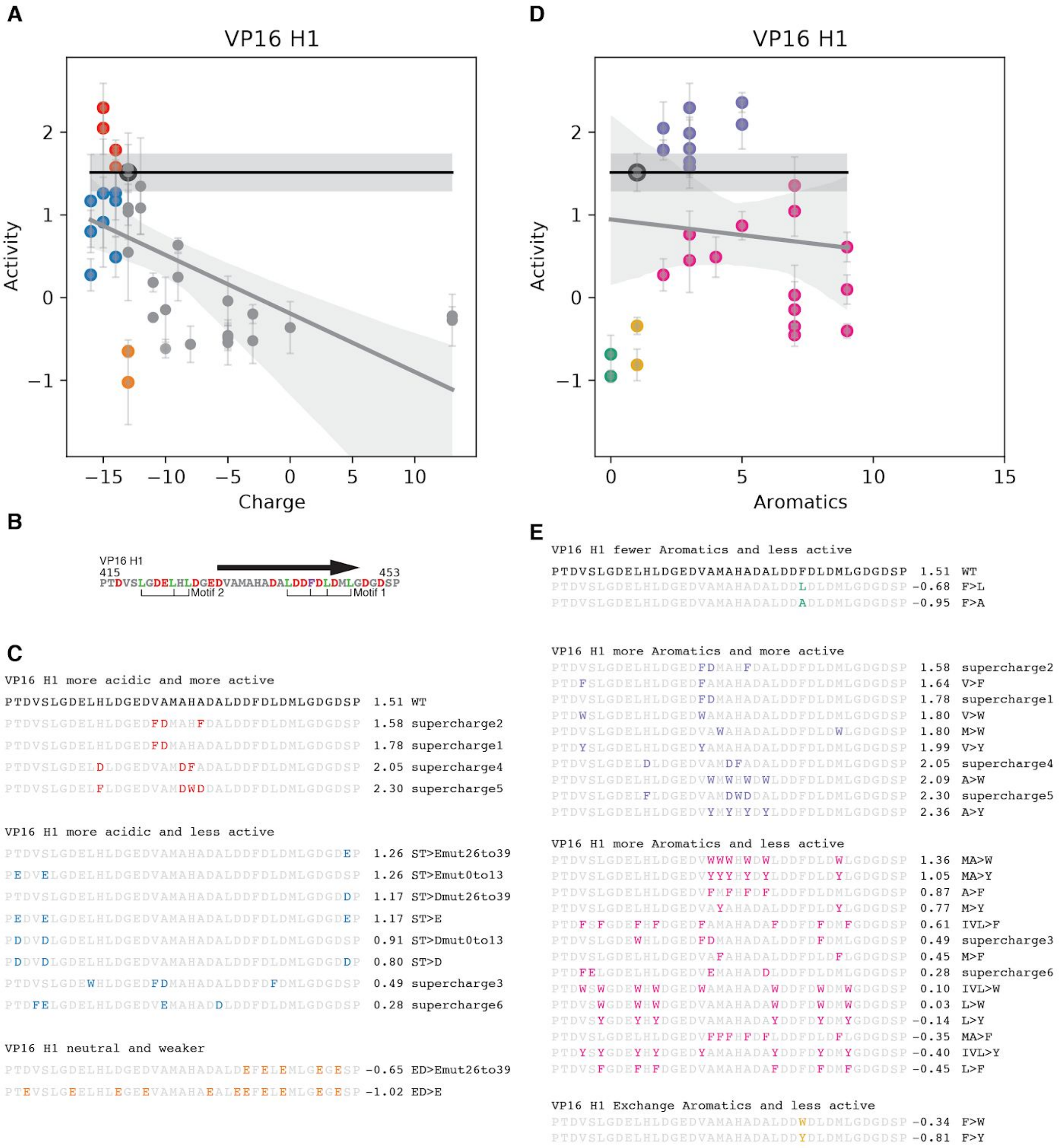


```
Hif1α more acidic and more active
RLLGQSMDESGLPQLTSYDCEVNAPIQGSRNLLQGEELLRALDQVN-0.31 WT
RLLGQSMDEEGLPQLEEYDCEVNAPIQGERNLLQGEELLRALDQVN 0.51 ST>Emut7to29
ALLGQSMDESGLPQLTSYDCEVNAPIQGSANLLQGEELLAALDQVN 0.57 RK>A
ELLGQSFDESGLPQLTSYDCEVNAPIQGSDNLLQGEELLRALDQVN 0.62 supercharge4
RLLGESMDESGLPELTSYDCEVEAPIEGSRELLEGEELLRALDEVE 0.74 QN>E
RLLGQSMDESGLPQLTSYDCEVNAPIQGSDNLLQGEELLDALDQVN 0.78 RK>Dmut29to46
RLLGQSMDESGLPQLTSYDCEVNAPIQGSRDLLDGEELLRALDDVD 0.90 QN>Dmut29to46
RLLGQSMDESGLPQLTSYDCEVNAPIQGSANLLQGEELLAALDQVN 1.16 RK>Amut29to46
RLLGEEMDEEGLPELEEYDCEVEAPIEGERELLEGEELLRALDEVE 1.25 STQN>E
ELLGQSMDESGLPQLTSYDCEVNAPIQGSENLLQGEELLEALDQVN 1.34 RK>E
RLLGQSMDESGLPQLTSYDCEVNAPIQGSENLLQGEELLEALDQVN 1.53 RK>Emut29to46
RLLGQSMDESGLPQLTSYDCEVNAPIQGSRELLEGEELLRALDEVE 1.61 QN>Emut29to46
ELLGQSMDESGLPQLTSYDCEVNAPIQGSEFLLQGEELLDALDQVN 1.74 supercharge6
DLLGQSMDESGLPQLTSYDCEVNAPIQGSDNLLQGEELLDALDQVN 1.89 RK>D
RLLGQSMDESGLPELTSYDCEVEAPIEGSRNLLQGEELLRALDQVN 1.90 QN>Emut7to29

Hif1α more acidic and less active
RLLGQSMDEDGLPQLDDYDCEVNAPIQGDRNLLQGEELLRALDQVN-0.19 ST>Dmut7to29
RLLGQSMDEDGLPDLDDYDCEVDAPIDGDRNLLQGEELLRALDQVN-0.23 STQN>Dmut7to29
ELLGQSMDESGLPQLTSYDFEVNAPIQGSRNLLFGEELLRALDQVN-0.27 supercharge2
RLLGQDMDEDGLPQLDDYDCEVNAPIQGDRNLLQGEELLRALDQVN-0.28 ST>D
ELLGQSFDESGLPQLTSYDCEVNAPIQGSRNLLQGEELLEWLDQVN-0.39 supercharge5
RLLGQSMDEFGLPQLTEYDFEVNAPIQGSRNLLQGEELLEALDQVN-0.43 superchargeB
RLLGQEMDEEGLPQLEEYDCEVNAPIQGERNLLQGEELLRALDQVN-0.46 ST>E
RLLGDSMDESGLPDLTSYDCEVDAPIDGSRDLLDGEELLRALDDVD-0.50 QN>D
FLLGQSMDESGLPQLTSYDCEVNAPWQGSRNLLQGEELLFALDQVN-0.53 supercharge3
ELLGQSMDESGLPQLTSYDWEVNAPIQGSRNLLQGEELLRALDQVN-0.59 supercharge1
RLLGQSMDESGLPDLTSYDCEVDAPIDGSRNLLQGEELLRALDQVN-0.76 QN>Dmut7to29
RLLGDDMDEDGLPDLDDYDCEVDAPIDGDRDLLDGEELLRALDDVD-0.85 STQN>D
RLLGQSMDEEGLPELEEYDCEVEAPIEGERNLLQGEELLRALDQVN-0.88 STQN>Emut7to29
RLLGQSMDESGLPQLTSYDFEVNAPIQGSRNLLQGEELLEFLDQVN-0.90 superchargeA

Hif1α neutral and weaker
RLLGQSMDDSGLPQLTSYDCDVNAPIQGSRNLLQGDDLLRALDQVN-1.05 ED>D
```

```
Hif1α fewer Aromatics and more active
RLLGQSMDESGLPQLTSYDCEVNAPIQGSRNLLQGEELLRALDQVN-0.31 WT
RLLGQSMDESGLPQLTSLDCEVNAPIQGSRNLLQGEELLRALDQVN 0.22 Y>L

Hif1α fewer Aromatics and approximantly WT
RLLGQSMDESGLPQLTSADCEVNAPIQGSRNLLQGEELLRALDQVN-0.27 Y>A

Hif1α more Aromatics and more active
RYYGQSMDESGYPQYTSYDCEVNAPIQGSRNYYQGEEYYRAYDQVN-0.28 L>Y
RLLGQSFDESGLPQLTSYDCEVNAPIQGSRNLLQGEELLRALDQVN-0.27 M>F
ELLGQSMDESGLPQLTSYDFEVNAPIQGSRNLLFGEELLRALDQVN-0.27 supercharge2
RWWGQSMDESGWPQWTSYDCEVNAPIQGSRNWWQGEEWWRAWDQVN-0.25 L>W
RLLGQSYDESGLPQLTSYDCEVNAPIQGSRNLLQGEELLRALDQVN-0.16 M>Y
RLLGQSMDESGLPQLTSYDCEFNAPIQGSRNLLQGEELLRALDQFN-0.05 V>F
RLLGQSMDESGLPQLTSYDCEVNFPIQGSRNLLQGEELLRFLDQVN-0.03 A>F
RLLGQSMDESGLPQLTSYDCEWNAPIQGSRNLLQGEELLRALDQWN 0.25 V>W
RLLGQSMDESGLPQLTSYDCEYNAPIQGSRNLLQGEELLRALDQYN 0.25 V>Y
ELLGQSFDESGLPQLTSYDCEVNAPIQGSDNLLQGEELLRALDQVN 0.62 supercharge4
ELLGQSMDESGLPQLTSYDCEVNAPIQGSEFLLQGEELLDALDQVN 1.74 supercharge6

Hif1α more Aromatics and less active
RYYGQSMDESGYPQYTSYDCEYNAPYQGSRNYYQGEEYYRAYDQYN-0.37 IVL>Y
ELLGQSFDESGLPQLTSYDCEVNAPIQGSRNLLQGEELLEWLDQVN-0.39 supercharge5
RLLGQSMDEFGLPQLTEYDFEVNAPIQGSRNLLQGEELLEALDQVN-0.43 superchargeB
RLLGQSWDESGLPQLTSYDCEVNAPIQGSRNLLQGEELLRALDQVN-0.48 M>W
RLLGFSMDESGLPFLTSYDCEVFAPIFGSRFLLFGEELLRALDFVF-0.49 NQ>F
RLLGQSMDESGLPQLTSYDCEVNYPIQGSRNLLQGEELLRYLDQVN-0.52 A>Y
FLLGQSMDESGLPQLTSYDCEVNAPWQGSRNLLQGEELLFALDQVN-0.53 supercharge3
RLLGQSMDESGLPQLTSYDCEVNAPFQGSRNLLQGEELLRALDQVN-0.53 I>F
RFFGQSMDESGFPQFTSYDCEFNAPFQGSRNFFQGEEFFRAFDQFN-0.54 IVL>F
RLLGQSFDESGLPQLTSYDCEVNFPIQGSRNLLQGEELLRFLDQVN-0.57 MA>F
RLLGQSMDESGLPQLTSYDCEVNAPWQGSRNLLQGEELLRALDQVN-0.57 I>W
RLLGQSYDESGLPQLTSYDCEVNYPIQGSRNLLQGEELLRYLDQVN-0.58 MA>Y
RLLGYSMDESGLPYLTSYDCEVYAPIYGSRYLLYGEELLRALDYVY-0.58 NQ>Y
ELLGQSMDESGLPQLTSYDWEVNAPIQGSRNLLQGEELLRALDQVN-0.59 supercharge1
RLLGQSMDESGLPQLTSYDCEVNWPIQGSRNLLQGEELLRWLDQVN-0.63 A>W
RFFGQSMDESGFPQFTSYDCEVNAPIQGSRNFFQGEEFFRAFDQVN-0.65 L>F
RWWGQSMDESGWPQWTSYDCEWNAPWQGSRNWWQGEEWWRAWDQWN-0.69 IVL>W
RLLGQSMDESGLPQLTSYDCEVNAPYQGSRNLLQGEELLRALDQVN-0.72 I>Y
RLLGQSWDESGLPQLTSYDCEVNWPIQGSRNLLQGEELLRWLDQVN-0.86 MA>W
RLLGQSMDESGLPQLTSYDFEVNAPIQGSRNLLQGEELLEFLDQVN-0.90 superchargeA
RLLGWSMDESGLPWLTSYDCEVWAPIWGSRWLLWGEELLRALDWVW-0.91 NQ>W

Hif1α Exchange Aromatics and more active
RLLGQSMDESGLPQLTSWDCEVNAPIQGSRNLLQGEELLRALDQVN 0.88 Y>W

Hif1α Exchange Aromaticsand less active
RLLGQSMDESGLPQLTSFDCEVNAPIQGSRNLLQGEELLRALDQVN-0.87 Y>F
```

Supplemental Figure S5: In Hif1α adding negative charges near the motifs is more likely to increase activity
A) Variants that added acidity and increased in activity are highlighted in red. The blue variants added acidity and had WT+2*SEM or lower activity. B) The locations of the motifs and alpha helices (arrows). C) The sequences of the red, blue and orange variants from A, in order of increasing activity. The red variants frequently remove the basic residue R820, or add acidic residues near to L795, L812, L813 or L819. The sequences are shown in order of increasing

activity. Adding E's increases activity more often than adding D's. Replacing E's with D's decreased activity (orange), the opposite of the trend observed for VP16 and CITED2. D-E) Adding aromatics to Hif1ɑ generally decreases activity. Replacing the Y with an L causes a small increase in activity. Adding aromatics can increase activity (purple) but more often decreases activity (pink). The two variants that add aromatics and have notably higher activity, supercharge4 and supercharge6, also added acidity. Supercharge6 increased activity in a statistically significant manner, but its similarity to RK>D (panel C) suggests the main effect comes from adding acidity.



```
VP16 H1 more acidic and more active
PTDVSLGDELHLDGEDVAMAHADALDDFDLDMLGDGDSP  1.51 WT
PTDVSLGDELHLDGEDFDMAHFDALDDFDLDMLGDGDSP  1.58 supercharge2
PTDVSLGDELHLDGEDFDMAHADALDDFDLDMLGDGDSP  1.78 supercharge1
PTDVSLGDELDLDGEDVAMDFADALDDFDLDMLGDGDSP  2.05 supercharge4
PTDVSLGDELFLDGEDVAMDWDDALDDFDLDMLGDGDSP  2.30 supercharge5

VP16 H1 more acidic and less active
PTDVSLGDELHLDGEDVAMAHADALDDFDLDMLGDGDEP  1.26 ST>Emut26to39
PEDVELGDELHLDGEDVAMAHADALDDFDLDMLGDGDSP  1.26 ST>Emut0to13
PTDVSLGDELHLDGEDVAMAHADALDDFDLDMLGDGDDP  1.17 ST>Dmut26to39
PEDVELGDELHLDGEDVAMAHADALDDFDLDMLGDGDEP  1.17 ST>E
PDDVDLGDELHLDGEDVAMAHADALDDFDLDMLGDGDSP  0.91 ST>Dmut0to13
PDDVDLGDELHLDGEDVAMAHADALDDFDLDMLGDGDDP  0.80 ST>D
PTDVSLGDEWHLDGEDFDMAHADALDDFDFDMLGDGDSP  0.49 supercharge3
PTDFELGDELHLDGEDVEMAHADDLDDFDLDMLGDGDSP  0.28 supercharge6

VP16 H1 neutral and weaker
PTDVSLGDELHLDGEDVAMAHADALDEFELEMLGEGESP -0.65 ED>Emut26to39
PTEVSLGEELHLEGEEVAMAHAEALEEFELEMLGEGESP -1.02 ED>E
```

```
VP16 H1 fewer Aromatics and less active
PTDVSLGDELHLDGEDVAMAHADALDDFDLDMLGDGDSP  1.51  WT
PTDVSLGDELHLDGEDVAMAHADALDDLDLDMLGDGDSP -0.68  F>L
PTDVSLGDELHLDGEDVAMAHADALDDADLDMLGDGDSP -0.95  F>A

VP16 H1 more Aromatics and more active
PTDVSLGDELHLDGEDFDMAHFDALDDFDLDMLGDGDSP  1.58  supercharge2
PTDFSLGDELHLDGEDFAMAHADALDDFDLDMLGDGDSP  1.64  V>F
PTDVSLGDELHLDGEDFDMAHADALDDFDLDMLGDGDSP  1.78  supercharge1
PTDWSLGDELHLDGEDWAMAHADALDDFDLDMLGDGDSP  1.80  V>W
PTDVSLGDELHLDGEDVAWAHADALDDFDLDWLGDGDSP  1.80  M>W
PTDYSLGDELHLDGEDYAMAHADALDDFDLDMLGDGDSP  1.99  V>Y
PTDVSLGDELDLDGEDVAMDFADALDDFDLDMLGDGDSP  2.05  supercharge4
PTDVSLGDELHLDGEDVWMWHWDWLDDFDLDMLGDGDSP  2.09  A>W
PTDVSLGDELFLDGEDVAMDWDDALDDFDLDMLGDGDSP  2.30  supercharge5
PTDVSLGDELHLDGEDVYMYHYDYLDDFDLDMLGDGDSP  2.36  A>Y

VP16 H1 more Aromatics and less active
PTDVSLGDELHLDGEDVWWWHWDWLDDFDLDWLGDGDSP  1.36  MA>W
PTDVSLGDELHLDGEDVYYYHYDYLDDFDLDYLGDGDSP  1.05  MA>Y
PTDVSLGDELHLDGEDVFMFHFDFLDDFDLDMLGDGDSP  0.87  A>F
PTDVSLGDELHLDGEDVAYAHADALDDFDLDYLGDGDSP  0.77  M>Y
PTDFSFGDEFHFDGEDFAMAHADAFDDFDFDMFGDGDSP  0.61  IVL>F
PTDVSLGDEWHLDGEDFDMAHADALDDFDFDMLGDGDSP  0.49  supercharge3
PTDVSLGDELHLDGEDVAFAHADALDDFDLDFLGDGDSP  0.45  M>F
PTDFELGDELHLDGEDVEMAHADDLDDFDLDMLGDGDSP  0.28  supercharge6
PTDWSWGDEWHWDGEDWAMAHADAWDDFDWDMWGDGDSP  0.10  IVL>W
PTDVSWGDEWHWDGEDVAMAHADAWDDFDWDMWGDGDSP  0.03  L>W
PTDVSYGDEYHYDGEDVAMAHADAYDDFDYDMYGDGDSP -0.14  L>Y
PTDVSLGDELHLDGEDVFFFHFDFLDDFDLDFLGDGDSP -0.35  MA>F
PTDYSYGDEYHYDGEDYAMAHADAYDDFDYDMYGDGDSP -0.40  IVL>Y
PTDVSFGDEFHFDGEDVAMAHADAFDDFDFDMFGDGDSP -0.45  L>F

VP16 H1 Exchange Aromatics and less active
PTDVSLGDELHLDGEDVAMAHADALDDWDLDMLGDGDSP -0.34  F>W
PTDVSLGDELHLDGEDVAMAHADALDDYDLDMLGDGDSP -0.81  F>Y
```

Supplemental Figure S6: In VP16 adding aromatic residues in the center is more likely to increase activity.

A) For VP16, no variants that added acidity increased activity. Some supercharge variants, that added both aromatic and acidic residues, had small increases in activity that were not statistically significant (red). Most variants that added acidity decreased activity (blue). VP16 is saturated for the effect of acidity on activity and this may be because nearly all aromatics and leucines are flanked by acidic residues. B) Motif locations and alpha helix (arrow). C) The locations of variants in A. D>E substitutions in the N terminal region decreased activity (orange). VP16 and CITED2 prefer D over E. D) The one aromatic residue, F442, is critical for activity because all substitutions caused lower activity. Adding 1-4 aromatics can increase activity (purple), but often does not (pink). E) The purple variants tended to add aromatics in the center of the AD, N-terminal to F442. Adding more aromatic residues or adding them C-terminal to F442 tended to decrease activity. The only variant that increased activity in a statistically significant manner was A>Y.

```
              vp16N_C_Sig_Up

PTDVSLGDELHLDGEDVAMAHADALDDFDLDMLGDGDSP  WT
PTDVSLGDELHLDGEDVYMYHYDYLDDFDLDMLGDGDSP   2.4   A>Y
```

```
              vp16N_C_Sig_Down

PTDVSLGDELHLDGEDVAMAHADALDDFDLDMLGDGDSP         WT
PTEVSLGEELHLEGEEVAMAHAEALEEFELEMLGEGESP  -1.0   ED>E
PTDVSLGDELHLDGEDVAMAHADALDDADLDMLGDGDSP  -1.0   F>A
PTDVSLGDELHLDGEDVAMAHADALDDYDLDMLGDGDSP  -0.8   F>Y
PTDVSLGDELHLDGEDVAMAHADALDDLDLDMLGDGDSP  -0.7   F>L
PTDVSLGDELHLDGEDVAMAHADALDEFELEMLGEGESP  -0.6   ED>Emut26to39
PKDVKLGDELHLDGEDVAMAHADALDDFDLDMLGDGDKP  -0.6   ST>K
PTDVSLGDELHLDGEDVAMAHADAADDADADMAGDGDSP  -0.6   killmotif1
PTDVSLGDELHLDGEDVAMAHADALDAFALAMLGAGASP  -0.6   ED>Amut26to39
PTRVSLGRRLHLRGEDVAMAHADALDDFDLDMLGDGDSP  -0.5   ED>Rmut0to13
PTDVSLGDELHLDGEDVAMAHADALDRFRLRMLGRGRSP  -0.5   ED>Rmut26to39
PTKVSLGKKLHLKGEDVAMAHADALDDFDLDMLGDGDSP  -0.5   ED>Kmut0to13
PTDVSLGDELHLDGKKVAMAHAKALKDFDLDMLGDGDSP  -0.5   ED>Kmut13to26
PTDVSFGDEFHFDGEDVAMAHADAFDDFDFDMFGDGDSP  -0.5   L>F
PTDVSLGDELHLDGEDVAMAHGDGLDDFDLDMLGDGDSP  -0.5   breakhelixG
PTDYSYGDEYHYDGEDYAMAHADAYDDFDYDMYGDGDSP  -0.4   IVL>Y
PTAVSLGAALHLAGAAVAMAHAAALAAFALAMLGAGASP  -0.4   ED>A
PTDVSLGDELHLDGEDVFFFHFDFLDDFDLDFLGDGDSP  -0.3   MA>F
PTDVSLGDELHLDGEDVAMAHADALDDWDLDMLGDGDSP  -0.3   F>W
PTKVSLGKKLHLKGKKVAMAHAKALKKFKLKMLGKGKSP  -0.3   ED>K
PKDVKLGDELHLDGEDVAMAHADALDDFDLDMLGDGDSP  -0.2   ST>Kmut0to13
PTDVSLGDELHLDGEDVAMAHPDPLDDFDLDMLGDGDSP  -0.2   breakhelixP
PTRVSLGRRLHLRGRRVAMAHARALRRFRLRMLGRGRSP  -0.2   ED>R
PTDVSLGDELHLDGEDVAMAHADALDKFKLKMLGKGKSP  -0.2   ED>Kmut26to39
PTDVSAGDEAHADGEDVAMAHADALDDFDLDMLGDGDSP  -0.2   killmotif2
PRDVRLGDELHLDGEDVAMAHADALDDFDLDMLGDGDRP  -0.1   ST>R
PTDVSYGDEYHYDGEDVAMAHADAYDDFDYDMYGDGDSP  -0.1   L>Y
PTDVSLGDELHLDGRRVAMAHARALRDFDLDMLGDGDSP  -0.0   ED>Rmut13to26
PTDVSWGDEWHWDGEDVAMAHADAWDDFDWDMWGDGDSP   0.0   L>W
PTDWSWGDEWHWDGEDWAMAHADAWDDFDWDMWGDGDSP   0.1   IVL>W
PRDVRLGDELHLDGEDVAMAHADALDDFDLDMLGDGDSP   0.2   ST>Rmut0to13
PTAVSLGAALHLAGEDVAMAHADALDDFDLDMLGDGDSP   0.2   ED>Amut0to13
PTDFELGDELHLDGEDVEMAHADDLDDFDLDMLGDGDSP   0.3   supercharge6
PTDVSLGDEWHLDGEDFDMAHADALDDFDFDMLGDGDSP   0.5   supercharge3
PTDFSFGDEFHFDGEDFAMAHADAFDDFDFDMFGDGDSP   0.6   IVL>F
PTDVSLGDELHLDGAAVAMAHAAALADFDLDMLGDGDSP   0.6   ED>Amut13to26
```

```
              Hif1-AD2_46_Sig_Up

RLLGQSMDESGLPQLTSYDCEVNAPIQGSRNLLQGEELLRALDQVN        WT
RLLGQSMDESGLPQLTSYDCEVNAPIQGSDNLLQGEELLDALDQVN  0.8   RK>Dmut29to46
RLLGQSMDESGLPQLTSYDCEVNAPIQGSRDLLDGEELLRALDDVD  0.9   QN>Dmut29to46
RLLGQSMDESGLPQLTSYDCEVNAPIQGSRELLEGEELLRALDEVE  1.6   QN>Emut29to46
ELLGQSMDESGLPQLTSYDCEVNAPIQGSEFLLQGEELLDALDQVN  1.7   supercharge6
DLLGQSMDESGLPQLTSYDCEVNAPIQGSDNLLQGEELLDALDQVN  1.9   RK>D
RLLGQSMDESGLPELTSYDCEVEAPIEGSRNLLQGEELLRALDQVN  1.9   QN>Emut7to29
```

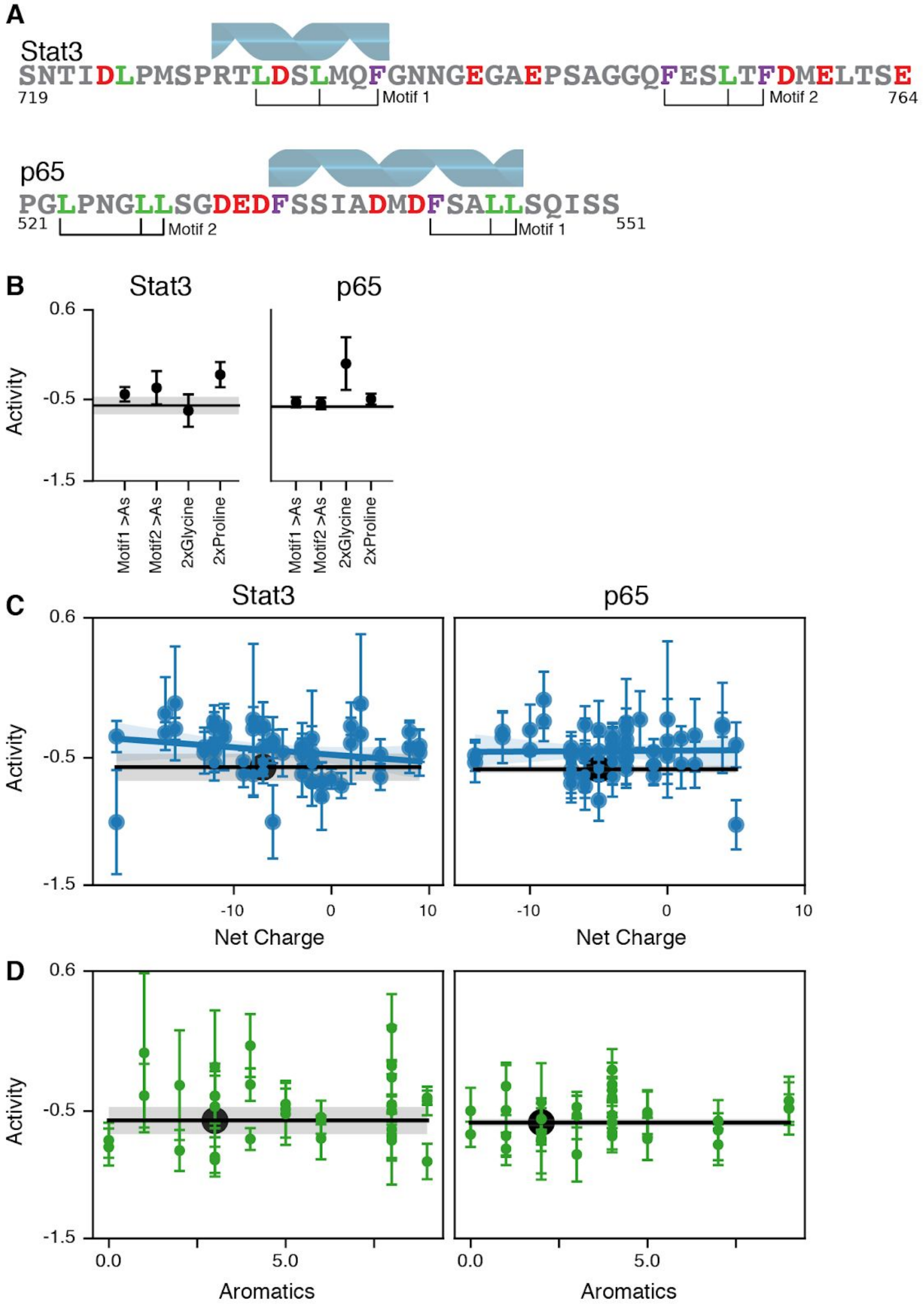
```
              CITED2_N_Sig_Up

TDFIDEEVLMSLVIEMGLDRIKELPELWLGQNEFDFMTD         WT
TDAIDEEVLMSLVIEMGLDRIKELPELWLGQNEFDFMTD   2.0   F>Amut0to20
TDFIDEEVLMSLVIEMGLDEIEELPELWLGQNEFDFMTD   2.0   RK>E
TDFIDEEYLMSLYIEMGLDRIKELPELWLGQNEFDFMTD   2.1   V>Y
TDFIDEEVLMSLVIEMGLDRIKELPELWLGQNEWDWMTD   2.3   F>Wmut20to3
TDWIDEEVLMSLVIEMGLDRIKELPELWLGQNEFDFMTD   2.3   F>Wmut0to20
EDFIDEEVLMELVIEMGLDRIKELPELWLGEEEFDFMED   2.3   STQN>E
```

```
              CITED2_N_Sig_Down

TDFIDEEVLMSLVIEMGLDRIKELPELWLGQNEFDFMTD         WT
TAFIAAAVLMSLVIAMGLARIKELPELWLGQNEFDFMTD  -1.6   ED>Amut0to20
TDFIDEEVLMSLVIEMGLDRIKKLPKLWLGQNKFKFMTK  -0.9   ED>Kmut20to39
TDFIDEEVLMSLVIEMGLDRIAALPALWLGQNAFAFMTA  -0.8   EDKR>Amut20to39
TDAIDEEVAMSAVIEMGLDRIKELPELWLGQNEFDFMTD  -0.8   killmotif1
TDFIDEEVFMSFVIEMGFDRIKEFPEFWFGQNEFDFMTD  -0.8   L>F
TAFIAAAVLMSLVIAMGLAAIKELPELWLGQNEFDFMTD  -0.8   EDKR>Amut0to20
TDFIDEEVYMSYVIEMGYDRIKEYPEYWYGQNEFDFMTD  -0.7   L>Y
TRFIRRRVLMSLVIRMGLRRIKELPELWLGQNEFDFMTD  -0.6   ED>Rmut0to20
TDFFDEEFFMSFFFEMGFDRFKEFPEFWFGQNEFDFMTD  -0.6   IVL>F
TAFIAAAVLMSLVIAMGLAAIAALPALWLGQNAFAFMTA  -0.6   EDKR>A
TDFIDEEVWMSWVIEMGWDRIKEWPEWWWGQNEFDFMTD  -0.6   L>W
TDFIDEEVLMSLVIEMGLDRIKALPALWLGQNAFAFMTA  -0.6   ED>Amut20to39
TDFYDEEYYMSYYYEMGYDRYKEYPEYWYGQNEFDFMTD  -0.5   IVL>Y
TDFIDEEVLMSLVIEMGLDRIKEGGGGWLGQNEFDFMTD  -0.5   kilmotif2_LPEL_G
TRFIRRRVLMSLVIRMGLRRIKRLPRLWLGQNRFRFMTR  -0.5   ED>R
TKFIKKKVLMSLVIKMGLKRIKKLPKLWLGQNKFKFMTK  -0.5   ED>K
RDFIDEEVLMRLVIEMGLDRIKELPELWLGRREFDFMRD  -0.5   STQN>R
TDFIDEEVLMSLVIEMGLDRIKRLPRLWLGQNRFRFMTR  -0.4   ED>Rmut20to39
TKFIKKKVLMSLVIKMGLKRIKELPELWLGQNEFDFMTD  -0.4   ED>Kmut0to20
TDFIDEEVLMSLVIEMGLDRIKEAAAAWLGQNEFDFMTD  -0.3   kilmotif2_LPEL_A
TDFIDEEVLMSLVIEMGLDRIKELPELWLGKKEFDFMKD  -0.3   STQN>Kmut20to39
TDFIDEEVLMSLVIEMGLDRIKELPELWLGFFEFDFMTD  -0.3   NQ>F
TDFWDEEWWMSWWWEMGWDRWKEWPEWWWGQNEFDFMTD  -0.3   IVL>W
TDFIDEEVLMSLVIEMGLDRIKELPELWLGRREFDFMRD  -0.2   STQN>Rmut20to39
TDFIDEEVLMSLVIEMGLDRIKEAPEAAAGQNEFDFMTD  -0.1   killmotif3
KDFIDEEVLMKLVIEMGLDRIKELPELWLGKKEFDFMKD  -0.1   STQN>K
TDFWDEEVLMSLVWEMGLDRWKELPELWLGQNEFDFMTD  -0.0   I>W
TDFIDEPVLMPLVIEMGLDRIKELPELWLGQNEFDFMTD  -0.0   breakhelixP
TDFIDEEVLMSLVIEMGLDRIKELPELWLGRREFDFMTD   0.2   QN>R
TEFIEEEVLMSLVIEMGLERIKELPELWLGQNEFDFMTD   0.2   ED>Emut0to20
TEFIEEEVLMSLVIEMGLERIKELPELWLGQNEFEFMTE   0.3   ED>E
TDFIDEEVLMSLVIEMGLDRIKELPELALGQNEADAMTD   0.3   WF>Amut20to39
TDFIDEEVLMSLVIEMGLDRIKELPELWLGWWEFDFMTD   0.4   NQ>W
TDFIDEEVLFSLVIEFGLDRIKELPELWLGQNEFDFMTD   0.7   aro3
TDYIDEEVLMSLVIEMGLDRIKELPELYLGQNEYDYMTD   0.7   WF>Y
TDFIDEEVLFSLVIEFGLDRIKELPELWLGQNEFDFFTD   0.7   M>F
```

```
              Hif1-AD2_46_Sig_Down

RLLGQSMDESGLPQLTSYDCEVNAPIQGSRNLLQGEELLRALDQVN        WT
RLLGQSMKKSGLPQLTSYKCKVNAPIQGSRNLLQGKKLLRALKQVN  -1.4  ED>K
RLLGQSMDEKGLPQLKKYDCEVNAPIQGKRNLLQGEELLRALDQVN  -1.3  ST>Kmut7to29
RLLGQSMDESGLPKLTSYDCEVKAPIKGSRNLLQGEELLRALDQVN  -1.2  QN>Kmut7to29
RLLGQSMDDSGLPQLTSYDCDVNAPIQGSRNLLQGDDLLRALDQVN  -1.1  ED>D
RLLGQSMDERGLPRLRRYDCEVRAPIRGRRNLLQGEELLRALDQVN  -1.0  STQN>Rmut7to29
```

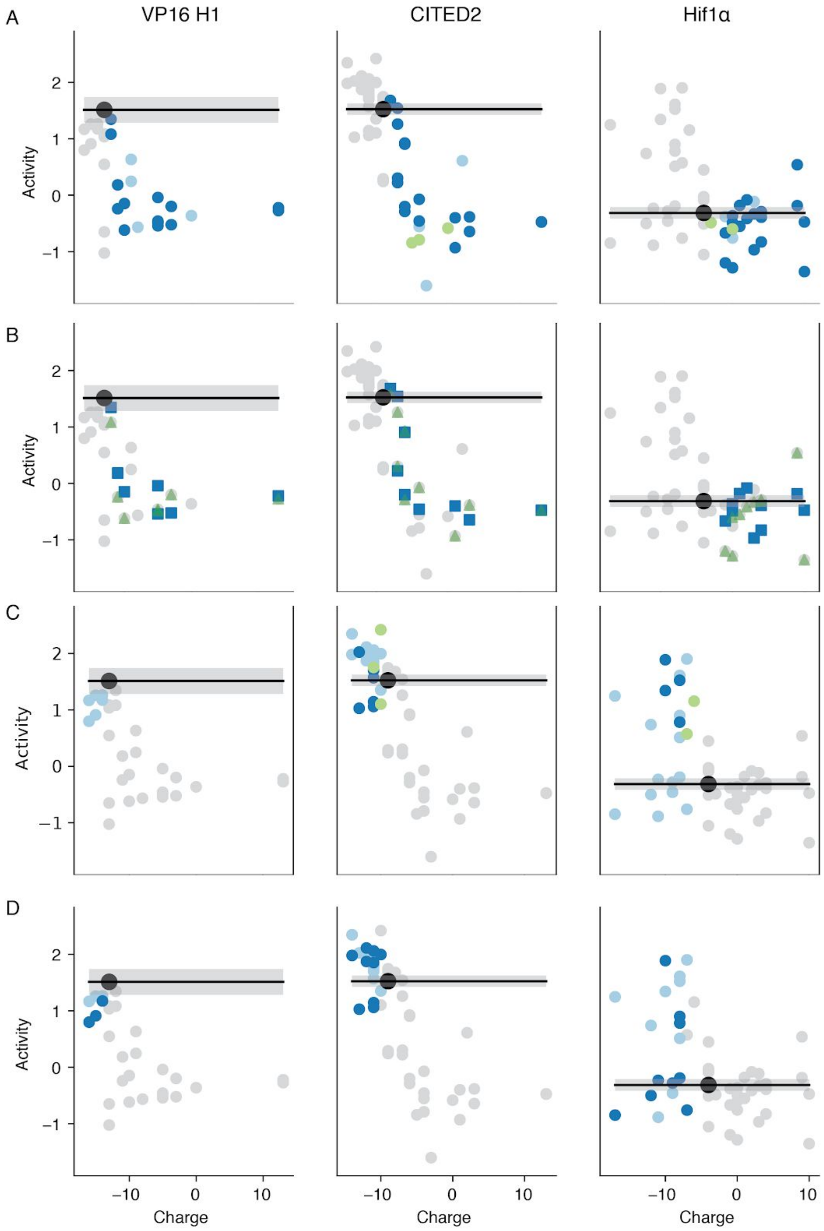
Supplemental Figure S7: Variants that had a statistically significant increase or decrease activity after correcting for multiple hypothesis testing with 5% FDR.

For each AD, variants that are stronger (cyan) or weaker (red) than WT. It is easier to make strong ADs weaker than to make them stronger.



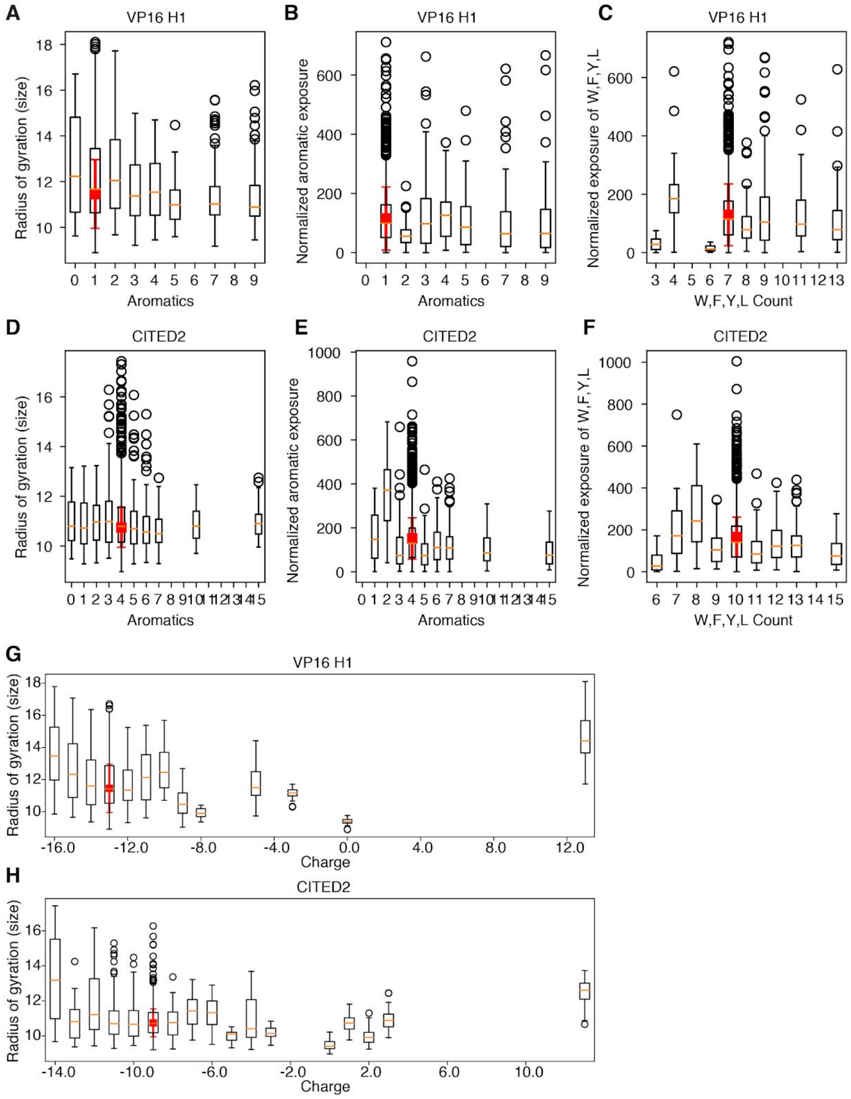
Supplemental Figure S8. Variants of Stat3 and p65 had very small changes in activity despite large sequence perturbations.

A) The motifs and alpha helices of Stat3 and p65. B) The motif mutants and the helix breaking mutants do not cause statistically significant changes in activity. C) The regression suggests that Stat3 activity has a very mild negative correlation with net charge. D) Adding or removing aromatic residues cause small changes in activity. None of the Stat3 or p65 variants had statistically significant changes in activity. They may belong to a different class of ADs.



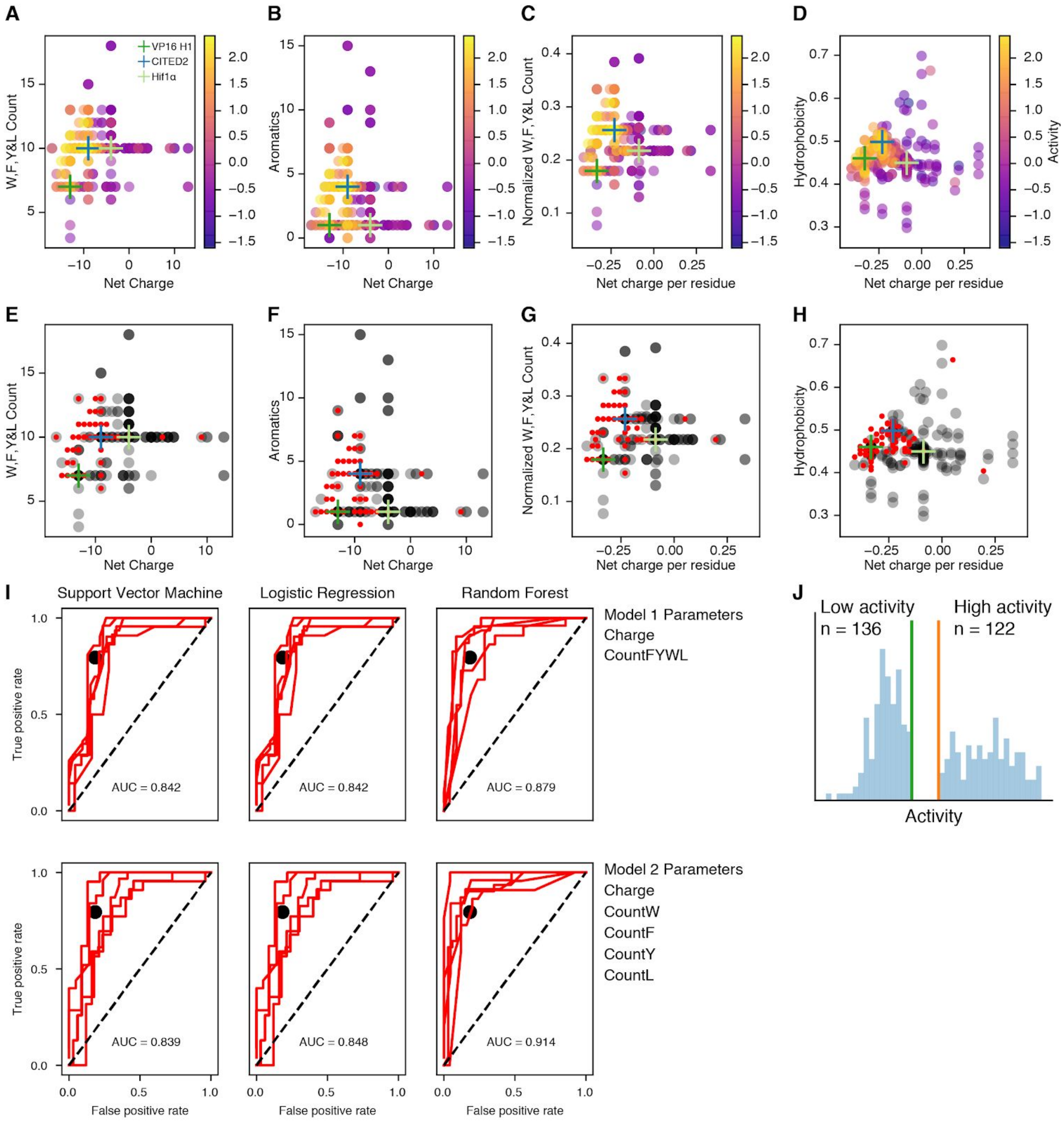
Supplemental Figure S9: Net charge more efficiently describes the effects of substitutions than amino acid identity.

A) Removing acidic residues (negative charges, light blue) has a similar effect as adding basic residues (positive charges, dark blue). B) The identity of added basic residues does not matter. Adding K's (green) has a similar effect on activity as adding R's (blue). C) Adding acidic residues (light blue) has similar effects to replacing basic residues with alanine (green) or replacing basic residues with acidic residues (dark blue). In Hif1α, removing positives always increases activity. D) The identities of added acidic residues frequently does not matter. Adding D's (dark blue) has a similar effect on activity as adding E's (light blue). In Hif1α, adding E's is more likely to increase activity than adding D's.



Supplemental Figure S10: Summary of all atom Monte Carlo simulations of VP16 and CITED2 variants.

These simulations are well-poised to capture the conformational ensemble of and residual structure in disordered proteins. Simulations of the Yeast AD, Gcn4, helped us develop the Acidic Exposure Model (Staller et al., 2018). For each variant, we ran 10 simulations starting in a helix and 10 simulations starting in a random coil. Here, each of the 20 simulations for each variant is included separately. For VP16 (A) and CITED2 (B) variants with more aromatic residues tended to have a smaller radius of gyration, a parameter that captures the diameter of the conformational ensemble. In both ADs, adding aromatic residues leads to smaller radius of gyration, consistent with chain collapse. (C-F) For each simulation, we computed the exposure of all aromatic residues (C-D) or W,F,Y&L residues (E-F) (See Methods) and normalized this total exposure by the number of aromatics or W,F,Y&L residues in each sequence. For VP16, adding 2-3 aromatics leads to a slight increased normalized exposure. For CITED2, adding aromatics generally does not increase normalized exposure. (G-H) Adding acidic residues frequently increased the radius of gyration, indicating a more expanded ensemble of conformations.

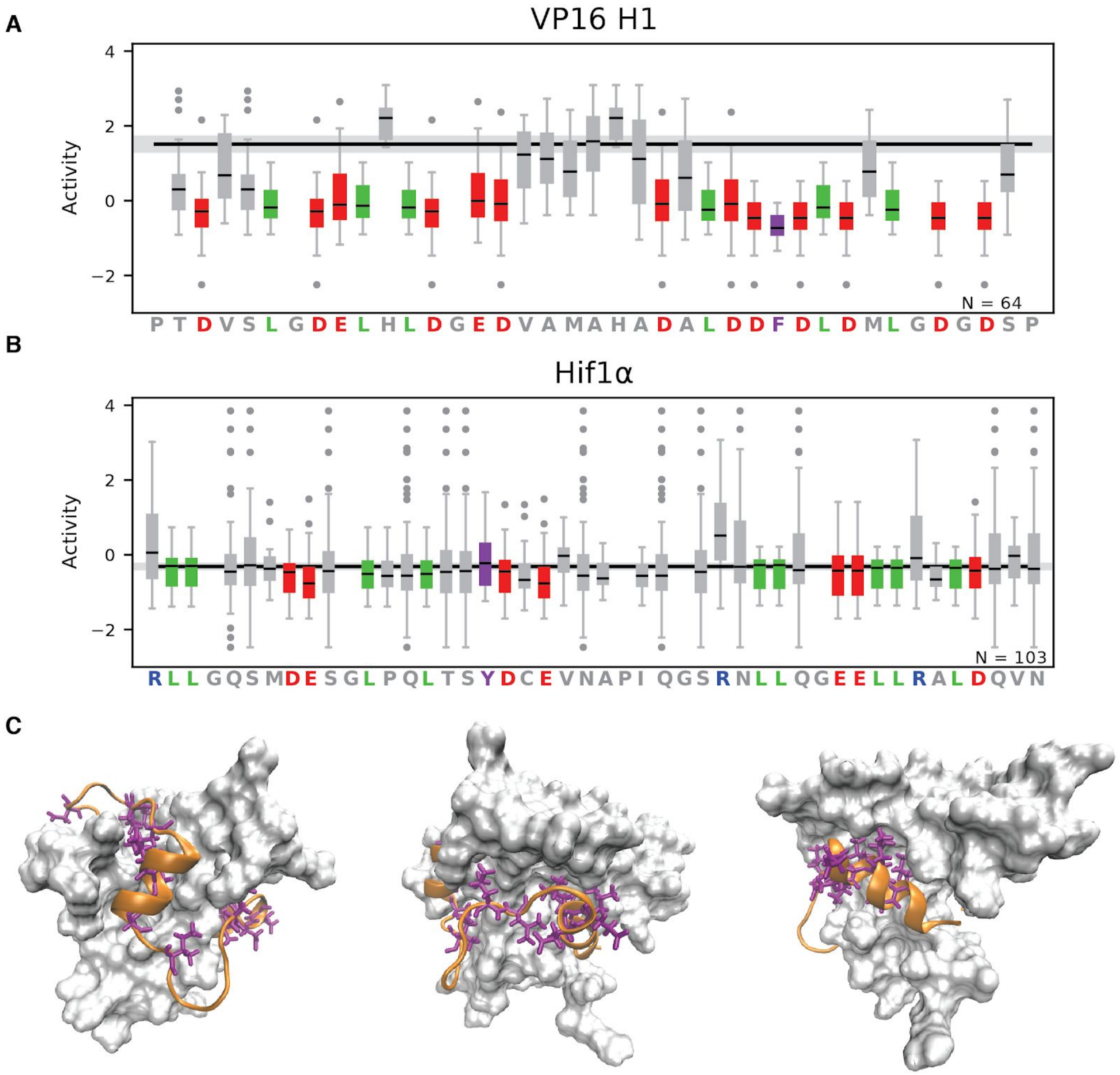


Supplemental Figure S11: Net charge and W,F,Y&L count can separate high and low activity variants. A) Repeat of Figure. 5A. The location of each point encodes its properties. Color indicates activity. B) Counting only aromatics does not separate high and low activity variants as well as counting W,F,Y&L residues. C) Normalizing the counts and charge by AD length preserves the separation between high and low activity variants. D) Using Kyte-Doolittle hydrophobicity(Kyte and Doolittle, 1982) instead of counting W,F,Y&L also separates high and low activity variants. WT ADs are indicated with crosses. E-H) Replotting A-D with binarized data. Many points on this grid are occupied by both high (red) and low (gray) activity variants. We split the data into high (Activity >0.5, N=125, red) and low (Activity <=0.5, N = 177, gray) activity variants and replotted them. When multiple low points overlap they appear black. Red points are on top of gary points to emphasize the overlap. I) We deployed Support Vector

Machines, Logistic Regression and Random Forest classifiers. We included all VP16, Hif1α and CITED2 variants except for the shuffle variants, because these composition-based classification models cannot distinguish the shuffle variants from WT. Receiver operator characteristic curves for 5-fold cross validation. The average area under the curve (AUC) of the 5 test sets is shown. For comparison, the performance of the proteome-wide AD predictor described in Fig. 3c is plotted as a black dot. J) There are 122 data points in the High Activity Set (Activity >0.5) and 136 in the Low Activity Set (Activity <0, see Methods).

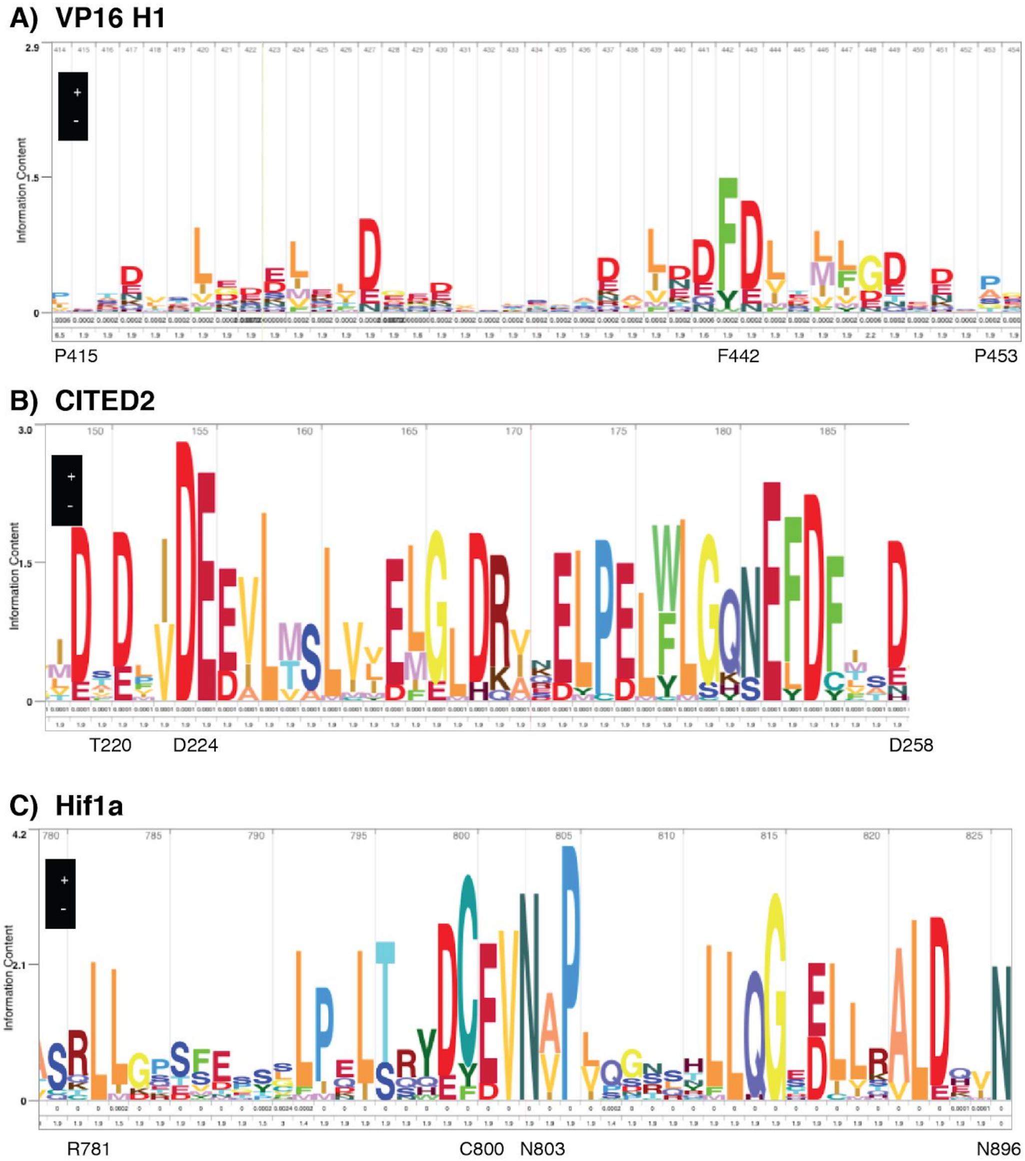
Model Parameters	Support Vector Machine Avg. AUC	Logistic Regression Avg. AUC	Random Forest Avg. AUC
Charge, WFYL	0.8242	0.8239	0.8715
Charge, Aromatics	0.8172	0.8238	0.8866
Charge, W, F, Y, L	**0.8351**	**0.8425**	**0.9088**
W, F, Y, L	0.6778	0.6710	0.7968
Charge, F, Y, L	0.8253	0.8416	0.9207
Charge, W, Y, L	0.8358	0.8383	0.9150
Charge, W, F, L	0.8301	0.8421	0.9176
Charge, W, F, Y	0.8222	0.8213	0.8784
W, F, Y, L, E, D, K, R	**0.8415**	**0.8651**	**0.9177**
F, Y, L, E, D, K, R	0.8388	0.8651	0.9076
W, Y, L, E, D, K, R	0.8372	0.8651	0.8883
W, F, L, E, D, K, R	0.8404	0.8725	0.8993
W, F, Y, E, D, K, R	0.8234	0.8549	0.8550

Supplemental Table S1: Comparison of machine learning models with different parameter sets. For each parameter set, the average AUCs from 5 fold cross validation are shown. Net Charge is the parameter that, when removed, caused the largest drop in model performance. Removing leucine residues caused a larger drop in model performance than removing any individual aromatic residue.



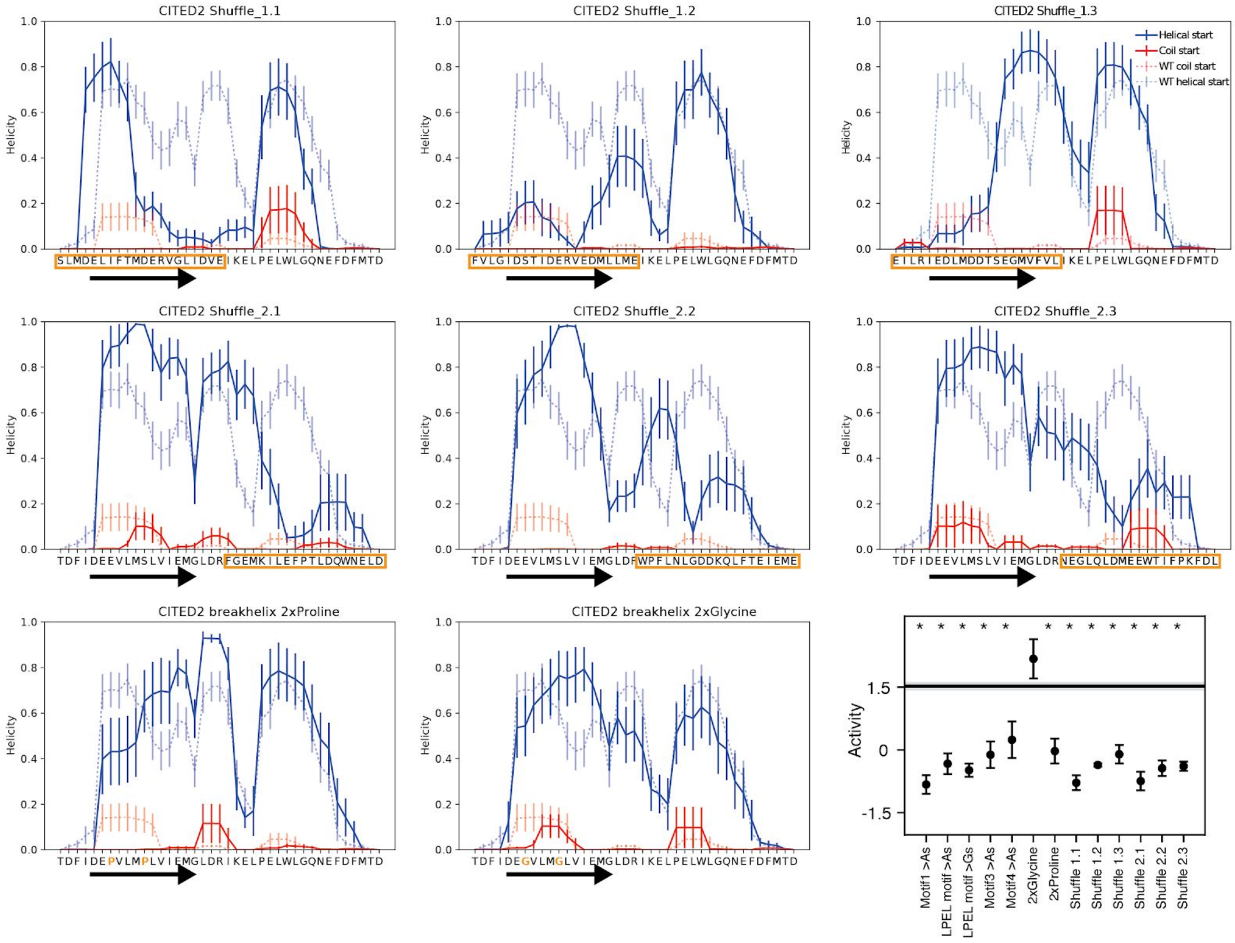
Supplemental Figure S12: Leucines and acidic residues contribute to activity of VP16 and Hif1α.

A-B) For each position, the activities of all variants that introduced a substitution at that position are summarized. The 4 biological replicates are included separately. Most variants changed multiple positions. Each residue was mutated a different number of times (1-22 variants, 4-88 measurements) and mutated to different amino acids. A) In VP16, F442 is critical for activity. The leucine and acidic residues also contribute to activity. B) For Hif1α, WT activity was low, reducing our ability to resolve decreases in activity. Leucines and acidic residues have low average activities. C) In the structure of the Hif1α-TAZ1 interaction, positions with the lowest average activities (purple) in Hif1α (orange) point towards the surface of the TAZ1 (white). Three rotations of the same snapshot are shown. The mutagenesis can detect positions that contact the coactivator.



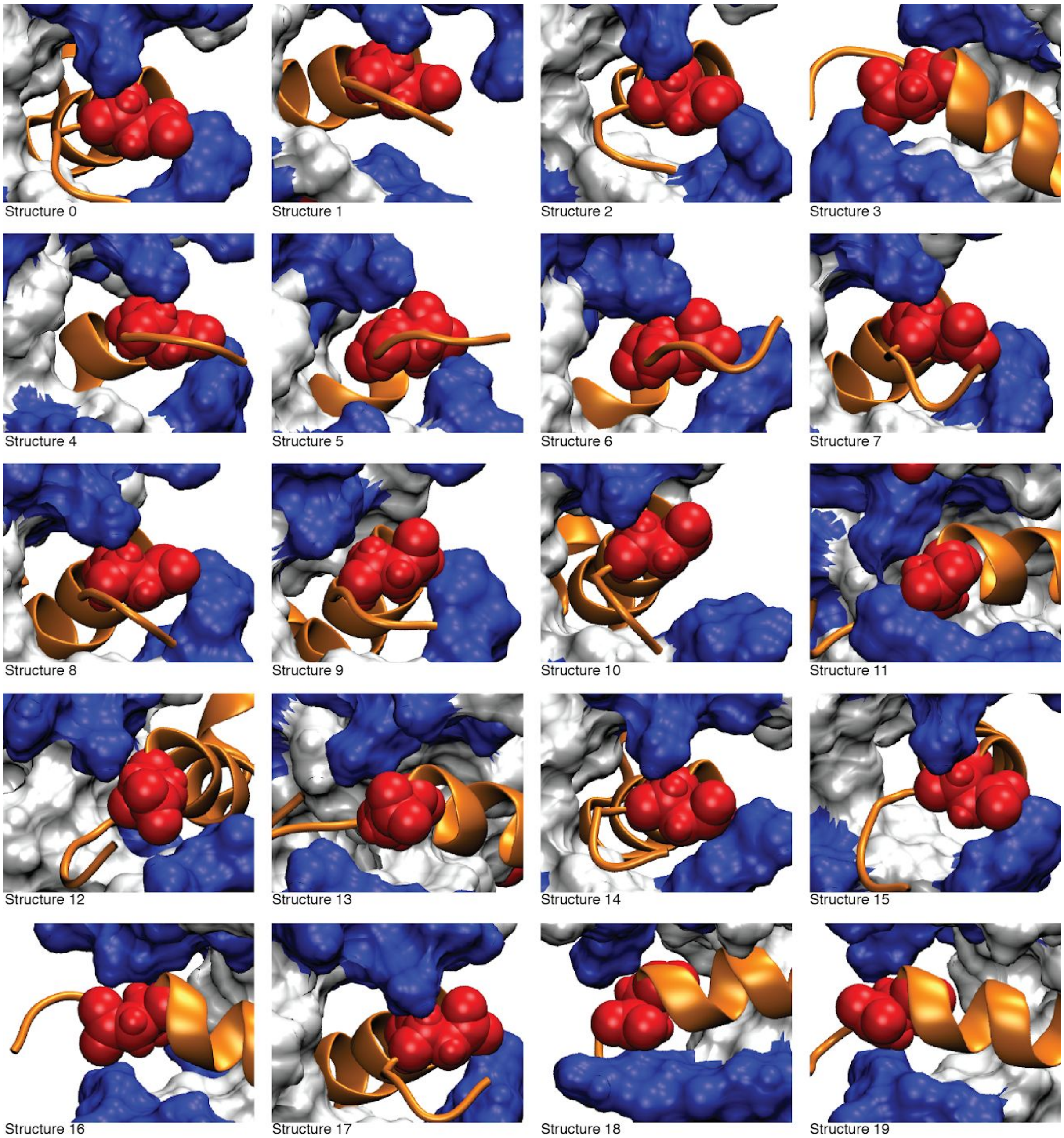
Supplemental Figure S13: The HMM logos from HMMER for each AD.

For each AD, we ran HMMER for 3 rounds, at which point VP16 and CITED2 had converged and Hif1ɑ had >126K sequences. We show the HMMER generated logo for the Hidden Markov Model (HMM) model for each AD region. The height of each letter encodes the information content of residue, a reasonable proxy for conservation. A) For VP16, F442, the leucine and aspartic acid residues are the positions with the highest information. In our data, these positions make large contributions to activity. B) For CITED2, D224 is the position in the HMM with the most information. In our data this residue contributes to activity. In addition, the leucine and acidic residues that contribute to activity in our data have high information content. C) In Hif1ɑ, C800 is important for sensing hypoxia and N803 is hydroxylated under normoxia, a modification that interferes with binding to p300 (Lando et al., 2002; Yasinska and Sumbayev, 2003). The leucine and acidic residues that contribute to activity in our data have moderate to high information content.



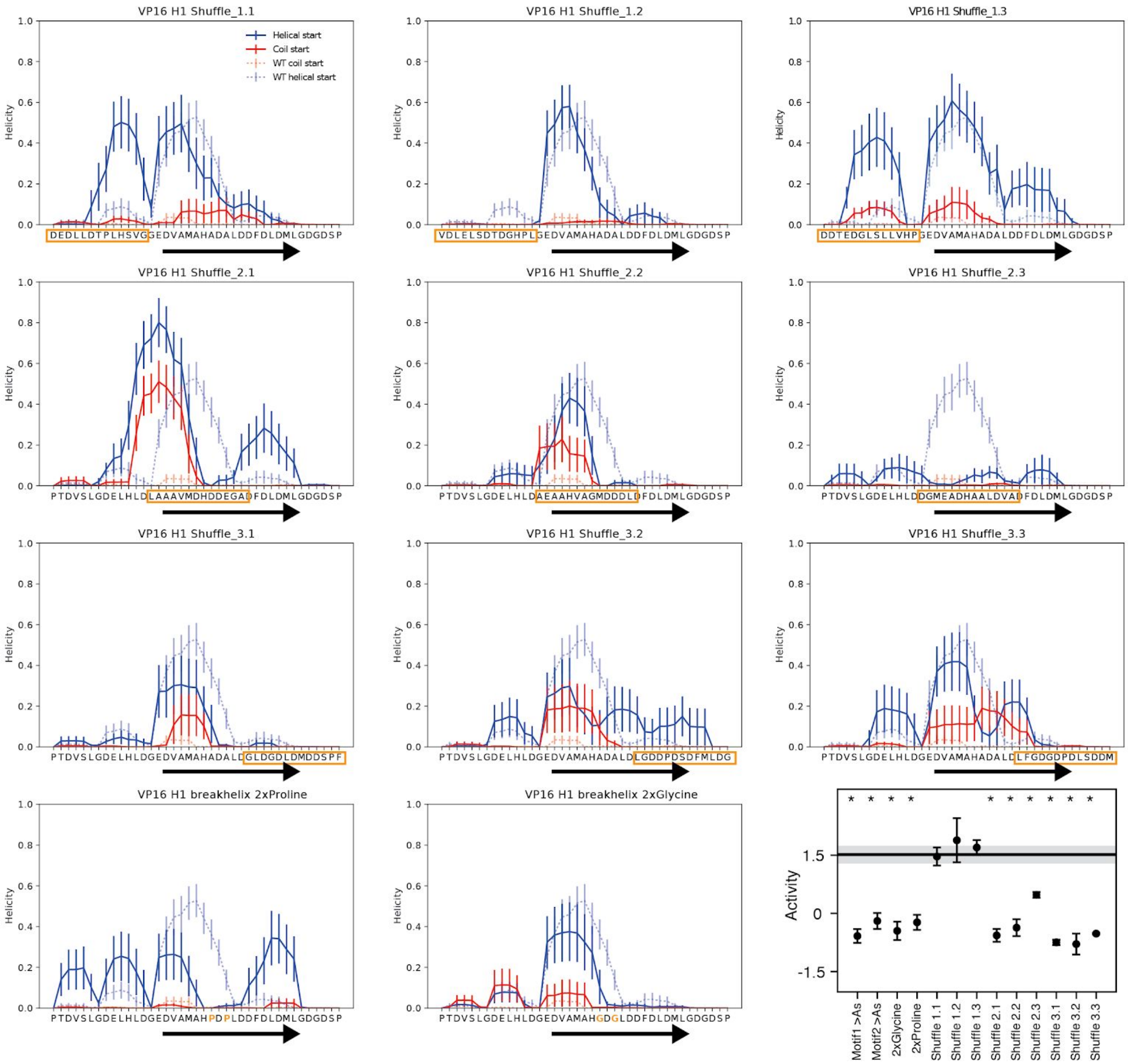
Supplemental Figure S14: Mutations designed to disrupt alpha helix formation in CITED frequently reduce helicity in the all-atom Monte Carlo simulations.

For each variant, we ran 10 simulations starting as a helix (solid blue) and 10 starting as a random coil (solid red). The arrow indicates the alpha helix observed in the NMR interaction structure with TAZ1. Not all simulations converged between these starting conditions, so it is only appropriate to compare each starting condition to its matched WT simulations (dashed lines). For example Shuffle variant 1.1 shuffled the N terminal region (orange box). In the random coil set, helicity in the NMR helix region disappeared (compare solid red line to dashed red line). In the helix set, helicity in the center was greatly reduced. We conclude that this variant reduced helicity. The 2xProline variant, in the random coil set, eliminated helicity over the NMR helix and increased helicity in the center. The 2xGlycine variant retained some helicity in the NMR region. The last panel shows the activities of these variants and motif disrupting mutants for comparison. Shuffling any region of this AD decreased activity as much as removing a motif, indicating that the arrangement (sequence) of residues is important for function.



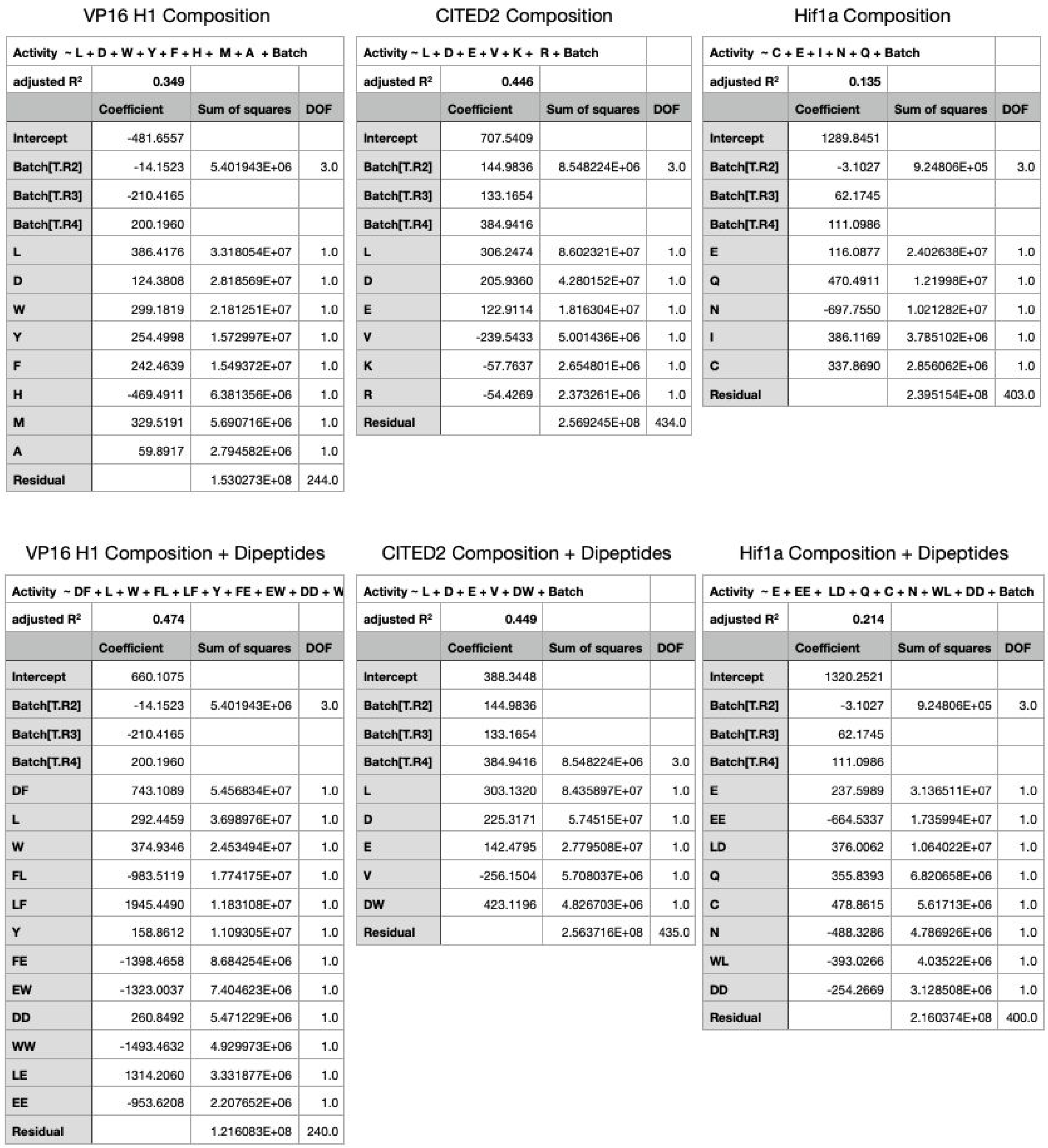
```
TDFIDEEVLMSLVIEMGLDRIKELPELWLGQNEFDFMTD  1.5  WT
TDFIDDDVLMSLVIDMGLDRIKDLPDLWLGQNDFDFMTD  1.6  ED>D
TDFIDDDVLMSLVIDMGLDRIKELPELWLGQNEFDFMTD  1.7  ED>Dmut0to20
TDFIDEEVLMSLVIEMGLDKIKELPELWLGQNEFDFMTD  1.7  RK>K
TDFIDEEVLMSLVIEMGLDRIKDLPDLWLGQNDFDFMTD  1.7  ED>Dmut20to39
TDFIDEEVLMSLVIEMGLDRIKELPELWLGQNEFEFMTE  1.6  ED>Emut20to39
TDFIDEEVLMSLVIEMGLDRIRELPELWLGQNEFDFMTD  1.5  RK>R
TEFIEEEVLMSLVIEMGLERIKELPELWLGQNEFDFMTD  0.2  ED>Emut0to20
TEFIEEEVLMSLVIEMGLERIKELPELWLGQNEFEFMTE  0.3  ED>E
```

Supplemental Figure S15: Snapshots of CITED2 bound to the TAZ1 domain of CBP. In 18/20 structures (1R8U), D224 (red) of CITED2 (orange) sits between the narrow, positively charged rims (blue) of the binding canyon on TAZ1 (white). CITED2 D224 is closest to R439 (upper blue) and K365 (lower blue) of TAZ1. The canyon is widest in structures 1 and 10, and D224 interacts primarily with R439. Structure 16 is shown in Figure 6C. Most variants that exchanged charged residues had nearly WT activity, but two variants with the D221E, D224E and D238 substitutions had decreased activity. D221 and D238 are not sterically constrained.



Supplemental Figure S16: Mutations designed to disrupt the alpha helix of VP16 frequently reduce helicity in simulations.

For each variant, we ran 10 simulations starting as a helix (solid blue) and 10 starting as a random coil (solid red). Not all simulations converged between these starting conditions, so it is only appropriate to compare each starting condition to its matched WT simulations (dashed lines). Shuffling the sequence of region 1 did not disrupt the helix and did not decrease AD activity. Shuffle variants 2.1, 2.2 and 3.1 moved the helix toward the N terminus. Variant 2.3 decreased helicity. Variants 3.2 and 3.3 perturbed helicity. The 2xProline and 2xGlycine variants disrupted helicity near the substitutions. The inset plot shows the activities of all the variants and motif disrupting variants for comparison. Variants that shuffled regions that overlapped the known helix (black arrow) reduced activity. Shuffling Region 1 did not reduce activity, but mutating the leucines in this region to alanines did reduce activity. This region appears to have fewer constraints on the arrangement of residues, and may exhibit fuzzy binding to a coactivator.



VP16 H1 Composition

Activity ~ L + D + W + Y + F + H + M + A + Batch			
adjusted R²	0.349		
	Coefficient	Sum of squares	DOF
Intercept	-481.6557		
Batch[T.R2]	-14.1523	5.401943E+06	3.0
Batch[T.R3]	-210.4165		
Batch[T.R4]	200.1960		
L	386.4176	3.318054E+07	1.0
D	124.3808	2.818569E+07	1.0
W	299.1819	2.181251E+07	1.0
Y	254.4998	1.572997E+07	1.0
F	242.4639	1.549372E+07	1.0
H	-469.4911	6.381356E+06	1.0
M	329.5191	5.690716E+06	1.0
A	59.8917	2.794582E+06	1.0
Residual		1.530273E+08	244.0

CITED2 Composition

Activity ~ L + D + E + V + K + R + Batch			
adjusted R²	0.446		
	Coefficient	Sum of squares	DOF
Intercept	707.5409		
Batch[T.R2]	144.9836	8.548224E+06	3.0
Batch[T.R3]	133.1654		
Batch[T.R4]	384.9416		
L	306.2474	8.602321E+07	1.0
D	205.9360	4.280152E+07	1.0
E	122.9114	1.816304E+07	1.0
V	-239.5433	5.001436E+06	1.0
K	-57.7637	2.654801E+06	1.0
R	-54.4269	2.373261E+06	1.0
Residual		2.569245E+08	434.0

Hif1a Composition

Activity ~ C + E + I + N + Q + Batch			
adjusted R²	0.135		
	Coefficient	Sum of squares	DOF
Intercept	1289.8451		
Batch[T.R2]	-3.1027	9.24806E+05	3.0
Batch[T.R3]	62.1745		
Batch[T.R4]	111.0986		
E	116.0877	2.402638E+07	1.0
Q	470.4911	1.21998E+07	1.0
N	-697.7550	1.021282E+07	1.0
I	386.1169	3.785102E+06	1.0
C	337.8690	2.856062E+06	1.0
Residual		2.395154E+08	403.0

VP16 H1 Composition + Dipeptides

Activity ~ DF + L + W + FL + LF + Y + FE + EW + DD + W			
adjusted R²	0.474		
	Coefficient	Sum of squares	DOF
Intercept	660.1075		
Batch[T.R2]	-14.1523	5.401943E+06	3.0
Batch[T.R3]	-210.4165		
Batch[T.R4]	200.1960		
DF	743.1089	5.456834E+07	1.0
L	292.4459	3.698976E+07	1.0
W	374.9346	2.453494E+07	1.0
FL	-983.5119	1.774175E+07	1.0
LF	1945.4490	1.183108E+07	1.0
Y	158.8612	1.109305E+07	1.0
FE	-1398.4658	8.684254E+06	1.0
EW	-1323.0037	7.404623E+06	1.0
DD	260.8492	5.471229E+06	1.0
WW	-1493.4632	4.929973E+06	1.0
LE	1314.2060	3.331877E+06	1.0
EE	-953.6208	2.207652E+06	1.0
Residual		1.216083E+08	240.0

CITED2 Composition + Dipeptides

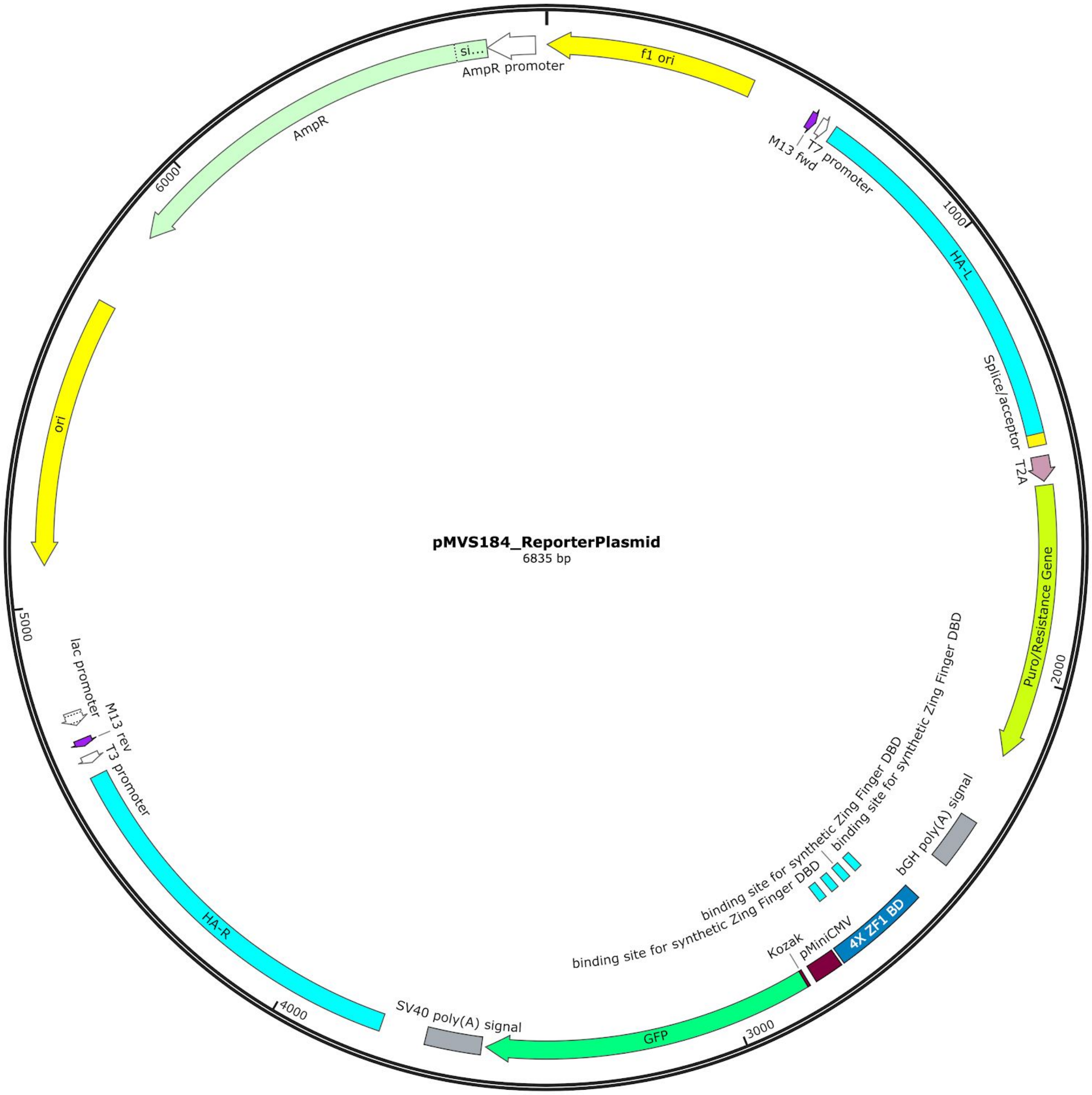
Activity ~ L + D + E + V + DW + Batch			
adjusted R²	0.449		
	Coefficient	Sum of squares	DOF
Intercept	388.3448		
Batch[T.R2]	144.9836		
Batch[T.R3]	133.1654		
Batch[T.R4]	384.9416	8.548224E+06	3.0
L	303.1320	8.435897E+07	1.0
D	225.3171	5.74515E+07	1.0
E	142.4795	2.779508E+07	1.0
V	-256.1504	5.708037E+06	1.0
DW	423.1196	4.826703E+06	1.0
Residual		2.563716E+08	435.0

Hif1a Composition + Dipeptides

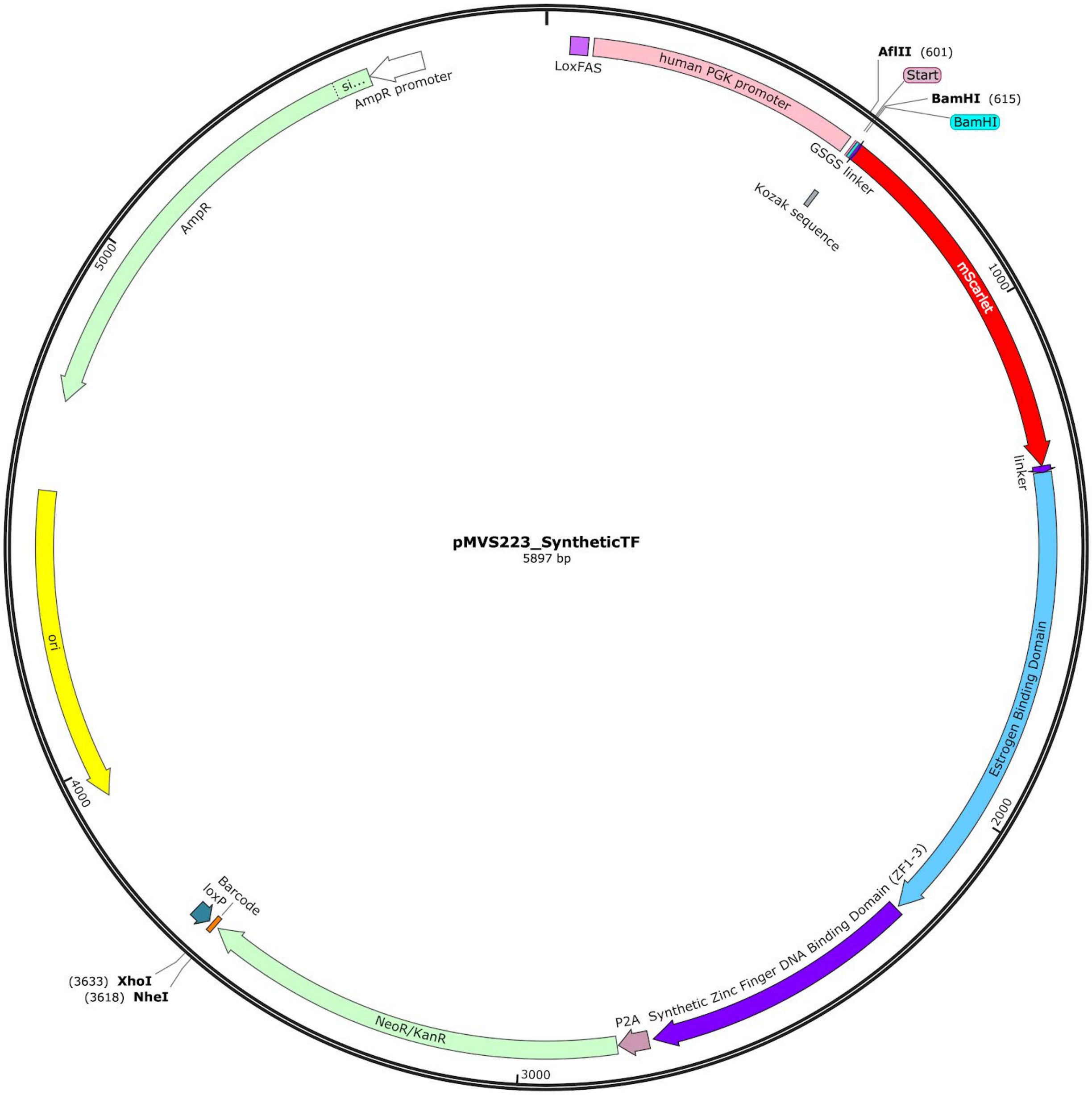
Activity ~ E + EE + LD + Q + C + N + WL + DD + Batch			
adjusted R²	0.214		
	Coefficient	Sum of squares	DOF
Intercept	1320.2521		
Batch[T.R2]	-3.1027	9.24806E+05	3.0
Batch[T.R3]	62.1745		
Batch[T.R4]	111.0986		
E	237.5989	3.136511E+07	1.0
EE	-664.5337	1.735994E+07	1.0
LD	376.0062	1.064022E+07	1.0
Q	355.8393	6.820658E+06	1.0
C	478.8615	5.61713E+06	1.0
N	-488.3286	4.786926E+06	1.0
WL	-393.0266	4.03522E+06	1.0
DD	-254.2669	3.128508E+06	1.0
Residual		2.160374E+08	400.0

Supplemental Figure S17: Composition based models can only account for a minority of the variance in activity.

We performed ANOVA on composition, regressing activity against all 20 amino acids. After iteratively removing parameters that were not significant, we arrived at the minimal composition model for each AD (top row). Separately, we trained a model with all dipeptides derived from ['D','E','W','F','Y','L'] and identified significant contributors. We added these significant dipeptide parameters to the composition model and iteratively removed parameters that were not significant. Adding dipeptides improved model performance.



Supplemental Figure S18: The GFP reporter, pMVS184.
This plasmid has four binding sites for the synthetic DBD and is integrated at the AAVS1 locus. The synthetic zinc finger DBD and cognate TF binding sites were designed and validated by Minhee Park and Ahmed Khalil (Park et al., 2019). The reporter plasmid (pMVS184) contains upstream homology to AAVS1 (804 bp), a splice acceptor, a T2A signal, puromycin resistance, the bGH poly(A) signal, four binding sites for the synthetic zinc finger DBD, the pMiniCMV minimal promoter, GFP, the SV40 3' UTR and downstream homology to AAVS1 (837 bp). The plasmid sequence is in Supplemental Dataset 6. The plasmid will be available from Addgene.



Supplemental Figure S19: Map of the synthetic TF plasmid, pMVS223.
The synthetic TF contains a loxFAS site, the human PGK promoter, a multiple cloning site for inserting AD variants, an mScarlet red fluorescent protein, an estrogen response domain, a synthetic zinc finger DNA binding domain (DBD), a P2A cleavage sequence in frame with a neomycin resistance gene, a stop codon, a barcode sequence in the 3' UTR, and a loxP site. The landing pad adds a WPRE sequence to the 3' UTR of the final transcript (Maricque et al., 2018). In the final library, the AD variants are located between the ATG START and the BamHI site, and the barcodes are between the NheI and XhoI sites. The plasmid sequence is in Supplemental Dataset 6. The plasmid will be available from Addgene.