SUPPORTING INFORMATION for
Quantifying the sources of kinetic frustration in folding simulations of small proteins

Andrej J. Savol and Chakra S. Chennubhotla
Dept. of Computational and Systems Biology, University of Pittsburgh School of Medicine
3501 Fifth Ave., Ste. 3064 BST3, Pittsburgh, PA, 15260, USA

Table S1. Correlation, ρ, between $\bar{f}_{\mathrm{nat}}$ and five structural parameters compared with bias values, β. Correlation values are weighted according to substate populations [2], whereas frustration biases are non-weighted. Statistically significant β values are indicated in bold ($p < 0.005$ according to permutation test). Cf. Fig. 5, main text.

	TRP	BBA	VHP	WW	BBL	PB	HMDM	PG	A3D	LAMDA
$\rho_{\bar{f}_{\mathrm{nat}},\mathsf{RMSD}}$	-0.00	-0.36	-0.02	-0.29	-0.14	0.02	-0.46	-0.08	-0.30	-0.35
β_{RMSD}	0.01	**-0.29**	0.04	**-0.32**	**-0.21**	0.01	**-0.30**	**-0.18**	**-0.28**	**-0.28**
$\rho_{\bar{f}_{\mathrm{nat}},H_{\mathrm{nn}}}$	0.02	-0.11	0.05	-0.08	0.01	0.04	-0.04	-0.08	-0.09	-0.21
$\beta_{H_{\mathrm{nn}}}$	-0.02	**-0.13**	0.14	-0.09	-0.09	-0.03	-0.03	**-0.19**	**-0.12**	**-0.10**
$\rho_{\bar{f}_{\mathrm{nat}},H_{\mathrm{n}}}$	-0.00	0.46	0.03	0.40	0.01	-0.21	0.06	-0.01	0.14	-0.08
$\beta_{H_{\mathrm{n}}}$	0.05	**0.39**	-0.05	**0.39**	**0.11**	-0.00	0.01	**0.13**	**0.14**	-0.07
$\rho_{\bar{f}_{\mathrm{nat}},Q_{\mathrm{nn}}}$	-0.45	-0.16	-0.07	-0.01	-0.31	-0.24	-0.09	-0.14	-0.14	-0.34
$\beta_{Q_{\mathrm{nn}}}$	**-0.33**	**-0.13**	-0.03	0.00	**-0.18**	-0.17	-0.05	**-0.15**	**-0.13**	**-0.28**
$\rho_{\bar{f}_{\mathrm{nat}},Q_{\mathrm{n}}}$	-0.21	0.17	0.13	0.56	-0.01	-0.22	0.30	0.06	0.23	0.16
$\beta_{Q_{\mathrm{n}}}$	**-0.16**	**0.15**	0.08	**0.58**	**0.16**	-0.04	**0.21**	**0.17**	**0.20**	**0.16**

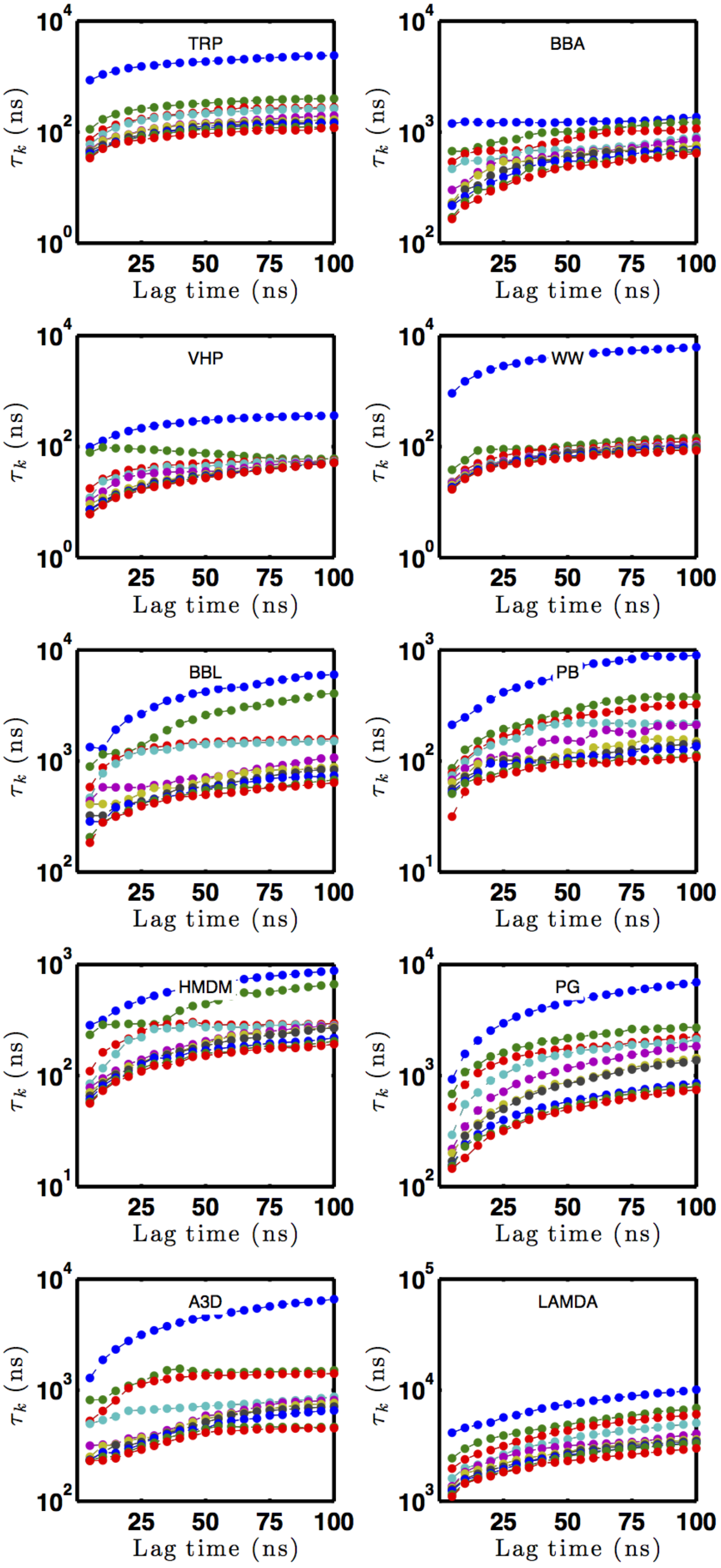

Figure S1. **Implied Timescales**. Ten slowest implied timescales as a function of lag time computed from MSMs constructed for each aggregate simulation.

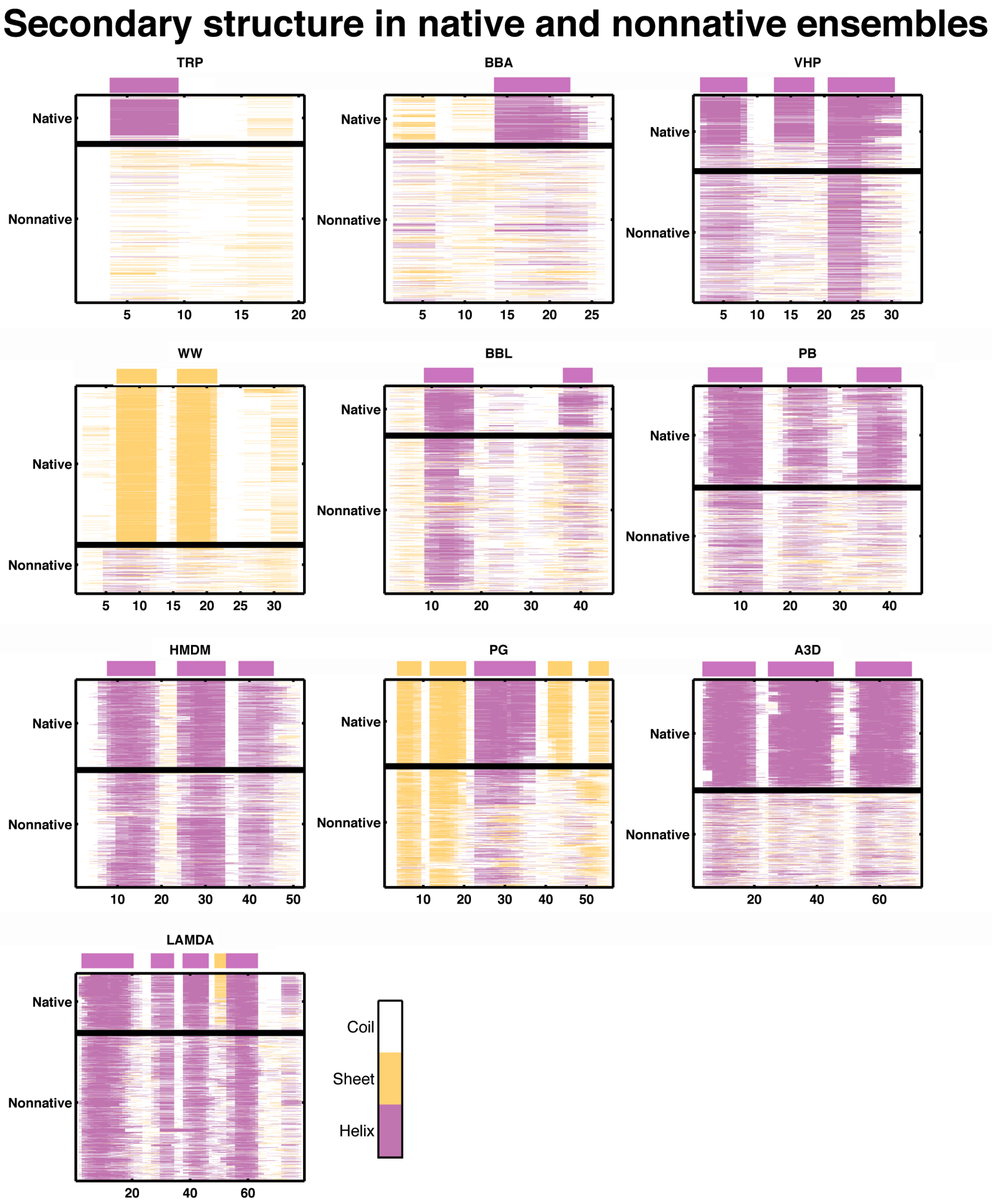

Figure S2. Secondary structure is shown for every trajectory frame used in the clustering and subsequent analysis, grouped according to substate. Frames classified as belonging to the nonnative ensemble are shown in panels' lower portions; native conformers are in upper divisions. The *structure sequence* (see Methods, main text) is depicted above the upper y-axis. Lower abscissa labels denote residue indices for each peptide. Residue-wise secondary structure assignment was performed on C_α coordinates with P-SEA [1].

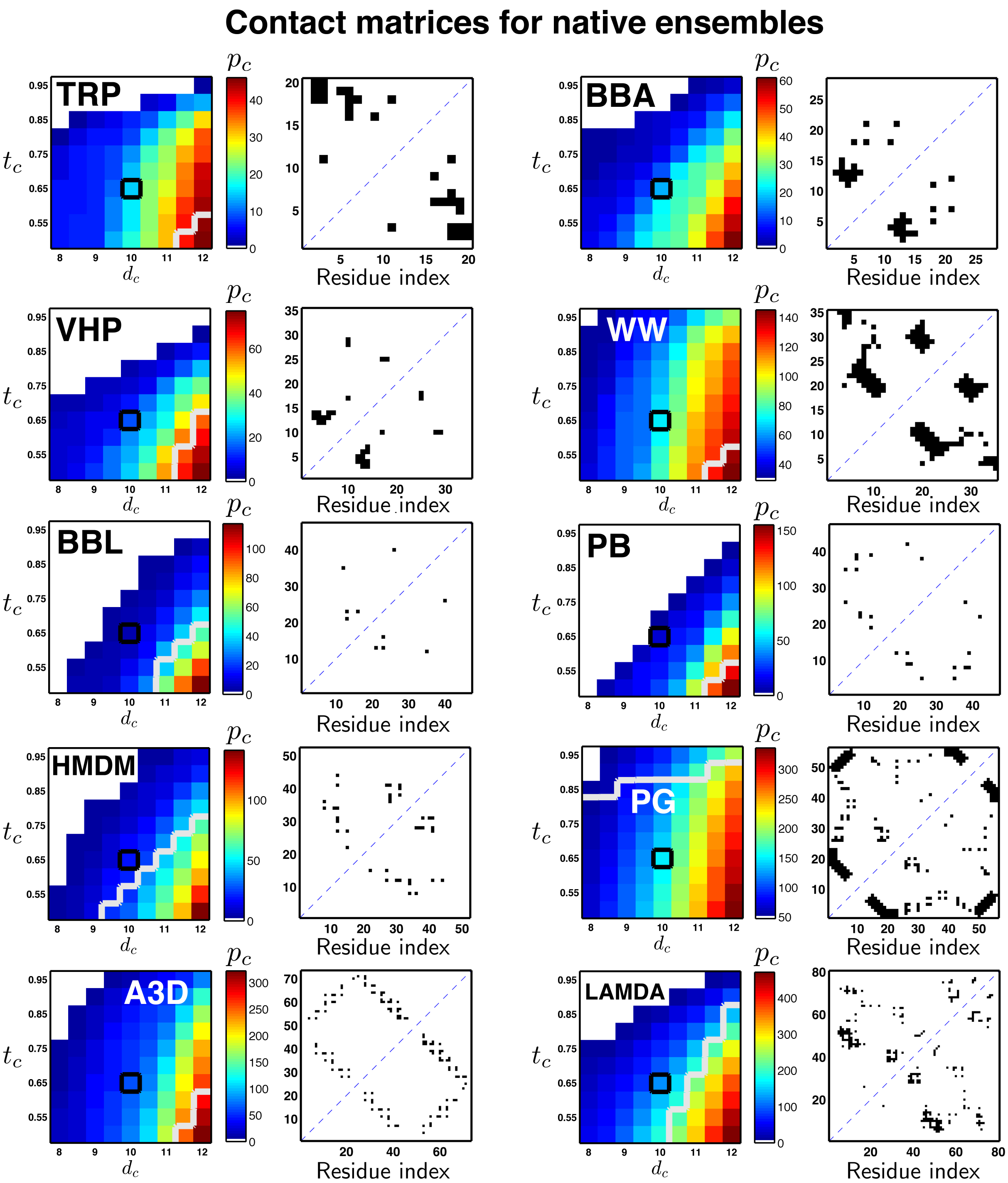

Figure S3. **Temporal and spacial thresholds for defining native contacts.** Heat maps (left columns) show the number of native contacts, p_c, defined for a range of interresidue distances, d_c (Å), and temporal thresholds, t_c (proportion of native frames). The black box indicates the selected threshold pair selected for all results obtained in main text: $d_c = 10$ Å and $t_c = 0.65$. Gray contours delineate threshold pairs that produce native contacts in the *nonnative* ensemble. Resulting contacts per protein are depicted in the contact maps (right columns).

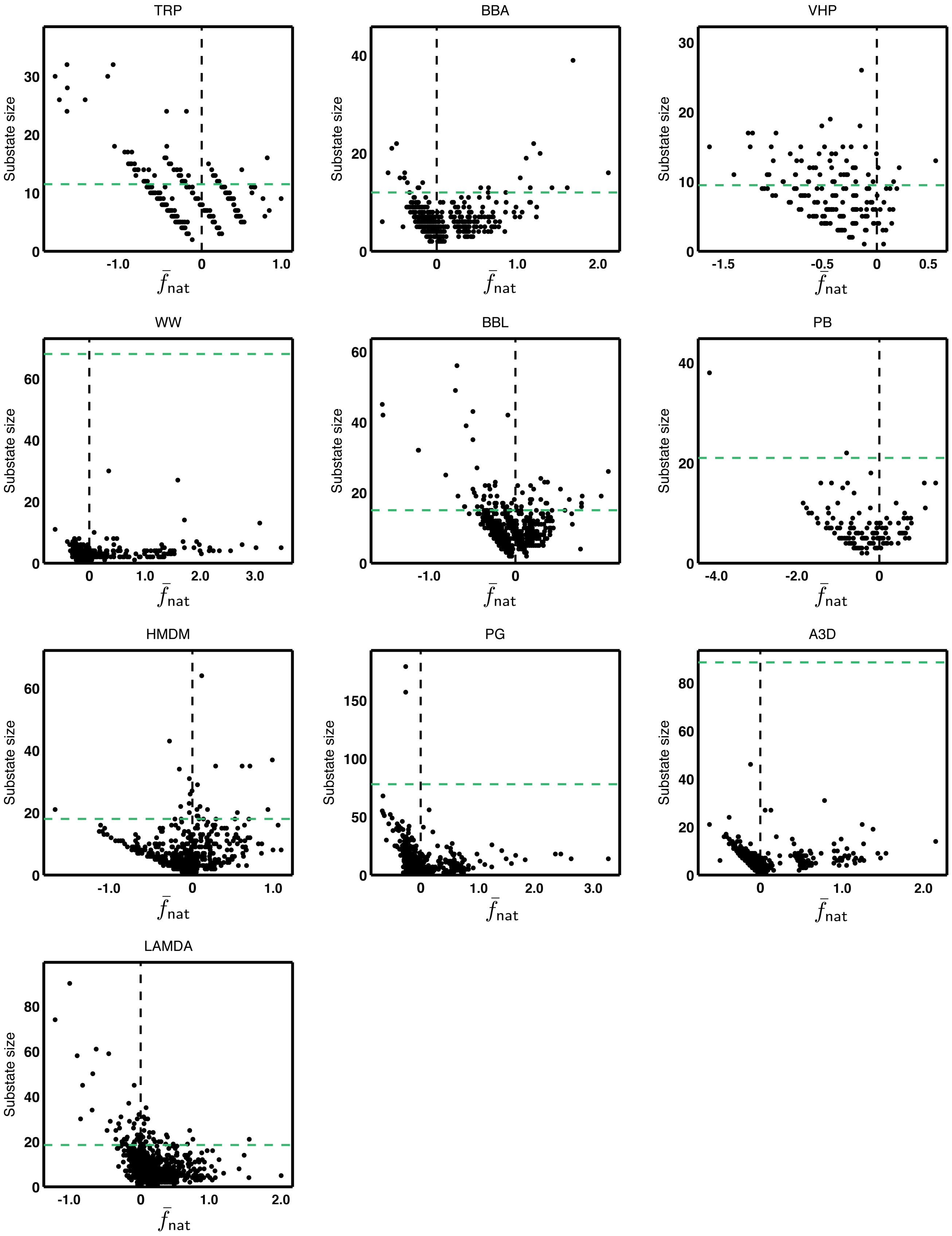

Figure S4. Frustration scores, $\bar{f}_{\text{nat}}$, versus substate populations. The green dashed line indicates the median substate size within the native ensemble.

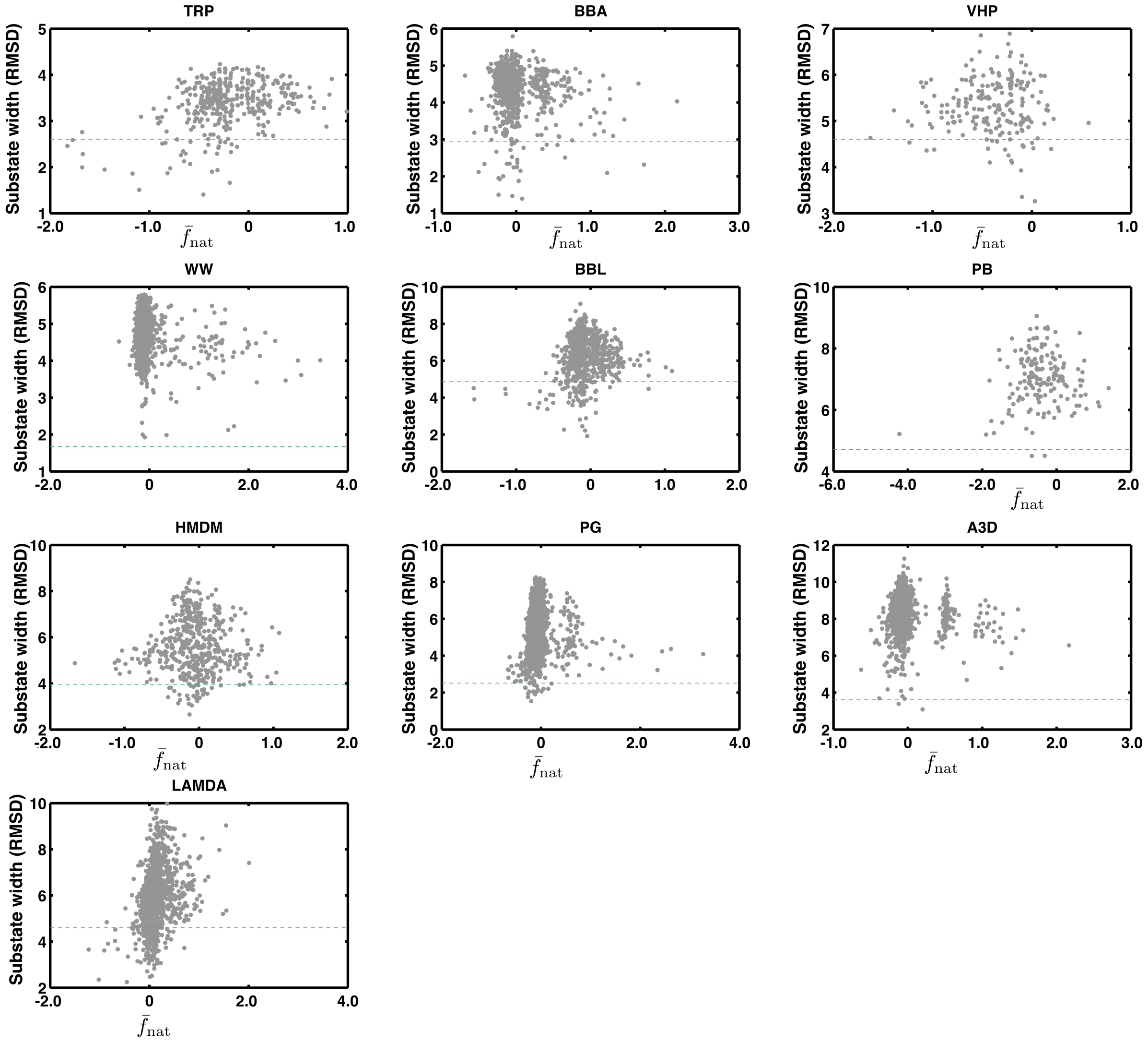

Figure S5. **Substate widths**. Frustration scores, $\bar{f}_{\text{nat}}$, plotted against substate widths, defined as average intra-substate pairwise RMSD. The green dashed line indicates the median cluster width of all conformational substates within the native ensemble. Singletons are excluded.

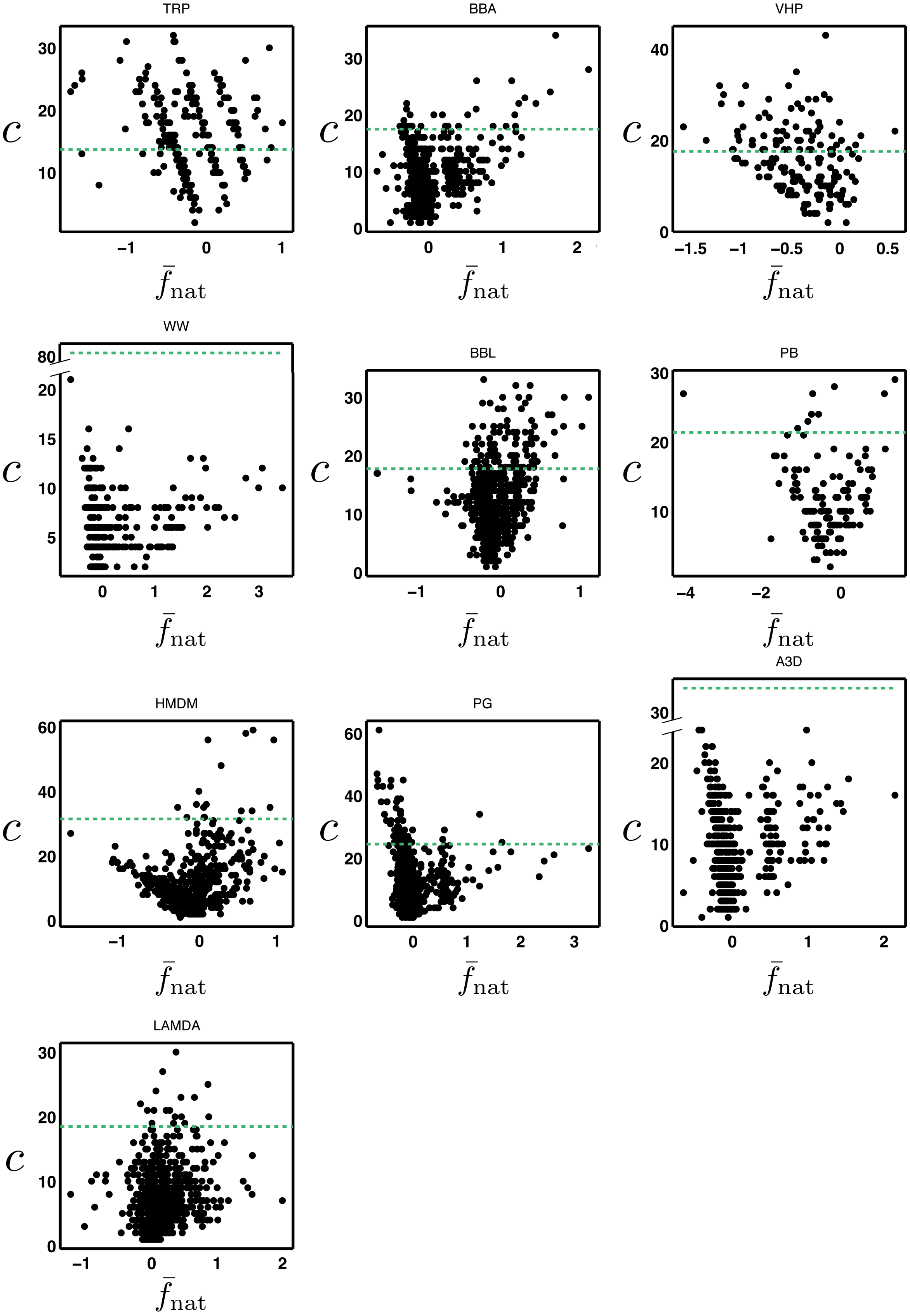

Figure S6. **Substate connectivities**. Frustration scores, $\bar{f}_{\text{nat}}$, plotted against total neighborhood connectivity (i.e., number of neighbors), c, for each nonnative substate. The green dashed line indicates the mean neighborhood connectivity for all substates within the native ensemble.

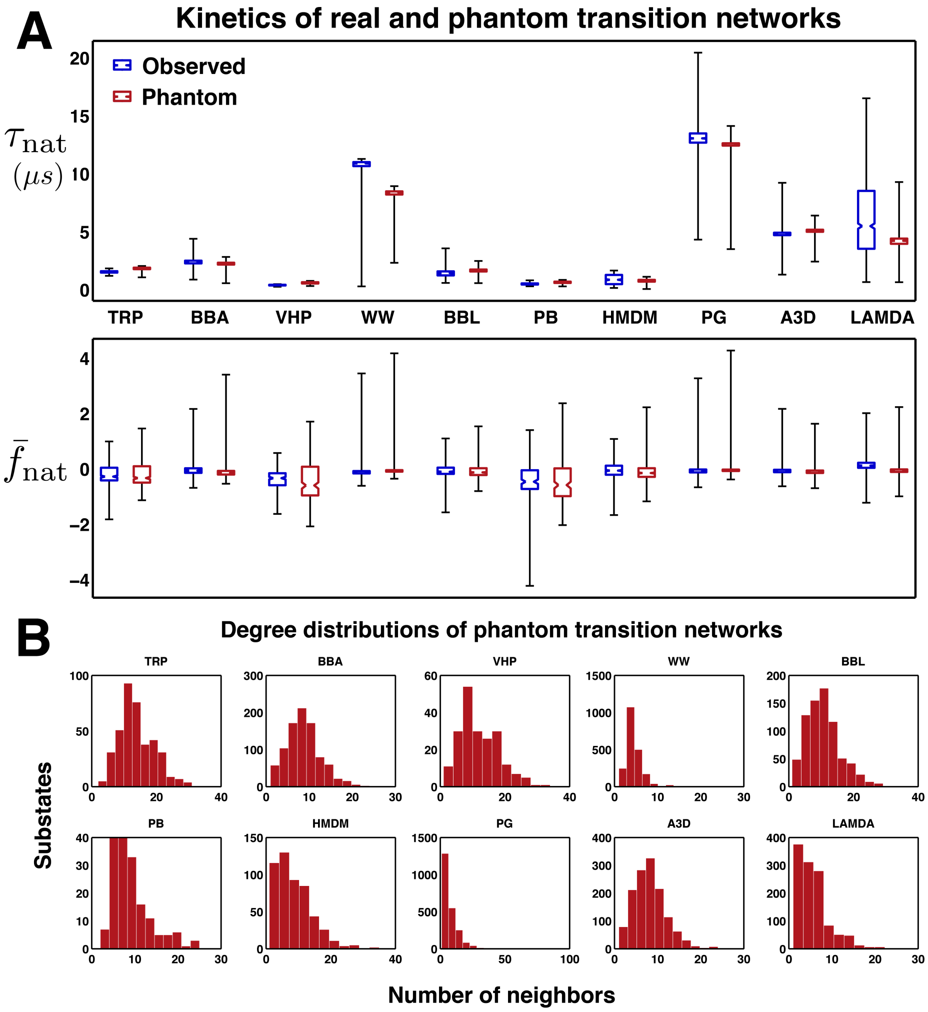

Figure S7. (A) A comparison of kinetic properties (transit times, top; frustration scores, bottom) for observed, blue, and phantom (i.e., synthesized), red, kinetic transition networks. Box notches indicate the median, box edges indicate the 25th and 75th percentiles, and whiskers denote data limits. (B) Degree distributions of phantom networks. Phantom nonnative ensembles with substate counts and degree distributions matching those of the observed networks were synthesized with Complex Networks Package for Matlab [3]. The resulting transition count matrix was symmetrized and edge weights were assigned based on corresponding distributions within observed networks. As in Fig. 1, a single substate was then added to represent the entire native ensemble, and edges connecting the native and nonnative ensembles were introduced in accordance with their prevalence in the observed networks, c.f. $\frac{l_{\rm nn\rightarrow n}}{l_{\rm nn}}$ in Table 2. Native ensemble self transitions were assigned to equate with total intra-ensemble transitions from the native ensembles in the observed networks. The resulting transition count matrix then underwent the perturbation process in Methods to yield $\bar{f}_{\rm nat}$ values.

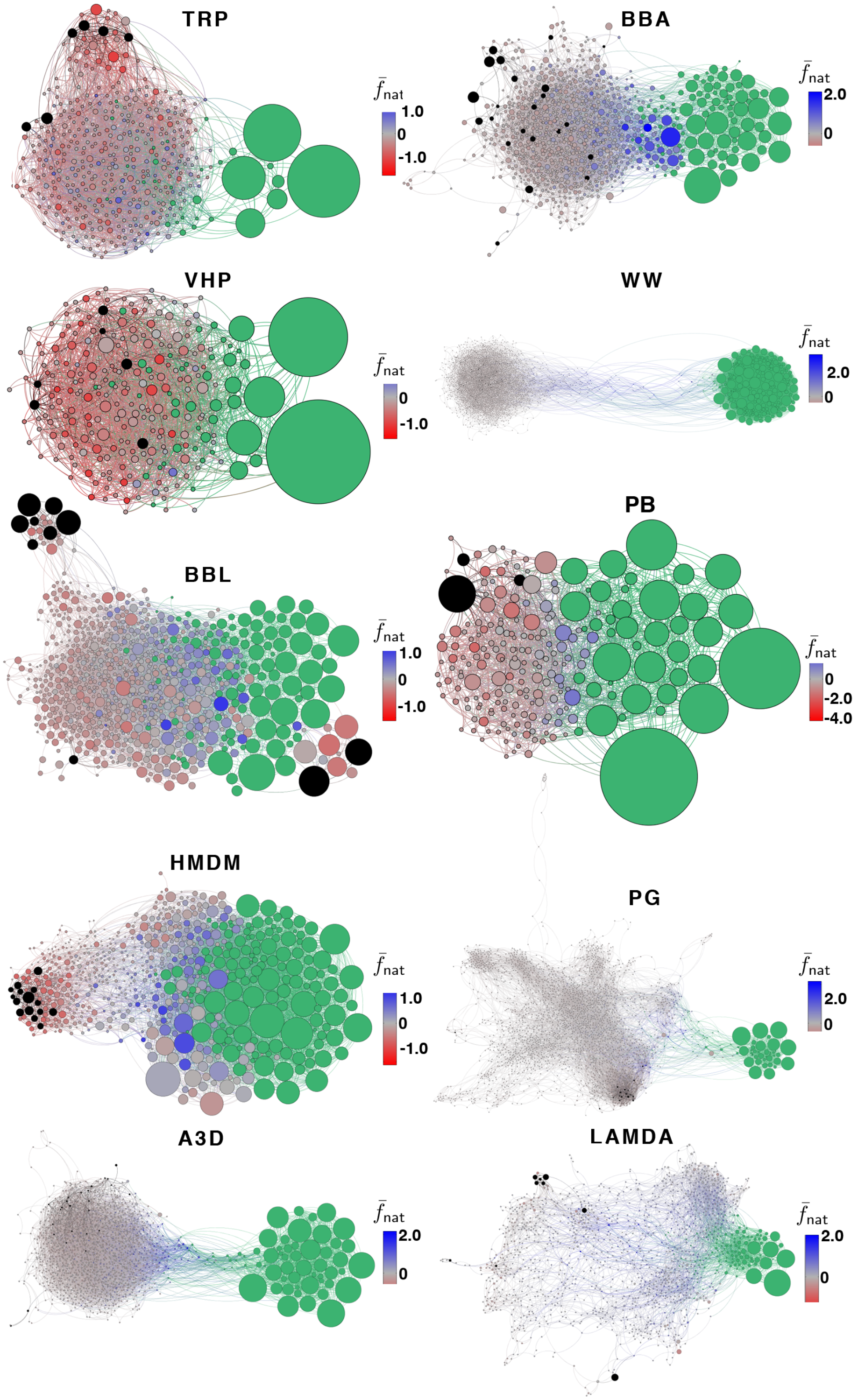

Figure S8. **Location of major inhibitor substates**. Conformational substates depicted in structural ensembles (main text, Fig. 6) are colored black.

---

[1] G. Labesse, N. Colloc'h, J. Pothier, and J.-P. Mornon, CABIOS **13**, 291 (1997).
[2] F. Pozzi, T. Di Matteo, and T. Aste, Eur. Phys. J. B (2012).
[3] L. Muchnik, *Complex Networks Package for MatLab (Version 1.6)* (www.levmuchnik.net).