Supporting information

Structure-based methyl resonance assignment with MethylFLYA

Iva Pritišanac,[1] Julia Würz,[1] T. Reid Alderson,[2] and Peter Güntert*[,1,3,4]

[1]Institute of Biophysical Chemistry, Center for Biomolecular Magnetic Resonance, Goethe University Frankfurt am Main, 60438 Frankfurt am Main, Germany
[2] Laboratory of Chemical Physics, NIDDK, National Institutes of Health, Bethesda, Maryland 20892-0520, United States
[3]Laboratory of Physical Chemistry, ETH Zürich, 8093 Zürich, Switzerland
[4]Graduate School of Science, Tokyo Metropolitan University, Hachioji, Tokyo 192-0397, Japan
*Correspondence to: guentert@em.uni-frankfurt.de

Table S2 Results of the CYPICK protocol application to the 3D CCH NOESY spectra (ATCase-r2, HSP90-N) and the 4D HCCH NOESY spectrum (EIN)

Table S3 FLYA's calculation speed

Supplementary methods

S1) cyana.lib

S2) Format of the input peak lists for FLYA

S3) Preparation of the input structure (PDB) file

S4) FLYA.cya

S5) peaklists.cya

S6) CONSOL.cya

S7) consol3g.tab

Supplementary references

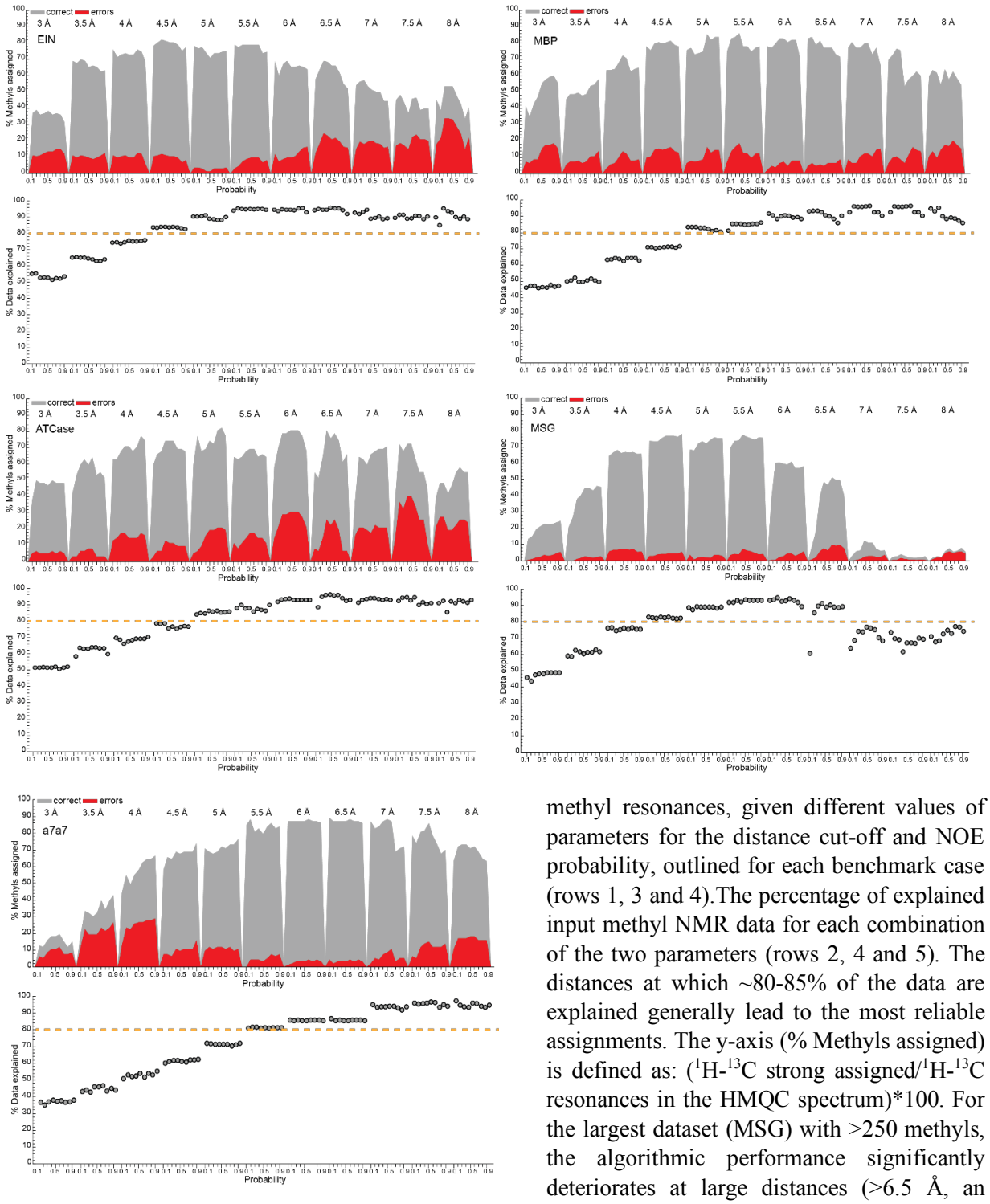


Figure S1 | Parametrization of the distance cut-off and probability of the expected methyl-methyl NOE contacts. The percentage of correctly (*grey*) and erroneously (*red*) strongly (i.e. confidently) assigned methyl resonances, given different values of parameters for the distance cut-off and NOE probability, outlined for each benchmark case (rows 1, 3 and 4).The percentage of explained input methyl NMR data for each combination of the two parameters (rows 2, 4 and 5). The distances at which ~80-85% of the data are explained generally lead to the most reliable assignments. The y-axis (% Methyls assigned) is defined as: (^{1}H-^{13}C strong assigned/^{1}H-^{13}C resonances in the HMQC spectrum)*100. For the largest dataset (MSG) with >250 methyls, the algorithmic performance significantly deteriorates at large distances (>6.5 Å, an equivalent of >11.5 Å C-C distance) resulting in an unreliable assignment. Note that such large distance cut-offs are unrealistic in practice and unlikely to be generally required.

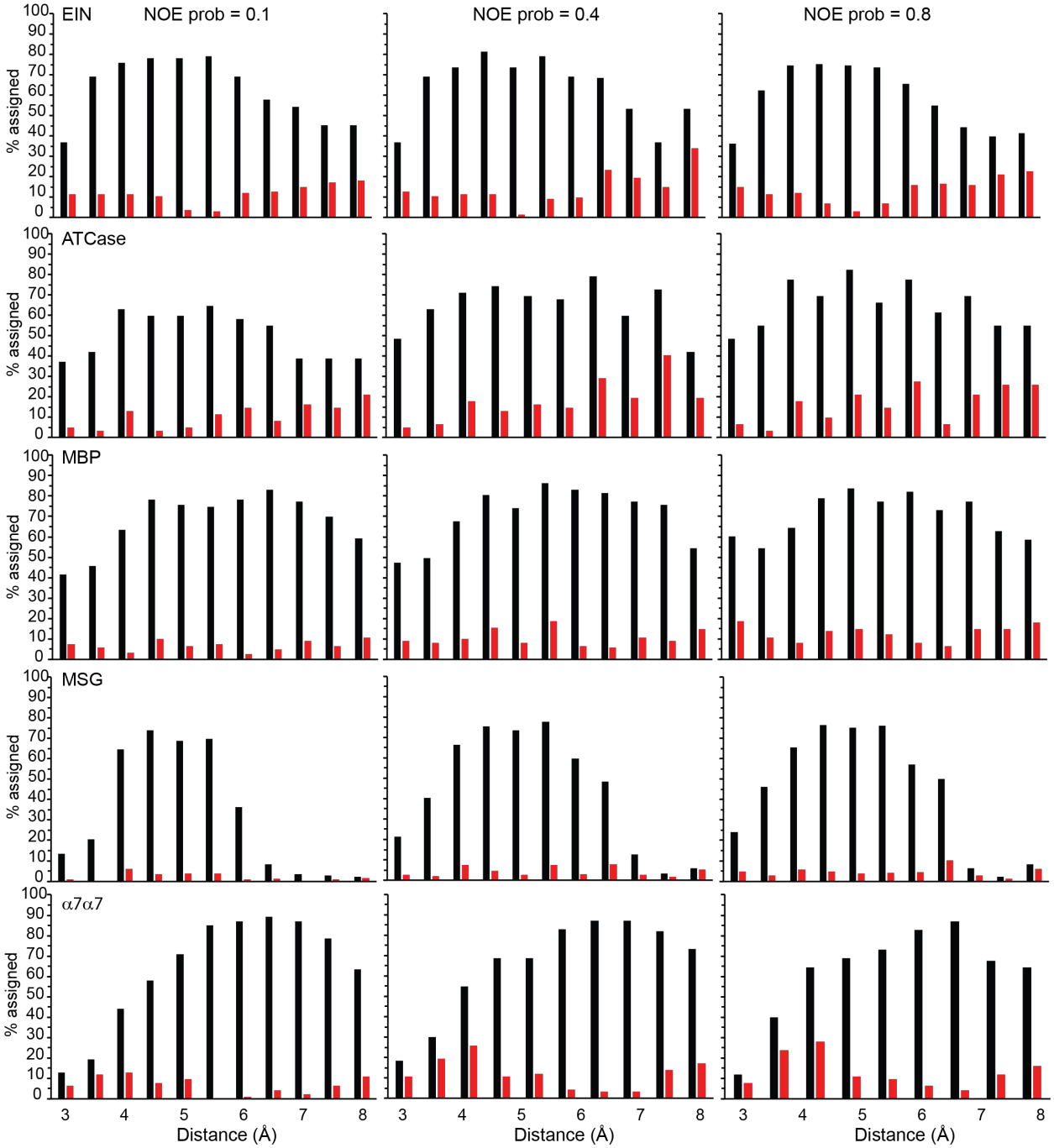


Figure S2 | Accuracy of the methyl assignments obtained for the different values of NOE probabilities over a range of distance thresholds. The percentage of accurately (*black*) and erroneously (*red*) assigned methyl groups is shown for the NOE probability values of 0.1, 0.4, and 0.8 in the first, second, and third column, respectively. The probabilities were tested for a range of $^{1}H_{M}$-$^{1}H_{M}$ distance cut-offs, from 3 to 8 Å, with 0.5 Å steps.

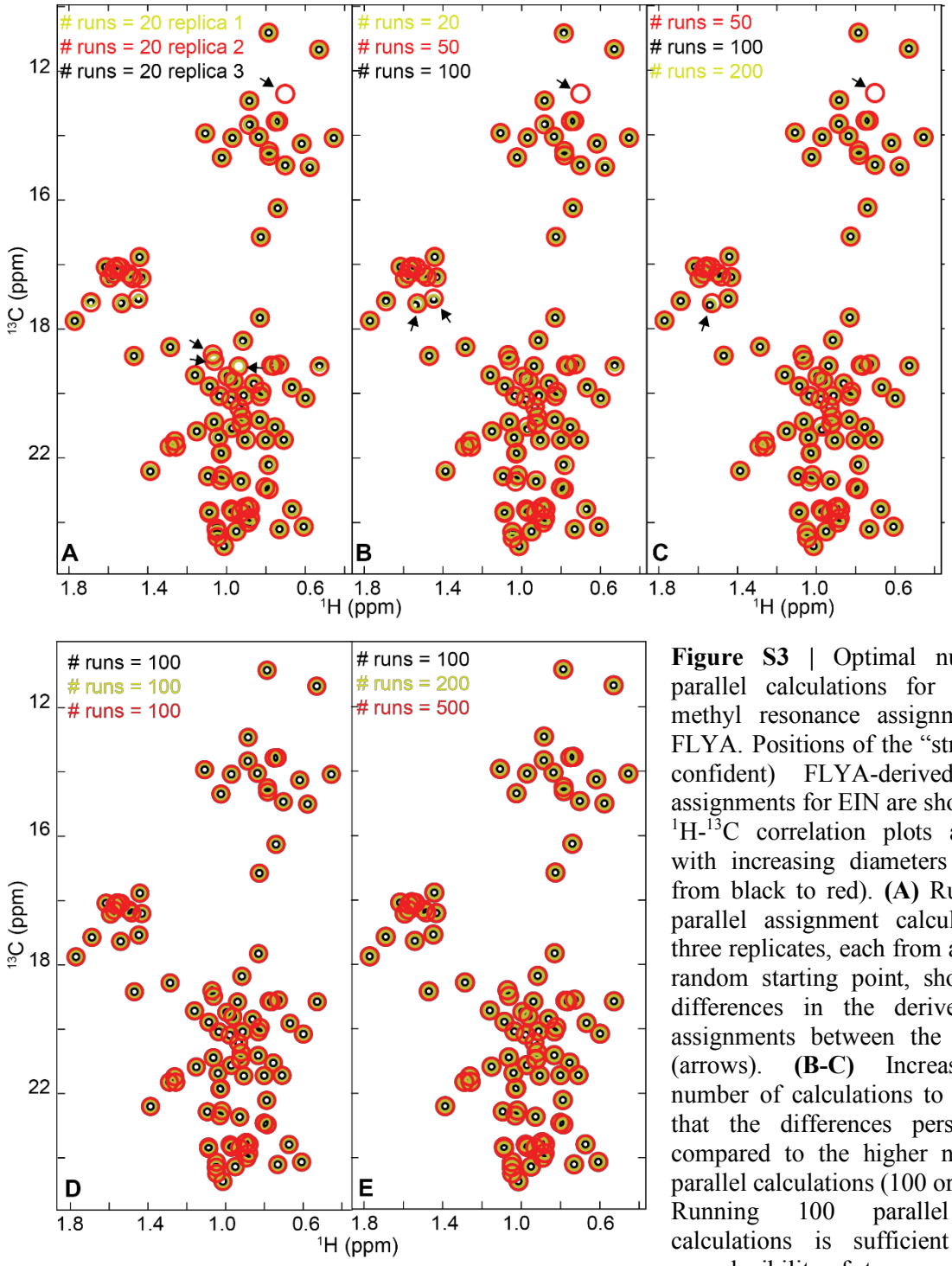


Figure S3 | Optimal number of parallel calculations for automatic methyl resonance assignment with FLYA. Positions of the "strong" (i.e. confident) FLYA-derived methyl assignments for EIN are shown in the ^{1}H-^{13}C correlation plots as circles with increasing diameters (increase from black to red). **(A)** Running 20 parallel assignment calculations in three replicates, each from a different random starting point, shows some differences in the derived strong assignments between the replicates (arrows). **(B-C)** Increasing the number of calculations to 50 shows that the differences persist when compared to the higher number of parallel calculations (100 or 200). **(D)** Running 100 parallel FLYA calculations is sufficient for the reproducibility of strong assignments. **(E)** A further increase in the number of parallel calculations (e.g. 200, 500) results in sets of strong assignments that are consistent with the set of 100 parallel calculations.

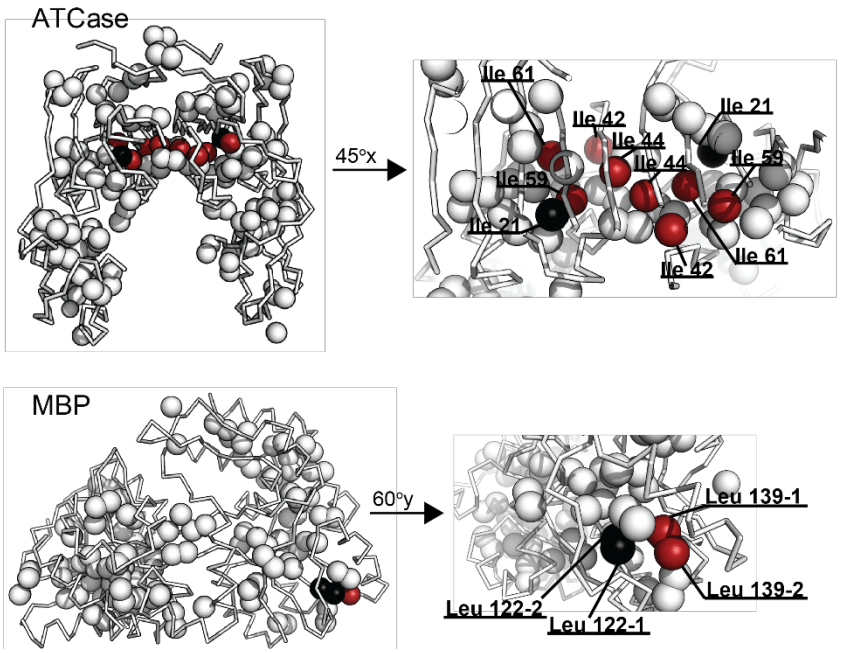


Figure S4 | Sources of errors in the automatic methyl resonance assignments generated by FLYA. Carbon atoms of the erroneously assigned methyls are shown as red spheres, whereas their correct assignment positions are given in black spheres, or exceptionally in red when the assignment at those positions is also incorrect (i.e. for assignment swaps such as Ile42↔Ile44). The mis-assigned resonances belong to nearby methyl groups. For ATCase, the assignment errors cluster at the interface of the two subunits of the homodimer.

Table S1 | Summary of errors in the automatic methyl resonance assignments generated by FLYA. The mis-assigned methyls are assigned to spatially proximal residues.

Protein	Assign error	Assigned to	Distance (Å)*
ATCase	Ile 42	Ile 44	5.5
	Ile 44	Ile 42	5.5
	Ile 59	Ile 21	4
	Ile 61	Ile 59	5.9
MBP	Leu 139_1	Leu 122_1	5.2
	Leu 139_2	Leu 122_2	6.9

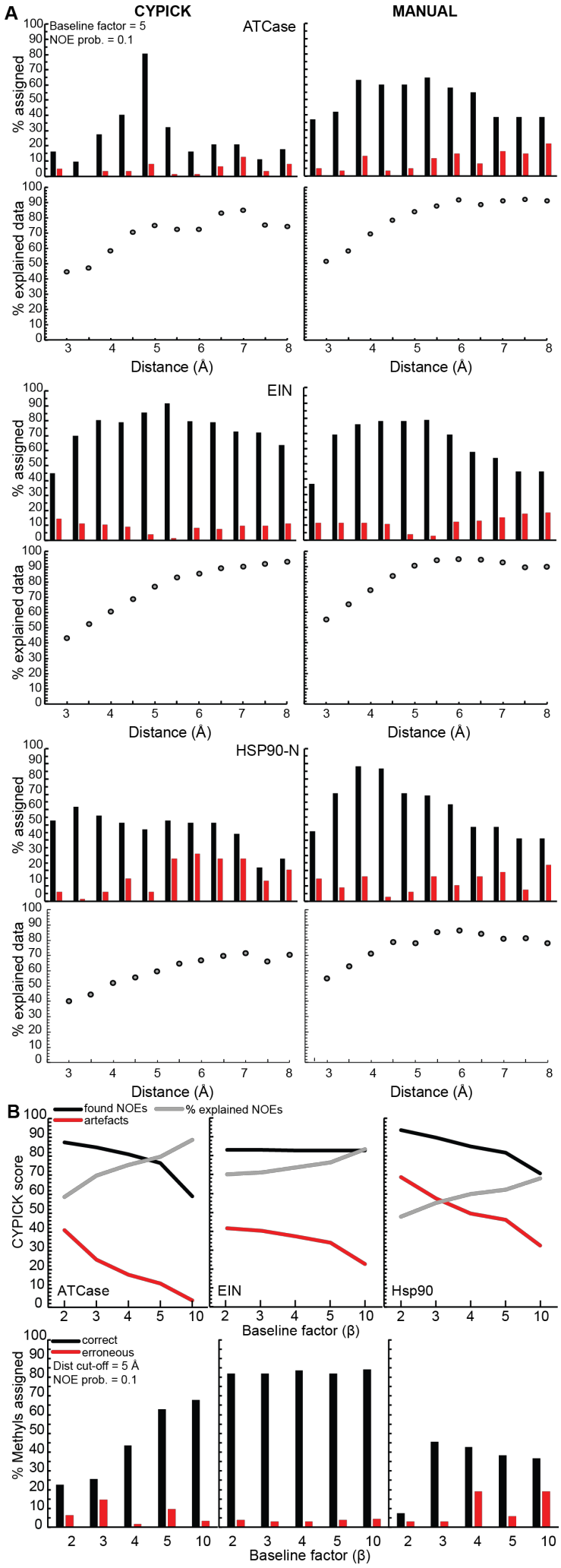


Figure S5 | Optimizing the parameters for automatic NOE signal picking with CYPICK on methyl-methyl NOESY spectra. **A)** The percentage of accurately (black) and erroneously (red) assigned methyls, and of explained NMR data, over a range of $^{1}H_{M}$-$^{1}H_{M}$ distance thresholds (3-8 Å) at a fixed NOE probability value (0.1). Results using the automatically generated CYPICK NOESY lists are in the left columns, and results obtained using the manually prepared NOESY lists are in the right column. **B)** Varying baseline factors for automatic peak picking of methyl-methyl NOESY spectra using CYPICK at a fixed distance cut-off. Variation in the found CYPICK score (*black*) with increasing baseline factor is monitored together with the "artefact" score (red) and the percentage of explained inter-methyl NOEs (*grey*). For all three proteins, a fixed distance cut-off of 5 Å with 0.1 probability was used to define the expected methyl-methyl NOEs. The y-axis % assigned is defined as: (^{1}H-^{13}C strong assigned/^{1}H-^{13}C resonances in the HMQC spectrum)*100. For more details see *Methods*.

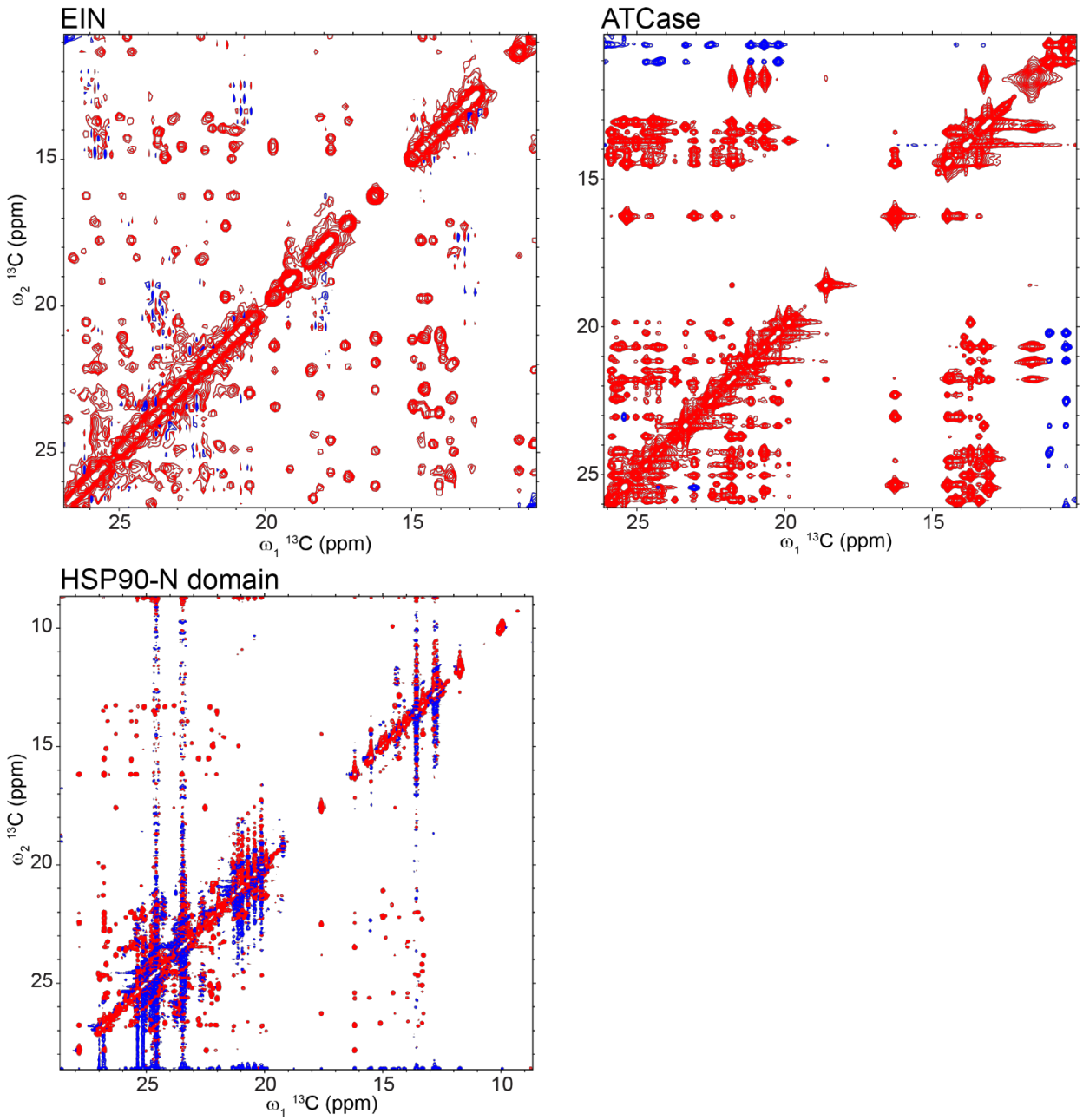


Figure S6 | 2D $^{13}C(\omega_1)$-$^{13}C(\omega_2)$ projections of 3D CCH NOESY (ATCase, HSP90-N domain) or 4D HCCH NOESY spectra (EIN). All spectra are plotted with contour levels at a signal-to-noise ratio of three, as determined by the software Sparky[1]. Positive and negative signals are colored in red and blue, respectively. The shown spectra were acquired previously by Venditti *et al.*[2] (EIN), Velyvis *et al.*[3] (ATCase), and Shah *et al.*[4](HSP90-N).

Table S2 | Results of the CYPICK application to the 3D CCH NOESY spectra (ATCase-r_2, HSP90-N) and the 4D HCCH NOESY spectrum (EIN). The results are given for the baseline factor β=5. A found score indicates which percentage of NOE peaks from the reference were also found by CYPICK. An artefact score is equal to 100 minus the found score, and the overall score is a combination of the two with the artefact score weight of 0.2, as defined by Würz *et al.*[5].

Protein	# peaks (ref)	# peaks (CYPICK)	Overall score	Found score	Artefact score
EIN	618	775	74.30%	82.80%	34%
ATCase-r2	563	495	74.4%	76.6%	12.9%
HSP90-N	409	624	67.8%	81.9%	46.3%

Table S3 | FLYA's calculation speed in sec. for different combinations of input NMR data, as in Fig 3.

Protein	# methyl groups	Filtered NOEs	Unfiltered NOEs	L, V=LV	2L, 2V	L, V=LV, 2LV
EIN	133	1.86E+03	1.81E+03	1.96E+03	1.72E+03	1.95E+03
ATCase-r2	62	1.35E+03	1.43E+03	1.44E+03	1.28E+03	1.46E+03
MBP	123	1.66E+03	1.63E+03	1.91E+03	1.88E+03	2.14E+03
MSG	268	4.43E+03	4.43E+03	5.49E+03	4.33E+03	4.47E+03
a7a7	93	1.81E+03	1.82E+03	2.16E+03	2.49E+03	2.85E+03

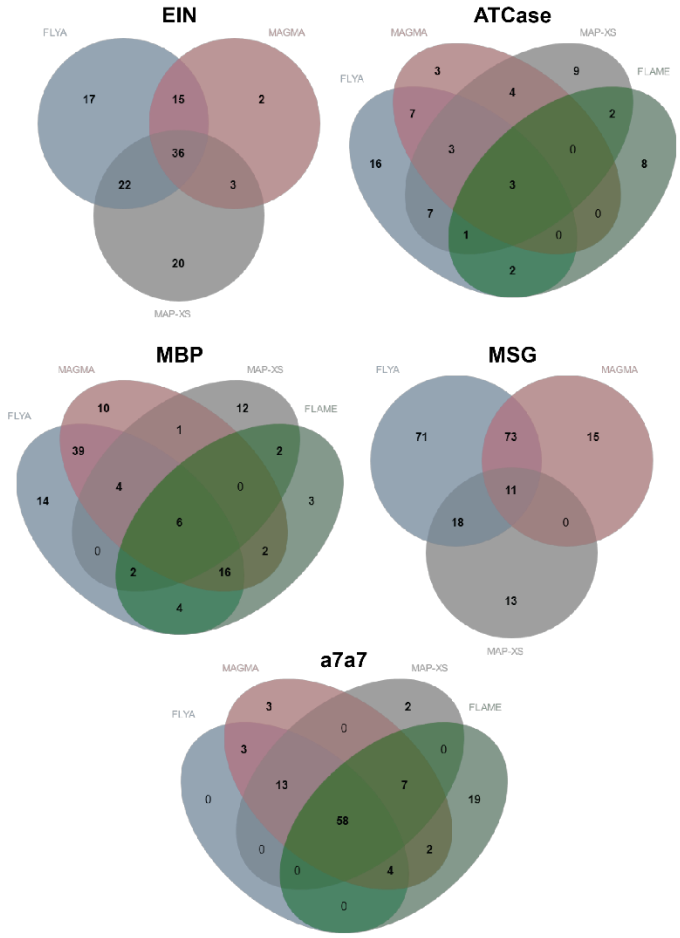
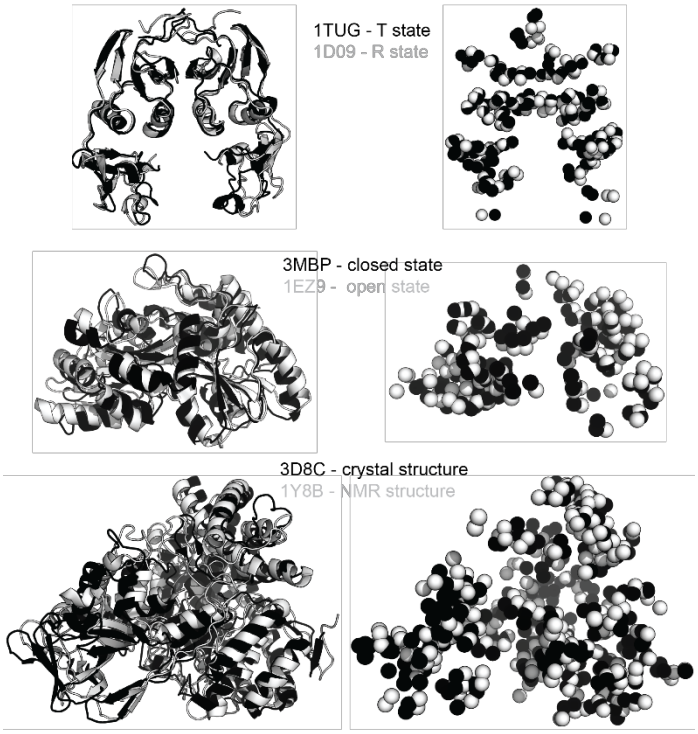


Figure S7 | Intersection of the assignment solutions generated with different automatic methyl assignment protocols. The illustration of the intersections of assignment solutions from FLYA, MAGMA, FLAMEnGO2.0, and MAP-XSII are shown for the indicated benchmark cases. In the case of FLAMEnGO2.0, no confident (100%) assignments were found for EIN and MSG.



Protein	PDB ID	correct	error
atcase	1d09 (Rstate)	36	3
	1tug (Tstate)	36	1
mbp	1ez9 (open)	83	2
	3mbp (closed)	75	2
msg	1y8b (NMR)	129	0
	1d8c (X-ray)	173	0

Figure S8 | FLYA's performance on different input structures for three enzymes in the benchmark. Differences in backbone conformations between the different protein states are shown with protein structures in cartoon representation (*right* column). In the *left* column, the positions of the methyl carbons are indicated with spheres for each of the conformers, with colors matching those assigned to the backbone (*right*). The total number of accurate and erroneous "strong" methyl assignments generated with FLYA using different conformers is summarized in the table (bottom row).

Supplementary methods

S1) cyana.lib

A section of the CYANA library (*cyana.lib*) containing definitions of experiments necessary for automatic methyl resonance assignment with FLYA. The headers include the name of the spectrum and the set of nuclei that constitute the direct and indirect dimensions. The rows specify the magnetization transfer pathways for the defined experiment, where semicolons indicate the nuclei for which the frequency is measured in the experiment. In the first column, the probability of observing an expected peak at the distance indicated in the column starting with "~" are listed. The specified atom names follow the CYANA residue library convention where, e.g., "C_A*" stands for 'aliphatic' carbon (see http://www.cyana.org/wiki/index.php/Residue_library_file).

```
SPECTRUM CCNOESY3D C1 C2 H1
 0.900 C1:C_A* H1:H_A* ~4.0 H_A* C2:C_A*
 0.800 C1:C_A* H1:H_A* ~4.5 H_A* C2:C_A*
 0.700 C1:C_A* H1:H_A* ~5.0 H_A* C2:C_A*
 0.600 C1:C_A* H1:H_A* ~5.5 H_A* C2:C_A*
 0.500 C1:C_A* H1:H_A* ~6.0 H_A* C2:C_A*

SPECTRUM CCNOESY H1 H2 C2 C1
 0.900 C1:C_A* H1:H_A* ~4.0 H2:H_A* C2:C_A*
 0.800 C1:C_A* H1:H_A* ~4.5 H2:H_A* C2:C_A*
 0.700 C1:C_A* H1:H_A* ~5.0 H2:H_A* C2:C_A*
 0.600 C1:C_A* H1:H_A* ~5.5 H2:H_A* C2:C_A*
 0.500 C1:C_A* H1:H_A* ~6.0 H2:H_A* C2:C_A*

SPECTRUM HCcCH H1 H2 C2 C1
1.0      H1:H_ALI C1:C_ALI C_ALI C2:C_ALI H2:H_ALI
```

If a customized CYANA library file is used in addition to the standard cyana.lib, the library file should be included in the directory as well and read into CYANA prior to the calculation (e.g. by adding the command: *$ read mylib.lib append* to the FLYA.cya macro - see below). This is useful in the instances where a user defines a new experiment type in the *mylib.lib* file. For more details on such customizations, please consult CYANA manual http://www.cyana.org/wiki/index.php/Main_Page and tutorials http://www.cyana.org/wiki/index.php/Tutorials.

S2) Format of the input peak lists for FLYA

An example CCNOESY peak list from 3D (top) and 4D (bottom) methyl-methyl NOESY experiments from which the network of measured peaks is constructed.

a) 3D CCNOESY

```
# Number of dimensions 3
#FORMAT xeasy3D
#INAME 1 C1
#INAME 2 C2
#INAME 3 H1
#SPECTRUM CCNOESY3D C1 C2 H1
1     13.732  13.734  0.904  1     U     9.03E+06      0.00e+00      e     0     -     -     -     0
3     21.887  21.889  1.161  1     U     5.34E+06      0.00e+00      e     0     -     -     -     0
4     21.889  22.433  1.162  1     U     1.22E+05      0.00e+00      e     0     -     -     -     0
6     27.063  27.066  0.917  1     U     3.12E+06      0.00e+00      e     0     -     -     -     0
7     27.064  21.159  0.915  1     U     -7.78E+04     0.00e+00      e     0     -     -     -     0
```

b) 4D methyl-methyl NOESY (cyana.lib: CCNOESY)

```
# Number of dimensions 4
#FORMAT xeasy4D
#INAME 1 H1
#INAME 2 H2
#INAME 3 C2
#INAME 4 C1
#SPECTRUM CCNOESY H1 H2 C2 C1
   872   0.230   0.655  13.602  19.382 1 U   1.000E+02  0.000E+00 e 0 -      -      -      -
   874  -0.834   0.655  13.602  17.697 1 U   1.000E+02  0.000E+00 e 0 -      -      -      -
   883   0.848   0.390  11.341  21.207 1 U   1.000E+02  0.000E+00 e 0 -      -      -      -
   887   0.924   0.390  11.341  22.805 1 U   1.000E+02  0.000E+00 e 0 -      -      -      -
   894   1.376   0.390  11.341  25.567 1 U   1.000E+02  0.000E+00 e 0 -      -      -      -
   901   0.407   0.746  16.119  21.750 1 U   1.000E+02  0.000E+00 e 0 -      -      -      -
```

c) 3D methyl-methyl NOESY (cyana.lib: C13NOESY)

```
# Number of dimensions 3
#FORMAT xeasy3D
#INAME 1 H
#INAME 2 HC
#INAME 3 C
#SPECTRUM C13NOESY H HC C
227    1.381   0.372   11.316  1  U   1115000.0      0.000E+00  a  0  0  0  0
153    1.431   0.824   10.659  1  U   2957000.0      0.000E+00  a  0  0  0  0
156    1.569   0.812   11.844  1  U   999600.0       0.000E+00  a  0  0  0  0
```

a) HCcCH TOCSY (substitute for a 4D short-mixing time NOESY)

```
# Number of dimensions 4
#FORMAT xeasy4D
#INAME 1 H1
```

```
#INAME 2 H2
#INAME 3 C2
#INAME 4 C1
#SPECTRUM HCcCH H1 H2 C2 C1
   1  0.879  0.870  20.219  21.153 1 U   1.000E+00  0.000E+00 e 0 -      -      -      -
   2  0.870  0.879  21.153  20.219 1 U   1.000E+00  0.000E+00 e 0 -      -      -      -
   3  0.902  0.790  21.088  23.701 1 U   1.000E+00  0.000E+00 e 0 -      -      -      -
   4  0.790  0.902  23.701  21.088 1 U   1.000E+00  0.000E+00 e 0 -      -      -      -
```

S3) Preparation of the input structure (PDB) file

FLYA uses the CYANA atom naming convention, as outlined here: http://www.cyana.org/wiki/index.php/Residue_library_file. To prepare the input structure (PDB) file for a FLYA calculation it can be read into CYANA and hydrogens can be added using the following set of commands:

```
$ path/to/cyana/executable cyana
cyana> read pdb name.pdb unknown=warn
cyana> atoms attach "H*"
cyana> write pdb name-H.pdb
```

S4) FLYA.cya

An example of the FLYA execution macro for the methyl resonance assignment calculation is given below (# indicate lines that contain comments and are ignored by the program):

```
peaklists
# define matching tolerances
tolerance:=0.04,0.04,0.4
assigncs_accH:=tolerance(1)
assigncs_accC:=tolerance(3)
# seed for the random number generator
randomseed = 101
# size of population for evolutionary algorithm (default methyls=200)
shiftassign_population:=200
# call flya;  n parallel runs (default methyls=100)
flya runs=100 stages=0 assignpeaks=$assignpeaks structure=mbp.pdb
```

Known, partial methyl assignments can be included in the calculation by specifying their shifts in a separate *fix.prot* file. For more details on specifying the partial assignments, comparing results to the known reference, or other generic FLYA input file requirements, macros, and output files please see: http://www.cyana.org/wiki/index.php/Automated_resonance_assignment_with_FLYA_(Brazil_2018).

S5) peaklists.cya

An example CYANA macro specifying the probability and distance cut-off for methyl-methyl NOESY spectra employed in the FLYA calculation.

```
# CCNOESY
nlist=nlist+1
assignpeaks(nlist):=CCNOESY_II
command CCNOESY_II_expect peaks
  peaks expected $peaks C1="CD1 @ILE" C2="CD1 @ILE" distance=5.0 probability=0.1 append
end

nlist=nlist+1
assignpeaks(nlist):=CCNOESY_IL
command CCNOESY_IL_expect peaks
  peaks expected $peaks C1="CD1 @ILE" C2="CD* @LEU" distance=5.0 probability=0.1 append
end
nlist=nlist+1
assignpeaks(nlist):=CCNOESY_IV
command CCNOESY_IV_expect peaks
  peaks expected $peaks C1="CD1 @ILE" C2="CG* @VAL" distance=5.0 probability=0.1 append
end
```

The macro can be modified to join the peak list entries of certain residue types, as needed when information about the amino-acid type of resonant frequencies is not available. For instance, to join Leu and Val resonance labels to an ambiguous "LV" label, only a simple modification in the macro needs to be made:

```
nlist=nlist+1
assignpeaks(nlist):=CCNOESY_ILV
command CCNOESY_ILV_expect peaks
  peaks expected $peaks C1="CD1 @ILE" C2="CD* @LEU" distance=5.0 probability=0.1 append
  peaks expected $peaks C1="CD1 @ILE" C2="CG* @VAL" distance=5.0 probability=0.1 append
end
```

In this case, measured "Ile-Leu/Val" methyl-methyl NOEs should be listed in the file *CCNOESY_ILV.peaks*.

S6) CONSOL.cya

For each value of the distance cut-off parameter, a separate subdirectory for FLYA calculation should be prepared that should contain all input peak list files (e.g. *CCNOESY_II.peaks, CCNOESY_ILV.peaks*, etc.), appropriately formatted PDB structure file (see *S3*), and appropriately formatted macros (*peaklists.cya, FLYA.cya*). To consolidate the results of calculations at three distance cut-offs (optimum±0.5 Å), the *CONSOL.cya* macro should be run in one of the subdirectories (e.g. the directory of the run at the optimum distance cut-off). An example *CONSOL.cya* macro is given below:

```
# define matching tolerances for assignment validation
tolerance:=0.04,0.04,0.4
# define the extent for strong assignments
shifts_consolidate_extent=100*0.8
# consolidate [1]H, [13]C resonances together
shifts_consolidate_heavy=.true.
# consolidate results from independent runs at different distance cut-offs
consolidate file=consol3g.tab prot=../[^P]*_*_*/details/a[0-9][0-9][0-9].prot
```

Note that the final command ('consolidate') assumes that subdirectories with results for the three calculations are in the same parent directory.

An additional command can be introduced to the *CONSOL.cya* macro to change the extent of self-consistency of methyl assignments, generated by parallel instances of the FLYA calculation, which is used to define strong assignments. Inserting the line *shifts_consolidate_concentration:=0.9* to the macro will increase the requirement by defining the strong assignments as those consistent across 90% of the parallel calculations (the FLYA default is 80%).

To get all ambiguous assignment options for every assigned resonance, an additional "ambiguity" flag can be added to the consolidate command.

```
consolidate reference=ref.prot plot=consol3.pdf file=consol3g.tab prot=../[^P]*_*_*/details/a[0-9][0-9][0-9].prot ambiguity=3
```

This addition specifies the number of ambiguous assignments, in this example 3, to be output to the *consol3g.tab* file. Note that the additional flags 'reference' and 'plot' can also be added if a reference methyl assignment is known, so that the accuracy of FLYA assignment can be assessed in the consolidation step.

S7) consol3g.tab

Upon the consolidation, the FLYA output file *consol3g.tab* is written, which contains the final methyl assignment results. An example *consol3g.tab* file is given below. The first three columns in the file list the atom type, residue type, and residue number for each assigned residue. When a reference assignment is known and specified in the consolidate command (see above), the value of the known reference chemical shift will be listed in the fourth column. The assigned chemical shift (i.e. the FLYA result) is given in the fifth ("Shift") column. When applicable, a deviation from the known assignment is listed in the 'Dev' column in ppm. The 'Extent' column refers to the number of parallel runs over which the consolidation is being performed. Given that a hundred calculations are run at each distance cut-off (optimum±0.5 Å), the consolidation runs over 300 calculations. The next two columns 'inside' and 'inref' respectively specify how frequently, in percentage, the listed chemical shift value is assigned to the atom across the parallel (300) runs and in the reference file. The final column determines if an assignment is 'strong' (i.e. confident). This is based on its self-consistency across multiple parallel calculations, as evident from the 'inside' value. In this example, the extent cut-off is set to 80% and as such, every assignment with the inside value >=80 will be defined as strong. When the reference assignment is known, as in the example below, the '=' is added to the FLYA assignments that match the reference.

```
Total number of shift values: 87600
Cutoff for extent          : 80.00

  Atom  Residue  Ref     Shift    Dev    Extent  inside   inref
  QD1   ILE   2           0.739           300.0    99.7     0.0    strong
  CD1   ILE   2          16.246           300.0    99.6     0.0    strong
  QD1   ILE   5           0.839           300.0    50.9     0.0
  CD1   ILE   5          12.866           300.0    53.2     0.0
  QD1   LEU   6   0.755   0.754   0.001   300.0   100.0   100.0    strong=
```

QD2	LEU	6	0.891	0.892	-0.001	300.0	98.3	100.0	strong=
CD1	LEU	6	23.060	23.050	0.010	300.0	100.0	100.0	strong=
CD2	LEU	6	25.480	25.539	-0.059	300.0	99.4	100.0	strong=
QB	ALA	7		1.031		300.0	67.7	0.0	
CB	ALA	7		22.204		300.0	67.7	0.0	

Supplementary references:

(1) Goddard, T. D.; Kneller, D. G.; University of California: San Francisco, 2001.
(2) Venditti, V.; Fawzi, N. L.; Clore, G. M. *J. Biomol. NMR* **2011**, *51*, 319–328.
(3) Velyvis, A.; Schachman, H. K.; Kay, L. E. *J. Am. Chem. Soc.* **2009**, *131*, 16534-16543.
(4) Shah, D. M.; Ab, E.; Diercks, T.; Hass, M. A. S.; van Nuland, N. A. J.; Siegal, G. *J. Med. Chem.* **2012**, *55*, 10786–10790.
(5) Würz, J. M.; Güntert, P. *J. Biomol. NMR* **2017**, *67*, 63–76.