

## Supplementary Information

### Supplementary Methods

#### Computational Techniques

##### *Overview of Anti-bZIP Design Using CLASSY*

CLASSY is a computational design procedure for optimizing the stability of a particular structural state as a function of sequence, under an arbitrary number of constraints. It is compatible with many types of potential functions. Any linear analytical function of sequence variables can be constrained; examples include energy gaps towards other structures, or properties such as amino-acid composition or hydrophobicity.

CLASSY is based on two components: cluster expansion (CE) and integer linear programming (ILP) optimization. CE provides a way to express the energy of a sequence adopting a particular backbone structure as an algebraic function of the sequence itself<sup>28</sup>. The formal basis of the technique is briefly described in the next section, but two properties of a cluster expansion are important for CLASSY: (1) it makes the evaluation of sequence energies many orders of magnitude faster than with direct structural methods, and (2) its simple functional form renders a new set of computational approaches applicable to protein design. We used CE in conjunction with ILP as a way to incorporate information about undesired states into design calculations.

##### *Theory of Cluster Expansion*

We have previously shown that the conformational energy of a protein sequence in a specified fold, defined numerically using structural calculations and optimization, can be expressed as a direct function of sequence using the method of cluster expansion<sup>28, 29</sup>. For completeness, we briefly describe this method here. Let  $E_{\min}(\vec{\sigma})$  be the energy of sequence  $\vec{\sigma}$  in a given backbone fold (subscript *min* stands for minimization over side-chain degrees of freedom). Let  $\vec{\sigma} = \{\sigma_1, \dots, \sigma_N\}$ , where  $\sigma_i$  is a discrete variable representing the amino acid at the  $i$ -th position of the sequence. For simplicity, and without loss of generality, assume that in our design problem there are  $M$  amino-acid possibilities at each position and  $\sigma_i$  can take on values from 0 to  $M-1$ . We can then express  $E_{\min}(\vec{\sigma})$  as a cluster expansion of the form:

$$E_{\min}(\vec{\sigma}) = J_o + \sum_{si=1}^N \sum_{i=1}^{M-1} J_{si}^i \cdot \varphi(si, i) + \sum_{si=1}^{N-1} \sum_{sj=si+1}^N \sum_{i=1}^{M-1} \sum_{j=1}^{M-1} J_{si,sj}^{i,j} \cdot \varphi(si, i) \cdot \varphi(sj, j) + \dots,$$

where  $\varphi(si, i)$  is a binary function that evaluates as 1 if site  $si$  is occupied with amino-acid  $i$  and zero otherwise. The summations are over sites and amino-acid identities. A collection of sites is referred to as a

cluster, and a cluster populated by a given set of amino acids is a cluster function (CF). Terms  $J$  are the effective contributions of each cluster function to the overall energy (effective cluster interactions, or ECI). The three terms shown correspond to the constant, point and pair cluster-function contributions. If the expansion is written out in its entirety (i.e. up to the  $N$ -tuple cluster functions), then by virtue of having exactly the same number of ECI as possible sequences ( $M^N$ ), it is exact. If the expansion is truncated at a given point, an approximation of  $E_{\min}$  can be derived by fitting the ECI to minimize the error between CE-estimated energies and structure-derived energies for a training set of sequence-energy pairs. Once this procedure is carried out, the process of estimating the energy of a sequence adopting the specified structure is made many orders of magnitude more efficient<sup>28</sup>.

### ***bZIP Models***

To model parallel dimeric coiled coils, we employed two variants of the energy function HP/S/C that was previously shown to perform well in predicting human bZIP interaction specificity<sup>26</sup>. This function evaluates the relative stability of coiled-coil dimers primarily as a function of the amino acids at **a**, **d**, **e** and **g** positions, based on predicted structures of coiled-coil complexes. One of the key features of model HP/S/C is that core **a-a'** and **d-d'** terms derived from structure-based calculations are replaced with statistical weights from a machine-learning algorithm<sup>26</sup>. These terms can alternatively be replaced by experimentally determined thermodynamic coupling energies. However, these were only available for 15 amino-acid pairs at **a-a'** at the time of our earlier study<sup>39</sup>, and using them gave inferior performance. Since then, Vinson and co-workers have measured coupling energies for 55 amino-acid pairs at **a-a'**<sup>22</sup>. Additionally, we recognized that almost all of the improvement upon replacing **d-d'** interactions with statistical weights can be attributed to Leu-Leu pairs, which are modelled as only slightly favourable in structure-based approaches, contrary to experimental data. As a result of these findings, we developed model HP/S/Cv. Structure-based **a-a'** interactions were replaced with **a-a'** coupling energies for 55 amino-acid combinations; the **d-d'** interaction for Leu-Leu was replaced with  $-2$  kcal/mol (no experimental value is available), and the resulting model was expanded using CE. Because effective self contributions from our structural models and experimental coupling energies may be on different scales, point ECI values for the **a** position were adjusted such that 100 folding free energies measured by Acharya *et al.* were predicted optimally (in the least squares sense) by the overall CE model – see Supplementary Fig. 10.

As a way to account for pair-wise interactions in the reference state, both variant models used in this study ignored the energy of intra-chain side-chain interactions in the final predicted structure (see reference<sup>26</sup>). Note, however, that because the process of placing side chains for structure prediction does take into

account all side-chain interactions, intra-chain interactions do make indirect contributions to the final energy, and corresponding ECI do emerge in cluster expansion.

### ***Integer Linear Programming***

Kingsford *et al.* have shown that the problem of finding the lowest-energy rotamer-based side-chain packing arrangement, in the context of protein design, can be expressed and solved as an ILP<sup>27</sup>. Given that CE provides the energies of the desired and undesired states as analytical functions of sequence, we introduced a similar approach for handling specificity in design. With notation as in Kingsford *et al.*<sup>27</sup>, we represent the sequence space in our problem of designing a peptide of length  $p$  as an undirected  $p$ -partite graph with node set  $V = V_1 \cup \dots \cup V_p$ . Set  $V_i$  contains one node for each amino-acid possibility at position  $i$ . For each state  $S$ , each node  $u \in V_i$  is assigned a weight  $E_{uu}^S$  corresponding to its contribution to the energy of that state. If  $S$  is a heterodimer state (i.e. a state in which the design is complexed with a protein of fixed sequence), this individual contribution is simply the sum of the point ECI corresponding to  $u$  and pair ECI corresponding to pairs between  $u$  and all amino acids of the partner sequence. If  $S$  is the design•design homodimer state, then  $E_{uu}^S$  is the sum of point ECI corresponding to  $u$  and pair ECI of  $u$  and its image on the opposite chain. The edges of the graph  $D = \{(u, v): u \in V_i \text{ and } v \in V_j, i \neq j\}$  are assigned weights  $E_{uv}^S$ . If  $S$  is a heterodimer state, then  $E_{uv}^S$  is the ECI of the corresponding intra-chain pair cluster function. If  $S$  is the design•design homodimer state, then additional contributions to  $E_{uv}^S$  come from the ECI between  $u$  and the image of  $v$  as well as  $v$  and the image of  $u$ . Given these definitions, the energy of the design sequence in any state  $S$  can be expressed as  $\mathcal{E}^S = \sum_{u \in V} E_{uu}^S x_{uu} + \sum_{u, v \in D} E_{uv}^S x_{uv}$ , where binary variables  $x_{uu}$  and  $x_{uv}$  determine which nodes and edges the sequence involves. Thus, the problem of optimizing the energy of state  $S$  can be expressed as an ILP seeking to minimize  $\mathcal{E}^S$ , under the constraint that the chosen nodes and edges correspond to one another. Further, because gaps between different states are also linear functions of decision variables  $x_{uu}$  and  $x_{uv}$ , arbitrary gap constraints can also be incorporated. Finally, any additional function of these decision variables, such as a PSSM score, can also be incorporated. With  $T$  as the target state and  $U_i$  representing undesired states, the ILP we used in this study is (where  $\mathbb{W}_j$  stands for the set difference between  $V$  and  $V_j$ ):

$$\text{Minimize : } \mathcal{E}^T = \sum_{u \in V} E_{uu}^T x_{uu} + \sum_{u,v \in D} E_{uv}^T x_{uv}$$

subject to :

$$\sum_{u \in V_j} x_{uu} = 1 \quad \text{for } j = 1, \dots, p$$

$$\sum_{u \in V_j} x_{uv} = x_{vv} \quad \text{for } j = 1, \dots, p \quad \text{and } v \in V \setminus V_j$$

$$\mathcal{E}^{U_1} - \mathcal{E}^T > gc_1, \quad \text{where } \mathcal{E}^{U_1} = \sum_{u \in V} E_{uu}^{U_1} x_{uu} + \sum_{u,v \in D} E_{uv}^{U_1} x_{uv}$$

...

$$\mathcal{E}^{U_k} - \mathcal{E}^T > gc_k, \quad \text{where } \mathcal{E}^{U_k} = \sum_{u \in V} E_{uu}^{U_k} x_{uu} + \sum_{u,v \in D} E_{uv}^{U_k} x_{uv}$$

$$\sum_{u \in V} W_u x_{uu} < pssmc$$

$$x_{uu}, x_{uv} \in \{0,1\}$$

Here  $k$  is the number of undesired states,  $gc_i$  is the gap constraint for  $i$ -th state,  $pssmc$  is the PSSM constraint and  $W_u$  is the PSSM weight corresponding to node  $u$ . We solved such ILPs with the *glpsol* tool from the GNU Linear Programming Kit (<http://www.gnu.org/software/glpk/>). Because of the simplicity of sequence-based expressions obtained through CE, solutions to these ILPs with as many as 46 undesired states were generally obtained within 1-5 minutes on a single 2.7 GHz CPU.

Note that although everything was formulated in this instance for energy functions that are pair-wise decomposable at the sequence level, in principle this approach can be easily generalized for higher-order terms. Clearly, the CE methodology is already capable of taking higher-order interactions into account, should there be a need<sup>28</sup>. The ILP formulation can be extended to handle higher-order terms by introducing additional decision variables. For example,  $x_{uvw}$  would be 1 if there is a triplet interaction between nodes  $u$ ,  $v$ , and  $w$ . Constraints for these new decision variables would also have to be imposed to ensure that higher-order interactions occur only between those nodes that are chosen (e.g. in this case  $x_{uu}$ ,  $x_{vv}$  and  $x_{ww}$  are 1). Note that these higher-order decision variables would have to be introduced only for those clusters of sites that do, in fact, participate in higher-order interactions. This allows the complexity of the ILP problem to grow naturally with the size of the system (i.e. the number of variables and constraints grows linearly with the number of interactions in the system).

### **PSSM Constraint**

To constrain CLASSY designs to favour a leucine-zipper fold, we derived heptad position-specific amino-acid frequencies from the multi-species alignment of 432 bZIP leucine zippers described above. These frequencies were then used to score all of the sequences in the alignment (taking into account only **a**, **d**, **e** and **g** positions), from which a length-normalized score distribution was derived. Based on this

distribution, a cutoff value of 0.247 was imposed in CLASSY such that all of the designed sequences had a PSSM score of at least 0.247. Although this is a stringent cutoff, with 84% of native sequences scoring below it, the sequence space remaining is still large. For example, for a six-heptad design sequence, where **a**, **d**, **e** and **g** positions are varied and 10 amino acids are allowed per position, the total sequence space is  $10^{24}$ , whereas after applying the PSSM cutoff of 0.247 it is still  $\sim 10^{18}$  (calculated by convolving score distributions at individual positions to obtain the final distribution of scores and integrating it from 0.247 up).

### *Choosing b, c and f Positions*

Positions **a**, **d**, **e** and **g** are assumed to encode most of the interaction specificity of the designed peptides<sup>19, 40</sup>. Thus, we chose the identities of the **b**, **c** and **f** positions such that they were appropriate for the already selected **a**, **d**, **e**, and **g** positions, given what is observed in the multi-species dataset of 432 bZIP sequences referenced above. Thus, for each **b**, **c**, and **f** position  $b_i$  we sought to optimize  $P(b_i|a_1, \dots, a_n)$ , where  $a_1 \dots a_n$  are the identities of the selected **a**, **d**, **e**, and **g** positions. We expressed this quantity in terms of probabilities we could measure from the dataset:

$$\begin{aligned} P(b_i|a_1, \dots, a_n) &= \frac{P(b_i, a_1, \dots, a_n)}{P(a_1, \dots, a_n)} = \frac{P(a_1|b_i, a_2, \dots, a_n) \cdot P(b_i, a_2, \dots, a_n)}{P(a_1, \dots, a_n)} \\ &= \frac{P(a_1|b_i, a_2, \dots, a_n) \cdot P(a_2|b_i, a_3, \dots, a_n) \cdot \dots \cdot P(a_n|b_i) \cdot P(b_i)}{P(a_1, \dots, a_n)} \\ &\approx \frac{P(a_1|b_i) \cdot P(a_2|b_i) \cdot \dots \cdot P(a_n|b_i) \cdot P(b_i)}{P(a_1, \dots, a_n)} \end{aligned}$$

The last step assumes that the pre-selected amino-acid decoration at positions **a**, **d**, **e**, and **g** represents well the natively observed decorations at these positions (i.e. probability  $P(a_k|b_i)$  measured in the **adeg** context of the designed peptide and the probability averaged over all native contexts is the same). Quantity  $P(a_1, \dots, a_n)$  is hard to estimate, but it is constant with respect to **b**, **c** and **f** and is therefore not important. Conditional probabilities  $P(a_k|b_i)$  can be easily measured from the native bZIP dataset, and for each **b**, **c** and **f** position the amino acid that optimizes the above probability can be found. Using this approach, we were able to obtain **b**, **c**, **f** decorations of natural content and distribution. However, we found that infrequently this procedure resulted in sequences with large charge and/or helix propensity (mostly due to the fact that the pre-selected **a**, **d**, **e**, and **g** amino acids already had high values of charge or helix propensity). Thus, we expressed the problem of finding the optimal **b**, **c** and **f** combination according to the above equation as an ILP (by taking the logarithm of the probability it can be decomposed into a sum of pre-computed probability logarithms) and incorporated constraints on total charge, charge content (number of charged residues) and

helix propensity. For each property, the range of acceptable values was defined as  $\mu \pm \sigma$ , where  $\mu$  and  $\sigma$  are the mean and standard deviation of the corresponding property in the native bZIP dataset. In a few instances this resulted in no solutions (i.e. the selected **a**, **d**, **e** or **g** were already outside of the range for one of the properties) and for these cases more liberal intervals were allowed (either  $\mu \pm 1.5\sigma$  or  $\mu \pm 2\sigma$ ). Finally, because we wanted to rely on UV absorbance for determining concentration, we imposed the additional constraint that the **b**, **c**, **f** positions contain at least one Y or W residue (unless there was one already present at **a**, **d**, **e** or **g**).

### ***Uncovering Specificity-encoding Features***

We analyzed the 8 designs determined to be most specific using the arrays to identify specificity-encoding features. First, we compared each design•target complex with the corresponding design•undesired heterocomplexes. For each such comparison, we computed the contribution of each amino acid in the  $i$ -th position of the design sequence ( $aa_i$ ) to the overall stability and specificity. This was done by computing the interaction of  $aa_i$  with the region of the target peptide from  $i-7$  to  $i+7$  (one heptad N- and C-terminal to  $aa_i$ ) as well as the interaction of  $aa_i$  with the same region of the undesired partner. The first value corresponded to the stability contribution of  $aa_i$  and the difference between the two was the specificity contribution. To further isolate specificity determinants, this difference was decomposed into contributions from different positions on the target sequence and the corresponding positions on the undesired partner sequence.

We performed a similar analysis to elucidate features encoding specificity against the design•design homodimer, except the contribution of each amino acid  $aa_i$  to specificity was considered as the difference between interaction of  $aa_i$  with the residue opposing it in the target sequence and its interaction with itself in the design homodimer. The same analysis was repeated for pairs of amino acids at all position pairs ( $i$  and  $j$ ) of the design sequence.

### ***Dividing Human bZIPs into 20 Families***

Human bZIPs were divided into 20 families based on the evolutionary analysis of Amoutzias *et al.*<sup>41</sup> with the exception of including CHOP and ZF as individual families, and condensing OASIS and OASISb into a single family based on the similarity of their interaction profiles<sup>4</sup>. The phylogenetic tree of human bZIPs shown in Supplementary Fig. 13 was made using only the leucine-zipper regions and was constructed with the PHYLIP (<http://evolution.genetics.washington.edu/phylip.html>) package using the Neighbour-Joining algorithm and the Jones-Taylor-Thornton (JTT) model of amino-acid replacements. TreeDyn (<http://www.treedyn.org/>) was used to visualize and annotate the tree.

### ***How Many Unique anti-bZIP Profiles Are There?***

Fig. 3A shows that our CLASSY designs exhibit many novel interaction profiles when binding human bZIPs, while the sequence diversity used to generate these profiles is rather limited (Fig. 3C). This suggests that there may be a very large number of different interaction profiles, of which our 48 designs have revealed only a very small portion. But how large is this number? To answer this question with high confidence we need either an extremely large number of designs and measurements or an extremely accurate model. At present, neither is available. However, if we have a good idea of a model's prediction accuracy and use this model to calculate the number of unique profiles that exist, we can then estimate a lower bound on the true number of profiles. Here, we used model HP/S/Cv for this purpose. Several steps were taken to ensure that our estimates were always below the true number of profiles.

The interaction profile of a peptide was defined as a binary vector indicating whether the peptide interacts (1) or does not interact (0) with each human bZIP. If two binary vectors are equal, the profiles are equivalent. In reality, there is a lot of space between such vectors, because interaction strength also plays a role in defining a profile. This is one way that we underestimated the total number of possible profiles. We also defined these vectors at the family level rather than the protein level – again, a significant underestimate of the real size of the profile space. We considered 19 out of the 20 families (due to difficulties assessing model performance on the ATF3 family), giving a total of 524,288 possible unique profiles. The following procedure was followed:

*Compute the total number of unique profiles predicted by HP/S/Cv.* For each human bZIP coiled coil  $P_i$  we defined a computational energy cutoff  $c_i$  to optimally discriminate interactions and non-interactions in the human bZIP interaction dataset (experimental interactions/non-interactions taken from Fong *et al.*<sup>25</sup>). To increase prediction confidence, we introduced a buffer parameter  $b$ , such that energy scores above  $c_i+b$  were considered non-interactions, below  $c_i-b$  were considered interactions, and scores between  $c_i-b$  and  $c_i+b$  were not considered as either ( $b$  was set to 3 kcal/mol by optimizing performance on the human bZIP interaction dataset). This parameter increases prediction confidence but reduces the number of peptides that can give rise to a profile, further reducing our final estimate. Next, we generated 1,000 random binary profile vectors and ran CLASSY to find the most stable sequence consistent with each profile (e.g. its interaction stability with each of the 40 bZIPs from the 19 considered families is either below  $c_i-b$  or above  $c_i+b$  in accordance to the profile). The bZIP PSSM constraint was applied. Out of these 1,000 cases, 5 produced a solution. Given that there are a total of 524,288 possible binary profiles, this translates into ~2,600 unique profiles that can be achieved in design.

*Estimate prediction rates.* The rates of true positive (TP), true negative (TN), false positive (FP) and false negative (FN) predictions were estimated from anti-bZIP•bZIP interaction data. Performance is expected to be worse than for the human•human dataset for several reasons. First, the process of design tends to

exacerbate errors in an energy function. Second, because designed sequences are different from native bZIPs in systematic ways, the ranges of HP/S/Cv scores for anti-bZIP•bZIP and bZIP•bZIP interactions will also be different, making cutoffs derived from the bZIP•bZIP dataset less applicable to anti-bZIP•bZIP interactions. Thus, although the prediction rates for the human•human interactions were  $TP = 0.84$ ,  $TN = 0.91$ ,  $FP = 0.16$ ,  $FN = 0.09$ , they were worse for the anti-bZIP•bZIP interactions:  $TP = 0.39$ ,  $TN = 0.94$ ,  $FP = 0.61$ ,  $FN = 0.06$ . The drastic difference between the two performance rates is a result of over-training optimal cutoffs to the case of human•human interactions, but since the most important goal here is not to over-estimate the performance rate, this approach is still valid. The performance predicting relative stabilities of two complexes of anti-bZIP•bZIP is much better than this.

*Given two predicted distinct profiles, find the probability that they are in fact the same.* This probability,  $ps$ , is a product of the probabilities that each individual element of the profile (interaction or non-interaction with each human bZIP) is the same. Formally,

$$ps = (TN \cdot TN + FN \cdot FN)^{zz} \cdot (TN \cdot FP + FN \cdot TP)^{oz} \cdot (TP \cdot FN + FP \cdot TN)^{zo} \cdot (TP \cdot TP + FP \cdot FP)^{oo},$$

where  $oo$ ,  $oz$ ,  $zo$ , and  $zz$  are the number of corresponding profile elements that are both 1, 0 and 1, 1 and 0, or both 0, respectively. Probability  $ps$  was estimated to be  $2.0 \cdot 10^{-4}$  by averaging over 1,000 pairs of randomly generated profiles.

*Calculate the probability distribution of the true number of profiles.* We predicted that there exist ~2,600 unique profiles. The first one we consider is certainly unique. The second one is predicted to be unique, but it is actually unique with probability  $1 - ps$ . The third one is also predicted to be unique, but it is truly different from the first and the second with probability  $(1 - ps)^2$ . In general, if  $P(k, n)$  is the probability of having  $k$  unique profiles after considering  $n$  predicted unique profiles, then we can give the recursive definition  $P(k, n) = P(k, n-1) \cdot k \cdot ps + P(k-1, n-1) \cdot (1 - ps)^{k-1}$ . Using this we generated the probability distribution of the true number of profiles after considering 2,600 profiles. This distribution had a sharp peak around 1,900 profiles and quickly fell to essentially zero before and after that (integral between 1,785 and 1950 is 0.9999). Based on this, there should exist at least ~1,900 unique peptide•human bZIP interaction profiles, and probably there are many more.

### ***A Picture of Multi-state Energy Phase Space***

Specificity-sweep calculations predict that designs selected solely for optimal binding to the target are often not specific, and are especially prone to homodimerization (see Supplementary Fig. 2A). Many specificity problems can be eliminated by sacrificing relatively small amounts of stability (Supplementary Fig. 2C). However, it is not clear how severe the specificity constraint is and how much it restricts the choice of sequences. We investigated this in a simplified case where design•design homodimers are the only



competing state. We constructed a 2D histogram of the entire design sequence space for several design problems, looking at the distribution of design•target energies versus design•design energies. In such a histogram, each 2D bin corresponds to energy ranges for the design•design and the design•target complexes and contains the number of sequences that satisfy these ranges.

If each amino acid at each site made an independent contribution to the total energy, this histogram could be built by convolving the 2D energy histograms of each individual site. However, amino acids at different sites interact with each other. To address this, we used the fact that amino acids more than a heptad apart do not interact in our CE energy expressions. As in the case for independent site contributions, sites were considered one-by-one and their histograms were convolved with the running total. However, at each step energy contributions from both single-residue and pair-wise interactions with residues in the preceding heptad were incorporated. In order to account for the pair-wise terms appropriately, individual histograms were maintained for each unique sequence combination in the preceding heptad. To limit memory usage, only 9 amino acids were considered at each site for this purpose. Note that because positions **b**, **c** and **f** were not explicitly considered in our models, there were a total of  $9^4 = 6,561$  possible heptad sequences and 6,561 running total histograms needed to be kept at each stage. In the last step these 6,561 histograms were added to produce the final 2D histogram.

The results for ATF-2 and MafG are shown in Supplementary Fig. 12 (other bZIPs produced similar results). The dashed lines show where the design•design and design•target energies are equal. Clearly, most stable sequences are even more stable as homodimers (i.e. are below the line; note log scale), indicating that destabilization of the design homodimer is an extremely severe constraint that limits sequence space by many orders of magnitude.

## **Experimental Characterization**

### ***Jun family constructs***

The following peptides were used for the Jun family, which have more uniform length than those previously constructed by Newman & Keating<sup>4</sup>.

cJun

MSYYHHHHHHLESTSLYKKAGSGSRKLERIARLEEKVKTLLKAQNSELASTANMLREQVAQLKQKVMNHLE ,

JunB

MSYYHHHHHHLESTSLYKKAGSGSRKLERIARLEDKVKTLLKAENAGLSSTAGLLREQVAQLKQKVMNHLE ,

JunD

MSYYHHHHHHLESTSLYKKAGSGSGSRKLERISRLEEKVKTLLKSQNTTELASTASLLREQVAQLKQKVMNHLE

### ***Data Analysis***

Scanned images of slides were analyzed using the program Digital Genome (Molecularware). For each probe the scan at the highest PMT voltage that did not show saturation was used for analysis. The signal in the red channel from the Alexa Fluor 633 hydrazide was used to identify the location of spots. The median signal and median background for each spot was determined, and signal less background for each spot was calculated. Missed spots and artifacts were manually flagged and removed from analysis; these represented less than 0.1% of all spots. For each pair of adjacent sub-arrays probed with the same labeled peptide, the average of 8 measurements for each protein on the surface was calculated and defined as  $a$ . These values are reported in Supplementary Tables 3 – 5.

Two other quantities were used in analyses. Because a small number of probes showed high background, a corrected fluorescence signal was defined as  $F = a - \tilde{a}$ , with  $\tilde{a}$  the median of all signals measured using a common probe. The maximum of this quantity for a given probe was designated  $F_{\max}$ . The quantity  $-\log(F/F_{\max})$  was used in Fig. 2, Fig. 3A, and Supplementary Figures 1 and 14 to indicate relative array signal differences.

To distinguish signal from noise, and thus put an approximate lower bound on the signal required as

evidence for an interaction, we defined the quantity  $S_{\text{array}}$  as 
$$S_{\text{array}}(a) = \frac{(a - \tilde{a})}{\sqrt{\frac{\sum_{i=1, a_i < \tilde{a}}^N (a_i - \tilde{a})^2}{N_{a < \tilde{a}}}}}$$
, where  $\tilde{a}$  is

again the median of  $a$ ,  $N$  is the number of unique printed proteins, and  $N_{a < \tilde{a}}$  is the number of proteins producing  $a$  below the median.  $N$  and  $N_{a < \tilde{a}}$  excluded other designed peptides on the surface when the solution probe was itself a designed peptide.  $S_{\text{array}}$  is a Z-score-like quantity, where the distribution of signals below the median was assumed to be primarily noise-driven and thus was used to correct stronger signals.  $S_{\text{array}}$  values are also provided in Supplementary Tables 3 – 5.

For the purpose of estimating the number of designs that homodimerize, and how many designs interacted with their target, the following criterion was used: A and B were judged to give signal above background, and thus to interact, if they produced an  $S_{\text{array}}$  score above 2.5 either when A was on the surface and B was the probe or when B was on the surface and A was the probe. This cutoff was chosen based on reported homodimerization of bZIP families as well as our solution measurements of stability<sup>4, 42, 43</sup>.

### ***Interaction-Profile Clustering***

An interaction profile was defined using  $-\log(F/F_{\max})$  scores derived from microarrays, and profiles were clustered using Euclidian distance as the dissimilarity metric. Average linkage clustering was performed using the *linkage* command in Matlab 6.5.

## ***Circular Dichroism***

Circular dichroism (CD) spectra were measured on AVIV 400 and 202 spectrometers in 12.5 mM potassium phosphate (pH 7.4)/150 mM KCl/0.25 mM EDTA/1M GuHCl/1 mM DTT. All mixtures of peptides were incubated at room temperature for several hours before measurement. Wavelength scans were performed at 40  $\mu$ M total peptide concentration and measured at 25 °C in a 1-mm cuvette. Scans were monitored from 280 nm to 195 nm in 1 nm steps averaging for 5 seconds at each wavelength. Three scans for each sample were averaged. Thermal unfolding curves were performed at 4  $\mu$ M total peptide concentration and measured in a 1-cm cuvette. Melting curves were determined by monitoring ellipticity at 222 nm with an averaging time of 30 seconds, an equilibration time of 1.5 minutes, and a scan rate of 2 °C/min. All samples were measured from 0 °C to 85 °C unless otherwise noted. All thermal denaturations were reversible.  $T_m$  values were estimated by fitting thermal denaturation data to a monomer-dimer equilibrium, assuming no change in heat capacity upon folding. Specifically, we fit the derivative of the CD signal with respect to temperature to the equation:

$$\frac{d(\text{signal})}{dT} = A \cdot \frac{\Delta H}{RT^2} \cdot \exp\left(-\frac{\Delta H}{R} \left[\frac{1}{T_m} - \frac{1}{T}\right]\right) \left[ \frac{4 \exp\left(\frac{\Delta H}{R} \left[\frac{1}{T_m} - \frac{1}{T}\right]\right) + 1}{\sqrt{8 \cdot \exp\left(\frac{\Delta H}{R} \left[\frac{1}{T_m} - \frac{1}{T}\right]\right) + 1}} - 1 \right].$$

Here  $A$ ,  $\Delta H$ , and  $T_m$  were fitting parameters, with  $\Delta H$  and  $T_m$  corresponding to the change in enthalpy upon folding and the apparent melting temperature, respectively. We fit the derivative of the CD signal to reduce the reliance of the fit on pre- and post-transition baselines<sup>44</sup>. For two-species mixtures AB, the difference between the melting curve of the AB mix and the average of melting curves of A and B ( $S_{AB-A-B}$ ) was calculated and treated as the signal for the purposes of fitting the above equation. No fitting was performed for mixtures where  $S_{AB-A-B}$  was positive at any point during the unfolding transition (i.e. the signal from the average was stronger than the signal from the mixture), as it was not clear which species was being melted. Those mixtures with  $S_{AB-A-B} > 0$  over the entire temperature range were assumed to show no evidence of interaction. Fitting was performed using the non-linear least squares method in Matlab 6.0. The 95% confidence intervals resulting from the fits are reported in Supplementary Table 2.

## ***Comparing CD and Array-based Stability Ordering***

Relative stability orders established by CD and microarray were compared conservatively. The arrays were only used to judge relative stabilities when two interactions involved the same solution probe interacting with partners on the same array surface. CD ranks were determined by visual inspection of thermal melts, with cases where the order was not clearly obvious being assigned the same rank. Array ranks

for interactions sharing a common probe were established based on the  $S_{\text{array}}$  measure, with ranks differing by only one unit in normalized  $S_{\text{array}}$  considered the same. All possible pair-wise comparisons of CD and array ranks were made, a total of 41 comparisons, 35 of which gave the same order by CD and microarray.

### ***Array Results were Highly Reproducible***

The array measurements were highly reproducible over replicate experiments and a range of concentrations, as shown in Supplementary Fig. 14. The complete array data (averaged background-corrected signals as well as  $S_{\text{array}}$  scores) are given in Supplementary Tables 3, 4, 5 and 6. Proteins listed in columns were fluorescently labelled and used in solution as probes against proteins on the surface, which are listed in rows. All protein probes were at 160 nM unless otherwise noted. Duplicates are labelled. Supplementary Tables 3, 4, and 5 contain values from experiments in rounds 1, 2, and 3 respectively. Supplementary Table 6 contains experimentally determined  $S_{\text{array}}$  scores for 33 human proteins.

## **Supplementary Discussion**

### ***Beyond bZIPs: Requirements for Applying CLASSY to Other Systems***

There are a variety of reasons that we selected bZIP transcription factors for this study. They comprise a biologically important class of proteins for which questions of interaction specificity are central to function. But also, interaction specificity is probably better understood for the bZIPs than for any other protein complex, and convenient properties of these proteins facilitate modelling and measurement. To what extent can CLASSY be applied to other problems in molecular recognition? To answer this it is important to distinguish between limitations that arise from CLASSY itself – of which there are few – and limitations that arise from our understanding of specificity in other protein complexes. The systematic study of protein interaction specificity is a new, expanding research area. There are already several complexes amenable to study using CLASSY, and this number will increase with advances in experimental screening technologies and computational modelling.

Below we outline three requirements that must be met to apply CLASSY to a specificity design problem. For each, we comment on how the bZIPs satisfy the requirement and discuss prospects for other complexes.

#### **1. Application of CLASSY requires that sets of desired and competing states be defined.**

To address interaction specificity explicitly, one must define the universe of relevant complexes. For many problems, competing states of particular interest can be identified as those that share structural and evolutionary similarity with the target. In our bZIP application, the competitors were other bZIPs. These can

be detected easily by sequence similarity. Many related interaction specificity problems can be posed. In the design of peptides to activate specific integrins, the competitors would be other integrins; in the design of specific PDZ domains the competitors would be undesired protein C-terminal peptides; in the design of BH3 peptides that bind specific Bcl-2 family members, the competitors would be other Bcl-2 proteins. Although criterion 2 (below) may not yet be satisfied for these examples, at least one prior example of a successful design calculation in each of these cases illustrates progress in modelling and highlights the types of applications where CLASSY may prove fruitful<sup>11, 12, 14</sup>. Similar examples can be constructed for any set of paralogous interaction domains; zinc-finger and homeodomain transcription factors as well as SH2, SH3 and PDZ domains are discussed below.

## 2. A scoring function must provide information about the relative stabilities of the states under consideration.

Specificity can be designed using CLASSY only if a model captures information about the relative favourability of different states. CLASSY can use many types of scoring functions. Physical/structure-based models and empirical/statistical models are equally compatible with the requirements of the method. The only formal requirement is that the scoring function be expressed as a linear function of sequence variables (not necessarily limited to amino-acid pair terms). We have demonstrated that cluster expansion can accomplish this for complex structure-based energy functions and for several different protein folds<sup>28, 29, 32</sup>. Cluster expansion can in theory also be applied directly to large experimental datasets, where available, to generate a predictive expression in the appropriate computational form.

In designing anti-bZIPs, we took advantage of experiments that elucidated some of the determinants of interaction specificity; we captured these in a hybrid structure-based/experiment-based model, which was tested using available peptide array data<sup>4, 26</sup>. Specificity-scoring functions published for other protein domains can now be tested using CLASSY. For example, models based on fitting residue interactions to experimental data have been developed for PDZ domains and zinc fingers. Such scoring functions typically have the functional form required for CLASSY<sup>1, 3, 34, 45</sup>. Scoring functions based on structural modelling have the greatest potential to be general. RosettaDesign has been used for many applications, including the design of specific protein-protein interactions<sup>8, 46</sup>. Other structure-based specificity models have been tested for PDZ<sup>12</sup>, SH2<sup>33</sup> and SH3<sup>47, 48</sup> domains. Structure-based models have also shown good performance for several transcription factor families<sup>49-52</sup>. Physical structure-based models face significant challenges, in particular capturing side-chain and backbone relaxation that can impact specificity. But as new methods for modelling structural relaxation are developed (and several groups report progress in this area<sup>36, 53, 54</sup>), there are no obvious barriers to employing them in conjunction with CLASSY. In fact, we recently demonstrated that cluster expansion works well when applied to models that incorporate backbone flexibility<sup>32</sup>. Finally,

structural approaches that use atom-based or residue-based statistical potentials can give good predictions of binding energies and can capture some interaction specificity trends<sup>37, 55, 56</sup>; such models may prove especially useful for negative design.

How good do the scoring functions need to be? Our bZIP scoring functions, while capable of distinguishing strong interactions from non-interactions, do not provide quantitative predictions of relative stability (they do not correlate strongly with experimental  $\Delta\Delta G$  estimates). Models can likely be effective for use in CLASSY if they (1) accurately capture some key specificity determinants and (2) are not under-defined. A model is under-defined if it has many missing or inappropriate weights; these can allow the design optimization calculations to proceed into non-sensible regions of sequence space. In our bZIP study, the experiments of Vinson and colleagues provided valuable data contributing to (1), though these experiments did not comprehensively assess all possible specificity determinants<sup>19, 22</sup>. To address (2), we used structural modelling to impose a physically realistic description of all amino-acid interactions that were not defined by experiments. A similar combined approach is likely to be appropriate for other domains. For example, for PDZ domains and zinc fingers, a small set of weights derived from experiments seem to predict much of the observed specificity<sup>3, 45</sup>. But structural modelling may be required to provide reasonable (even if not highly accurate) estimates for the many amino-acid interactions that are not constrained by experiments. Also important for addressing (2) is the ability of CLASSY to incorporate sequence property constraints (e.g. the PSSM constraint used in this study), which can be used to ensure that only the sequence space that is reasonably well described by the underlying model is considered in design.

Finally, energy gaps in CLASSY can be chosen according to the estimated accuracy of the underlying energy function. Thus, if errors in predicted energies are known to be large, the user can choose to impose large energy gaps as constraints, ensuring that any designs returned are predicted to have a significant preference for the desired state over others (at the risk of finding either no solutions or only poorly stable solutions).

In summary, while we do not yet know if breakthroughs in predicting specificity will come primarily from improvements in modelling or from fitting to large experimental data sets, this likely does not matter in terms of applying CLASSY. Designing specific PDZ/SH2/SH3 domains or specific PDZ/SH2/SH3 ligands, or zinc-finger transcription factors with specialized binding profiles, are already good candidate applications for testing this method more broadly.

### 3. An experimental assay appropriate for testing the specificity of the proteins under study is required.

It is impossible to know the quality of the scoring function, or the quality of CLASSY designs, without experiments that report on interaction specificity. Assessing specificity profiles generally involves testing

many possible complexes. For the bZIPs, we took advantage of a previously validated peptide microarray assay<sup>4</sup>. Similar large data sets exist for SH2, SH3, PTB, and PDZ domains, as well as for many transcription factors<sup>2, 57-61</sup>. Exciting advances using SPOT arrays, protein microarrays, protein-binding DNA arrays, phage-display/phage ELISA, protein complementation assays and plate-based fluorescence assays expand the possibilities in this area, and suggest that many moderately sized binary complexes will be amenable to analysis<sup>1-5, 59-63</sup>.

### ***CLASSY Introduces Negative Design Using Familiar bZIP Features***

CLASSY designs employed a range of strategies to achieve specificity, but some trends were evident. Designs optimized for stability alone often had **a** and **d** positions with medium-to-large hydrophobic residues<sup>22</sup>, and CLASSY initially improved specificity by maintaining these cores and modulating electrostatic **g-e'** interactions in early iterations of the specificity sweeps (see Fig. 1C for definitions of coiled-coil heptad positions; a prime indicates a residue on the opposite helix). To achieve greater specificity ( $\Delta$ ), at a greater price in stability, CLASSY introduced core substitutions such as pairing of Ile with Ala (e.g. to destabilize homodimers using Ala-Ala pairs). The sequences selected for testing typically included additional elements, such as charged amino acids in core positions. Such interactions imparted large amounts of specificity but were also predicted to be quite destabilizing. They were chosen for analysis because we judged specificity to be relatively more important; generic strategies such as ACID extensions could be used to improve stability if necessary<sup>64</sup>.

Our 8 most specific designs exhibit canonical bZIP specificity determinants (Supplementary Fig. 15A): there is a strong preference for Asn at an **a** position to be paired with Asn at the opposing **a'**, and electrostatic complementarity is exploited at **g-e'** positions<sup>19, 26</sup>. Interestingly, a less recognized complementarity between **g-a'** positions is predicted to make a comparable, if not larger, contribution to specificity; this feature was extensively used in our designs (Supplementary Fig. 15A)<sup>65</sup>. A strong preference for Leu-Leu over all other amino-acid pairs at **d-d'** positions was also exploited<sup>66</sup>. Finally, our model predicts that interactions between **a** and **d'** can contribute significantly to specificity. In particular, a beta-branched residue at an **a** position strongly prefers a non-beta branched residue at the next **d** position of the opposing strand. Similar effects have been noted in anti-parallel coiled coils<sup>31</sup>.

### ***Off-target Interactions May Form via Structures That Were Not Modelled***

In our computational modelling, we considered only parallel coiled-coil dimer structures with a unique axial alignment of helices. For the designs that bound to their targets, it is likely that the interaction occurred as modelled because the designs were restrained to have leucine zipper-like sequences, frequently

retained buried Asn and Lys residues to favour dimers over other oligomers, and retained paired Asn residues at **a-a'** positions to favour particular parallel alignments<sup>67, 68</sup>. These features were selected automatically by CLASSY in most cases, and where they were not present in all candidate designs, we imposed a bias for such solutions when choosing examples for experimental testing. Further supporting the formation of dimers, interactions of designs with their targets were observed to occur irrespective of which peptide was printed on the array and which was labelled in solution, which is unlikely for some alternate stoichiometries.

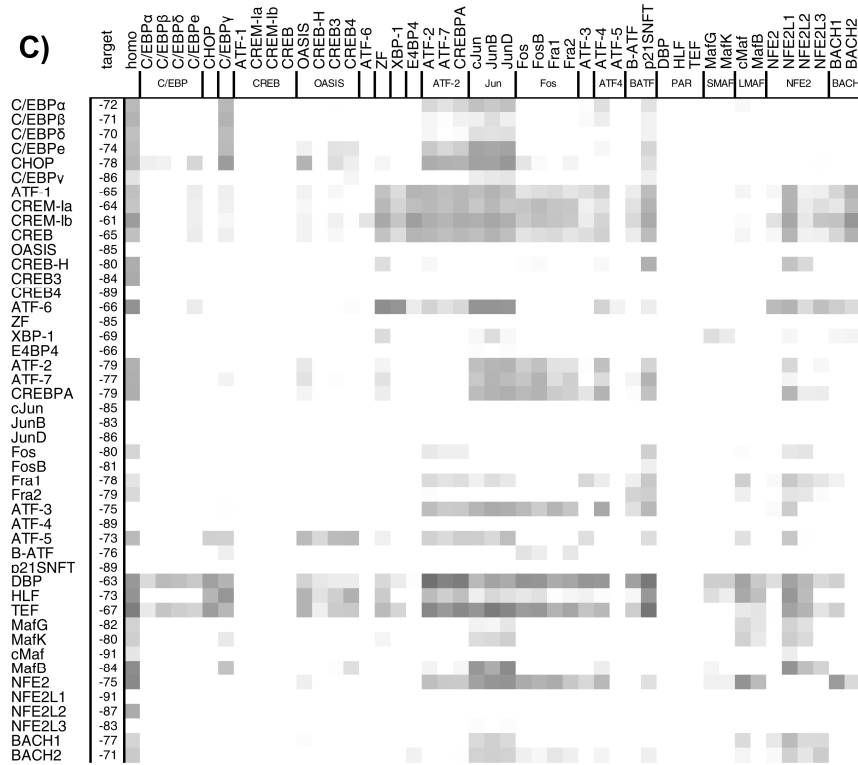
When unexpected design•off-target interactions occurred, it is less clear what the structures of those complexes were. In several instances, we suspect that the complex formed was not one that was modelled as an undesired state. For example, the strong interaction between anti-SMAF-2 and ATF-4 (Supplementary Fig. 1) was predicted to be very unfavourable relative to anti-SMAF-2•MafG (Supplementary Fig. 16A-B). However, because the SMAF family has an Asn in a different heptad than most human bZIPs, the alignment used to model anti-SMAF-2 paired with ATF-4 left two asparagines at **a** positions unpaired (see Supplementary Fig. 16A). Asn residues have a strong preference to occur in pairs in coiled-coil dimers<sup>22</sup>, and it is unlikely that the anti-SMAF-2•ATF-4 interaction would occur in this way. More likely, the complex would adopt a shifted axial alignment (though this is also predicted to be unfavourable, Supplementary Fig. 16C), an anti-parallel helix orientation, or some other structure. Anti-BACH2-2, which showed strong homo-association on the array, illustrates another case where the complex formed may not be the one that was modelled as an undesired state. Anti-BACH-2 homodimer was predicted to be much less stable than anti-BACH-2•BACH1. However, although anti-BACH-2 has very strong anti-homodimerization features, they are heavily concentrated in the first two N-terminal heptads (see Supplementary Fig. 17). It is likely that this portion of the homodimer simply does not fold, and the rest of the sequence forms a stable association. Of course, if such problems can be anticipated, additional constraints can be incorporated into CLASSY, where alternative alignments, coiled-coil lengths and orientations can be explicitly considered.



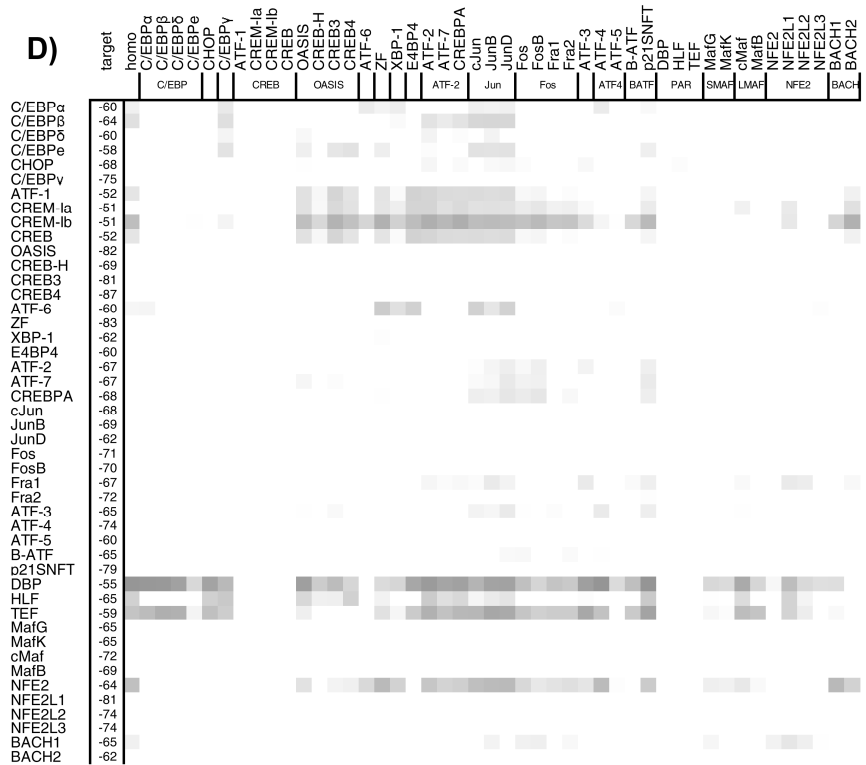


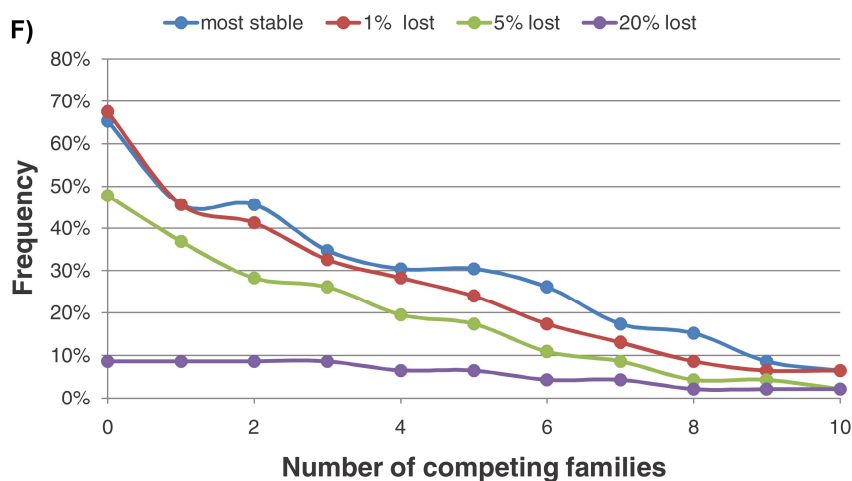
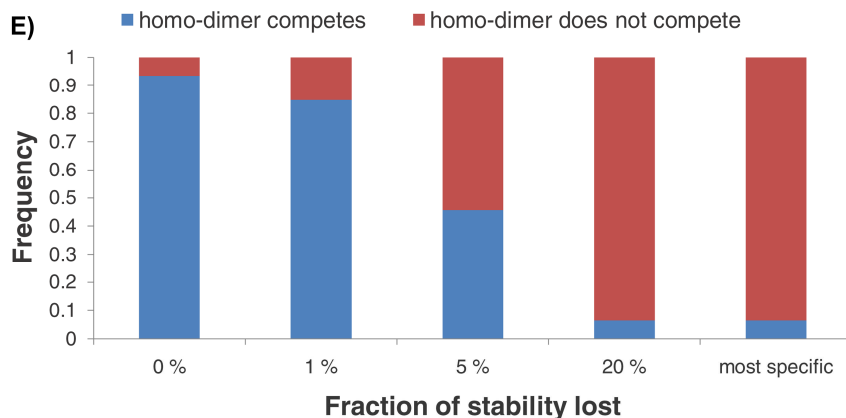


C)

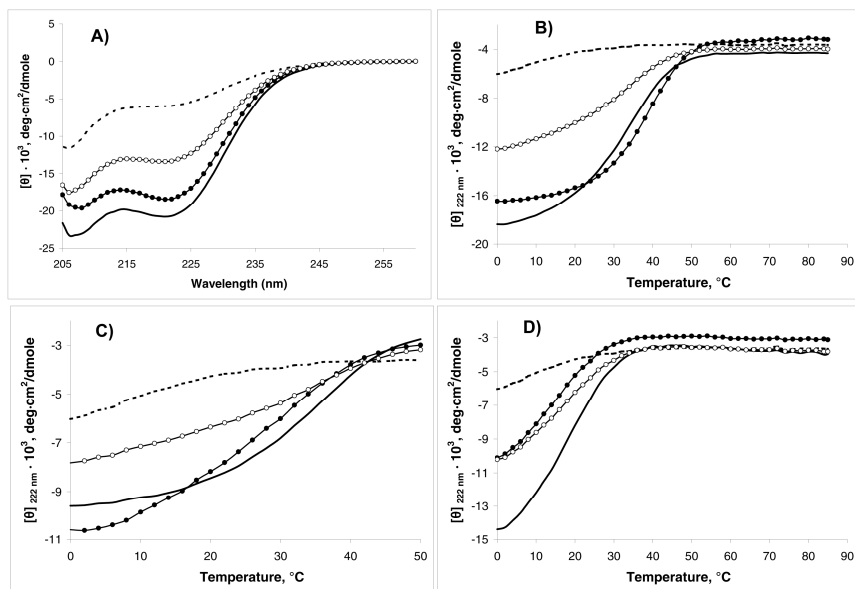


D)

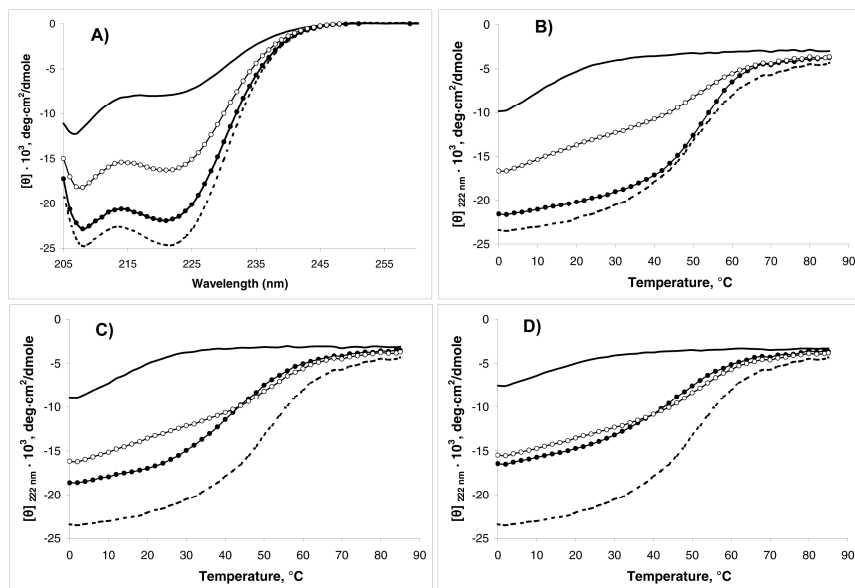




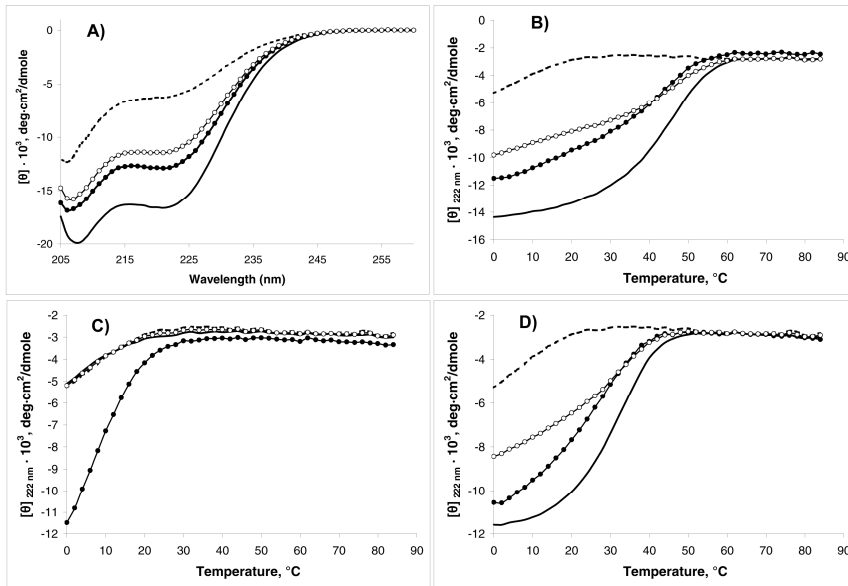
**Supplementary Figure 2** A global view of specificity sweeps with each human bZIP coiled coil as a target. In each row, the protein indicated at left is the target. The first column contains the score of the optimal design•target complex, whereas each subsequent column contains the energy gaps between the design•target complex and the corresponding design•competitor complex, including the design homodimer in the second column. A positive energy gap corresponds to design•target being more favorable than design•competitor. The color bar gives the energy scale. (A), (B), (C) and (D) correspond to designs from different stages of specificity sweeps. In (A) the design producing the most stable complex for each target was used to compute energies (first iteration). In (B) up to 1% of the stability score was sacrificed to gain specificity. In (C) up to 5% of stability was sacrificed and in (D) the most specific designs were considered. In (E) and (F) the specificity data are summarized as a function of decreasing stability. (E) shows the proportion of anti-human designs for which the design•design homodimer has a gap of less than 6 kcal/mol, and (F) shows the proportion of designs predicted to compete with a non-target-family human bZIP by the same criterion. Energies were computed using model HP/S/Cv.



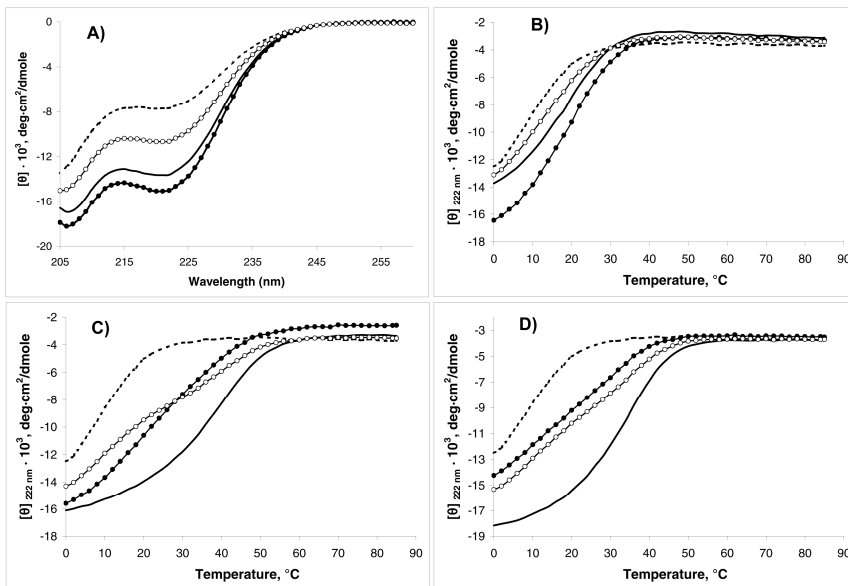
**Supplementary Figure 3** Solution characterization of anti-ATF2 by CD. Format and presentation is the same as in Fig. 2B-E for anti-SMAF. The target protein is ATF-7 (which is in the same family as ATF-2) (in **A** and **B**), the closest off-target competitor is p21SNFT (in **C**), and the bZIP related to the target by sequence is cJun (in **D**).  $T_m$  values are given in Supplementary Table 2.



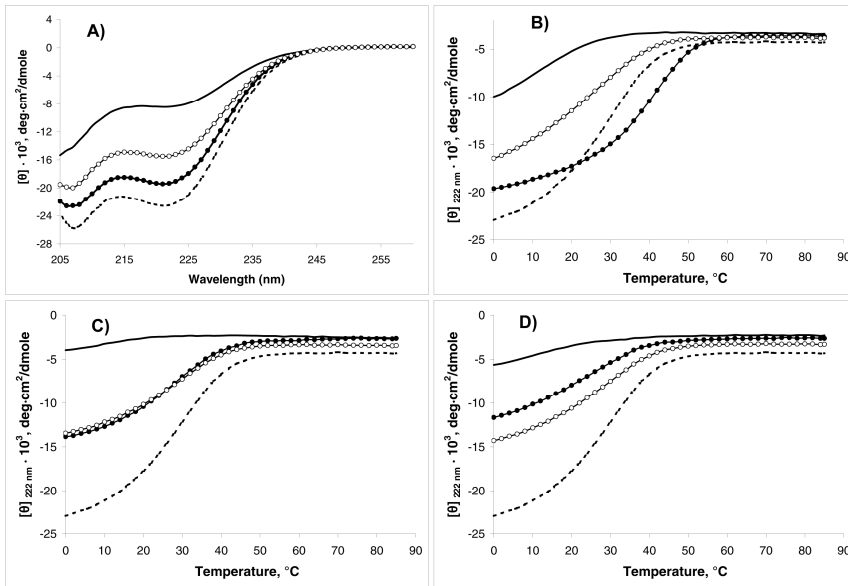
**Supplementary Figure 4** Solution characterization of anti-ATF4 by CD. Format and presentation is the same as in Fig. 2B-E for anti-SMAF. The target protein is ATF-4 (in **A** and **B**), the closest off-target competitor is Fos (in **C**), and the bZIP related to the target by sequence is ATF-3 (in **D**).  $T_m$  values are given in Supplementary Table 2.



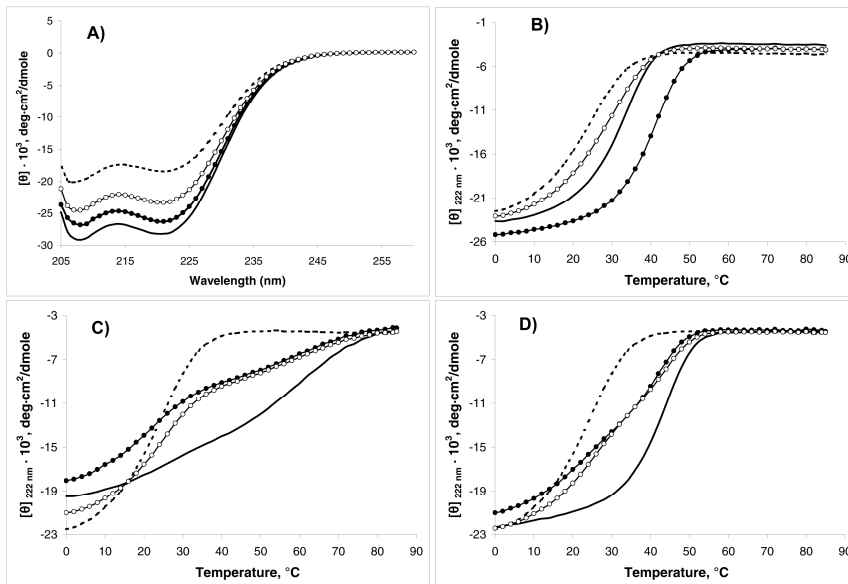
**Supplementary Figure 5** Solution characterization of anti-LMAF by CD. Format and presentation is the same as in Fig. 2B-E for anti-SMAF. The target protein is cMaf (in **A** and **B**), the closest off-target competitor is Fra2 (in **C**), and the bZIP related to the target by sequence is MafG (in **D**).  $T_m$  values are given in Supplementary Table 2.



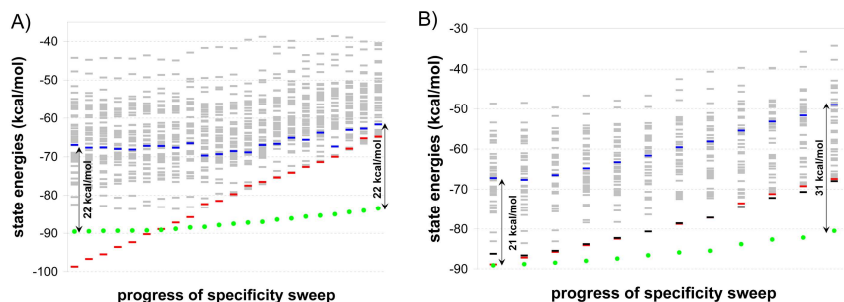
**Supplementary Figure 6** Solution characterization of anti-JUN by CD. Format and presentation is the same as in Fig. 2B-E for anti-SMAF. The target protein is cJun (in **A** and **B**), the closest off-target competitor is CHOP (in **C**), and the bZIP related to the target by sequence is ATF-7 (in **D**).  $T_m$  values are given in Supplementary Table 2.



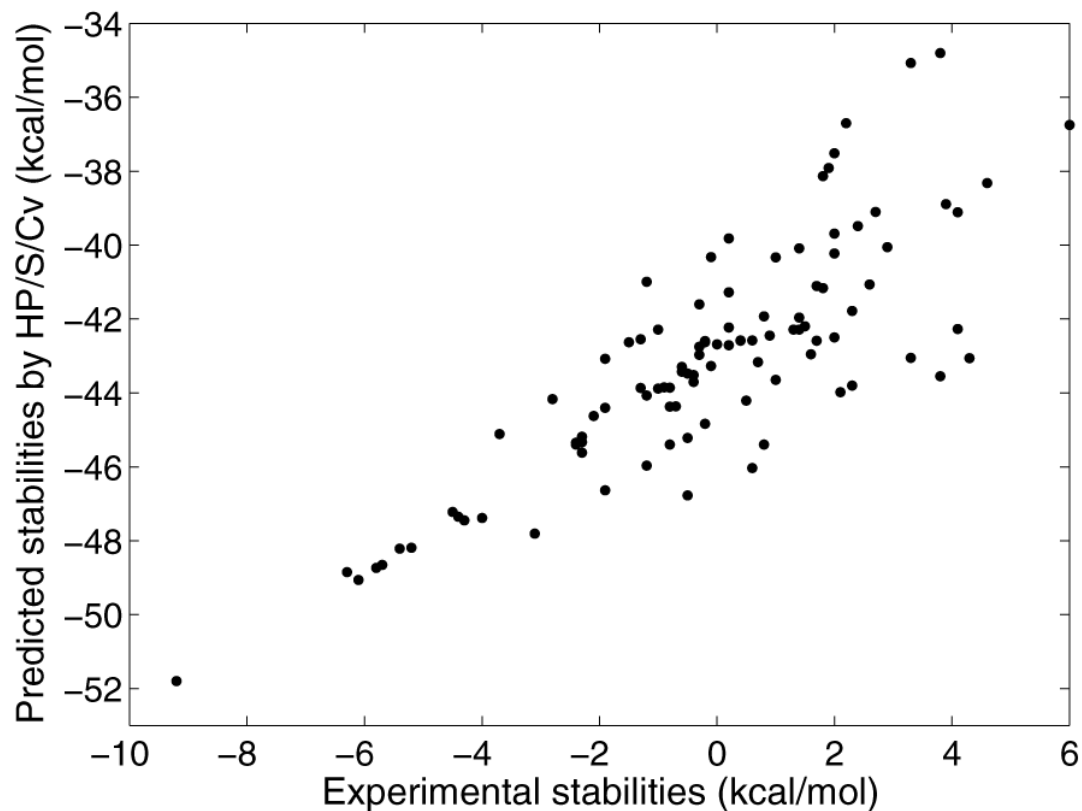
**Supplementary Figure 7** Solution characterization of anti-FOS by CD. Format and presentation is the same as in Fig. 2B-E for anti-SMAF. The target protein is Fos (in **A** and **B**), closest off-target competitor is BACH1 (in **C**), and bZIP related to the target by sequence is ATF-3 (in **D**).  $T_m$  values are given in Supplementary Table 2.



**Supplementary Figure 8** Solution characterization of anti-ZF by CD. Format and presentation is the same as in Fig. 2B-E for anti-SMAF. The target protein is ZF (in **A** and **B**), closest off-target competitor is NFE2 (in **C**), and the bZIP related to the target by sequence is XBP-1 (in **D**).  $T_m$  values are given in Supplementary Table 2.

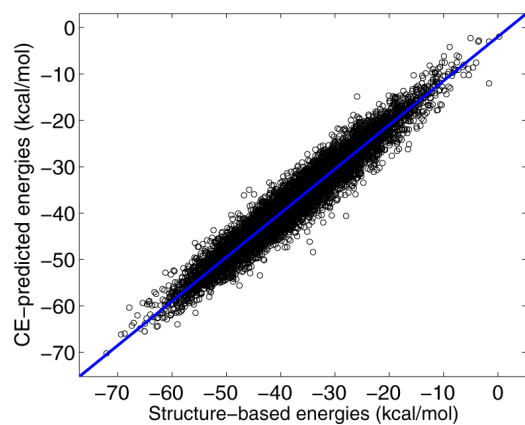
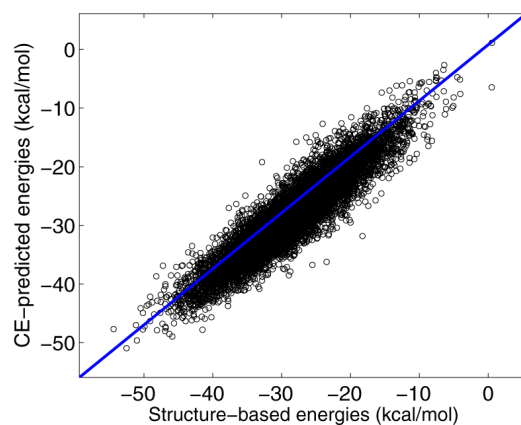


**Supplementary Figure 9** Specificity sweep (A) and biased specificity sweep (B) diagrams for the design of a peptide to bind the leucine-zipper region of ZF. Green dots correspond to the design•target complex and red bars to the design•design complex. Blue bars in A) correspond to the energy of the design•XBP-1 complex, which contrary to the prediction of the model showed evidence of strong interaction on the microarray. As a way of addressing this issue, a biased specificity sweep was conducted for ZF, where the gap between the energies of the design•ZF and design•XBP-1 complexes was shifted by 19 kcal/mol. This is shown in (B) with blue bars corresponding to the actual model-predicted design•XBP-1 energy, while the black bars are the energies used in the biased specificity sweep. Whereas in the regular specificity sweep there is no competition with the design•XBP-1 state, due to its incorrectly predicted high energy, in the biased specificity sweep this competition is imposed. This procedure generated a successful, highly specific design: anti-ZF.

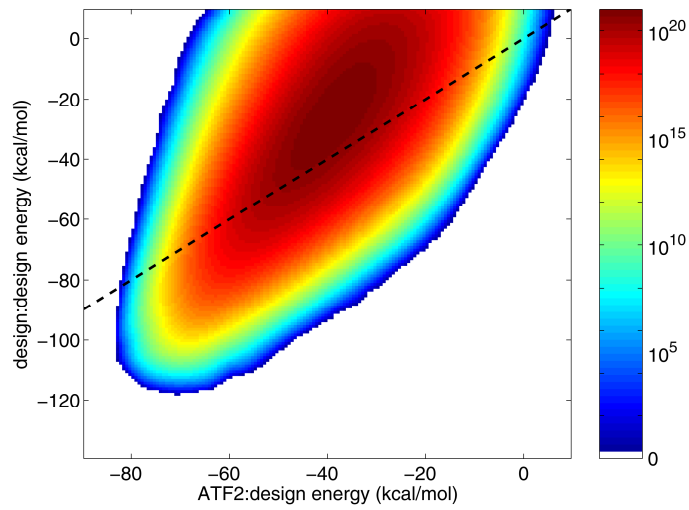
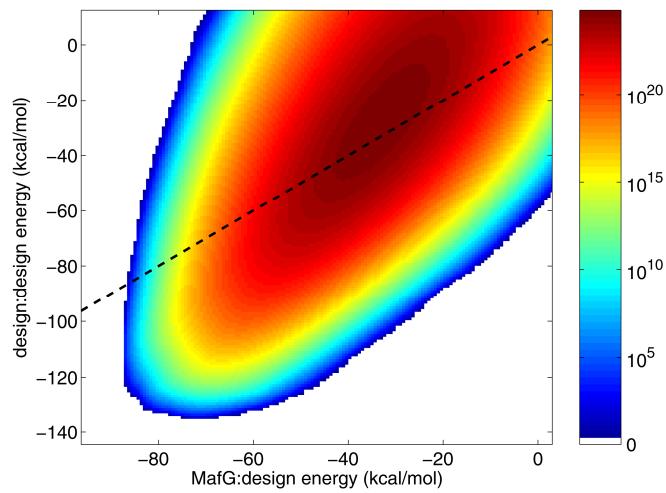


**Supplementary Figure 10** Adjusting the 9 a-position point ECI in model HP/S/Cv to optimally fit 100 stabilities experimentally measured by Vinson and co-workers<sup>22</sup>. R for the final fit is 0.83.

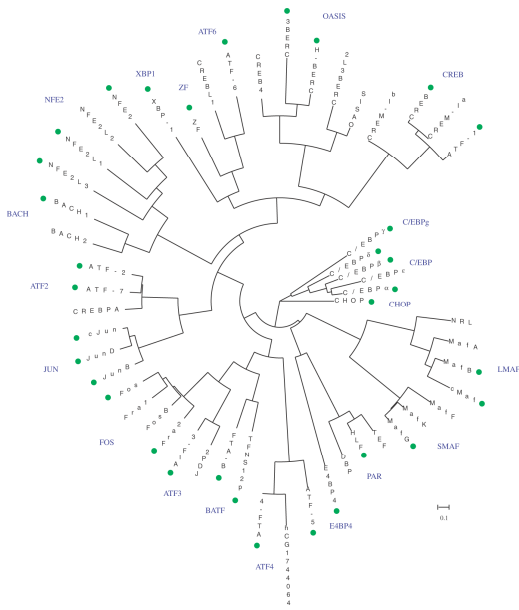


**A****B**

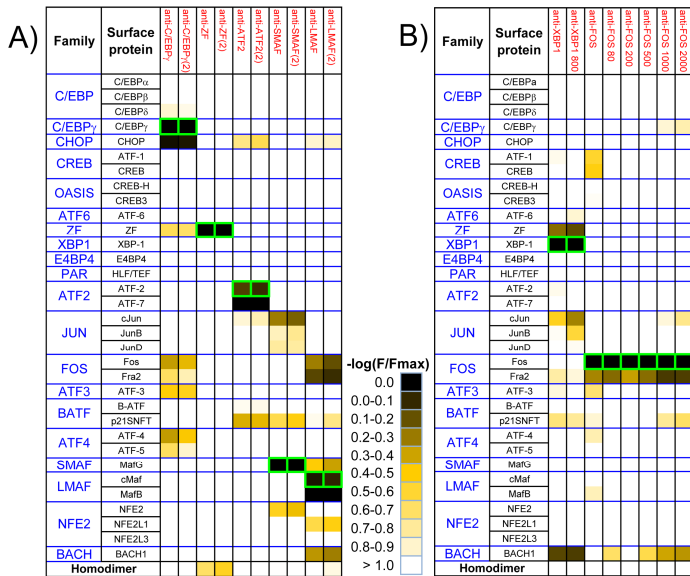
**Supplementary Figure 11** The performance of cluster-expanded versions of models HP/S/Ca and HP/S/Cv (panels **A** and **B**, respectively) on a randomly generated set of 10,000 sequences not present in the training set. Root mean square deviations between CE-predicted and structure-based energies are 2.4 and 2.6 kcal/mol for HP/S/Ca and HP/S/Cv, respectively. The cluster expansions contain 2,544 ECI for HP/S/Ca and 2,470 ECI for HP/S/Cv.

**A****B**

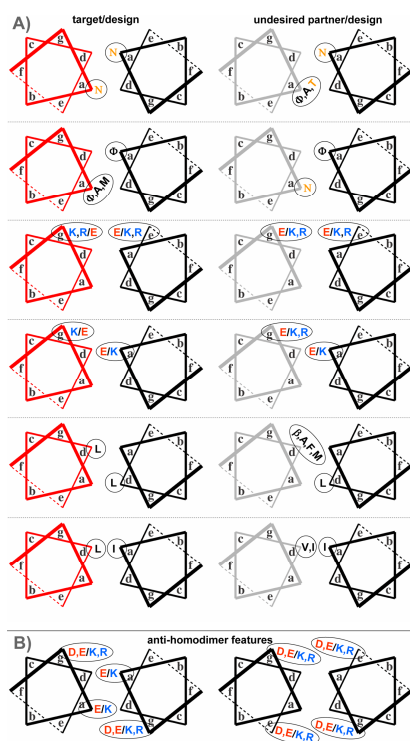
**Supplementary Figure 12** 2D energy histograms of two states – the design•target state and the design•design homodimer state. Color represents the total number of possible sequences in each bin (bin sizes are  $\sim 1$  kcal/mol). The targets are ATF-2 and MafG in (A) and (B), respectively. The line where design•target and design•design scores are equal is shown. By optimizing only the design•target energy, sequences with high homodimerization propensity will be obtained in these examples. The specificity sweep procedure run with only one disfavoured state (design•design) locates the top boundary of this phase space.



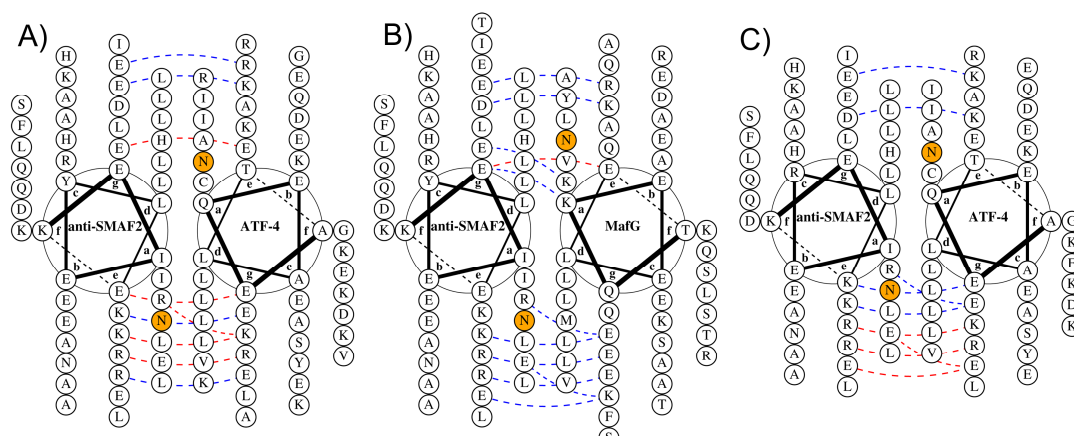
**Supplementary Figure 13** Phylogenetic tree constructed using the leucine-zipper regions of all human bZIP proteins. Protein names are in black and family names are in blue. Green dots indicate the 33 proteins used in the experiments in this study. The scale refers to amino-acid replacements per site.



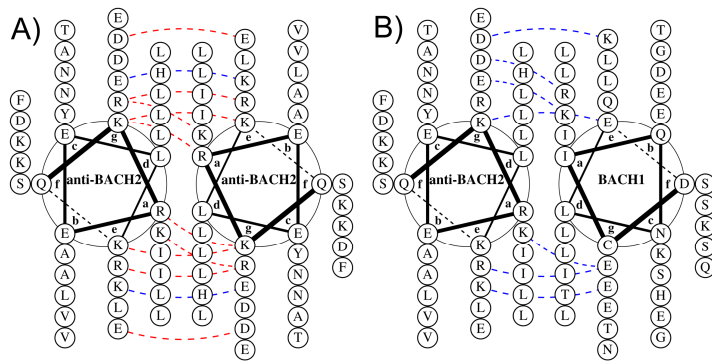
**Supplementary Figure 14** Reproducibility of protein-microarray measurements of design interactions probed in duplicate in (A) and at different concentrations in (B) (probe concentration in nM is shown as part of the probe name in the top row and is 160 nM where not indicated). Data are displayed in the same format as for Supplementary Fig. 1.



**Supplementary Figure 15** Common specificity mechanisms in successful designed peptides. **A)** Specificity features used for discriminating between design•target and design•off-target interactions. The design is in black, the target in red and the undesired partner in gray. Amino acids listed with single-letter codes are the residues comprising the specificity pattern. Slashes delineate subgroups of residues, with corresponding subgroups delineated similarly at the interacting position.  $\Phi$  designates hydrophobic residues Ile, Val or Leu and  $\beta$  stands for beta-branched residues Ile or Val. In the last row, the **a-d'** interaction is between an **a** residue and the more C-terminal **d'** residue on the opposite helix. **B)** Specificity features commonly used in designed peptides to disfavor the design•design homodimer, using the same notation.



**Supplementary Figure 16** Helical-wheel diagrams for anti-SMAF-2 complexes with ATF-4 and MafG. **(A)** The anti-SMAF-2•ATF-4 complex is predicted to be much weaker than the anti-SMAF-2•MafG complex shown in **(B)**, in large part due to the misaligned asparagines at **a** positions in anti-SMAF-2•ATF-4. **(C)** A different alignment of anti-SMAF-2•ATF-4, where the asparagines match up, may be more favorable, although it is not predicted to be much stronger computationally. Diagrams made with DrawCoil 1.0 (<http://www.gevorggrigoryan.com/drawcoil/>).



**Supplementary Figure 17** Helical-wheel diagrams of the anti-BACH-2 homodimer complex, shown in **(A)**, and the anti-BACH-2•BACH1 complex shown in **(B)**. The strong anti-homodimerization features of anti-BACH-2 are concentrated at the N-terminus of the sequence, leaving open the possibility that this portion simply does not fold, while the remainder of the coiled coil forms a stable complex. Diagrams made with DrawCoil 1.0 (<http://www.gevorggrigoryan.com/drawcoil/>).

## Supplementary Tables

**Supplementary Table 1** All designed sequences tested. For each design, listed in columns are: the name of the design, the name of the bZIP target for that design, the family of the target bZIP, the round of design/testing in which this sequence was produced, the count of attempts to design a partner for the given target, the energy function used and the designed sequence. Note that designs are named after the family of the target rather than the individual protein. There were three rounds of experiments. Attempts are different than rounds because not all targets were attempted in the first (or second) rounds. An attempt involved testing one or two designs (in one case, three) for each target considered in a set of experiments. When the first experimental attempt to identify a specific design was unsuccessful, alternative solutions from the specificity sweep were selected for testing in subsequent rounds (constituting further attempts). In a few cases, listed in the footnotes, these additional designs were created with a modified procedure aimed at addressing experimentally identified shortcomings of previous designs.

Design name	Target	Family	Round	Attempt	Method	Design sequence
anti-C/EBP-2	C/EBP $\alpha$	C/EBP	2	1	HP/S/Ca	f g a b c d e f g a b c d e f g a b c d e f g a b c d e f g a b c d e f g a b c d e f g a b c
anti-C/EBP	C/EBP $\alpha$	C/EBP	3	2	HP/S/Cv	FENVTHEFILATLEENENAKLRRLLEAKLERELARLRNEVAWL
anti-C/EBP-3	C/EBP $\alpha$	C/EBP	3	2	HP/S/Cv	AENQYVEDLIQYLEKENARLKKEVQRLVRELSYFRRIAEIA
anti-C/EBP $\gamma$	C/EBP $\gamma$	C/EBP $\gamma$	2	1	HP/S/Ca	NDLDAYEREAEKLEKKNVLRNRLAALENELATLRQEVASMKQELQS
anti-C/EBP $\gamma$ -2	C/EBP $\gamma$	C/EBP $\gamma$	2	1	HP/S/Cv	RDLQNVREIQSLEKKNESLKKKIASLENELEATLTKQEIAYFKRELAY
anti-CHOP	CHOP	CHOP	3	1	HP/S/Cv	DRLAVKENRVAVLKNENAKLRNI IANLKDRIAYFRRELAYLELEEEQLA
anti-CREB	CREB	CREB	2	1	HP/S/Cv	QLVAQLRNSKVEQLVNRNQAALKNKLEYLRQEIAYETEQ
anti-CREB-2	CREB	CREB	3	2	HP/S/Cv <sup>[1]</sup>	NKVEQLKKNKVEQLKRNRAALKNLRLARLERE IAYAE
anti-CREB-3	CREB	CREB	3	2	HP/S/Cv <sup>[1]</sup>	QKVESLQKQIEELKQKKAQLKNDIANLEKEIAYAE
anti-OASIS	CREB3	OASIS	2	1	HP/S/Cv	QKVEQLKKNKVEQLKKNESLENKVAELKRNREYLKKNKIENLINDITNLENDVAR
anti-OASIS-2	CREB3	OASIS	2	1	HP/S/Cv	QKVAELKKNRVAVLKNRNELQKKNKVEELKRNRAYLKNELEATLENEVARLENDVAE
anti-OASIS-3	CREB3	OASIS	3	2	HP/S/Cv <sup>[5]</sup>	QKVAQLKKNRVAYLKNENAKLENI VARLENDNANLEKDIANLEKDIANLERDVAR
anti-OASIS-4	CREB3	OASIS	3	2	HP/S/Cv <sup>[5]</sup>	QKVAQLKKNK IAKKEDENAVLENLVALENEENAYLEKELARLERDIARAERDVKV
anti-ATF6	ATF-6	ATF6	3	1	HP/S/Cv	EKIQLKRRRLAYFRRENATLKNNDNATLENELASVEAENEALRK
anti-ZF-2	ZF	ZF	2	1	HP/S/Cv	QKIAYLRDR IAAALKAENEALRAKNEALRSKIEELKKEEELRDKIAQKKDR
anti-ZF	ZF	ZF	3	2	HP/S/Cv <sup>[6]</sup>	NLVAQLENEVASLENNENETLKKKNLHKKDLIAYLEKEIANLRKKIEE
anti-XBP1-2	XBP-1	XBP1	2	1	HP/S/Ca	SKYDALRNKLEALKNRINAQLRKNELRLEEAFLVLRNEVL
anti-XBP1	XBP-1	XBP1	2	1	HP/S/Cv	QKIAYLRDR IAAALKAENEALRAKNEALRSKIEELKKEEELRDKIAQKKDR
anti-E4BP4-2	E4BP4	E4BP4	2	1	HP/S/Ca	QKRQELKQRLAVLENDNARLKNDLAQLEVEEAYIE
anti-E4BP4	E4BP4	E4BP4	2	1	HP/S/Cv	NKNVNVKNRLAVLENNENATLRLNLAWLRLAAME
anti-E4BP4-3	E4BP4	E4BP4	3	2	HP/S/Cv <sup>[3]</sup>	EKNQELKKNRLAVLENDNAAALRNDLRLARLERE IAYME
anti-ATF2-2	ATF-2	ATF2	1	1	HP/S/Ca	QKLQTLRDLAVLENRNQLKQRLRQHLKDLKYLEDELEATLEKE
anti-ATF2-3	ATF-2	ATF2	2	2	HP/S/Cv	STVEELLRATIQLEKRNNAELKKNKKEELKNLVAHLRQELAAHKYE
anti-ATF2	ATF-2	ATF2	3	3	HP/S/Cv	NTVCKELKNIQLEELERNALKNLKEHLKFAKAELEFELAAHKFE
anti-ATF2-4	ATF-7	ATF2	3	3	HP/S/Cv	QKVEELKKNIAELENNRVAVKNRVAHLKQEIAYLKDELAHEFE
anti-JUN	cJun	JUN	1	1	HP/S/Ca	STAATLENDLRLARLENNARLEKDIANLERDLAKLEREEAYF
anti-FOS	Fos	FOS	1	1	HP/S/Ca	NEKEELKSKKAEALRNRIEQLKQKREQLKQKI ANLRKEIEAYK
anti-ATF3	ATF-3	ATF3	1	1	HP/S/Ca	ELTDELKKNKEALRKDAAALINELASLENEIANLEKEIAYFK
anti-ATF3-2	ATF-3	ATF3	1	1	HP/S/Ca	NETEQLLINKKQQLKNDNAALEKDAASLEKEIANLEKEIAYFK
anti-ATF3-3	ATF-3	ATF3	3	3	HP/S/Cv <sup>[7]</sup>	NILASLENKKEELKKNLNAHLKEIENLEKEIANLEKEIAYFK
anti-ATF4	ATF-4	ATF4	2	1	HP/S/Cv	KRIAYLRKKIAALKKDNANLEKDIANLENEIERLIKEIKTLENEVASHEQ
anti-ATF4-2	ATF-4	ATF4	2	1	HP/S/Cv	ARNAYLRKKIARLKKDNLQLEERDEQNLEKI IANLRDEIARLENEVASHEQ
anti-BATF	p21SNFT	BATF	2	1	HP/S/Ca	NELESLENKKEELKKNRNEELKQKREQLKQKLAALRNKLDAYKNRL
anti-BATF-2	p21SNFT	BATF	3	2	HP/S/Cv	NDIENLKDKEELKQKREELKQKIEYLKQKIEALRQKLAALKQRIA
anti-BATF-3	p21SNFT	BATF	3	2	HP/S/Cv	EKIIEELKDKIAELSRNAALRNKIEALKQKLEALRQKIEYLKDRIA
anti-PAR	HLF	PAR	3	1	HP/S/Cv	NRLQLEENKNEVLEKRAELRNEVATLEQELAAHRYELAAIEKEIA
anti-SMAF-2	MafG	SMAF	1	1	HP/S/Ca	KEIEYLEKEIEERLKDRLREHLQDNAAHRQELNARLEEAKELEFLAHLST
anti-SMAF-3	MafG	SMAF	1	1	HP/S/Ca	KEIERLEKEIKITLINLLTTLRQDNAAHRKEAAALEKEEANLERDIQNLLRY
anti-SMAF	MafG	SMAF	2	2	HP/S/Cv	KEIANLEKEIASLEKAVLQKRNAAHKQEVAAALRKEIAYVEDEIQYVEDE
anti-LMAF-2	cMaf	LMAF	3	1	HP/S/Cv	NKNETLKNINARLRNDVARLKNRIARLKDDIENVEDEIQYLE
anti-LMAF-3	cMaf	LMAF	3	1	HP/S/Cv	LENAQIKKEIAQLRKEVAQLKQKIEELKNDNARVEREIQYLE
anti-LMAF	cMaf	LMAF	3	1	HP/S/Ca	KDIANLKKETIAHLKNDLQRLRESIRERLKFDILNHEQEEYALE
anti-NFE2	NFE2	NFE2	1	1	HP/S/Ca	QKRQQLKQKLAALRRDIENLQDEIAYKEDEIANLKDKIEQLLS
anti-NFE2-2	NFE2	NFE2	3	2	HP/S/Cv	QKIEESLKDKLANKRDKIALLRSEVASFEKEIAYLEKEIANLEN
anti-NFE2-3	NFE2	NFE2	3	2	HP/S/Cv <sup>[4]</sup>	EKIEYLKDKLAHKNRVAQLRKEVTHKVDLETSLNEVAQLLK
anti-BACH-2	BACH1	BACH	2	1	HP/S/Ca	QKREELKSRKAYLRKETIANLKKDIINLDDLVAFELVTL
anti-BACH	BACH1	BACH	2	1	HP/S/Cv	QKIYQLKQRIAEALRKKIANLRKDIANLEDDAAVKEDELVHL
anti-BACH-3	BACH1	BACH	3	2	HP/S/Cv <sup>[2]</sup>	EKIEYLKDRIAELRSKIAALRNDLTHLKNDKAHKENLAHLA

[1] The only strong off-target interaction for design anti-CREB, produced in round 2, was the design\*design homodimer. However, the specificity sweep produced no solutions that were significantly more specific against the homodimer. Thus, in the next round we sought to remove design homodimerization by considering only the homodimer as a competitor. In the resulting designs anti-CREB-2 and anti-CREB-3,

homodimerization was indeed no longer a problem, but global specificity was reduced. This indicates that maintaining gaps to many states simultaneously can be important.

[2] The two strong off-target competitors for anti-BACH in round 2 were Fos and NFE2. The latter was deemed too close in sequence to effectively discriminate with our models. To improve specificity against Fos, a biased specificity sweep was used with a gap offset of -10 kcal/mol for Fos (making gaps with Fos more negative than they would be, which caused competition with Fos to be more stringent). However, anti-BACH-3 still interacted with Fos more strongly than with BACH-1.

[3] The initial two designs against E4BP4 were not very stable, and this was not predicted by the models. HP/S/Cv predicted that the most stable design against E4BP4 had a Lys at the N-terminal **d** position. To address this, we temporarily adjusted the ECI for Leu-Leu at **d-d'** in HP/S/Cv to be more favorable by 2 kcal/mol and reran the specificity sweep procedure. Anti-E4BP4-3 was picked from this list. Although this resulted in a more hydrophobic core, there was no detectable increase in stability according to the microarray assay.

[4] The only strong off-target competitor for anti-NFE2 was ATF-4, so in this design we used a biased specificity sweep approach with a gap offset of -3 kcal/mol for ATF-4 (making gaps with ATF-4 more negative). However this design interacted with Fos, which had not previously been a strong competitor.

[5] To eliminate the only significant competitor of anti-OASIS, p21SNFT, a biased specificity sweep was run with a gap offset of -10 kcal/mol for p21SNFT. This did indeed eliminate p21SNFT as a competitor, but MafG emerged as a new strong competitor.

[6] Because the only significant competitor for the first design, anti-ZF-2, was XBP-1, we applied a biased specificity sweep approach with a gap offset of -10 kcal/mol for XBP-1. This successfully removed XBP-1 as a competitor and resulted in a very specific and stable design.

[7] Significant competitors for designs against ATF-3 were Fos and ATF-4, whereas the models considered JUN and ATF2 families more likely to interact. To bias the specificity sweep against the relevant competitors, gap offsets of +8 and +2 kcal/mol for JUN and ATF2 families respectively were imposed (making gaps with JUN and ATF2 family members less important in the optimization).

**Supplementary Table 2** Melting temperature ( $T_m$ ) values estimated by fitting to CD-monitored melting curves. Corresponding 95% confidence intervals are given in brackets (see section Circular dichroism). Some measurements were made in duplicate to evaluate reproducibility; duplicate measurements are marked with a number two in parentheses.

<b>bZIP•bZIP homodimers</b>	<b><math>T_m</math> (° C)</b>	<b>95% CI</b>	<b>design•design homodimers</b>	<b><math>T_m</math> (° C)</b>	<b>95% CI</b>
CHOP	36.4	[35.8 36.9]	anti-SMAF	11.6	[11.1 12.1]
BACH1	8.4	[6.9 9.9]	anti-ATF2	5.2	[1.7 8.7]
XBP-1	42.0	[41.7 42.3]	anti-ATF4	48.6	[48 49.3]
NFE2	multiple transitions		anti-LMAF	3.0	[-3.4 9.3]
ZF	31.6	[31.3 31.8]	anti-ZF	22.1	[21.7 22.4]
MafB	19.8	[19.1 20.6]	anti-JUN	7.3	[6.6 8.1]
cMaf	43.1	[42.5 43.8]	anti-FOS	27.2	[26.8 27.6]
Fra2	<0	[-13.4 5.7]			
p21SNFT	33.0	[32.6 33.4]	<b>design•bZIP heterodimers</b>	<b><math>T_m</math> (° C)</b>	
ATF-4	7.9	[6.1 9.7]	anti-ATF4:ATF-4	52.1	[51.4 52.8]
ATF-3	9.4	[6.4 12.3]	anti-ATF2:ATF-7	41.0	[40.4 41.6]
ATF-3(2)	6.6	[4.3 9]	anti-SMAF:MafG	37.9	[37 38.7]
Fos	10.6	[8.9 12.4]	anti-JUN:cJun	24.2	[23.4 24.9]
Fos(2)	9.0	[8.1 9.9]	anti-FOS:FOS	43.6	[42.7 44.4]
cJun	16.6	[16.0 17.3]	anti-ZF:ZF	43.0	[42.7 43.4]
cJun(2)	16.2	[15.7 16.8]			
ATF-7	31.4	[31 31.8]			
ATF-7(2)	31.7	[31.3 32.1]			
MafG	30.2	[29.7 30.8]			
MafG(2)	31.8	[31.5 32.2]			

**Supplementary Table 3** Average background-corrected fluorescence values (top panel) and  $S_{array}$  values (bottom panel) from round 1 of array measurements. Peptides on the surface are in rows, those in solution in columns. Duplicate measurements are marked with a number two in parentheses. The anti-FOS peptide was tested at concentrations ranging from 80 nM to 2000 nM, as indicated in the probe names.

	ATF-2	cJun	Fos	Fra2	ATF-3	ATF-4	p21SNFT	MatG	NFE2	NFE2L1	BACH1	anti-ATF2-2	anti-ATF3	anti-ATF3-2	anti-JUN	anti-FOS	anti-SMAF-2	anti-SMAF-3	anti-NFE2	anti-FOS80	anti-FOS200	anti-FOS500	anti-FOS1000	anti-FOS2000
C/EBPα	-1210	-220	291	-790	4573	16459	2170	-341	-1513	92	-207	-819	1158	277	-406	-599	-294	80	719	204	270	542	1006	765
C/EBPβ	-2318	-543	-1388	-2407	7687	5320	2156	338	-4575	-3091	-721	-2532	-311	367	-283	357	-401	-1885	-2307	327	292	394	1885	2025
C/EBPδ	-3236	-540	-1080	-978	7598	15172	2626	-245	-1144	-575	-1533	-566	955	351	-118	-617	-131	-274	153	216	119	-174	-263	47
C/EBPγ	5156	346	733	3662	21941	34209	5172	-8	-861	-717	82	8475	2588	390	616	2685	-112	1142	492	277	409	2204	6819	5538
CHOP	5110	997	5691	3996	24898	5879	18419	331	-980	-830	3778	14295	3111	438	2289	1453	-155	-139	380	285	317	1676	5433	3921
ATF-1	-1962	-947	-880	-861	-2213	-2297	-2073	-319	-3054	-2447	1971	-2297	-687	410	1454	-923	60	-1207	832	407	388	263	913	597
CREB	-6257	-1355	-1371	-2093	-4313	-3194	-3907	-671	-5053	-2797	-3077	-2383	-290	420	756	-991	-316	-2506	-2083	420	374	-534	-2069	-1908
CREB-H	-933	-831	-6	-351	-235	-91	-376	-284	-708	-596	210	3	-890	220	5	1466	-113	217	51	197	217	1359	4159	3986
CREB3	-821	-279	-1117	-1014	1397	-1810	438	-75	-816	-1099	-599	915	-1117	357	444	1225	38	212	39	346	330	927	4479	3427
ATF-6	-3780	-1124	-980	-1252	-1826	-1502	-2021	-867	-2016	-2256	642	-992	-1229	302	-105	-897	-369	-995	-1414	289	199	-574	908	-257
ZF	-790	296	-360	-385	3405	5529	1020	862	2730	21548	2383	4059	973	337	-48	-372	494	3841	-862	292	246	-124	-415	-416
XBP-1	-2049	-359	-727	-2537	445	-815	-1979	-294	-1705	-1383	696	-1393	-3462	335	-687	-904	-885	-970	-1292	300	296	464	1520	2020
E4BP4	-1943	-513	-932	-546	1080	-1776	306	116	-817	-1685	101	1511	-1435	306	532	-113	-154	-486	-475	248	205	302	1728	1681
ATF-2	6995	4899	5724	4636	20295	1121	5203	-266	-2236	3164	4399	6768	486	322	-303	-87	-332	433	5345	303	260	226	1282	1361
ATF-7	9594	8100	6785	7510	22271	4420	7512	-453	724	7212	6092	13903	828	239	223	462	58	1555	3605	241	223	556	2135	2013
cJun	14760	1997	27052	24951	24562	-320	11769	-334	-2590	-1941	42	532	806	285	4862	3658	1305	561	5047	267	419	2437	6728	5978
JunB	4759	449	16151	16857	18106	-758	9729	17	-219	-1815	-299	-457	-226	332	1163	1084	889	207	1225	289	273	908	3121	2901
JunD	9823	789	22889	22692	22719	-1332	11304	98	-712	-1082	203	-588	259	272	2164	1755	1268	390	2837	251	274	1211	4347	3817
Fos	13984	35015	2121	6143	6143	7327	704	-855	-1298	5427	2638	201	24396	4055	-197	35148	4274	4563	15804	842	3726	23274	38006	25664
Fra2	9788	19126	3893	609	9627	1023	266	-634	-847	1584	200	-858	6044	1450	-107	17466	240	609	13442	512	1444	11485	26425	17567
ATF-3	19928	12050	2675	6100	195	11043	13952	181	-4022	-2556	647	-120	29694	969	1028	1568	1084	-519	382	408	1836	5583	4559	
ATF-4	1729	-1060	8751	-223	21845	-1508	14458	-338	-1788	27265	2517	8517	26002	595	1520	-547	43710	5986	45654	362	249	-547	-1360	-1113
ATF-5	-4832	-1941	-2288	-1926	-2946	-3322	3955	-945	-704	-997	1437	4261	-129	374	126	-1422	2217	1429	-2845	312	221	-958	-3448	-3212
B-ATF	2150	13378	463	-713	8148	4416	2425	-675	-52	3341	4338	3926	1459	322	-214	819	-118	371	360	234	254	1175	3755	3799
p21SNFT	9069	12668	298	-21	23989	8109	5699	1393	-1526	-536	3288	11316	6628	489	249	2910	26	-42	765	354	490	2749	7858	6503
HLF	-2699	-188	-1099	-829	-628	-3384	122	-474	-4778	-1520	21	-1036	-371	254	23	-634	-408	-1248	-2333	225	145	-126	-459	-203
MatG	-254	268	768	-532	3180	-193	2433	1388	11187	49005	25075	908	1654	568	595	1147	7203	11537	13977	274	235	1283	4001	3615
cMaf	-253	93	212	128	2222	1816	-1081	-39	-899	3789	5397	-1087	-167	392	35	848	283	2257	3369	334	337	1224	3534	2918
MatB	-734	-218	2279	1685	4347	-1943	-936	-472	-572	4157	11696	-968	-433	368	66	1441	324	1584	1353	360	399	1642	5263	4429
NFE2	-2475	-545	-444	-1246	-891	-115	-1827	346	15668	-1367	759	-1075	434	317	-177	161	1185	10083	35514	269	247	629	2240	1951
NFE2L1	99	-285	771	-13	-1683	3979	-770	38492	-329	4830	-126	66	3531	389	269	-75	51690	4758	1654	270	226	106	1112	1050
NFE2L3	-4487	-1629	-3241	-2831	-4049	-415	-3341	31186	13034	-3371	-1015	-1228	-1316	189	-1338	-1048	6395	-428	-1981	283	218	-589	-1905	-1728
BACH1	6186	457	1933	2078	-50	3253	896	18424	-618	332	2580	-788	2616	374	706	6630	3044	4954	11213	341	614	5078	14231	10864
anti-ATF2-2	735	-410	-874	-1497	-754	-731	3897	-171	-2764	13088	-1593	6187												
anti-ATF3	7223	1718	19633	15128	30646	16602	9409	3310	15014	44240	11233		-906											
anti-ATF3-2	579	562	13724	7513	8260	839	936	532	306	8928	295													
anti-JUN	2506	6966	2211	511	6447	7163	2592	926	674	15897	6820													
anti-FOS	-177	2685	39045	22696	4549	-975	4052	1183	-683	-6	17631													
anti-SMAF-2	-2070	1877	1930	-1085	-448	21908	-1262	7927	29607	42287	5227							-274						
anti-SMAF-3	-812	-280	-80	-744	-1135	-915	-802	-37	1143	11089	-95													
anti-NFE2	67	-305	6307	3093	479	10427	45	2958	32984	4472	7008													

	ATF2	cJun	Fos	Fra2	ATF3	ATF4	p21SNFT	MatG	NFE2	NFE2L1	BACH1	anti-ATF2-2	anti-ATF3	anti-ATF3-2	anti-JUN	anti-FOS	anti-SMAF-2	anti-SMAF-3	anti-NFE2	anti-FOS80	anti-FOS200	anti-FOS500	anti-FOS1000	anti-FOS2000
C/EBPα	-0.4	-0.3	-0.2	0.4	0.3	6.3	0.5	-0.6	-0.3	0.1	-0.9	-0.6	0.5	-0.9	-1.1	-1.0	-0.3	0.0	-1.9	-0.1	-0.2	-0.5	-0.6	
C/EBPβ	-0.8	-0.7	-1.2	-1.7	0.9	1.8	0.5	0.8	-1.9	-1.6	-1.2	-1.9	-0.5	0.2	-0.9	-0.2	-1.2	-1.8	1.5	0.9	0.3	-0.4	-0.3	-0.2
C/EBPδ	-1.2	-0.7	-1.0	-0.6	0.9	5.7	0.7	-0.4	-0.2	-0.3	-1.7	-0.4	0.3	0.0	-0.5	-1.0	-0.5	-0.5	-0.3	-1.7	-2.4	-1.0	-1.0	-0.9
C/EBPγ	2.0	0.3	0.0	0.4	4.1	13.4	1.8	0.1	0.0	-0.4	-0.8	6.6	1.4	0.5	0.8	1.6	-0.4	0.6	-0.1	-0.3	2.1	1.5	1.3	1.1
CHOP	2.0	1.0	2.8	3.3	4.7	2.0	7.4	0.8	-0.1	-0.4	1.3	11.1	1.8	1.0	3.9	0.6	-0.6	-0.4	-0.2	-0.1	0.7	1.0	0.9	0.5
ATF-1	-0.7	-1.1	-0.9	-0.5	-1.2	-1.3	-1.3	-0.6	-1.1	-1.3	0.3	-1.7	-0.8	0.7	2.4	-1.2	0.0	-1.3	0.0	2.7	1.8	-0.5	-0.6	-0.7
CREB	-2.3	-1.5	-1.2	-1.5	-1.7	-1.6	-2.1	-1.3	-2.1	-1.5	0.9	-1.8	-0.5	0.8	1.1	-1.3	-1.0	-2.3	-1.4	3.0	1.5	-1.4	-1.5	-1.6
CREB-H	-0.3	-1.0	-0.4	-0.1	-0.8	-0.4	-0.6	-0.5	0.1	-0.3	-0.9	0.0	-0.9	-1.6	-0.3	0.6	-0.4	-0.2	-0.3	-2.1	-0.9	0.6	0.5	0.6
CREB3	-0.2	-0.4	-1.0	-0.6	-0.4	-1.1	-0.2	-0.1	0.0	-0.6	-1.1	0.8	-1.1	0.0	0.5	0.5	0.0	-0.2	0.4	1.3	0.9	0.2	0.6	0.4
ATF-6	-1.4	-1.3	-1.0	-0.8	-1.2	-0.9	-1.3	-1.7	-0.6	-1.2	-0.9	-0.7	-1.1	-0.6	-0.5	-1.2	-1.1	-1.1	-1.1	0.0	-1.1	-1.4	-1.2	-1.0
ZF	-0.2	0.2	-0.6	-0.1	0.0	1.9	0.0	1.8	1.4	0.5	3.2	0.3	-0.2	-0.4	-0.8	1.2	2.7	-0.8	0.1	-0.4	-1.0	-1.0	-1.0	-1.0
XBP-1	-0.7	-0.5	-0.8	-1.8	-0.7	-0.7	-1.3	-0.5	-0.4	-0.7	-0.4	-1.0	-2.7	-0.2	-1.6	-1.2	-2.5	-1.1	-1.0	0.2	0.4	-0.3	-0.4	-0.2
E4BP4	-0.7	-0.6	-0.9	-0.3	-0.5	-1.0	-0.3	0.3	0.0	-0.9	-0.7	1.2	-1.3	-0.6	0.7	-0.6	-0.6	-0.7	-0.6	-0.9	-1.1	-0.5	-0.3	-0.3
ATF-2	2.7	5.1	2.8	3.8	3.7	0.1	1.8	-0.5	-0.7	1.7	1.7	5.3	0.0	-0.4	-0.9	-0.6	-1.0	0.0	2.3	0.3	-0.2	-0.6	-0.5	-0.4
ATF-7	3.7	8.5	3.4	6.1	4.2	1.4	2.8	-0.9	0.8	3.8	2.6	10.8	0.3	-0.7	0.1	-0.1	0.0	0.9	1.4	-1.1	-0.8	-0.2	-0.2	-0.2
cJun	5.7	2.0	14.6	19.8	4.7	-0.5	4.6	-0.6	-0.9	-1														





Supplementary Table 5 Average background-corrected fluorescence values (top panel) and S<sub>array</sub> values (bottom panel) from round 3 of array measurements. Peptides on the surface are in rows, those in solution in columns. Duplicate measurements are marked with a number two in parentheses.

Table with 30 columns for array measurements (e.g., CEBPA, CEBPB, CEBPD) and 30 rows for peptides (e.g., CEBPA, CEBPB, CEBPD). The table contains numerical values for fluorescence and S\_array, with some cells containing numbers in parentheses indicating duplicate measurements.



50. Siggers, T. W. & Honig, B. Structure-based prediction of C2H2 zinc-finger binding specificity: sensitivity to docking geometry. *Nucleic Acids Res* 35, 1085-97 (2007).
51. Paillard, G., Deremble, C. & Lavery, R. Looking into DNA recognition: zinc finger binding specificity. *Nucleic Acids Res* 32, 6673-82 (2004).
52. Morozov, A. V., Havranek, J. J., Baker, D. & Siggia, E. D. Protein-DNA binding specificity predictions with structural models. *Nucleic Acids Res* 33, 5781-98 (2005).
53. Smith, C. A. & Kortemme, T. Backrub-like backbone simulation recapitulates natural protein conformational variability and improves mutant side-chain prediction. *J Mol Biol* 380, 742-56 (2008).
54. Friedland, G. D., Linares, A. J., Smith, C. A. & Kortemme, T. A simple model of backbone flexibility improves modeling of side-chain conformational variability. *J Mol Biol* 380, 757-74 (2008).
55. Aloy, P. & Russell, R. B. Interrogating protein interaction networks through structural biology. *Proc Natl Acad Sci U S A* 99, 5896-901 (2002).
56. Apgar, J. R., Gutwin, K. N. & Keating, A. E. Predicting helix orientation for coiled-coil dimers. *Proteins* 72, 1048-65 (2008).
57. Matys, V. et al. TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res* 31, 374-8 (2003).
58. Spaller, M. R. Act globally, think locally: systems biology addresses the PDZ domain. *ACS Chem Biol* 1, 207-10 (2006).
59. Tonikian, R. et al. A specificity map for the PDZ domain family. *PLoS Biol* 6, e239 (2008).
60. Noyes, M. B. et al. Analysis of homeodomain specificities allows the family-wide prediction of preferred recognition sites. *Cell* 133, 1277-89 (2008).
61. Berger, M. F. et al. Variation in homeodomain DNA binding revealed by high-resolution analysis of sequence preferences. *Cell* 133, 1266-76 (2008).
62. Tarassov, K. et al. An in vivo map of the yeast protein interactome. *Science* 320, 1465-70 (2008).
63. Remy, I. & Michnick, S. W. A highly sensitive protein-protein interaction assay based on Gaussia luciferase. *Nat Methods* 3, 977-9 (2006).
64. Ahn, S. et al. A dominant-negative inhibitor of CREB reveals that it is a general mediator of stimulus-dependent transcription of c-fos. *Mol Cell Biol* 18, 967-77 (1998).
65. McClain, D. L., Gurnon, D. G. & Oakley, M. G. Importance of potential interhelical salt-bridges involving interior residues for coiled-coil stability and quaternary structure. *J Mol Biol* 324, 257-70 (2002).
66. Moitra, J., Szilak, L., Krylov, D. & Vinson, C. Leucine is the most stabilizing aliphatic amino acid in the d position of a dimeric leucine zipper coiled coil. *Biochemistry* 36, 12567-73 (1997).
67. Oakley, M. G. & Kim, P. S. A buried polar interaction can direct the relative orientation of helices in a coiled coil. *Biochemistry* 37, 12603-10 (1998).
68. Gonzalez, L., Jr., Woolfson, D. N. & Alber, T. Buried polar residues and structural specificity in the GCN4 leucine zipper. *Nat Struct Biol* 3, 1011-8 (1996).