Supporting Information for

Simulations of the regulatory ACT domain of human PAH unveil the mechanism of phenylalanine binding

Yunhui Ge[†], Elias Borne[‡], Shannon Stewart[‡], Michael R. Hansen[‡], Emilia C. Arturo[§], Eileen K. Jaffe[‡] and Vincent A. Voelz[†*]

[†]Department of Chemistry, Temple University, Philadelphia, PA 19122
[‡]Fox Chase Cancer Center, Temple University Health System, Philadelphia, PA 19111
[§]Drexel University College of Medicine, Philadelphia, PA 19129

Corresponding Author *voelz@temple.edu

Table of Contents

Supporting Methods

page
Molecular Simulation Methods 3
Simulations of the human PAH ACT domain dimer
Simulations of the human PAH ACT domain monomer
Analysis of structural observables 5
Time-lagged Independent Component Analysis 7
Markov State Model construction 7
MSM of the ACT domain dimer in the presence of free Phe
Transition Path Theory Analysis 9
Estimates of rates and equilibria of free phenylalanine binding to ACT domain residue Phe80 12
Estimates of transition rates between RS-PAH-like and A-PAH-like monomer without free Phe. 18
Estimates of transition rates from a Markov State Model (MSM)
Estimates of transition rates from Transition Path Theory

Supporting Tables

		page
Table S1.	Number of particles and periodic box sizes for ACT domain dimer simulations.	4
Table S2.	Number of particles and periodic box sizes for ACT domain monomer simulations.	4
Table S3.	Summary of trajectory data obtained by distributing computing.	5
Table S4.	Binding times for the 29 *ab initio* binding events observed in the trajectory data.	11
Table S5.	Estimates of binding rates of free phenylalanine (Phe) to residue Phe80 on the ACT domain dimer.	14
Table S6.	Pairwise distances selected to show slow ACT domain dimer motions along tIC1 and tIC2, in the presence of free Phe.	15
Table S7.	Pairwise distances selected to show slow ACT domain dimer motions along tIC1 and tIC2, in the absence of free Phe.	17
Table S8.	Estimates of transition rates between RS-PAH-like and A-PAH-like monomer in the absence of Phe.	19
Table S9.	Estimates of binding rates of free phenylalanine (Phe) to the ACT domain monomer.	19

Supporting Figures

		page
Figure S1	Sequence differences in human versus rat PAH (rPAH) regulatory ACT domain.	3
Figure S2	Twenty-one initial structures of the ACT domain obtained by interpolation.	3
Figure S3	An example of an A-PAH-like monomer conformation.	6
Figure S4	A comparison between atomic partial charges for our zwitterionic Phe model with standard Phe residues in AMBER ff99sb-ildn-nmr.	6
Figure S5	Pairwise distances for residues (shown in cyan) selected for tICA analysis of Phe binding simulations.	7
Figure S6	Implied timescales plots for MSMs of Phe binding to ACT domain dimer.	8
Figure S7	GMRQ results for MSMs of Phe binding to ACT domain dimer.	8
Figure S8	Phe binding trajectory data projected to the 2D tICA landscape, shown with conformational clusters used to define MSM states, and the 80 highest-flux binding pathways calculated using TPT.	10
Figure S9	Phe binding rates predicted by TPT as a function of the number of sink states considered.	10
Figure S10	Posterior distribution of Phe binding rates inferred from observed binding times.	11
Figure S11	Distributions of trajectory lengths for *ab initio* binding simulations.	11
Figure S12	Expected numbers of binding events given a known binding rate k_{on} and the distribution of trajectory lengths.	12
Figure S13	Volume density maps reveal a "hot spot" of binding propensity at residue Phe80 in both ACT domains.	13
Figure S14	A Markov State Model of free Phe binding to F80.	13
Figure S15	Pathways of two trajectories showing free Phe association with both binding sites of the ACT domain dimer.	14
Figure S16	Changes in interresidue distances for the slowest (i.e. most time-correlated) motions associated with Phe binding of the ACT domain dimer.	15
Figure S17	Changes in solvent accessible surface area (SASA) along tIC1 are consistent with opening of a binding gate.	16
Figure S18	Changes in interresidue distances for the slowest (i.e. most time-correlated) motions in simulations of the ACT domain dimer in the absence of free Phe.	17
Figure S19	Examples of ACT domain monomer trajectories that make transitions between RS-PAH-like to A-PAH-like transitions.	18
Figure S20	A trajectory with transitions between A-PAH-like conformations and RS-PAH-like conformations.	19
Figure S21	Adaptive seeding simulations of the ACT domain monomer.	20
Figure S22	Per-residue root-mean-squared fluctuations (RMSF) observed in simulations ACT domain monomer and dimer, in the presence and absence of Phe.	21
Figure S23	Simulations of ACT domain monomer and dimers show similar per-residue root-mean-squared fluctuations, both the presence and absence of Phe.	22
Figure S24	The backbone dihedral angles of Lys74 in the Phe-bound dimer crystal structure (PDB: 5FII) show differences across chains A, B, C, and D.	23
Figure S25	Implied timescale plots for MSMs of the ACT domain monomer.	24
Figure S26	GMRQ plots for MSMs of the ACT domain monomer in the absence of free Phe.	25
Figure S27	Annotated heat maps of adaptive seeding trajectory data for the ACT domain monomer projected to the 2D tICA landscape.	25

Supporting References 26

Molecular Simulation Methods

Simulations of the human PAH ACT domain dimer

Simulations (Figure S1) were prepared from the protein coordinates of the crystal structure of the Phe-bound human PAH ACT domain dimer (PDB: 5FII).[1] The crystal structure contains four chains in the asymmetric unit; chain A has atomic coordinates deposited for 34-109, while chains B, C and D has coordinates for residues coordinates 34-111. Simulations of biological dimers (A+C) and (B+D) were prepared only for available residues, and therefore we ignored residues 100 and 111 in our subsequent analysis. The two dimer crystal poses are virtually identical (0.38 Å root mean square deviation, rmsd-C_α) with the exception of some variation in the hairpin loop near residues 71-72, which are missing for chain C. Loop conformations were reconstructed using the Modeller algorithm.[2] Missing sidechain coordinates were reconstructed using the most probable backbone-dependent rotamer.[3] A series of twenty-one alternative dimer poses were generated using the rigid-body morphing algorithm of Krebs and Gerstein,[4] interpolating between the crystal pose and a previously published homology model of the ACT domain dimer[5] (Figure S2).

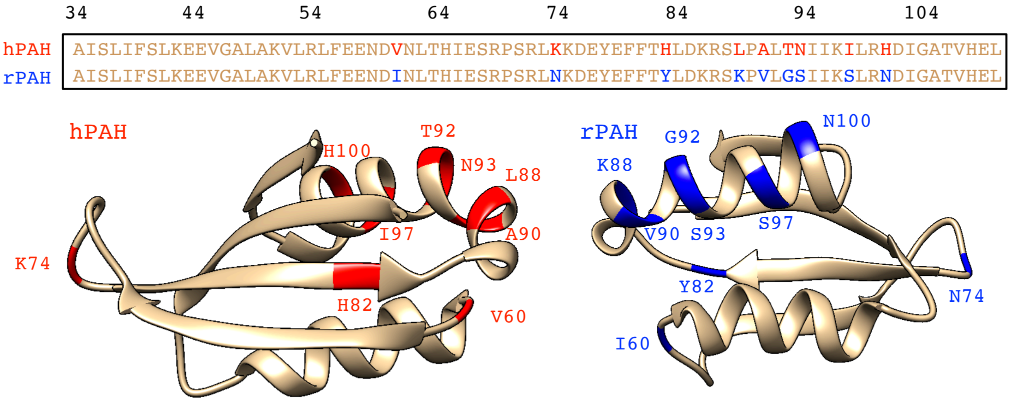
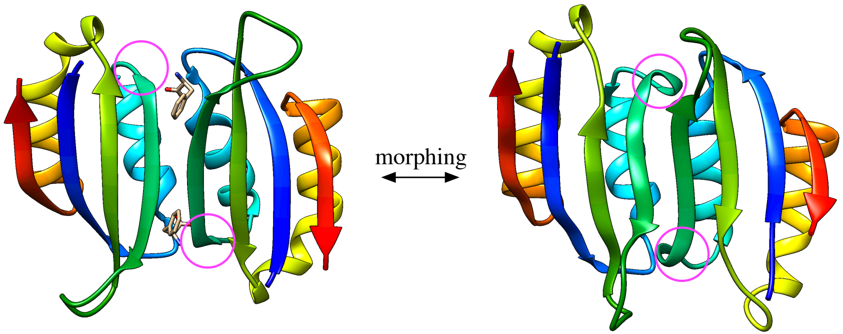

34
44
54
64
74
84
94
104
hPAH
AISLIFSLKEEVGALAKVLRLFEENDVNLTHIESRPSRLKKDEYEFFTHLDKRSLPALTNIIKILRHDIGATVHEL
rPAH
AISLIFSLKEEVGALAKVLRLFEENDINLTHIESRPSRLNKDEYEFFTYLDKRSKPVLGSIIKSLRNDIGATVHEL
hPAH
T92
H100
N93
L88
I97
A90
K74
H82
V60
rPAH
G92
N100
K88
V90
S93
S97
Y82
N74
I60

Figure S1. Sequence differences in human versus rat PAH (rPAH) regulatory ACT domain (residue 34 to 109).



Crystal structure of the human PAH ACT domain dimer PDB:5FII

Homology model of the human PAH ACT domain dimer

Figure S2. Interpolation between two structures was used to generate twenty-one initial structures. The first structure (left) is the crystal structure of the Phe-bound human PAH ACT domain dimer (PDB:5FII). The second structure is a previously published human PAH homology model of an ACT domain dimer built from a ligand-free monomeric form of the rat PAH ACT domain in the RS-PAH conformation.[5] The conformation of each monomer in the homology is similar to that seen in the full length structure of RS-PAH (PDB:5DEN). Key conformational differences (residues 61-64) are circled in magenta.

Simulations of ACT domain dimers were performed according the protocols described in the main text. System sizes and numbers of particles are described in Table S1. Over 346 µs of aggregate trajectory data was generated (Table S3) for three simulation systems: (1) the crystallographic dimer bound by two Phe ligands, (2) the twenty-one dimer poses in the absence of free Phe ligand, and (3) the twenty-one dimer poses in the presence of 19 free (un-bound) Phe, at an effective concentration of 99.5 mM.

Table S1. Number of particles and periodic box sizes for ACT domain dimer simulations.

Folding@home project number	Chains (5FII)	No. atoms	No. Phe Ligands	Na+ atoms	Cl- atoms	No. water molecules	Cubic box length (nm)	Box size (nm^3)
p8617	B,D	35966	*2	22	22	11094	7.147	365.07
p8621	A,C	31308	0	20	19	9570	6.824	317.77
p8662	A,C	31247	**19	20	19	9404	6.819	317.08

*Bound ligands and **free ligands.
p8617: ACT domain dimer with two bound Phe ligands
p8621: *apo* ACT domain dimer
p8662: ACT domain dimer with 19 free Phe ligands

Simulations of the human PAH ACT domain monomer

The same initial twenty-one conformations generated for the dimer simulations were used as starting points for simulations of the ACT domain monomer in the presence and absence of Phe. System sizes and numbers of particles are described in Table S2. Over 286 µs of aggregate trajectory data (Table S3) was simulated for the following systems: (1) ACT domain monomer in the absence of Phe, (2) the twenty-one dimer poses in the presence of 19 free (un-bound) Phe, at an effective concentration of 96.5 mM, and (3) an ACT domain monomer in the crystallographic pose of the dimer (PDB: 5FII), with one Phe ligand "bound" at one of the two binding sites (Figure S3). In practice, we found that simulations initiated with Phe ligands bound at either of the two sites quickly dissociated, eventually becoming what could be considered "free" Phe conditions at a concentration of 5.34 mM.

After obtaining over 30 µs of trajectory data for each ACT domain monomer simulation, a preliminary analysis indicated that spontaneous transitions were occurring between A-PAH-like and RS-PAH-like states, although we didn't have sufficient statistics to accurately estimate the rates of these transitions. Therefore, we performed an adaptive seeding procedure, where by a series of new simulations were initiated from twelve different conformational states along the transition pathway. These initial states were derived from an rmsd-based *k*-centers clustering of the tICA coordinates (see below) from a trajectory that traversed RS-PAH- and A-PAH-like conformations (see an example in Figure S20). From these new simulations, a total of 186.24 µs aggregate trajectory data were collected, about three times more than the previous data set. Figure S21 shows the locations of initial structures of adaptive sampling on the tICA plot generated using both old and new data.

Table S2. Number of particles and periodic box sizes for ACT domain monomer simulations.

Folding@home project number	No. atoms	No. Phe Ligands	Na+ atoms	Cl- atoms	No. water molecules	Cubic box length (nm)	Box size (nm^3)
p13717	30404	0	20	19	9701	6.765	309.60
p13718	30406	*1	20	19	9694	6.774	310.84
p13719	30347	**18	20	19	9544	6.765	309.60
p14041	30404	0	20	19	9701	6.765	309.60
p14042	30347	**18	20	19	9544	6.765	309.60

*Bound ligands and **free ligands.
p13717: ACT domain monomer only
p13718: ACT domain monomer with 1 "bound" Phe ligand
p13719: ACT domain monomer with 18 free Phe ligands
p14041 and p14042: adaptive seeding of p13717 and p13719

Table S3. Summary of trajectory data obtained by distributing computing.

Folding@home project	Number of trajectories	Total simulation time (μs)	Mean trajectory length (ns)	Longest trajectory length (μs)
p8617	466	98.430	211.2	1.074
p8621	446	95.735	214.6	1.316
p8662	484	152.175	314.4	1.196
p13717	351	33.530	95.5	0.380
p13718	373	34.027	91.2	0.425
p13719	357	33.161	92.9	0.355
p14041	2226	91.568	41.1	0.210
p14042	2235	94.669	42.4	0.210

Analysis of structural observables

Calculation of atom distances and solvent accessible surface area (SASA) was performed using the MDTraj python library.[6]

For the dimer simulations, productive binding events were detected by monitoring the average of three atomic distances between free phenylalanine (Phe) and ACT domain residues in the hydrophobic core: (Phe-C)—(Leu48-N), (Phe-C_ζ)—(Ile65-C_β), and (Phe-N)—(Leu62-O). We defined a binding event to occur when the average distance went below the threshold of 0.375 nm.

For the monomer simulations, productive binding events were detected by monitoring the average of four (two for each binding site, see Figure S3) atomic distances between free phenylalanine (Phe) and ACT domain residues in the hydrophobic core. For binding site 1: (Phe-C)—(Leu48-N), (Phe-C_ζ)—(Tyr77-C_β). For binding: (Phe-C_ζ)—(Ile65-C_β), and (Phe-N)—(Leu62-O). We defined a binding event to occur when the average distance was less than then threshold of 0.493 and 0.351 nm, for binding sites 1 and 2, respectively.

Calculation of the root mean-square fluctuation (RMSF) seen in ACT domain dimer and monomer simulations was performed using the MDTraj python library. First each frame of the trajectory was superposed upon the reference (PDB: 5FII). RMSF was calculated by

$$RMSF = \sqrt{\frac{1}{N}\sum_{i}^{N}[(x_i - x^*)^2 + (y_i - y^*)^2 + (z_i - z^*)^2]}$$

where N is the total number of trajectory frames, (x, y, z) are the coordinates of atoms in trajectory frame i, and (x^*, y^*, z^*) are the average atomic coordinates. Only backbone and C_β atoms were selected. In this way, we can assess the fluctuation of each residue and check mobility of different parts of dimer and monomer (check more details about how we prepared these dimer and monomer simulations).

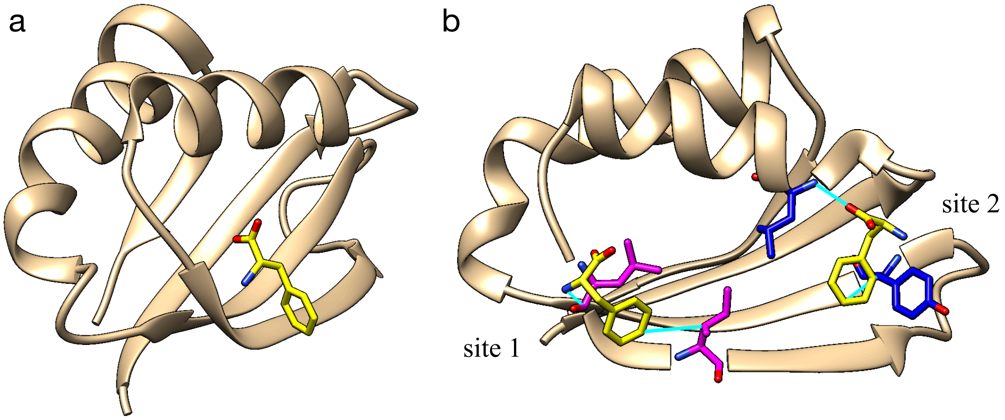


Figure S3. (a) An example of an A-PAH-like monomer conformation with one bound Phe. The conformation shown is from the crystal structure of the Phe-bound human PAH ACT domain dimer (PDB: 5FII). (b) Distances (cyan) used for binding events monitoring in monomer simulations.

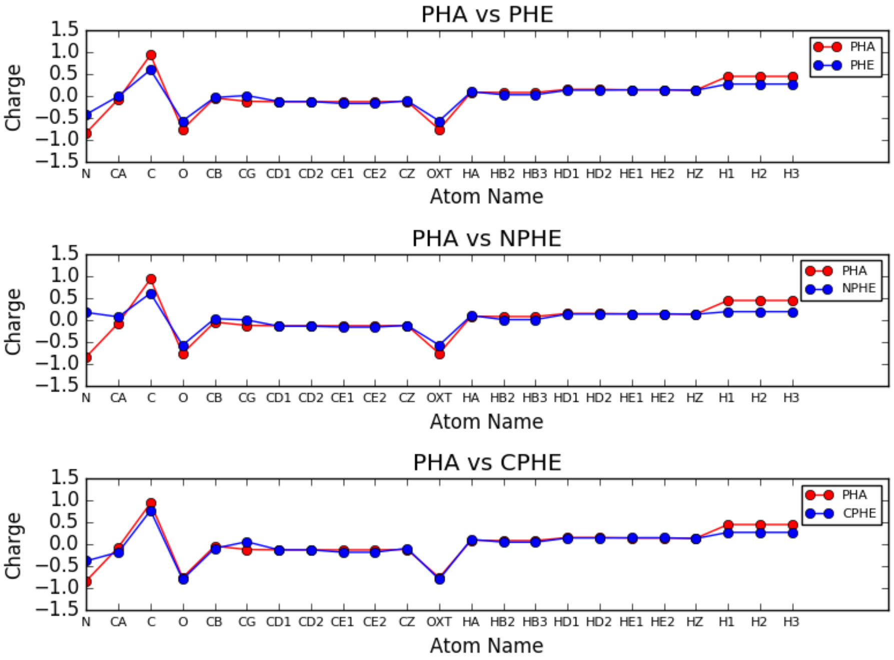


Figure S4. Comparison of atomic partial charges of our PHA residue topology (free Phe with zwitterions) parametrized using antechamber with other Phe residue topologies available in the AMBER ff99sb-ildn-nmr force field: (top) NPHE, the positively charged N terminal Phe residue); (middle) CPHE, the negatively charged C terminal Phe residue; (bottom) PHE, the neutral non-terminal Phe residue.

Time-lagged Independent Component Analysis (tICA)

tICA analysis was performed as described in the main text. To analyze the slowest motions of the unbound ACT domain dimer (see Figure 1h in the main text) we used as structural observables the set of all 11781 pairwise distances for backbone C_α atoms in both ACT domains. To analyze the slowest motions of the unbound monomer, we used the set of all 11175 pairwise distances for all backbone C_α and C_β atoms in monomer.

To analyze the slowest motions associated with ligand binding (see Figure 2 in the main text), we used the set of all 1829 pairwise distances for backbone C_α atoms and sidechain C_β atoms for a selected free Phe ligand and selected residues in both ACT domains surrounding the binding site (Figure S5): residues from domain 1 are Leu41, Lys42, Glu43, Glu44, Val45, Gly46, Ala47, Leu48, Ala49, Ile65, Glu66, Ser67, Arg48, Pro69, Ser70, Arg71, Leu72, Lys73, Lys74, Asp75, Glu76, Tyr77, Glu78; residues from domain 2 are Leu52, Asp59, Val60, Asn61, Leu62, Thr63, His64.

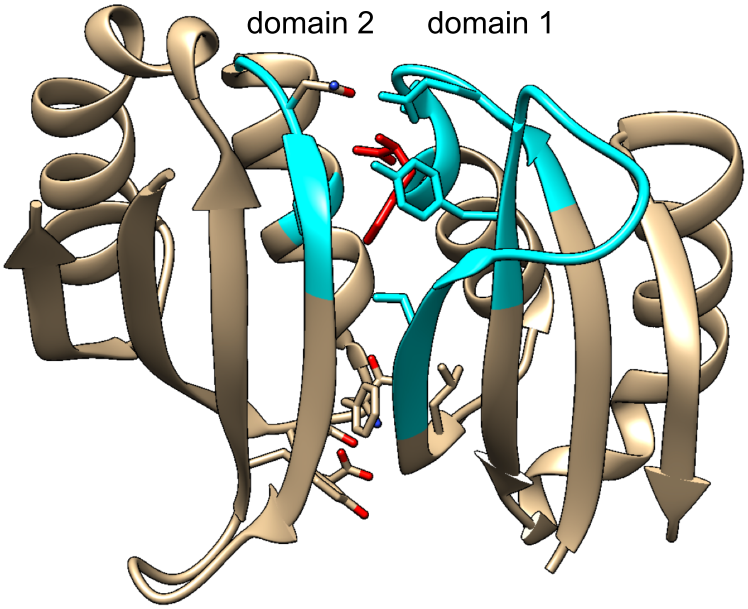


Figure S5. Pairwise distances for residues (shown in cyan) selected for tICA analysis of Phe binding simulations. Shown is the Phe-bound crystal conformation of the dimer (PDB: 5FII), with the Phe ligand in red.

Markov State Model construction

MSM of the ACT domain dimer in the presence of free Phe

MSMs were constructed as described in the main text, with additional details provided here. Implied timescale plots are shown in Figure S6, and GMRQ results are shown in Figure S7. For GMRQ analysis, MSM states were defined by *k*-centers clustering of trajectory data projected onto the 4 largest tICA components (tICs), where the tICA correlation time was 5 ns. An MSM lag time of 20 ns was used. We used five-fold cross-validation, training the model on 4/5 of the data (training score) and computing the GMRQ score using the remaining 1/5 of data (testing score). While the training score continues to increase with the increasing number of states, the test score achieves a maximum at 75 states (marked with a star). The GMRQ results indicate that more than 75 states would lead to an MSM potentially affected by overfitting.

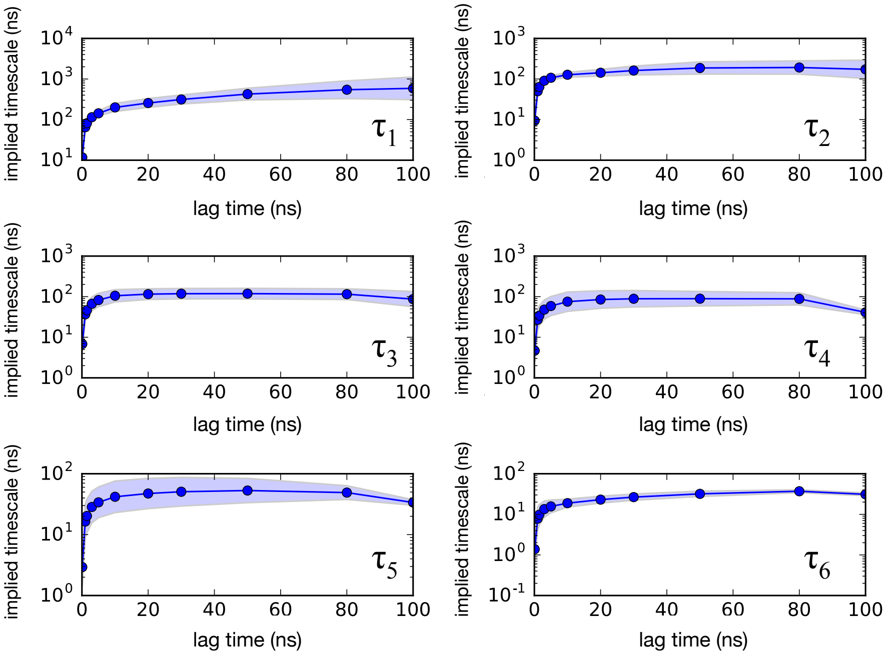


Figure S6. Implied timescales versus MSM lag time for MSMs constructed of Phe binding to the ACT domain dimer (Figure 2 in main text). A bootstrap analysis was performed to explore sensitivities to finite sampling of the six slowest implied timescales (τ_1- τ_6). The error estimates were calculated using a bootstrap procedure, whereby 20 different MSMs were constructed by sampling the input trajectories with replacement.

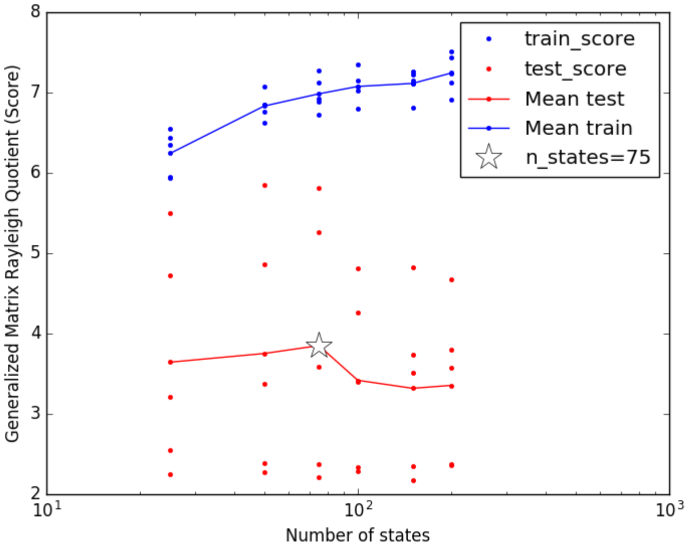


Figure S7. The generalized matrix Rayleigh quotient (GMRQ) method was used to optimize the number of states for constructing an MSM of the ACT domain dimer. Here, other model construction parameters are held fixed (i.e. 4 tICA components, tICA lag time of 5 ns).

Our preliminary MSM results identified an interesting artifact that needed to be filtered from the final data set we analyzed. Inspection of the deposited crystal structure of the Phe-bound ACT domain dimer (PDB:5FII) shows that residue Lys74 in chain B of differs from chains A, C and D; it has a beta-sheet backbone conformation, while the other chains have alpha-helical backbone conformations. Only chain A has coordinates deposited for the entire Lys74 sidechain (the rest have only the C_β atom of the sidechain), indicative of loop flexibility and/or conformational variation, perhaps due to crystal packing artifacts. Simulations of bound-state ACT domain dimer (chains B+D) and unbound ACT domain dimer in the presence of free Phe (chains A+C) show slow interconversion between these backbone states (Figure S24). Therefore, to construct self-consistent MSMs of the binding site, we limited our selection of bound-state trajectories to those with a binding site containing Lys74 from chain D.

We constructed an MSM of the ACT domain monomer from the ~ 91.6-μs dataset of trajectories collected after adaptive seeding. The final MSM consisted of 40 metastable states obtained by *k*-centers clustering of the trajectory data projected to the four largest tICA components. Implied timescales calculated as a function of lag time plateau beyond a lag time of 20 ns, indicating Markovian dynamics (Figure S25). The optimal number of metastable states (40) for MSM construction was determined using the GMRQ variational cross-validation method (Figure S26).

Transition Path Theory Analysis

Transition Path Theory (TPT) analysis was performed as described in the main text, with additional details provided here.

The effective flux between two states *i* and *j* along $A \rightarrow B$ is given by $f_{ij}^{+} = \max(0, f_{ij} - f_{ji})$, where $f_{ij} = \pi_i(1-q_i^{+})T_{ij}\, q_j^{+}$. The total flux through any state *i* is conserved (the total incoming and outgoing flux must be equal), which enables the decomposition of fluxes into specific pathways. Consider a pathway as a sequence of $i \rightarrow j$ edges. We define the pathway of maximum *net flux* as the pathway with the largest "bottleneck flux", i.e. the minimum-f_{ij}^{+} edge. This pathway is not necessarily unique. Regardless, a series of pathways ranked from largest- to smallest-net flux can be selected iteratively, by subtracting the bottleneck flux from all edges in the top-ranked pathway, and repeating the calculation.

To compute Phe binding rates and pathways from our MSM, we chose as our source (*A*) a collection of five unbound MSM states (5, 8, 12, 43, 48) and as the sink (*B*) a collection of 11 ligand-bound MSM states (13, 15, 28, 32, 39, 41, 54, 58, 63, 67, 68) (see Figure S8). Since the source and sink states are subjective, we estimated uncertainty in TPT rate estimates by calculating predicted rates as a function of the number of sink states, and examining the variation across random selections from the 11 sink states (Figure S9). We found that the TPT rate prediction increases as more states are included in the definition, reaching 6.0×10^{7} s^{-1} M^{-1} near 10 or 11 sink states.

To compare the relative flux of pathways involving bent versus non-bent hairpin loop conformations, we computed the net flux for the subset of pathways passing through bent-hairpin intermediate states (24, 39, 49, 57, 67). The bent-hairpin pathways comprise 7.8% of the total binding flux.

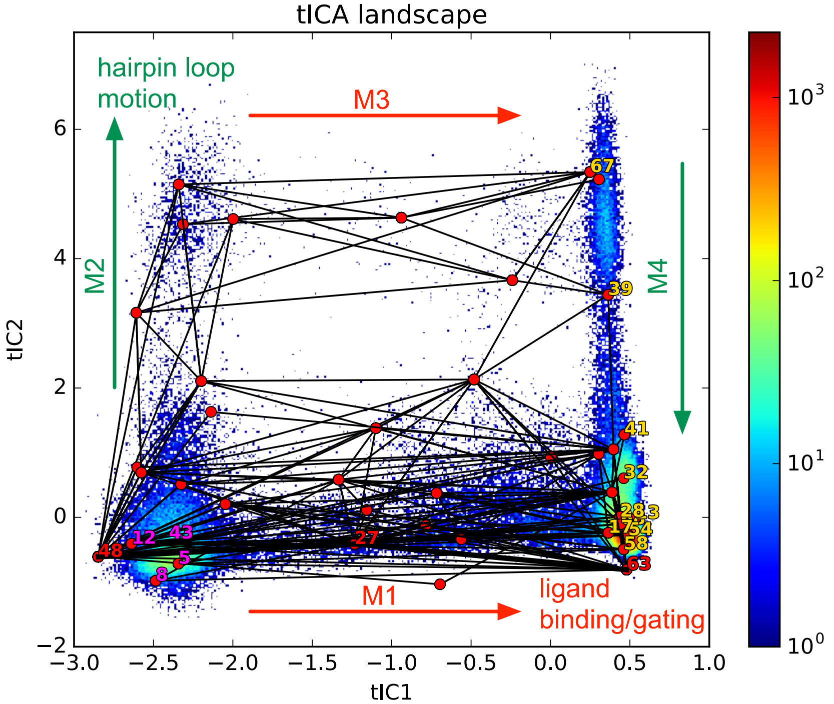


Figure S8. Phe binding trajectory data (28 binding-event trajectories and 29 bound-state trajectories) projected to the 2D tICA landscape, shown with conformational clusters used to define MSM states (red circles). TPT was used to calculate pathway fluxes between an unbound source state (4 states, magenta labels) and a bound sink state (11 states, yellow labels). The 80 highest-flux binding pathways (black lines) fall into two groups: one group of pathways (M1) directly connects unbound and Phe-bound dimers in crystal-like poses (M1); the other group of pathways is indirect (M2→M3→M4), involving the opening (M2) and then closing (M3) of the hairpin loop.

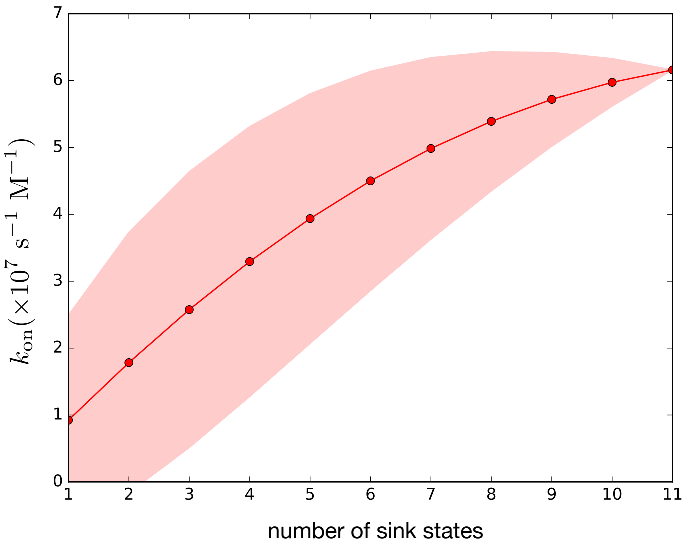


Figure S9. TPT predictions of Phe binding rates as a function of the number of sink states considered. Sink states are defined as bound states which at least one Phe is bound to the dimer. Uncertainties (shaded region) were estimated using a bootstrap procedure, drawing upon the set of 11 sink states.

Table S4. The binding times for the 29 *ab initio* biding events observed in the trajectory data.

Binding times (ns)				
15.1,	50.6,	96.8,	196.8,	267.8,
18.3,	68.5,	154.4,	200.3,	289.3,
25.9,	73.2,	164.4,	215.7,	358.0,
43.2,	76.7,	164.8,	242.9,	396.5,
46.5,	84.2,	178.5,	246.6,	399.4
50.1	89.1,	181.5,	257.2	

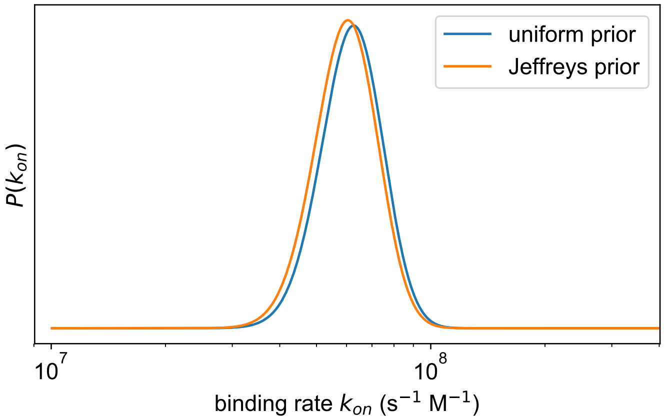


Figure S10. Posterior distribution of Phe binding rates inferred from observed binding times, using (blue) a uniform prior and (red) a Jeffreys prior.

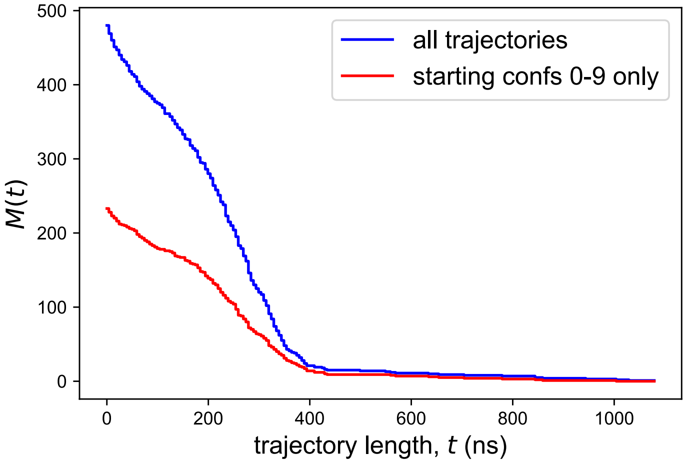


Figure S11. Distributions of trajectory lengths for *ab initio* binding simulations, shown as $M(t)$, the number of trajectories that reach a given length of time t, for (blue) all trajectory data, and (red) a subset of the trajectory data initiated from a crystal-like dimer pose (starting conformations 0 through 9) used for analysis.

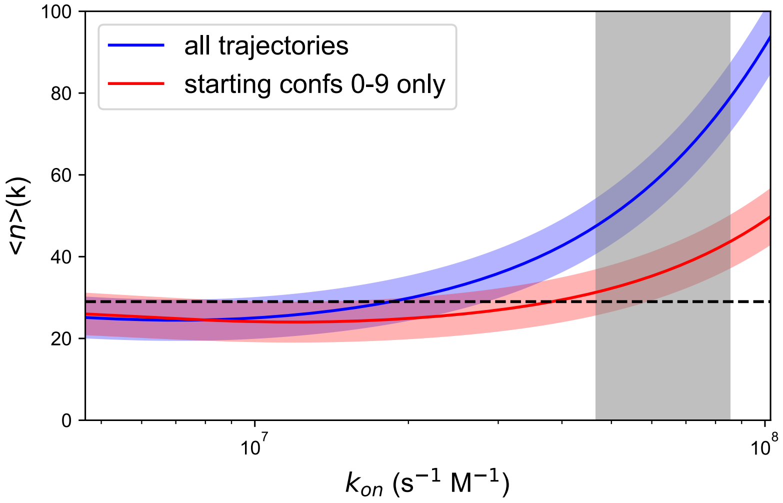


Figure S12. Expected numbers of binding events given a known binding rate k_{on} and the distribution of trajectory lengths. Shown are predictions using all trajectories (blue) and a subset of trajectories starting from crystal-like poses (RUNS 0-9, red), with uncertainties (shaded regions) of ± 5.2 calculated as the standard deviation of a binomial distribution with p=29/480. The dashed line shows the number of binding events (29) observed in the simulations (480 total trajectories).

Estimates of rates and equilibria of free Phe binding to ACT domain residue Phe80

Volume density maps computed for free Phe in *ab initio* binding simulations reveal a "hot spot" of binding propensity at residue Phe80 in both ACT domains (Figure S13a). Inspection of this site shows Phe ligands bound between Phe80 and Arg68 on the surface of the ACT domain beta sheet (Figure S13b). Histograms of (F80-Cγ)-(Phe-Cγ) distances shows a bound population of Phe that can be distinguished from bulk using a 0.7 nm distance threshold (Figure S13c). Trajectory traces show that multiple free Phe molecules can bind within 0.7 nm of Phe80 to the site over time (Figure S13d). Therefore, to estimate binding on- and off-rates, we constructed a five-state Markov State Model (MSM) where each state is defined by the number Phe molecules (0, 1, 2, 3 or 4) bound to Phe80 (Figure S14a). For a series of lag times τ, we compiled a transition count matrix $\mathbf{C}^{(\tau)}$ from the trajectory dataset with elements

$$C_{ij}^{(\tau)} = \Sigma_t\, \chi_i(t)\chi_j(t+\tau),$$

where χ_i and χ_j are state indicator functions. We estimate the transition matrix $\mathbf{T}^{(\tau)}$ as a row-normalized matrix $T_{ij}^{(\tau)} = C_{ij}^{\mathrm{sym}}/\Sigma_j C_{ij}^{\mathrm{sym}}$, where $\mathbf{C}^{\mathrm{sym}} = [\mathbf{C}^{(\tau)} + (\mathbf{C}^{(\tau)})^{\mathrm{T}}]/2$ is a symmetrized count matrix enforcing detailed balance. The implied timescales τ_n are calculated from the eigenvalues μ_n of $\mathbf{T}^{(\tau)}$, as $\tau_n = -\tau/(\ln \mu_n)$. The implied timescales plateau for lag times τ > 50 ns, indicating Markovian dynamics (Figure S14c). We chose a lag time of τ = 200 ns to calculate estimated relaxation rates, and used five-fold partitioning of the trajectory data to compute error estimates.

Equilibrium populations of the five states were estimated from the stationary eigenvector of the transition matrix dynamics (Figure S14b). The equilibrium populations and the eigenvector corresponding to the slowest eigenmode relaxation (Figure S14d) are dominated by 0- and 1-bound states, in accordance with two-state binding.

Assuming unbound and bound populations of (1-p) and p, respectively, (where p is the fractional population of bound states) we estimate binding and unbinding rates of free phenylalanine (Phe) to Phe80 using (1) the fact that the observed two-state binding rate is $k_{\mathrm{obs}} = k_{\mathrm{on}} + k_{\mathrm{off}} = 1/\tau_1$ where τ_1 is the slowest implied timescale, and (2) detailed balance, which requires that $k_{\mathrm{on}}/k_{\mathrm{off}} = p/(1\text{-}p)$, leading to $k_{\mathrm{on}} = p/\tau_1$, $k_{\mathrm{off}} = (1\text{-}p)/\tau_1$. Estimated uncertainties in k_{on}, k_{off} and dissociation constant $K_{\mathrm{D}} = k_{\mathrm{off}}/k_{\mathrm{on}}$ are propagated from uncertainties in τ_1 (Table S5).

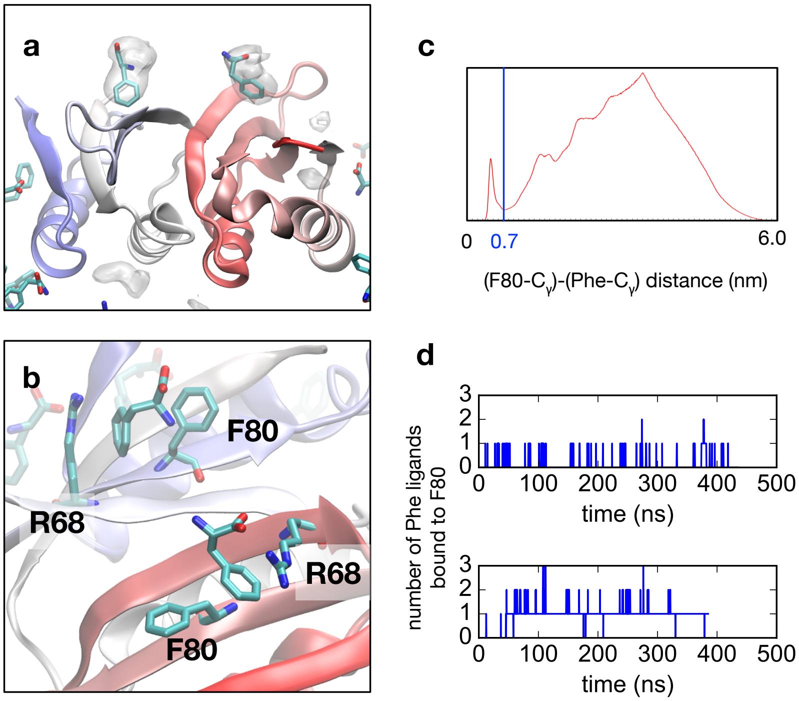


Figure S13. Volume density maps reveal a "hot spot" of binding propensity at residue Phe80 in both ACT domains. (a) Density isosurface showing prevalent binding modes. (b) A typical configuration of free Phe molecules interacting with Phe80 and Arg68 from both ACT domains. (c) Histogram of (F80-Cγ)-(Phe-Cγ) distances observed in the simulation data. (d) Typical trajectory of the occupancy number of free Phe bound to F80 over time.

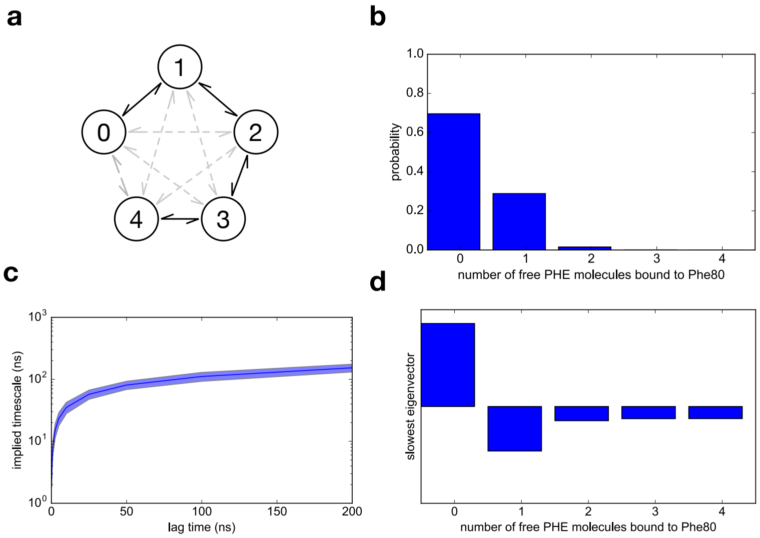


Figure S14. A Markov State Model of free Phe binding to F80. (a) A schematic drawing of a five-state Markov State Model (MSM) where each state is defined by the number Phe molecules (0, 1, 2, 3 or 4) bound to Phe80. (b) Histogram of F80 occupancies. (c) Implied timescale plot of the five-state MSM. (d) The slowest relaxation eigenmode of the MSM.

Table S5. Estimates of binding rates of free phenylalanine (Phe) to residue PHE80 on the ACT domain dimer

simulation	[Phe] (mM)	k_{on} ($\times 10^7$ s^{-1} M^{-1})	k_{off} ($\times 10^7$ s^{-1})	K_D (M)
Phe80, chain 1	99.52	1.9 (1.6–2.2)	0.47 (0.41–0.54)	0.25 (0.18–0.33)
Phe80, chain 2	99.52	1.95 (1.7–2.2)	0.39 (0.35–0.43)	0.20 (0.16–0.25)

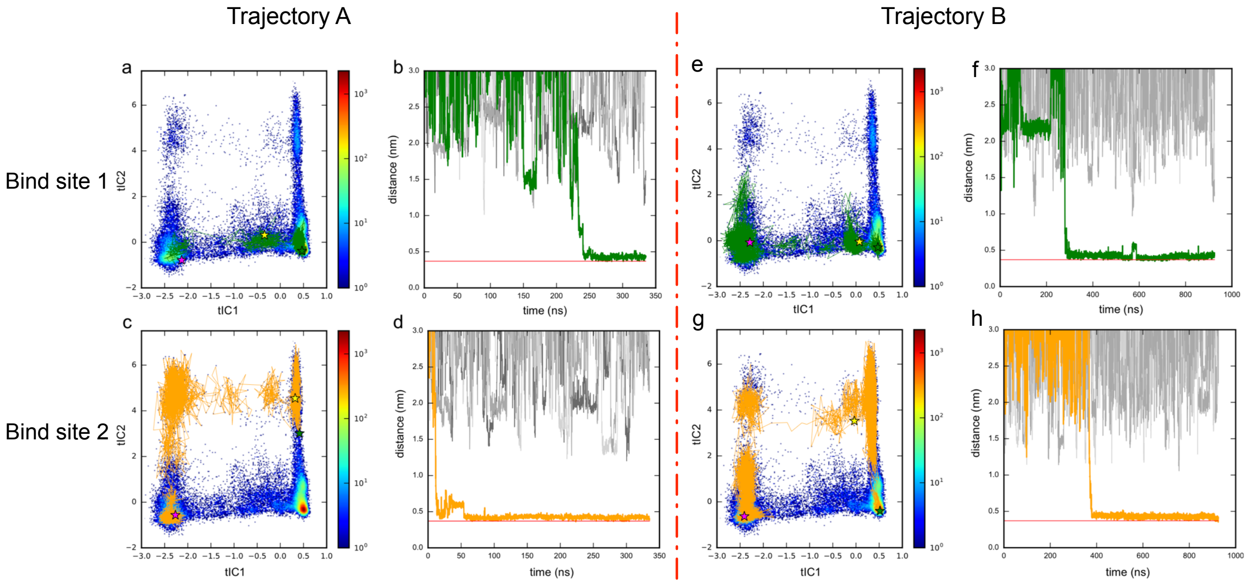


Figure S15. Pathways of two trajectories (A and B) showing free Phe association with both binding sites of the ACT domain dimer. Traces for binding to site 1 (green) and site 2 (orange) are shown projected to the 2D tICA landscape, along with a heatmap of the total trajectory data (a, c, e, and g). Colored stars indicate trajectory starting points (magenta), trajectory end points (green), and ligand binding events (yellow). Corresponding distance traces (b, d, f, and h) are shown for same the Phe binding trajectories (orange, green), along with an example of a non-binding trajectory (gray).

Table S6. Pairwise distances selected to show dimer slow motions along tIC1 and tIC2 (see Figure S16).

tIC1			tIC2		
distance index	Domain 1	Domain 2	distance index	Domain 1	Domain 1
1	Glu44 (Cα)	Asp59 (Cα)	1	Leu72 (Cβ)	Glu78 (Cα)
2	Glu44 (Cα)	Val60 (Cα)	2	Leu72 (Cβ)	Leu41 (Cα)
3	Glu44 (Cα)	Asn61 (Cα)	3	Leu72 (Cβ)	Ser67 (Cα)
4	Glu44 (Cα)	Leu62 (Cα)	4	Glu78 (Cβ)	Arg71 (Cα)
5	Val45 (Cα)	Asp59 (Cα)	5	Leu72 (Cα)	Glu78 (Cβ)
6	Val45 (Cα)	Val60 (Cα)	6	Leu72 (Cβ)	Ile65 (Cα)
7	Val45 (Cα)	Asn61 (Cα)	7	Leu72 (Cα)	Ile65 (Cα)
8	Val45 (Cα)	Leu62 (Cα)			

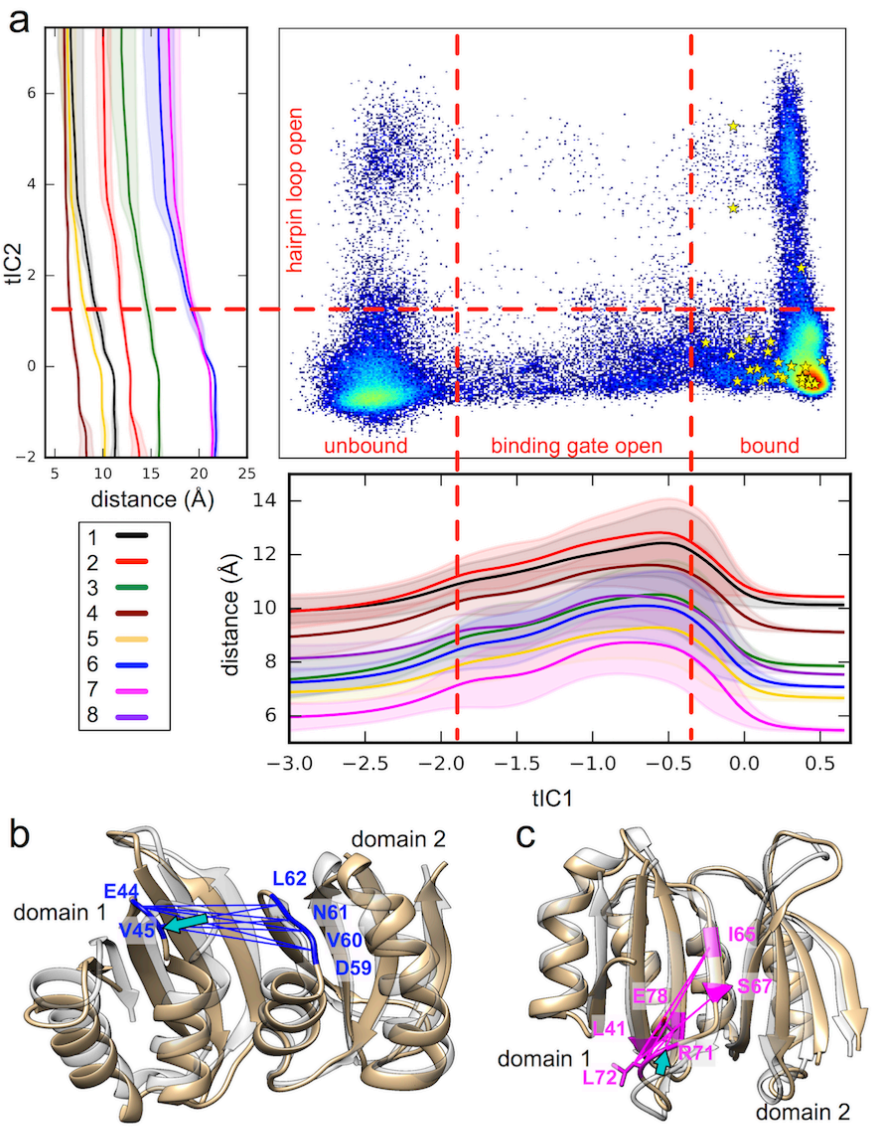


Figure S16. Changes in inter-residue distances for the slowest (i.e. most time-correlated) motions associated with Phe binding of the ACT domain dimer. The slowest motion involves the "binding gate" motion along tIC1. Inter-residue distances that change greatly during these motions are shown at the bottom of panel (a) and in panel (b) (blue lines). The next-slowest motions involve the hairpin loop (Leu72) shown in the left of panel (a) and in panel (c) (magenta lines). Table S8 contains the complete list of inter-residue distances shown.

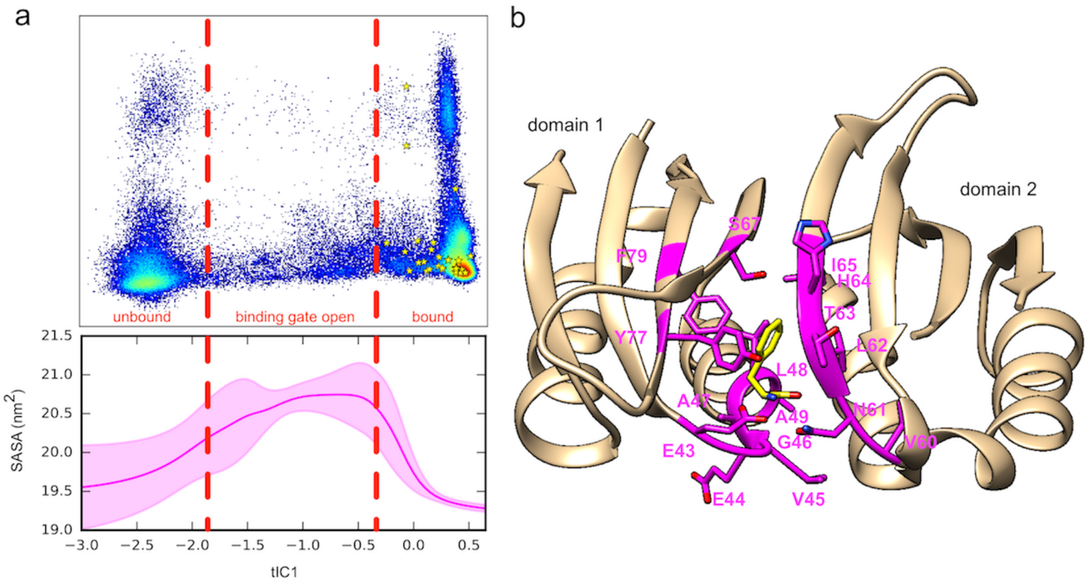


Figure S17. Changes in solvent accessible surface area (SASA) along tIC1 are consistent with opening of a binding gate. Residues (shown in magenta) within 5 Å from the ligand (shown in yellow) were selected for SASA calculation. Residues from domain 1: Glu43, Glu44, Val45, Gly46, Ala47, Leu48, Ala49, Ser67, Tyr77, Phe79; residues from domain 2: Val60, Asn61, Leu62, Thr63, His64, Ile65.

Table S7. Pairwise distances selected to show slow motions of the ACT domain (in the absence of free Phe) along tIC1 and tIC2 (see Figure S18).

	tIC1			tIC2	
distance index	Domain 1	Domain 2	distance index	Domain 1	Domain 1
1	Val45 (Cα)	Asp59 (Cβ)	1	Ala47 (Cβ)	Thr63 (Cα)
2	Val45 (Cα)	Asp59 (Cα)	2	Ala47 (Cβ)	Leu62 (Cβ)
3	Val45 (Cα)	Leu52 (Cα)	3	Gly46 (Cα)	Val60 (Cβ)
4	Val45 (Cβ)	Val60 (Cα)	4	Ala47 (Cα)	Lys42 (Cα) (Domain 1)
5	Val45 (Cβ)	Asp59 (Cα)	5	Ala47 (Cα)	Asn61 (Cα)
6	Val45 (Cα)	Val60 (Cα)	6	Ala47 (Cβ)	Thr63 (Cβ)
7	Val45 (Cα)	Leu48 (Cα) (Domain 1)	7	Ala47 (Cα)	Tyr77 (Cβ) (Domain 1)
8	Val45 (Cβ)	Asp59 (Cβ)	8	Ala47 (Cα)	Glu44 (Cβ) (Domain 1)

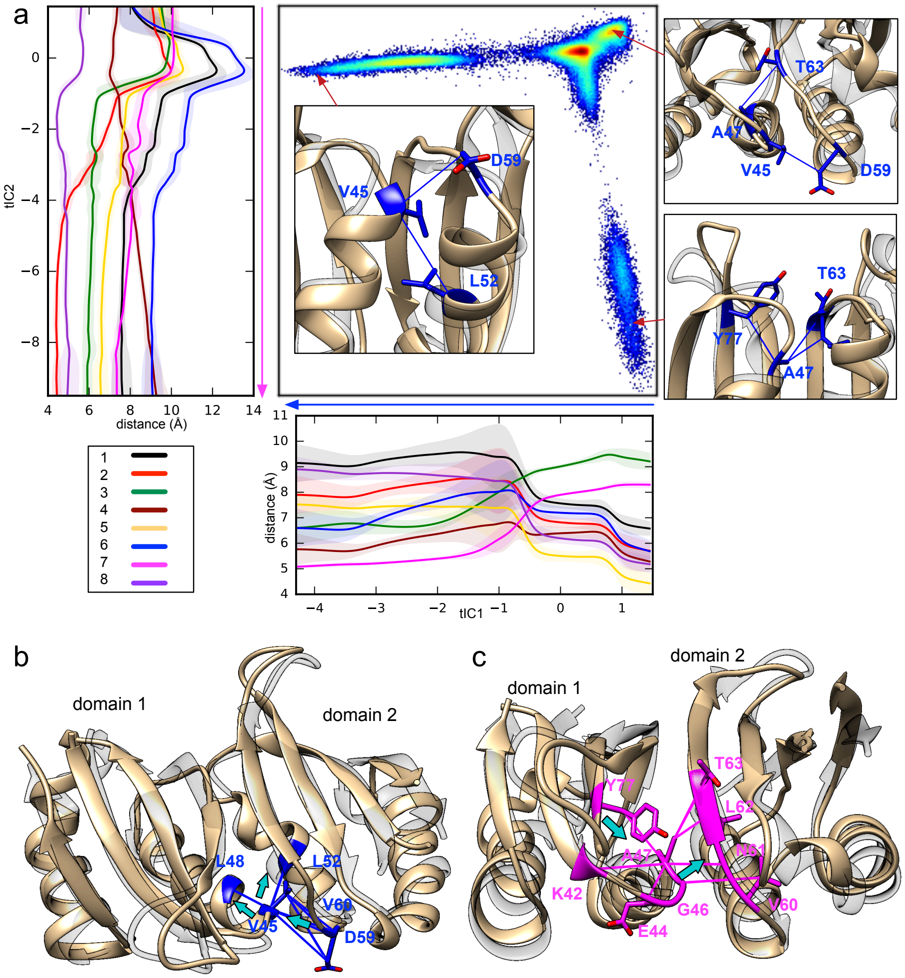


Figure S18. Changes in interresidue distances for the slowest (i.e. most time-correlated) motions in simulations of the ACT domain dimer in the absence of free Phe. (a) Selected pairwise distances (see Table S9) change greatly along tIC1 (blue) and tIC2 (magenta). (b) The slowest motion along tIC1 involves Val45 (domain 1) moving closer (cyan arrows) to Leu48 and Leu52. (c) The second-slowest motion along tIC2 involves residues 60-63 on domain 2 moving closer (cyan arrows) to Ala47. Ribbon structures (tan) show conformations belong to selected states on the tICA landscape, superimposed the crystal structure of the Phe-bound ACT domain dimer (transparent grey, PDB: 5FII). Only a subset of trajectory data (those initiated from the five starting structures most similar to the crystal structure) were used in this analysis.

Estimates of transition rates between RS-PAH-like and A-PAH-like monomer in the absence of free Phe

Estimates of transition rates from a Markov State Model (MSM)

The slowest MSM implied timescale τ_1 = (4.2 +/- 2.1 μs) corresponds to the transition between RS-PAH-like and A-PAH-like monomer (see Figure 4 in the main text). According to a two-state kinetic model, the observed relaxation rate k_{obs} is related to the transitions rates between RS-PAH-like and A-PAH-like states by

$$k_{\mathrm{obs}} = 1/\tau_1 = k_{\mathrm{RS}\rightarrow\mathrm{A}} + k_{\mathrm{A}\rightarrow\mathrm{RS}}.$$

Using the equilibrium MSM populations π_{RS} and π_{A}, and detailed balance, the rates are determined as $k_{\mathrm{RS}\rightarrow\mathrm{A}} = \pi_{\mathrm{A}}/\tau_1$, $k_{\mathrm{A}\rightarrow\mathrm{RS}} = \pi_{\mathrm{RS}}/\tau_1$. Uncertainty estimates come from a bootstrap procedure for standard error in [ln τ_1] (shaded region in Figure S25). See Figure S27 for our selection of RS-PAH-like and A-PAH-like states.

Estimates of transition rates from Transition Path Theory

Similar to what we've done for binding rate estimation using TPT analysis, we chose as our source (*A*) a collection of seven RS-PAH-like MSM states (4, 9, 16, 20, 22, 29, 35) and as the sink (*B*) a collection of four A-PAH-like MSM states (0, 12, 28, 33) (see Figure S27). As we mentioned above, the source and sink states are subjective and the predicted rates are dependent of random selections. Therefore, a bootstrap procedure is performed, whereby randomly picked up N_{RS} and N_{A} states (N_{RS} = 1, 2, …, 7 and N_{A} = 1, 2, …, 4) without replacement as sink and source states and did the rate estimation. 20 rounds such calculation was performed. The results are shown in Table S8.

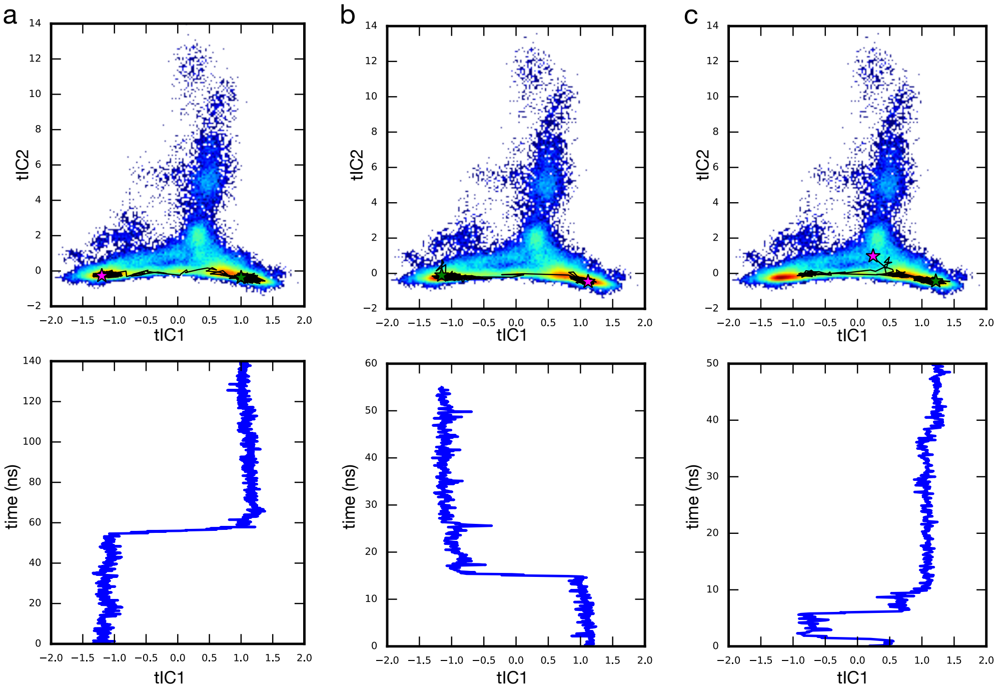


Figure S19. Examples of ACT domain monomer trajectories (in the presence of free Phe) observed to make RS-PAH-like to A-PAH-like transitions. Top panels show trajectory traces in the 2D tICA projection, along with a heat map of the total trajectory data. Magenta and green stars represent the starting and end points of the trajectories, respectively. Bottom panels show that the variation of tIC1 of each trajectory over time.

Table S8. Estimates of transition rates between RS-PAH like monomer and A-PAH like monomer

Method	$\log_{10}(k_{RS\rightarrow A}$ *)	$\log_{10}(k_{A\rightarrow RS}$ **)
MSM implied timescales	5.22 ± 0.21	4.84 ± 0.21
Transition Path Theory	5.59 ± 0.57	5.40 ± 0.27

*the rate of transition from RS-PAH-like to A-PAH-like monomer
**the rate of transition from A-PAH-like to RS-PAH-like monomer

Table S9. Estimates of binding rates of free phenylalanine (Phe) to the ACT domain monomer

simulation	[Phe] (mM)	k_{on} (× 10^7 s^{-1} M^{-1})	k_{off} (× 10^7 s^{-1})	K_D (M)
bound monomer, site 1	5.34	47.7 (30.1–75.5)	14.2 (8.9–22.4)	0.30 (0.11–0.75)
bound monomer, site 2	5.34	6.6 (3.9–10.9)	5.9 (3.5–9.9)	0.90 (0.32–2.5)
unbound monomer, site 1	96.5	16.8 (15.8–17.8)	26.4 (24.8–28.0)	1.6 (1.3–1.8)
unbound monomer, site 2	96.5	2.9 (1.4–6.0)	9.0 (4.3–18.7)	3.1 (0.7–13)

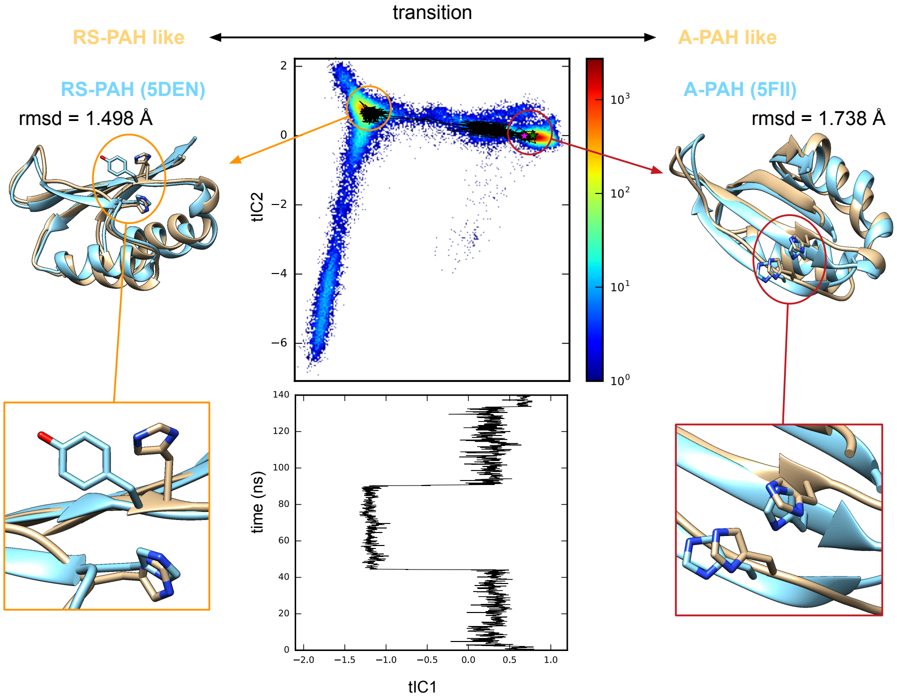


Figure S20. A trajectory with transitions between A-PAH-like conformations and RS-PAH-like conformations (see main text for definition). Colored stars represent the start (magenta) and end (green) of the trajectory. (left) Selected RS-PAH-like snapshots of the ACT domain monomer taken from the trajectory data (tan) are superposed on the RS-PAH crystal structure (PDB: 5DEN). (right) Selected A-PAH-like snapshots of the ACT domain monomer taken from the trajectory data (tan) are superposed on the crystal structure of the Phe-bound ACT domain dimer (PDB: 5DEN). In each case, backbone-rmsd values to the crystal structures are shown.

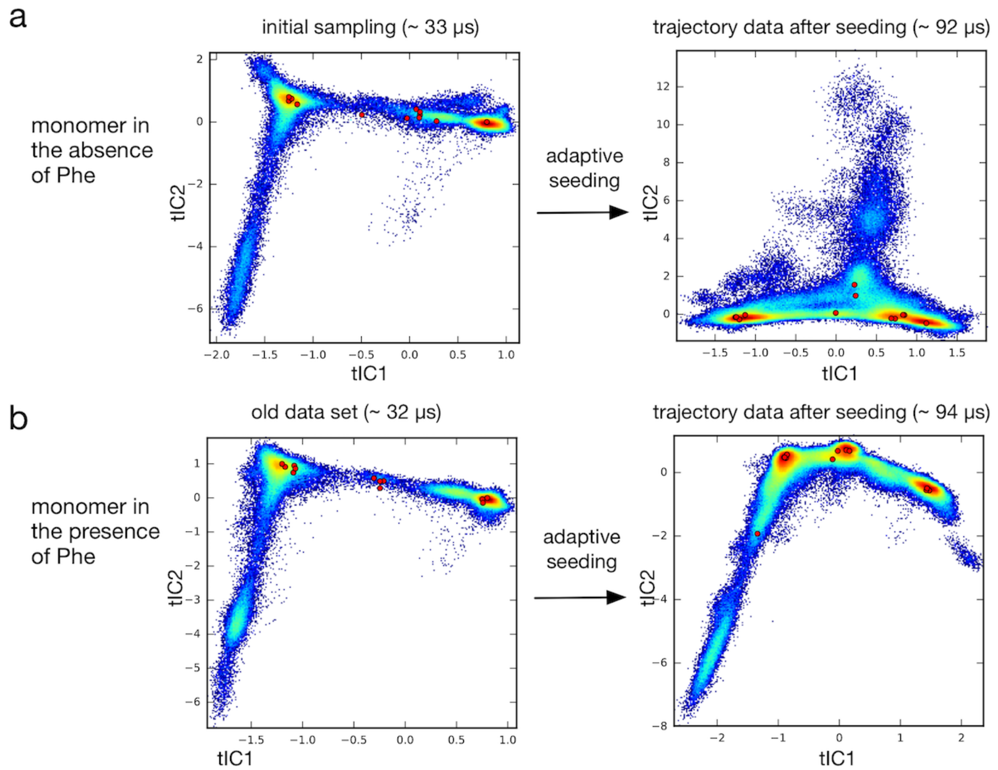


Figure S21. Adaptive seeding simulations of the ACT domain monomer. Trajectory data for ACT domain monomers projected to the 2D tICA landscape, shown with the locations (red circles) of initial structures used for seeding. Results are shown for simulations performed in absence of Phe (a) and in the presence of Phe (b). Initial structures were obtained through rmsd-based *k*-centers clustering from the initial sampling.

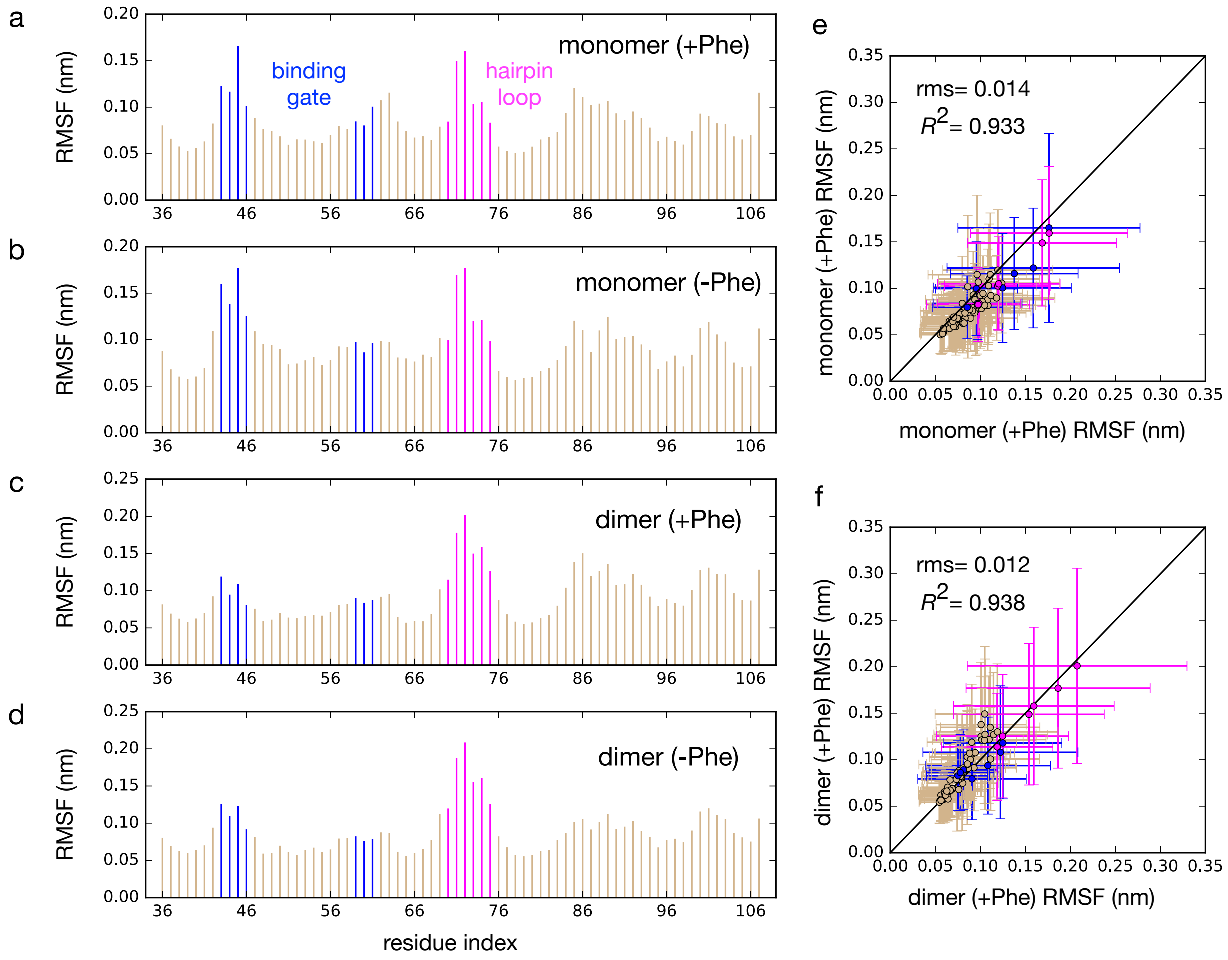


Figure S22. Per-residue root-mean-squared fluctuations (RMSF) observed in simulations ACT domain monomer and dimer, in the presence and absence of Phe. (a) RMSF profile of ACT domain monomer in the presence of free Phe. (b) RMSF profile of ACT domain monomer in the absence of free Phe. (c) RMSF profile of ACT domain dimer (only one chain shown) in the presence of free Phe. (d) RMSF profile of ACT domain dimer (only one chain shown) in the absence of free Phe. (e) Comparison on monomer RMSFs in the presence and absence of free Phe. (f) Comparison on dimer RMSFs in the presence and absence of free Phe. Residues corresponding to the hairpin loop region are shown in pink; binding gate residues are shown in blue.

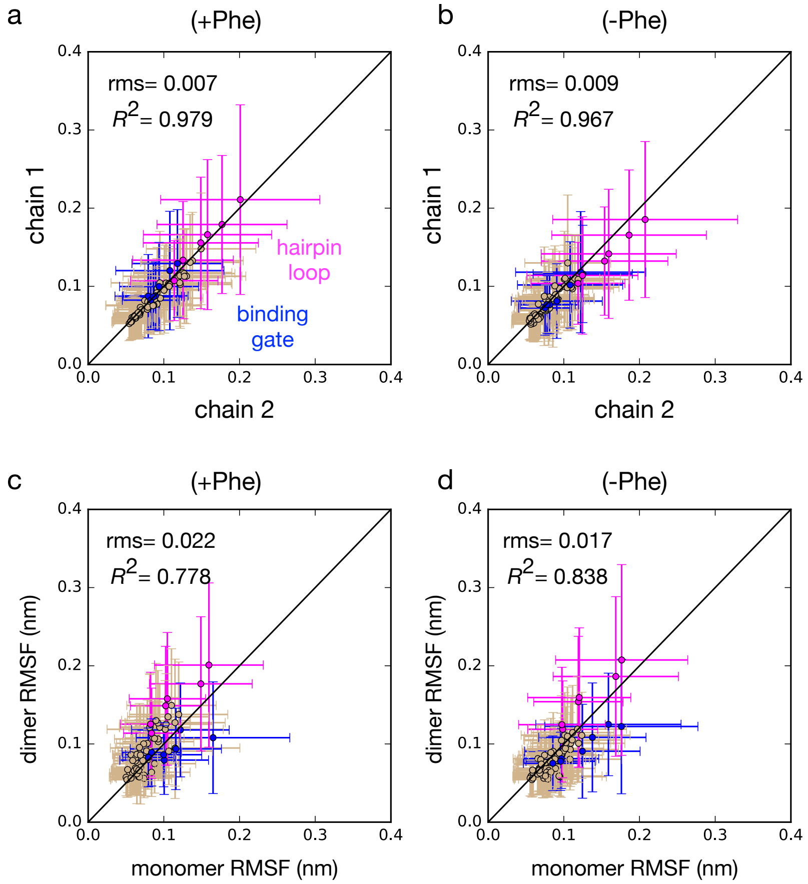


Figure S23. Comparison between ACT domain monomer and dimer per-residue RMSFs. (a) and (b): Comparison of per-residue RMSFs for chain 1 and chain 2 of the ACT domain dimer, in the presence and absence of free Phe. (c) and (d): Comparison of per-residue RMSFs for ACT domain monomer and dimer simulations, in the presence and absence of free Phe.

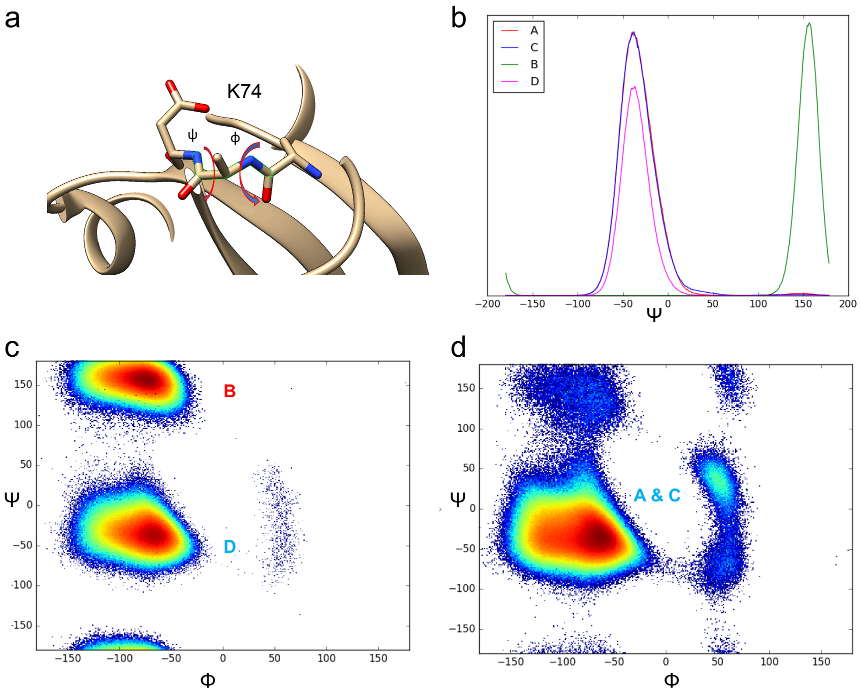


Figure S24. The backbone dihedral angles of Lys74 in the Phe-bound dimer crystal structure (PDB: 5FII) show differences across chains A, B, C, and D. (a) A visualization of Lys74 (sidechain is incomplete) and its corresponding dihedral angles (ψ, φ). (b) Distributions of ψ-angle values for AC and BD dimer simulations show Lys74 on chain B (green) to occupy a β-sheet backbone conformation, different from other three chains. (c) Backbone dihedral distributions for Lys74 in the BD dimer. (d) Backbone dihedral distributions for Lys74 in the AC dimer.

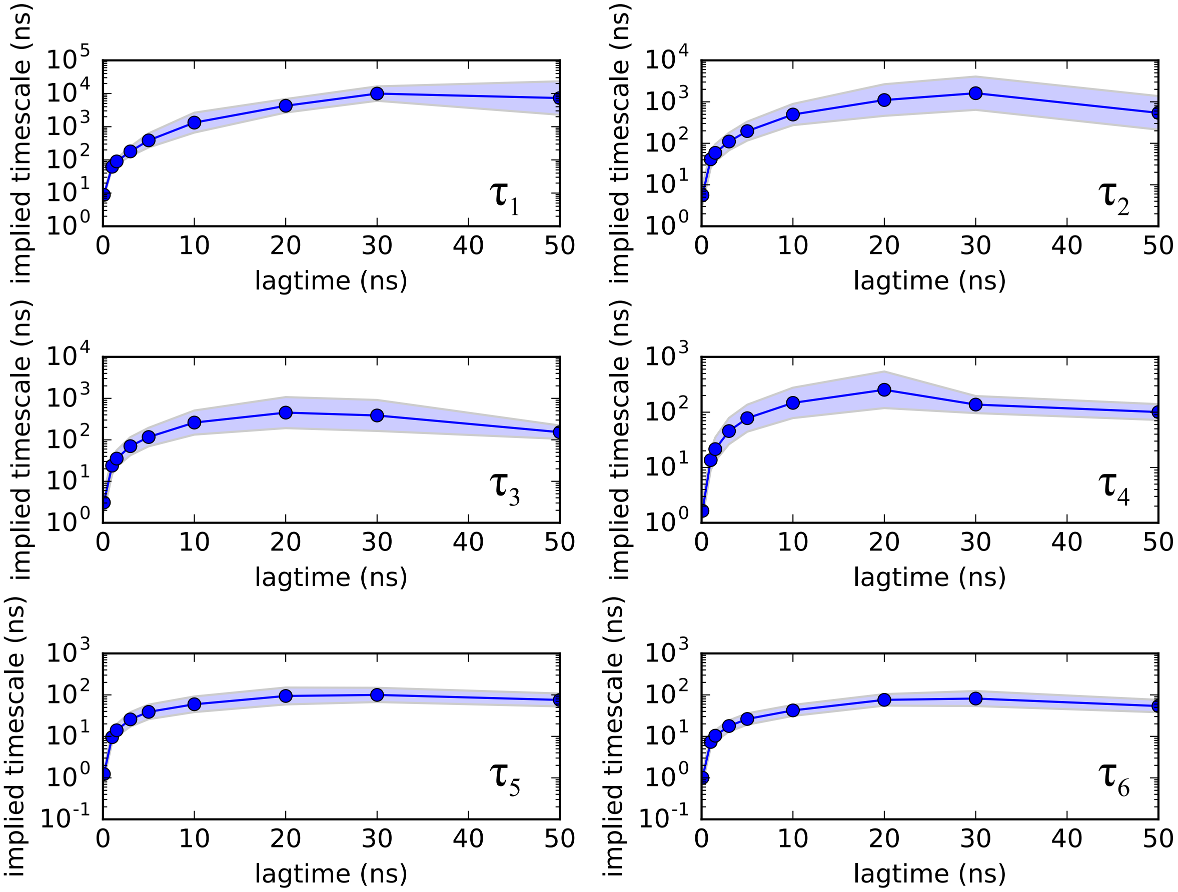


Figure S25. Implied timescale plots for MSMs of the ACT domain monomer in the absence of free Phe. Separate panels show implied timescale plot for the six slowest implied timescales (τ_1- τ_6). Error estimates were calculated using a 10-fold bootstrap procedure, sampling the 2226 input trajectories with replacement.

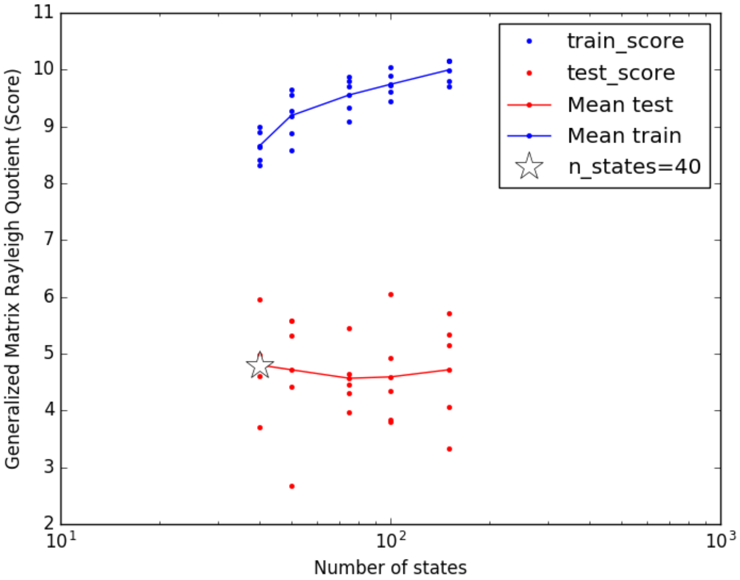


Figure S26. GMRQ plots for MSMs of the ACT domain monomer in the absence of free Phe.

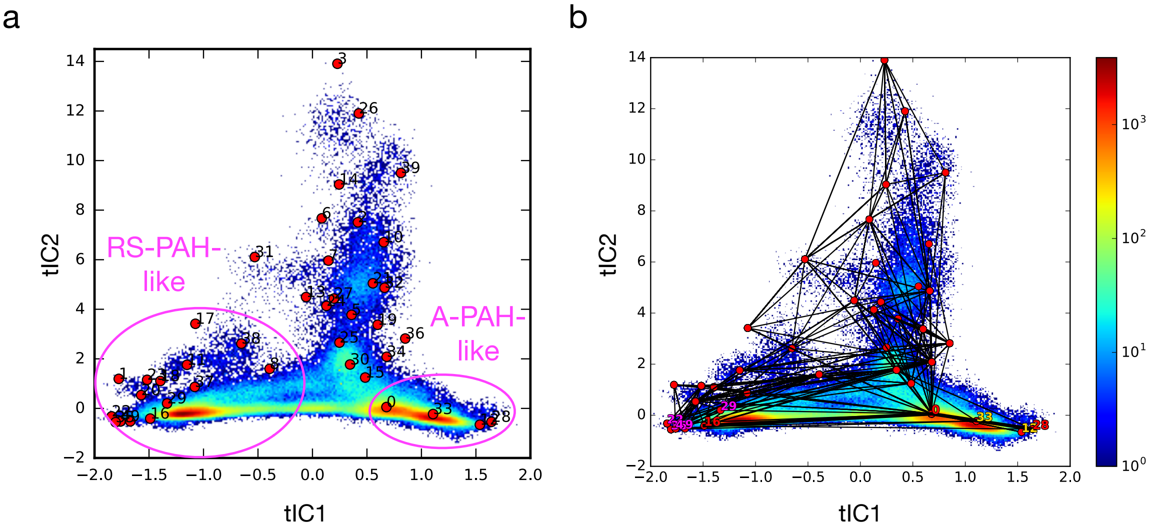


Figure S27. Annotated heat maps of adaptive seeding trajectory data for the ACT domain monomer projected to the 2D tICA landscape. Red circles denote the conformational cluster centers corresponding to the 40 metastable states of the MSM. (a) RS-PAH-like and A-PAH-like (source and sink) states selected for TPT analysis. (b) The highest-flux pathways predicted by TPT. The highest-flux transition pathway (bold black) is between state 16 and state 0.

Supporting References

1. Patel, D.; Kopec, J.; Fitzpatrick, F.; McCorvie, T. J.; Yue, W. W., Structural basis for ligand-dependent dimerization of phenylalanine hydroxylase regulatory domain. *Sci. Rep.* **2016**, 6, 23748.
2. Webb, B.; Sali, A. Comparative Protein Structure Modeling Using Modeller. In *Current Protocols in Bioinformatics*; John Wiley & Sons, Inc.: 2014; Chapter 5.6.1-5.6.32.
3. Dunbrack, R. L., Rotamer libraries in the 21(st) century. *Current opinion in structural biology* **2002**, 12, 431-440.
4. Krebs, W. G.; Gerstein, M., The morph server: a standardized system for analyzing and visualizing macromolecular motions in a database framework. *Nucleic Acids Res* **2000**, 28, 1665-75.
5. Jaffe, E. K.; Stith, L.; Lawrence, S. H.; Andrake, M.; Dunbrack Jr, R. L., A new model for allosteric regulation of phenylalanine hydroxylase: Implications for disease and therapeutics. *Archives of Biochemistry and Biophysics* **2013**, 530, 73-82.
6. McGibbon, R. T.; Beauchamp, K. A.; Harrigan, M. P.; Klein, C.; Swails, J. M.; Hernández, C. X.; Schwantes, C. R.; Wang, L.-P.; Lane, T. J.; Pande, V. S., MDTraj: A Modern Open Library for the Analysis of Molecular Dynamics Trajectories. *Biophys. J.* **2015**, 109, 1528-1532.