Supporting information for: Development and Validation of the QUBE Protein Force Field

Alice E. A. Allen,[†] Michael J. Robertson,[‡,¶] Michael C. Payne,[†] and Daniel J. Cole[*,§]

†*TCM Group, Cavendish Laboratory, 19 JJ Thomson Ave, Cambridge CB3 0HE, United Kingdom*

‡*Department of Molecular and Cellular Physiology, Stanford University School of Medicine, 279 Campus Drive, Stanford, California 94305, USA.*

¶*Department of Structural Biology, Stanford University School of Medicine, 279 Campus Drive, Stanford, California 94305, USA.*

§*School of Natural and Environmental Sciences, Newcastle University, Newcastle upon Tyne NE1 7RU, United Kingdom*

E-mail: daniel.cole@ncl.ac.uk

Contents

S1 Preliminary Work

This section summarizes the preliminary work carried out to determine appropriate weighting and regularization values. MD simulations were carried out using the backbone torsional parameters. A set of six dipeptides were tested.

The parameters and the RMSD error values have units of kcal/mol throughout. The error is calculated using eq.5 in the main text with $\lambda = 0$ and no weighting used.

S1.1 Alanine/Glycine Backbone Torsional Parameters

Table S1: The preliminary backbone torsional parameters along with their corresponding errors are stated. This shows the change in the the error and torsional parameter values with the type of weights and strength of regularization applied.

Regularization and Weighting	**Dihedral Angle**	**Torsional Parameters**				**Error QUBE**
$\lambda = 0.00$	ϕ	-0.60	1.18	-2.76	0.00	1.254
No Weighting	ψ	1.26	2.88	-2.08	0.00	
	ϕ'	-3.04	-0.04	1.56	0.00	
	ψ'	1.96	-0.42	0.93	0.00	
$\lambda = 0.01$	ϕ	0.02	0.41	-0.09	0.00	2.880
No Weighting	ψ	0.09	0.43	-0.29	0.00	
	ϕ'	-0.37	-0.16	0.00	0.00	
	ψ'	-0.26	-0.32	-0.24	0.00	
$\lambda = 0.00$	ϕ	-0.67	1.13	-3.02	0.00	1.231
1000K	ψ	1.21	2.70	-1.28	0.00	
	ϕ'	-3.08	-0.07	1.85	0.00	
	ψ'	1.88	-0.50	0.11	0.00	
$\lambda = 0.01$	ϕ	0.08	0.99	-0.33	0.00	1.804
1000K	ψ	0.16	1.00	-0.66	0.00	
	ϕ'	-1.00	-0.32	-0.08	0.00	
	ψ'	-0.34	-0.87	-0.58	0.00	

Regularization and Weighting	**Dihedral Angle**	**Torsional Parameters**				**Error QUBE**
$\lambda = 0.00$ *2000K*	ϕ	0.04	0.75	-0.18	0.00	2.069
	ψ	0.14	0.86	-0.51	0.00	
	ϕ'	-0.80	-0.28	-0.04	0.00	
	ψ'	-0.40	-0.64	-0.46	0.00	
$\lambda = 0.01$ *2000K*	ϕ	0.03	0.60	-0.12	0.00	2.360
	ψ	0.12	0.64	-0.42	0.00	
	ϕ'	-0.60	-0.24	-0.02	0.00	
	ψ'	-0.37	-0.51	-0.37	0.00	

Table S2: The J coupling error for a subset of the dipeptides tested. Side torsional parameters were fit using varying levels of weights and regularization in order to observe the change in the J Coupling error.

Regularization and Weighting	**J coupling Error / Hz**						
	Arg	**Asn**	**Cys**	**Gln**	**Ile**	**Phe**	**RMSE**
$\lambda = 0.00$ *No Weighting*	0.89	0.63	0.60	0.48	0.61	0.16	0.60
$\lambda = 0.01$ *No Weighting*	0.93	0.82	0.39	0.11	0.69	0.01	0.60
$\lambda = 0.00$ *1000K*	0.07	1.47	0.72	0.39	0.36	0.63	0.75
$\lambda = 0.01$ *1000K*	0.50	2.20	1.63	1.18	0.92	1.40	1.41
$\lambda = 0.00$ *2000K*	0.69	0.93	0.38	0.14	0.00	0.22	0.51
$\lambda = 0.01$ *2000K*	0.45	1.62	1.01	0.42	0.04	0.77	0.88

S1.2 Serine Backbone Torsional Parameters

Regularization was shown to be necessary for the serine backbone torsional parameters. As the backbone terms are dependent on the sidechain values, a second iteration of backbone torsional parameter fitting was performed.

Table S3: The change in the J coupling and rotamer error (the error is the difference between the experimental and simulated rotamer populations) with the regularization and weighting scheme used to fit the torsional parameters.

Regularization and Weighting	**Dihedral Angle**	**Torsional Parameters**					**J Coup. Error**	**Rotamer Error**
$\lambda = 0.00$	ϕ	0.94	0.21	-2.93	0.00	*Ser*	4.75	0.32
No Weighting	ψ	-1.81	2.42	2.57	0.00			
	ϕ'	-3.07	4.74	-1.61	0.00			
	ψ'	0.33	2.29	-2.34	0.00			
$\lambda = 0.05$	ϕ	0.62	-0.50	-1.03	0.00	*Ser*	0.66	0.37
No Weighting	ψ	0.09	0.43	0.18	0.00			
	ϕ'	-0.79	1.51	-0.80	0.00			
	ψ'	0.69	0.30	0.20	0.00			
$\lambda = 0.05$	ϕ	0.51	-0.52	-0.96	0.00	*Ser*	0.27	0.11
No Weighting	ψ	0.28	0.4	0.19	0.00	*Thr*	0.22	0.08
Second Iteration	ϕ'	-0.69	1.32	-0.79	0.00			
	ψ'	0.77	0.32	0.31	0.00			

S1.3 Alanine Backbone Torsional Parameters

Table S4: The alanine torsional parameters as a function of the regularization strength used and the corresponding change in the error. The decrease in the magnitude of the torsional parameters can clearly be seen as the regularization strength is increased. OPLS error values are also given for comparison.

Regularization λ Value	**Dihedral Angle**	**Torsional Parameters**				**Error QUBE**	**Error OPLS**
0.00	ϕ	-0.67	1.13	-3.02	0.00	1.231	0.927
	ψ	1.21	2.70	-1.28	0.00		
	ϕ'	-3.08	-0.07	1.85	0.00		
	ψ'	1.88	-0.50	0.11	0.00		
0.10	ϕ	0.08	0.99	-0.33	0.00	1.804	
	ψ	0.16	1.00	-0.66	0.00		
	ϕ'	-1.00	-0.32	-0.08	0.00		
	ψ'	-0.34	-0.87	-0.58	0.00		
0.20	ϕ	0.04	0.75	-0.18	0.00	2.069	
	ψ	0.14	0.86	-0.51	0.00		
	ϕ'	-0.80	-0.28	-0.04	0.00		
	ψ'	-0.40	-0.64	-0.46	0.00		
0.30	ϕ	0.03	0.60	-0.12	0.00	2.360	
	ψ	0.12	0.64	-0.42	0.00		
	ϕ'	-0.60	-0.24	-0.02	0.00		
	ψ'	-0.37	-0.51	-0.37	0.00		
0.40	ϕ	0.02	0.49	-0.09	0.00	2.589	
	ψ	0.10	0.51	-0.34	0.00		
	ϕ'	-0.47	-0.20	-0.01	0.00		
	ψ'	-0.32	-0.41	-0.31	0.00		

Regularization λ Value	**Dihedral Angle**	**Torsional Parameters**				**Error QUBE**	**Error OPLS**
0.50	ϕ	0.02	0.41	-0.07	0.00	2.761	
	ψ	0.09	0.42	-0.29	0.00		
	ϕ'	-0.38	-0.17	0.00	0.00		
	ψ'	-0.28	-0.34	-0.26	0.00		
0.60	ϕ	0.01	0.35	-0.06	0.00	2.889	
	ψ	0.08	0.36	-0.25	0.00		
	ϕ'	-0.33	-0.14	0.00	0.00		
	ψ'	-0.24	-0.29	-0.22	0.00		

Table S5: The change in the J coupling error for Ala_5 with the regularization used in the torsional parameter fitting process. The value in brackets excludes the $J^3(H_N, C_\beta)$ coupling value which is known to be difficult to reproduce with simulation.

Regularization λ Value	**J Coupling Error**		
	Set 1	**Set 2**	**Set 3**
0.00	7.74(5.57)	7.62(7.29)	9.33(7.95)
0.10	3.67(2.64)	2.41(2.79)	3.36(3.45)
0.20	2.26(1.71)	1.96(1.68)	1.77(2.06)
0.30	1.16 (1.00)	3.03(0.87)	1.25(1.04)
0.40	0.97(0.89)	3.74(0.81)	1.38(0.90)
0.50	0.87(0.80)	3.59(0.63)	1.28(0.78)
0.60	0.93(0.90)	4.15(0.82)	1.54(0.90)

S2 Methods

S2.1 Simulation Details

Simulations of the dipeptides, the alanine and glycine peptides, and the protein structures were all carried out with the following simulation setup. The molecules were placed in a cubic water box, with the total length ranging from 25 Å, for small dipeptides, to 58 Å, for the largest protein. The TIP3P force field is used to model water molecules.[S1,S2] The charge of the system was then neutralized with chlorine and sodium ions at a salt concentration of around 150 mM. The temperature was set to 300 K and the pressure to 1 atm. These parameters were maintained throughout the simulation by using a Nose-Hoover Langevin piston barostat. The piston period was set to 100 fs and the damping timescale was 500 fs. The Langevin thermostat had a damping coefficient of 1ps^{-1}. Long ranged electrostatics were treated with particle mesh Ewald. The cutoff for non-bonded terms was set to 11Å with the smoothing function starting at 9Å.

The system was first equilibrated for 5ns using a timestep of 2fs. The simulation was then run for a further 200 ns, again using a timestep of 2fs with use of the SHAKE and SETTLE algorithms. These properties are the same as those used in Ref. S3. The protein simulations additionally employed a 400ps equilibration period in which the system was gradually heated to 300K. This heating period and the 5ns of equilibration were not used in the simulation analysis.

S2.2 ONETEP Calculation Setup

Four nonorthogonal generalized Wannier functions (NGWFs) were used for all atoms with the exception of hydrogen which used one. The NGWFs had radii of 10 Bohr. The periodic cardinal sine (psinc) basis was used to describe the NGWFs, with a grid size employed that corresponds to a plane wave cutoff energy of 1020 eV. Convergence of DDEC atomic charges with psinc cutoff energy and NGWF radius is reported in Ref. S4. The PBE

exchange-correlation functional was used with PBE OPIUM norm-conserving pseudopotentials.[S5] Partitioning of the polarized ground-state electron density was performed using the DDEC scheme in ONETEP with the mixing parameter (γ) set to 0.02. Additionally, hydrogen atoms attached to polar atoms were assigned Lennard-Jones terms equal to zero, with the polar atoms having a new B coefficient given by:

$$\sqrt{B'_x} = \sqrt{B_x} + n_H \sqrt{B_H} \tag{1}$$

where B'_x, B_x are the new and old terms respectively, n_H is the number of hydrogen atoms and B_H is the dispersion coefficient on the hydrogen atom.

S2.3 Bond and Angle Parameters for Cysteine

The S–S bond parameters and S–S–CT angle parameters needed to describe the disulfide bridge between two cysteine residues were derived from dimethyl disulfide. A structural optimization and frequency calculation was performed using Gaussian 09 with the ωB97XD functional and 6-311++G(d,p) basis set. The QM Hessian matrix was extracted from the Gaussian output files and the modified Seminario method used to find the bond and angle terms. A scaling factor of $(0.957)^2$ was used to scale the bond and angle terms, as discussed in Ref S6. The terms calculated were for S–S, a force constant of 146.29 kcal/mol/Å^2 and bond length of 2.070Å, and for S–S–CT, a force constant 99.50 kcal/mol/rad^2 and equilibrium angle of 102.05°.

S2.4 Experimental NMR J Couplings

The dipeptide experimental J coupling values for the backbone are taken from Ref. S7 in which coupling constants were measured for blocked dipeptides. Experimental J coupling terms for ubiquitin and GB3 were extracted from a number of sources and all the values used are listed in the SI of Ref. S3.[S8–S12] Within Ref. S8 experimental measurement were collated,

with ubiquitin measurement supplied there from unpublished work by Ad Bax and John L. Marquardt and GB3 experimental values taken from Ref. S13. Sidechain experimental measurements for the methyl bearing side chains were taken from Ref. S11 with further sidechain values for ubiquitin taken from Ref. S9.

S3 Torsional Parameters

The energies, parameters and the RMSD error have units of kcal/mol throughout.

S3.1 QM Energy Scans

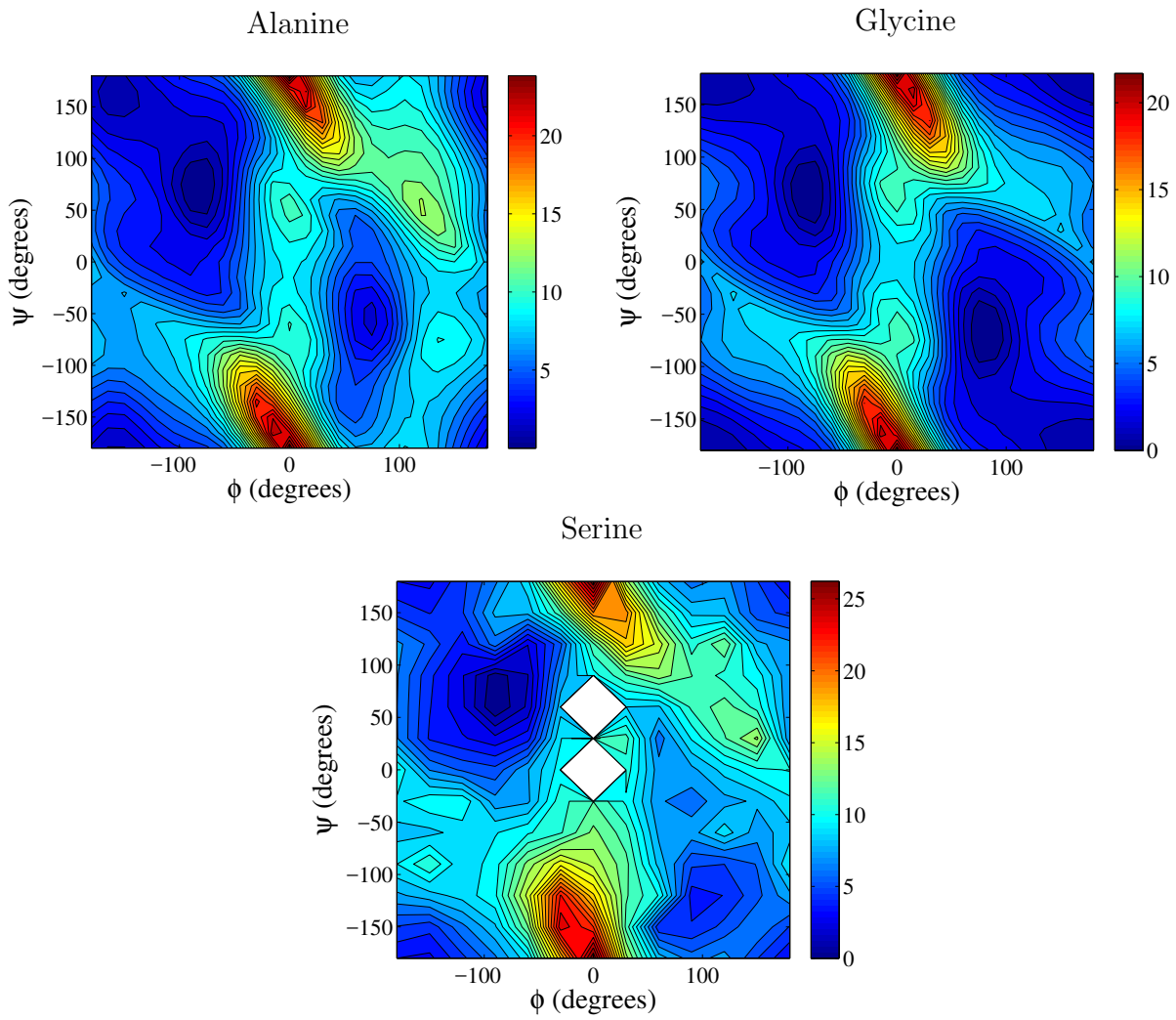


Figure S1: The 2D energy scan of ψ and ϕ dihedral angles for a set of dipeptides (units of kcal/mol). For serine the minimum value of the three sidechain scans is shown. The white areas correspond to regions with unavailable data due to convergence issues with the QM calculations.

S3.2 Backbone Torsional Parameters

This section details the final torsional parameters and corresponding errors . The error is calculated using eq.5 in the main text with $\lambda = 0$ and no weighting used. The error shown uses all the scans included in the OPLS-AA/M fitting process, even when the QUBE terms are fit to a subset of scans. The OPLS-AA/M force field is fit using a 2000K QM weighting scheme, whilst the error used here is not weighted.

Table S6: The backbone torsional parameters used in the QUBE force field.

Dipeptide Name	**Dihedral Angle**	**Torsional Parameters**				**Error QUBE**	**Error OPLS**
Ala/Gly							
λ= 0.00	ϕ	-0.60	1.18	-2.76	0.00	1.254	0.929
	ψ	1.26	2.88	-2.08	0.00		
	ϕ'	-3.04	-0.04	1.56	0.00		
	ψ'	1.96	-0.42	0.93	0.00		
λ= 0.50	ϕ	0.02	0.41	-0.09	0.00	2.880	
	ψ	0.09	0.43	-0.29	0.00		
	ϕ'	-0.37	-0.16	0.00	0.00		
	ψ'	-0.26	-0.32	-0.24	0.00		
Ala							
λ= 0.00	ϕ	-0.67	1.13	-3.02	0.00	1.231	0.927
	ψ	1.21	2.70	-1.28	0.00		
	ϕ'	-3.08	-0.07	1.85	0.00		
	ψ'	1.88	-0.50	0.11	0.00		
λ= 0.50	ϕ	0.02	0.41	-0.07	0.00	2.761	
	ψ	0.09	0.42	-0.29	0.00		
	ϕ'	-0.38	-0.17	0.00	0.00		
	ψ'	-0.28	-0.34	-0.26	0.00		

Dipeptide Name	**Dihedral Angle**	**Torsional Parameters**				**Error QUBE**	**Error OPLS**
Gly							
λ= *0.00*	ϕ	-0.38	1.39	-2.01	0.00	1.307	
	ψ	1.33	4.32	-1.62	0.00		
λ= *0.50*	ϕ	0.07	0.42	-0.37	0.00	4.008	
	ψ	0.14	0.52	-0.33	0.00		
Pro							
λ= *0.00*	ψ	2.11	0.95	-2.36	0.00	1.128	1.116
	ψ'	1.57	-1.09	1.32	0.00		
λ= *0.50*	ψ	0.10	0.22	-0.21	0.00	1.598	
	ψ'	-0.01	-0.28	-0.17	0.00		
Ser							
Reg=0.05	ϕ	0.51	-0.52	-0.96	0.00	2.541	2.834
	ψ	0.28	0.4	0.19	0.00		
	ϕ'	-0.69	1.32	-0.79	0.00		
	ψ'	0.77	0.32	0.31	0.00		
λ= *0.50*	ϕ	0.08	-0.16	-0.13	0.00	3.109	
	ψ	-0.03	0.00	0.12	0.00		
	ϕ'	-0.02	0.19	-0.12	0.00		
	ψ'	0.17	0.12	0.12	0.00		

S3.3 Sidechain Torsional Parameters

Table S7: The sidechain torsional parameters used in the QUBE force field with the manual changes to the fitting process stated.

Dipeptide Name	**Dihedral Angle**	**Torsional Parameters**				**Error QUBE**	**Error OPLS**
Arg	X1	1.91	-0.29	0.56	0.00	0.606	0.662
	X1'	-1.1	-0.35	0.44	0.00		
	X2	-2.10	-0.50	0.00	0.00	2.302	0.759
Asn	X1	-7.51	0.53	-0.56	0.00	1.904	1.246
	X1'	4.79	0.60	0.21	0.00		
	X2	0.81	0.51	0.00	0.00	1.819	0.759
	X2'	1.59	-0.10	-0.10	0.00		
Asp	X1	-8.04	-0.57	0.73	0.00	1.244	1.807
β_{180} *only*	X1'	3.58	0.73	-0.12	0.00		
Set to zero	X2	0.00	0.00	0.00	0.00	1.897	
Cys	X1	3.31	0.29	1.72	0.00	1.143	1.062
β_{60} *only*	X1'	-2.81	-0.03	-0.09	0.00		
	X2	0.93	2.90	-0.32	0.00	0.661	0.813
	X2'	0.78	1.04	2.09	0.00		

Dipeptide Name	Dihedral Angle	Torsional Parameters				Error QUBE	Error OPLS
Gln	X1	2.13	-0.37	1.58	0.00	0.947	0.910
	X1'	-1.62	0.24	0.08	0.00		
	X2	-1.16	1.01	-0.06	0.00	2.884	1.384
Glu	X1	-0.57	-1.45	1.33	0.00	3.120	2.132
β Scans only	X1'	-2.26	-1.87	1.04	0.00		
	X2	-8.81	-0.11	-0.80	0.00	3.105	2.819
Hid	X1	-0.70	0.19	0.50	0.00	1.238	1.177
	X1'	0.20	0.60	0.20	0.00		
	X2	2.91	-1.61	0.40	0.00	1.523	
Hie	X1	-3.12	0.39	0.92	0.00	0.803	1.113
	X1'	1.90	0.50	0.10	0.00		
	X2	-1.19	-0.50	0.03	0.00	0.749	
Hip	X1	-1.31	-0.37	0.98	0.00	1.482	1.747
β Scans only	X1'	-0.28	-0.91	0.01	0.00		
	X2	-0.38	0.50	0.19	0.00	1.135	
Ile	X1	7.51	-0.59	-0.21	0.00	0.694	1.066
	X1'	5.01	-0.19	0.99	0.00		
	X2	0.10	-0.14	0.13	0.00	0.855	0.661

Dipeptide Name	**Dihedral Angle**	**Torsional Parameters**				**Error QUBE**	**Error OPLS**
Leu	X1	-2.08	-0.65	0.58	0.00	0.797	1.238
	X1'	0.53	0.34	0.12	0.00		
	X2	-1.09	0.71	0.20	0.00	0.623	0.733
Lys	X1	0.00	0.00	0.00	0.00	1.683	1.024
	X1'	0.00	0.00	0.00	0.00		
OPLS-AA/M	X2	1.30	-0.20	0.20	0.00	2.689	
Met	X1	0.19	-0.83	-0.74	0.00	0.916	0.553
β_{-60} *only*	X1'	0.81	0.03	1.63	0.00		
	X2	-1.15	-0.45	-0.23	0.00	1.586	1.196
Phe	X1	-0.94	0.76	0.91	0.00		
	X1'	0.46	0.58	0.32	0.00	0.576	0.990
	X2	-5.21	0.19	0.00	0.00	0.828	0.616
Pro	X1	2.11	0.95	-2.36	0.00	1.128	1.116
	X1'	1.57	-1.09	1.32	0.00		
Ser	X1	5.32	0.75	0.23	0.00	1.526	1.388
β_{60} *only*	X1'	-6.93	0.95	1.17	0.00		
	X2	0.21	-1.21	0.41	0.00	0.961	0.929

Dipeptide Name	**Dihedral Angle**	**Torsional Parameters**				**Error QUBE**	**Error OPLS**
Thr	X1	-3.34	1.18	-1.72	0.00	1.879	1.455
β_{60} *only*	X1'	-3.87	3.12	-0.05	0.00		
	X1	3.64	2.84	3.91	0.00		
	X1'	-1.61	2.13	-0.89	0.00		
Trp	X1	1.09	0.59	0.88	0.00	0.661	0.970
	X1'	-0.40	0.90	0.12	0.00		
	X2	0.48	-0.58	-0.19	0.00	0.875	1.295
Tyr	X1	-0.12	0.70	1.07	0.00	0.764	1.014
	X1'	-0.13	1.02	-0.01	0.00		
	X2	-4.40	0.19	0.59	0.00	0.864	0.662
Val	X1	3.81	-0.55	0.79	0.00	0.540	0.550
	X1'	-1.74	-0.11	0.03	0.00		
Mean Error						1.29	1.12

S4 Dipeptides

S4.1 J Coupling Values

Table S8: The J coupling error for the set of dipeptides.

Dipeptide Name	Backbone J Coupling MUE Hz			
	First Run	Second Run	Third Run	Average
Arg	1.17	1.15	0.81	1.04
Asn	0.36	0.41	0.36	0.38
Asp	0.10	0.07	0.07	0.08
Cys	0.43	0.46	0.45	0.45
Gln	0.42	0.42	0.45	0.43
Glu	0.39	0.26	0.18	0.28
Hid	N/A	N/A	N/A	N/A
Hie	N/A	N/A	N/A	N/A
Hip	0.50	0.44	0.53	0.49
Ile	0.71	0.67	0.68	0.69
Leu	0.10	0.07	0.09	0.09
Lys	0.22	0.24	0.24	0.23
Met	0.38	0.45	0.39	0.41
Phe	0.09	0.06	0.04	0.06
Ser	0.27	0.38	0.31	0.32
Thr	0.22	0.12	0.09	0.14
Trp	0.14	0.16	0.14	0.15
Tyr	0.04	0.01	0.05	0.03
Val	0.47	0.44	0.41	0.44
RMSE				0.42

Table S9: The error in the rotamer populations for the set of dipeptides.

Dipeptide Name	**Sidechain Rotamer Error**			
	First Run	**Second Run**	**Third Run**	**Average**
Arg	0.08	0.09	0.10	0.09
Asn	0.17	0.16	0.17	0.17
Asp	0.28	0.29	0.28	0.28
Cys	0.08	0.11	0.13	0.11
Gln	0.16	0.15	0.15	0.15
Glu	0.31	0.17	0.10	0.19
Hid	0.10	0.16	0.10	0.12
Hie	0.13	0.13	0.06	0.11
Hip	0.26	0.24	0.23	0.24
Ile	0.13	0.11	0.13	0.12
Leu	0.10	0.11	0.10	0.10
Lys	0.14	0.13	0.17	0.15
Met	0.13	0.15	0.14	0.14
Phe	0.07	0.09	0.08	0.08
Ser	0.11	0.09	0.07	0.09
Thr	0.08	0.08	0.10	0.09
Trp	0.14	0.16	0.19	0.16
Tyr	0.08	0.09	0.08	0.08
Val	0.18	0.20	0.21	0.20
MUE				0.14

S4.2 Dihedral Angle Distributions ϕ and ψ

The dihedral distributions are given in this section. Regions of low probability are shown in dark red as are regions that are not sampled during the simulation.

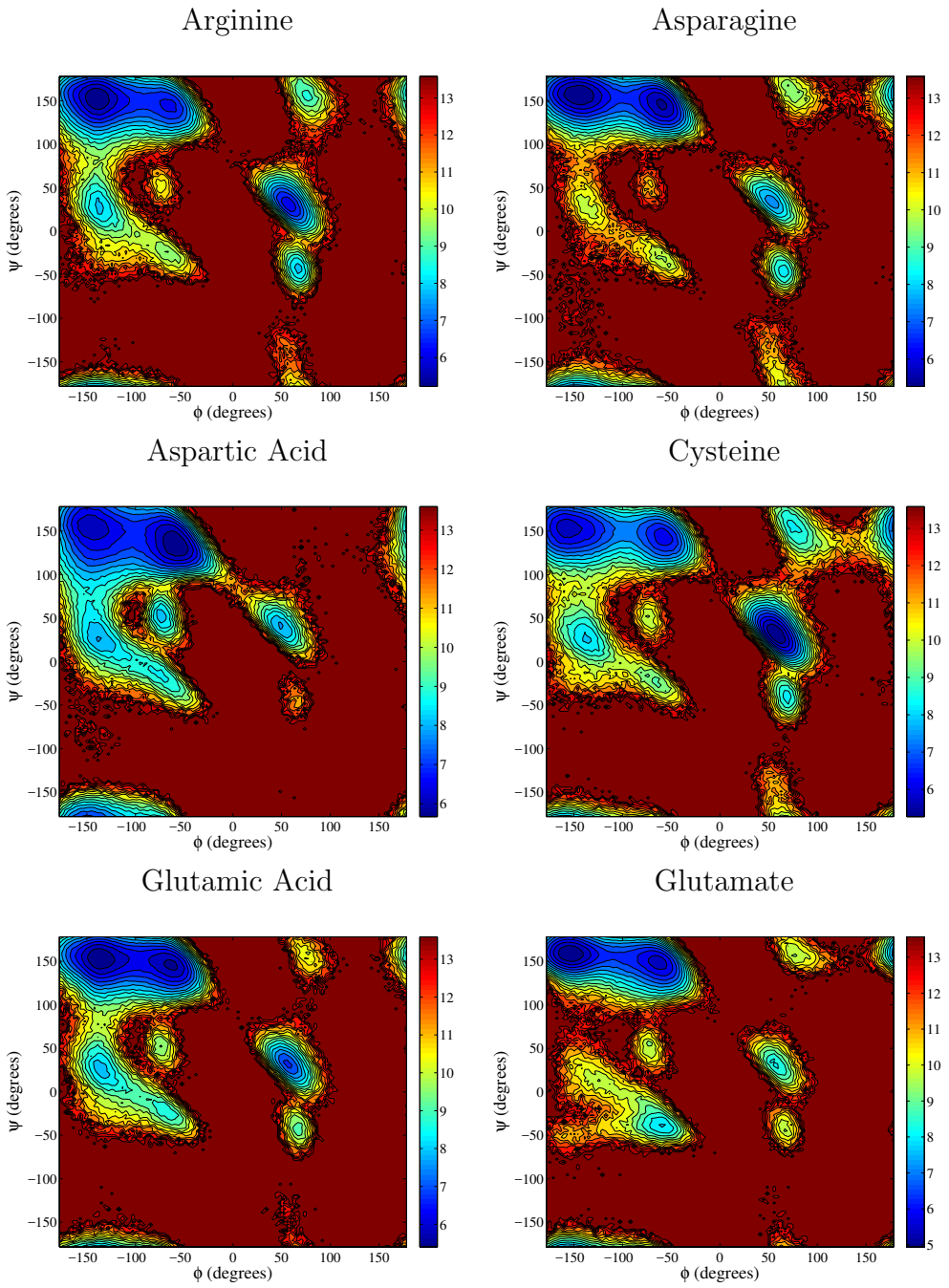


Figure S2: The ψ and ϕ distribution for the dipeptide simulations, plotted in the form $-log(p_{\psi,\phi})$ (where $p_{\psi,\phi}$ is the probability of a region being occupied). The dark red regions correspond to low probability areas including conformations that are not sampled during the simulation.

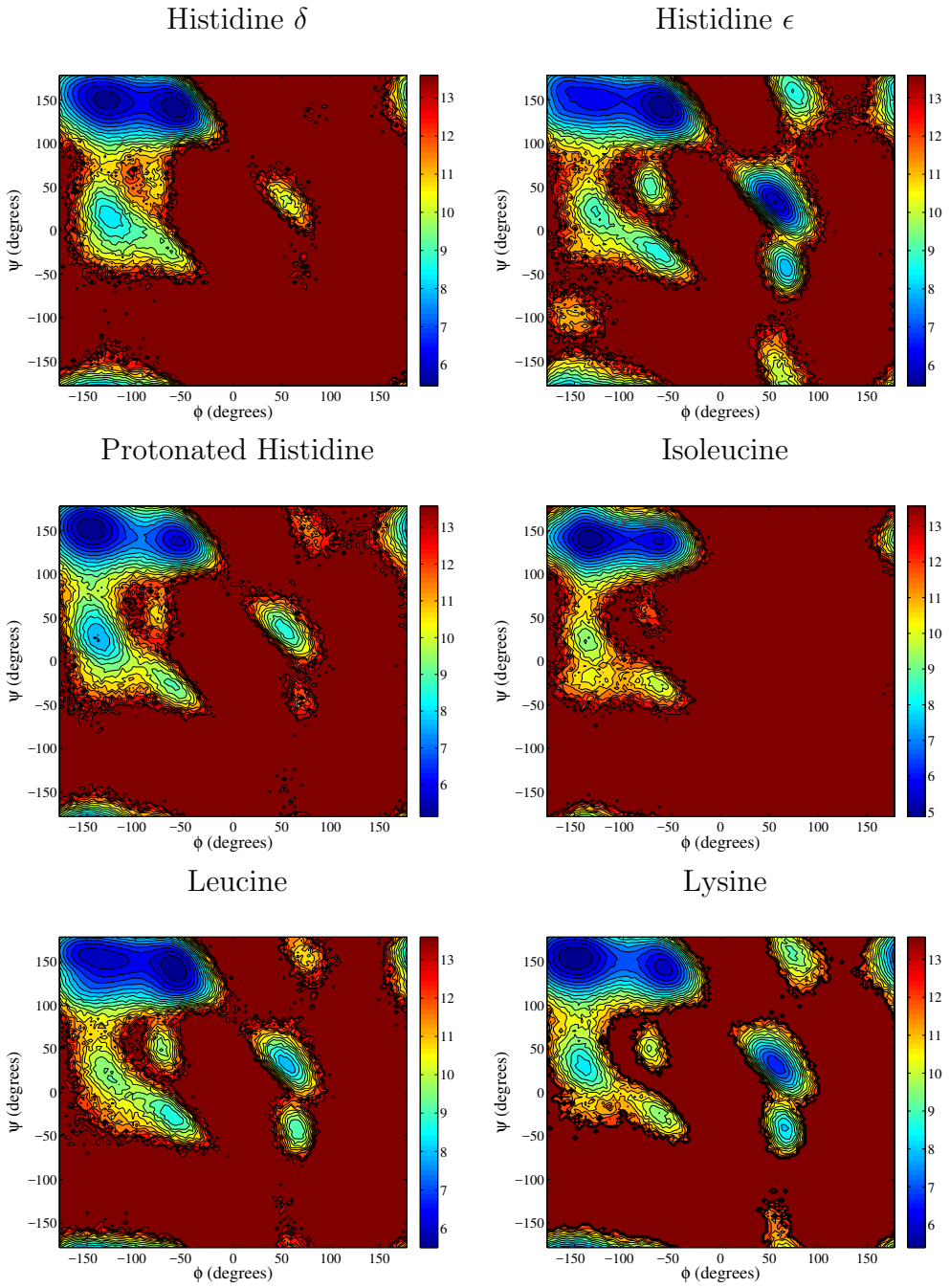


Figure S2 continued.

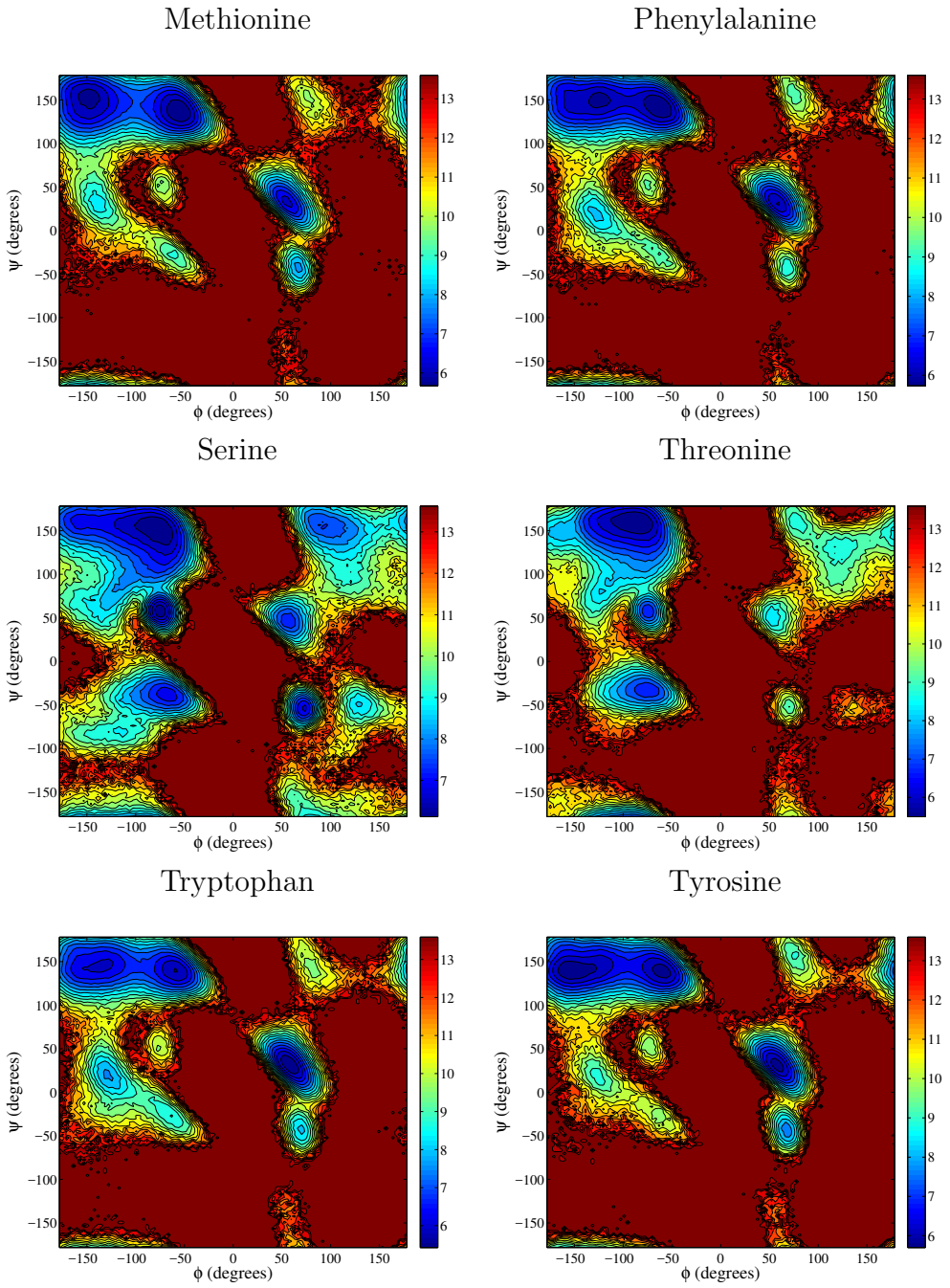


Figure S2 continued.

Valine

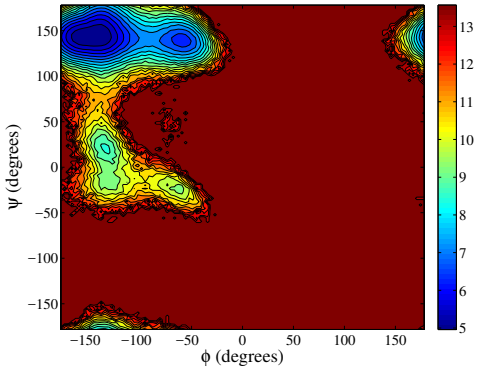


Figure S2 continued.

S5 Peptides

S5.1 J Coupling Errors

The backbone J coupling errors and conformations populated by Ala_5 are presented in this section.

Table S10: The J coupling error for Ala_5 using the three Karplus parameters from Ref. S14.

a) First Set of Karplus Parameters

Peptide Name	**J Coupling Error**			
	Run 1	**Run 2**	**Run 3**	**Average**
Ala_5	0.93(0.89)	0.91(0.87)	0.87(0.83)	0.90 ± 0.03 (0.86 ± 0.03)
Gly_3	5.69	5.90	5.48	5.69 ± 0.21

b) Second Set of Karplus Parameters

Peptide Name	**J Coupling Error**			
	Run 1	**Run 2**	**Run 3**	**Average**
Ala_5	4.16(0.84)	4.16(0.82)	4.15(0.78)	4.16 ± 0.01(0.81 ± 0.03)
Gly_3	7.13	7.25	6.96	7.11 ± 0.15

a) Third Set of Karplus Parameters

Peptide Name	**J Coupling Error**			
	Run 1	**Run 2**	**Run 3**	**Average**
Ala_5	1.52(0.89)	1.52(0.88)	1.49(0.83)	1.51 ± 0.02(0.87 ± 0.03)
Gly_3	6.82	6.96	6.61	4.52 ± 0.18

Table S11: The population of each conformation present in the MD simulation of Ala_5

Peptide Name	**Conformation Percentages**		
	PPII	$\boldsymbol{\beta}$	$\boldsymbol{\alpha}$
Ala_5	62.10 ± 2.23	23.5 ± 1.20	12.93 ± 2.82
Gly_3	45.93 ± 1.46	29.77 ± 7.74	24.30 ± 8.57

S6 Protein Non-Bonded Parameters

The QUBE and OPLS non-bonded parameters for each atom of the five proteins tested are compared.

S6.1 Comparisons between QUBE and OPLS

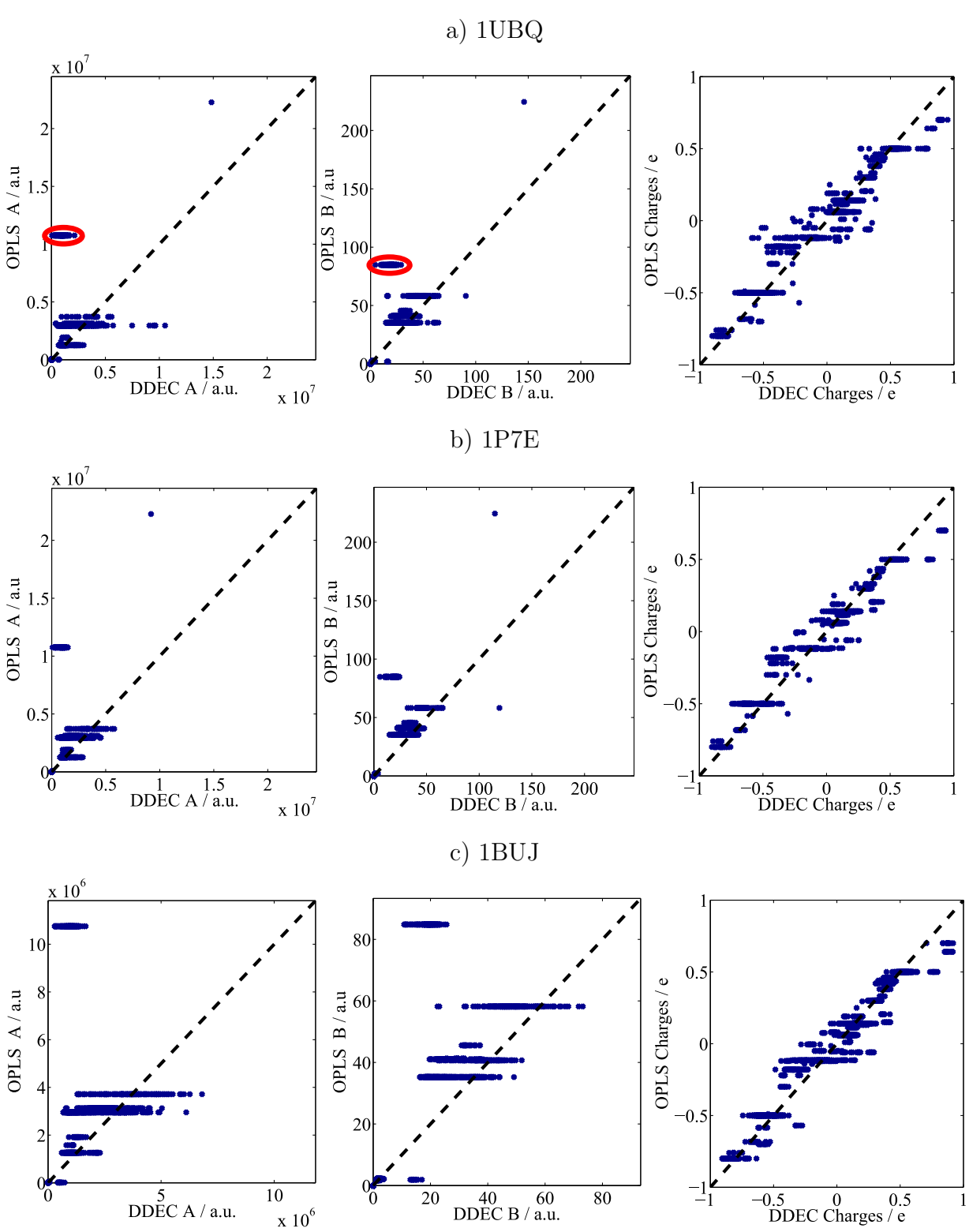
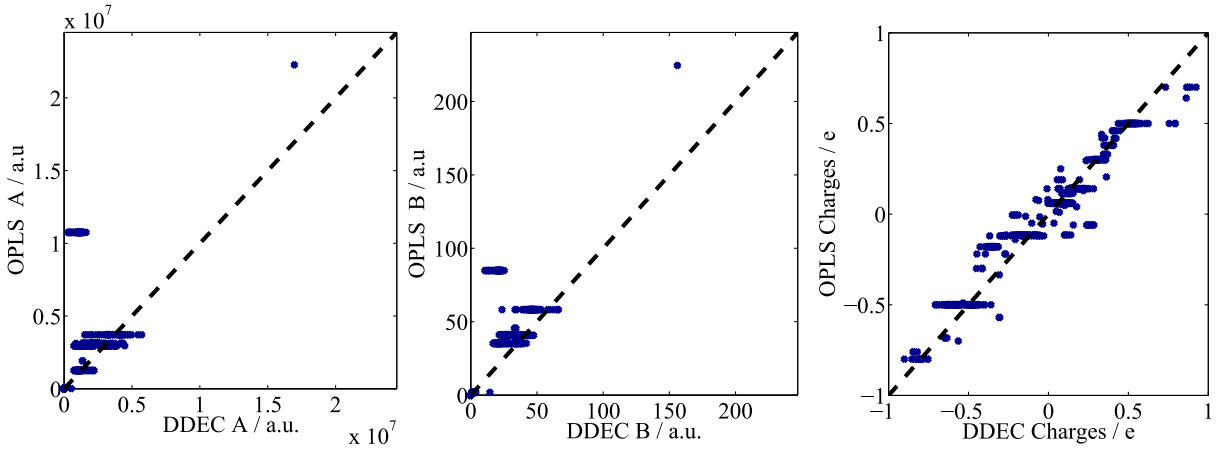


Figure S3: A comparison between the QUBE and OPLS non-bonded parameters. The dashed line corresponds to $y = x$.

d) 2F4K



e) 5PTI

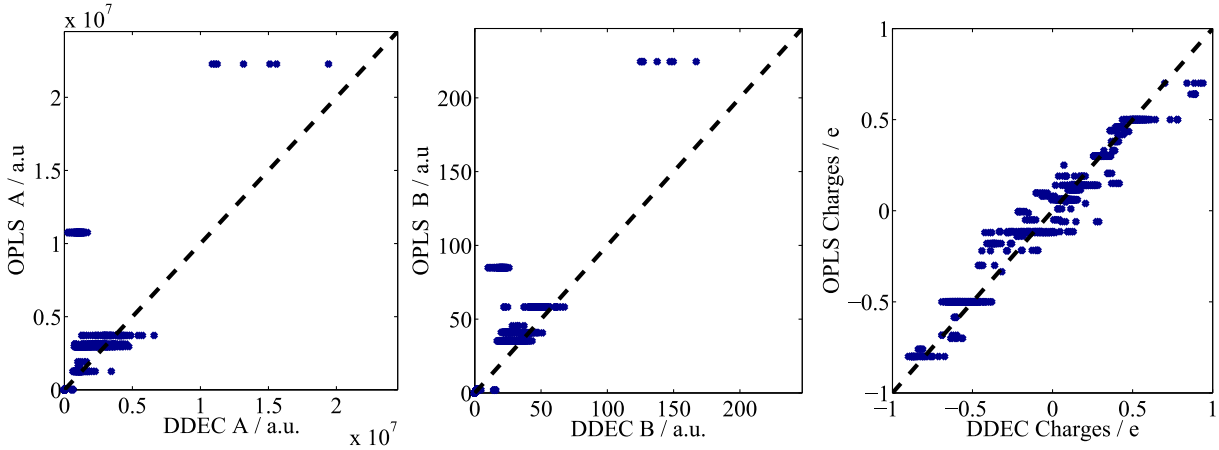


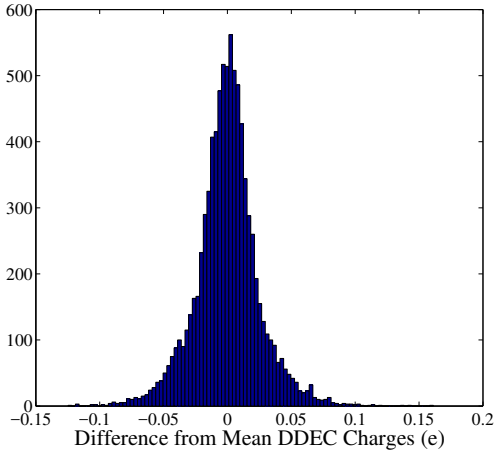
Figure S3 continued.

S6.2 Change in Non-bonded Parameters with Protein Conformation

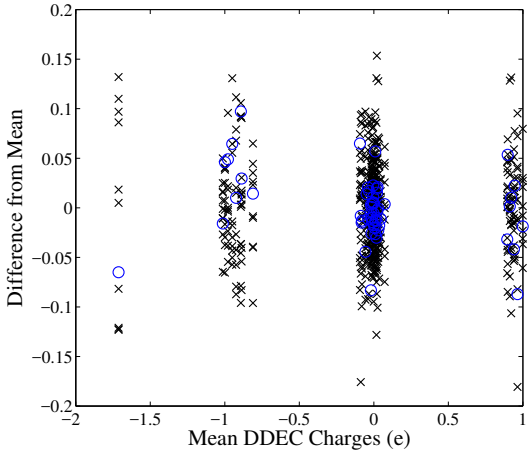
The non-bonded parameters used in the QUBE force field are calculated for one specific input structure. However, during the course of the MD simulation many different conformations will occur. In order to study the change in non-bonded parameters with the structure, a short MD simulation of GB3 in a water box was carried out using OPLS-AA/M. Ten structures were then extracted from this simulation and the non-bonded parameters for

these ten structure were calculated.

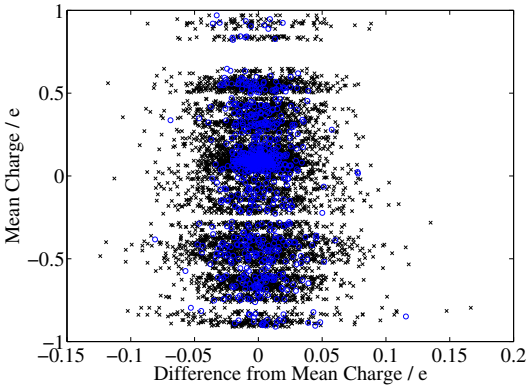
a) Charge Distribution



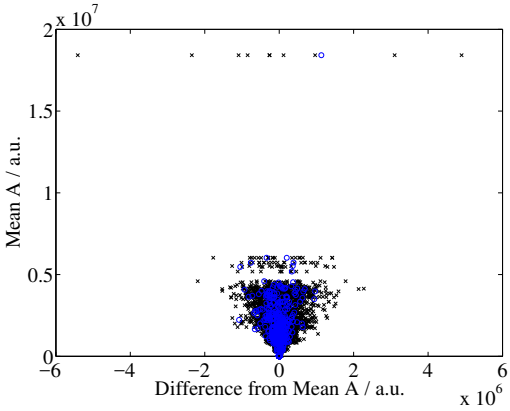
b) Charge per Residue Difference



c) Charge Difference



d) A_i Difference



e) B_i Difference

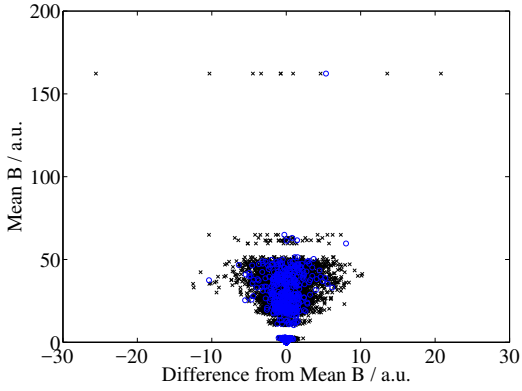


Figure S4: The change in non-bonded QUBE parameters with structure. The black crosses correspond to the ten structures from the MD simulation whilst the blue circles are the non-bonded terms for the structure minimized in water.

Figure S4a) shows the distribution of the atom charges about their mean values. It can be seen that some charges show discrepancy from the mean values, with a maximum difference of 0.161e. However, as the distribution has a standard deviation of 0.0243e the majority of atomic charges do not change considerably. This shows that whilst the non-bonded parameters do respond to their conformation the changes are not excessive, which would be problematic. Figure S4c) demonstrates that this difference is independent of the mean charge. Both figures clearly show the importance of considering the input structure used. The net charge of each residue is considered in Fig. S4b) and shows the charge fluctuations are not entirely due to a redistribution of the electron density within the residue.

The difference in the Lennard-Jones parameters from the mean value is also considered in Fig. S4. The non-bonded parameters clearly change slightly with the input structure. Again, this shows the importance of the input structure used for the ONETEP calculation.

Table S12: The standard deviation of the difference between the non-bonded parameters from one structure and the mean values of the non-bonded parameters.

	Run 1	**Run 2**	**Run 3**	**Run 4**	**Run 5**	**Run 6**
Charge / e	0.026	0.026	0.023	0.026	0.025	0.025
$\mathbf{A_i}$ / a.u.	2.93×10^5	2.37×10^5	2.87×10^5	2.58×10^5	2.35×10^5	2.33×10^5
$\mathbf{B_i}$ / a.u.	1.87	1.74	1.81	1.83	1.68	1.65

	Run 7	**Run 8**	**Run 9**	**Run 10**	**Minimum**
Charge / e	0.025	0.025	0.025	0.025	0.017
$\mathbf{A_i}$ / a.u.	2.44×10^5	2.14×10^5	2.34×10^5	2.44×10^5	1.71×10^5
$\mathbf{B_i}$ / a.u.	1.69	1.54	1.59	1.70	1.29

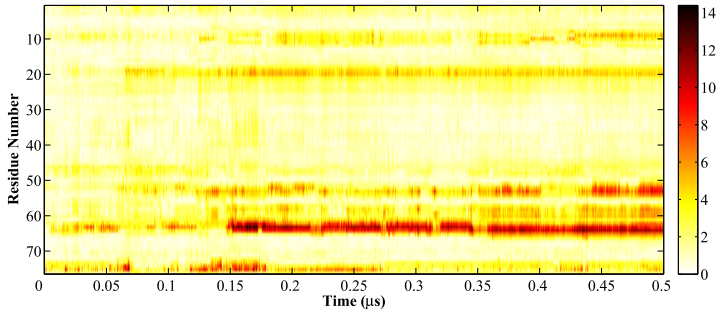
The standard deviation of the difference in the nonbonded parameters from the mean, is shown in Table S12. A low standard deviation is considered preferable as we require nonbonded terms to be representative of as large a range of structures as possible, and therefore close to the mean values. For all three coefficients, the minimum structure non-bonded parameters have the lowest standard deviation of all of the input structures. This is encouraging as it shows that the minimum structure does not have many extreme values and the standard deviations are low.

To choose the most appropriate input structure, it is important to consider not only the accuracy of the non-bonded terms but also the time required to find these non-bonded terms. If for every protein force field we wanted to create, it was necessary to take multiple snapshots of a MD simulation, calculate the non-bonded values and then average these terms, the cost of the QM calculations would start to become prohibitive. Although, as the computational cost of these calculations decreases further this may be the preferred option. Additionally, taking structures from an MD simulation will weight the structures sampled according to the force field used to carry out the simulation. Therefore, we chose the input structure to the ONETEP calculation to be the experimental structure minimised with OPLS-AA/M in a water box. This provides a quick and consistent way to obtain an input structure. In preliminary work, the mean non-bonded parameters of the ten structures of GB3 were used to perform a 200 ns MD simulation. The conformational preference of the molecule did not greatly vary using the minimized structure as the input.

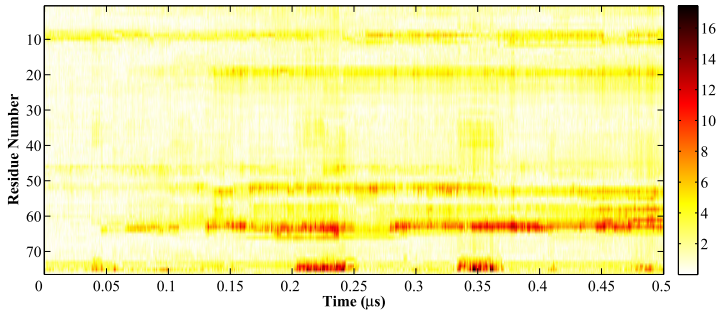
S7 Protein MD Results

S7.1 1UBQ

a) Run 1



b) Run 2



c) Run 3

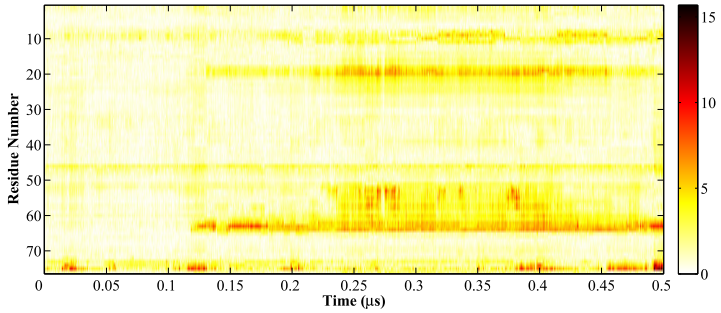


Figure S5: The RMSD per residue ($\mathring{A}$), relative to the crystal structure with PDB code 1UBQ, of three simulations of the ubiquitin protein.

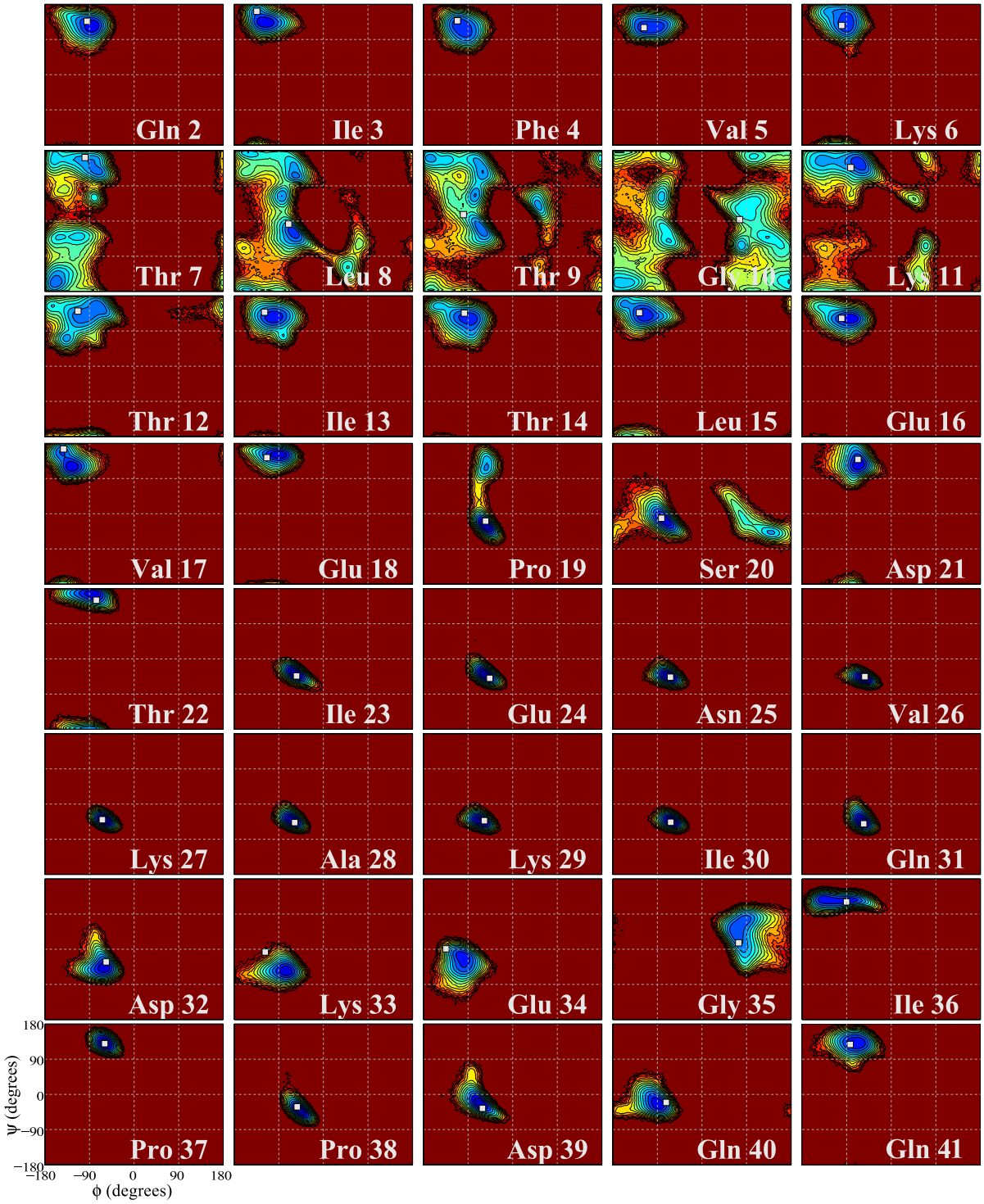


Figure S6: The ψ and ϕ distributions observed in the simulation of the ubiquitin protein, plotted in the form $-log(p_{\psi,\phi})$. The white square corresponds to the experimental structure given in the PDB file 1UBQ. The dark red regions correspond to low probability regions that include conformations that are not sampled during the simulation.

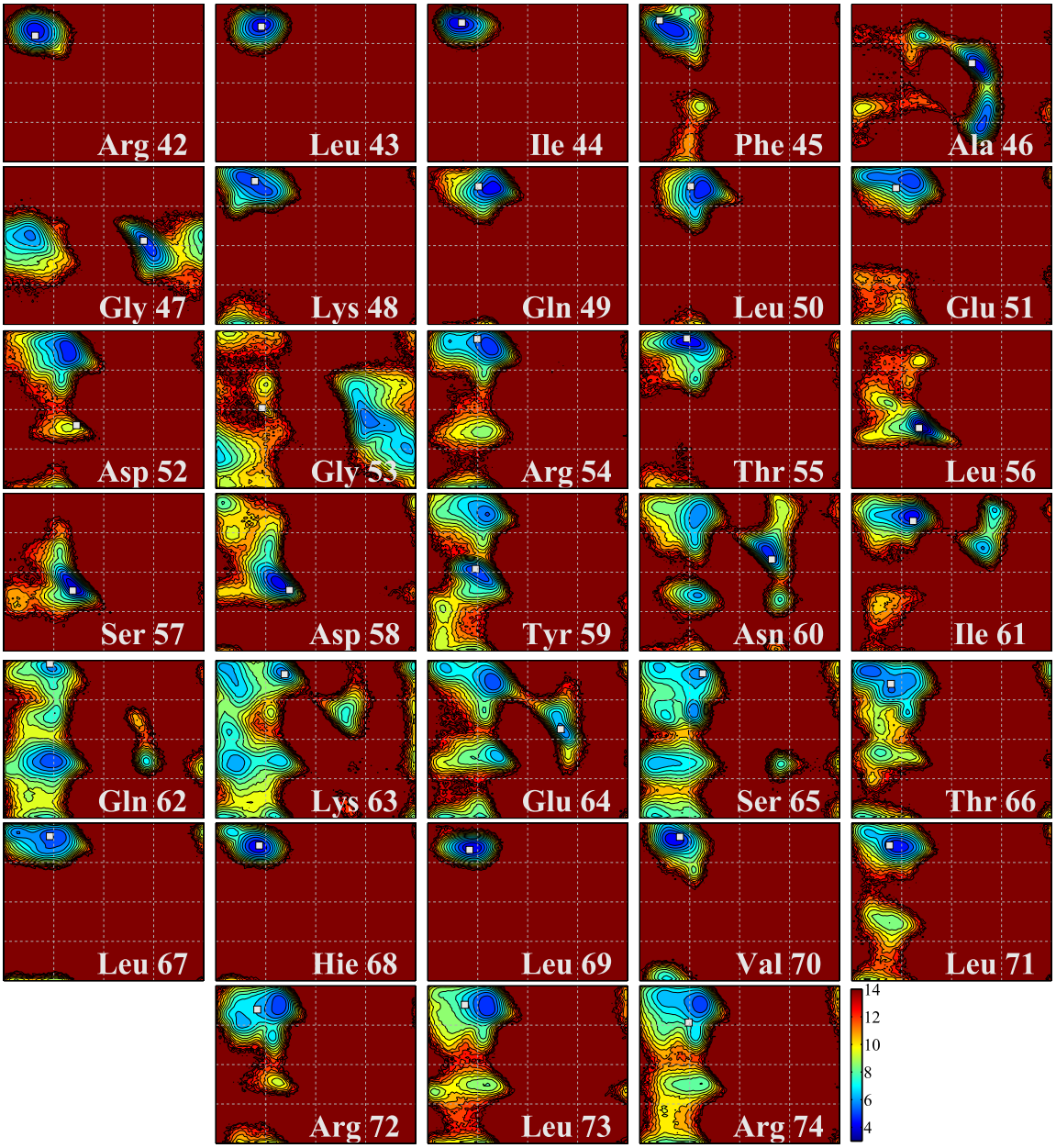
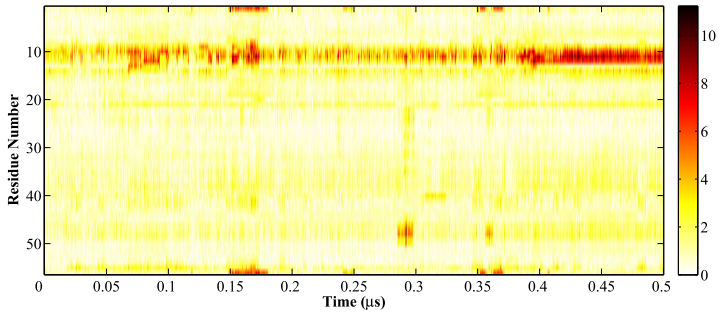


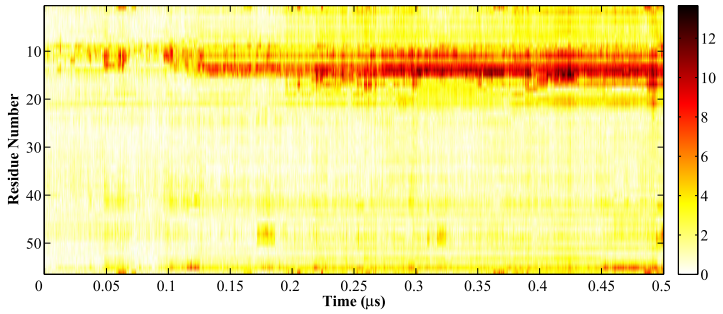
Figure S6 continued.

S7.2 1P7E

a) Run 1



b) Run 2



c) Run 3

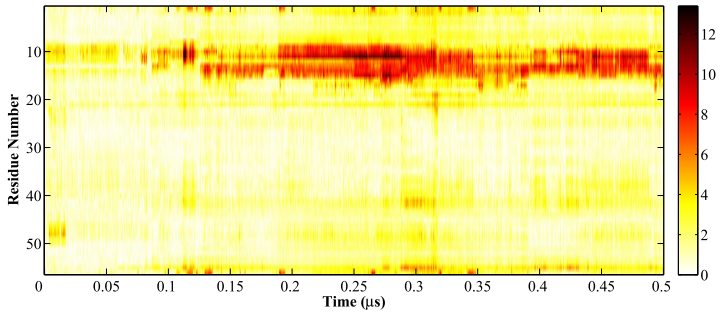


Figure S7: The RMSD per residue, relative to the crystal structure with PDB code 1P7E, of three simulations of GB3.

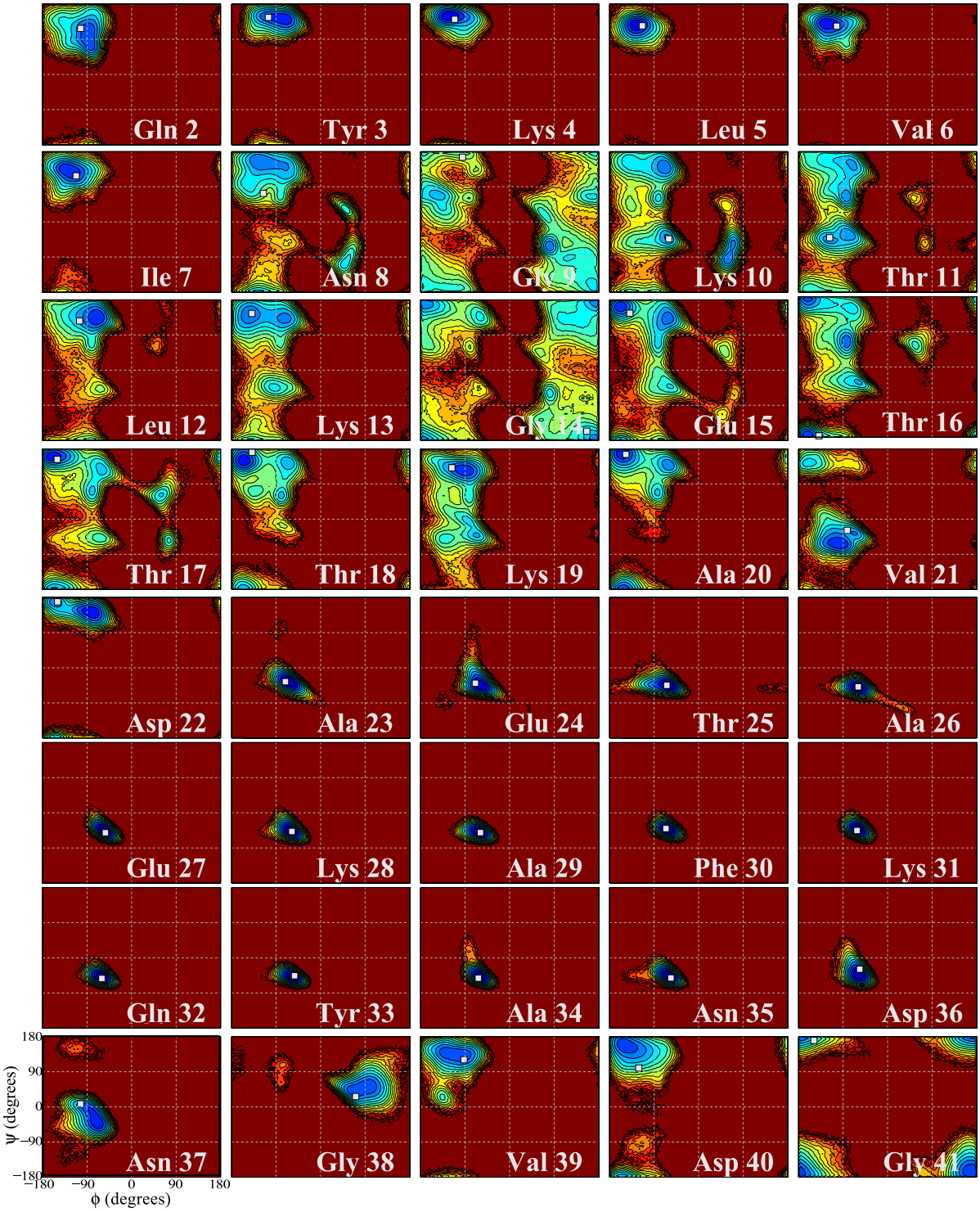


Figure S8: The ψ and ϕ distributions observed in the simulation of the GB3 protein, plotted in the form $-log(p_{\psi,\phi})$. The white square corresponds to the experimental structure given in the PDB file 1P7E. The dark red regions correspond to low probability regions that include conformations that are not sampled during the simulation.

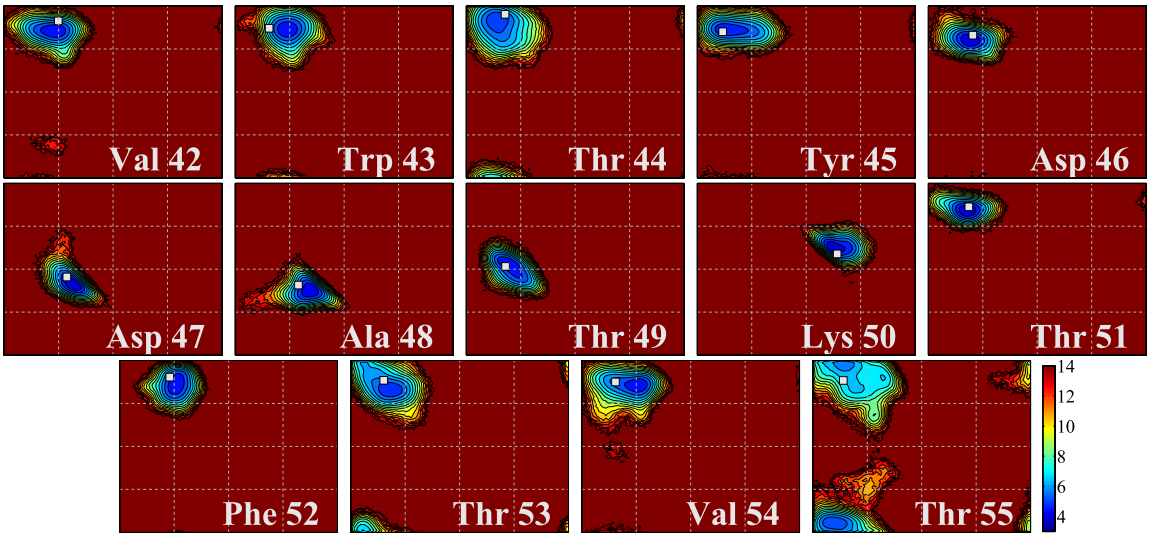
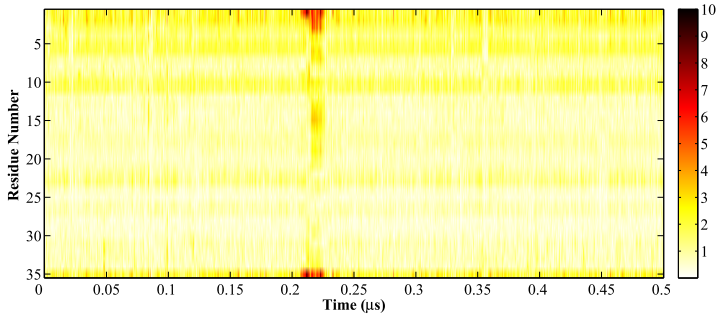


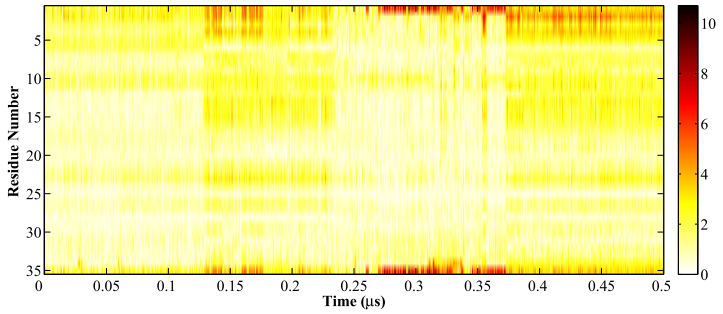
Figure S8 continued.

S7.3 2F4K

a) Run 1



b) Run 2



c) Run 3

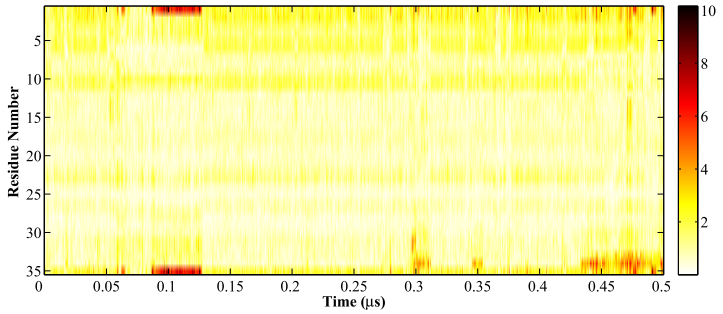


Figure S9: The RMSD per residue ($\mathring{A}$), relative to the crystal structure with PDB code 2F4K, of three simulations of the villin headpiece.

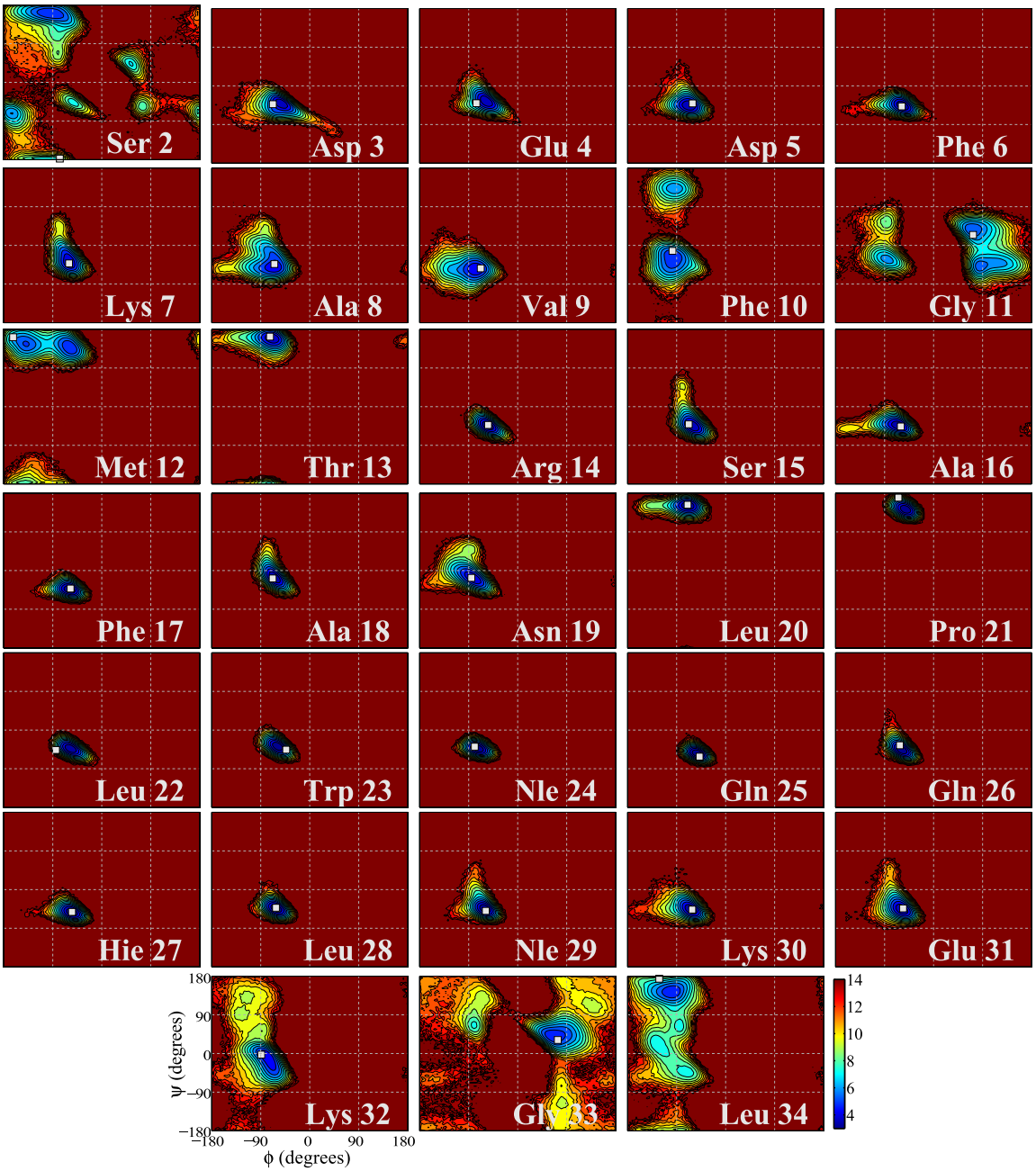
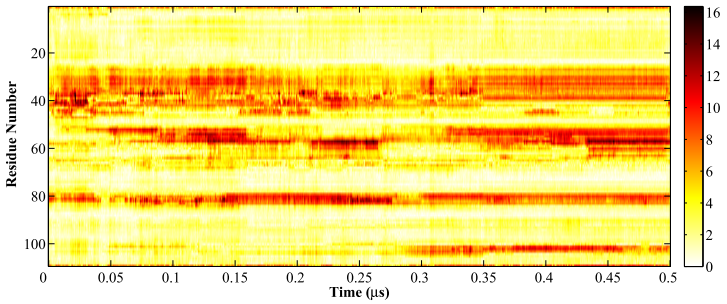


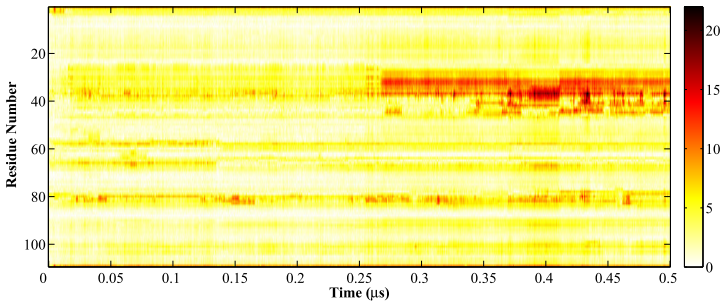
Figure S10: The ψ and ϕ distributions observed in the simulation of the villin headpiece, plotted in the form $-log(p_{\psi,\phi})$. The white square corresponds to the experimental structure given in the PDB file 2F4K. The dark red regions correspond to low probability regions that include conformations that are not sampled during the simulation.

S7.4 1BUJ

a) Run 1



b) Run 2



c) Run 3

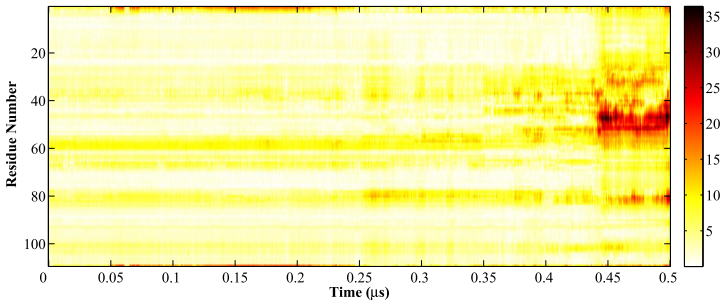


Figure S11: The RMSD per residue ($\mathring{A}$), relative to the crystal structure with PDB code 1BUJ, of three simulations of the binase protein.

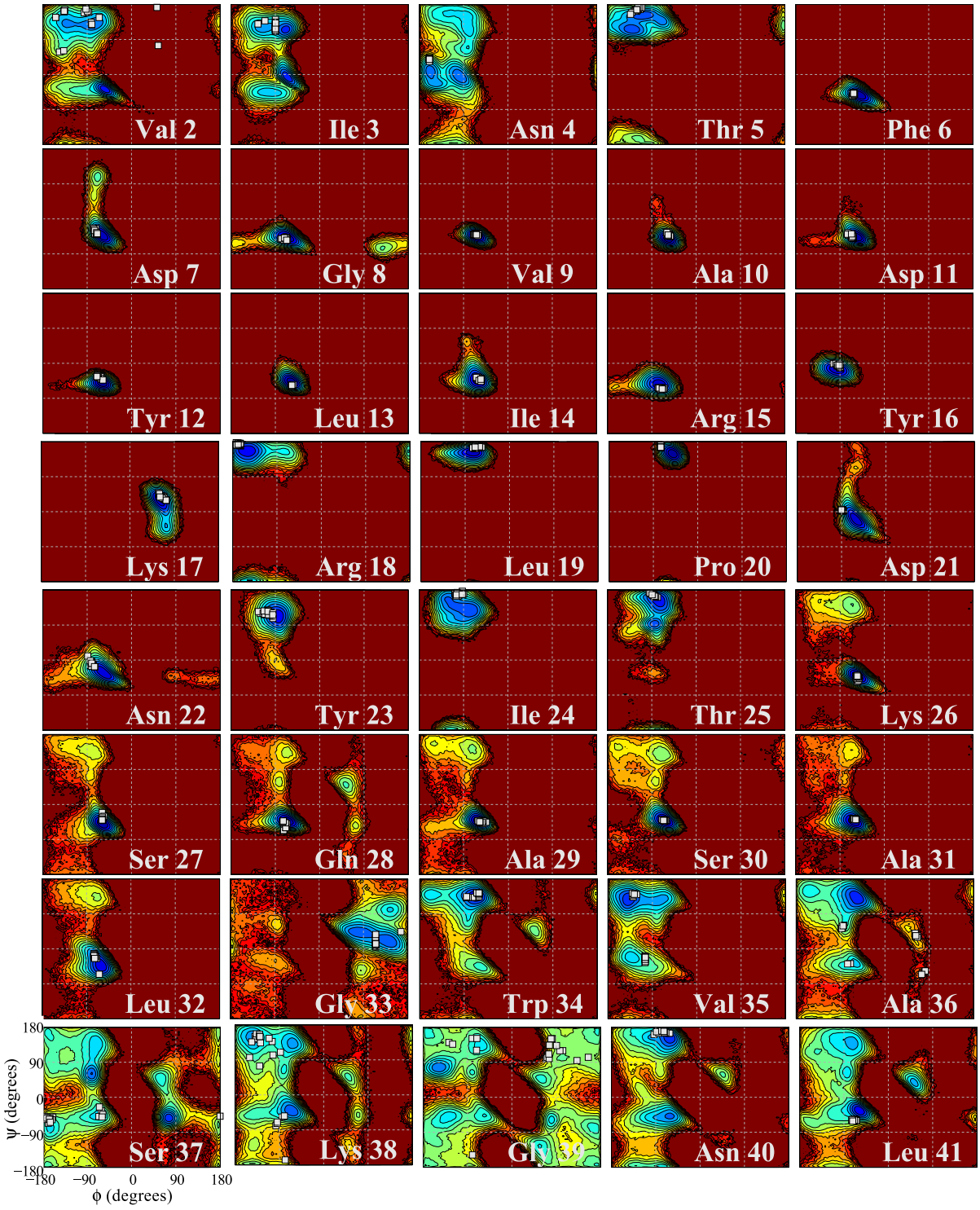


Figure S12: The ψ and ϕ distributions observed in the simulation of the binase protein, plotted in the form $-log(p_{\psi,\phi})$. The white square corresponds to the experimental structure given in the PDB file 1BUJ. The dark red regions correspond to low probability regions that include conformations that are not sampled during the simulation.

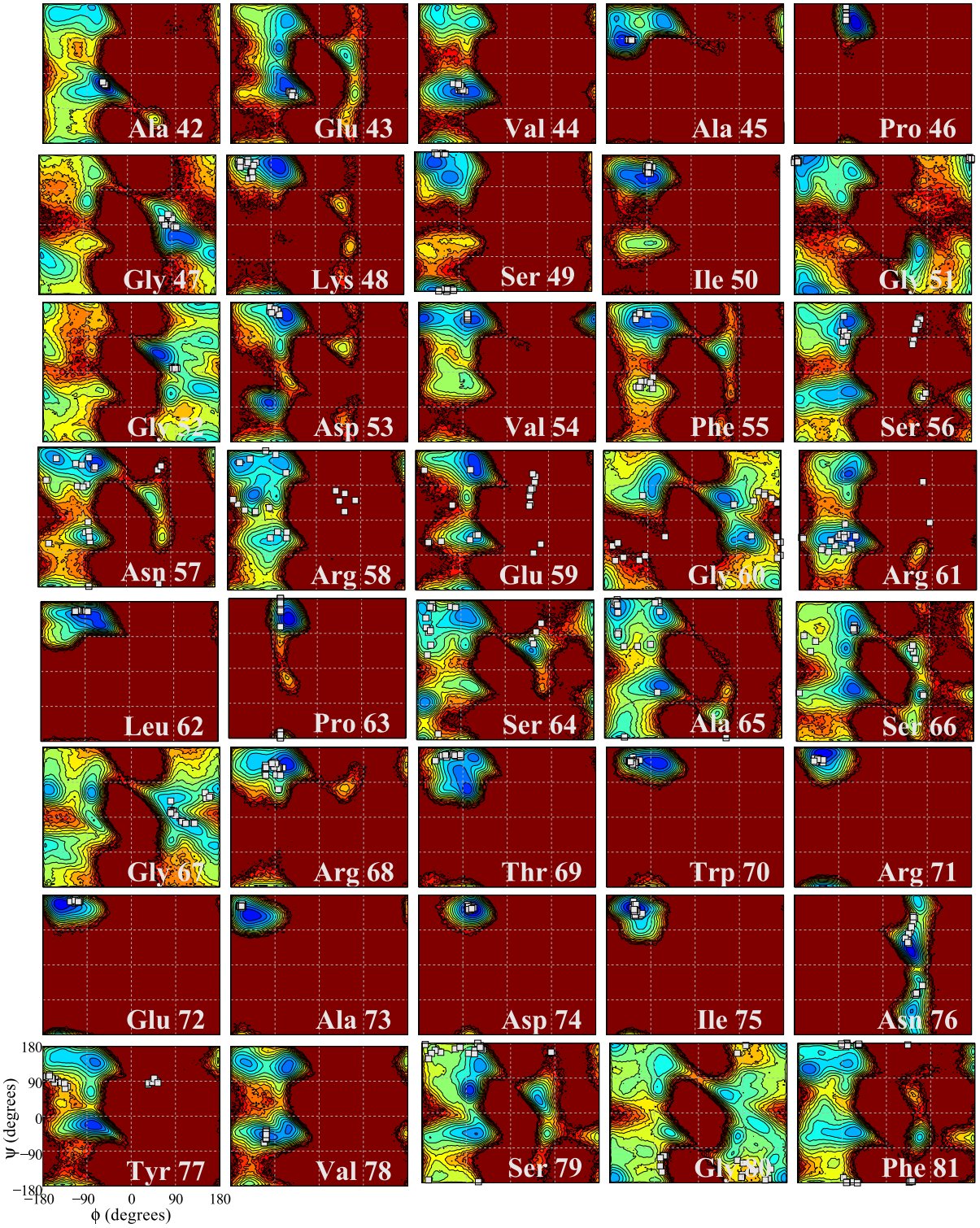


Figure S12 continued.

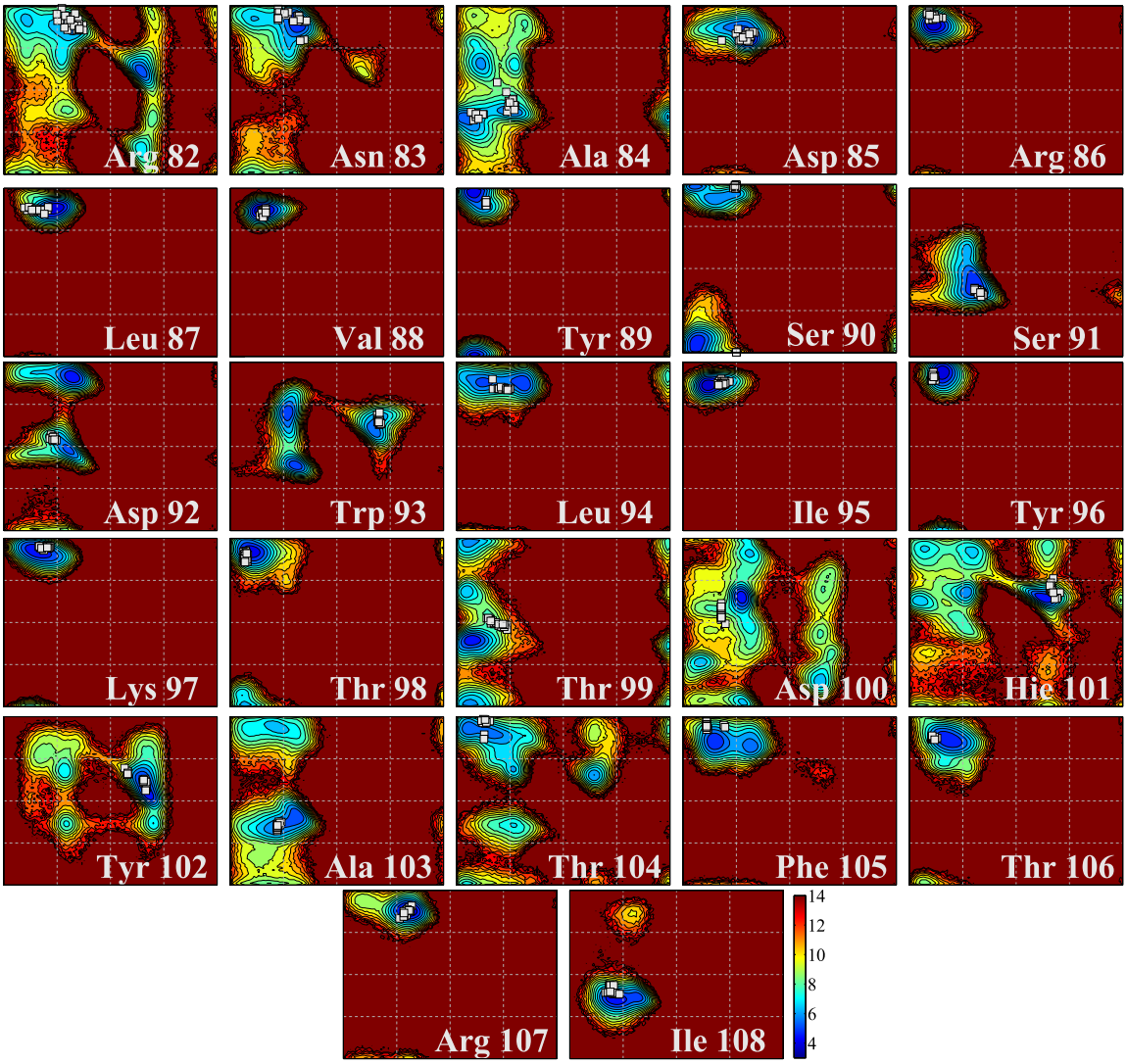
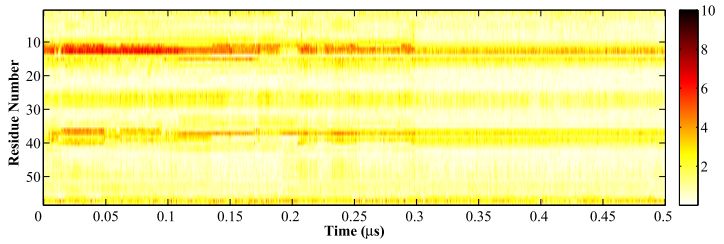


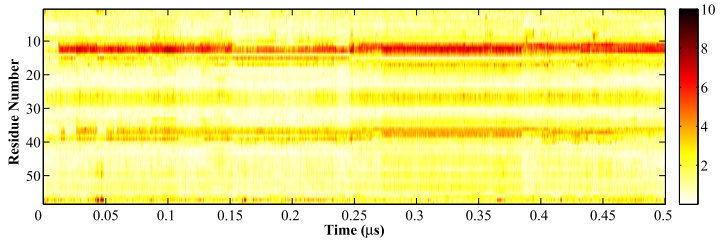
Figure S12 continued.

S7.5 5PTI

a) Run 1



b) Run 2



c) Run 3

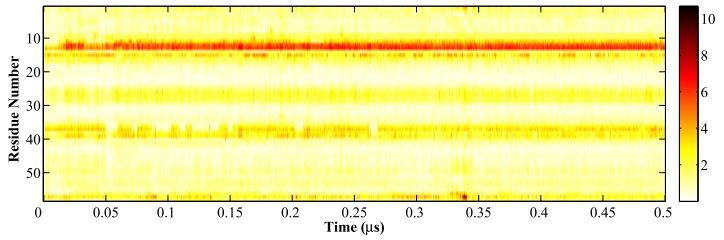


Figure S13: The RMSD per residue ($\mathring{A}$), relative to the crystal structure with PDB code 5PTI, of three simulations of the BPTI protein.

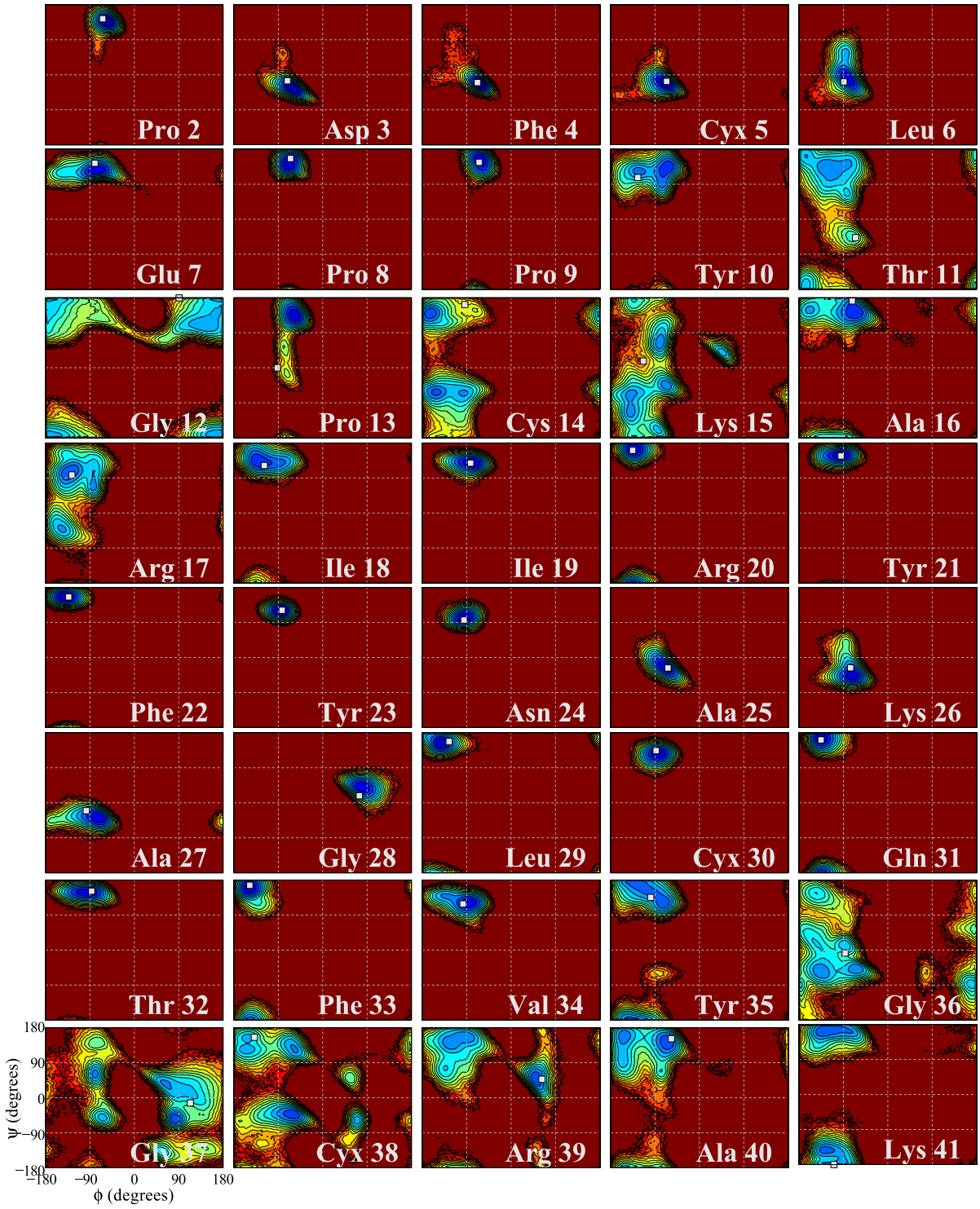


Figure S14: The ψ and ϕ distributions observed in the simulation of the BPTI protein, plotted in the form $-log(p_{\psi,\phi})$. The white square corresponds to the experimental structure given in the PDB file 5PTI. The dark red regions correspond to low probability regions that include conformations that are not sampled during the simulation.

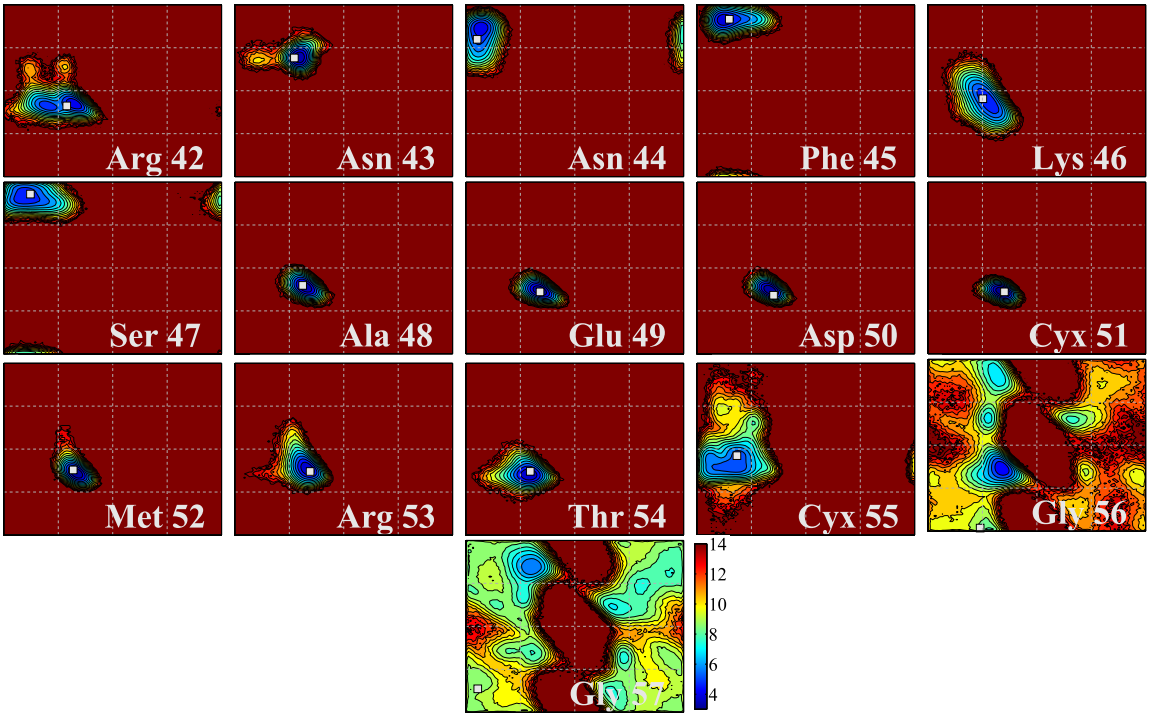


Figure S14 continued.

S8 Ramachandran Plots

The following plots are produced using data given from Ref. S15 and give the probability density of the ϕ and ψ dihedral angles of the residue. This probability density is neighbor dependent and the plots shown are when all neighboring residue types are used. These densities are derived from experimental electron density data. The following plots were used in the preliminary work as weighting functions for torsional parameter fitting.

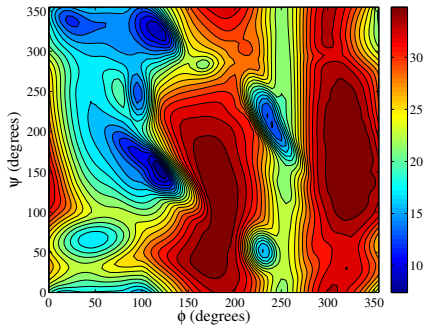


Figure S15: The probability density of ψ and ϕ dihedral angles for alanine, plotted in the form $-log(p_{\psi,\phi})$.

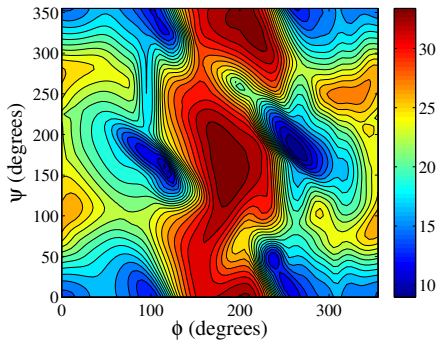


Figure S16: The probability density of ψ and ϕ dihedral angles for glycine, plotted in the form $-log(p_{\psi,\phi})$.

S9 J Coupling Analysis

J coupling values are commonly used to validate the accuracy of force fields.[S3,S16] However, if they are used in isolation as a validation test (without ϕ/ψ distribution analysis) results can be misleading. For example, the J Coupling value for the left-handed α-helical conformation is close to the experimental J Coupling for the dipeptides. Therefore, a low J coupling error can occur if there is a high proportion of left-handed α-helical conformations. However, experimental results show that left handed helical populations should remain low,[S17] and this demonstrates the need to analyze the ϕ/ψ distribution. Additionally, the experimental J coupling value is similar for all the dipeptides in the set. Therefore, similar ϕ/ψ distributions for the set of dipeptides could result in low J coupling errors. However, experimental measurements indicate that this is not the case and different dipeptides occupy varying ϕ/ψ regions.[S18] This points to additional issues with using the J coupling values as a means to analyze a force field's accuracy.

Additionally, there may be problems with using the Karplus parameters with glycine, as previously suggested in Ref. S3. To demonstrate the unsuitability of these Karplus parameters, ${}^2J(N, C_\alpha)$ as a function of the dihedral angle is shown in Fig. S17. The dashed red line is the experimental J coupling, and it can be seen that no dihedral angle can recreate the experimental measurement. This issue has also been previously discussed in Ref. S19.

In addition to the issues with the glycine J coupling values discussed above, the second set of Karplus parameters used in Ref. S14 has also been problematic for a variety of force fields.[S3,S16] Simulations of Ala_5 struggle to recreate the experimental ${}^2J(N, C_\alpha)$ coupling with the second set of Karplus parameters. Figure S18 shows the J coupling for all three sets of Karplus parameters along with the experimental value. From this figure, the conformation populations that would result in a low J coupling error can be estimated. For the first set of Karplus parameters a high PPII conformation would result in a low error. For the second and third set a sizeable β population would be required along with the PPII conformation, however this would negatively impact other J coupling terms. For instance, with

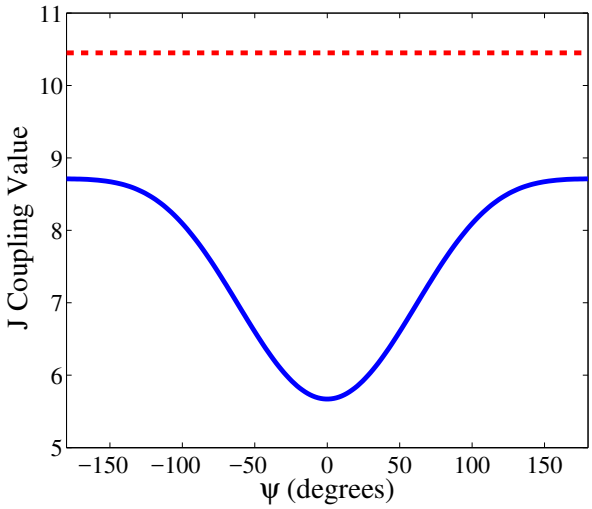


Figure S17: The J coupling value for ${}^2J(N, C_\alpha)$ as a function of ψ is shown in blue. The red dashed line shows the experimental value.

the second set of Karplus parameters ${}^3J(C', C')$ has an error of 160 Hz at $\phi = -135°$ (the β conformation). This points to problems with the Karplus parameters used for ${}^2J(N, C_\alpha)$ as also discussed in Ref. S20. It is useful to consider the issues discussed in this section when analyzing the results of the simulations.

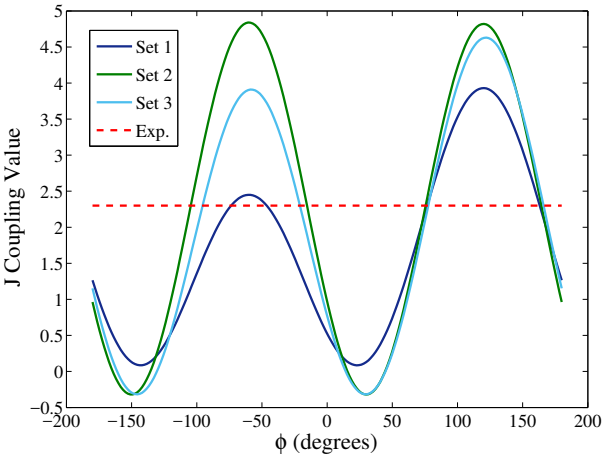


Figure S18: The J coupling value for ${}^2J(N, C_\alpha)$ as a function of ψ is shown in blue. The red dashed line shows the experimental value.

References

(S1) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **1983**, *79*, 926–935.

(S2) Mahoney, M. W.; Jorgensen, W. L. A five-site model for liquid water and the reproduction of the density anomaly by rigid, nonpolarizable potential functions. *J. Chem. Phys.* **2000**, *112*, 8910–8922.

(S3) Robertson, M. J.; Tirado-Rives, J.; Jorgensen, W. L. Improved Peptide and Protein Torsional Energetics with the OPLS-AA Force Field. *J. Chem. Theory Comput.* **2015**, *11*, 3499–3509.

(S4) Lee, L. P.; Cole, D. J.; Skylaris, C.-K.; Jorgensen, W. L.; Payne, M. C. Polarized Protein-Specific Charges from Atoms-in-Molecule Electron Density Partitioning. *J. Chem. Theory Comput.* **2013**, *9*, 2981–2991.

(S5) Perdew, J. P.; Burke, K.; Ernzerhof, M. Generalized gradient approximation made simple. *Phys. Rev. Lett.* **1996**, *77*, 3865–3868.

(S6) Allen, A. E. A.; Payne, M. C.; Cole, D. J. Harmonic Force Constants for Molecular Mechanics Force Fields via Hessian Matrix Projection. *J. Chem. Theory Comput.* **2018**, *14*, 274–281.

(S7) Avbelj, F.; Grdadolnik, S. G.; Grdadolnik, J.; Baldwin, R. L. Intrinsic backbone preferences are fully present in blocked amino acids. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 1272–1277.

(S8) Lindorff-Larsen, K.; Piana, S.; Palmo, K.; Maragakis, P.; Klepeis, J. L.; Dror, R. O.; Shaw, D. E. Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins* **2010**, *78*, 1950–1958.

(S9) Hu, J.-S.; Bax, A. Determination of ϕ and $\chi 1$ Angles in Proteins from 13C-13 Three-Bond J Couplings Measured by Three-Dimensional Heteronuclear NMR. *J. Am. Chem. Soc.* **1997**, *119*, 6360–6368.

(S10) Vögeli, B.; Ying, J.; Grishaev, A.; Bax, A. Limits on Variations in Protein Backbone Dynamics from Precise Measurements of Scalar Couplings. *J. Am. Chem. Soc.* **2007**, *129*, 9377–9385.

(S11) Chou, J. J.; Case, D. A.; Bax, A. Insights into the Mobility of Methyl-Bearing Side Chains in Proteins from 3JCC and 3JCN Couplings. *J. Am. Chem. Soc.* **2003**, *125*, 8959–8966.

(S12) Pérez, C.; Löhr, F.; Rüterjans, H.; Schmidt, J. M. Self-Consistent Karplus Parametrization of 3J Couplings Depending on the Polypeptide Side-Chain Torsion χ_1. *J. Am. Chem. Soc.* **2001**, *123*, 7081–7093.

(S13) Miclet, E.; Boisbouvier, J.; Bax, A. Measurement of eight scalar and dipolar couplings for methine–methylene pairs in proteins and nucleic acids. *J. Biomol. NMR* **2005**, *31*, 201–216.

(S14) Best, R.; Buchete, N.-V.; Hummer, G. Are Current Molecular Dynamics Force Fields too Helical? *Biophys. J.* **2008**, *95*, L07–L09.

(S15) Ting, D.; Guoli, W.; Maxim, S.; Rajib, M.; I., J. M.; L., D. J. R. Neighbor-Dependent Ramachandran Probability Distributions of Amino Acids Developed from a Hierarchical Dirichlet Process Model. *PLOS Comput. Biol.* **2010**, *6*, e1000763.

(S16) Debiec, K. T.; Cerutti, D. S.; Baker, L. R.; Gronenborn, A. M.; Case, D. A.; Chong, L. T. Further along the Road Less Traveled: AMBER ff15ipq, an Original Protein Force Field Built on a Self-Consistent Physical Model. *J. Chem. Theory Comput.* **2016**, *12*, 3926–3947.

(S17) Huang, J.; Rauscher, S.; Nawrocki, G.; Ran, T.; Feig, M.; Bert, L. d. G.; Grubmüller, H.; MacKerell, A. D. CHARMM36m: An Improved Force Field for Folded and Intrinsically Disordered Proteins. *Nat. Methods.* **2016**, *14*, 71–73.

(S18) Grdadolnik, J.; Mohacek-Grosev, V.; Baldwin, R. L.; Avbelj, F. Populations of the three major backbone conformations in 19 amino acid dipeptides. *Proc. Natl. Acad. Sci. U.S.A.* **2011**, *108*, 1794–1798.

(S19) Nerenberg, P. S.; Head-Gordon, T. Optimizing Protein-Solvent Force Fields to Reproduce Intrinsic Conformational Preferences of Model Peptides. *J. Chem. Theory Comput.* **2011**, *7*, 1220–1230.

(S20) Graf, J.; Nguyen, P. H.; Stock, G.; Schwalbe, H. Structure and Dynamics of the Homologous Series of Alanine Peptides: A Joint Molecular Dynamics/NMR Study. *J. Am. Chem. Soc.* **2007**, *129*, 1179–1189.