The intrinsically disordered transcriptional activation domain of CIITA is functionally tuneable by single substitutions: An exception or a new rule?

Shwetha Sreenivasan[1], Paul Heffren[1], Kyung-Shin Suh[2], Mykola V. Rodnin[1], Edina Kosa[1], Aron W. Fenton[1], Alexey S. Ladokhin[1], Paul E. Smith[2], Joseph D. Fontes[1], and Liskin Swint-Kruse[1]

[1] Department of Biochemistry and Molecular Biology, University of Kansas Medical Center, Kansas City, Kansas, 66160, United States

[2] Department of Chemistry, Kansas State University, Manhattan, Kansas, 66506, United States

* To whom correspondence should be addressed. Tel: 1 913 588 0399; Fax: +1 913 588 9896; Email: lswint-kruse@kumc.edu

Present Address: Paul Heffren, Department of Cell Biology and Physiology, University of Kansas Medical Center, Kansas City, Kansas, 66160, United States

Present Address: Edina Kosa, Department of Biosciences, Kansas City University, Kansas City, Missouri, 64106, United States

ABSTRACT

During protein evolution, some amino acid substitutions modulate protein function ("tuneability"). In most proteins, the accessible tuneable range is wide and can be sampled by a set of protein variants that each contain multiple amino acid substitutions. In other proteins, the full range can be accessed by a set of variants that each contains a single substitution. Indeed, site-saturating substitutions at individual "rheostat" positions can sample the full range for some folded globular proteins. In proteins with intrinsically disordered regions (IDRs), many functional studies – which would also detect tuneability – used multiple substitutions or deleted small regions. These results have led to proposed mechanisms such as the "acidic exposure" model for transcriptional activation domains (AD). Only a few IDRs have been assessed with single substitutions. Results have been mixed: (i) The disordered ADs of two full-length transcription factors did *not* show tuneability, yet (ii) a fragment of another AD was tuneable by single substitutions. Here, we tested tuneability in the AD of a full-length, non-DNA-binding transcription factor, human class II transactivator (CIITA). Sequence analyses and experiments showed that CIITA's AD is an IDR. Functional assays of singly-substituted variants showed that this IDR was highly tuneable, and some outcomes were not predicted by the acidic exposure model. Instead, four tested positions showed rheostat behaviour, with a wide range of effects on transcriptional activation. Thus, tuneability of different IDRs can vary widely; future studies are needed to illuminate the biophysical features that govern whether an IDR can be tuned by single substitutions.

INTRODUCTION

Amino acid substitutions enable protein evolution and serve as a means for bioengineering functional variation. Among their outcomes, many amino acid substitutions tune protein function. A protein's full tuneable range spans from wild-type (WT), to partially-impaired, to "dead"; gain-of-function outcomes (better than WT) can also occur. To access the full tuneable range for function, some proteins require variants with multiple substitutions. In other proteins, the full tuneable range can be accessed by a variety of single substitutions (*e.g.,* [1]). Furthermore, site-saturating and semi-saturating mutagenesis (SSM and semiSM, respectively) of single positions revealed that, in some globular proteins, substituting individual "rheostat" positions can sample 50-100% of the tuneable functional range (*e.g.,* [2-9]). It is still unknown whether a classically folded protein structure is a pre-requisite for the presence of rheostat positions or whether they can also be present in intrinsically disordered regions (IDRs).

Indeed, most historical studies of IDRs, including those found in many eukaryotic transcription factors, have assessed functional tuning *via* multiple, simultaneous substitutions (*e.g.,* [1; 10-16]) or deleting small regions (*e.g.,* [17-19]). Justification for this approach was further bolstered by early SSM studies of the IDR transcriptional activation domains in p53[20] and peroxisome proliferator–activated receptor ɣ (PPARɣ)[21]: These studies showed few significant functional effects for single substitutions and did not identify any rheostat positions.[20; 21] The IDRs of p53 and PPARɣ represent the broader class of acidic activation domains; their SSM results have been extrapolated to other activation domains to guide "rational mutagenesis" studies that explored contributions on function that arise from altering sets of acidic and aromatic/leucine residues.[13-16; 22-24] These results led to the "acidic exposure model" of transcriptional activation.[13; 15]

Although dominated by combinatorial substitutions, the data sets arising from rational mutagenesis did contain a handful of single substitutions.[13; 15] Intriguingly, the set of single substitutions for the Hifα-derived activation domain showed broad functional tuning (Supplementary Figure 1; [15]), in contrast to the results for p53 and PPARɣ. The presence of single substitutions with partially-impaired or enhanced function suggests that (i) the IDR activation domains from different transcription factors can have different substitution sensitivities, (ii) the SSM/semiSM approaches to studying IDRs should not be abandoned,

and (iii) rheostat positions might be present in some IDRs. These hypotheses are further supported by earlier, albeit limited, alanine scans of IDRs, which also reveal singly-substituted protein variants with partially-diminished or enhanced function [18; 25-31].

To test these hypotheses, we chose a transcription factor with different domain arrangements than p53/PPARɣ – the class II major histocompatibility complex transactivator protein (CIITA) from vertebrates.[32; 33] Although CIITA *lacks* a DNA binding domain, it contains an acidic activation domain at its N-terminus that is critical for transcriptional activation (positions 1-161).[34; 35] To carry out its function, CIITA interacts with multiple DNA-bound complexes at the promoters of genes encoding the major histocompatibility class II proteins; CIITA also recruits other factors necessary to modulate chromatin structure and initiate transcription.[36] Several of these binding interactions are mediated through CIITA's activation domain. Furthermore, the G58P and A94P variants of CIITA retained 71% and 43%, respectively, of WT's transcriptional activation[19]; these partially-diminished activities suggest that these two positions could be rheostat positions.

Here, we first showed that the activation domain of CIITA lacked persistent secondary structure using sequence analyses, circular dichroism spectropolarimetry and molecular dynamics simulations. Within this domain, we next selected four positions from regions with high probabilities for disorder and three positions from a region with low probability. For these seven positions, we measured the effect of 10-14 single amino acid substitutions on cellular protein concentrations and normalized transcriptional activation. An important feature of our study design is that it used unbiased selection of amino acid side chains. Our agnostic approach was previously used to explore the roles of rheostat positions in globular proteins, which often showed *no* correlation between substitution outcomes and either side chain chemical similarities or structural properties.[2; 5; 6; 37; 38] As such, this study expands beyond the rational mutagenesis approach used to formulate the acidic exposure model for activation domains.

Our results showed that transcriptional activation could be tuned across the full functional range; gain-of-function substitutions (better than WT) were also observed. All seven tested positions were sensitive to individual substitutions, *i.e.,* no position was neutral. Furthermore, the sets of substitutions for four positions sampled at least half of the range

observed for transcriptional activation, which classifies these positions as rheostat positions. Strikingly, we did not observe any functional trends related to amino acid side chain chemistries, including outcomes expected from the acidic exposure model. However, for two positions, transcriptional activation did correlate with amino acid structural propensities, suggesting that this region of CIITA undergoes coupled folding and binding. Overall, modulating structural change could be one means by which the various functions of some IDRs are tuned by single amino acid substitutions.

RESULTS

Sequence and structure analyses indicate that the activation domain of CIITA is intrinsically disordered.

Most transcriptional activation domains contain intrinsically disordered regions (IDRs). IDRs were first identified by their unique sequence features [39], which include higher fractions of charged residues and structure-breakers (proline and glycine) and fewer hydrophobic residues than globular proteins (reviewed in [40]). Since these early studies, several computational algorithms have been developed to predict disorder from sequence. Thus, for the transcriptional activation domain of CIITA (canonical isoform I; accession NM_001286402; 1130 amino acids;123.5 kDa), we obtained its disorder probabilities using Predictor of Natural Disordered regions (PONDR) (45). Much of the N-terminus, which contains the activation domain (residues 1 to 161), has a high likelihood (>0.5) for intrinsic disorder (Figure 1; Supplementary figure 2a; Supplementary table 1). Results from alternative IDR predictors[41-44] show highly similar trends across the whole protein (Supplementary figure 2).

Another general feature of IDRs is that they appear to lack secondary structures in their unbound states,[45] although some may fold upon binding ligand.[46-48] To experimentally confirm that the activation domain of CIITA is intrinsically disordered, we expressed and purified fragment 1-210 of CIITA and performed circular dichroism (CD) studies. CD spectra are consistent with this region being primarily random coil (Figure 2), even in the presence of the structure-promoting agent, trimethylamine N-oxide (TMAO).

To investigate the possibility of persistent residual structure at the atomic level, 5 μs of classical molecular dynamics (MD) simulations of the fragment containing positions 56-94 of CIITA were performed with three different force fields specifically developed and tested for IDRs (Figure 3 and Supplementary figure 3). The radius of gyration and end-to-end distances observed for all three force fields were indicative of peptides with little to no residual long-range structure, as expected for an IDR. This result was consistent for multiple independent simulations using different starting structures for each force field. The only exception was a single simulation using the CHARMM36m force field, which predicted a more collapsed structure (Supplementary figure 3). However, this collapsed structure did not display any significant secondary structure elements as provided by analyses with Define Secondary Structure of Proteins (DSSP)[49; 50] of the trajectory. An analysis of the helical content for each force field, averaged over all three simulations, is shown in Figure 3. The KBFF20 force field suggested some transient helical content for several regions of the peptide. Together, the results suggest the simulated CIITA fragment displays the characteristics of an IDR with possible transient helix formation.

Towards the end of our project, the predicted structure for full-length CIITA became available from AlphaFold2 (Supplementary figure 4).[51] In agreement with sequence analyses, AlphaFold2 predicts that most of the activation domain lacks secondary or tertiary structure; the region tested in this study (between amino acids 61 and 92) was predicted to contain a small helix, albeit with low confidence.

The intrinsically disordered activation domain of CIITA is functionally tuneable by single substitutions.

Functionally, experimental evidence shows that CIITA binds to several protein complexes through its activation domain: BRG1[52], CBP[19], TAF9[53], and SRC1[54]. These binding interactions are similar to those of CIITA's homologs in the nucleotide-binding and leucine-rich-repeat family,[55] even though CIITA lacks a DNA binding domain. These protein-protein interactions might be sensitive to single amino acid changes, even though the activation domain is intrinsically disordered. Thus, we chose seven CIITA positions from regions with high (I62, Y65, T70, T91) and low (N74, Q77, F78) probabilities for intrinsic disorder (Table 1). This region contains binding sites for the four protein complexes listed above as well as an

additional 10 predicted short linear motifs (SLiMs)[56] that are predicted to be involved in putative protein-protein interactions (Supplementary Table 2).

To identify the overall role of a given amino acid position for tuning protein function, multiple substitutions must be experimentally assessed. Neutral behavior is assigned to amino acid positions for which >70% of its substitutions maintain wild-type-like function for all measured parameters[57]; this is the behavior previously observed for most positions in the activation domains of p53[20] and PPARγ[21]. Rheostat behavior is assigned to amino acid positions for which the substituted variants sample ≥50% of the accessible functional range;[2] this behavior has been observed for numerous positions in globular proteins[1-3; 5; 6; 9; 37; 58; 59]. Toggle behavior is assigned to amino acid positions for which >64% of the substituted variants abolish function;[38] this is the text-book substitution behavior frequently observed at important, conserved positions in globular proteins.

These three types of outcomes can be quantified using a modified histogram analysis that returns three scores (neutral, rheostat, and toggle) for each position.[59] When assigning a position's overall functional behavior, it would be ideal but expensive to substitute and assess all 20 amino acids (SSM). Fortunately, we empirically determined that ~10 substitutions (semiSM) are sufficient to categorize each position as rheostat or toggle and can provide a reasonable estimate of neutrality.[2; 7; 9; 37; 57; 59] The current, semiSM CIITA experiments used 10-14 variants (including WT) for each of the seven studied positions (Table 1).

First, we measured the effects of substitutions on CIITA's overall *in vivo* function in regulating an MHC class II promoter controlling expression of a luciferase reporter gene (see Methods, equation 1; Supplementary figure 5). The observed range for the set of substitutions spanned ~80-fold (Supplementary figure 6). Using this parameter, all seven positions were non-neutral, and three positions (position 62, 65, and 77) met the threshold for being strong rheostat positions (rheostat scores >0.5). However, the effects measured *in vivo* could arise from changes in either cellular concentration, or the events required for transcriptional activation, or both.

Thus, we also measured the effects of the amino acid substitutions on cellular protein concentration of CIITA in HEK293 cells (see Methods, equation 2; Supplementary figure 5).

These concentrations can be affected by changes in numerous processes, including transcription of the CIITA cDNA, mRNA translation, proteolytic susceptibility, and protein stability; one or more of these processes could be altered by amino acid substitution. Nonetheless, many of the 81 substituted CIITA variants (58 variants from one-way ANOVA; 48 variants from FDR analyses) showed WT-like levels (Figure 4, Supplementary table 3). However, when effects were considered at individual positions, only position 91 could be classified as neutral for cellular concentration (RheoScale neutral score >0.7; Table 1). All other positions had a few variants with diminished concentration and, as such, were non-neutral. No positions met the threshold of a strong rheostat position (RheoScale rheostat score >0.5) (Figure 4; Table 1).

Additionally, we used protein concentration to normalize the *in vivo* functional data and thereby determined the ability of each CIITA variant to activate transcription (see Methods, equation 3; Supplementary figure 7). Again, a wide range of outcomes was observed (Figure 5), with some variants diminishing and others enhancing transcriptional activation. The overall range spanned >30-fold. Strikingly, no positions were neutral; all had at least a few variants that altered transcriptional activation (Figure 5). Four of the seven amino acid positions (62, 65, 77, and 78) met the threshold for rheostat behaviour (rheostat score ≥0.5): These included the three positions with rheostatic effects on the overall *in vivo* function as well as position 78. The latter was not detected in the *in vivo* data because effects on transcriptional activation were compensated by effects on protein concentration. For these four positions, we think it most likely that substitutions alter protein-protein interactions, but they might also alter CIITA's sub-cellular localization[60].

Finally, we note that the experimental design using single substitutions revealed information that could not be gleaned from conventional deletion studies. For example, the change observed in the *in vivo* function of CIITA (the parameter most comparable to previous CIITA deletion studies[19]) ranged from 5% (Y65H) to 400% (I62V) of WT. As such, transcriptional activation for Y65H was lower than that of the previously-assessed Δ51-71 deletion variant, which retained 27% WT activity[19]. Furthermore, the variants with enhanced activation could not be detected at all in deletion studies.

The locations of these CIITA rheostat positions cannot be inferred from phylogenetic patterns of change.

In two globular proteins, the set of positions that showed subfamily-specific evolutionary conservation were enriched with rheostat positions,[9] and extremely nonconserved positions were neutral.[57] Although disordered proteins were initially assumed to be highly non-conserved, increases in sequencing information have shown that disordered sequences have both conserved and non-conserved positions.[61] To ascertain whether the rheostat/phylogeny correlation holds for CIITA, we used ConSurf analyses of a multiple sequence alignment (MSAs) to predict position-wise conservation scores [62].

For these analyses, input MSAs for CIITA were generated by two independent approaches (see Methods); both yielded similar conservation trends (Supplementary table 4). The N-terminus contains alternating regions of high and low conservation, with no apparent correlation with disorder probabilities (Supplementary figure 7). For the seven tested positions, no correlation was observed between ConSurf scores and substitution outcomes. However, CIITA is only present in vertebrates; our other studies correlating substitution outcomes with evolutionary scores were for pyruvate kinase, which are present in all domains of life[7] and LacI/GalR homologs, which are present in almost all bacteria[37; 63]. Both of these families are much older than CIITA, providing more opportunities to explore sequence space. As such, MSA-based predictions about CIITA positions (and perhaps other taxonomically-restricted proteins) may not be meaningful.

Tuneability from single substitutions in the activation domain of CIITA are not explained by amino acid side chain chemistries.

Next, we assessed whether chemically similar side chains had similar outcomes on transcriptional activation. For rheostat positions in globular proteins, one puzzling feature was that substitution outcomes were not well explained by the "textbook" biophysical properties of amino acid side chains.[2; 5; 6; 37] For positions in CIITA, when the rank order of substitution outcomes was color-coded according to amino acid side chain chemistry, we again did not observe any obvious relationship (Figures 4 and 5, Supplementary figure 6, bar

colours). For example, the transcriptional activation of I62N was 35% of WT, whereas the "similar" I62Q substitution increased transcriptional activation to 200%.

Next, we assessed whether changing the balance between acidic and hydrophobic residues in this region of CIITA altered function. Activation domains typically contain hydrophobic motifs flanked by acidic residues, and studies using engineered transcriptional activator proteins showed that the balance of these two features can be altered to vary transcriptional activation (the acidic exposure model).[13-16] In this study, although we did not mutate positions with acidic residues in the WT sequence, we did introduce at least one acidic residue at the chosen positions by our unbiased substitution strategy. All of the acidic substitutions either reduced or maintained transcriptional activation; this outcome is consistent with the acidic exposure model. On the other hand, most of the single aromatic substitutions maintained WT function; only T70F showed the enhanced function predicted by the acidic exposure model. However, substitutions with other amino acid side chains showed bigger effects on CIITA function than those that altered the charge balance.

Changes in transcriptional activation may be explained by altered structural features.

Since side chain chemistries do not explain changes in functional outcomes, we next assessed other structural features. To this end, we correlated the experimentally-measured parameters with: (i) side chain size, as calculated from G-X-G tripeptides;[64] (ii) amino acid helical propensities, which were expressed as free energies (*i.e.,* lower values indicate higher helical propensity);[65] and (iii) changes in disorder probabilities arising from substitution (Supplementary figure 8 (control computations for this approach are described in Supplementary Figure 9)).

When the measured cellular concentrations of each CIITA variant were compared to the three structural features, no meaningful correlations were observed (Supplementary figures 10-12). In contrast, when transcriptional activation was assessed (Supplementary figures 13-15), three positions showed intriguing correlations (Figure 6). Rheostat position 78 exhibited a relationship between transcriptional activation and amino acid size (Spearman r = 0.56), with an increase in size corresponding to increased transcriptional activation.

Rheostat position 62 exhibited a relationship between transcriptional activation and helical propensities (Spearman r = -0.68), which suggests that an increase in helix formation favours transcriptional activation (Figure 6). Rheostat position 77 showed a negative correlation between transcriptional activation and disorder probability (Spearman r = -0.67), indicating that higher disorder probabilities led to lower transcriptional activation (Figure 6, Supplementary figure 15).

The changes in transcriptional activation levels for positions 62 and 77 are especially intriguing in light of the MD simulations, which predicted that transient helices occur in this region of the activation domain (average per residue helicity: KBFF20 = 16%, Charmm = 5%, and Amber = 4%; Figure 3). Indeed, the average helicity at position 62 was as high as 43% (in KBFF20). Thus, this region of CIITA may undergo coupled folding and binding as it interacts with the binding partners required for transcriptional activation. Furthermore, substitutions in this region might tune transcriptional activation by altering the helical content.

DISCUSSION

Our results show that individual substitutions in the intrinsically disordered activation domain of CIITA can alter function: Substitutions at just seven of the 1130 CIITA positions sampled an 80-fold range in overall *in vivo* function and a >30-fold range in normalized transcriptional activation (Figure 7). These changes meet our definition of "tuning" – the set of substitutions sampled most of the range bounded by WT and the "no activation" control. In addition, comparable numbers of diminishing and enhancing substitutions were identified (Supplementary Table 3). As such, the CIITA activation domain was more tuneable than the activation domains of p53[20] and PPARγ[21].

Furthermore, substitution outcomes did *not* follow the patterns predicted by the acidic exposure model, which proposes that altering the balance of net charge and hydrophobicity in disordered activation domains provides the molecular mechanism by which their functions are tuned: The charged residues prevent the hydrophobic residues from collapsing onto themselves and thereby allow them to interact with partner proteins. Thus, according to the acidic exposure model, decreasing the net charge (by increasing the

proportion of acidic residues) should enhance activation, whereas increasing the net charge/hydrophobicity should diminish activation. A tempting extrapolation of this model is to assume that substituting amino acid side chains with other chemistries will have no effect on transcriptional activation and, thus, rheostat positions would not exist. However, this was *not* observed in CIITA: (i) Substituting a variety of side chain chemistries altered transcriptional activation over the tuneable range, (ii) several substitutions had greater impact than changes that altered the acidic/hydrophobic balance, and (iii) rheostat positions did exist.

The simplest explanation for the discrepancy between the results observed for CIITA and those expected from the acidic exposure model is that different types of activation domains exist, each associated with different sensitivities to substitution. For example, CIITA and full-length p53/PPARɣ clearly exhibit different sensitivities to single substitutions. This demonstrates that substitution outcomes should *not* be generalized across all activation domains and allows for the possibility that the acidic/hydrophobic balance controls function in some proteins, whereas rheostat positions are present in others.

On the other hand, none of the existing studies rule out the possibility that the two means of functional tuning can co-occur. First, the presence (or absence) of rheostat positions, which requires a semiSM/SSM experimental design, could not have been detected by the "rational mutagenesis" design used to develop the acidic exposure model. Second, the acidic exposure studies were carried out with fragments of activation domains linked to a non-native DNA binding domain; the presence of rheostat positions might require the context of full-length protein. Third, the current work with CIITA did not exhaustively test the whole activation domain, and – although various substitutions changed the charge or hydrophobicity at individual positions – the single substitutions only caused minor perturbations to the overall charge balance. Thus, future studies will be required to determine whether rheostat positions and acidic exposure behaviour co-occur or are mutually exclusive.

We must also determine the molecular mechanisms by which substitutions at IDR rheostat positions exert their effects on transcriptional activation. Since the existence of rheostat positions is not fold-dependent, perhaps the biophysical changes associated with

their substitutions are similar in both IDRs and globular proteins. In the globular protein LacI, the modulating effects of substitutions were well-described by perturbations to an all-atom dynamics model comprising an asymmetric, weighted network of balls (nodes/atoms) and springs (edges/bonds);[8] this computation has also been successfully correlated with single substitution outcomes in other proteins (*e.g.,* [4]). Notably, such a model does *not* require the presence of ordered secondary structure. Instead, any atom added or deleted from the protein has the potential to perturb the asymmetric weighted network. CIITA function requires binding of its activation domain to partner proteins and thus CIITA participates in a structural network. As such, single amino acid substitutions could have the same potential to alter functionally-relevant dynamics.

In conclusion, substitution sensitivities cannot be generalized among IDR activation domains. Furthermore, aside from the p53 and PPARγ studies, IDRs have been largely passed over in SSM studies; of the 161 proteins with IDRs in MaveDB[70], only their globular domains have been mutated as of this writing. Thus, future SSM/semiSM studies of full-length transcription factors are required to illuminate mechanisms of functional tuning in IDRs.

MATERIAL AND METHODS

Sequence analyses of full-length CIITA

The human CIITA sequence exists as multiple isoforms that arise from alternative promoter usage and mRNA splicing (44). The full-length human CIITA sequence for isoform 1 (Uniprot P22076.1), comprising 1130 amino acids, was used in all analyses.

Intrinsic disorder prediction. The CIITA primary amino acid sequence was analysed with the Predictor of Natural Disorder Region (PONDR) software using its VL-XT subroutine to obtain disorder probabilities [71] (Supplementary figure 2a, Supplementary table 1). We also computationally edited the WT CIITA sequence to create full-length sequences corresponding to all possible single substitutions at the seven chosen positions (133 sequences) and all possible substitutions at a different position with high disorder probability and repeated the PONDR analyses (Supplementary figures 8 and 9). As an alternative computational approach, we used Rapid Intrinsic Disorder Analysis Online (RIDAO) [72], which integrates disorder probability prediction outcomes from different IDR

predictors and their default analysis parameters (Supplementary figure 2b; Supplementary table 1; Supplementary figure 9).

Conservation scores. To compute evolutionary information for CIITA positions, we used the human CIITA sequence in an automated ConSurf computation using default parameters: This webserver first used an HMMER-based[73] search to build a multiple sequence alignment (MSA; named "CIITA_ConSurf_default_MSA", Supplementary data) using a MAFFT-LINS-i alignment [74], which was then subjected to phylogenetic/conservation analysis. This MSA was limited to 150 sequences, which might lead to biased ConSurf scores due to insufficient sampling; larger MSAs could not be constructed using the default ConSurf parameters because the CIITA sequence is so long that the process timed out.

Thus, we also created a larger CIITA MSA by manual curation: The CIITA primary amino acid sequence was used as a BLAST query against the nr-database in the NCBI server [75]; results were filtered using 80 to 100% query coverage to yield 1102 unique sequences. These sequences were aligned using clustalOmega [76]. This initial MSA (i) was manually curated to remove sequences with insertions >100 amino acids or >100 sequential gaps, leaving 1017 sequences, and (ii) a python script[77] (https://github.com/ncbi/FixJ_KdpE/tree/main/clusters/cull_MSA.py) was used to exclude other long and short sequences: To filter out sequences with long inserts (which can disrupt alignments of other sequences), sequences were searched to identify windows of 8-positions for which <5% of the other sequences were occupied with an amino acid. To filter out sequences that were too short (which could bias the gap penalty calculations for the region and slow down ConSurf calculations), sequences were searched to identify windows of 10-positions for which >90% of the other sequences contained amino acids. The remaining 603 sequences were re-aligned in clustalOmega to produce the MSA named "CIITA_manual_MSA" (Supplementary data).

CIITA_manual_MSA was input into ConSurf[62] to identify position-wise conservation scores. However, ConSurf was unable to analyze this MSA, perhaps again because it was too large. Thus, from the CIITA_manual_MSA, we randomly extracted five subsets for each of 250, 300, 350, 400, 450 sequences using python scripts (https://github.com/ShwethaSreenivasan/CIITA_Swint-Kruse_Lab_2023). Except for one of the

450-sequence MSA subsamples, all ConSurf analyses were successful; position raw scores were in good agreement with each other for subsets ≥300 sequences (Spearman, r = 0.96-0.99, Supplementary figure 16; ConSurf's transformation of raw scores into 9 categories is further described in Supplementary table 4). Thus, we used the 19 largest subsets to calculate an average normalized ConSurf score for each position (Supplementary table 3).

ConSurf results for the manual and default MSAs were similar (Spearman, r = 0.85, Supplementary figure 16i). ConSurf scores of the experimentally-assessed positions ranged from 2 (low) to 9 (high) (Table 1, Supplementary table 4). ConSurf scores from 3-8 were previously associated with a high prevalence of rheostat[9] positions in soluble globular proteins. Also in soluble globular proteins, positions which showed the least pattern of change within an MSA (*i.e.,* changes did not follow any of several phylogenetic or co-evolutionary patterns) were associated with a prevalence of neutral positions;[57] such positions usually have ConSurf scores of 1-2.[9] However, for CIITA, comparison of ConSurf and PONDR scores showed no correlation (Spearman, r = 0.16-0.17, Supplementary figure 7).

Predictions for short linear motifs (SLiMs). SLiMs are 3-10 amino acid long regions found especially in intrinsically disordered sequences that contain putative protein-protein interaction sites. Using the full-length CIITA, analyses with Eukaryotic Linear Motif[56] predicted 43 SLiMs in the activation domain (Supplementary table 2).

Spectroscopic characterization of CIITA protein fragment 1-210

Based on the PONDR analysis (Figure 1), a fragment of the CIITA protein containing amino acids 1-210 with an N-terminal HIS6-tag was expressed utilizing a pET28b expression plasmid. This region spans all seven positions chosen for experimental substitution studies. Briefly, *E. coli* Rosetta cells (Millipore Sigma, Catalog #70954) were transformed with pET28b-CIITA 1-210 and a single colony was used to inoculate 5 mL LB medium with 50 µg/mL kanamycin and grown overnight at 37°C. From that overnight culture, 100 µL were inoculated in 1 L of LB medium supplied with kanamycin (50 µg/mL) and grown to OD_{600} of 0.6. The temperature was reduced to 20°C and 0.25 mM IPTG was added. After 22 h growth, cells were harvested by centrifugation and frozen at -80 °C until purification.

For purification of the CIITA protein fragment, cells were resuspended in binding buffer (50 mM Tris-HCl pH 8.0, 300 mM NaCl, 5 mM imidazole, and 1 mM PMSF) and lysed by sonication. Cleared supernatant was mixed with Ni-NTA agarose (GE Healthcare, MO), bound overnight with mixing, and washed with binding buffer. Protein was eluted with 0.5 M imidazole in the same buffer. Additional purification was performed using AKTA FPLC system on the Superose 12 column (GE Healthcare, MO) equilibrated with 50 mM sodium phosphate buffer (pH 8.0).

CIITA fragment 1-210 concentration was determined by absorbance measurements at 280 nm using molecular extinction coefficient of 14000 mol^{-1}, which was estimated from the one tryptophan and seven tyrosine residues in this sequence. Circular dichroic (CD) spectra of the CIITA 1-210 fragment were obtained using a Jasco-720 spectropolarimeter (Japan Spectroscopic, Tokyo, Japan). Samples containing 4 µM of protein in 50 mM sodium phosphate buffer pH 8.0 were placed in 1-mm optical path cuvette at room temperature in the absence or presence of 1 M trimethylamine N-oxide hydrate (TMAO) (Hampton Research, CA). Ellipticity data were collected for 200-260 nm with a 1 nm step. Data from 100-150 scans were averaged and the background was subtracted to generate the final spectra.

Molecular dynamics simulations of CIITA fragment 56-95

Based on the PONDR analysis (Figure 1), a truncated sequence of CIITA encompassing residues L56 to A95 (39 residues total) was selected for molecular dynamics simulations (MD). This region spans all seven positions chosen for experimental substitution studies. First, the coordinates for a blocked (N-acetylated and C-N methylated) sequence in a fully extended conformation were generated using the Avogadro software [78]. Then, three random initial configurations were selected from preliminary 100 ns implicit solvent simulations using a simple dielectric constant of 80. Finally, explicit solvent simulations were performed starting from the three different initial conformations.

All simulations were performed with the GROMACS package version 2016.4 [79]. Simulations of each configuration were performed with three different force fields: KBFF20 [80] with SPC/E water model [81]; CHARM36m [82] with TIP3P water model [83]; and AMBER99sb-ildn [84]

with TIP4P-D water model [85]. Each configuration of the CIITA fragment was solvated in a rhombic dodecahedral box with an image distance of 7.5 nm using periodic boundary conditions. The systems contained approximately 9400 water molecules, after neutralization with 12 Na^+ ions and the addition of 25 more Na^+ and 25 Cl^- ions to achieve a physiological (0.15 M) salt concentration.

The equations of motion were numerically integrated using the Verlet leapfrog algorithm [86] with a time step of 2 fs. The short-range interactions were calculated using a nonbonded pair list with a single cut-off of 1 nm with a Verlet list update. PME was applied for the long-range electrostatic interactions [87]. Nose-Hoover temperature coupling was used to maintain a target temperature of 300 K [88], and Parrinello−Rahman pressure coupling was used to maintain a target pressure of 1 bar [89]. All peptide bond lengths were constrained using the LINCS algorithm [90], while all solvent bonds were constrained using SETTLE [91]. After minimization using the steepest descents method, a 1 ns equilibration was performed using Berendsen temperature and pressure coupling at 300 K and 1 bar [92]. Production simulations of 5μs were performed for each initial configuration and each force field.

Generation of variants in full-length CIITA

The plasmid pUNO1-hCIITA (InVivogen, Inc., San Diego, CA; catalog code pUno1-ciita), which encodes the CIITA cDNA, was used to generate CIITA variants. For each targeted amino acid position, 9-13 unique mutations were generated by random, site-directed mutagenesis using the QuikChange II XL kit (Agilent Technologies, Inc., Santa Clara, CA). For each targeted codon, four sets of primers were synthesized (Integrated DNA Technologies, Inc. (Coralville, IA) with each set substituting one of the random codons ANN, CNN, GNN, or TNN for the target codon. Plasmids purified from individual clones were sequenced to identify mutations in the CIITA cDNA (ACGT, Wheeling, IL). All plasmid isolations used the ZymoPURE Plasmid Miniprep Kit (Zymo Research Corp., Irvine, CA). Unless otherwise stated, the manufacturer's protocols were followed for all procedures.

Transfection and Luciferase Assays

HEK293 cells (American Type Culture Collection, Manassas, VA; catalog number CRL-1573) do not express the MHC class II genes, making this cell line suitable for reporter assays of CIITA transcriptional activation. To that end, HEK293 cells were cultured in complete growth medium consisting of Dulbecco's Modified Eagle Medium with 4.5 g/L glucose, supplemented with 10% fetal bovine serum, 100 µg/mL penicillin, and 100 U/mL streptomycin. Cells were maintained at 37 °C, under a humidified atmosphere containing 5% CO_2. Growth was monitored daily, and the cells were split once they reached 80-90% confluence. Medium was replaced every two to three days, as required. Cells were used for a maximum of 15 passages. Subcultures for transfection were prepared in 24-well plates with 1.5×10^5 cells and 1 mL complete medium per well. Subcultures were incubated overnight prior to transfection.

All plasmid transfections were performed in quadruplicate. Each transfected well received 100 µL Opti-MEM (Thermo Fisher Scientific, Waltham, MA), 100 ng of reporter plasmid pDRA-Gluc (the *Gaussia* luciferase gene under the control of the HLA-DRA promoter, responsive to CIITA) [93], 100 ng pSV40-Cluc (expressing *Cyrpridina* luciferase, which served as a transfection efficiency control), 800 ng pUNO1-hCIITA (WT or variant), and 2 µL TurboFect transfection reagent (Thermo Fisher Scientific). Culture media was changed four hours post transfection. Cells were incubated overnight before performing the luciferase assay and harvesting cells for western blot analysis.

Luciferase assays were carried out in 24-well culture plates. For each plate, one transfection mixture was prepared with WT pUNO1-hCIITA as a positive control for activation of expression from the pDRA-GLuc plasmid and one transfection mixture was prepared with pCMV-LacZ as a negative control. Luciferase assays were performed on aliquots of cleared media harvested from transfected cells using the Pierce *Gaussia* and *Cypridina* luciferase flash assay kits (catalog numbers 16159 and 16169 respectively; Thermo Fisher Scientific) and a GloMax 20/20 luminometer (Promega) per manufacturer's instructions.

Cellular concentrations of CIITA variant proteins

To assess the relative *in vivo* concentrations of CIITA variant proteins, cells were harvested and lysed in western loading buffer as described. Proteins were separated on precast 4-15% gradient polyacrylamide gels (Bio-Rad, Hercules, CA) and transferred to PVDF membranes (Bio-Rad, Hercules, CA). Membranes were blocked in 1X TBST and 1X fish gelatin (Biotium, Fremont, CA) and blotted with E-12 mouse monoclonal anti-human CIITA (catalog number sc-376174, Santa Cruz Biotechnology, Inc., Santa Cruz, CA) at a dilution of 1:100 and 9F3 rabbit monoclonal anti-human β-tubulin from (catalog number 2128, Cell Signaling Technology, Inc., Danvers, MA) at a dilution of 1:5000, incubating overnight at 4°C. After washing, secondary blotting was carried out in the same buffer with donkey anti-mouse IgG H+L Alexa Fluor 488 at a dilution of 1:2000 and donkey anti-rabbit IgG H+L Alexa Fluor 647 at a dilution of 1:5000 (Catalog #150079, Abcam, PLC, Cambridge, UK). Membranes were imaged using a GE Typhoon FLA9500 biomolecular imager in fluorescence mode. Two images were obtained for each membrane, using the 473 nm and the 635 nm excitation wavelengths. Quantification relative concentrations of protein in the CIITA and β-tubulin bands was accomplished using ImageJ version 1.52a (Rasband, W.S., U. S. National Institutes of Health, Bethesda, MD).

Analyses of experimental data

Four experimental measurements were made for each CIITA WT and variant sample (Supplementary Figure 5): (i) *Transcriptional activation by CIITA* was measured using the *Gaussia* luciferase activity from the reporter plasmid (pDRA-GLuc) as a proxy. (ii) Variations in *transfection efficiency* were controlled by measuring the *Cypridina* luciferase signal intensity from the transfection control plasmid (pSV40-CLuc). (iii) *Cellular CIITA protein concentration* was determined from the intensity of the WT or variant CIITA western blot bands. (iv) Experimental variation in *cell density* in the western blot was controlled by the intensity of the β-tubulin band. For each variant, two or more technical replicates of four or more biological replicates were assayed. All measurements were then normalized to the value for a WT CIITA measured in parallel, to control for day-to-day technical variation. The four normalized, independent measurements were used to determine the different functional outcomes available in https://github.com/ShwethaSreenivasan/CIITA_Swint-Kruse_Lab_2023.

In vivo function = (CIITA reporter activity/Transfection efficiency)/(Cell density) *(1)*

Cellular concentration = (CIITA protein levels)/(Transfection efficiency/Cell density) *(2)*

Transcriptional activation = (CIITA reporter activity/Transcription efficiency)/(CIITA protein levels/Cell density) *(3)*

To determine which variants significantly differed from WT, statistical analyses were carried out using GraphPad Prism version 7.05.237. D'Agostino-Pearson tests showed that the replicates for several variants followed a log normal distribution, rather than a normal distribution. Hence, all values were transformed to log10 prior to statistical tests. One-way ANOVA was performed with Dunnett's correction for multiple comparison of each substitution at a position to the value collected for WT of that position's data set; statistically significant variants were identified ($P < 0.0001$, significant). Similar analyses were performed using correction for multiple comparisons by controlling for False Discovery Rate (FDR) analysis using the Two-stage step-up method of Benjamini, Krieger, and Yekutieli to identify "discoveries" significantly different from the WT. All parameters were set to Prism's default values. Complete reports of the statistical analyses are included in Supplementary tables 5a-u and 6a-u.

Assigning overall substitution outcomes to each CIITA position

For each position, a set of 10-14 variants (including WT) was used determine overall substitution outcomes. To that end, we used the RheoScale calculator, which uses a modified histogram analysis to quantify how a set of variants samples the empirically observed range of possible functional outcomes.[59] The full range is determined by the full dataset – here, 81 variants and WT CIITA – in addition to the negative control that determines the value for "no detectable activity" (or "no detectable protein"). Using this range, the Rheoscale calculator reports three scores for each position: (1) the fraction of variants equivalent to WT (neutral score); (2) the fraction of variants that showed no activity (toggle score), and (3) the fraction of the accessible functional range that is sampled by the set of variants (rheostat score). For the latter, the observed range is divided into histogram bins; bins occupied by at least one variant are counted as filled; to account for potential error

in the WT measurement, bins farther from WT are given higher weights than the WT-containing and adjacent bins. The number of bins used to analyse each of the measured outcomes were set by the default recommendations of the RheoScale calculator: Cellular concentration - 11 bins; transcriptional activation - 9 bins; combined effects - 6 bins.

DATA AVAILABILITY. The number of replicates, averages and standard deviations for the values shown in Figures 4 and 5 are available in https://github.com/ShwethaSreenivasan/CIITA_Swint-Kruse_Lab_2023.

SUPPLEMENTARY DATA. Supplementary Data are available at Protein Science.

FUNDING. This work was supported by funding from the W.M Keck Foundation (to LSK, AWF, JDF, PES, and ALL) and the National Institutes of Health: GM118589 (to LSK and AWF) and GM147635 (to LSK and JDF).

CONFLICT OF INTEREST. The authors have no conflict of interest to declare.

ACKNOWLEDGEMENTS. We thank Kristen Schwingen, Ibtihal Alghusen, and Halya Fedosyuk (KUMC) for assistance with initial experiments and Braelyn Page (KUMC) and Dr. Sarah Bondos (Texas A&M Health Science Center) for discussions on the manuscript. We also thank Jordan Baker (KUMC) for providing suggestions for data analyses.

REFERENCES

1. Tang Q, Fenton AW (2017) Whole-protein alanine-scanning mutagenesis of allostery: A large percentage of a protein can contribute to mechanism. Hum Mutat 38:1132-1143.
2. Wu T, Swint-Kruse L, Fenton AW (2019) Functional tunability from a distance: Rheostat positions influence allosteric coupling between two distant binding sites. Scientific Reports 9:16957.
3. Bantis LE, Parente DJ, Fenton AW, Swint-Kruse L. "Multiplex" rheostat positions cluster around allosterically critical regions of the lactose repressor protein. (2020). Cold Spring Harbor Laboratory.
4. Campitelli P, Swint-Kruse L, Ozkan S (2020) Substitutions at Non-Conserved Rheostat Positions Modulate Function by Re-Wiring Long-Range, Dynamic Interactions. Molecular biology and evolution 38:201-214.
5. Fenton AW, Page BM, Spellman-Kruse A, Hagenbuch B, Swint-Kruse L (2020) Rheostat positions: A new classification of protein positions relevant to pharmacogenomics. Med Chem Res 29:1133-1146.
6. Ruggiero MJ, Malhotra S, Fenton AW, Swint-Kruse L, Karanicolas J, Hagenbuch B (2021) A clinically relevant polymorphism in the Na(+)/taurocholate cotransporting polypeptide (NTCP) occurs at a rheostat position. J Biol Chem 296:100047.

7. Page BM, Martin TA, Wright CL, Fenton LA, Villar MT, Tang Q, Artigues A, Lamb A, Fenton AW, Swint-Kruse L (2022) Odd one out? Functional tuning of Zymomonas mobilis pyruvate kinase is narrower than its allosteric, human counterpart. Protein Science 31:e4336.
8. Campitelli P, Ozkan SB, Swint-Kruse L, Ross D (2023) Dynamics-based network analysis identifies laci rheostats with high accuracy. Biophysical Journal 122:189a-190a.
9. Swint-Kruse L, Martin TA, Page BM, Wu T, Gerhart PM, Dougherty LL, Tang Q, Parente DJ, Mosier BR, Bantis LE, Fenton AW (2021) Rheostat functional outcomes occur when substitutions are introduced at nonconserved positions that diverge with speciation. Protein Sci 30:1833-1853.
10. Holyoak T, Zhang B, Deng J, Tang Q, Prasannan CB, Fenton AW (2013) Energetic coupling between an oxidizable cysteine and the phosphorylatable N-terminus of human liver pyruvate kinase. Biochemistry 52:466-476.
11. Gill G, Ptashne M (1987) Mutants of GAL4 protein altered in an activation function. Cell 51:121-126.
12. Thukral SK, Morrison ML, Young ET (1991) Alanine scanning site-directed mutagenesis of the zinc fingers of transcription factor ADR1: residues that contact DNA and that transactivate. Proc Natl Acad Sci U S A 88:9188-9192.
13. 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.e446.
14. 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. Molecular Cell 78:890-902.e896.
15. Staller MV, Ramirez E, Kotha SR, Holehouse AS, Pappu RV, Cohen BA (2022) Directed mutational scanning reveals a balance between acidic and hydrophobic residues in strong human activation domains. Cell Syst 13:334-345.e335.
16. Sanborn AL, Yeh BT, Feigerle JT, Hao CV, Townshend RJL, Lieberman Aiden E, Dror RO, Kornberg RD (2021) Simple biochemical features underlie transcriptional activation domain diversity and dynamic, fuzzy binding to Mediator. eLife 10:e68068.
17. Unger T, Nau MM, Segal S, Minna JD (1992) p53: a transdominant regulator of transcription whose function is ablated by mutations occurring in human cancer. Embo j 11:1383-1390.
18. Ma W, Kong Q, Grix M, Mantyla JJ, Yang Y, Benning C, Ohlrogge JB (2015) Deletion of a C–terminal intrinsically disordered region of <scp>WRINKLED</scp> 1 affects its stability and enhances oil accumulation in Arabidopsis. The Plant Journal 83:864-874.
19. Fontes JD, Kanazawa S, Jean D, Peterlin BM (1999) Interactions between the class II transactivator and CREB binding protein increase transcription of major histocompatibility complex class II genes. Mol Cell Biol 19:941-947.
20. 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. Nature Genetics 50:1381-1387.
21. 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, Kathiresan S, O'Rahilly S, Chatterjee K, Florez JC, Mikkelsen T, Savage DB, Altshuler D, Consortium UKMD, Myocardial Infarction Genetics C, Consortium UKCL (2016) Prospective functional classification of all possible missense variants in PPARG. Nature Genetics 48:1570-1575.

22. Karlsson E, Schnatwinkel J, Paissoni C, Andersson E, Herrmann C, Camilloni C, Jemth P (2022) Disordered Regions Flanking the Binding Interface Modulate Affinity between CBP and NCOA. Journal of Molecular Biology 434:167643.
23. Song L, Yao X, Li H, Peng B, Boka AP, Liu Y, Chen G, Liu Z, Mathias KM, Xia L, Li Q, Mir M, Li Y, Li H, Wan L (2022) Hotspot mutations in the structured ENL YEATS domain link aberrant transcriptional condensates and cancer. Molecular Cell 82:4080-4098.e4012.
24. Hu X, Wu X, Berry K, Zhao C, Xin D, Ogurek S, Liu X, Zhang L, Luo Z, Sakabe M, Trubicka J, Łastowska M, Szulzewsky F, Holland EC, Lee L, Hu M, Xin M, Lu QR (2023) Nuclear condensates of YAP fusion proteins alter transcription to drive ependymoma tumourigenesis. Nature Cell Biology.
25. Bulman AL, Hubl ST, Nelson HC (2001) The DNA-binding domain of yeast heat shock transcription factor independently regulates both the N- and C-terminal activation domains. J Biol Chem 276:40254-40262.
26. Kumar MB, Ramadoss P, Reen RK, Vanden Heuvel JP, Perdew GH (2001) The Q-rich subdomain of the human Ah receptor transactivation domain is required for dioxin-mediated transcriptional activity. J Biol Chem 276:42302-42310.
27. Ruas JL, Poellinger L, Pereira T (2002) Functional analysis of hypoxia-inducible factor-1 alpha-mediated transactivation. Identification of amino acid residues critical for transcriptional activation and/or interaction with CREB-binding protein. J Biol Chem 277:38723-38730.
28. Schweiger M-R, You J, Howley PM (2006) Bromodomain Protein 4 Mediates the Papillomavirus E2 Transcriptional Activation Function. Journal of Virology 80:4276-4285.
29. Berlow RB, Dyson HJ, Wright PE (2017) Hypersensitive termination of the hypoxic response by a disordered protein switch. Nature 543:447-451.
30. Brown C, Campos-León K, Strickland M, Williams C, Fairweather V, Brady RL, Crump MP, Gaston K (2010) Protein flexibility directs DNA recognition by the papillomavirus E2 proteins. Nucleic Acids Research 39:2969-2980.
31. Li J, Motlagh HN, Chakuroff C, Thompson EB, Hilser VJ (2012) Thermodynamic dissection of the intrinsically disordered N-terminal domain of human glucocorticoid receptor. J Biol Chem 287:26777-26787.
32. Steimle V, Otten LA, Zufferey M, Mach B (1993) Complementation cloning of an MHC class II transactivator mutated in hereditary MHC class II deficiency (or bare lymphocyte syndrome). Cell 75:135-146.
33. Steimle V, Siegrist CA, Mottet A, Lisowska-Grospierre B, Mach B (1994) Regulation of MHC class II expression by interferon-gamma mediated by the transactivator gene CIITA. Science 265:106-109.
34. Sears C, Olesen J, Rubin D, Finley D, Maniatis T (1998) NF-kappa B p105 processing via the ubiquitin-proteasome pathway. J Biol Chem 273:1409-1419.
35. Scholl T, Mahanta SK, Strominger JL (1997) Specific complex formation between the type II bare lymphocyte syndrome-associated transactivators CIITA and RFX5. Proceedings of the National Academy of Sciences 94:6330-6334.
36. Masternak K, Muhlethaler-Mottet A, Villard J, Zufferey M, Steimle V, Reith W (2000) CIITA is a transcriptional coactivator that is recruited to MHC class II promoters by multiple synergistic interactions with an enhanceosome complex. Genes Dev 14:1156-1166.
37. Meinhardt S, Manley MW, Parente DJ, Swint-Kruse L (2013) Rheostats and Toggle Switches for Modulating Protein Function. PLoS ONE 8:e83502.
38. Miller M, Bromberg Y, Swint-Kruse L (2017) Computational predictors fail to identify amino acid substitution effects at rheostat positions. Scientific Reports 7:41329.
39. Tompa P (2002) Intrinsically unstructured proteins. Trends Biochem Sci 27:527-533.
40. Dyson HJ (2016) Making Sense of Intrinsically Disordered Proteins. Biophysical Journal 110:1013-1016.

41. Emenecker RJ, Griffith D, Holehouse AS (2022) Metapredict V2: An update to metapredict, a fast, accurate, and easy-to-use predictor of consensus disorder and structure. bioRxiv:2022.2006.2006.494887.
42. Xue B, Dunbrack RL, Williams RW, Dunker AK, Uversky VN (2010) PONDR-FIT: a meta-predictor of intrinsically disordered amino acids. Biochim Biophys Acta 1804:996-1010.
43. Hu G, Katuwawala A, Wang K, Wu Z, Ghadermarzi S, Gao J, Kurgan L (2021) flDPnn: Accurate intrinsic disorder prediction with putative propensities of disorder functions. Nature Communications 12.
44. Erdős G, Pajkos M, Dosztányi Z (2021) IUPred3: prediction of protein disorder enhanced with unambiguous experimental annotation and visualization of evolutionary conservation. Nucleic Acids Research 49:W297-W303.
45. Kosol S, Contreras-Martos S, Cedeño C, Tompa P (2013) Structural Characterization of Intrinsically Disordered Proteins by NMR Spectroscopy. Molecules 18:10802-10828.
46. Robustelli P, Piana S, Shaw DE (2020) Mechanism of Coupled Folding-upon-Binding of an Intrinsically Disordered Protein. J Am Chem Soc 142:11092-11101.
47. Lawrence CW, Kumar S, Noid WG, Showalter SA (2014) Role of Ordered Proteins in the Folding-Upon-Binding of Intrinsically Disordered Proteins. The Journal of Physical Chemistry Letters 5:833-838.
48. Dogan J, Mu X, Engström Å, Jemth P (2013) The transition state structure for coupled binding and folding of disordered protein domains. Scientific Reports 3.
49. Joosten RP, te Beek TAH, Krieger E, Hekkelman ML, Hooft RWW, Schneider R, Sander C, Vriend G (2011) A series of PDB related databases for everyday needs. Nucleic acids research 39:D411-419.
50. Kabsch W, Sander C (1983) DICTIONARY OF PROTEIN SECONDARY STRUCTURE - PATTERN-RECOGNITION OF HYDROGEN-BONDED AND GEOMETRICAL FEATURES. Biopolymers 22:2577-2637.
51. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, Bridgland A, Meyer C, Kohl SAA, Ballard AJ, Cowie A, Romera-Paredes B, Nikolov S, Jain R, Adler J, Back T, Petersen S, Reiman D, Clancy E, Zielinski M, Steinegger M, Pacholska M, Berghammer T, Bodenstein S, Silver D, Vinyals O, Senior AW, Kavukcuoglu K, Kohli P, Hassabis D (2021) Highly accurate protein structure prediction with AlphaFold. Nature 596:583-589.
52. Pattenden SG, Klose R, Karaskov E, Bremner R (2002) Interferon-gamma-induced chromatin remodeling at the CIITA locus is BRG1 dependent. Embo j 21:1978-1986.
53. Devaiah BN, Singer DS (2013) CIITA and Its Dual Roles in MHC Gene Transcription. Front Immunol 4:476.
54. Zika E, Ting JP (2005) Epigenetic control of MHC-II: interplay between CIITA and histone-modifying enzymes. Curr Opin Immunol 17:58-64.
55. León Machado JA, Steimle V (2021) The MHC Class II Transactivator CIITA: Not (Quite) the Odd-One-Out Anymore among NLR Proteins. International Journal of Molecular Sciences 22:1074.
56. Kumar M, Michael S, Alvarado-Valverde J, Mészáros B, Sámano‐Sánchez H, Zeke A, Dobson L, Lazar T, Örd M, Nagpal A, Farahi N, Käser M, Kraleti R, Davey Norman E, Pancsa R, Chemes Lucía B, Gibson Toby J (2021) The Eukaryotic Linear Motif resource: 2022 release. Nucleic Acids Research 50:D497-D508.
57. Martin TA, Wu T, Tang Q, Dougherty LL, Parente DJ, Swint-Kruse L, Fenton AW (2020) Identification of biochemically neutral positions in liver pyruvate kinase. Proteins 88:1340-1350.
58. Tang Q, Alontaga AY, Holyoak T, Fenton AW (2017) Exploring the limits of the usefulness of mutagenesis in studies of allosteric mechanisms. Human Mutation 38:1144-1154.

59. Hodges AM, Fenton AW, Dougherty LL, Overholt AC, Swint-Kruse L (2018) RheoScale: A tool to aggregate and quantify experimentally determined substitution outcomes for multiple variants at individual protein positions. Hum Mutat 39:1814-1826.
60. Hake SB, Masternak K, Kammerbauer C, Janzen C, Reith W, Steimle V (2000) CIITA leucine-rich repeats control nuclear localization, in vivo recruitment to the major histocompatibility complex (MHC) class II enhanceosome, and MHC class II gene transactivation. Mol Cell Biol 20:7716-7725.
61. Zhou J, Oldfield CJ, Yan W, Shen B, Dunker AK (2019) Intrinsically disordered domains: Sequence ➔ disorder ➔ function relationships. Protein Science.
62. Ashkenazy H, Abadi S, Martz E, Chay O, Mayrose I, Pupko T, Ben-Tal N (2016) ConSurf 2016: an improved methodology to estimate and visualize evolutionary conservation in macromolecules. Nucleic Acids Res 44:W344-350.
63. Meinhardt S, Manley MW, Jr., Becker NA, Hessman JA, Maher LJ, 3rd, Swint-Kruse L (2012) Novel insights from hybrid LacI/GalR proteins: family-wide functional attributes and biologically significant variation in transcription repression. Nucleic Acids Res 40:11139-11154.
64. Bendell CJ, Liu S, Aumentado-Armstrong T, Istrate B, Cernek PT, Khan S, Picioreanu S, Zhao M, Murgita RA (2014) Transient protein-protein interface prediction: datasets, features, algorithms, and the RAD-T predictor. BMC Bioinformatics 15:82.
65. Pace CN, Scholtz JM (1998) A helix propensity scale based on experimental studies of peptides and proteins. Biophys J 75:422-427.
66. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, Collins RL, Laricchia KM, Ganna A, Birnbaum DP, Gauthier LD, Brand H, Solomonson M, Watts NA, Rhodes D, Singer-Berk M, England EM, Seaby EG, Kosmicki JA, Walters RK, Tashman K, Farjoun Y, Banks E, Poterba T, Wang A, Seed C, Whiffin N, Chong JX, Samocha KE, Pierce-Hoffman E, Zappala Z, O'Donnell-Luria AH, Minikel EV, Weisburd B, Lek M, Ware JS, Vittal C, Armean IM, Bergelson L, Cibulskis K, Connolly KM, Covarrubias M, Donnelly S, Ferriera S, Gabriel S, Gentry J, Gupta N, Jeandet T, Kaplan D, Llanwarne C, Munshi R, Novod S, Petrillo N, Roazen D, Ruano-Rubio V, Saltzman A, Schleicher M, Soto J, Tibbetts K, Tolonen C, Wade G, Talkowski ME, Aguilar Salinas CA, Ahmad T, Albert CM, Ardissino D, Atzmon G, Barnard J, Beaugerie L, Benjamin EJ, Boehnke M, Bonnycastle LL, Bottinger EP, Bowden DW, Bown MJ, Chambers JC, Chan JC, Chasman D, Cho J, Chung MK, Cohen B, Correa A, Dabelea D, Daly MJ, Darbar D, Duggirala R, Dupuis J, Ellinor PT, Elosua R, Erdmann J, Esko T, Färkkilä M, Florez J, Franke A, Getz G, Glaser B, Glatt SJ, Goldstein D, Gonzalez C, Groop L, Haiman C, Hanis C, Harms M, Hiltunen M, Holi MM, Hultman CM, Kallela M, Kaprio J, Kathiresan S, Kim B-J, Kim YJ, Kirov G, Kooner J, Koskinen S, Krumholz HM, Kugathasan S, Kwak SH, Laakso M, Lehtimäki T, Loos RJF, Lubitz SA, Ma RCW, Macarthur DG, Marrugat J, Mattila KM, McCarroll S, McCarthy MI, McGovern D, McPherson R, Meigs JB, Melander O, Metspalu A, Neale BM, Nilsson PM, O'Donovan MC, Ongur D, Orozco L, Owen MJ, Palmer CNA, Palotie A, Park KS, Pato C, Pulver AE, Rahman N, Remes AM, Rioux JD, Ripatti S, Roden DM, Saleheen D, Salomaa V, Samani NJ, Scharf J, Schunkert H, Shoemaker MB, Sklar P, Soininen H, Sokol H, Spector T, Sullivan PF, Suvisaari J, Tai ES, Teo YY, Tiinamaija T, Tsuang M, Turner D, Tusie-Luna T, Vartiainen E, Vawter MP, Ware JS, Watkins H, Weersma RK, Wessman M, Wilson JG, Xavier RJ, Neale BM, Daly MJ, Macarthur DG (2020) The mutational constraint spectrum quantified from variation in 141,456 humans. Nature 581:434-443.
67. Walkiewicz K, Benitez Cardenas AS, Sun C, Bacorn C, Saxer G, Shamoo Y (2012) Small changes in enzyme function can lead to surprisingly large fitness effects during adaptive evolution of antibiotic resistance. Proceedings of the National Academy of Sciences 109:21408-21413.
68. Swint-Kruse L (2016) Using Evolution to Guide Protein Engineering: The Devil IS in the Details. Biophysical Journal 111:10-18.
69. Bromberg Y, Kahn PC, Rost B (2013) Neutral and weakly nonneutral sequence variants may define individuality. Proceedings of the National Academy of Sciences 110:14255-14260.

70. 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 Biology 20.
71. Romero P, Obradovic Z, Dunker AK (2004) Natively disordered proteins: functions and predictions. Appl Bioinformatics 3:105-113.
72. Dayhoff GW, 2nd, Uversky VN (2022) Rapid prediction and analysis of protein intrinsic disorder. Protein Sci 31:e4496.
73. Finn RD, Clements J, Eddy SR (2011) HMMER web server: interactive sequence similarity searching. Nucleic Acids Res 39:W29-37.
74. Katoh K, Standley DM (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol 30:772-780.
75. Sayers EW, Bolton EE, Brister JR, Canese K, Chan J, Comeau DC, Connor R, Funk K, Kelly C, Kim S, Madej T, Marchler-Bauer A, Lanczycki C, Lathrop S, Lu Z, Thibaud-Nissen F, Murphy T, Phan L, Skripchenko Y, Tse T, Wang J, Williams R, Trawick BW, Pruitt KD, Sherry ST (2022) Database resources of the national center for biotechnology information. Nucleic Acids Res 50:D20-d26.
76. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Söding J, Thompson JD, Higgins DG (2011) Fast, scalable generation of high‐quality protein multiple sequence alignments using Clustal Omega. Molecular Systems Biology 7:539.
77. Chakravarty D, Sreenivasan S, Swint-Kruse L, Porter LL (2022) Identification of a covert evolutionary pathway between two protein folds. BioRxiv https://doi.org/10.1101/2022.12.08.519646.
78. Hanwell MD, Curtis DE, Lonie DC, Vandermeersch T, Zurek E, Hutchison GR (2012) Avogadro: an advanced semantic chemical editor, visualization, and analysis platform. Journal of Cheminformatics 4:17.
79. Abraham MJ, van der Spoel D, Lindahl E, Hess B, GROMACS, development team. GROMACS User Manual version 2016. (2018).
80. Ploetz EA, Karunaweera S, Bentenitis N, Chen F, Dai S, Gee MB, Jiao YF, Kang YS, Kariyawasam NL, Naleem N, Weerasinghe S, Smith PE (2021) Kirkwood-Buff-Derived Force Field for Peptides and Proteins: Philosophy and Development of KBFF20. Journal of Chemical Theory and Computation 17:2964-2990.
81. Berendsen HJC, Grigera JR, Straatsma TP (1987) THE MISSING TERM IN EFFECTIVE PAIR POTENTIALS. Journal of Physical Chemistry 91:6269-6271.
82. Huang J, Rauscher S, Nawrocki G, Ran T, Feig M, de Groot BL, Grubmueller H, MacKerell AD, Jr. (2017) CHARMM36m: an improved force field for folded and intrinsically disordered proteins. Nature Methods 14:71-73.
83. Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, Klein ML (1983) COMPARISON OF SIMPLE POTENTIAL FUNCTIONS FOR SIMULATING LIQUID WATER. Journal of Chemical Physics 79:926-935.
84. Lindorff-Larsen K, Piana S, Palmo K, Maragakis P, Klepeis JL, Dror RO, shaw DE (2010) Improved side-chain torsion potentials for the Amber ff99SB protein force field. Proteins-Structure Function and Bioinformatics 78:1950-1958.
85. Piana S, Donchev AG, Robustelli P, Shaw DE (2015) Water Dispersion Interactions Strongly Influence Simulated Structural Properties of Disordered Protein States. Journal of Physical Chemistry B 119:5113-5123.
86. Hockney RW. The potential calculation and some applications. (1970).
87. Darden T, York D, Pedersen L (1993) PARTICLE MESH EWALD - AN N.LOG(N) METHOD FOR EWALD SUMS IN LARGE SYSTEMS. Journal of Chemical Physics 98:10089-10092.
88. Nose S (1984) A MOLECULAR-DYNAMICS METHOD FOR SIMULATIONS IN THE CANONICAL ENSEMBLE. Molecular Physics 52:255-268.

89. Bussi G, Donadio D, Parrinello M (2007) Canonical sampling through velocity rescaling. Journal of Chemical Physics 126.
90. Hess B, Bekker H, Berendsen HJC, Fraaije J (1997) LINCS: A linear constraint solver for molecular simulations. Journal of Computational Chemistry 18:1463-1472.
91. Miyamoto S, Kollman PA (1992) SETTLE - AN ANALYTICAL VERSION OF THE SHAKE AND RATTLE ALGORITHM FOR RIGID WATER MODELS. Journal of Computational Chemistry 13:952-962.
92. Berendsen HJC, Postma JPM, Vangunsteren WF, Dinola A, Haak JR (1984) MOLECULAR-DYNAMICS WITH COUPLING TO AN EXTERNAL BATH. Journal of Chemical Physics 81:3684-3690.
93. Mudhasani R, Fontes JD (2002) The class II transactivator requires brahma-related gene 1 to activate transcription of major histocompatibility complex class II genes. Mol Cell Biol 22:5019-5026.

FIGURES LEGENDS

Figure 1. Sequence analyses of the N-terminal transcriptional activation domain of CIITA. The probabilities for intrinsic disordered were predicted for CIITA positions 1-161 by PONDR. Predictions for the full-length protein and predictions by other algorithms are in Supplementary figure 2. Values >0.5 indicate positions with high probability of disorder (e.g., positions 62, 65, 70, 91; solid vertical red lines) and values <0.5 indicate positions with low probability of disorder (e.g., positions 74, 77, 78; dashed vertical blue lines). Individual values for these positions are listed in Supplementary table 1.

Figure 2. Circular dichroism (CD) of purified CIITA 1-210. The CD spectra of the purified CIITA 1-210 fragment was measured in the absence (black) and presence of 1 M TMAO (red) at room temperature. Samples contained 4 µM CIITA in 50 mM sodium phosphate buffer, pH 8.0. Results shown are the average of 100-150 replicate scans with background subtraction.

Figure 3. Simulated properties of the 56-94 residue fragment of CIITA. (a) The radius of gyration and end-to-end probability distributions of CIITA 56-94 from three independent simulations (Runs 1-3) using the KBFF20 protein force field predict little-to-no persistent structure. Similar plots for simulations using the CHARMM and AMBER force fields are shown in Supplementary figure 3. (b) Helix residue probability distributions averaged over the three simulations for each of three different force fields.

Figure 4. Cellular concentrations for position-wise variants of CIITA. All data were normalized to a WT value that was set to 1. (a) Reference range for cellular concentrations: The background signal ("No CIITA") was estimated as ~1% of the WT CIITA concentration (dotted

line at 0.01); the variant with the highest cellular concentration of CIITA was set as "Max Observed" (dotted line at 1.85). All values for the biological and technical replicates of WT are shown as dots; the average of these values was normalized to 1 and the standard error of means is shown with error bars. (b)-(h) The normalized cellular concentrations of CIITA variants are shown for positions with low (b) N74, (c) Q77 (d) F78 and high disorder probabilities (e) I62, (f) Y65, (g) T70, (h) T91. Individual measurements are shown as black dots, and the standard error of the means are indicated with error bars. The variant corresponding to WT is circled on the X-axis legend, and the normalized WT value of 1 is indicated with a dashed horizontal line; the WT samples shown on (b)-(h) were measured in parallel with each position's variants. Bars are coloured according to amino acid type: acidic (D, E; red, checked), basic (R, H, K; blue, slanted lines), polar uncharged (S, T, N, C, Q; yellow, vertical lines), hydrophobic (M, F, W, Y, A, V, I, L; green, horizontal lines), and structure breakers (P, G; grey, solid). Substitution variants with protein levels statistically different from WT were identified with (i) one-way ANOVAs with Dunnett's correction (****$p < 0.0001$, ***$p<0.001$, **$p<0.01$, *$p<0.1$), and (ii) false discovery rate analyses (underlined variants on the X-axis legend).

Figure 5. Transcriptional activation for position-wise variants of CIITA. All data were normalized to a WT value that was set to 1. (a) Reference range for transcriptional activation: At the lower end of the range, the transcriptional activation level for a hypothetical "Dead" CIITA was determined from a control sample that lacked CIITA (dotted line at 0.08; calculations in Supplementary table 7). The upper end of the range was defined by the variant with the greatest transcriptional activity, Y65R ("Max Observed", dotted line at 5.5). All values for the biological and technical replicates of WT are shown with dots; the average of these values was normalized to 1 and the standard error of means is shown with error bars. (b)-(h) Normalized CIITA transcriptional activation levels are shown for each substitution variant at positions with low (b) N74, (c) Q77, (d) F78 and high disorder probabilities (e) I62, (f) Y65, (g) T70, (h) T91. Individual measurements are shown with black dots, and standard error of the means are indicated with error bars. The variant corresponding to WT is circled on the X-axis legend, and the normalized WT value of 1 is indicated with a dashed line; the WT samples shown on (b)-(h) were measured in parallel with each position's variants. Bars are coloured according to amino acid type: acidic (D, E; red, checked), basic (R, H, K; blue,

slanted lines), polar uncharged (S, T, N, C, Q; yellow, vertical lines), hydrophobic (M, F, W, Y, A, V, I, L; green, horizontal lines), and structure breakers (P, G; grey, solid). Variants with activity statistically different from WT were identified with (i) one-way ANOVAs with Dunnett's correction (****$p < 0.0001$, ***$p<0.001$, **$p<0.01$, *$p<0.1$), and (ii) false discovery rate analyses (underlined variants in the X-axis legend).

Figure 6. Correlation of transcriptional activation with structural features. Of the 28 correlations of transcriptional activation with structural features, three showed Spearman coefficients >0.5. (a) At position 78, transcriptional activation positively correlated with the amino acid side chain size obtained from the relative, solvent-accessible amino acid side chain surface area ($Å^2$) in a G-X-G tripeptide, where X is the amino acid. (b) At position 62, transcriptional activation negatively correlated with amino acid helical propensities obtained from host-guest peptide studies (reported in units of kcal/mol); lower values correspond to higher helical propensities. Proline's measured propensity was 3.16 kcal/mol; to aid visual interpretation of these plots, this value was set to 1.2 kcal/mol; this change did not alter the Spearman correlation coefficient, which compares rank order of two parameters. (c) At position 77, transcriptional activation positively correlated with disorder probabilities predicted by PONDR.

Figure 7. Overall tunability of CIITA's transcriptional activation from 82 variants. Numbers on the X-axis correspond to log(upper limits) for transcriptional activation in a histogram analysis for the observed values. The overall range of observed change is >30-fold. The green dot indicates the bin containing WT. The magenta dot at the far left indicates the level of a hypothetical "dead" variant (see Supplementary table 7 for these calculations).

Table 1. Position-wise sequence properties and experimental outcomes for CIITA variants.

Position	Disorder probability[a]	ConSurf score[b]		No. of variants[c]	Cellular concentration			Transcriptional activation			*In vivo* function		
		Manual[d]	Default[e]		N[f]	R[g]	T[h]	N	R	T	N	R	T
62	**0.66**	2	4	13	0.58	0.42	0.00	0.25	**0.70**	0.00	0.33	**0.55**	0.00
65	**0.86**	3	4	10	0.22	0.42	0.00	0.44	**0.55**	0.00	0.00	**0.70**	0.00
70	**0.53**	4	5	13	0.42	0.19	0.00	0.42	0.40	0.00	0.50	0.25	0.00
74	0.22	8	9	14	0.62	0.12	0.00	0.31	0.40	0.00	0.54	0.40	0.00
77	0.15	9	9	12	0.09	0.31	0.00	0.18	**0.60**	0.09	0.64	**0.55**	0.00
78	0.16	5	5	14	0.54	0.31	0.00	0.23	**0.55**	0.00	0.54	0.25	0.00
91	**0.69**	7	8	12	**0.82**	0.12	0.00	0.45	0.25	0.00	0.54	0.15	0.00

[a] Values >0.5 (bold text) indicate high disorder probabilities; values <0.5 indicate low disorder probabilities
[b] Consurf scores range from 1 (least conserved) to 9 (most conserved)
[c] The number of variants for each position includes WT
[d] ConSurf scores averaged from 19 random sequence subsets extracted from "CIITA_manual_MSA"
[e] ConSurf scores for "CIITA_default_MSA
[f] Positions with RheoScale neutral scores >0.7 (bold text) are defined as "neutral" for that parameter; a neutral position must have neutral RheoScale scores >0.7 for all parameters measured (12)
[g] Positions with RheoScale rheostat scores >0.5 (bold text) are defined as "rheostat" positions (18)
[h] Positions with RheoScale toggle scores >0.64 (bold text) are defined as "toggle" positions (18)

Figure 1. Sequence analyses of the N-terminal transcriptional activation domain of CIITA. The probabilities for intrinsic disordered were predicted for CIITA positions 1-161 by PONDR. Predictions for the full-length protein and predictions by other algorithms are in Supplementary figure 2. Values >0.5 indicate positions with high probability of disorder (e.g., positions 62, 65, 70, 91; solid vertical red lines) and values <0.5 indicate positions with low probability of disorder (e.g., positions 74, 77, 78; dashed vertical blue lines). Individual values for these positions are listed in Supplementary table 1.

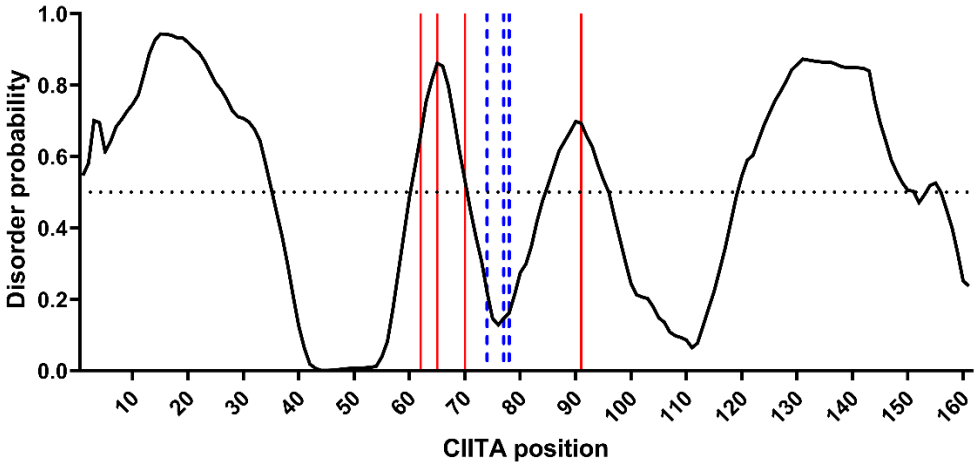


Figure 2. Circular dichroism (CD) of purified CIITA 1-210. The CD spectra of the purified CIITA 1-210 fragment was measured in the absence (black) and presence of 1 M TMAO (red) at room temperature. Samples contained 4 μM CIITA in 50 mM sodium phosphate buffer, pH 8.0. Results shown are the average of 100-150 replicate scans with background subtraction.

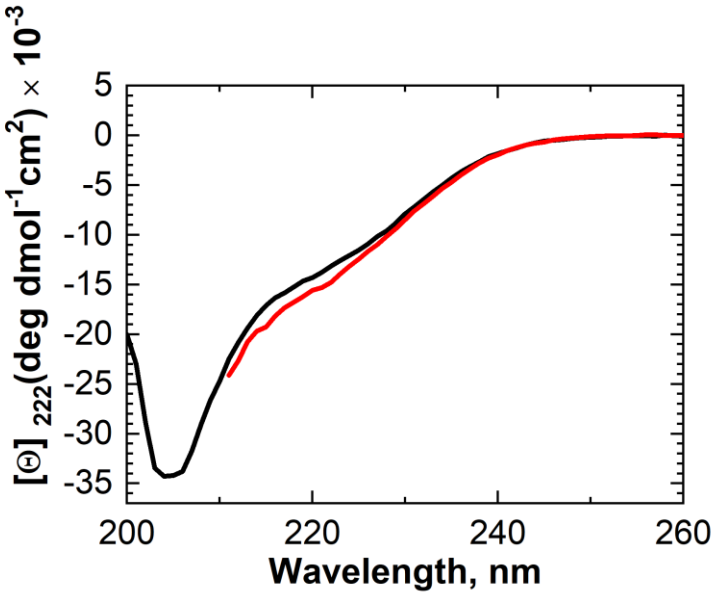


Figure 3. Simulated properties of the 56-94 residue fragment of CIITA. (a) The radius of gyration and end-to-end probability distributions of CIITA 56-94 from three independent simulations (Runs 1-3) using the KBFF20 protein force field predict little-to-no persistent structure. Similar plots for simulations using the CHARMM and AMBER force fields are shown in Supplementary figure 3. (b) Helix residue probability distributions averaged over the three simulations for each of three different force fields.

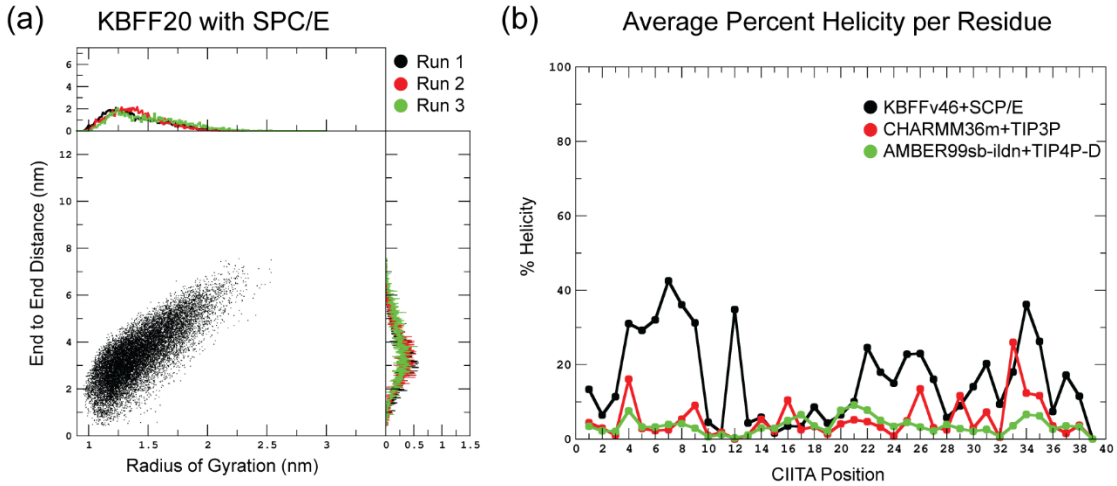


Figure 4. Cellular concentrations for position-wise variants of CIITA. All data were normalized to a WT value that was set to 1. (a) Reference range for cellular concentrations: The background signal ("No CIITA") was estimated as ~1% of the WT CIITA concentration (dotted line at 0.01); the variant with the highest cellular concentration of CIITA was set as "Max Observed" (dotted line at 1.85). All values for the biological and technical replicates of WT are shown as dots; the average of these values was normalized to 1 and the standard error of means is shown with error bars. (b)-(h) The normalized cellular concentrations of CIITA variants are shown for positions with low (b) N74, (c) Q77 (d) F78 and high disorder probabilities (e) I62, (f) Y65, (g) T70, (h) T91. Individual measurements are shown as black dots, and the standard error of the means are indicated with error bars. The variant corresponding to WT is circled on the X-axis legend, and the normalized WT value of 1 is indicated with a dashed horizontal line; the WT samples shown on (b)-(h) were measured in parallel with each position's variants. Bars are coloured according to amino acid type: acidic (D, E; red, checked), basic (R, H, K; blue, slanted lines), polar uncharged (S, T, N, C, Q; yellow, vertical lines), hydrophobic (M, F, W, Y, A, V, I, L; green, horizontal lines), and structure breakers (P, G; grey, solid). Substitution variants with protein levels statistically different from WT were identified with (i) one-way ANOVAs with Dunnett's correction (****p <0.0001, ***p<0.001, **p<0.01, *p<0.1), and (ii) false discovery rate analyses

(underlined variants on the X-axis legend).

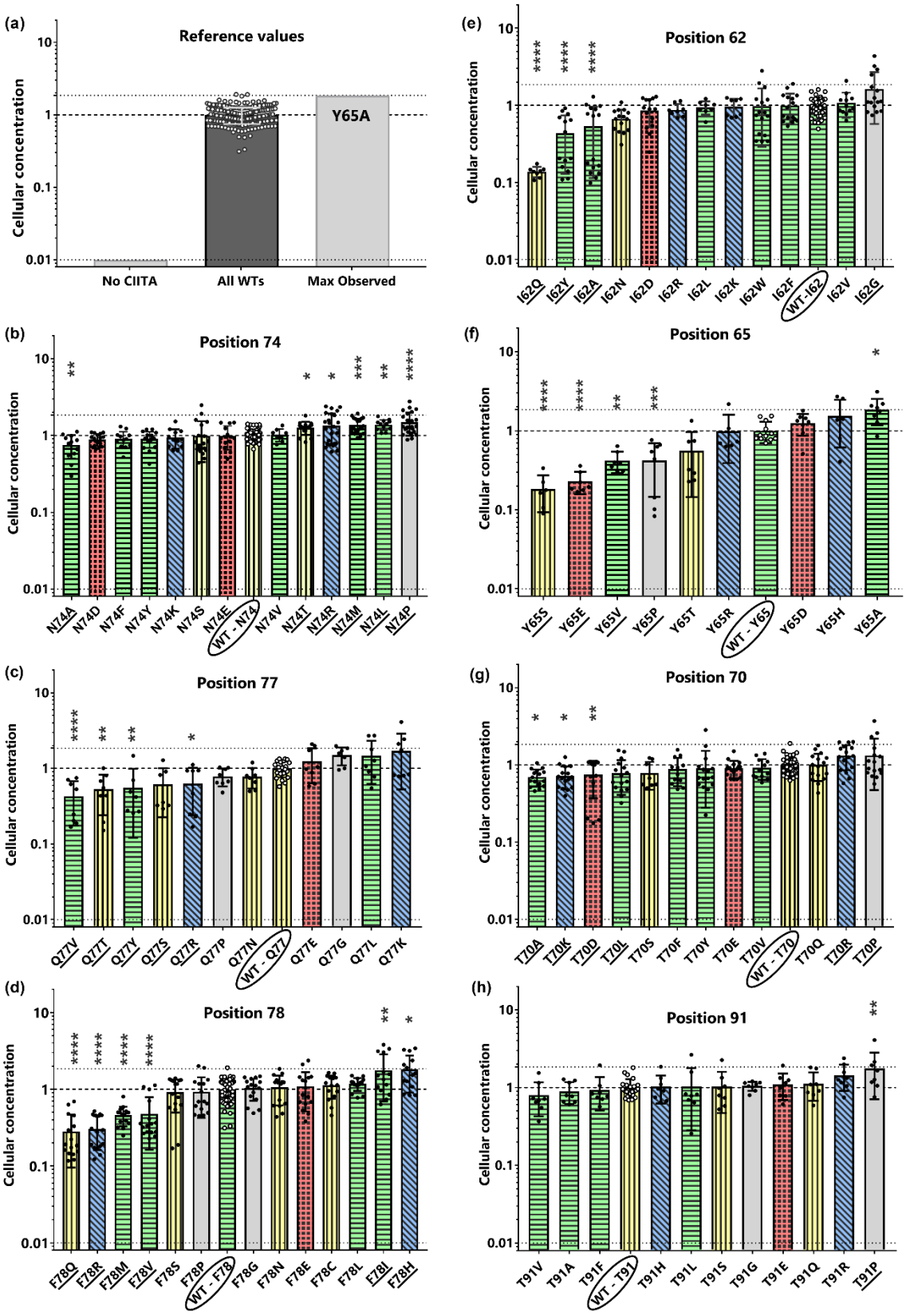

(a)
Reference values
Cellular concentration
Y65A
No CIITA
All WTs
Max Observed
(b)
Position 74
N74A
N74D
N74F
N74Y
N74K
N74S
N74E
WT - N74
N74V
N74T
N74R
N74M
N74L
N74P
(c)
Position 77
Q77V
Q77T
Q77Y
Q77S
Q77R
Q77P
Q77N
WT - Q77
Q77E
Q77G
Q77L
Q77K
(d)
Position 78
F78Q
F78R
F78M
F78V
F78S
F78P
WT - F78
F78G
F78N
F78E
F78C
F78L
F78I
F78H
(e)
Position 62
I62Q
I62Y
I62A
I62N
I62D
I62R
I62L
I62K
I62W
I62F
WT-I62
I62V
I62G
(f)
Position 65
Y65S
Y65E
Y65V
Y65P
Y65T
Y65R
WT - Y65
Y65D
Y65H
Y65A
(g)
Position 70
T70A
T70K
T70D
T70L
T70S
T70F
T70Y
T70E
T70V
WT - T70
T70Q
T70R
T70P
(h)
Position 91
T91V
T91A
T91F
WT - T91
T91H
T91L
T91S
T91G
T91E
T91Q
T91R
T91P

Figure 5. Transcriptional activation for position-wise variants of CIITA. All data were normalized to a WT value that was set to 1. (a) Reference range for transcriptional activation: At the lower end of the range, the transcriptional activation level for a hypothetical "Dead" CIITA was determined from a control sample that lacked CIITA (dotted line at 0.08; calculations in Supplementary table 7). The upper end of the range was defined by the variant with the greatest transcriptional activity, Y65R ("Max Observed", dotted line at 5.5). All values for the biological and technical replicates of WT are shown with dots; the average of these values was normalized to 1 and the standard error of means is shown with error bars. (b)-(h) Normalized CIITA transcriptional activation levels are shown for each substitution variant at positions with low (b) N74, (c) Q77, (d) F78 and high disorder probabilities (e) I62, (f) Y65, (g) T70, (h) T91. Individual measurements are shown with black dots, and standard error of the means are indicated with error bars. The variant corresponding to WT is circled on the X-axis legend, and the normalized WT value of 1 is indicated with a dashed line; the WT samples shown on (b)-(h) were measured in parallel with each position's variants. Bars are coloured according to amino acid type: acidic (D, E; red, checked), basic (R, H, K; blue, slanted lines), polar uncharged (S, T, N, C, Q; yellow, vertical lines), hydrophobic (M, F, W, Y, A, V, I, L; green, horizontal lines), and structure breakers (P, G; grey, solid). Variants with activity statistically different from WT were identified with (i) one-way ANOVAs with Dunnett's correction (****$p < 0.0001$, ***$p < 0.001$, **$p < 0.01$, *$p < 0.1$), and (ii) false discovery rate analyses (underlined variants in the X-axis legend).

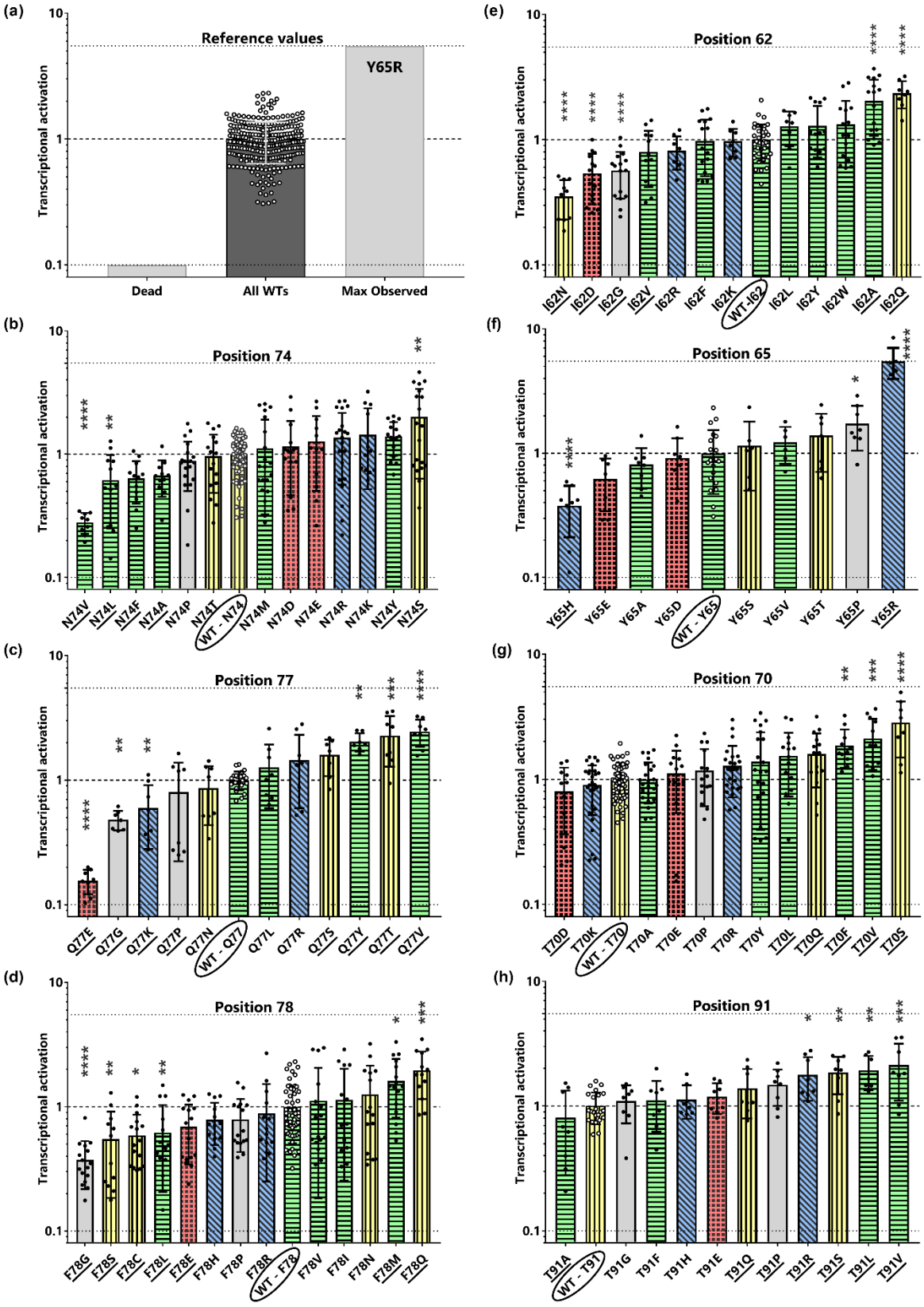
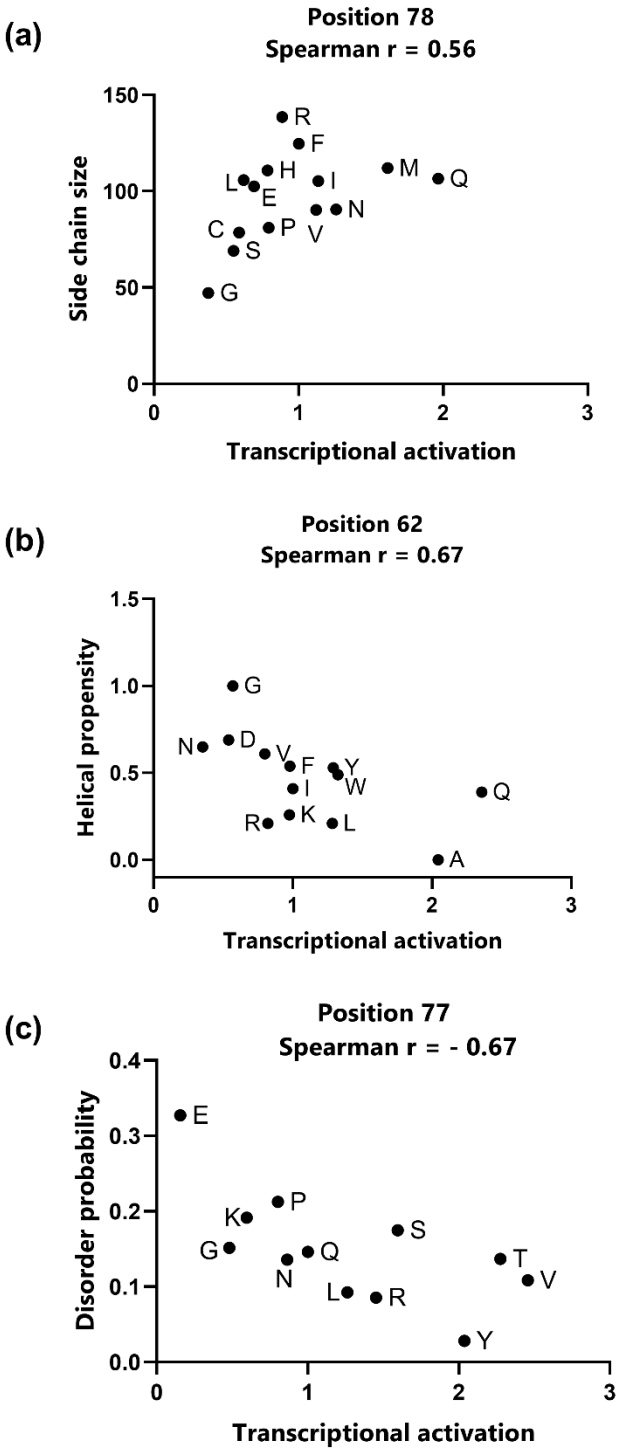

(a)
Reference values
Y65R
Transcriptional activation
10
1
0.1
Dead
All WTs
Max Observed
(b)
Position 74
N74V
N74L
N74F
N74A
N74P
N74T
WT - N74
N74M
N74D
N74E
N74R
N74K
N74Y
N74S
(c)
Position 77
Q77E
Q77G
Q77K
Q77P
Q77N
WT - Q77
Q77L
Q77R
Q77S
Q77Y
Q77T
Q77V
(d)
Position 78
F78G
F78S
F78C
F78L
F78E
F78H
F78P
F78R
WT - F78
F78V
F78I
F78N
F78M
F78Q
(e)
Position 62
I62N
I62D
I62G
I62V
I62R
I62F
I62K
WT-I62
I62L
I62Y
I62W
I62A
I62Q
(f)
Position 65
Y65H
Y65E
Y65A
Y65D
WT - Y65
Y65S
Y65V
Y65T
Y65P
Y65R
(g)
Position 70
T70D
T70K
WT - T70
T70A
T70E
T70P
T70R
T70Y
T70L
T70Q
T70F
T70V
T70S
(h)
Position 91
T91A
WT - T91
T91G
T91F
T91H
T91E
T91Q
T91P
T91R
T91S
T91L
T91V

Figure 6. Correlation of transcriptional activation with structural features. Of the 28 correlations of transcriptional activation with structural features, three showed Spearman coefficients >0.5. (a) At position 78, transcriptional activation positively correlated with the amino acid side chain size obtained from the relative, solvent-accessible amino acid side chain surface area ($Å^2$) in a G-X-G tripeptide, where X is the amino acid. (b) At position 62, transcriptional activation negatively correlated with amino acid helical propensities obtained from host-guest peptide studies (reported in units of kcal/mol); lower values correspond to higher helical propensities. Proline's measured propensity was 3.16 kcal/mol; to aid visual interpretation of these plots, this value was set to 1.2 kcal/mol; this change did not alter the Spearman correlation coefficient, which compares rank order of two parameters. (c) At position 77, transcriptional activation positively correlated with disorder probabilities predicted by PONDR.


(a)
Position 78
Spearman r = 0.56
Side chain size
Transcriptional activation
(b)
Position 62
Spearman r = 0.67
Helical propensity
Transcriptional activation
(c)
Position 77
Spearman r = - 0.67
Disorder probability
Transcriptional activation

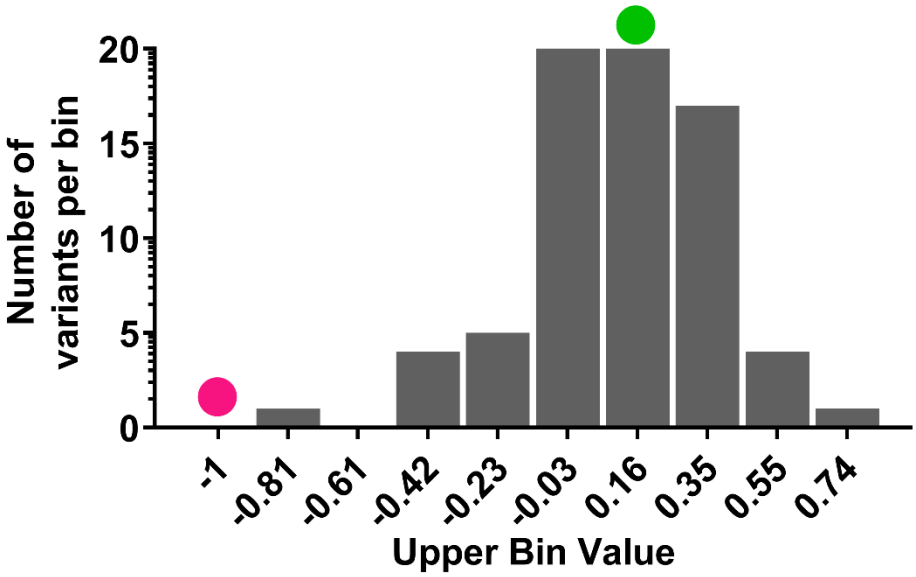



Figure 7. Overall tunability of CIITA's transcriptional activation from 82 variants. Numbers on the X-axis correspond to log(upper limits) for transcriptional activation in a histogram analysis for the observed values. The overall range of observed change is >30-fold. The green dot indicates the bin containing WT. The magenta dot at the far left indicates the level of a hypothetical "dead" variant (see Supplementary table 7 for these calculations).