Supporting Material

Partial least squares functional mode analysis

Tatyana Krivobokova[1]
Institute for Mathematical Stochastics and Courant Research Center PEG,
Georg-August-University Göttingen, Germany

Rodolfo Briones[1]
Computational Biomolecular Dynamics Group,
Max-Planck Institute for Biophysical Chemistry, Göttingen, Germany

Jochen S. Hub
Computational Molecular Biophysics Group
Dept. of Molecular Structural Biology
Georg-August-University Göttingen, Germany

Axel Munk
Institute for Mathematical Stochastics,
Georg-August-University Göttingen and Statistical inverse problems group,
Max-Planck Institute for Biophysical Chemistry, Göttingen, Germany

Bert L. de Groot[2]
Computational Biomolecular Dynamics Group,
Max-Planck Institute for Biophysical Chemistry, Göttingen, Germany

[1]equal contribution

[2]Corresponding author. Address: Computational Biomolecular Dynamics Group, Max-Planck Institute for Biophysical Chemistry, Am Fassberg 11, 37077, Göttingen, Germany. Tel.: +(49)551-2012308, Fax: +(49)551-2012302

Supporting Material

PLS algorithm as a regularization scheme

In contrast to PCA-based FMA, partial least squares is an iterative procedure and the resulting PLS estimator is highly non-linear in $\boldsymbol{f}$. In the literature there are several versions of the PLS algorithm available, which are, in fact, equivalent, as shown in (1). To keep the notation simple, in the following we assume that both $\boldsymbol{X}$ and $\boldsymbol{f}$ are centered. The idea of the original algorithm of (2) (known as NIPALS) is to find $k \leq p$ orthogonal n-dimensional components $\boldsymbol{t}_1, \ldots, \boldsymbol{t}_k$, such that $\boldsymbol{t}_i = X\boldsymbol{w}_i$ for some p-dimensional weights $\boldsymbol{w}_i$, $i = 1, \ldots, k$. Thereby, weights $\boldsymbol{w}_i$ are chosen to maximize the empirical covariance between the data $\boldsymbol{f}$ and $\boldsymbol{t}_i$. We describe the construction of $\boldsymbol{t}_i$ following (3). The first component $\boldsymbol{t}_1 = \boldsymbol{X}\boldsymbol{w}_1$ is found solving

$$\boldsymbol{w}_1 = \arg\max_{\boldsymbol{w}} \frac{\mathrm{cov}^2(\boldsymbol{X}\boldsymbol{w}, \boldsymbol{f})}{\boldsymbol{w}^t\boldsymbol{w}} = \arg\max_{\boldsymbol{w}} \frac{\boldsymbol{w}^t\boldsymbol{X}^t\boldsymbol{f}\boldsymbol{f}^t\boldsymbol{X}\boldsymbol{w}}{\boldsymbol{w}^t\boldsymbol{w}}, \tag{1}$$

which gives (up to a scalar) $\boldsymbol{w}_1 = \boldsymbol{X}^t\boldsymbol{f}$. Further components $\boldsymbol{t}_i = \boldsymbol{X}\boldsymbol{w}_i$ are found from Eq. 1, subject to mutual orthogonality of all $\boldsymbol{t}_j$, $j = 1, \ldots, i-1$. One possible way to do this is to set

$$\boldsymbol{t}_i = \boldsymbol{X}\boldsymbol{w}_i = \boldsymbol{X}\boldsymbol{X}^t \left\{ \boldsymbol{f} - \boldsymbol{T}_{i-1}(\boldsymbol{T}_{i-1}^t\boldsymbol{T}_{i-1})^{-1}\boldsymbol{T}_{i-1}^t\boldsymbol{f} \right\}, \tag{2}$$

for $\boldsymbol{T}_{i-1} = (\boldsymbol{t}_1, \ldots, \boldsymbol{t}_{i-1})$, $i = 2, \ldots, k$. Hence, for $\boldsymbol{W}_k = (\boldsymbol{w}_1, \ldots, \boldsymbol{w}_k)$ and $\boldsymbol{T}_k = \boldsymbol{X}\boldsymbol{W}_k$, we can write the partial least squares estimator of order k for $\boldsymbol{f}$ as

$$\widehat{\boldsymbol{f}}_{PLS}^k = \boldsymbol{X}\hat{\boldsymbol{\beta}}_{PLS}^k = \boldsymbol{X}\boldsymbol{W}_k\hat{\boldsymbol{\alpha}}_k = \boldsymbol{T}_k\hat{\boldsymbol{\alpha}}_k = \boldsymbol{T}_k(\boldsymbol{T}_k^t\boldsymbol{T}_k)^{-1}\boldsymbol{T}_k\boldsymbol{f}. \tag{3}$$

This iterative definition (that is, $\boldsymbol{t}_i$ obtained from $(\boldsymbol{t}_1, \ldots, \boldsymbol{t}_{i-1})$) of the PLS algorithm gives a good intuition how the method works: one is looking for mutual orthogonal predictors $\boldsymbol{t}_i$, that have the maximal covariance with $\boldsymbol{f}$.

The ewMCM is the collective mode that, estimated from the given structural ensemble, has the highest probability to achieve a specific alteration of the order parameter. From the presented PLS definition one can also easily verify why ewMCM weights are equivalent to the scaled $\boldsymbol{W}_1$. Simple algebra allows to rewrite the definition of the ewMCM weights given in Eq. 12 of (4) as $\boldsymbol{W}_k\boldsymbol{W}_k^t\boldsymbol{X}^t\boldsymbol{f}\widehat{\mathrm{var}}^{-1}(\hat{\boldsymbol{f}}_{PLS}^k)/n$, with $\widehat{\mathrm{var}}(\hat{\boldsymbol{f}}_{PLS}^k)$ as the sample variance of $\boldsymbol{f}_{PLS}^k$. Note that the first basis vector $\boldsymbol{W}_1$ is given by $\boldsymbol{w}_1 = \boldsymbol{X}^t\boldsymbol{f}$, while subsequent vectors are chosen so that they are mutually orthogonal. Thus, in vector $\boldsymbol{W}_k^t\boldsymbol{X}^t\boldsymbol{f}$ the first element equals $\boldsymbol{f}^t\boldsymbol{X}\boldsymbol{X}^t\boldsymbol{f}$, whereas the others are zero, so that ewMCM weights result in

$$\boldsymbol{W}_1 \frac{\boldsymbol{f}^t\boldsymbol{X}\boldsymbol{X}^t\boldsymbol{f}}{n\widehat{\mathrm{var}}(\hat{\boldsymbol{f}}_{PLS}^k)}. \tag{4}$$

Another – non-iterative – formulation of PLS algorithm makes obvious the connection of partial least squares to the conjugate gradient method. It has been for a long time known, see e.g., (5), that PLS is equivalent to the approximate solution of $(\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{\beta} = \boldsymbol{X}^t\boldsymbol{f}$ by the conjugate gradient method with early stopping k. Denote $\boldsymbol{H} = \boldsymbol{X}^t\boldsymbol{X}$, $\boldsymbol{s} = \boldsymbol{X}^t\boldsymbol{f}$, $\boldsymbol{K}_k = (\boldsymbol{s}, \boldsymbol{H}\boldsymbol{s}, \ldots, \boldsymbol{H}^{k-1}\boldsymbol{s})$. Then, (1) has shown that

$$\widehat{\boldsymbol{f}}_{PLS}^k = \boldsymbol{X}\hat{\boldsymbol{\beta}}_{PLS}^k = \boldsymbol{X}\boldsymbol{K}_k(\boldsymbol{K}_k^t\boldsymbol{X}^t\boldsymbol{X}\boldsymbol{K}_k)^{-1}\boldsymbol{K}_k^t\boldsymbol{X}^t\boldsymbol{f}. \tag{5}$$

Early stopping

To understand the excellent predictive power of the PLS method, it is pertinent to recall that this algorithm has to be stopped at an early stage k, the dimension of the basis. Besides its computational speed up this is necessary to avoid overfitting. In fact, from Eq. 5 it becomes obvious that as the iteration k increases, $\hat{\boldsymbol{\beta}}^k_{PLS}$ converges essentially to the unrestricted least squares fit $(\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{f}$ which overfits the data and hence yields insufficient predictive power. Hence, the iteration depth k serves as a regularization parameter, which has to be chosen properly to terminate the algorithm. This early stopping phenomenon is well-known for a variety of other learning algorithms, such as iterative Tikhonov or boosting, see (6) and (7), to mention a few. Recently, (8) could show that the PLS is able to obtain optimal rates of reconstruction, if combined with an early stopping rule based on the discrepancy principle. However, this rule merely is of theoretical interest and it can hardly be employed in practice because it finally depends on the unknown true parameter $\boldsymbol{\beta}$ itself. Hence, for practical purposes we suggest a cross-validation method to yield the optimal k.

Trp-cage unfolding

The Trp-cage is a 20 amino acid miniprotein designed by Neidigh et al. (9) with a very short folding time of 4 μs. Here we applied FMA to identify unfolding pathways. Therefore, as the functional property $\boldsymbol{f}$ we chose the hydrophobic solvent accessible surface (hSAS) that increases during unfolding. Protein atoms excluding hydrogens were used to perform the PLS- and PCA-based FMA analysis. This represents a particularly challenging case, firstly because the hSAS is a highly non-linear function of the coordinates and secondly because we only use trajectory parts of the folded peptide for training the PLS- and PCA-based FMA models, and assess the predictive power by cross-validating against initial unfolding trajectories (4). The input trajectory was constructed such that the training part consists of a concatenation of folded trajectories of 100 ns combined length, and the cross-validation part consists of a combination of initial unfolding trajectories (100−180 ns).

Fig. S2, *A* and *B* show that the R_m converges to around 0.8 for PLS and to approx. 0.7 for PCA-based FMA. PLS R_c converges to a vaue of around 0.6 after 20 components, suggesting that a linear model can only partly describe the hSAS, whereas atomic fluctuations that are not captured by the MCM substantially contribute to the hSAS. For PCA-based FMA a maximum R_c is reached at about 40 EVs, whereas for PLS-based FMA a basis of dimensionality six yields the highest correlation in cross validation. Fig. S2, *C* and *D* show the data and model for both the training and cross-validation parts. 6 components and 40 PCA eigenvectors were used for the PLS- and PCA-based FMA analysis, respectively. The corresponding ewMCM are displayed in Figs. S2, *E* and *F*. Concerning the backbone, the ewMCM of PLS- and PCA-based FMA show a similar opening motion, with a scalar product of 0.985 between the two ewMCMs.

References

1. Helland, I. S., 1988. On the structure of partial least squares regression. *Commun. Stat. Simulat.* 17:581–607.

2. Hold, H. O. A., 1973. Nonlinear iterative partial least squares (NIPALS) modelling: some current developments. *In* P. Krishnaiah, editor, Multivariate analysis III, Academic Press, New York, 383–407.

3. Krämer, N., A.-L. Boulesteix, and G. Tutz, 2008. Penalized partial least squares with applications to B-spline transformations and functional data. *Chemometr. Intell. Lab.* 94:60–69.

4. Hub, J. S., and B. L. de Groot, 2009. Detection of functional modes in protein dynamics. *PLoS Comput. Biol.* 5:e1000480.

5. Manne, R., 1987. Analysis of two partial least squares algorithms for multivariate calibration. *Chemometr. Intell. Lab.* 2:187–197.

6. Bühlmann, P., and B. Yu, 2003. Boosting with the L2 loss: regression and classification. *J. Am. Stat. Assoc.* 98:324–339.

7. Bissantz, N., T. Hohage, A. Munk, and F. Ruymgaart, 2007. Convergence rates of general regularization methods for statistical inverse problems and applications. *SIAM J. Numer. Anal.* 45:2610–2636.

8. Blanchard, G., and N. Krämer, 2010. Optimal learning rates for kernel conjugate gradient regression. *In* Adv. Neur. In. volume 23, 226–234.

9. Neidigh, J. W., R. M. Fesinmeyer, and N. H. Andersen, 2002. Designing a 20-residue protein. *Nat. Struct. Biol.* 9:425–430.

10. Fischer, G., U. Kosinska-Eriksson, C. Aponte-Santamaría, M. Palmgren, C. Geijer, K. Hedfalk, S. Hohmann, B. L. de Groot, R. Neutze, and K. Lindkvist-Petersson, 2009. Crystal structure of a yeast aquaporin at 1.15 Å reveals a novel gating mechanism. *PLoS Biol.* 7:e1000130.

Figure Legends

Figure S1.

Experimental ensemble prediction and fitting structure effect on PLS-based FMA method for Glu11-Asp20 distance (d_{ED}) of T4 Lysozyme (T4L). (A) Scatter plot and linear regression of the experimental vs. predicted d_{ED} for 38 T4L x-ray structures. The PLS-based FMA model was built as in Fig.1 B using 10 components. (B) Pearson correlation coefficients between data and model for PLS-based FMA as function of the number of components calculated for the model training subset (*black*, R_m, 38 T4L x-ray structures) and the cross-validation

subset (*red*, R_c, 4601 T4L MD frames). (C) Cartoon representation of 3 T4L structures (with 0.76, 1.01 and 1.24 nm of d_{ED}) used to test the effect of the fitting structure on the PLS-based FMA. (D) Pearson correlation coefficients between data and model for PLS-based FMA as function of the number of PLS components calculated for the model training (R_m) and the cross-validation subsets (R_c). The correlation coefficients were calcuted using the reference structures of B. In the inset it is possible to observe the subtle differences for R_c among these PLS-based FMA models.

Figure S2.

Comparison of PLS- and PCA-based FMA for hydrophobic solvent accessible surface (hSAS) applied to Trp-cage folded trajectories. (A/B) Pearson correlation coefficients between data and model for PLS/PCA based FMA as function of the number of PLS components/PCA vectors calculated for the model training subset (*black*, R_m, folded trajectories) and the cross-validation subset (*red*, R_c, initial unfolding trajectories). (C/D) Overlay of data and model for the hSAS as function of time. The *black lines* correspond to the MD data, the *green* to the model training subset and *red* to the cross-validation subset. The models were calculated using 6 components for PLS-based (C) and 40 PCA vectors for PCA-based FMA (D). (E/F) Cartoon and stick representation of the ewMCM contributing to the change in the hSAS. The color-scale (*red-white-blue*) represents the interpolation between the extreme projections along the ewMCMs. The PLS- and PCA-based FMA models used to plot the molecular representations have the same number of components or PCA vectors as the models used in panels C and D.

Figure S3.

Comparison of the PLS-based FMA and partial PCA analysis (10) of the helices 4, 5, 6 and loop D of yeast Aquaporin (Aqy1). (A) Backbone representations and overlay of the modes along the first PCA eigenvector (*red*) and the PLS-based FMA ewMCM (*blue*). (B) Comparison of the modes in A in terms of root mean-square fluctuation. The color-scale (*blue-green-red*) and the line thickness represents the RMSF along the modes.

Figure S4.

Distribution of the E148 Ψ angle for doubly occupied CLC-ec1 monomers at S_{cen} and S_{int}. The simulations corresponded to: S_{int} and S_{cen} restrained; S_{int} restrained and S_{cen} free, and free S_{int} and S_{cen}. They show different proportions of the three Ψ peaks.

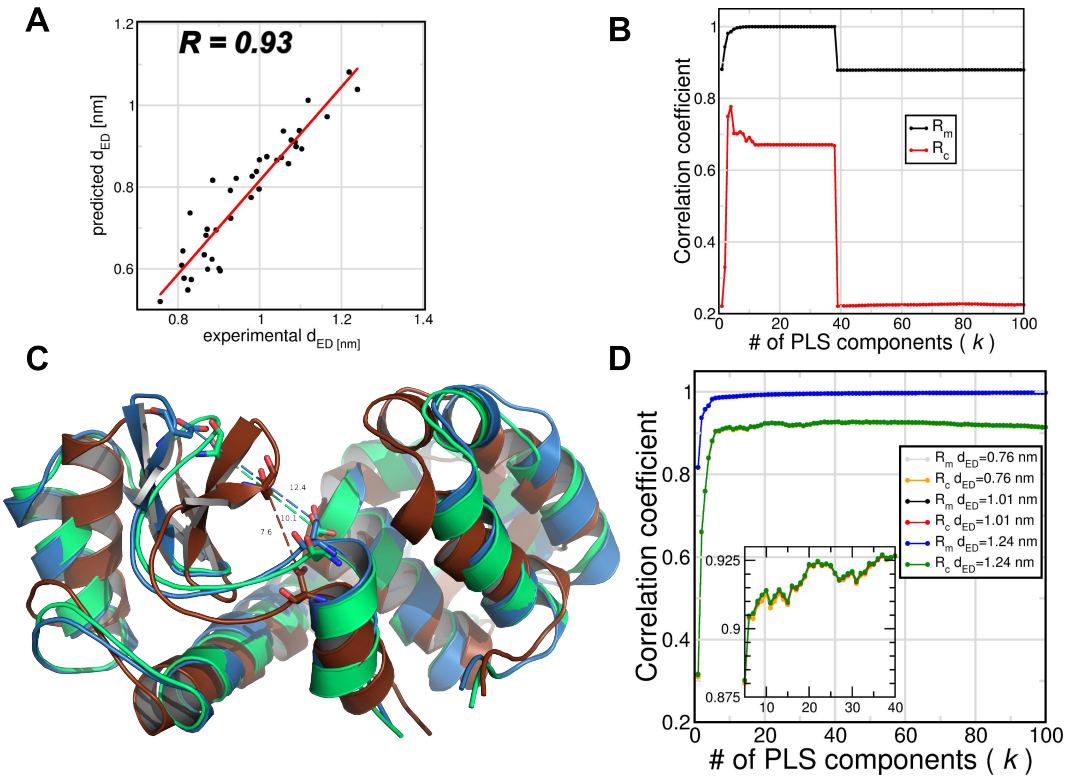
A
R = 0.93
predicted d_ED [nm]
experimental d_ED [nm]
B
Correlation coefficient
# of PLS components (k)
R_m
R_c
C
12.4
10.1
7.6
D
Correlation coefficient
# of PLS components (k)
R_m d_ED=0.76 nm
R_c d_ED=0.76 nm
R_m d_ED=1.01 nm
R_c d_ED=1.01 nm
R_m d_ED=1.24 nm
R_c d_ED=1.24 nm

Figure S1:

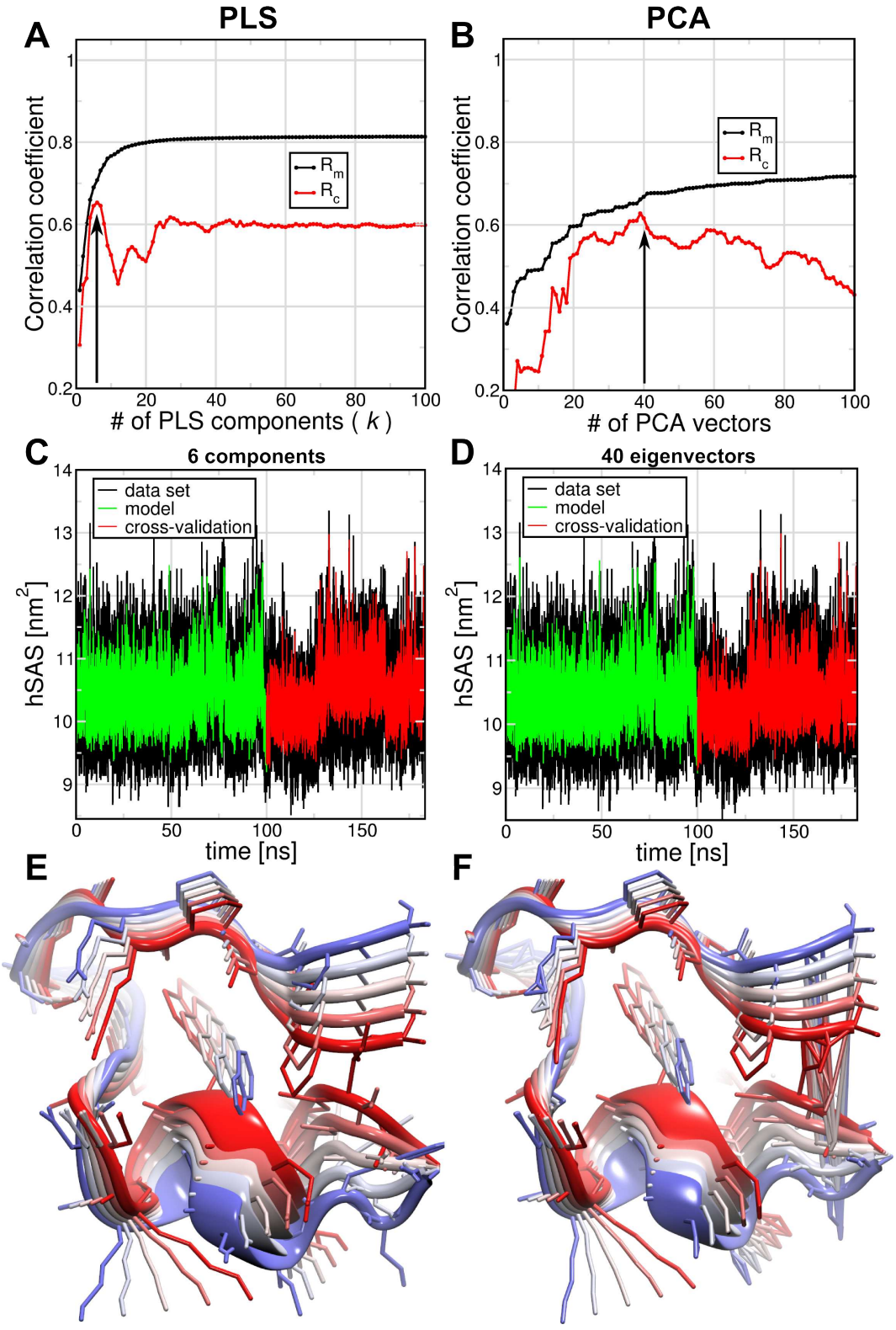
A
PLS
Correlation coefficient
# of PLS components (k)
B
PCA
Correlation coefficient
# of PCA vectors
C
6 components
data set
model
cross-validation
hSAS [nm²]
time [ns]
D
40 eigenvectors
data set
model
cross-validation
hSAS [nm²]
time [ns]
E
F

Figure S2:

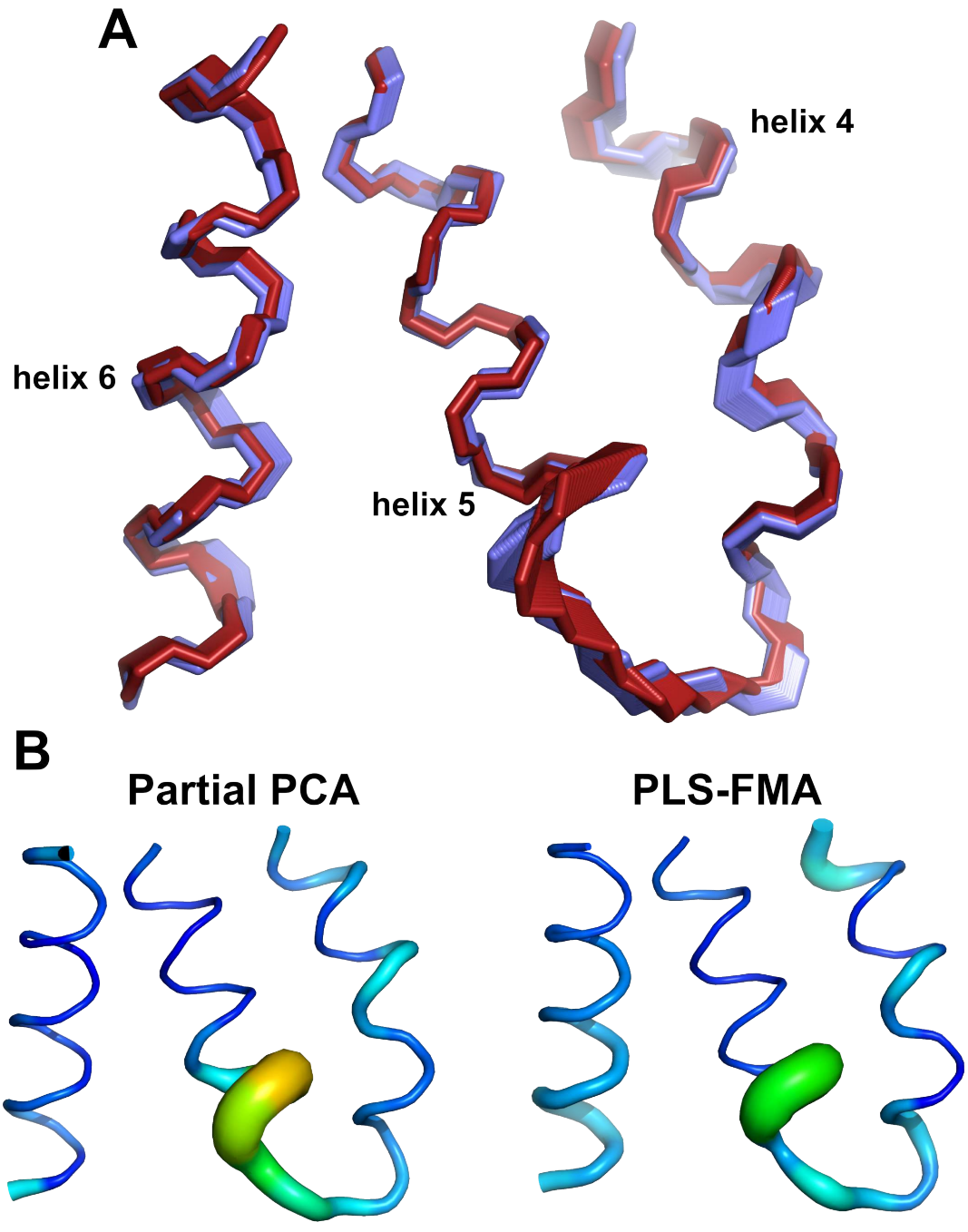

Figure S3:

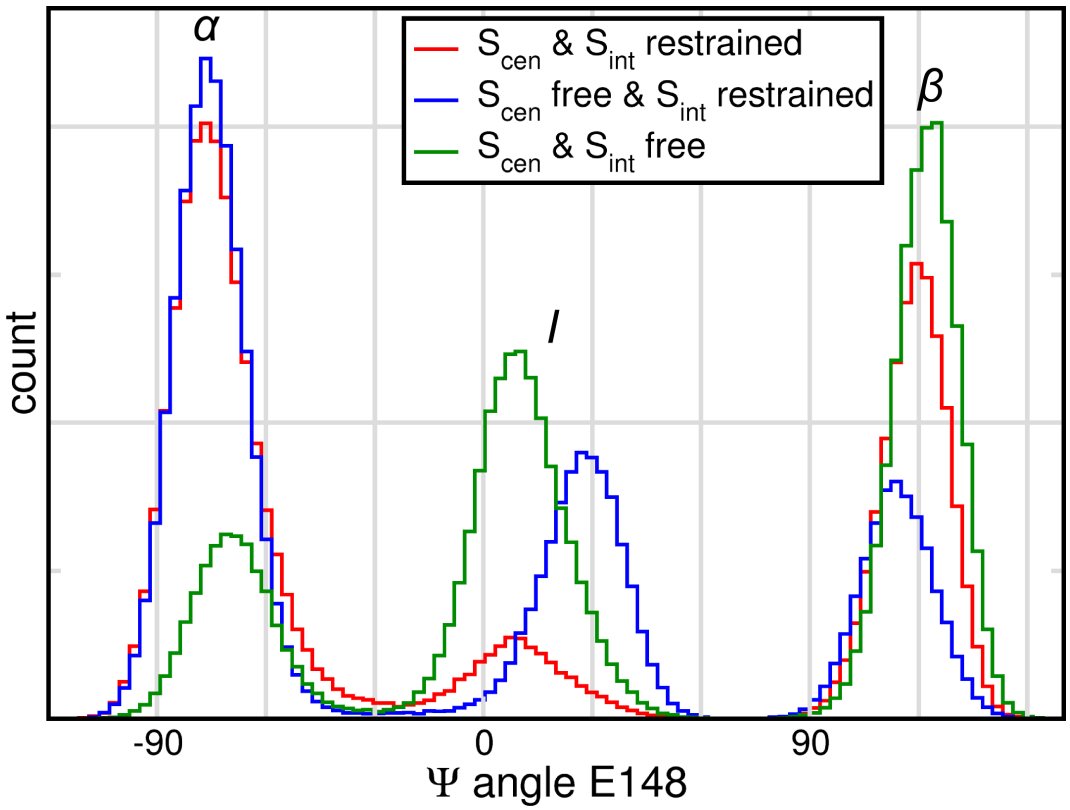
α
β
l
S_cen & S_int restrained
S_cen free & S_int restrained
S_cen & S_int free
count
-90
0
90
Ψ angle E148

Figure S4: