Supporting information: How evolution designs functional free energy landscapes of proteins? A case study on emergence of regulation in CDK family kinases.

Zahra Shamsi[†] and Diwakar Shukla[*,†,‡,¶,§,‖,⊥]

†*Department of Chemical and Biomolecular Engineering, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA*

‡*Department of Plant Biology, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA*

¶*Center for Biophysics and Quantitative Biology, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA*

§*National Center for Supercomputing Applications, Urbana, IL 61801, USA*

‖*NIH Center for Macromolecular Modeling and Bioinformatics, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA*

⊥*Beckman Institute for Advanced Science and Technology, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA*

E-mail: diwakar@illinois.edu

List of Figures

Accelerated molecular dynamics simulations.

Accelerated molecular dynamics is a biasing potential method, derived to lower the local barriers between different states and make calculations much faster. Amber 14 implementation of aMD was used for boosting torsional terms independently.[1] The modified potentials are defined as following:

$$V^*(r) = V(r) + \Delta V(r)$$

$$\Delta V(r) = \frac{(E_p - V(r))^2}{(\alpha_p + E_p - V(r))} + \frac{(E_d - V_d(r))^2}{(\alpha_D + E_d - V_d(r))}$$

where $V(r)$, $V_d(r)$, E_p and E_d are the normal potential, normal torsion potential, average potential and dihedral energies. Parameters of αP and αD are factors inversely control the strength with which the boost is applied.[1] Using a trial short unbiased MD simulation of different systems, energy terms and other parameters were calculated as shown in Table S1 and S2.

Table S1: Accelerated MD parameters for starting structures in CDK2.

System structure (PDB ID)	αD	E_d	αP	E_p
1FIN	223.13	4834.55	8317.40	-124339.53
3PXR	222.08	4811.80	8400.20	-125495.11
3PXF	221.37	4796.44	8359.80	-125030.85
4GCJ	226.55	4908.60	8723.20	-130328.57

Table S2: Accelerated MD parameters for starting structures in CMGI. .

System structure (model template PDB ID)	αD	E_d	αP	E_p
1FIN	222.19	4814.19	9478.20	-141969.11
3PXR	223.79	4848.71	9902.40	-148231.35
3PXF	222.37	4818.08	9591.60	-143668.54
4GCJ	223.00	4831.64	8927.20	-141717.94

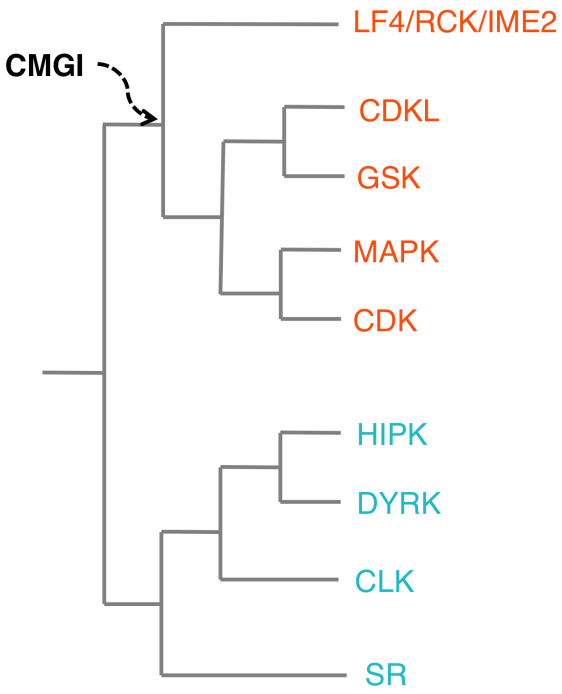

Figure S1: **Phylogenetic tree.** Cyclin Dependent Kinase (CDK), Mitogen Activated Protein Kinase (MAPK), Glycogen Synthase Kinase (GSK), and Casein Kinase (CK) group of kinases. The CMGC group also contains the CDK-Like kinases (CDKL), SR-kinases, Homeodomain-Interacting Kinases (HIPKs), CDC-Like Kinases (CLKs), Dual-Specificity Tyrosine Regulated Kinases (DYRKs), and a paralogous superfamily of kinases including LF4, the mammalian RCK kinases (ICK, MOK, and MAK), and the fungal IME2 kinases.

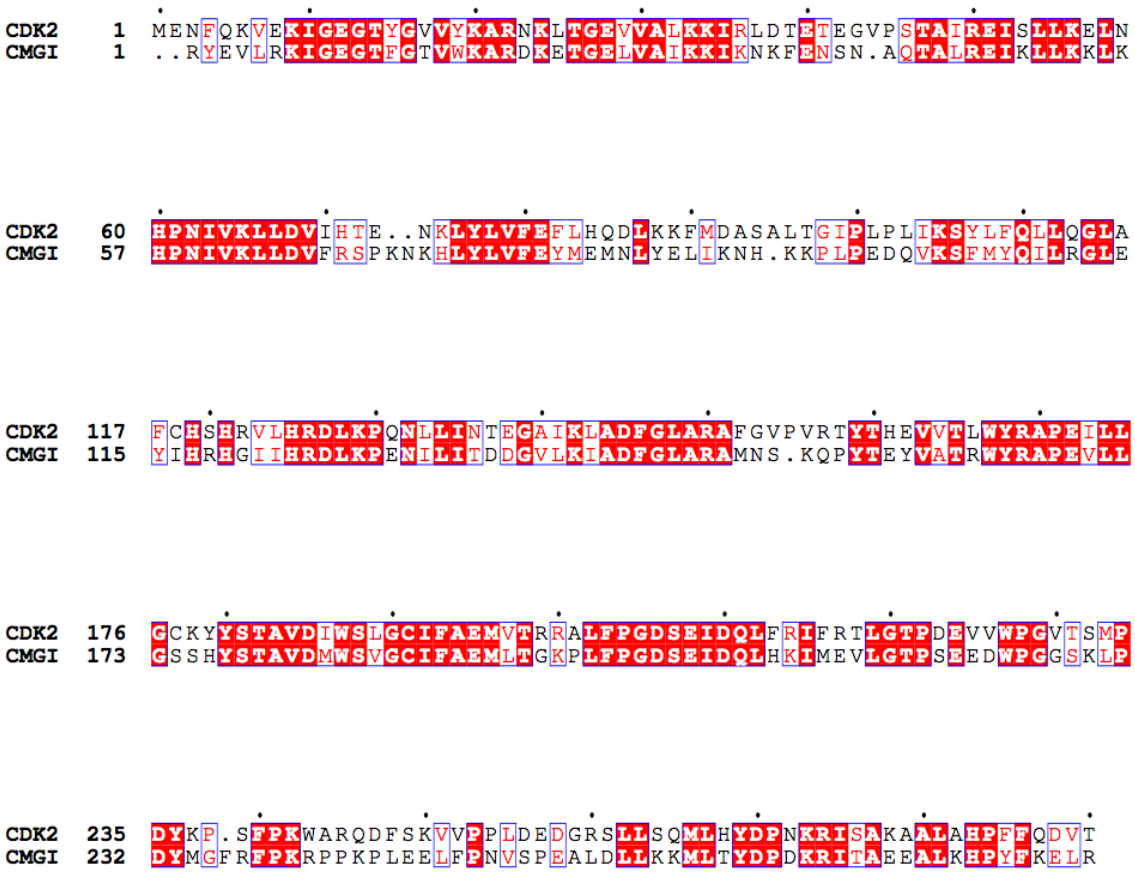

Figure S2: **Amino acid sequence alignment of CDK2 and CMGI.** The red box with white character shows strict identity and blue frame with red character means similarity across groups between the sequences. ESPript 3 web server is used to generate the alignment.[2]

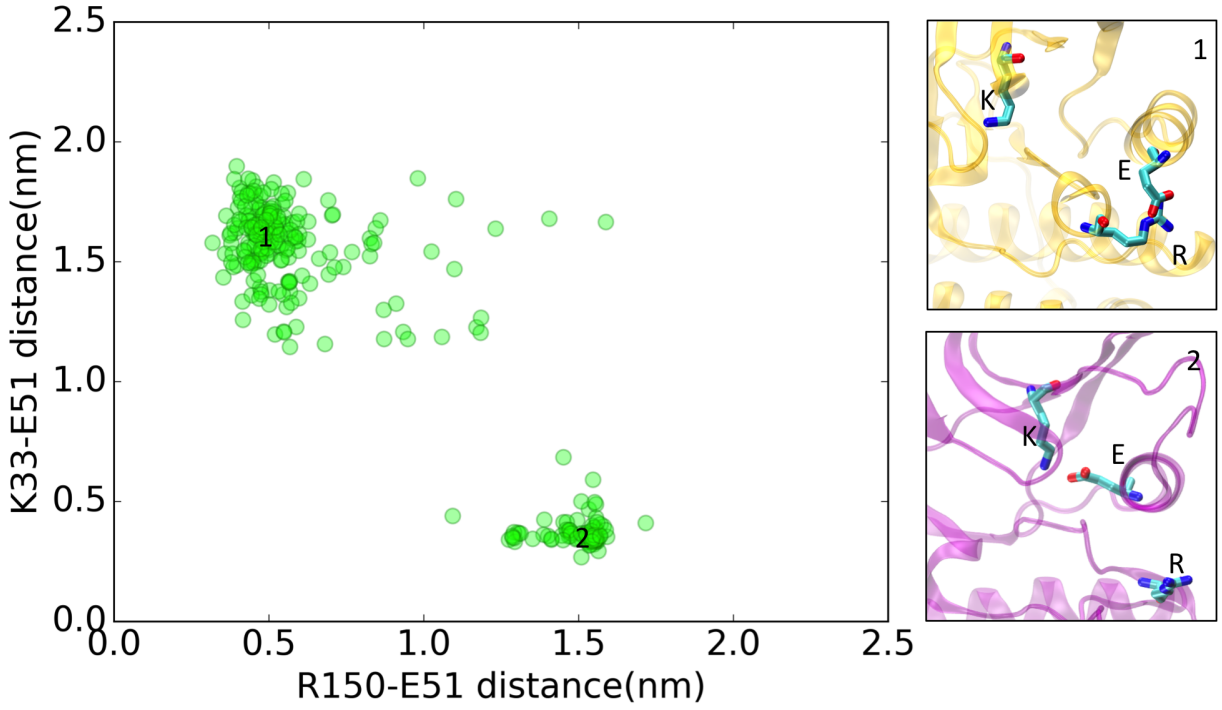

Figure S3: **Scatter plot of R150-E51 versus K33-E51 distances calculated using all available crystal structures of CDK2 kinase.** The scatter plot separates structures into two regions of active and inactive. Region 1 highlights inactive crystal structures with broken K-E bond, and mostly formed R-E bond. A crystal structure representative of region 1 (PBD ID: 2VTP[3]) is shown in **box 1**. Region 2 highlights active crystal structures with formed K-E and broken R-E bond. In **box 2** a crystal structure (PBD ID: 1FIN[4]) representative of region 2 is shown.

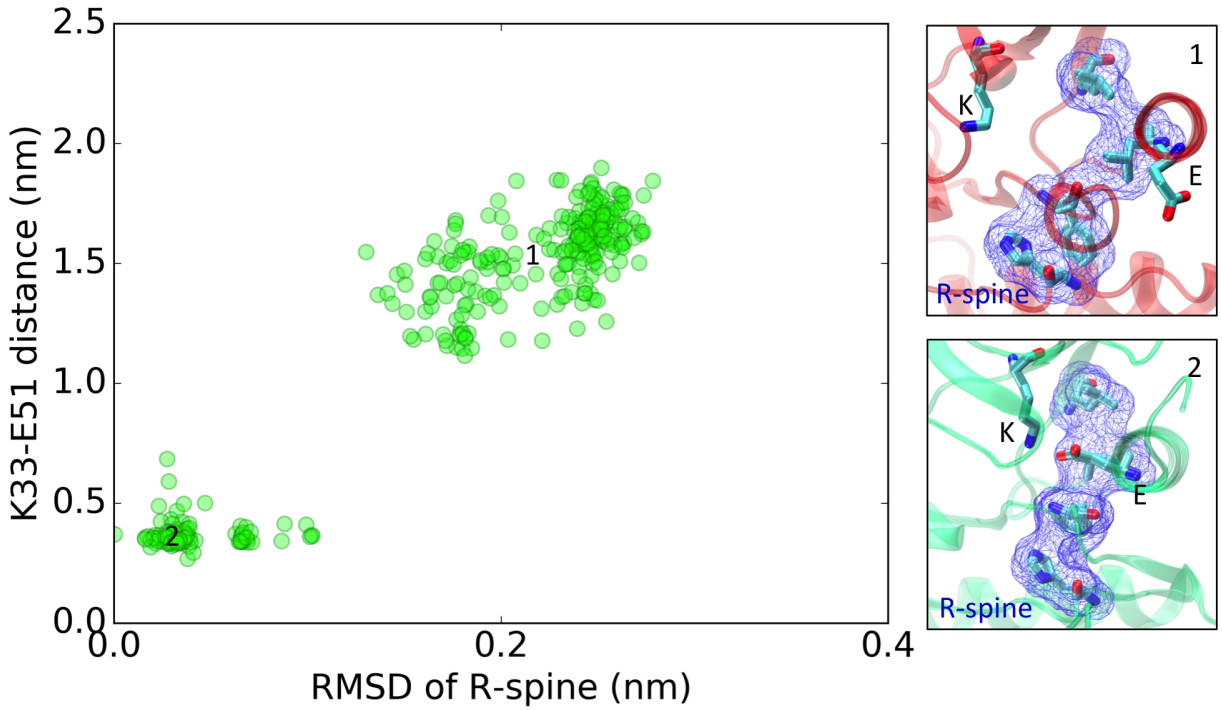

Figure S4: **Scatter plot of R-spine RMSD versus K33-E51 distances calculated using all available crystal structures of CDK2 kinase.** Scatter plot of R-spine RMSD versus K33-E51 distances calculated using all available crystal structures of CDK2 kinase. The scatter plot of the R-spine residues (Leu66, Leu55, Phe146, and His125) RMSD and K33-E51 distance shows two distinct regions. A crystal structure representative for region 1 (PBD ID: 2VTP) with broken R-spine and K-E bond is shown in **box 1**. Region 2 highlights crystal structures with formed R-spine and K-E bond. A crystal structure representative for region 2 (PBD ID: 1FIN) is shown in **box 2**. All R-spine RMSDs were calculated with respect to active crystal structure (PDB ID: 1FIN).

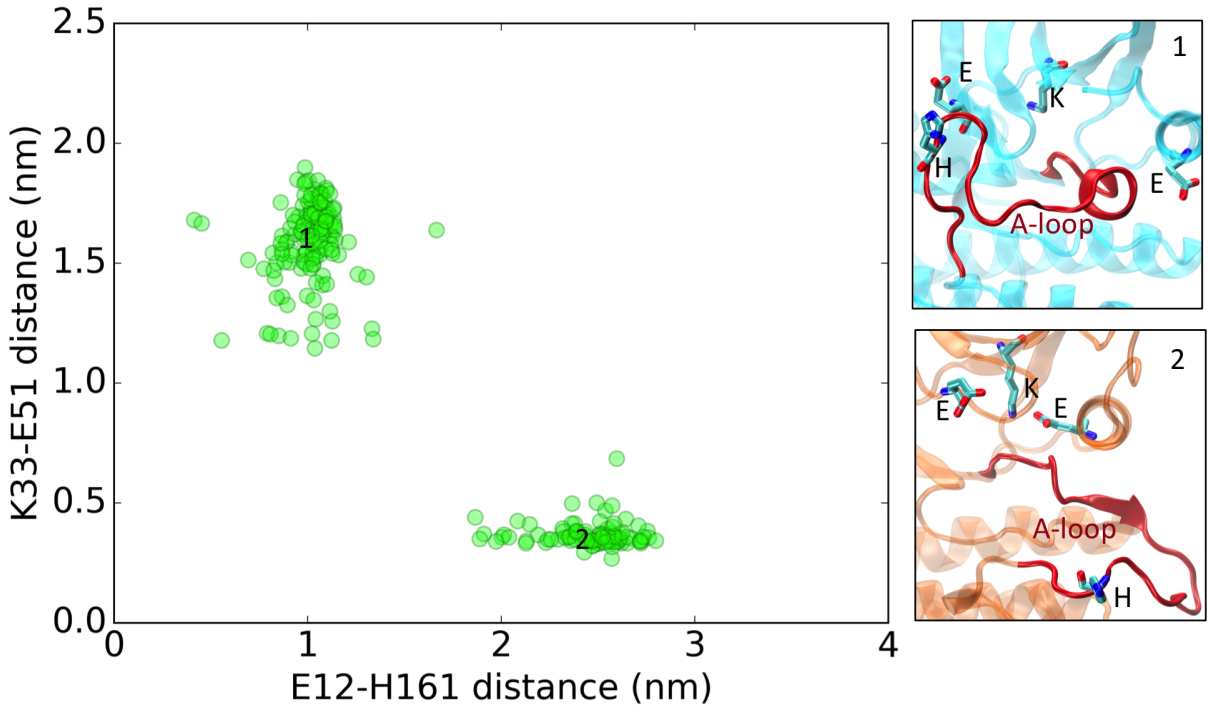

Figure S5: **Scatter plot of helical turn RMSD versus K33-E51 distances calculated using all available crystal structures of CDK2 kinase.** The scatter plot of the helical turn in A-loop residues (Gly147, Leu148, Ala149, Arg150, Ala151, Phe152) RMSD and K33-E51 distance separates the crystal structures into two regions of active and inactive. A crystal structure representative for region 1 (PBD ID: 3PXR[5]) with broken R-spine and K-E bond is shown in **box 1**. Region 2 highlights crystal structures with formed R-spine and K-E bond. A crystal structure representative for region 2 (PBD ID: 1FIN) is shown in **box 2**. Helical turn RMSDs were calculated with respect to inactive crystal structure (PDB ID: 3PXR[5]).

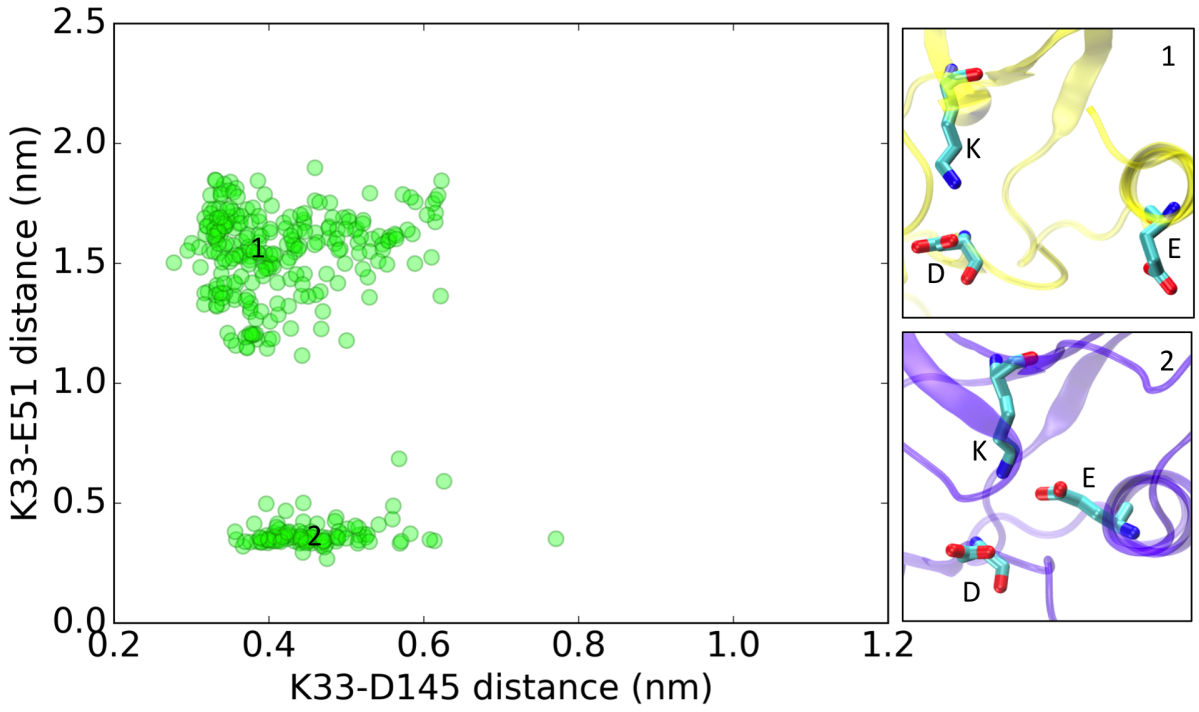

Figure S6: **Scatter plot of catalytic base versus K33-E51 distances calculated using all available crystal structures of CDK2 kinase.** Projection of available crystal structures of CDK2 in K33-E51 versus K33-D145 distances classifies the structures into two regions. Region 1 with broken K33-E51 bond and formed K33-D145 bond is depicted in the **box 1** (PBD ID: 1AQ1[4]). In **box 2** a crystal structure (PBD ID: 1FIN[6]) from region 2 is shown with formed K-E bond and broken K-D bond.

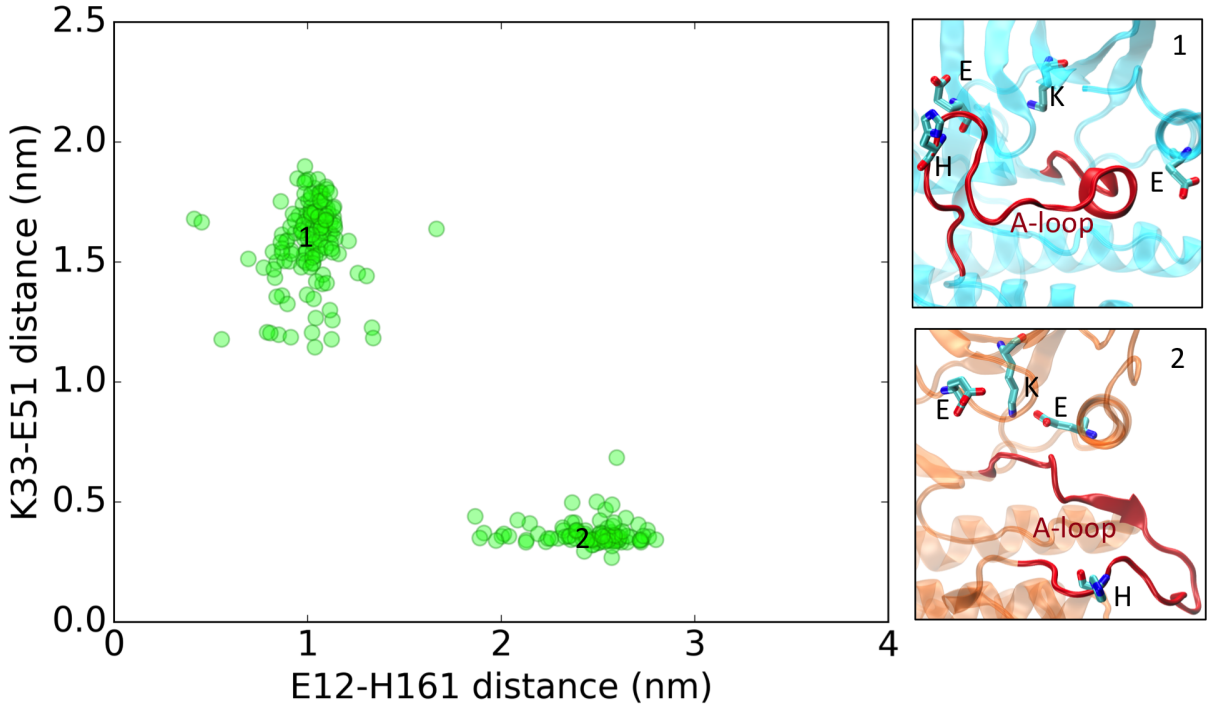

Figure S7: **Scatter plot of E12-H161 versus K33-E51 distances calculated using all available crystal structures of CDK2 kinase.** The distance between E12 in p-loop and H161 in A-loop can capture the distance between two loops and classifies the crystal structures into two distinct regions. Region 1 highlights crystal structures with broken K-E and formed E-H, which is shown in **box 1** (PBD ID: 3PXR[5]). Region 2 highlights crystal structures with formed K-E and broken E-H bond. A crystal structure representative for region 2 (PBD ID: 1FIN[4]) is shown in **box 2**.

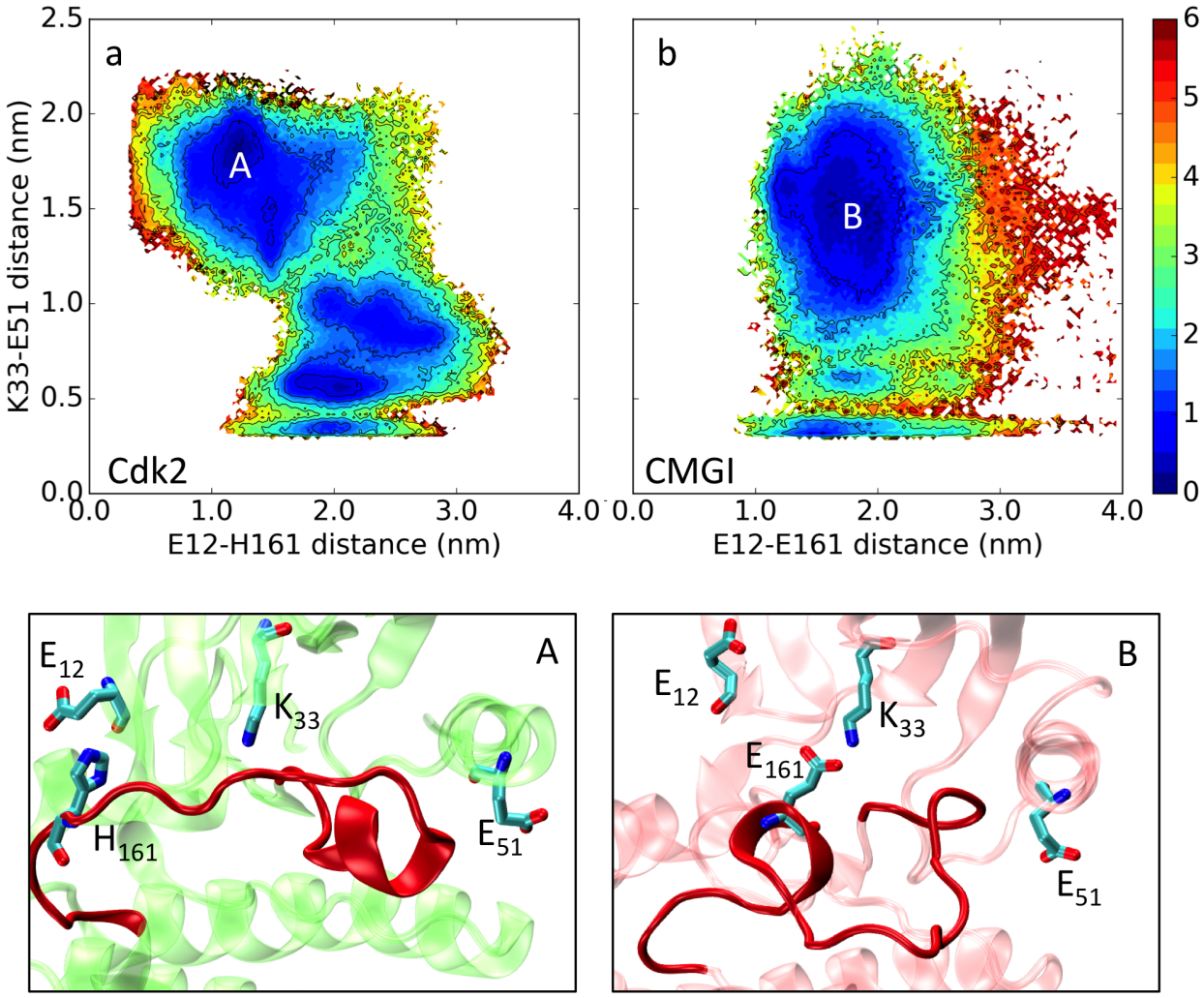

Figure S8: **Free energy landscape of K-E and E-H distances based on equilibrium weighted simulation data.** Comparing two-dimensional conformational landscapes of the K-E versus E-H residue distances shows emergence of a barrier of 3 kcal/mol in CDK2 for breakage of E-H bond. Colors show the free energy in kcal/mol.

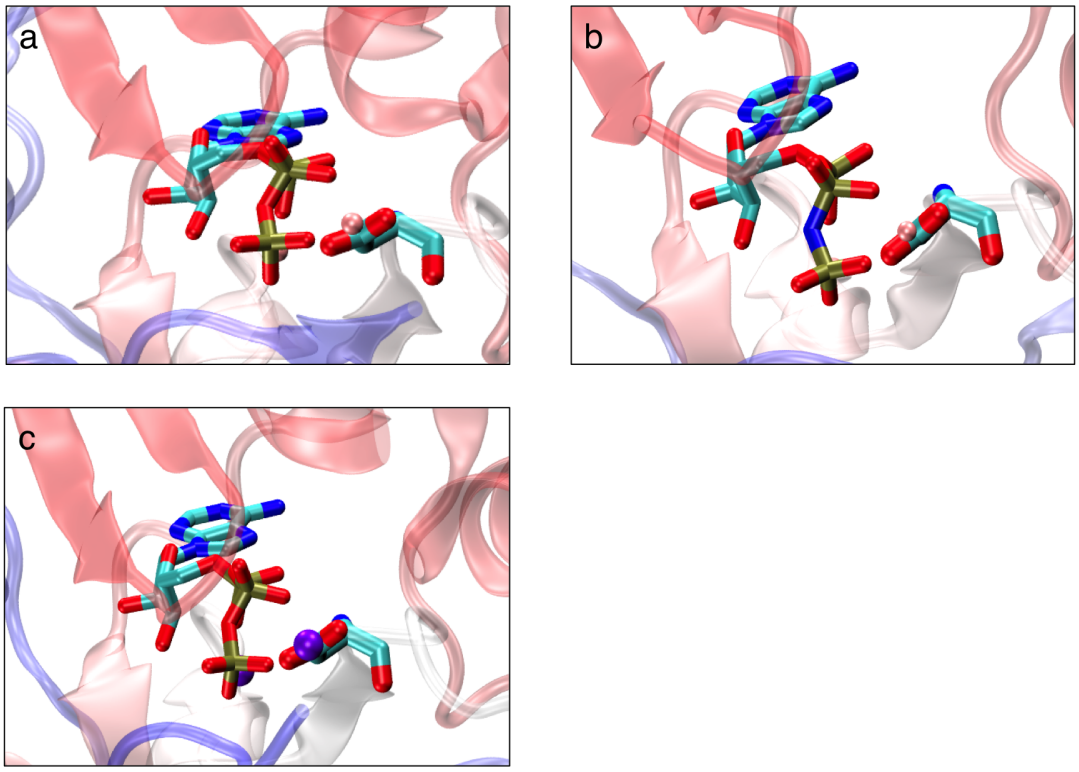

Figure S9: **Crystal structures of kinases bound to ATP and ions.** **a** is 1RDQ, Crystal Structure of a cAMP-dependent Protein Kinase Mutant at 1.26A: New Insights into the Catalytic Mechanism. , **b**, is 4HPU, Crystal structure of the catalytic subunit of cAMP-dependent protein kinase displaying partial phosphoryl transfer of AMP-PNP onto a substrate peptide . **c** is 1ATP, crystal structure of the catalytic subunit of cAMP-dependent protein kinase complexed with MnATP and a peptide inhibitor.

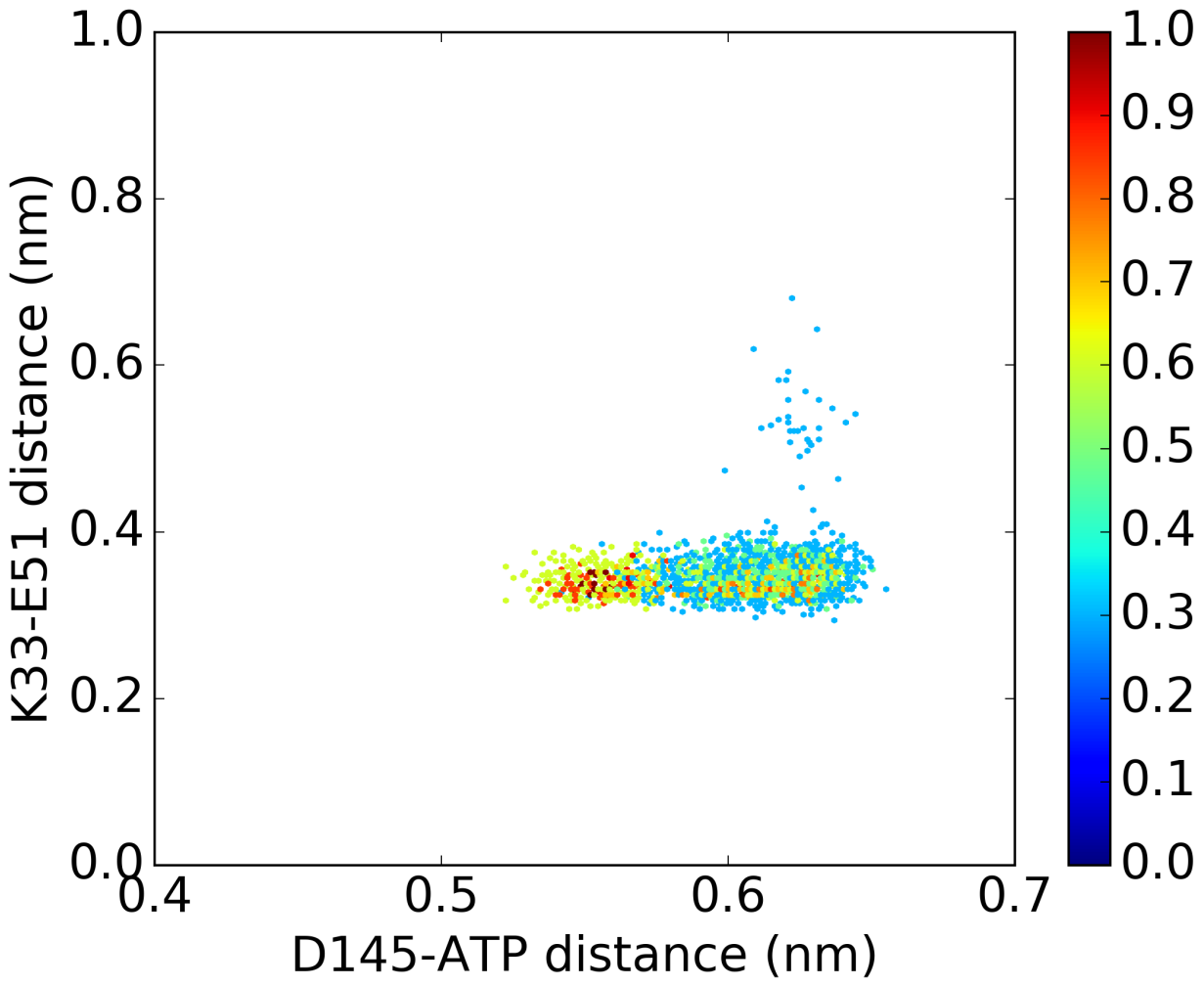

Figure S10: **Probability density map of K33-E51 and D145-ATP distances based on raw simulation data in CDK2 bound to cyclin.** Colors show the logarithm of the frequency in MD simulations.

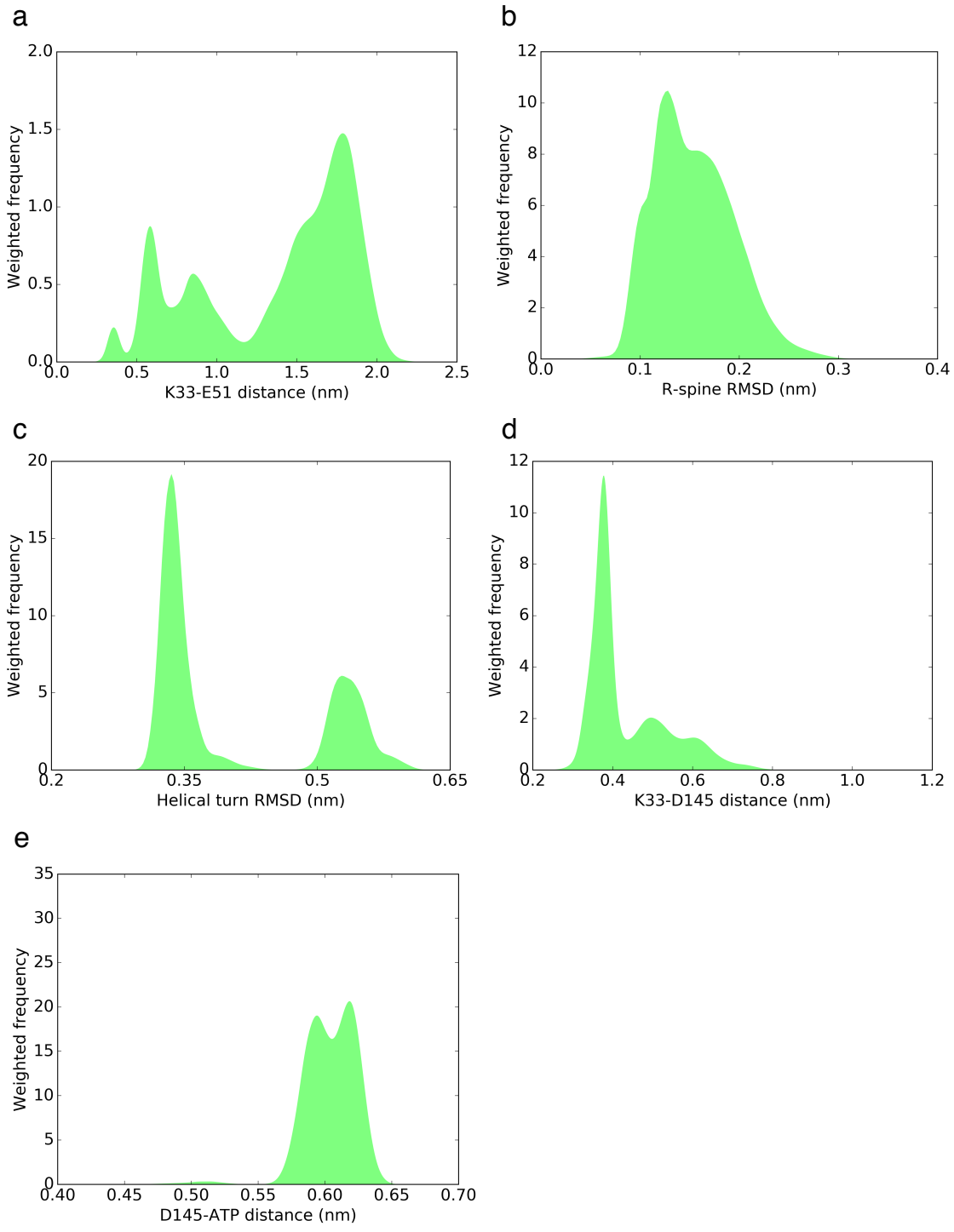

Figure S11: **Probability density map of four switches based on equilibrium weighted simulation data in CDK2.** R-spine RMSDs were calculated with respect to its active crystal structure (PDB ID: 1FIN[4]) and helical turn RMSDs were calculated with respect to inactive crystal structure (PDB ID: 3PXR[5]).

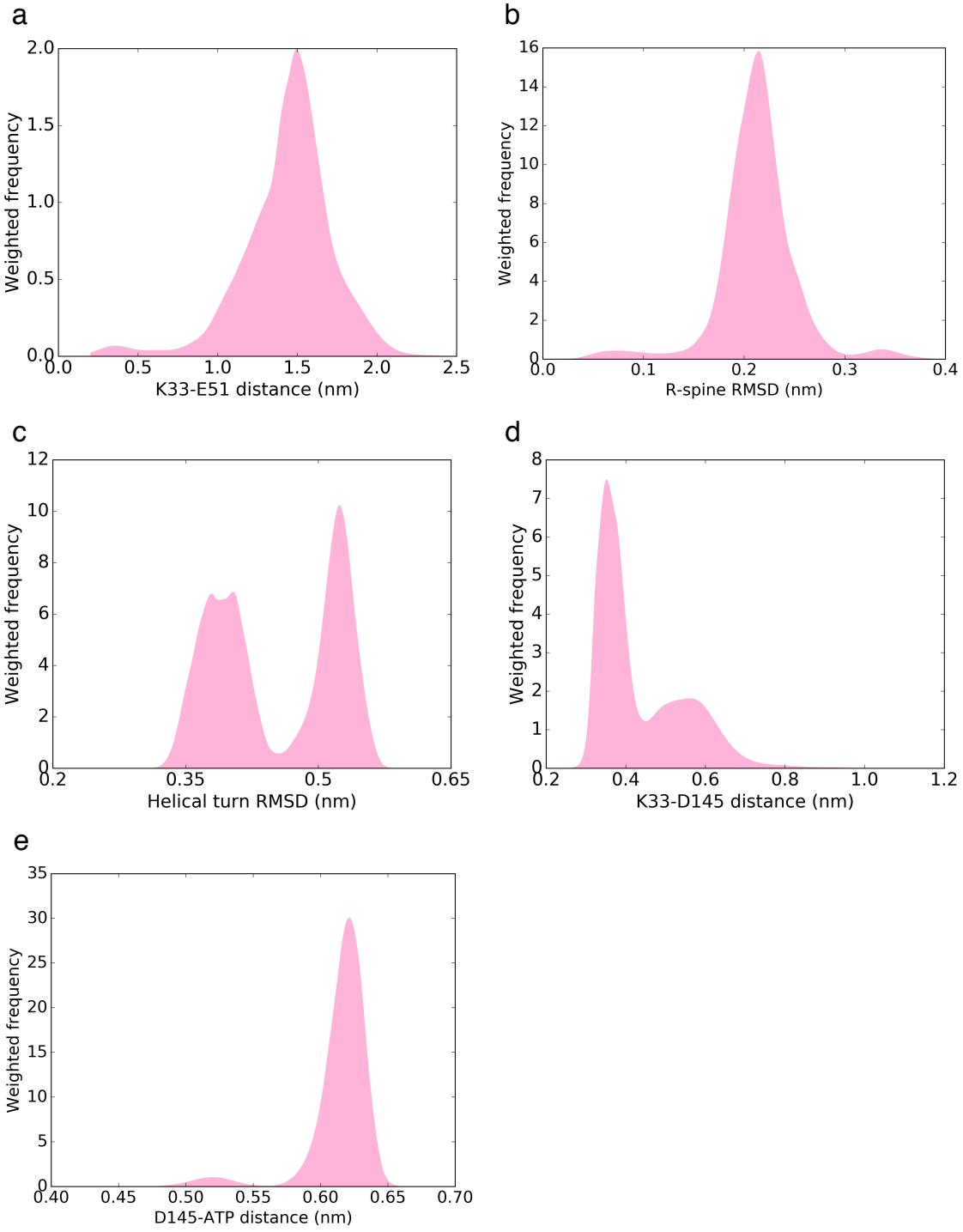

Figure S12: **Probability density map of four switches based on equilibrium weighted simulation data in CMGI.** R-spine RMSDs were calculated with respect to an active structure from simulation and helical turn RMSDs were calculated with respect to inactive crystal structure of CDK2 (PDB ID: 3PXR[5]).

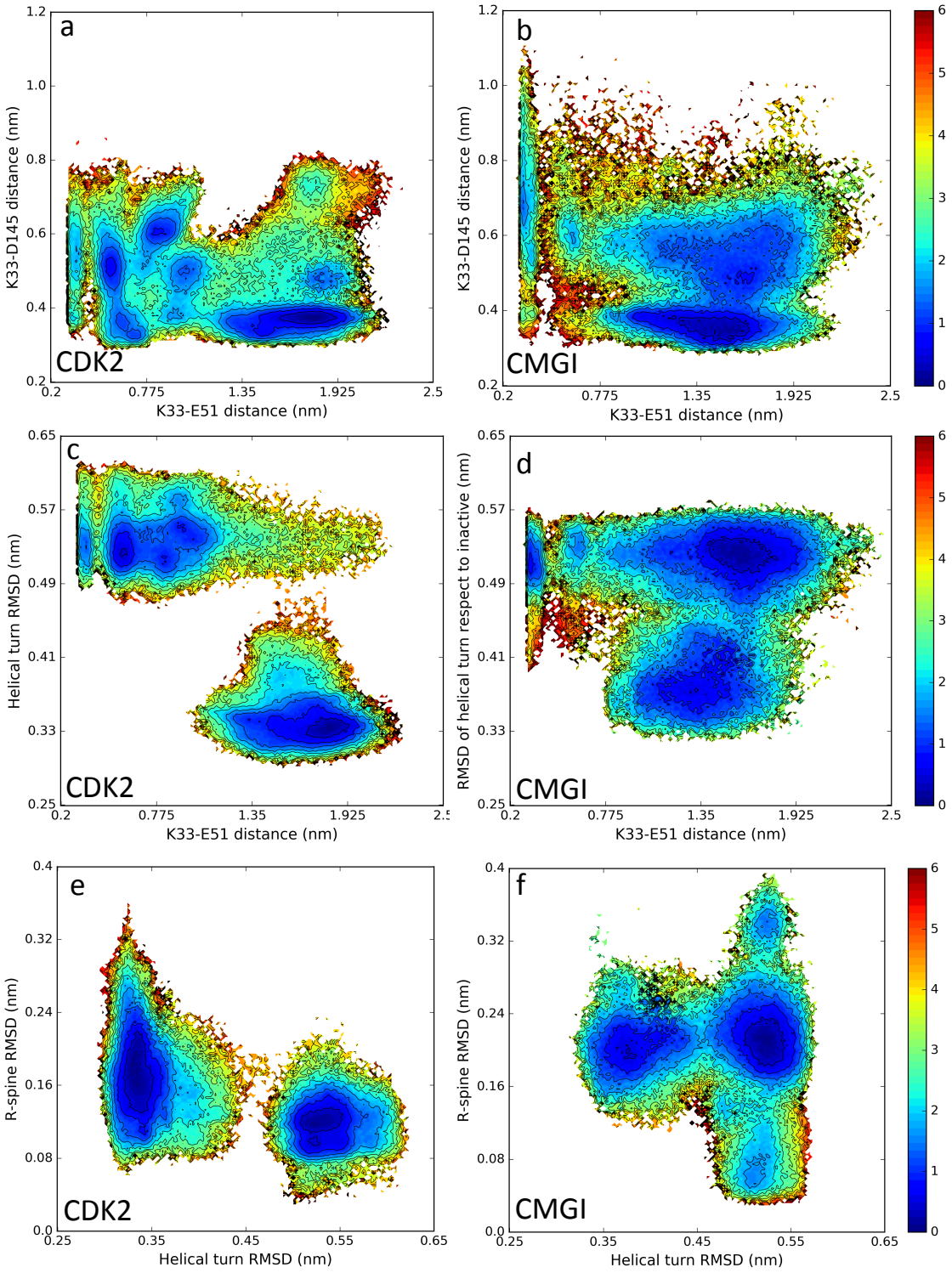

Figure S13: **Free energy landscape of CDK2 and CMGI based on equilibrium weighted simulation data.** R-spine RMSDs were calculated with respect to active crystal structure (PDB ID: 1FIN[4]) in CDK2 and a active structure from simulation in CMGI. Helical turn RMSDs were calculated with respect to inactive crystal structure (PDB ID: 3PXR[5]) in both CDK2 and CMGI. Colors show the free energy in kcal/mol.

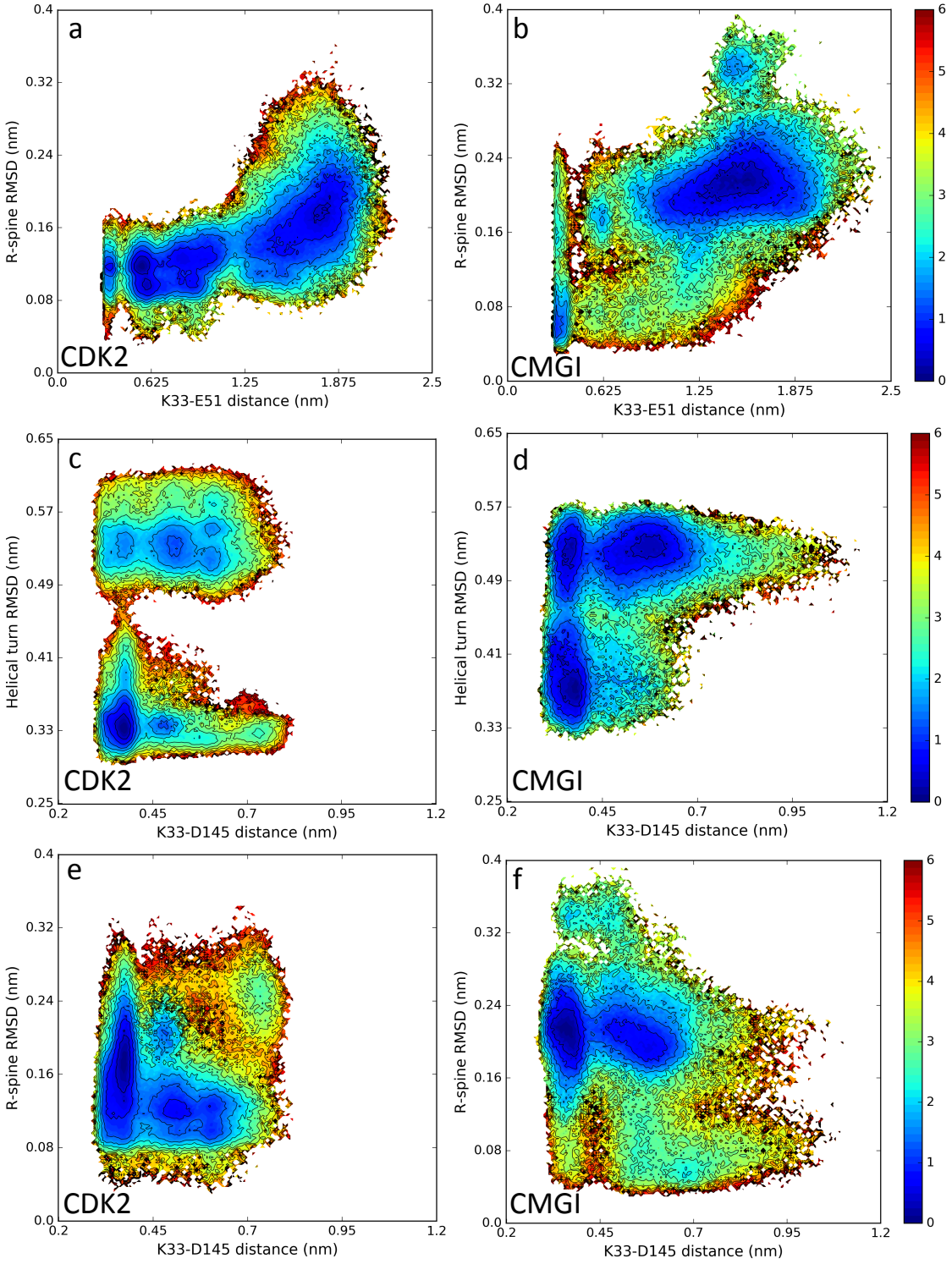

Figure S14: **Cont. free energy landscape of CDK2 and CMGI based on equilibrium weighted simulation data.** R-spine RMSDs were calculated with respect to active crystal structure (PDB ID: 1FIN[4]) in CDK2 and a active structure from simulation in CMGI. Helical turn RMSDs were calculated with respect to inactive crystal structure (PDB ID: 3PXR[5]) in both CDK2 and CMGI. Colors show the free energy in kcal/mol.

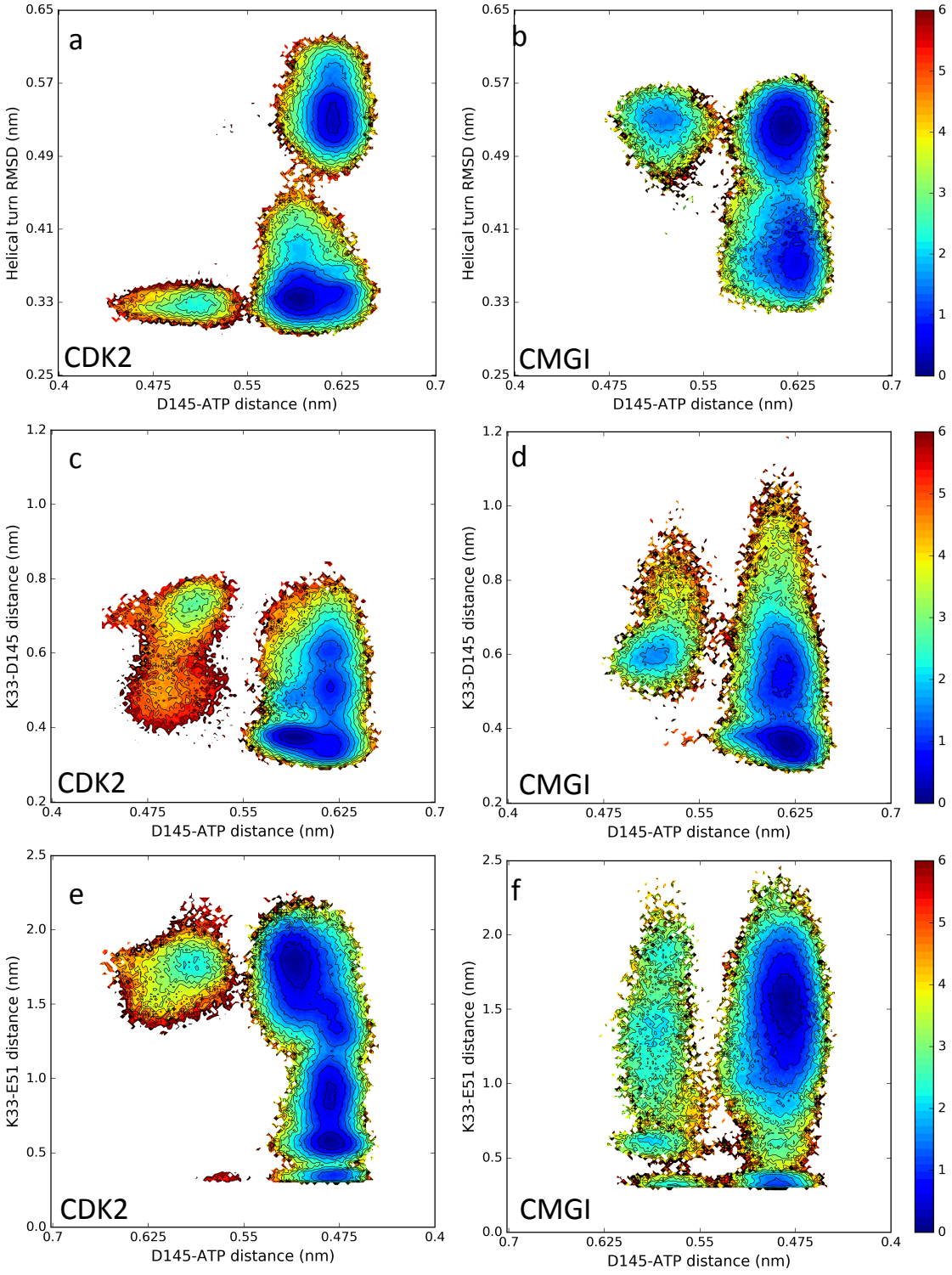

Figure S15: **Cont. free energy landscape of CDK2 and CMGI based on equilibrium weighted simulation data.** R-spine RMSDs were calculated with respect to active crystal structure (PDB ID: 1FIN[4]) in CDK2 and a active structure from simulation in CMGI. Helical turn RMSDs were calculated with respect to inactive crystal structure (PDB ID: 3PXR[5]) in both CDK2 and CMGI. Colors show the free energy in kcal/mol.

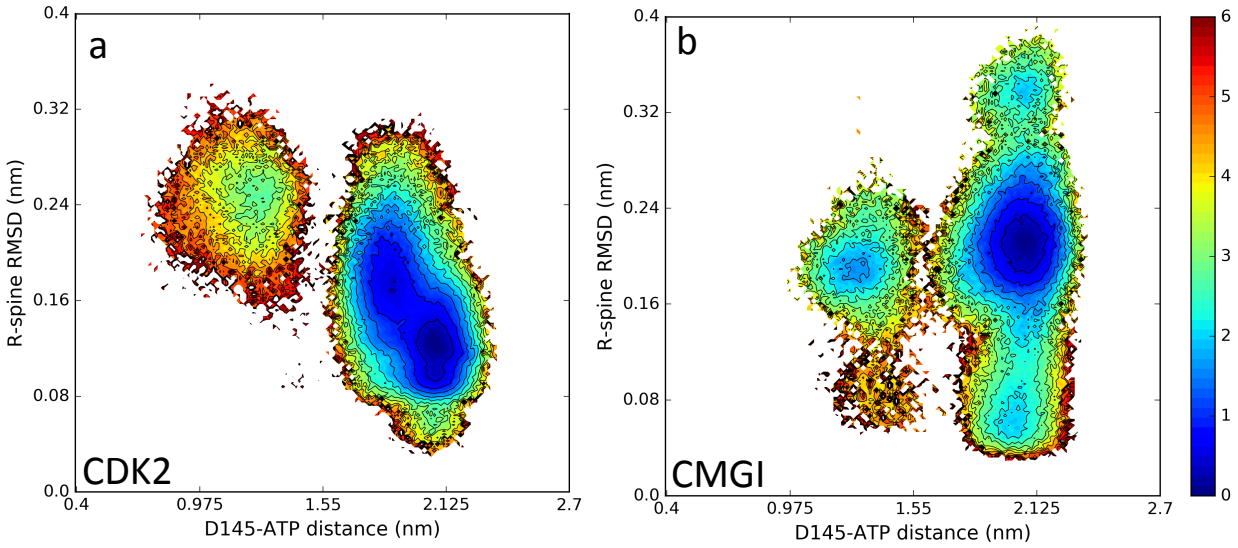

Figure S16: **Cont. free energy landscape of CDK2 and CMGI based on equilibrium weighted simulation data.** R-spine RMSDs were calculated with respect to active crystal structure (PDB ID: 1FIN[4]) in CDK2 and a active structure from simulation in CMGI. Helical turn RMSDs were calculated with respect to inactive crystal structure (PDB ID: 3PXR[5]) in both CDK2 and CMGI.Colors show the free energy in kcal/mol.

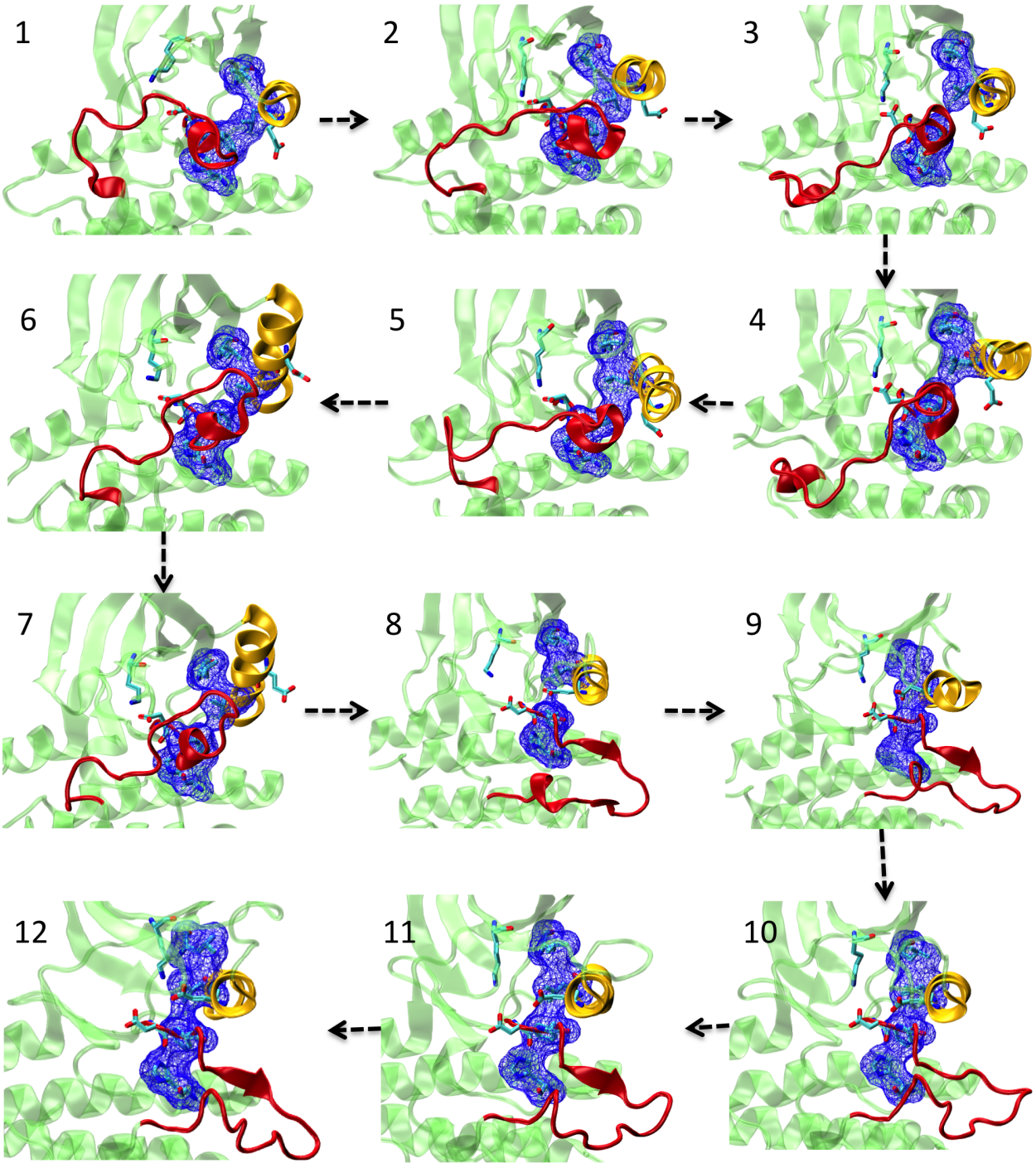

Figure S17: **Shortest path connecting closest state to inactive crystal structure (PDB ID: 3PXR[5]) to closest state to active crystal structure (PDB ID: 1FIN[4]) in CDK2's MSM.** Using transition path theory (TPT) the shortest activation pathway in CDK2 is captured.

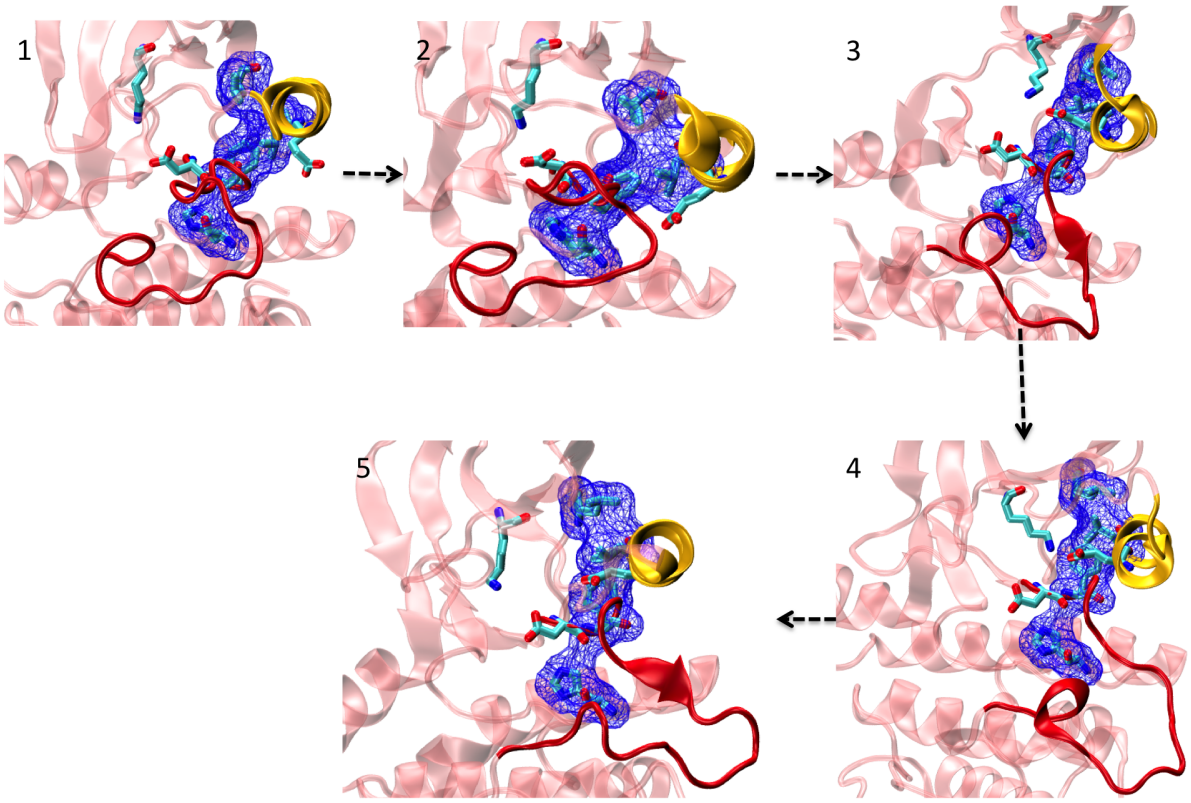

Figure S18: **Shortest path connecting closest state to inactive crystal structure (PDB ID: 3PXR[5]) to closest state to active crystal structure (PDB ID: 1FIN[4]) in CMGI's MSM.** Using transition path theory (TPT) the shortest activation pathway in CMGI is captured.

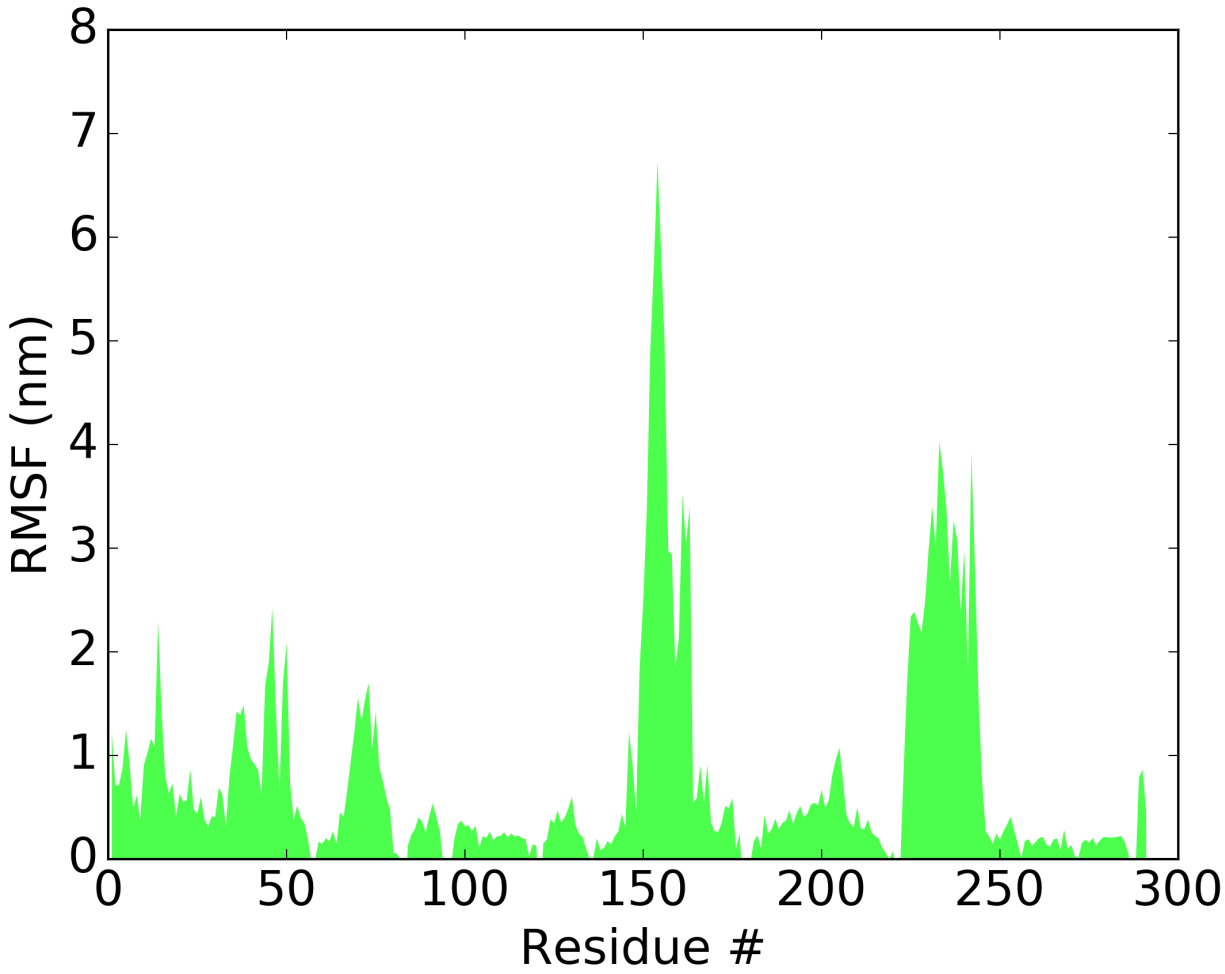

Figure S19: **Root mean square fluctuations (RMSF) of residues in CDK2.**

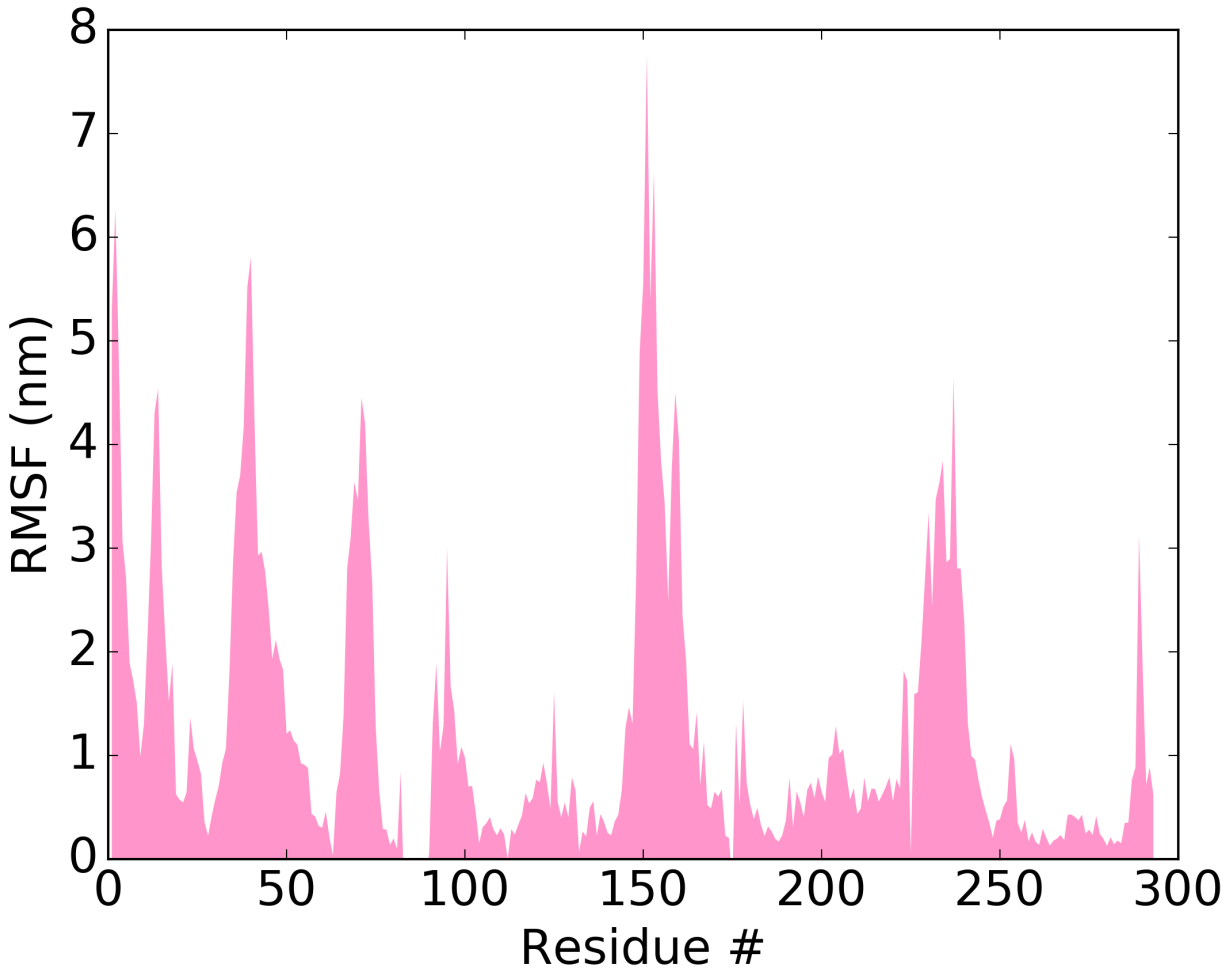

Figure S20: **Root mean square fluctuations (RMSF) of residues in CMGI.**

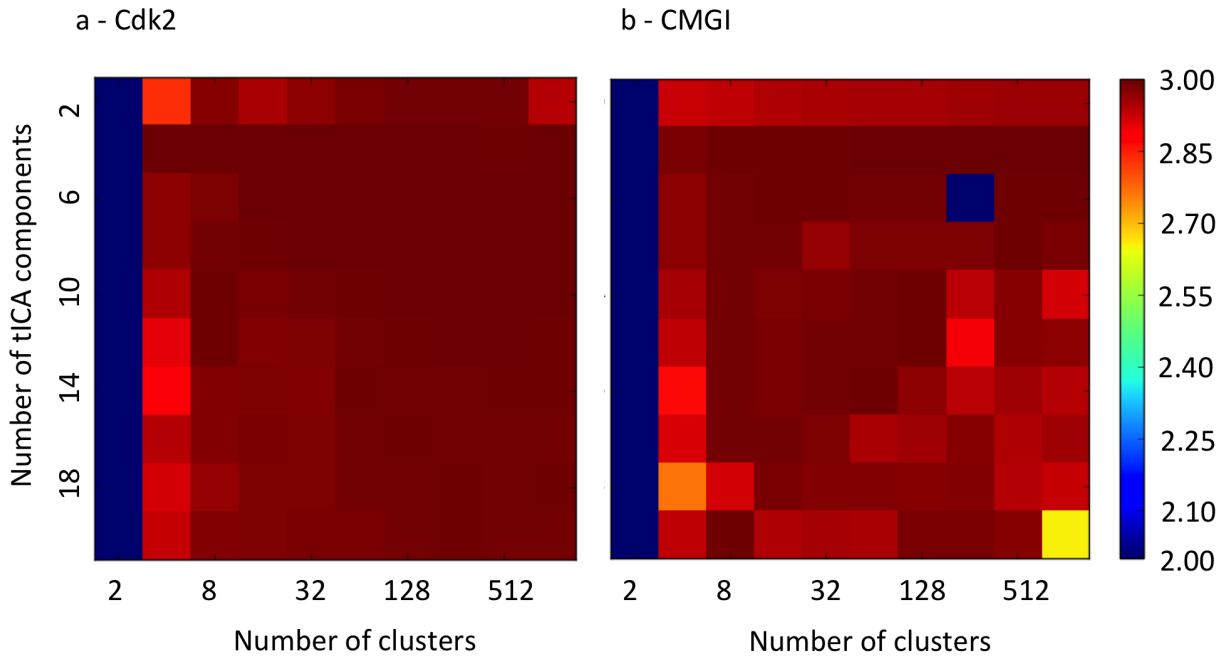

Figure S21: **GMRQ score for MSM as a function of number of clusters and tICA components.**

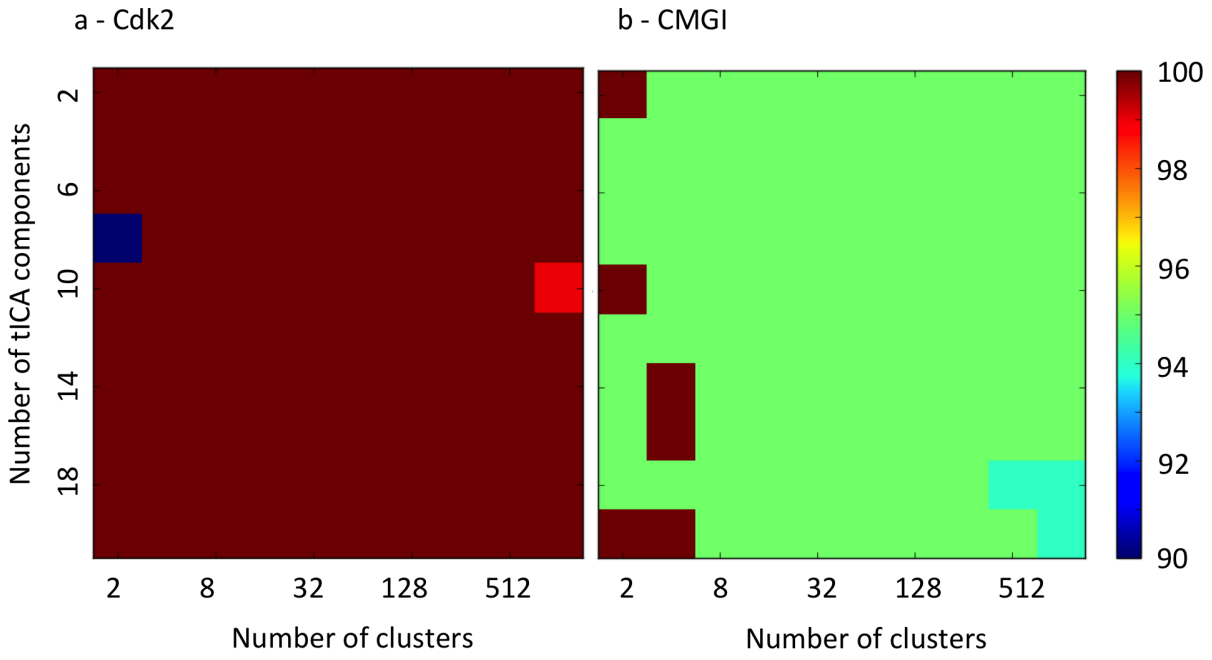

Figure S22: **Population of simulation data used in MSM as a function of number of clusters and tICA components.**

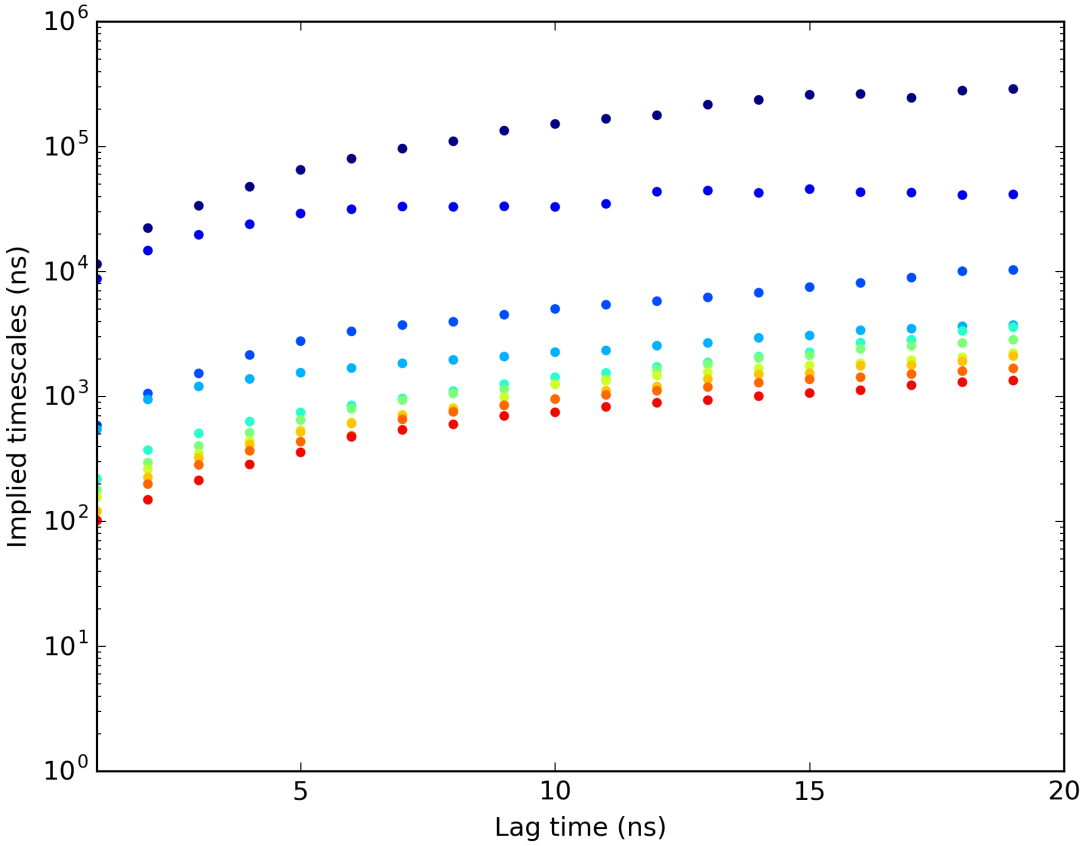

Figure S23: **CDK2's eigenvalues of the transition probability matrix.** The eigenvalues represent the timescales associated with the dynamical processes on the conformational landscape of CDK2 for the 300 state MSM. The longest timescale is in the range of 0.1-1 ms. The convergence of the eigenvalues as a function of the lag time indicates that the MSM is Markovian.

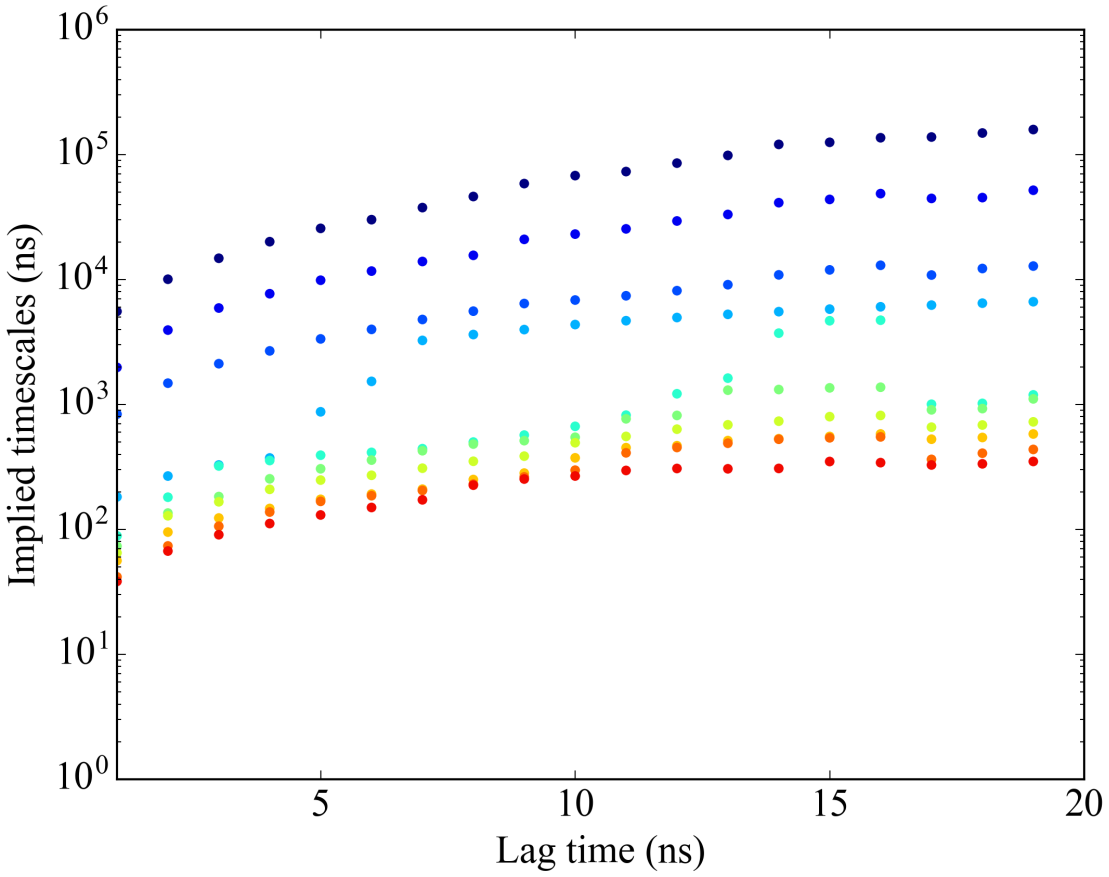

Figure S24: **CMGI's eigenvalues of the transition probability matrix.** The eigenvalues represent the timescales associated with the dynamical processes on the conformational landscape of CMGI for the 1000 state MSM. The longest timescale is in the range of 0.1-0.2 ms. The convergence of the eigenvalues as a function of the lag time indicates that the MSM is Markovian.

References

(1) Hamelberg, D.; Mongan, J.; McCammon, J. A. Accelerated molecular dynamics: A promising and efficient simulation method for biomolecules. *The Journal of Chemical Physics* **2004**, *120*, 11919–11929.

(2) Robert, X.; Gouet, P. Deciphering key features in protein structures with the new ENDscript server. *Nucleic Acids Research* **2014**, *42*, W320–W324.

(3) Wyatt, P. G. et al. Identification ofN-(4-Piperidinyl)-4-(2, 6-dichlorobenzoylamino)-1H-pyrazole-3-carboxamide (AT7519), a Novel Cyclin Dependent Kinase Inhibitor Using Fragment-Based X-Ray Crystallography and Structure Based Drug Design†. *Journal of Medicinal Chemistry* **2008**, *51*, 4986–4999.

(4) Jeffrey, P. D.; Russo, A. A.; Polyak, K.; Gibbs, E.; Hurwitz, J.; Massagué, J.; Pavletich, N. P. Mechanism of CDK activation revealed by the structure of a cyclinA-CDK2 complex. *Nature* **1995**, *376*, 313–320.

(5) Betzi, S.; Alam, R.; Martin, M.; Lubbers, D. J.; Han, H.; Jakkaraj, S. R.; Georg, G. I.; Schonbrunn, E. Discovery of a Potential Allosteric Ligand Binding Site in CDK2. *ACS Chemical Biology* **2011**, *6*, 492–501.

(6) Lawrie, A. M.; Noble, M. E.; Tunnah, P.; Brown, N. R.; Johnson, L. N.; Endicott, J. A. Protein kinase inhibition by staurosporine revealed in details of the molecular interaction with CDK2. *Nature Structural Biology* **1997**, *4*, 796–801.