Supporting Information

Recovering a Representative Conformational Ensemble from Underdetermined Macromolecular Structural Data

Konstantin Berlin, Carlos A. Castañeda, Dina Schneidman-Duhovny, Andrej Sali, Alfredo Nava-Tudela, and David Fushman

A Limitations of the Single Alignment Tensor Model

Consider RDC data for a N-state system:

$$\mathbf{d}_{exp} = w_1\mathbf{d}_1 + \ldots + w_N\mathbf{d}_N = w_1\mathbf{V}_1\mathbf{s}_1 + \ldots + w_N\mathbf{V}_N\mathbf{s}_N = \mathbf{V}\begin{bmatrix} w_1\mathbf{s}_1 \\ \vdots \\ w_N\mathbf{s}_N \end{bmatrix} = \mathbf{V}\mathbf{S}, \tag{S1}$$

where for state j, $\mathbf{V}_j$ is the associated matrix containing products of direction cosines for each relevant bond, $\mathbf{s}_j$ is a column vector comprising the 5 unique alignment tensor elements, and w_j is the associated population weight. Our linear system potentially spans up to $5N$ dimensional space, since there are $5N$-fitting parameters.

We compare this to the single-alignment-tensor model

$$\mathbf{d}_{exp} = w_1\mathbf{V}_1\hat{\mathbf{s}} + \ldots + w_N\mathbf{V}_N\hat{\mathbf{s}} = \mathbf{V}\begin{bmatrix} w_1\hat{\mathbf{s}} \\ \vdots \\ w_N\hat{\mathbf{s}} \end{bmatrix} = \mathbf{V}\hat{\mathbf{S}}. \tag{S2}$$

Even though $\hat{\mathbf{S}}$ has $5N$ values, there are now a large number of constraints on them (e.g. $\hat{S}_1\hat{S}_7 = w_1\hat{s}_1 w_2\hat{s}_2 = \hat{S}_2\hat{S}_6$), leaving only $(N-1)+5$ fitting parameters. Given the possibility that the rank of $\mathbf{V}$ is larger than the number of fitting parameters, there exist numerous possible sets of experimental data explained by the general N-state model that cannot be explained by the single-alignment-tensor model, even when allowing for incorrect values for the population weights.

B Orthogonal Matching Pursuit

The standard OMP algorithm[1,2] that solves the problem

$$\min_{\mathbf{x}} \|\mathbf{x}\|_0, \text{ s.t. } \|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2 \leq \epsilon, \tag{S3}$$

is given in Alg. S1.

Algorithm S1 Orthogonal Matching Pursuit (OMP)

Input: An $L \times N$ matrix $\mathbf{A}$, observed data vector $\mathbf{y}$, and error threshold ϵ. Let $\mathbf{a}_j$ be the jth column of $\mathbf{A}$.

Output: $\mathbf{x}^*$, the solution to Eq. (S3).

1: Set the initial solution, $\mathbf{x}^0 = 0$
2: Set the initial residual, $\mathbf{r}^0 = \mathbf{y}$
3: Set the initial list of non-zero columns to empty, $S^0 = \emptyset$
4: Set the loop counter, $m = 0$
5: **for** $\|\mathbf{r}^m\|_2 > \epsilon$ **do**
6: Compute the residual estimates for each column j, $\xi_j = \min_{z_j} \|\mathbf{r}^m - \mathbf{a}_j z_j\|_2$, by using the optimal choice $z_j = \mathbf{a}_j^T \mathbf{r}^m / \|\mathbf{a}_j\|_2^2$
7: Select the column j with the minimum value ξ_j, and add it to the list of non-zero columns, $S^{m+1} = S^m \cup \{j\}$
8: $\mathbf{x}^{m+1} = \arg\min_{\mathbf{x}} \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2$, s.t. $x_i = 0, \forall i \notin S^{m+1}$
9: $\mathbf{r}^{m+1} = \mathbf{y} - \mathbf{A}\mathbf{x}^{m+1}$
10: $m = m + 1$
11: **end for**
12: **return** $\mathbf{x}^m$

Our modified Multi-OMP algorithm solves

$$\min_{\mathbf{x}} \chi^2(\mathbf{x}), \text{ s.t. } \|\mathbf{x}\|_0 = M, \mathbf{x} \geq 0, \tag{S4}$$

where $\chi^2(\mathbf{x}) = \|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2^2$, and is given in Alg. S2.

The Multi-OMP algorithm is a modified OMP algorithm that returns K best nonnegative solutions, instead of just the best solution in the standard OMP algorithm. This is done by selecting the top K solutions at each iteration of Multi-OMP, instead of selecting only the best solution (see Alg. S2, Line 6), and then propagating them to the next m-loop iteration as multiple starting values. These top K solutions should be analyzed to determine if an alternative solution exists with a close enough χ^2 relative to the best solution. By propagating K solutions during each iteration we also improve the χ^2 of our $\mathbf{x}^*$ solution relative to the original OMP algorithm.

One of the more computationally intensive tasks in our Multi-OMP algorithm is recomputing a least-squares solution $\mathbf{x}^{m+1}$ after selecting a new column during the mth iteration (see Alg. S2, Line 18). This task can be formulated as a least-squares solution under a rank-one update of $\mathbf{A}$ from the previous iteration. $\mathbf{x}^{m+1}$ can be updated by a single iteration of the Gram-Schmidt orthogonalization algorithm[3]. If $\mathbf{x}^{m+1}$ contains negative weights, it is removed from the list of top K solutions.

In order to speed up the computation for large values of K, the current best K values are maintained using a priority queue[4], and the algorithm is parallelized. Computing K top sparse solutions for $M = 1, \ldots, \text{rank}(A)$ is relatively quick, since the Multi-OMP algorithm has an $O(KMLN)$ complexity and M typically is small. The selected value of K should be the largest possible value such that all the computations fit in the computer memory and the algorithm terminates in the desired amount of time.

Algorithm S2 Multi Orthogonal Matching Pursuit (M-OMP)

Input: An $L \times N$ matrix $\mathbf{A}$, observed data vector $\mathbf{y}$, size of desired sparsity M, and the number of top solutions to retain between iterations K.

Output: $\mathbf{x}^*$, the solution to Eq. (S4).

Set the initial solution, $\mathbf{x}^0 = 0$
Set the initial residual, $\mathbf{r}^0 = \mathbf{y}$
Set the initial list of non-zero columns to empty, $S^0 = \emptyset$
Set the initial list of lists, $T^0 = \{\{S^0, \mathbf{r}^0, \mathbf{x}^0\}\}$
Set the loop counter, $m = 0$
for $m < M$ **do**
  Initialize the set of top K guesses for the mth iteration, $\Xi = \emptyset$
  for all $\{S^m, \mathbf{r}^m, \mathbf{x}^m\} \in T^m$ **do**
    for $j = 1$ to N **do**
      Compute $\xi_j = \min_{z_j} \|\mathbf{r}^m - \mathbf{a}_j z_j\|_2$
      add element $\{\{S^m, \mathbf{r}^m\}, j, \xi_j\}$ to Ξ
      if size of Ξ greater than K, remove the element with the largest ξ value
    end for
  end for
  Initialize the storage of least-squares solution for all guesses in Ξ, $T^{m+1} = \emptyset$
  for all $\{\{S^m, \mathbf{r}^m, \mathbf{x}^m\}, j, \xi_j\} \in \Xi$ **do**
    $S^{m+1} = S^m \cup \{j\}$
    $\mathbf{x}^{m+1} = \arg\min_{\mathbf{x}} \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2$, s.t. $x_i = 0, \forall i \notin S^{m+1}$
    if $x_i^{m+1} > 0$ for all $i \in S^{m+1}$ **then**
      $\mathbf{r}^{m+1} = \mathbf{y} - \mathbf{A}\mathbf{x}^{m+1}$
      $T^{m+1} = T^{m+1} \cup \{\{S^{m+1}, \mathbf{r}^{m+1}, \mathbf{x}^{m+1}\}\}$
    end if
  end for
  $m = m + 1$
end for
return $\mathbf{x}^m$ with the smallest $\|\mathbf{r}^m\|_2$ in T^m

B.1 Advantages of Multi-OMP over Previous Minimum-Ensemble Algorithms

There are several reasons why Multi-OMP is a more reliable method for computing sparse solutions than previously suggested stochastic genetic programming and simulated-annealing algorithms. First, stochastic algorithms require careful selection of parameters such that the computations do not terminate too early or exceed the specified running time. These parameters would change depending on the input size and energy landscape of the function being minimized, however, it is difficult to know what they should be *a priori*.

In contrast, Multi-OMP has no adjustable parameters and its computational time linearly scales with the input size and its only parameter K. Second, OMP has a computationally efficient heuristic for evaluating the fitness of a conformer j to the ensemble by determining how correlated the data for that conformer, $\mathbf{a}_j$, are to the remaining residuals, $\mathbf{r}$, as measured by the dot-product $\mathbf{a}_j \cdot \mathbf{r}$. This fitness can be quickly evaluated for all conformers. By contrast, in previously suggested genetic programming algorithms (ASTEROIDS, MES) it is assumed that the fitness of an updated solution

is improved (in a probabilistic sense) by randomly mixing two previous-iteration solutions. This tends to be a less accurate heuristic, since the two previous solutions could be largely explaining the same observations, in which case mixing them would yield only marginal improvement. The simulated-annealing algorithms presented previously (e.g.,[5]) do not rely on a heuristic at all (though they can be modified to use a thresholding heuristic[6]), therefore each new ensembles fitness must be directly evaluated, which is computationally expensive.

For all methods, once the heuristic is applied and a potential candidate ensemble is selected, the associated weights need to be computed. In our Multi-OMP algorithm we take advantage of the fact that we add one conformation at a time to our previously computed weighted ensemble. This is equivalent to a rank-one update of a linear system, and therefore the weights can be efficiently updated using an $O(LM)$ Gram-Schmidt algorithm, instead of being directly recomputed for a newly generated M-sized ensemble, which has a computation $O(LM^2)$ complexity.[3]

Together with parallelization, our heuristic and rank-one update of the solution allow us to efficiently filter a larger number of unlikely ensembles, while also directly evaluating orders of magnitude more potential ensembles than in the previously suggested genetic-programming or simulated-annealing algorithms, with the added advantage that, while computing the best M-sized ensemble, we simultaneously also compute all best smaller-sized ensembles at no additional cost. Computing all smaller-sized ensembles, rather than a subsampling is critical for the ℓ-curve analysis, since skipping even a single ensemble size could result in incorrect regularization. Therefore, when combined with ℓ-curve regularization, our algorithm provides substantial improvement in computation time, as compared to previous approaches.

C Preconditioning of χ^2 to Improve Computational Complexity

As we have demonstrated in the manuscript, a large part of the experimental data contains redundant structural restraints, because the effective rank of $\mathbf{A}$ is significantly below the actual number of experimental observables. Since the computational complexity of the algorithm and the memory requirements are proportional to the number of rows in the $\mathbf{A}$ matrix, we compress our linear system such that all the redundancy is removed. To achieve this we rotate our linear system by multiplying by an orthonormal matrix $\mathbf{U}^T$, where $\mathbf{U}^T$ is the transpose of the $\mathbf{U}$ matrix defined as $\mathbf{A} = \mathbf{U}\mathbf{L}\mathbf{V}^T$ in the SVD decomposition of $\mathbf{A}$, such that

$$\chi^2(\mathbf{x}) = \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 = \|\mathbf{y} - \mathbf{U}\mathbf{L}\mathbf{V}^T\mathbf{x}\|_2^2 = \|\mathbf{U}^T\left(\mathbf{y} - \mathbf{U}\mathbf{L}\mathbf{V}^T\mathbf{x}\right)\|_2^2 = \|\mathbf{U}^T\mathbf{y} - \mathbf{L}\mathbf{V}^T\mathbf{x}\|_2^2. \quad \text{(S5)}$$

We then just remove the rows from $\mathbf{U}^T\mathbf{y}$ and $\mathbf{L}\mathbf{V}^T$ in Eq. S5 associated with the relative singular values below a certain threshold (10^{-4}). This computation reduces the number of rows in our linear system, L, from around 90 to 10 for RDC and from 200 to just 21 for SAXS. Depending on size of $\mathbf{A}$, the computation of the SVD might be intractable, in which case alternative matrix sketching algorithms could potentially be used.[7] Our proposed compression method can be extended to the more general preconditioning idea.[8,9] Note that preconditioning (or compression) can be combined with any sparse recovery algorithm, not just Multi-OMP. We will explore the idea of preconditioning of χ^2 in order to improve solution recovery in future work.

D Stability of SES Solution Under Noise

One of the primary advantages of describing the ensemble selection problems in terms of Eq. 6 is that that stability of weights (in individual ensembles) under noise in data is well understood in terms of a matrix condition number, easily computed as $\text{cond}(\mathbf{A}) = \|\mathbf{A}\|_2\|\mathbf{A}^*\|_2$, where $\mathbf{A}^*$ is a Moore-Penrose pseudoinverse of $\mathbf{A}$.

For a given M-sized ensemble, we refer to the associated $L \times M$ matrix of the M-state ensemble as $\tilde{\mathbf{A}}$, where $\tilde{\mathbf{A}}$ has only M $\mathbf{a}_j$ columns, for all j where $x_j > 0$. The non–zero weights of $\mathbf{x}$ for such an ensemble, are computed from the experimental values $\mathbf{y}$ in Line 18 of Alg. S2, where

$$\mathbf{x} = \arg\min_{\mathbf{x}} \|\mathbf{y} - \tilde{\mathbf{A}}\mathbf{x}\|_2 = \tilde{\mathbf{A}}^*\mathbf{y}. \tag{S6}$$

When noise ϵ is added to the linear system, we get a different solution,

$$\mathbf{x}' = \arg\min_{\mathbf{x}} \|(\mathbf{y} + \epsilon) - \tilde{\mathbf{A}}\mathbf{x}\|_2 = \tilde{\mathbf{A}}^*(\mathbf{y} + \epsilon). \tag{S7}$$

The difference in weights of the two solutions, $\|\mathbf{x}' - \mathbf{x}\|_2$, will be small given a low condition number of the matrix $\tilde{\mathbf{A}}$, since

$$\frac{\|\mathbf{x} - \mathbf{x}'\|_2}{\|\mathbf{x}\|_2} \leq \frac{\|\tilde{\mathbf{A}}^*(\mathbf{y} + \epsilon) - \tilde{\mathbf{A}}^*\mathbf{y}\|_2}{\|\mathbf{x}\|_2} = \frac{\|\tilde{\mathbf{A}}^*\epsilon\|_2}{\|\mathbf{x}\|_2} \leq \frac{\|\tilde{\mathbf{A}}\|_2\|\tilde{\mathbf{A}}^*\|_2\|\epsilon\|_2}{\|\tilde{\mathbf{A}}\|_2\|\mathbf{x}\|_2} \leq \text{cond}\,(\tilde{\mathbf{A}})\frac{\|\epsilon\|_2}{\|\mathbf{y}\|_2}. \tag{S8}$$

Meaning that the relative error in the weights is bounded by $\text{cond}\,(\tilde{\mathbf{A}})$ multiplied by the relative error in the experimental data. $\text{cond}\,(\tilde{\mathbf{A}})$ for our 3-state SES at pH 4.5, 6.8, and 7.6 are small, 1.4, 2.0, and 1.9 respectively, and hence our SES solution is robust to noise. Performing n-fold cross-validation on the 3-state ensemble is not useful, since our 3-state linear model is already overdetermined, with approximately 90 experimental observables, but only 3 fitting parameters, and low matrix condition numbers.

E Ensemble Generation

We generate our initial ensemble by adapting the Rapidly-exploring Random Trees (RRT) algorithm[10] for ensemble generation. The RRT algorithm is a robotics approach for motion planning that was previously used for sampling large-scale motion in proteins.[11–13] Our algorithm samples the conformational space by leveraging an iteratively constructed nearest neighbor linked tree. At each iteration, we generate a random conformation, q_{rand}, by uniform random sampling of the degrees of freedom spanning the desired conformational space, followed by finding its closest neighbor, q_{near}, in the tree. If it is feasible to move without a steric clash from q_{near} to q_{rand}, through the linear path between the two conformations, we add the new conformation by connecting it to q_{near} by a new edge. Otherwise, we attempt to expand the tree by exploring the linear path in small step sizes until a clashing conformation is reached. The last clash-free conformation, if it exists, is then connected to q_{near}. This iterative strategy expands the tree towards unexplored regions, since the probability that a node will be the nearest neighbor of a randomly generated conformation is proportional to its Voronoi region in the conformational space. Thus, using the RRT sampling strategy, we significantly improve the sampling of the good scoring regions of the conformational space compared to random sampling.

The conformational space sampling of the 20000 conformer initial ensemble, generated using the RRT algorithm from the initial PDB structure 1AAR and 3NS8, is shown in Fig. S1.

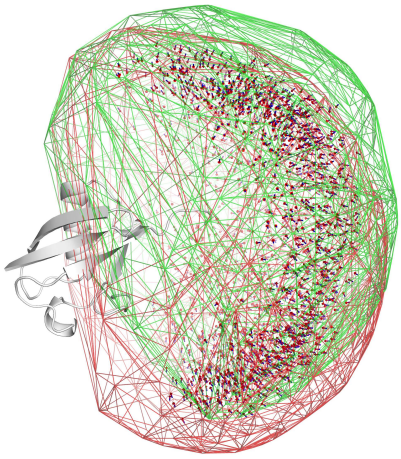

Figure S1: Visual representation of the 20000-conformer initial ensemble of K48-Ub_2. All conformers were superimposed by their distal Ub. The distal Ub is shown as a cartoon on the left, while shown on the right are the convex hulls of the C_α atoms of all 10000 conformers generated from PDB ID 1AAR (green) and all 10000 conformers generated from PDB ID 3NS8 (red). The red/blue vectors represent 2000 randomly selected proximal Ubs: each vector's position is the center of mass of the proximal Ub, with the vector oriented parallel to the amide NH bond of Lys29, such that it represents the orientation of the α-helix in the respective Ub unit.

F Applicability of Different Restrains for Ensemble Recovery

In testing the suitability of a structural restraint for ensemble recovery, it is helpful to assess the correlation of predicted data for different conformations in the initial ensemble. In Fig. S2, predicted data and their correlation are visualized for three different structural restraints (RDCs, SAXS, and PREs) using 100 randomly-selected conformations from the 20000-member K48-Ub_2 ensemble. In other words, 100 column vectors of the $\mathbf{A}$ matrix are plotted for each restraint. The paramagnetic relaxation enhancement (PRE) data (ΔR_2, increase in ^{1}H transverse relaxation rate) were predicted based on the spin label's position provided in[14]. In Fig. S2 (bottom panels), the correlation matrix was calculated for predicted data from pairwise combinations of 100 randomly-selected conformations.

It is evident from the plots that RDCs best discriminate among different structures (large spread in RDC values) and that the RDCs are the least correlated (correlation plot has the most blue). In contrast, SAXS profiles are all very similar and therefore, highly correlated to each other. PREs are badly conditioned, in that the predicted ΔR_2 data values span several orders of magnitude, with a number of residues having ΔR_2 of 0. For PREs, multiple datasets using different paramagnetic spin label positions will be necessary to improve the ability of PREs to discriminate among different structures.

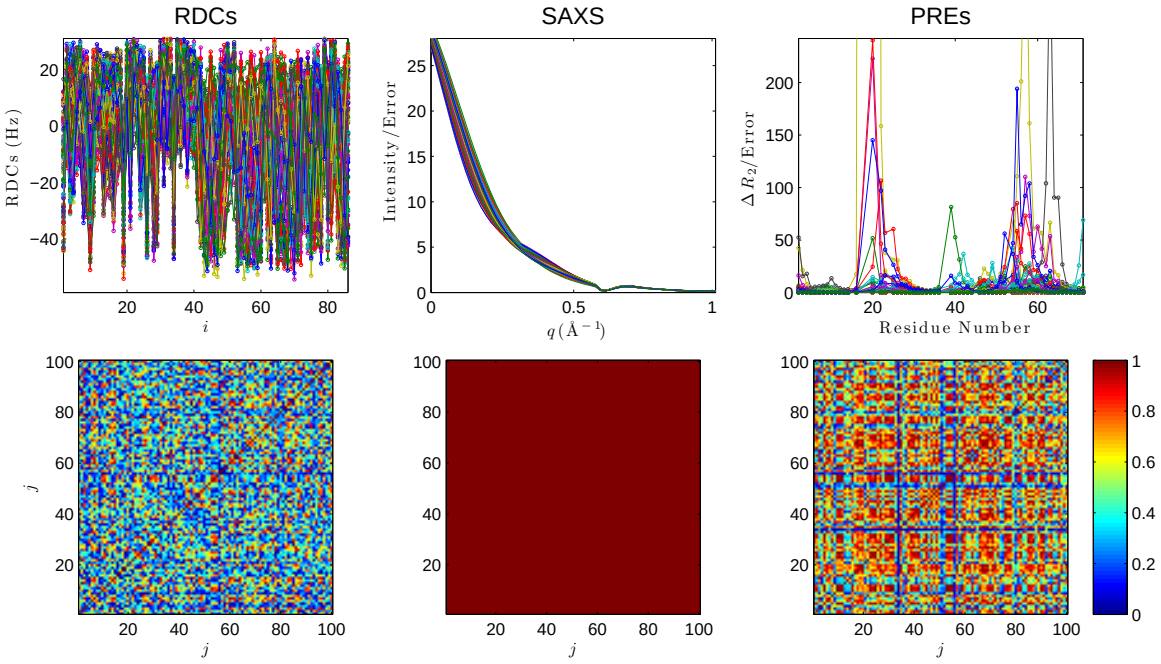

Figure S2: (Top Panels) For 100 randomly-selected conformations, $j = 1 \ldots 100$, predicted data (RDCs, SAXS profile, PREs) are plotted, and each conformations data colored differently. RDCs were predicted as described in the main text, for residues (x-axis) in elements of secondary structure of both distal and proximal Ubs. SAXS profiles were calculated for q values between 0 Å^{-1} and 1.0 Å^{-1}, and also detailed in the main text. PREs (ΔR_2) were calculated for residues in the proximal Ub, assuming spin label's placement on residue 48 in the distal Ub (as detailed in [14]), and using the overall rotational correlation time (τ_c) of 9.0 ns. (Bottom Panels) Matrix of the absolute uncentered correlation coefficient values, calculated from the predicted data for all pairwise combinations of the 100 randomly-chosen conformations (the same as in the top panels). The lowest correlation in the SAXS correlation matrix (for the 100 SAXS profiles shown in the above panel) is 0.993, hence that plot has uniform dark red color.

G Alternative Solutions

G.1 Clustering Methodology

The methodology for representing alternative solutions is given in Fig S3.

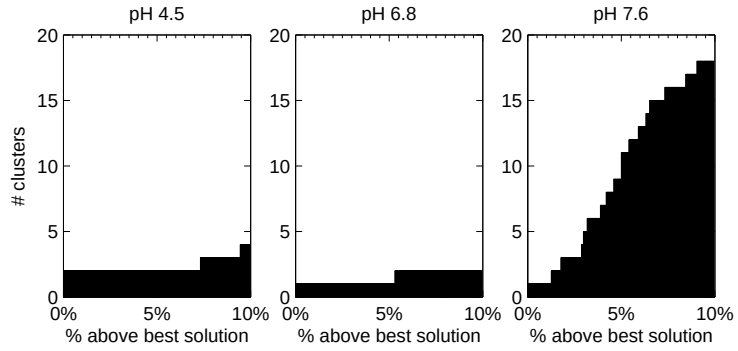

Figure S3: The number of alternative clusters of ensembles for $M = 3$ solution, as a function of their $\|\mathbf{r}\|_2$ values as a percent above the $\|\mathbf{r}\|_2$ value for the best $M = 3$ solution. Alternative ensemble solutions were clustered together in order to remove duplicate or almost identical solutions from the analysis. All the structures that are members of any alternative ensemble were hierarchically clustered by 8 Å RMSD. Any ensembles whose 3 states share the same 3 clusters were merged together, and this new cluster is represented by its best ensemble. For easier visualization, a 3% cutoff is used in the manuscript.

G.2 Top Clusters

The alternative top 3% solutions for pH 4.5, 6.8, and 7.6 are shown in Figs. S4 and S5.

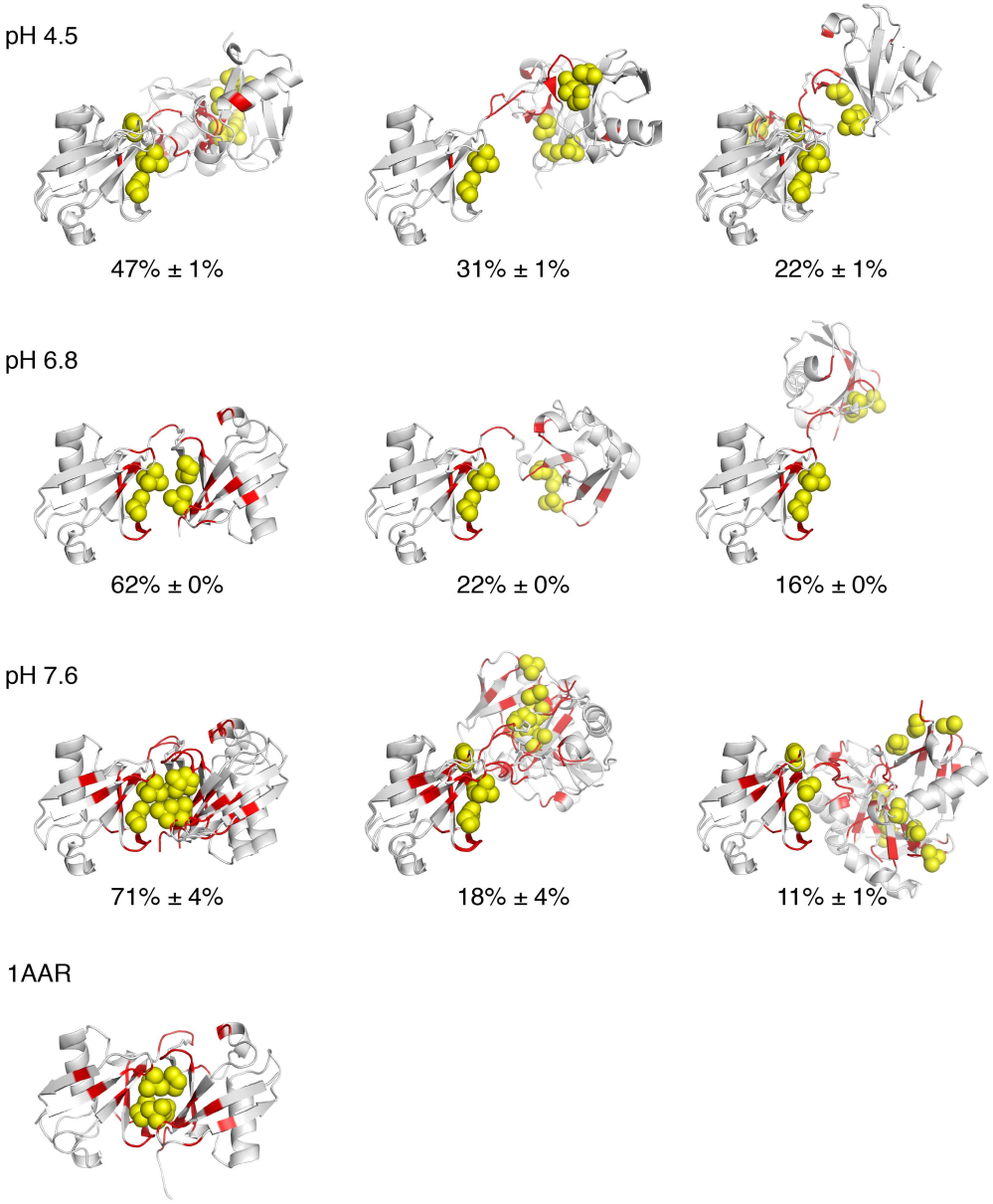

Figure S4: The top 3% $M = 3$ cluster centroid solutions, overlaid on top of each other, for each pH value analyzed here. Red coloring of the ribbon marks residues that exhibited significant spectral differences (CSPs ≥ 0.05 ppm) between the Ub_2 and the corresponding Ub monomers; the spheres (yellow) represent the side chains of the hydrophobic patch residues Leu8, Ile44, and Val70 in both Ub units. For comparison, crystal structure of the closed state of Lys48-linked Ub_2 (PDB ID 1AAR) is also shown, colored according to CSPs at pH 7.6.

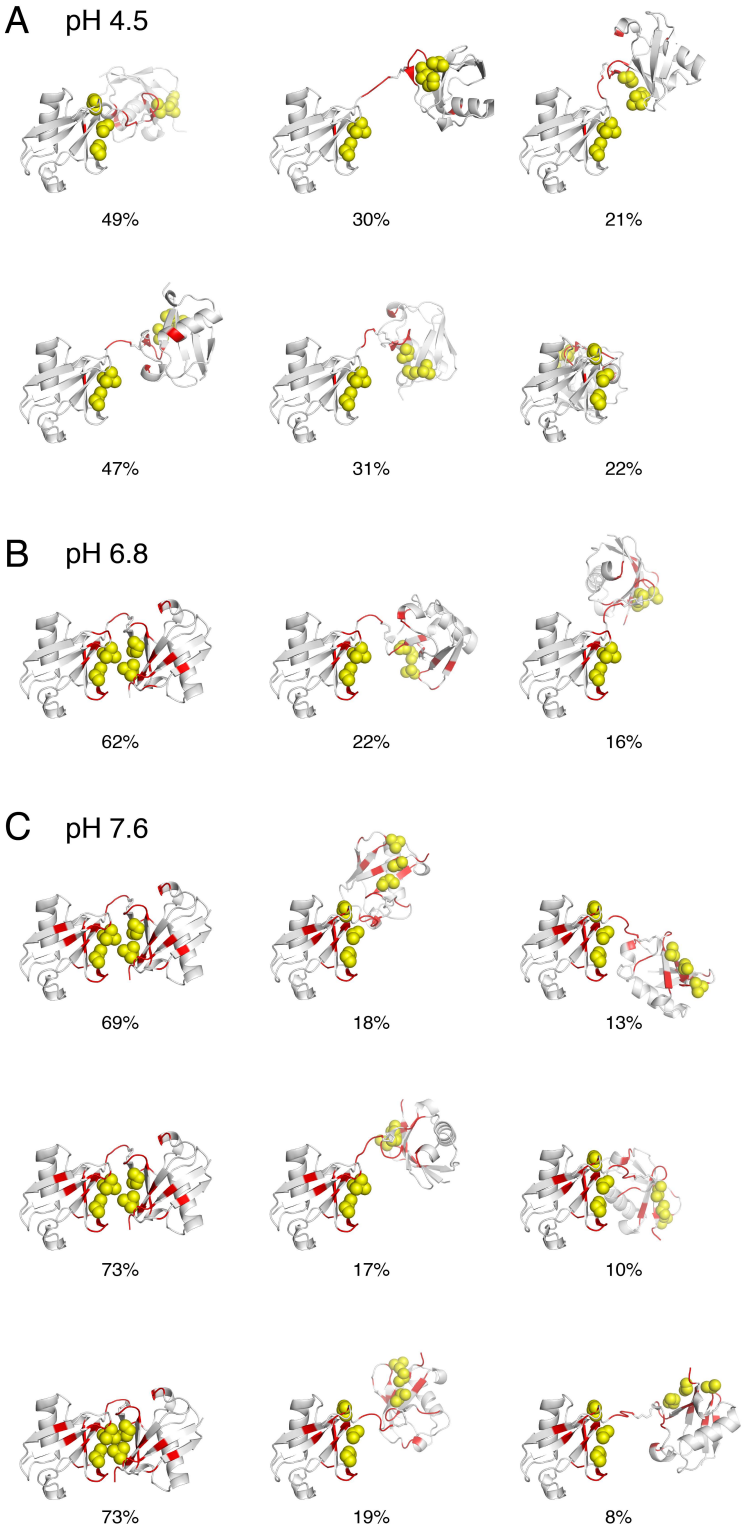

Figure S5: The top 3% $M = 3$ cluster centroid solutions for all pHs. Red coloring of the ribbon marks residues that exhibited significant spectral differences (CSPs ≥ 0.05 ppm) between the Ub_2 and the corresponding Ub monomers; the spheres (yellow) represent the side chains of the hydrophobic patch residues Leu8, Ile44, and Val70 in both Ub units.

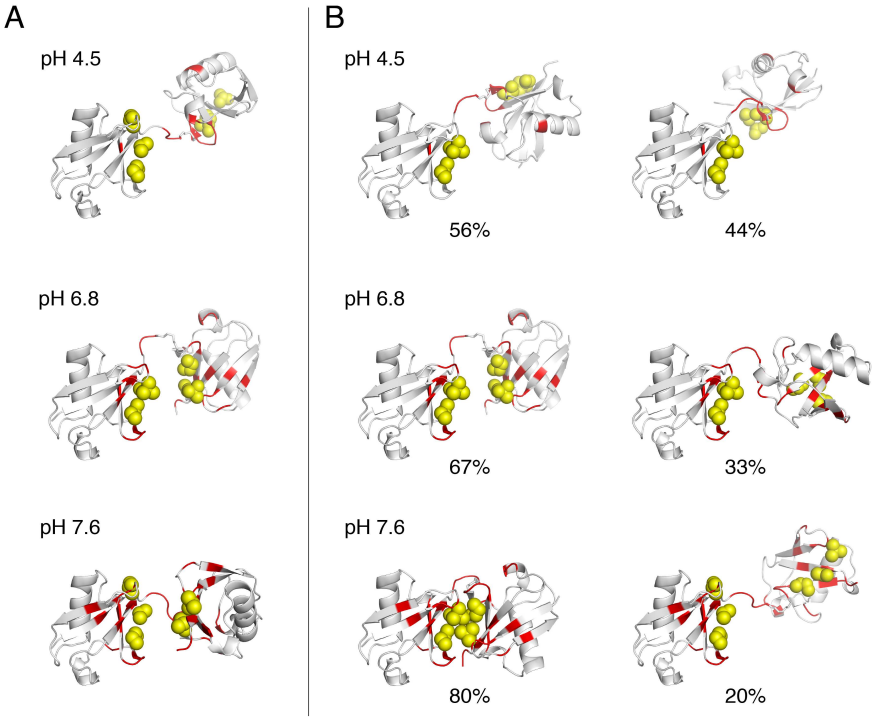

Figure S6: (A) The top $M = 1$ SES ensembles for all pHs. (B) The top $M = 2$ SES ensembles for all pHs. Red coloring of the ribbon marks residues that exhibited significant spectral differences (CSPs ≥ 0.05 ppm) between the Ub_2 and the corresponding Ub monomers; the spheres (yellow) represent the side chains of the hydrophobic patch residues Leu8, Ile44, and Val70 in both Ub units.

H Maximum Entropy Computation and Results

In the Results section we compare our SES solution with the maximum-entropy approach. Maximum entropy computation is a convex optimization problem, and so for all values of λ and all pH conditions we computed the solution in MATLAB using an interior-point method, while avoiding explicit computation of the Hessian by using the L-BFGS approximation.[15]

The appropriate λ regularization value was selected by using the described ℓ-curve approach. The ℓ-curve, $\log \chi^2(\mathbf{w})$ vs. $\log \sum_j w_j \log(N w_j)$, was computed for a discrete set of λ values, and then interpolated using a cubic smoothing spline ($p = 0.995$) in MATLAB. The spline was then twice differentiated, and the point with the maximum value was used to estimate the point of maximum curvature. The solution associated with that point was selected as the regularized final solution. The comparison between experimental RDC data and the back-calculated values for various combinations of substates is shown in Fig. S7. Note that for our MaxEnt solution $\chi^2(\mathbf{w}^*) \approx \varepsilon_r$, meaning that we have properly optimized Eq. 8. The fit of the experimental RDC data to the back-calculated values of the MaxEnt solution is given in Fig. S7A.

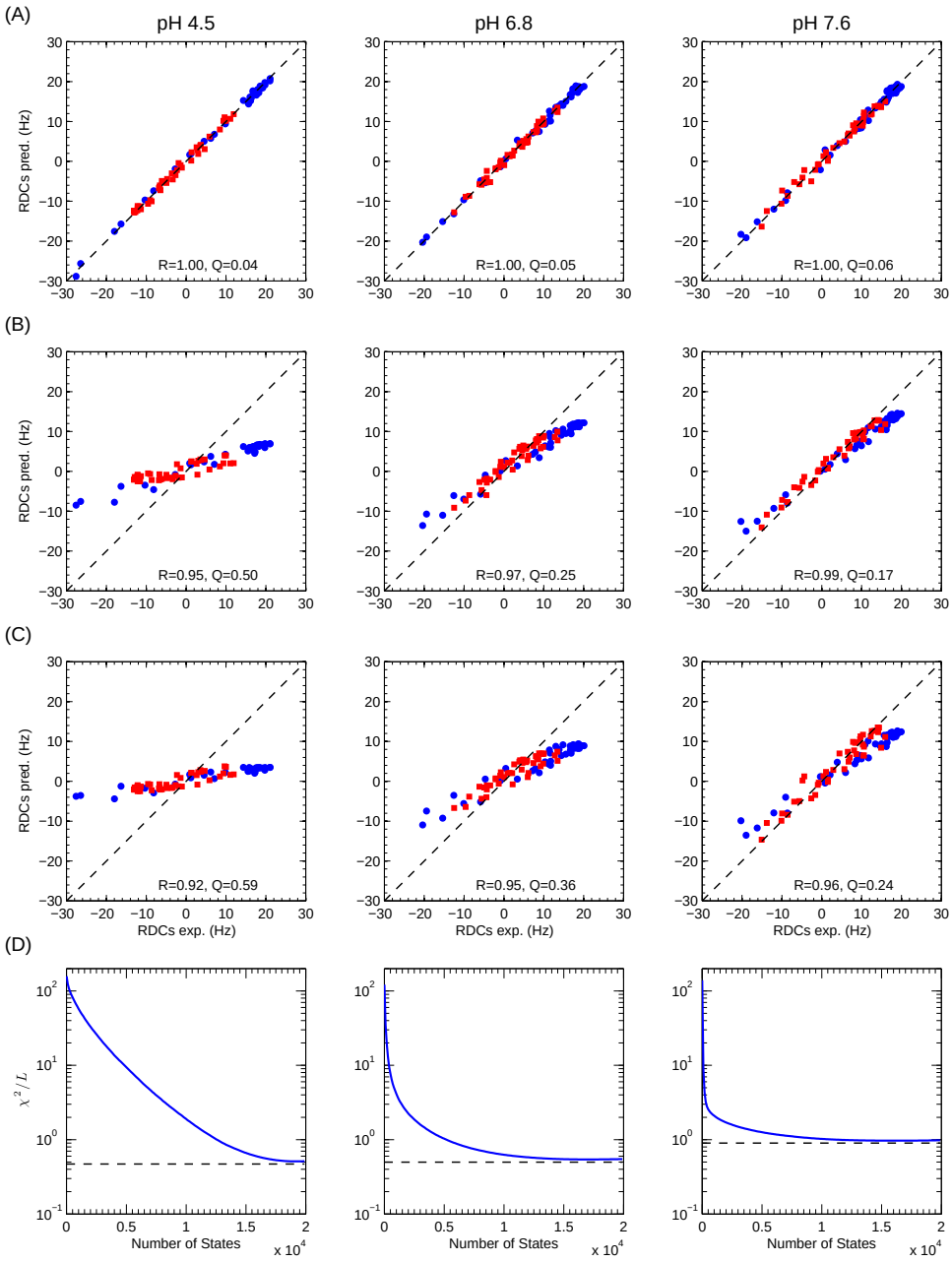

Figure S7: (A-C) Comparison between the experimental RDC data and the back-calculated values using MaxEnt solutions, at various pH conditions for the distal (blue circles) and proximal (red squares) Ubs in Lys48-linked Ub_2. Several subsets of the 20000 computed weights used to compute $\mathbf{d}_{pred}$: (A) All the weights from the 20000 states; (B) Only the weights of the significant states (states with weights of more than two standard deviations above the average weight), representing 31%, 62%, and 75% of the total weight, and corresponding to 559, 239, and 103 states, for pH 4.5, 6.8, and 7.6, respectively; (C) Only the weights of the states in the 4 most populated clusters of the significant states shown in Fig. 8 (main text). The significant states were clustered by hierarchical clustering with 4 Å C_α-RMSD cutoff. The 4 most populated clusters represent 15%, 47%, 66% the total weight, corresponding to 268, 182, 83 states, for pH 4.5, 6.8, and 7.6, respectively. The dashed lines in panels (A-C) represent the absolute agreement, R is the Pearson's correlation coefficicent, and Q is the quality factor. (D) The improvement in the quality of fit as a function of the number of most populated states included. The states are sorted in descending order by their MaxEnt solution weights. The dashed line shows the best possible χ^2/L value, ε_r/L, computed by minimizing Eq. 5.

I RDC Data for Lys48-linked Ub_2

Table S1: RDC data for Lys48-linked Ub_2 at pH 4.5 for distal (A) and proximal (B) units.

Residue	Unit	Exp. RDC (Hz)	Excluded(*)
2	A	-16.343	
4	A	19.785	
5	A	16.157	
6	A	18.456	
8	A	12.336	*
10	A	19.523	*
11	A	-17.340	*
12	A	13.782	*
13	A	9.840	
14	A	19.217	
15	A	18.572	
16	A	4.504	
17	A	0.889	
20	A	-15.356	*
23	A	18.802	*
25	A	9.993	*
26	A	17.878	
29	A	16.159	
32	A	7.077	
33	A	16.806	
34	A	16.017	
35	A	-23.653	*
36	A	-27.602	
39	A	21.018	
40	A	-2.738	
41	A	17.180	
43	A	18.254	
44	A	17.623	
45	A	6.173	
46	A	9.697	*
47	A	16.834	*
48	A	-26.460	
49	A	1.185	
51	A	-10.313	
52	A	-35.801	*
54	A	-8.138	
55	A	15.504	
56	A	18.507	*
57	A	19.730	

58	A	16.362	
59	A	16.757	
60	A	17.294	
61	A	15.843	
62	A	5.648	*
63	A	-17.967	
64	A	21.065	
67	A	15.657	
68	A	17.980	
69	A	14.376	
70	A	14.725	*
71	A	15.444	*
73	A	18.795	*
74	A	13.873	*
75	A	4.365	*
76	A	5.323	*
2	B	5.860	
3	B	1.357	
4	B	-5.472	*
5	B	-9.017	
6	B	-12.888	
7	B	-6.089	
8	B	10.779	*
10	B	-1.366	*
11	B	2.400	*
12	B	-7.544	*
14	B	-9.557	
15	B	3.050	
16	B	2.622	
17	B	11.074	
20	B	-2.649	
23	B	-12.741	*
25	B	-12.307	*
26	B	-12.729	
29	B	-12.160	
30	B	-13.063	
32	B	-13.095	
33	B	-12.243	
34	B	-12.161	
35	B	-3.357	
36	B	2.840	
39	B	4.651	
40	B	-1.399	
41	B	9.348	
42	B	-4.758	

43	B	-1.352	*
44	B	-11.618	
45	B	-6.436	
46	B	-8.657	
47	B	-12.055	
48	B	-1.063	
49	B	-6.703	
50	B	-11.558	
51	B	-4.944	
52	B	-1.889	
54	B	-6.590	
55	B	-11.353	
56	B	3.556	
57	B	9.635	
58	B	3.695	
59	B	-2.292	
60	B	8.474	
61	B	11.944	
62	B	5.859	*
63	B	-6.821	
64	B	9.999	
65	B	5.664	*
66	B	1.263	
68	B	-12.070	
69	B	-3.604	
70	B	-2.256	*
71	B	5.962	*
72	B	-15.413	*
73	B	1.375	*
74	B	0.493	*
75	B	3.262	*
76	B	-0.937	*

Table S2: RDC data for Lys48-linked Ub_2 at pH 6.8 for distal (A) and proximal (B) units.

Residue	Unit	Exp. RDC (Hz)	Excluded(*)
2	A	0.464	
6	A	19.595	
7	A	11.395	
8	A	17.000	*
10	A	17.528	*
11	A	-4.079	*

13	A	10.318	
14	A	20.102	
15	A	18.051	
16	A	13.851	
17	A	13.011	
18	A	-4.685	
20	A	-15.897	*
21	A	14.802	
25	A	13.530	*
26	A	18.422	
27	A	18.496	
28	A	7.203	
29	A	17.219	
31	A	11.695	
32	A	7.885	
33	A	17.274	
34	A	15.553	
35	A	-31.093	*
36	A	-12.567	
39	A	17.748	
40	A	-10.085	
41	A	3.375	
42	A	9.711	*
43	A	16.691	
45	A	11.584	
46	A	16.642	
47	A	17.158	*
48	A	-19.412	
49	A	-1.265	*
50	A	14.385	
51	A	-0.886	
52	A	-32.860	*
54	A	-15.377	
56	A	8.842	
58	A	18.783	
59	A	11.731	
60	A	3.809	*
61	A	18.634	
62	A	-11.706	*
63	A	-20.399	
64	A	16.877	
65	A	-5.861	
66	A	18.910	
67	A	11.188	
68	A	18.967	

69	A	17.359	
71	A	1.835	*
72	A	-10.592	*
73	A	7.579	*
74	A	3.893	*
75	A	0.404	*
76	A	1.233	*
2	B	6.157	
3	B	1.419	
5	B	-0.782	
6	B	5.825	
7	B	8.653	
8	B	11.017	*
10	B	11.709	*
11	B	5.373	*
12	B	9.986	*
13	B	0.087	
14	B	2.579	
15	B	-0.934	
16	B	4.445	
17	B	9.203	
18	B	5.969	
20	B	-9.729	*
21	B	13.457	
23	B	6.559	*
25	B	5.035	*
26	B	4.858	
28	B	5.622	
29	B	3.865	
30	B	8.009	
31	B	8.308	
32	B	5.550	
33	B	4.042	
34	B	9.381	
35	B	-11.149	*
36	B	2.188	
39	B	-2.075	
40	B	-12.486	
41	B	-1.275	*
42	B	9.492	*
43	B	12.755	
44	B	10.016	
45	B	7.421	
46	B	8.256	
47	B	4.631	

48	B	-5.611	
49	B	-4.960	
50	B	9.862	
51	B	2.077	
52	B	-9.606	
54	B	-4.305	
55	B	4.776	
56	B	-8.699	
57	B	-7.949	*
58	B	0.887	
59	B	-6.157	
60	B	0.323	*
61	B	9.850	
62	B	-11.482	*
63	B	-4.305	
64	B	-1.301	
65	B	-4.377	
66	B	-2.088	
67	B	-3.438	
68	B	8.822	
69	B	13.411	
70	B	6.662	*
71	B	1.207	*
72	B	-15.862	*
73	B	-4.552	*
74	B	1.567	*
75	B	1.227	*
76	B	-1.420	*
77	B	-0.364	*

Table S3: RDC data for Lys48-linked Ub_2 at pH 7.6 for distal (A) and proximal (B) units.

Residue	Unit	Exp. RDC (Hz)	Excluded(*)
2	A	3.891	
4	A	18.980	
5	A	13.451	
6	A	19.596	
7	A	11.631	
8	A	17.371	*
11	A	-1.965	*
13	A	9.342	
14	A	19.932	

15	A	18.544	
16	A	16.733	
17	A	15.670	
18	A	-0.327	
20	A	-16.850	*
21	A	16.456	
25	A	24.776	*
26	A	17.615	
27	A	18.606	
28	A	8.326	
29	A	18.121	
31	A	11.729	
32	A	8.795	
33	A	17.289	
34	A	14.990	
35	A	-31.729	*
36	A	-9.043	
39	A	16.396	
40	A	-12.022	
41	A	0.828	
42	A	9.712	*
43	A	16.429	
44	A	17.597	
45	A	10.820	
47	A	16.612	*
48	A	-20.277	
49	A	-3.714	*
50	A	14.624	
51	A	2.097	
52	A	-31.272	*
54	A	-16.176	
55	A	17.724	
56	A	5.983	
58	A	18.989	
59	A	9.690	
60	A	1.119	*
61	A	16.386	
62	A	-14.836	*
63	A	-18.968	
64	A	16.182	
65	A	-8.627	
66	A	17.958	
67	A	9.968	
68	A	18.887	
69	A	17.315	

70	A	4.478	*
71	A	-0.553	*
72	A	-15.314	*
73	A	3.165	*
2	B	6.739	
5	B	1.064	
6	B	10.284	
7	B	10.587	
8	B	12.282	*
11	B	5.604	*
12	B	12.525	*
13	B	1.551	
14	B	6.538	
15	B	0.438	
16	B	6.997	
17	B	10.367	
18	B	5.689	
20	B	-12.824	*
21	B	14.900	
23	B	10.800	*
25	B	8.541	*
26	B	8.715	
28	B	10.337	
29	B	8.355	
30	B	12.282	
31	B	12.545	
32	B	9.727	
33	B	7.849	
34	B	14.057	
35	B	-14.663	*
36	B	1.573	
39	B	-0.972	
40	B	-15.073	
41	B	-2.710	*
42	B	11.856	*
43	B	15.921	
44	B	14.178	
45	B	8.657	
47	B	8.128	
48	B	-8.558	
49	B	-6.912	
50	B	14.364	
51	B	4.433	
52	B	-13.782	
54	B	-4.390	

55	B	9.626	
56	B	-10.109	
57	B	-9.932	
58	B	2.879	
59	B	-5.624	
60	B	-0.903	*
61	B	10.596	
62	B	-15.074	*
63	B	-4.842	
64	B	-1.628	
65	B	-6.429	*
66	B	-0.745	
67	B	-2.620	
68	B	13.243	
70	B	7.759	*
71	B	-0.009	*
72	B	-20.862	*
73	B	-7.193	*
74	B	-1.721	*
77	B	-0.422	*

References

1. Pati, Y. C.; Rezaiifar, R.; Krishnaprasad, P. S. *Proc. 27th Annu Asilomar Conf Signals, Systems, and Computers* **1993**, 40–44.
2. Bruckstein, A.; Donoho, D.; Elad, M. *SIAM Rev.* **2009**, *51*, 34–81.
3. O'Leary, D. P. *Scientific Computing with Case Studies*; Society for Industrial Mathematics, 2009.
4. Cormen, T. H.; Stein, C.; Rivest, R. L.; Leiserson, C. E. *Introduction to Algorithms*; McGraw-Hill Higher Education, 2001.
5. Salmon, L.; Bascom, G.; Andricioaei, I.; Al-Hashimi, H. M. *J Am Chem Soc* **2013**, *135*, 5457–5466.
6. Blumensath, T.; Davies, M. E. *J Fourier Anal Appl* **2008**, *14*, 629–654.
7. Liberty, E. *arXiv preprint arXiv:1206.0594* **2012**,
8. Bruckstein, A. M.; Elad, M.; Zibulevsky, M. *IEEE Trans Inf Theory* **2008**, *54*, 4813–4820.
9. Jia, J.; Rohe, K. *arXiv preprint arXiv:1208.5584* **2012**,
10. Lavalle, S. M.; Kuffner Jr, J. J. *Algorithmic and Computational Robotics: New Directions*; 2000.

11. Cortés, J.; Siméon, T.; De Angulo, V.; Guieysse, D.; Remaud-Siméon, M.; Tran, V. *Bioinformatics* **2005**, *21*, i116–i125.

12. Enosh, A.; Raveh, B.; Furman-Schueler, O.; Halperin, D.; Ben-Tal, N. *Biophys J* **2008**, *95*, 3850–3860.

13. Raveh, B.; Enosh, A.; Schueler-Furman, O.; Halperin, D. *PLOS Comput Biol.* **2009**, *5*, e1000295.

14. Fushman, D.; Varadan, R.; Assfalg, M.; Walker, O. *Prog. NMR Spectrosc.* **2004**, *44*, 189–214.

15. Byrd, R. H.; Hribar, M. E.; Nocedal, J. *SIAM J Optim* **1999**, *9*, 877–900.