

ARTICLE

<https://doi.org/10.1038/s41467-019-12837-8>

OPEN

Automatic structure-based NMR methyl resonance assignment in large proteins

Iva Pritišanac ¹, Julia M. Würz¹, T. Reid Alderson ² & Peter Güntert ^{1,3,4*}

Isotopically labeled methyl groups provide NMR probes in large, otherwise deuterated proteins. However, the resonance assignment constitutes a bottleneck for broader applicability of methyl-based NMR. Here, we present the automated MethylFLYA method for the assignment of methyl groups that is based on methyl-methyl nuclear Overhauser effect spectroscopy (NOESY) peak lists. MethylFLYA is applied to five proteins (28–358 kDa) comprising a total of 708 isotope-labeled methyl groups, of which 612 contribute NOESY cross peaks. MethylFLYA confidently assigns 488 methyl groups, i.e. 80% of those with NOESY data. Of these, 459 agree with the reference, 6 were different, and 23 were without reference assignment. MethylFLYA assigns significantly more methyl groups than alternative algorithms, has an average error rate of 1%, modest runtimes of 0.4–1.2 h, and can handle arbitrary isotope labeling patterns and data from other types of NMR spectra.

¹Institute of Biophysical Chemistry, Center for Biomolecular Magnetic Resonance, Goethe University Frankfurt am Main, Max-von-Laue-Str. 9, 60438 Frankfurt am Main, Germany. ²Laboratory of Chemical Physics, NIDDK, National Institutes of Health, Bethesda, MD 20892-0520, USA. ³Laboratory of Physical Chemistry, ETH Zürich, Vladimir-Prelog-Weg 2, 8093 Zürich, Switzerland. ⁴Department of Chemistry, Tokyo Metropolitan University, 1-1 Minami-Osawa, Hachioji, Tokyo 192-0397, Japan. *email: guentert@em.uni-frankfurt.de

The last decade of structural biology has seen growing interest in biologically relevant large protein assemblies, as witnessed by an explosion of high- and low-resolution structural studies of macromolecular machines¹. NMR spectroscopy is the principal experimental method for the simultaneous analysis of both the structures and dynamics of biomolecules at atomic resolution. The traditional size-limit of solution-state NMR spectroscopy, typically placed below 30 kDa, was overcome by Transverse Relaxation-Optimized Spectroscopy (TROSY)². The TROSY enhancement, initially established for amide groups, was subsequently also realized for selectively methyl-labeled proteins (methyl-TROSY)^{3,4}. Methyl-TROSY has since enabled studies of protein complexes in excess of 1 MDa⁵ in unprecedented detail, revealing the mechanisms of dynamic molecular machines^{6–8}.

For optimal gains in the signal enhancement and resolution of methyl-TROSY spectra, selectively protonated, ¹³C-labelled methyl groups are introduced into an otherwise highly deuterated background⁹. To this end, cost-effective and robust biosynthetic strategies have been established for the selective or simultaneous labelling of all methyl-containing amino acids in *Escherichia coli*^{10,11}. Selective labeling of methyl groups is also possible in eukaryotic expression systems^{12–14}. The labeled methyl groups have favorable spectroscopic properties that render them observable also in large proteins and protein assemblies. Methyl groups are effective site-specific probes of molecular dynamics, structure, and interactions, as they are found both throughout the hydrophobic core of a protein and on its surface^{10,15}.

The major bottleneck for NMR studies with selective methyl-labeled proteins is the resonance assignment, i.e. relating ¹H/¹³C signals in the NMR spectra to specific methyl groups in the protein (Fig. 1)¹⁶. In small and medium-size proteins, NMR signals from the protein backbone can be observed and used in triple-resonance, “through-bond” experiments for the sequence-specific resonance assignment of the backbone¹⁷, to which side-chain methyl resonances can be linked¹⁸. In contrast, for large proteins, backbone resonances and triple-resonance spectra cannot be observed, and, unless the protein is modified, only nuclear Overhauser effects (NOEs) between methyl groups remain accessible as NMR input data for assignment.

Assignment strategies for large proteins or proteins assemblies include divide-and-conquer approaches wherein sufficiently small individual protein domains or subunits are produced separately, such that their backbone resonance assignment can be determined using standard methods¹⁹. This approach requires that the resonance frequencies of the subsystems coincide closely with those of the complete system. To complete the assignment, the approach is often supplemented with site-directed mutagenesis of individual methyl-bearing residues^{20,21}. As an alternative, a high-resolution structure of the studied protein or complex can be utilized in combination with NMR experiments that reveal spatial proximity between methyl groups^{22,23}, or between methyls and site-specifically attached paramagnetic probes²⁴.

The laborious and time-consuming nature of these assignment strategies prompted automation efforts. Presently, two groups of structure-based, automatic assignment approaches are available: NOE spectroscopy (NOESY) and paramagnetism-based methods. Both rely on NMR-derived, sparse distance measurements that are compared to a known three-dimensional (3D) structure. Paramagnetic approaches require the site-specific introduction of paramagnetic probes and estimates of the magnetic susceptibility tensors. These approaches define the optimal methyl assignments as those that minimize the difference between the measured and the calculated paramagnetic observables^{25–27}. For instance, PRE-ASSIGN²⁷ uses paramagnetic relaxation enhancements (PREs),

whereas PARAssign²⁶ relies on pseudo-contact shifts (PCs). NOESY-based automatic approaches match a network of measured methyl–methyl NOE contacts to the network of short inter-methyl distances predicted from the protein structure, using Monte Carlo^{28–31} or graph-based^{32,33} algorithms. For example, MAGMA³² uses exact graph matching algorithms to generate confident assignments for a subset of well inter-connected methyls. For the remaining methyls, MAGMA reports all ambiguous assignment possibilities, which may be used for further experimental investigation.

Automated methods for structure-based methyl resonance assignments can be characterized by the experimental requirements for measuring the input data, and by the completeness and accuracy of the assignments that they produce. An optimal algorithm functions with data that can be measured readily, tolerates experimental imperfections, is computationally efficient, and yields confident assignments for a large fraction of all methyls. To minimize the amount of error or subsequent manual checking, the algorithm (not the user) should distinguish confident assignments, which are almost certainly correct, from other, tentative or ambiguous ones. Existing algorithms fall short of this ideal in different ways.

Therefore, we here adopt the FuLIY Automated assignment algorithm FLYA³⁴, (Fig. 1) which is integrated in the CYANA structure calculation software³⁵ and has been shown capable to assign proteins exclusively from NOESY data³⁶, for structure- and NOESY-based methyl resonance assignment. We apply the resulting MethylFLYA algorithm to a benchmark³² of five large proteins and protein complexes and show that, on the basis of methyl–methyl NOEs alone, MethylFLYA can assign significantly more methyl resonances with high accuracy than the previously introduced methods MAGMA³², MAP-XSII²⁹, FLAMEnG02.0³¹, and MAGIC³³ operating on the same input data. To demonstrate its robustness with respect to ambiguous and imperfect experimental information, we apply MethylFLYA also to unrefined peak lists, reduced input data sets, and peak lists obtained by automated peak picking with the CYPICK algorithm³⁷.

Results

Benchmark data. MethylFLYA was applied to the five largest proteins of a benchmark data set that was originally prepared for evaluating the MAGMA algorithm for automated methyl assignment, as described in the original publication³². In addition, methyl NMR data for the 20 kDa N-terminal domain of Heat Shock Protein 90 (called HSP90 in this paper)³⁸, which has also been used previously with MAGMA, were used for evaluating MethylFLYA in combination with automated peak picking with CYPICK³⁷. The main benchmark data set comprised five proteins of varying molecular mass and shape for which NOESY data from specifically methyl-labeled samples, assignments, and 3D structures are available (Table 1):³² the N-terminal domain of *E. coli* Enzyme I (called EIN in this paper; molecular mass 28 kDa)³⁹, a dimer of regulatory chains of aspartate transcarbamoylase from *E. coli* (ATCase; 34 kDa)²⁴, maltose binding protein (MBP; 41 kDa)⁴⁰, malate synthase G (MSG; 81 kDa)^{15,18}, and the “half-proteasome” 20S core particle, a 14-mer ($\alpha_7\alpha_7$; 358 kDa)⁴¹.

The following experimental data were taken from the MAGMA benchmark³²: (i) Assigned [¹H,¹³C]-HMQC peak lists providing reference assignments, which were used only to evaluate the accuracy of the MethylFLYA results, while unassigned versions of these [¹H,¹³C]-HMQC peak lists were supplied to MethylFLYA. (ii) Filtered and unfiltered (see below) NOESY peak lists from 3D (ATCase, $\alpha_7\alpha_7$) or 4D (EIN, MBP, MSG) methyl–methyl NOESY spectra. (iii) Solution or crystal structures of the proteins, taken

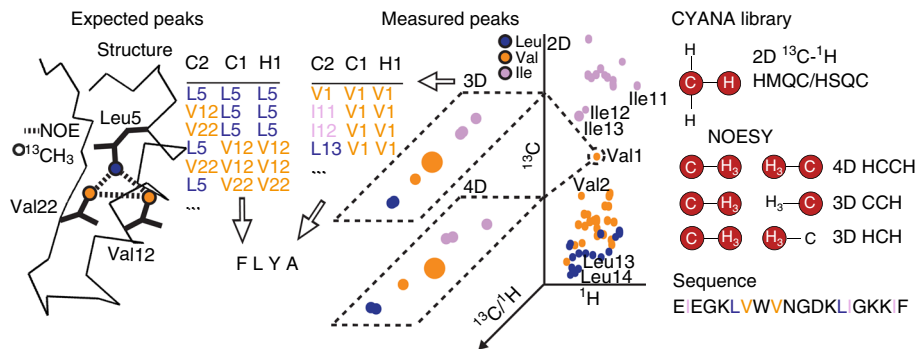


Fig. 1 Automatic methyl resonance assignment with MethylFLYA. The expected methyl-methyl contacts are computed from a 3D structure of the protein (left). The contacts are written to a list of expected NOE peaks with the amino acid type of each contact indicated (e.g. I-V, L-L). A list of measured NOESY peaks is obtained by manual (or automatic) inspection of the 3D or 4D methyl-methyl NOESY spectrum (center) guided by information from the 2D $[^1\text{H},^{13}\text{C}]$ -HMQC spectrum. If known, the amino acid types of the methyl peaks that give rise to measured NOEs are included in the peak list. The $[^1\text{H},^{13}\text{C}]$ -HMQC peak list and the expected and measured NOE peak lists are supplied to MethylFLYA for the automatic methyl assignment calculation. In addition to the peak lists, the MethylFLYA calculation requires the protein sequence and knowledge of the magnetization transfer pathways for the employed NMR experiments, which are provided in the CYANA library (right)

Table 1 Methyl resonance assignment statistics

	EIN	ATCase	MBP	MSG	$\alpha_7\alpha_7$
Protein properties:					
Residues per monomer	259	153	370	731	233
Multimeric state	monomer	dimer	monomer	monomer	14-mer
Molecular mass of multimer (kDa)	28.3	34.2	40.6	81.4	358.4
Experimental NMR data:					
Labeled amino acids	AILV	ILV	ILV	ILV	ILV
NOESY dimensions	HCCH	CCH	HCCH	HCCH	CCH
Labeled methyl ^1H - ^{13}C groups:					
All	146	66	123	276	97
With reference assignment	133	62	123	268	93
With NOESY peaks	116	54	118	236	88
Methyl resonances confidently assigned by MethylFLYA:					
All ^a	101	45	86	176	80
Correct	90	35	83	173	78
Erroneous	0	4	2	0	0
Assigned by MethylFLYA using unfiltered peak lists:					
All ^a	107	42	89	183	82
Correct	98	35	87	174	80
Erroneous	0	3	2	3	0
Assigned by MethylFLYA using a single distance cutoff:					
All ^a	118	44	96	194	84
Correct	101	35	85	176	80
Erroneous	4	3	8	8	1
Methyl resonances assigned by MAGMA: ³²					
Correct	56	18	78	97	84
Erroneous	0	0	0	2	0
Methyl resonances assigned by MAP-XSII: ²⁹					
Correct	64	24	16	33	79
Erroneous	17	5	11	9	1
Methyl resonances assigned by FLAMEnGO2.0: ³¹					
Correct	0	8	35	0	70
Erroneous	0	4	0	0	18
Methyl resonances assigned by MAGIC: ³³					
Correct	71	39	-	-	-
Erroneous	13	6	-	-	-

Filtered input NOESY peak lists were used, unless noted otherwise. See text for details

^aAll strong assignments, including correct, erroneous, and additional ones for methyl groups without reference assignment

from the Protein Data Bank with accession codes [1EZA](#) for EIN, [1D09](#) for ATCase, [1EZ9](#) for MBP, [1D8C](#) for MSG, and [1YAU](#) for $\alpha_7\alpha_7$. In addition, MethylFLYA calculations were performed for the alternative structural forms [1TUG](#) for ATCase, [3MBP](#) for MBP, and [1Y8B](#) for MSG. Automated peak picking with CYPICK

was performed for NOESY spectra in Sparky⁴² format for EIN, ATCase, and HSP90. Information about Leu/Val geminal methyl pairs, which was available in the MAGMA benchmark³², was incorporated into the MethylFLYA calculations in the form of simulated HCcCH TOCSY peak lists.

Two sets of experimental methyl–methyl NOESY peak lists were used for the five proteins. The first set (“filtered peak lists”) comprised peak lists from the MAGMA study that were filtered for reciprocity of donor and acceptor NOE cross peaks (only the reciprocated peaks were kept), and signal-to-noise ratios (only the peaks with $S/N \geq 2$ were kept)³². The second set comprised unfiltered (“raw”) peak lists, generated by manual analysis of NOESY spectra using Sparky⁴² software, which were not manually modified before the assignment calculation.

MethylFLYA parameter optimization. While most parameters of the MethylFLYA algorithm could be kept at the values that had been found optimal in earlier applications of the original FLYA algorithm^{34,36,43–46}, specific optimization of a small number of parameters that are of relevance to structure-based methyl assignment was advantageous.

MethylFLYA considers only methyl–methyl distances shorter than a user-defined cutoff d_{cut} for generating expected methyl–methyl NOESY cross peaks based on a protein structure (see Methods). In addition, each expected peak is attributed a probability value to (roughly) reflect the probability of observing it in the corresponding measured spectrum. For expected NOESY cross peaks, we tested a range of distance cutoffs and distance-dependent observation probabilities (Supplementary Fig. 1). Across these parameter values, we monitored the fraction of correct and incorrect strong (i.e. confident) methyl assignments and the percentage of explained input NMR data (methyl–methyl NOEs). Even though protein-specific profiles can be observed in Supplementary Fig. 1, the fractions of assigned methyl resonances generally plateaued around $d_{\text{cut}} = 5 \text{ \AA}$ for EIN, ATCase, MBP, and MSG, or $d_{\text{cut}} = 6 \text{ \AA}$ for $\alpha_7\alpha_7$ (Supplementary Fig. 1). These plateaus coincided with about 80% explained input NMR data, which was determined as optimal for these data sets. Increasing the observation probabilities generally diminished the quality of the results, as more incorrect assignments were obtained (Supplementary Fig. 2). Predictably, more of the observed NOEs were assigned using higher distance cutoffs for generating the expected NOEs, but assignment errors also increased. In most cases, the assignment accuracy peaked around the plateaus of assigned methyl fractions and decreased at higher ($\geq 7 \text{ \AA}$) and lower ($\leq 4 \text{ \AA}$) distance cutoffs. To reduce the dependency on small variations of the distance cutoff, we always performed three assignment calculations using the given d_{cut} as well as a slightly lower and a slightly higher value, i.e. $d_{\text{cut}} - 0.5 \text{ \AA}$, d_{cut} , and $d_{\text{cut}} + 0.5 \text{ \AA}$, and we required assignments to be self-consistent over the three runs (see Methods). As such, based on Supplementary Figs. 1 and 2, we used d_{cut} values of 4.5, 5.0, and 5.5 \AA for EIN, ATCase, MBP, and MSG, and 5.5, 6.0, and 6.5 \AA for $\alpha_7\alpha_7$, as well as a NOESY cross peak observation probability of 0.1 for all following MethylFLYA calculations.

The number of independent assignment optimization runs that is necessary for obtaining reproducible, virtually seed-independent strong assignments was also optimized (Supplementary Fig. 3). All further MethylFLYA runs comprised 100 independent assignment optimization runs.

Assignment completeness and accuracy. We evaluated the performance of MethylFLYA on manually (expert) picked methyl NOE signals that were either (i) filtered to keep only the NOESY cross peaks that are observed reciprocally between two methyl groups and that are above a defined signal-to-noise threshold³² (“filtered” peak lists); or (ii) used without any subsequent editing or filtering (“unfiltered”/“raw” peak lists). Using manually analyzed and filtered NOE data (i)³², MethylFLYA assigned between 63% (ATCase) and 84% ($\alpha_7\alpha_7$) of the methyl resonances for

which reference assignments are available (Fig. 2b; Table 1, Supplementary Table 1), with no assignment errors for EIN, MSG, and $\alpha_7\alpha_7$. Two incorrect methyl assignments were found for MBP, and four for ATCase (Fig. 2b). In the 3D structures, all incorrectly assigned methyls are located in proximity to their correct assignment positions (Supplementary Fig. 4, Supplementary Table 2). Such spatially localized assignment errors are expected to have minor impact on studies that do not require very high-resolution information, for instance, when identifying an interaction interface.

We also note that more stringent criteria can be applied to define the confident (strong) methyl assignments, which further reduce errors. For instance, increasing the requirement for self-consistency of assignments from multiple parallel runs of the algorithm from 80% to 90% (see Methods), results in a decrease in error for ATCase from 6% to 1%. This is achieved at the expense of reducing the percentage of strong assignments on average by 6%. It is thus possible to ensure a higher accuracy by “sacrificing” some of the strong assignments.

On the other hand, using a single distance cutoff (5 \AA for EIN, ATCase, MBP, MSG; 6 \AA for $\alpha_7\alpha_7$) instead of three cutoffs spaced by 0.5 \AA for generating the expected NOESY cross peaks in MethylFLYA increases the overall number of assignment errors about four-fold (Table 1). It is thus not advisable even though the total number of strong assignments rises by about 10%.

Importantly, MethylFLYA is robust with respect to the presence of ambiguous or incorrect methyl–methyl NOEs, as judged by its comparable, or in some cases even better, performance on “raw” NOESY peak lists that were not filtered for NOE cross peak reciprocities and signal-to-noise ratios and retained any ambiguous and tentative NOE cross peaks (Fig. 2).

A spatial clustering of strong assignments can be discerned in the structures of EIN, ATCase, and MSG (Fig. 2c). This is likely due to the low number of long-range NOEs between the clusters. In addition to the strong assignments, MethylFLYA outputs ambiguous assignment options for all resonances to which at least one inter-methyl NOE is attributed. The number of ambiguous assignment possibilities to be displayed can be specified by the user.

Reduced data sets. We tested the performance of MethylFLYA on the benchmark when experimental information provided to the algorithm was reduced (Fig. 3). In the best-case scenario, both knowledge of the amino acid types of methyl resonances and linkage of the two geminal methyl groups of Leu and Val is available (Fig. 3a, c, black). The Ile- δ_1 resonances are usually readily identified due to their upfield shifted ^{13}C frequencies. To discriminate between Leu and Val resonances, separate protein samples can be prepared using selective labelling schemes. For instance, selective Leu labelling can be achieved by using ^{13}C -labeled α -ketoisocaproate⁴⁷, whereas the combined addition of unlabeled α -ketoisocaproate and labeled α -ketoisovalerate leads to exclusive labeling of Val⁴⁸. To connect resonances from the two geminal Leu/Val methyl groups, an additional protein sample can be prepared in which both Leu/Val-methyl groups are protonated and ^{13}C -labelled. A short-mixing time NOESY experiment can then be used to record cross peaks between geminal methyl groups^{21,32} (Fig. 3a). Without discrimination between Leu and Val resonances, MethylFLYA performed very similarly as in the best-case scenario for EIN, MSG, and $\alpha_7\alpha_7$, confidently assigning 68, 62, and 79% of the methyl resonances, respectively, with complete accuracy (Fig. 3c, dark gray). For ATCase and MBP, the percentage of accurate confident assignments decreased by 3%. However, for ATCase the percentage of errors was also reduced simultaneously by 3%.

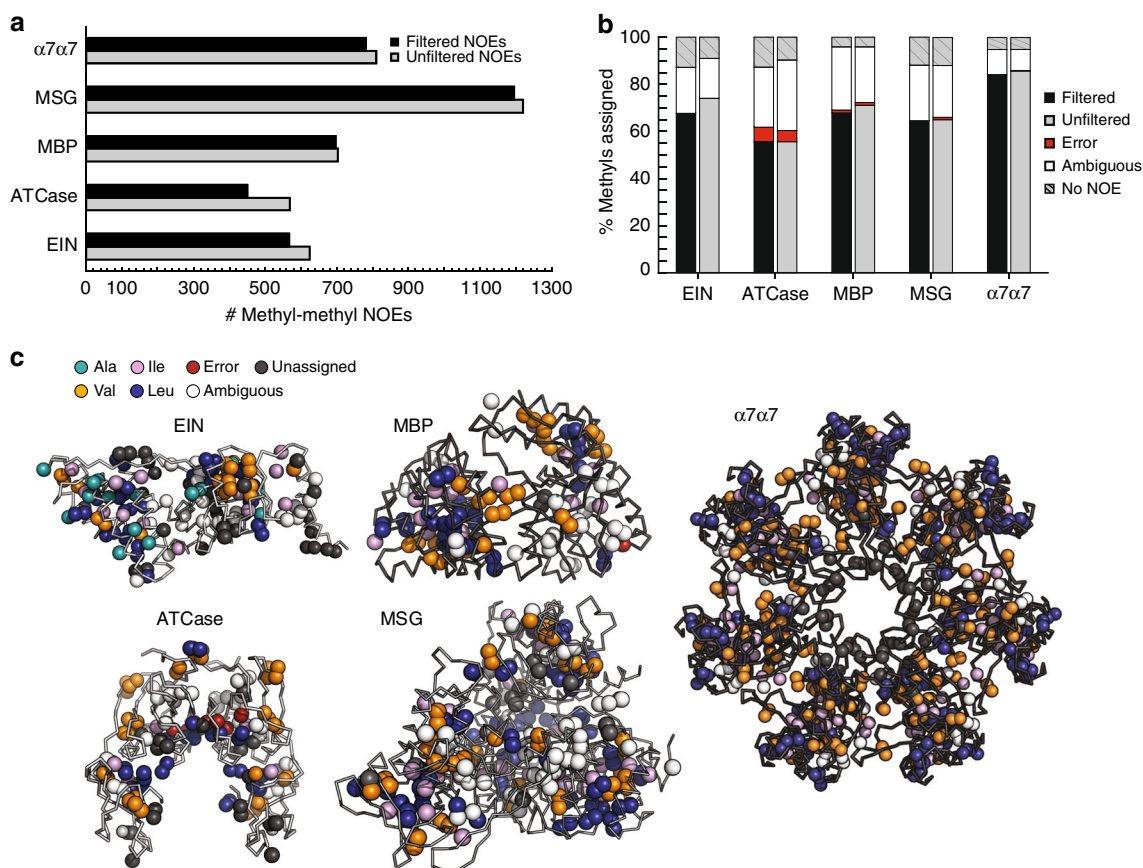


Fig. 2 MethylFLYA performance on the benchmark. **a** Number of methyl-methyl NOEs before and after filtering of manually picked NOESY peak lists (see Results, Benchmark data). **b** MethylFLYA performance on filtered (black) and “raw” (unfiltered) manually picked NOESY peak lists (gray). **c** Methyl groups assigned as strong (confident) by MethylFLYA with filtered NOESY peak lists are indicated with colored spheres in the 3D structures of the benchmark proteins. The colors indicate the amino-acid types of the assigned groups, with non-assigned groups colored white

Removing the geminal Leu/Val pairing had a more significant impact, reducing the percentage of assigned methyls by ~19% for EIN, ATCase, MBP, and MSG, and up to 30% for $\alpha_7\alpha_7$ (Fig. 3c, light gray). The overall accuracy, however, remained high. The critical importance of this restraint for automatic methyl assignment was reported previously in the MAGMA study³². In the MAGIC study, a four-fold decrease in computational time and a somewhat improved assignment accuracy were noted as benefits of the restraint³³. As an alternative, the information about Leu/Val geminal pairs can be substituted with stereospecific labelling schemes that restrict isotopic labeling to only pro-*R* or pro-*S* methyl groups, and thus reduce the number of methyl resonances in the $[^1\text{H}, ^{13}\text{C}]$ -HMQC spectrum⁴⁹. For MethylFLYA, removing both the Leu/Val-geminal pairing and discrimination between Leu/Val methyl resonances led to a similar outcome as geminal pairing removal alone (Fig. 3c, silver), and led overall only to a slight further increase in erroneous assignments (1–2%). Interestingly, for ATCase, removing the Leu/Val resonance discrimination always improved the accuracy (Fig. 3c, dark gray, silver). We conclude that, especially for smaller proteins (<80 kDa), Leu/Val residue discrimination is not crucial for MethylFLYA.

The computation time of MethylFLYA scaled approximately linearly with the number of methyl groups in the protein. The complete protocol took between 0.36 and 1.53 h (Supplementary Table 3). Negligible differences in speed were noted for the calculations with lower input information content (Fig. 3, Supplementary Table 3). This illustrates the ability of

MethylFLYA to efficiently deliver high-quality assignments even from considerably reduced input data.

Combination with automated peak picking. All currently available automatic methyl resonance assignment strategies rely, to different extents, on a manual analysis and interpretation of the NMR data. The NOE-based methods, for instance, require manual, expert, inspection of methyl–methyl NOESY spectra to generate peak lists as input to the assignment software^{28–33}. We investigated whether an automatic peak picking algorithm, CYPICK³⁷, could be used in combination with MethylFLYA to fully automate methyl resonance assignment⁵⁰. We tested the CYPICK-MethylFLYA combination on three proteins from the MAGMA study for which methyl–methyl NOESY spectra were available (Fig. 4). For these spectra, CYPICK found 77–83% of the manually identified methyl–methyl NOEs (Supplementary Fig. 5, Supplementary Table 4), which is comparable to its performance previously reported on 3D ^{13}C -edited and ^{15}N -edited NOESY spectra³⁷. The somewhat high CYPICK artifact scores for EIN (34%) and HSP90 (46%) did not result in assignment errors, as only one methyl group misassignment was found for EIN and three for HSP90. Moreover, for EIN, even slightly more methyls were confidently and accurately assigned when the automatically generated CYPICK peak lists (78%) were used compared to the manually prepared lists (68%).

Despite the relatively large number of assignments for EIN, similar success was not found for the HSP90 and ATCase CYPICK datasets. In the case of HSP90, the considerably smaller

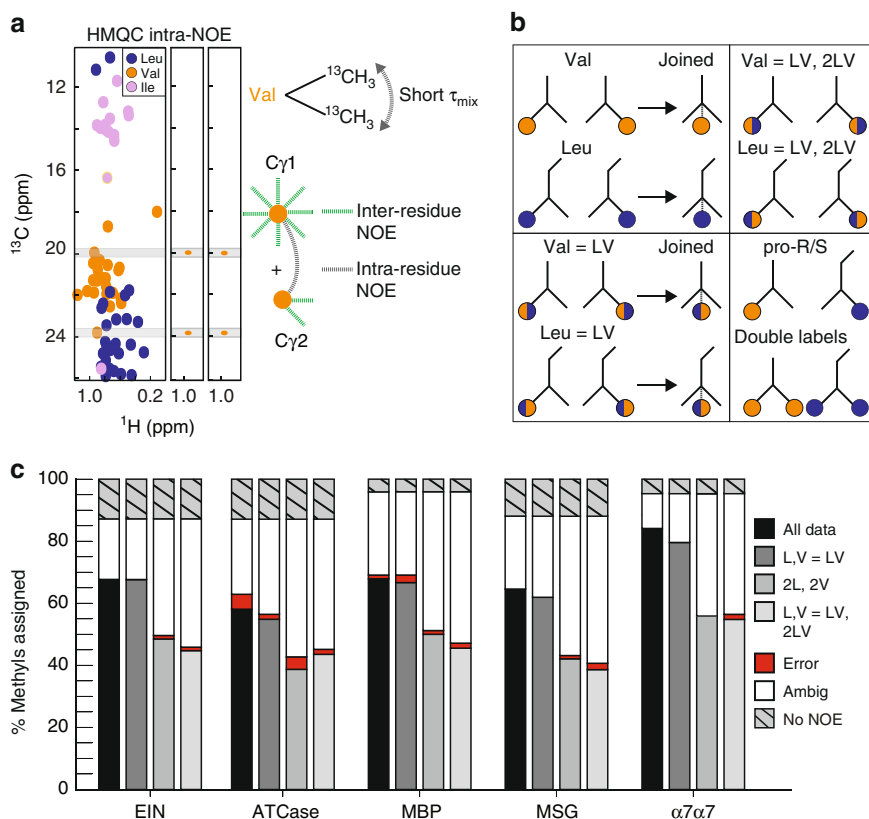


Fig. 3 MethylFLYA performance with reduced data input. **a** Geminal methyl groups of Leu/Val residues can be linked with a short-mixing time NOESY experiment on an exclusively double methyl-labeled ($^{13}\text{C}_{\delta_1}/^{13}\text{C}_{\delta_2}$ -Leu, $^{13}\text{C}_{\gamma_1}/^{13}\text{C}_{\gamma_2}$ -Val) protein sample. In the NOESY plane of each Leu/Val methyl resonance, a strong signal from its geminal methyl resonance is observed (right). **b** Different possibilities for treating Leu/Val methyl resonances in MethylFLYA. Top, left: differentiation of the Leu- and Val-type of methyl resonances and known connectivity between geminal Leu or Val methyl groups. Bottom, left: no differentiation between methyl resonances of the Leu- or Val-type, but known connectivity between geminal methyl groups. Top, right: no differentiation between methyl resonances of the Leu/Val-type, nor knowledge of the geminal methyl connectivity. Bottom, right: stereospecific labelling of Leu/Val-methyl groups (pro-R or pro-S), or double labelling (both pro-R and pro-S). **c** MethylFLYA results with reduced data input. The percentage of assigned methyl groups given knowledge of all methyl amino acid types and the geminal Leu/Val-methyl connectivity is shown in black (all data). In dark gray (L, V = LV), knowledge of Ile- and Ala-methyl resonance types is assumed, but there is no discrimination between Leu or Val methyl resonance types. The geminal (Leu/Val) methyl resonances are connected. In light gray (2L, 2V), knowledge of all amino-acid types is assumed, but the geminal pairing of Leu/Val resonances is omitted. In silver (L, V = LV, 2LV), neither the amino-acid type nor the geminal pairing is known for Leu/Val methyl resonances. Knowledge of the amino acid types of other resonances (Ala, Ile) is still assumed

amount of assigned methyls could be attributed to the lower percentage of explained NOE data when using the CYPICK lists (Fig. 4b). When the manually generated NOE list was used, the MethylFLYA assignments explained roughly 80% of the NOE data at a 5 Å distance cutoff (Supplementary Fig. 5), consistent with the results presented above for the five proteins of the benchmark. In contrast, at the same distance cutoff, only about 60% of the NOE data were explained for the CYPICK-derived list. For ATCase, less than 40% of the methyl groups could be assigned, except for a single d_{cut} value (Supplementary Fig. 5). The considerably worse performance of CYPICK-MethylFLYA on ATCase and HSP90 suggests that some methyl–methyl NOEs are more critical determinants of assignment success than the others. Overall, manual peak picking of the NOESY spectra (or manual screening and adaptation of automatically prepared peak lists) remains the best approach for preparing the input data for MethylFLYA.

MethylFLYA using minimal input information. All automatic methyl resonance assignment protocols that are presently available assume that the resonance frequency positions of ^1H - ^{13}C correlations from the 2D [^1H - ^{13}C]-HMQC spectrum are known

and that each methyl resonance is associated unambiguously or, in some cases, ambiguously with a methyl residue type specified by the user. The benchmark data set of proteins from the MAGMA study³², which was reused here, provides ^1H - ^{13}C resonance frequency positions based on the known reference assignment. However, the knowledge of these exact positions offers an additional source of information to the automatic assignment protocols, as it immediately resolves some inherent uncertainties, e.g., peak duplications or overlaps, and subsequently aids during the algorithmic attribution of NOEs to specific methyl-bearing residues.

Therefore, we sought to address the performance of MethylFLYA starting solely from 2D HMQC and 4D methyl–methyl NOESY spectra, whilst being ‘blind’ to the known ^1H - ^{13}C resonance frequencies and methyl residue types. To this end, we performed both manual and automated peak picking of both the 2D [^1H , ^{13}C]-HMQC and 4D methyl–methyl NOESY spectra of EIN (see Methods, Supplementary Table 5, Supplementary Figs. 6, 7). The methyl residue types (Ala, Ile, and ambiguous Leu/Val), were assigned based on the BMRB chemical shift statistics⁵¹ and the known number of expected peaks of different residue types (see Methods, Supplementary Fig. 7). The high degree of

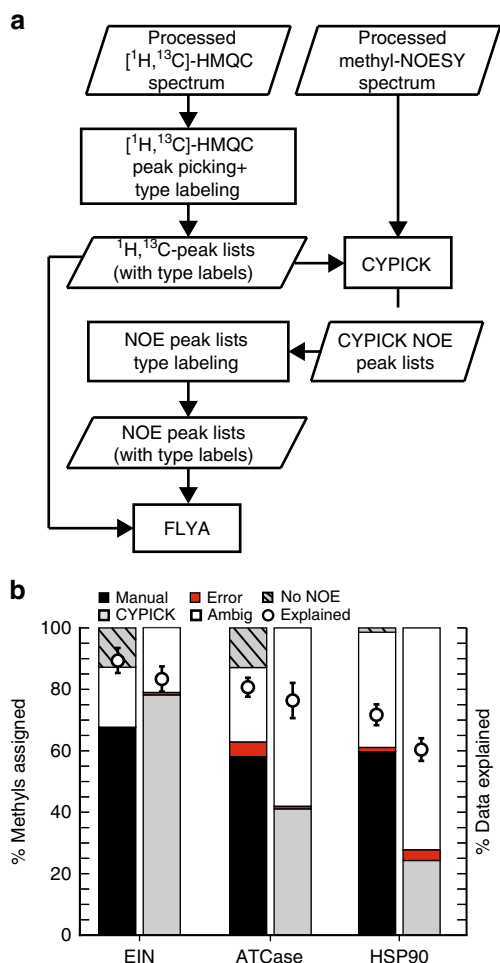


Fig. 4 MethylFLYA performance with automatic peak picking using CYPICK. **a** An outline of the combined CYPICK-MethylFLYA assignment protocol (see Methods). “Type labeling” refers to the attribution of methyl resonances to amino-acid types (e.g. Ala, Ile, Leu, Val). **b** Comparison of MethylFLYA performance on manually and automatically picked NOESY data. The percentage of assigned methyl resonances and the percentage of explained input NOESY data, given the MethylFLYA assignment, are shown. The error-bars are standard deviations of the percentage of data explained over the three distance cutoffs (optimal cutoff ± 0.5 Å, see Methods). Note that CYPICK does not assign methyl–methyl NOEs, and hence the identity of the methyl resonances with no NOEs (“no NOE”) is not indicated for the CYPICK bars

overlap between the average methyl chemical shifts of Leu and Val rendered them difficult to separate, and therefore MethylFLYA treated them as ambiguous. The methyl peaks falling in overlapping regions for other residue types, i.e. Ala and Leu/Val, were either assigned a type based on the “best guess” (see Methods, Supplementary Fig. 7), or assigned ambiguously to the three possible types (Ala or ambiguous Leu/Val). As anticipated, when provided with the minimum amount of information, the percentage of strong (confident) assignments dropped significantly, from 74% attained when using the unfiltered methyl–methyl NOESY peak list, the known ^1H - ^{13}C frequencies with known residue types, and the geminal Leu/Val resonance pairing (Fig. 2), to 30% or 24% when using only the manually or CYPICK-analyzed ^1H , ^{13}C -HMQC and 4D methyl–methyl NOESY spectra, respectively (Supplementary Fig. 7D). Nonetheless, the accuracy of the strong assignments remained high with only three and four methyls assigned incorrectly for

manually- and CYPICK-acquired peak lists, respectively (Supplementary Fig. 7E). The robustness of MethylFLYA to the highly ambiguous and partially incorrect input information is notable, especially when considering that the BMRB-derived assignment of residue types led to seven resonances with incorrectly assigned methyl residue type labels (Supplementary Fig. 7A, B). In fact, the assignment reflected this erroneous input information, as Ala12 and Ala160, that were attributed the wrong residue type, were assigned incorrectly by both approaches (Supplementary Fig. 7A, E). We here note that the user can use structure-based methyl resonance predictors, such as SHIFTX2⁵², to flag the methyls that are expected to deviate significantly from BMRB chemical shift statistics and therefore likely to acquire an incorrect residue type label and subsequent misassignment (Supplementary Fig. 8). In addition, the user can exclude from consideration any strong assignments to the methyl resonances with a highly ambiguous residue type (i.e., Ala or ambiguous Leu/Val). If structure-based methyl resonance prediction is applied to exclude from consideration any strong assignments of the methyls that are expected to have a misassigned residue type (Supplementary Fig. 8), the assignment errors are reduced to only one or two methyls using manually or CYPICK-acquired peak lists, respectively (Supplementary Fig. 7E (i), (iii)). Attributing the methyl peaks in the overlapped regions of Ala/Val methyl residue types to both Ala and Leu/Val resulted in a somewhat higher percentage of strong assignments, but a further increase in errors (Supplementary Fig. 7D), and is therefore not recommended with the present implementation of MethylFLYA.

We next considered the performance of MethylFLYA when additional information in the form of the geminal Leu/Val methyl pairing is provided, albeit only for the well-resolved Leu/Val resonances in the spectrum (Supplementary Table 5, Supplementary Fig. 7C), as an unambiguous pairing of the geminal pairs is expected to be challenging or impossible for overlapping peaks. This assumes the preparation of an additional protein sample with both Leu/Val methyl groups labeled simultaneously. Next to providing an additional restraint for every pair of Leu/Val methyl resonances, such a sample would additionally distinguish unambiguously between Ala and Leu/Val types based on the 2D ^1H , ^{13}C -HMQC spectrum. Accordingly, in these calculations, the methyl residue type annotation was corrected for Ala and Leu/Val types. This resulted in the geminal pairing of 62 out of the 92 Leu/Val resonances expected based on the protein sequence. Introducing the additional information more than doubled the fraction of strong methyl assignments from 30% to 64% or 24% to 59% for manually or CYPICK-acquired peak lists, respectively. The errors mostly mapped to positions that are spatially proximal to the correct assignment (Supplementary Fig. 7E (ii), (iv)). We noted that the error in assignment of Val176 γ_2 to Val156 γ_1 occurred both when using manually- and CYPICK-picked peak lists, being the sole error in the latter case. The reference assignment showed that the methyl resonance of Val176 γ_2 is overlapped with that of Leu123 δ_1 in the 2D ^1H , ^{13}C -HMQC spectrum. As such, neither of the resonances were paired with their geminal methyls in the calculations. Resonance overlap is expected to prevent an unambiguous assignment of methyl–methyl NOEs using the automatic methyl NOESY assignment protocol of MethylFLYA, which, combined with the lack of any additional restraints for the resonances, such as the geminal methyl pairing, likely underlies the assignment error. This example illustrates how overlapping peaks can constitute a challenge for automatic methyl resonance assignment, which represents an important aspect for future improvement. Overall, the results provide a fair estimate of the lower bound of the performance of MethylFLYA given minimal data input and maximal data uncertainty, and demonstrate how additional

restraints introduced into the assignment search can significantly improve its outcome (Supplementary Table 5, Supplementary Fig. 7).

Performance comparison. The MAGMA study³² included a performance comparison with the available NOE-based automatic methyl assignment software packages, MAP-XSII²⁹, and FLAMEnGO2.0³¹. For a comparison of the available methods, we used here the results for all proteins³², apart from MSG, for which a different structure of the protein was used (Fig. 5, Supplementary Table 1, Supplementary Fig. 9). The recently introduced MAGIC³³ method requires the knowledge of signal intensities for all methyl–methyl NOE cross peaks, information that was not available for three out of the five proteins of our benchmark set: methyl–methyl NOESY spectra were available for EIN and ATCase, and in addition for HSP90. The performance of the MAGIC method on these datasets is summarized in Supplementary Table 6 and Supplementary Fig. 10.

Discussion

Compared to the alternatives, MethylFLYA generated more confident and correct methyl assignments in all cases except for $\alpha_7\alpha_7$ (Table 1, Fig. 5), where all methods assigned more than 85% of the methyls. For the other proteins, MethylFLYA generated on average 18% more assignments than the next best performing software. Overall, MethylFLYA generated the highest number of confident and correct methyl ¹H and ¹³C resonance assignments on this benchmark (confident and correct/total = 459/465), followed by MAGMA (333/335), MAP-XSII (216/259), and FLAMEnGO2.0 (113/135). Across the entire benchmark, MethylFLYA made assignment errors for six methyls. Based on the error rate on this benchmark, MethylFLYA is the second most accurate method after MAGMA, which made assignment errors only for two methyls. The latter two errors result from the use of a crystal structure for MSG (PDB 1D8C) instead of the NMR-derived structure (PDB 1Y8B) that had been used in the original MAGMA benchmark³². In the original study, MAGMA was reportedly sensitive to the structural difference between the two forms, which is likely due to the presence of the ligand in the crystal structure³². Here, we tested the performance of all methods exclusively on crystal structures to omit the need for NMR structures, which are anticipated to be unavailable for most proteins for which methyl resonance assignment is sought.

For the subset of the data for which a comparison to MAGIC was possible, MethylFLYA generated more strong assignments

with higher accuracy (FLYA correct: 168, errors: 5; MAGIC correct: 130, errors: 50). The error rate of MAGIC on the two benchmark cases, EIN and ATCase, was ~10% when using the parameters that resulted in the highest MAGIC score (Supplementary Table 6, Supplementary Fig. 10). However, a significantly worse performance of MAGIC was found on the HSP90 data (Supplementary Fig. 10), which likely reflects a reduced quality of the methyl–methyl NOESY data (Supplementary Fig. 6). We note that the HSP90 dataset used in this study features Ile- δ_1 , dimethyl Leu- $\delta_{1/2}$ and Val- $\gamma_{1/2}$ labeling, and, consequently, a significantly sparser methyl–methyl NOESY network (78 labeled methyl groups and 330 3D NOE peaks), when compared to the HSP90 dataset employed in the original MAGIC study (111 labeled methyl groups and 686 3D NOE peaks)³³. The latter data were obtained on the N-terminal domain of HSP90 ¹H,¹³C-labeled on Ala- β , Met- ϵ , Thr- γ_2 , Ile- δ_1 , dimethyl Leu- $\delta_{1/2}$ and Val- $\gamma_{1/2}$ methyls, for which the authors report confident assignments of 88% of methyls with high accuracy (94%)³³.

A comparison of the assignments found by the different methods reveals that MAGMA and MethylFLYA produce the most similar solutions, which agree on 288 of the methyl assignments on this benchmark (Fig. 5, Supplementary Fig. 11, Supplementary Table 1). In contrast, MethylFLYA shares only 184 and 96 assignments with MAP-XSII and FLAMEnGO2.0, respectively. The intersection profiles are protein-specific (Supplementary Fig. 11), however, overall, a high degree of overlap with MethylFLYA solutions is seen for MAGMA (Supplementary Fig. 11, Supplementary Table 1). There are instances of confident assignments by MAGMA that are deemed tentative or ambiguous by MethylFLYA, and vice versa (Supplementary Table 1). Given that both protocols were given the same input data, a possible explanation for such assignment differences could be algorithm-specific parameters. The distance cutoffs used to generate the expected NOE contacts were similar for the two methods. Nonetheless, distance cutoff for MethylFLYA is applied as an r^{-6} sum over the methyl proton distances, whereas MAGMA considers methyl carbon distances and, in addition, averages two methyl carbon positions for the geminal methyl groups of Leu and Val, which are treated separately by MethylFLYA. Therefore, the exact composition of the expected NOE contacts differs between the two methods, resulting in differences in restraint matching. Furthermore, MAGMA provides assignment results for one distance cutoff, whereas, for its confident assignments, MethylFLYA requires assignment consistency over three distance cutoff values separated by 0.5 Å (see Methods). Finally, MAGMA uses exact graph comparison algorithms to exhaustively sample

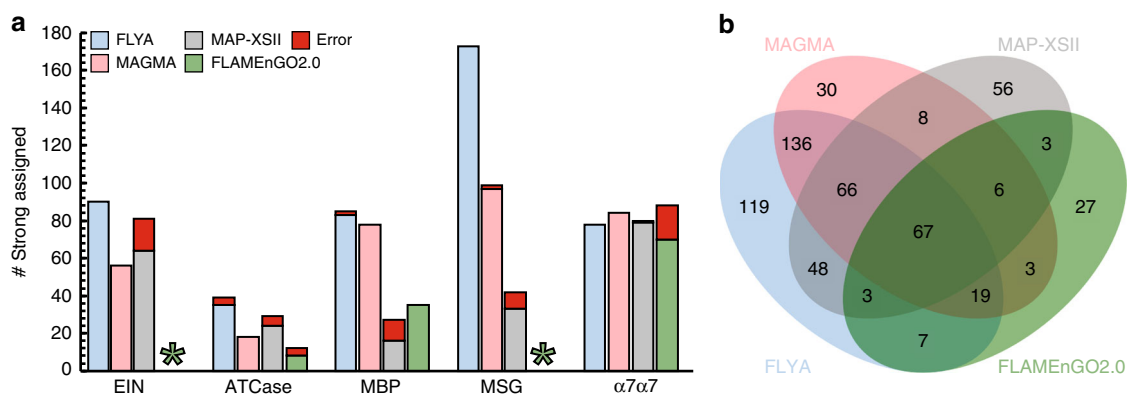


Fig. 5 Comparison of MethylFLYA, MAGMA, MAP-XSII, and FLAMEnGO2.0. **a** The number of correctly and erroneously strongly (i.e. confidently) assigned methyl resonances for each of the cases is shown. Asterisks are given in the places where no confident (100%) assignments could be obtained with the FLAMEnGO2.0 software. **b** Mutual agreement of the methyl group resonance assignments among the four methods. Numbers of assigned methyls are given in the intersection sets

all assignment solutions that maximize the number of explained NOEs. In contrast, the evolutionary algorithm in MethylFLYA uses a heuristic to converge on a subset of most likely solutions, relying on differences between parallel runs of the algorithm to assess assignment self-consistency. Despite the listed differences, the high overlap in assignment solutions between MethylFLYA and MAGMA and their high accuracies demonstrate the complementarity of these two methods. Comparing the solutions from the two methods therefore constitutes a useful cross-validation approach, as the methyl assignments in the intersection of the two methods are completely accurate (Supplementary Table 1). Moreover, an agreement in erroneous assignment is rare across the tested methods, suggesting the utility of all existing protocols for assignment cross-validation (Supplementary Table 1).

In conclusion, we have presented an NOE-based approach to automatic methyl resonance assignment that is a significant advance over existing methods. Even though the general FLYA algorithm underlying MethylFLYA (Fig. 1) was originally designed to deal with through-bond, or a combination of through-bond and through-space information³⁴, the method proved powerful also for the assignment of methyl groups exclusively from NOESY and structural data (Fig. 2). This confirms earlier findings showing that FLYA is effective in assigning small proteins exclusively from ¹³C and ¹⁵N-resolved NOESY data³⁶. However, the assignment of methyl resonances in proteins as large as 360 kDa ($\alpha_7\alpha_7$), based on exclusively methyl–methyl NOEs, presents a considerably greater challenge because of data sparsity and minimal redundancy in data content. Nonetheless, MethylFLYA could generate as many, and in most cases significantly more, correct methyl assignments than existing algorithms (Fig. 5a). Only a very small number of assignments from MethylFLYA were erroneous, and all of these were to methyls spatially proximal to the correct assignment in the 3D structure (Supplementary Fig. 4), thus limiting their impact on studies relying on methyl assignments to deduce lower-resolution information, e.g., a protein–protein or protein–ligand interface. In other cases that rely strongly on site-specific interpretations of methyl resonance assignments, the user can apply stricter criteria on assignment confidence by requiring higher self-consistency of assignments from multiple parallel runs of the algorithm, e.g. from the default 80% to 90% or higher. Furthermore, the complementarity between MethylFLYA and MAGMA could be exploited. The user could also combine the automatic assignment with site-directed mutagenesis in regions of special interest. Any methyl resonance assignments previously known or newly acquired through site-directed mutagenesis can be fixed in the protocol of MethylFLYA, which will further aid its performance.

MethylFLYA is fast and robust in coping with ambiguous and erroneous NOEs, showing nearly identical performance on raw and refined NOESY data (Fig. 2, Table 1), and robustness to differences in input protein structures (Supplementary Fig. 9). MethylFLYA is also tolerant to ambiguity in the identity of Leu and Val resonances, whereas it significantly benefits from experimentally linking the methyl resonances from the geminal Leu/Val methyl groups (Fig. 3, Supplementary Fig. 7). The latter is further confirmed by the results of the MethylFLYA assignment of EIN using exclusively the 2D [¹H,¹³C]-HMQC and the 4D methyl–methyl NOESY spectrum (Supplementary Fig. 7). We strongly advise running MethylFLYA with this information provided, which was also noted as critical and beneficial in the MAGMA³² and MAGIC³³ studies, respectively. Stereospecific labeling of Leu and Val methyls is another promising approach⁴⁹ which is expected to further enhance the performance of MethylFLYA as it reduces the number of methyl resonances to be

assigned, removes the need for the geminal Leu/Val methyl pairing, and provides longer methyl–methyl NOE restraints.

We emphasize that MethylFLYA can provide reliable assignments for approximately a quarter of methyl resonances based solely on manually or CYPICK-derived peak lists from 2D [¹H,¹³C]-HMQC and methyl–methyl NOESY spectra, and the “best guess” assignment of methyl residue types from, e.g., the BMRB chemical shift statistics. In such cases, we advise using a structure-based methyl chemical shift prediction⁵² to identify any outliers of the characteristic residue type-based chemical shifts, which are likely to be assigned incorrectly based on BMRB statistics (Supplementary Fig. 8).

Given high-quality 2D [¹H,¹³C]-HMQC and methyl–methyl NOESY spectra, automatic peak picking using CYPICK combined with MethylFLYA can provide assignment results of the same quality as the expert-prepared peak lists (Fig. 4, Supplementary Fig. 7). In fact, the combination with CYPICK can, in some cases, even lead to a higher number of strong assignments, or improved assignment accuracy (Fig. 4, Supplementary Fig. 7D, E). This finding is surprising considering that CYPICK peak lists show differences to the expert-generated ones and omit a fraction of the expert-picked signals, even for the highest quality spectra available (17% expert-picked peaks omitted for EIN, Supplementary Table 4). Given that successful applications of CYPICK with MethylFLYA are presently restricted to only two examples⁵⁰ (Fig. 4), wider applications of CYPICK with MethylFLYA will be required in the future to realistically judge its potential for fully-automated methyl resonance assignment.

A high fraction of overlap in confident methyl assignments between MAGMA and MethylFLYA indicates the complementarity of the two methods and can be useful in de novo assignment cross-validation (Fig. 5b). The utility of rapid, accurate methyl assignments is highlighted by recent studies that used NOEs between an unlabeled ligand and a methyl-labeled protein as restraints to generate models of the docked complex^{32,38,53,54} and PCSs to measure reorientation of methyl groups upon ligand binding⁵⁵. In the future, MethylFLYA could be extended to incorporate paramagnetic restraints, such as PREs or PCSs, or be combined with existing software packages that predominantly rely on these restraints^{26,27}. Furthermore, MethylFLYA can straightforwardly be used to assign methyl resonances in solid-state NMR spectra⁵⁶.

Methods

Overview of the MethylFLYA algorithm. The FLYA algorithm³⁴ determines resonance assignments by establishing an optimal mapping between expected peaks that are derived from knowledge of the protein sequence, types of NMR experiments, and, if available, 3D structure, and the observed peaks that are identified in the corresponding measured spectra. This mapping, and hence the assignments, are optimized by an evolutionary algorithm coupled to a local optimization routine^{34,57}. MethylFLYA adopts the general FLYA algorithm for the assignment of methyl groups based on methyl–methyl NOEs and a known 3D structure. MethylFLYA uses the atom positions from the input protein structure and magnetization transfer pathways defined for each NMR experiment type to compute a network of expected peaks (Fig. 1). The mapping of expected peaks to measured ones starts from an initial population of random assignment solutions, which are optimized through successive generations by an evolutionary algorithm. To select the best individuals for recombination, a scoring function is employed, which takes into account the alignment of peaks assigned to the same atom, the completeness of the assignment, and the minimization of chemical shift degeneracy³⁴. In each generation, a local optimization routine reassigns a subset of expected peaks through a defined number of iterations. This protocol is repeated multiple times starting from different random initial assignments. Details of the MethylFLYA algorithm are given in the following sections.

MethylFLYA scripts. Automated methyl assignment with MethylFLYA is performed by four scripts (CYANA macros written in the INCLAN⁵⁸ programming language) as described in Supplementary Methods. The initialization macro, `init.cya`, is executed when CYANA starts and reads the library of residues and NMR experiment types, as well as the protein sequence. The preparation macro, `PREP.`

cya, prepares the input data for the subsequent automated assignment calculations. This includes the splitting of experimental peak lists according to amino acid type (see below) and the setup for generating the corresponding expected peaks, which is saved in the expected peak list generation macro, `peaklists.cya`. `PREP.cya` may also include other preparatory steps, such as attaching hydrogen atoms to an input 3D structure from X-ray crystallography. The calculation macro, `FLYA.cya`, performs the actual automated assignment calculations using the `peaklists.cya` macro to generate the expected peaks with different values for the NOE distance cutoff (see below). After completion of the automated assignment calculations, the consolidation macro, `CONSOL.cya`, consolidates the assignment results from all individual optimization runs into a single consensus resonance assignment³⁴, which is the main result of `MethylFLYA`.

Library of NMR experiments. The types of NMR experiments that contribute input peak lists to `MethylFLYA` are defined in the `CYANA` library^{34,36} (Fig. 1 and Supplementary Methods). For each spectrum type, a library entry defines the types of atoms that are observed in each spectral dimension and one or several magnetization transfer pathways that give rise to peaks. A magnetization transfer path is given by a probability for observing the corresponding experimental peak and a linear list of atom types that defines a molecular fragment, in which atoms must be of a given type (e.g. ¹H_{amide}, ¹H_{aliphatic}, ¹H_{aromatic}, ¹³C_{aliphatic}, ¹³C_{aromatic}, ¹⁵N, etc.) and connected to the next atom in the list either by a covalent bond or by an NOE, i.e. a distance shorter than a given cutoff in the 3D structure. An expected peak is generated whenever a molecular fragment matches the covalent structure and, in case of NOEs, the 3D protein structure.

The following NMR experiments were used for `MethylFLYA` calculations in this paper: 2D [¹H,¹³C]-HMQC (formally called C13HSQC in the `CYANA` library), 3D CCH-NOESY (CCNOESY3D; ¹³C₁, ¹³C₂, ¹H₂ dimensions), 3D HCH-NOESY (C13NOESY; ¹H₁, ¹³C₂, ¹H₂ dimensions), 4D HCCH NOESY (CCNOESY; ¹H₁, ¹³C₁, ¹³C₂, ¹H₂ dimensions), and, optionally, 4D short-mixing time HCCH NOESY. The latter experiment can be recorded on a doubly methyl-labelled ([¹³C_{δ1}¹H₃/¹³C_{δ2}¹H₃]-Leu, [¹³C_{γ1}¹H₃/¹³C_{γ2}¹H₃]-Val) protein sample to correlate the geminal methyl groups of Leu and Val to each other. It is formally treated as an HCCCH-TOCSY experiment in the `CYANA` library for `MethylFLYA`. The experiment entries in the library are given in Supplementary Methods.

Input peak lists. `MethylFLYA` operates on peak lists with observed peaks from the measured NMR spectra that contribute data for the resonance assignment. The peak lists can be supplied in `XEASY`⁵⁹ format (Supplementary Methods), or other formats supported by `CYANA`. If residue type-specific information is available, e.g. from appropriately isotope-labeled samples, the [¹H,¹³C]-HMQC peak list can be split into separate files containing only the methyl peaks of a certain residue type (called, for example, “C13HSQC_V.peaks” for Val peaks). The NOESY peak lists can be split similarly according to the two amino acid types involved in an NOE. In the `MAGMA` study, this information was available from manually assigned NOESY peak lists³². Here, unassigned NOESY peak lists are used as input, and each NOESY peak is automatically attributed to the amino acid types of the two [¹H,¹³C]-HMQC peaks with the closest matching chemical shifts. Separate peak lists are written for each pair of amino acid types (called, for example, “CCNOESY_LL.peaks” and “CCNOESY_LV.peaks” for NOEs between two Leu residues or between Leu and Val, respectively). Splitting peak lists by residue types is optional. `MethylFLYA` also supports joint lists for the resonances of Leu/Val type, as well as for any other amino acid type combinations.

Expected peak lists. Lists of expected peaks are generated by `MethylFLYA` for a given set of experiments based on the protein sequence, the 3D structure, the library of NMR experiments, and the isotope labeling pattern. The input 3D structure file must contain hydrogen atoms. For all calculations in this paper, hydrogens were added to the input X-ray structures using the `CYANA` command ‘atoms attach’. If residue type-specific experimental peak lists are available, `MethylFLYA` generates a separate expected [¹H,¹³C]-HMQC peak list for each amino acid type and separate NOESY peak lists for each pair of amino acid types. Splitting the measured and expected peak lists by residue type(s) restricts the matching of expected peaks to measured peaks of the same amino acid type(s) in the automated assignment algorithm (Fig. 1).

The distance cutoff d_{cut} for NOEs is an important parameter for generating expected NOESY cross peaks because the number of expected NOEs is approximately proportional to d_{cut}^3 . `MethylFLYA` computes the effective distance for a pair of methyl groups as the r^{-6} -sum over the nine individual ¹H-¹H distances, i.e.

$$d_{\text{eff}} = \left(\sum_{i=1}^3 \sum_{j=1}^3 d_{ij}^{-6} \right)^{-1/6} \quad (1)$$

where d_{eff} stands for the effective distance, the sum includes all ¹H atoms of two methyl groups, and d_{ij} is the Euclidean distance between individual methyl protons i and j that belong to two different methyl groups in the input structure. For the case that all d_{ij} distances are assumed to be approximately equal, this yields $d_{\text{eff}} \approx 9^{-1/6} d_{ij} = 0.693 d_{ij}$. It should be noted that applying, for instance, a 5 Å cutoff to the

effective distance d_{eff} , allows inter-carbon distances between the two methyl groups of up to $5.0/0.693 + 2 \times 1.1 \approx 9.4$ Å, which includes twice the C–H bond length of 1.1 Å. To avoid giving high confidence to methyl assignments that are affected by minor changes of the NOE distance cutoff parameter d_{cut} , `MethylFLYA` performs assignment calculations with the three slightly different cutoffs of $d_{\text{cut}} - 0.5$ Å, d_{cut} and $d_{\text{cut}} + 0.5$ Å, and determines the consensus assignments from the results obtained with the three cutoffs (see below).

For the calculations in this paper, the NOESY cross peak observation probability was optimized (see below) and then set to 0.1 for expected NOESY peaks, and to 1 for expected C13HSQC and short-mixing time NOESY (HCCH) peaks for the calculations in this paper.

Optimization of assignments. Assignments are optimized by `MethylFLYA` using the same algorithm as the original `FLYA` method³⁴. `MethylFLYA` uses chemical shift tolerances for the assignment calculations and results evaluation. These were set to 0.4 ppm for ¹³C and 0.04 ppm for ¹H chemical shifts for all calculations of this paper. The population size for the evolutionary optimization algorithm³⁴ was set to 200, the value that was previously found to be optimal for exclusively NOESY-based `FLYA` calculations³⁶. The number of iterations of the local optimization routine that is coupled to the evolutionary algorithm was kept at the default value of 15,000. For each distance cutoff value, `MethylFLYA` performs 100 independent runs of the optimization algorithm with identical input data and parameters that start from different initial random assignments.

Consensus assignments. It is important for an assignment algorithm to distinguish reliable assignments, in which the algorithm has a high confidence, from others that are tentative or ambiguous. To establish the confidence of the assignment of an individual atom, `MethylFLYA` analyzes the chemical shift values obtained in a series of independent runs of the optimization algorithm. The global maximum of the sum of Gaussians centered at the chemical shift values of the given atom in the individual optimization runs defines the consensus chemical shift value of the atom³⁴. The standard deviation of these Gaussians is set to the chemical shift tolerance value of the atom (0.4 ppm for ¹³C and 0.04 ppm for ¹H). A consensus assignment is classified as “strong” (reliable) if more than 80% of the integral of the sum of Gaussians is concentrated in the region of the consensus chemical shift \pm tolerance, i.e. if more than 80% of the individual runs yielded (within the tolerance) the same chemical shift value. It has been shown for the original `FLYA` algorithm that strong assignments are much more accurate than the remaining “weak” ones³⁴.

In `MethylFLYA`, consolidation into consensus assignments is enhanced in three ways over the original `FLYA` algorithm. (i) Three series of 100 individual runs are performed with three slightly different distance cutoffs for the generation of expected NOESY peaks (see above), and the consolidation is performed over all 3×100 individual runs of the optimization algorithm. This makes the algorithm less susceptible to the, necessarily somewhat arbitrary, choice of the NOE distance cutoff value, thereby reducing the number of erroneous strong assignments. (ii) Special measures are necessary for the geminal methyls of Leu and Val, for which the stereospecific assignment is unknown a priori. In this case, the chemical shift values obtained for the two methyls in the individual runs are redistributed such that the consensus assignments of the first/second methyl group are determined from the smaller/larger of the two chemical shift values in each run, and `FLYA` does not attempt to determine stereospecific assignments. In the original `FLYA` algorithm³⁴ this approach was applied independently to the ¹H pair and the ¹³C pair of geminal Leu or Val methyl groups. This could result in inconsistent consensus assignments for the ¹H and ¹³C resonances of Leu and Val geminal methyl groups, even though the underlying ¹H and ¹³C assignments from the individual runs were always consistent with each other. To avoid this problem, the ¹H and ¹³C chemical shifts of Leu and Val geminal methyl groups are consolidated jointly in `MethylFLYA`. (iii) Methyl assignments are only accepted as strong if at least one methyl-methyl NOE is assigned to the methyl group. This excludes assignments for which no experimental basis exists.

MethylFLYA output. At the end of an assignment run, `MethylFLYA` outputs the list of consensus chemical shifts (`consol.prot`) and a table with assignment results (`consol.tab`). In the `consol.tab` file, strong (reliable) assignments are marked with the label ‘strong’. Other, tentative and ambiguous assignments are also reported for possible manual inspection. Further assignment statistics are given in the `flya.txt` file. It reports the number of expected, measured, and assigned peaks for each peak list, which are useful to detect problems with individual spectra or the assignment as a whole. In addition, more detailed information about the reliability of each resonance assignment is given, and, for each assignable atom, the expected and mapped measured peaks that have been used to establish its assignment are reported.

Optimization of MethylFLYA parameters. To establish optimal parameters for the `MethylFLYA` calculations, we tested a range of values for the methyl ¹H-¹H distance cutoffs for the generation of expected NOESY cross peaks, $d_{\text{cut}} = 3.0, 3.5, \dots, 8.0$ Å (Supplementary Figs. 1, 2), observation probabilities for expected methyl-methyl NOESY peaks, $p_{\text{NOE}} = 0.1, 0.2, \dots, 0.9$ (Supplementary Figs. 1, 2),

and the number of independent assignment optimization runs (Supplementary Fig. 3).

Automated peak picking with CYPICK. The CYPICK³⁷ algorithm for automated peak picking was applied to the NOESY spectra of EIN, ATCase, and HSP90. CYPICK relies on analyzing 2D contour lines of the spectrum, which are placed at intensity levels $I_i = \beta L^i$, where $i = 0, 1, \dots$ and L is the noise level of the spectrum that is estimated automatically by CYPICK. In this study, we used baseline factors $\beta = 2, 3, 4, 5, 10$ while keeping γ fixed at 1.3. The scaling factors for the spectral dimensions³⁷ were set to 0.18 and 0.16 ppm for the first and second ¹³C dimension, and 0.036 ppm for the ¹H dimension. The manually prepared or CYPICK-generated 2D [¹H,¹³C]-HMQC peak lists were used as a frequency filter in CYPICK, restricting peak picking in the ¹³C/¹³C-separated NOESY spectrum to locations within 0.01/0.1 ppm ¹H/¹³C chemical shift from a [¹H,¹³C]-HMQC peak position. Local maxima within the tolerance range that fulfilled the circularity and convexity criteria³⁷ were considered as peaks and stored in the peak list.

The peak picking performance was assessed by computing the find, artifact, and overall scores (with an artifact weight of 0.2) with respect to manually prepared reference peak lists³² using a tolerance of 0.04 ppm for ¹H and 0.4 ppm for ¹³C chemical shifts, as described in the CYPICK publication³⁷.

MethylFLYA calculations using minimal input information. The 2D HMQC spectrum was picked manually using Sparky software⁴². The BMRB chemical shift statistics⁵¹ and the known number of Ile, Ala, Leu, Val residues in the protein sequence were used to generate the “best-guess” assignment of the peaks in the 2D ¹H-¹³C HMQC spectrum to three methyl residue types: Ile, Ala, and Leu/Val (Supplementary Fig. 7A). The average methyl ¹H and ¹³C chemical shifts \pm one standard deviation were used to define regions associated to each methyl residue type. To assign residue types to peaks in the region of the spectrum where Ala and Leu/Val types overlap (Supplementary Fig. 7B), two strategies were followed: (i) the number of expected Ala peaks was maximized by attributing all the peaks in the overlapped regions to the Ala type; (ii) the peaks in the overlapped and border regions between the two types were assigned to both Ala and Leu/Val type (Supplementary Table 5). For the MethylFLYA calculations, the 4D methyl-methyl NOESY spectrum was reanalyzed based on the newly acquired 2D ¹H-¹³C HMQC peak list. The NOESY peaks in the planes of overlapped resonances in the 2D spectrum were repicked. NOESY cross peaks were attributed automatically by CYANA to the closest methyl groups from the 2D [¹H,¹³C]-HMQC spectrum (see above; section Input peak lists). SHIFTX2⁵² was used to predict methyl resonances based on the EIN protein structure (PDB ID 1EZA) in order to identify any significant deviations from the BMRB statistics⁵¹.

The Leu/Val geminal methyl group pairing was performed based on the reference assignment, restricted to the resolved methyl resonances in the Leu/Val region of the 2D [¹H,¹³C]-HMQC spectrum (Supplementary Table 5 (iii)). If the geminal pair of any given Leu/Val resonance from the reference assignment mapped to an overlapped peak in the spectrum, both methyl resonances were removed from the geminal pairing. The geminal Leu/Val methyl pairs were supplied to MethylFLYA calculations in the form of an HCCCH peak list (see above; section Library of NMR experiments).

Comparison with other assignment algorithms. The performance of the alternative structure-based methyl assignment algorithms MAGMA³², MAP-XSII²⁹, and FLAMENGO2.0³¹ has been compared earlier³². Here, we used the available results and identical parameters³², with the exception of the MSG dataset, for which the calculations were repeated using the crystal structure (PDB ID 1D8C). A comparison to the MAGIC method was performed on a subset of the proteins for which NOESY spectra were available (EIN, ATCase, and HSP90). Both filtered and unfiltered manually picked methyl NOE peak lists were tested, as well as a range of distance thresholds (lower 4, 5, 6, 7 Å and higher 7, 8, 9, 10 Å, respectively; see Supplementary Table 6) for computing the inter-methyl connectivity network from the X-ray structure, and peak matching tolerance values (0.1, 0.2, 0.4 ppm for ¹³C; 0.01, 0.02, 0.04 ppm for ¹H; see Supplementary Table 6). The mutual agreement between the resonance assignments generated by the different methods was visualized using an online tool available at the GPCRdb web interface (<http://www.gpcrdb.org/signprot/statistics>).

Data availability

Experimental input data and corresponding MethylFLYA output data are available at <http://www.cyana.org/methylflya.tgz>. Other data are available from the corresponding author upon reasonable request.

Code availability

MethylFLYA scripts for CYANA are available in <http://www.cyana.org/methylflya.tgz>.

Received: 1 April 2019; Accepted: 2 October 2019;

Published online: 29 October 2019

References

- Steven, A. C., Baumeister, W., Johnson, L. N. & Perham, R. N. *Molecular Biology of Assemblies and Machines*. (Garland Science, 2016)
- Pervushin, K., Riek, R., Wider, G. & Wüthrich, K. Attenuated T_2 relaxation by mutual cancellation of dipole-dipole coupling and chemical shift anisotropy indicates an avenue to NMR structures of very large biological macromolecules in solution. *Proc. Natl Acad. Sci. USA* **94**, 12366–12371 (1997).
- Tugarinov, V., Hwang, P. M., Ollerenshaw, J. E. & Kay, L. E. Cross-correlated relaxation enhanced ¹H-¹³C NMR spectroscopy of methyl groups in very high molecular weight proteins and protein complexes. *J. Am. Chem. Soc.* **125**, 10420–10428 (2003).
- Ollerenshaw, J. E., Tugarinov, V. & Kay, L. E. Methyl TROSY: explanation and experimental verification. *Magn. Reson. Chem.* **41**, 843–852 (2003).
- Religa, T. L., Sprangers, R. & Kay, L. E. Dynamic regulation of archaeal proteasome gate opening as studied by TROSY NMR. *Science* **328**, 98–102 (2010).
- Rosenzweig, R. & Kay, L. E. Bringing dynamic molecular machines into focus by methyl-TROSY NMR. *Annu. Rev. Biochem.* **83**, 291–315 (2014).
- Boswell, Z. K. & Latham, M. P. Methyl-based NMR spectroscopy methods for uncovering structural dynamics in large proteins and protein complexes. *Biochemistry* **58**, 144–155 (2019).
- Xing, Q. et al. Structures of chaperone-substrate complexes docked onto the export gate in a type III secretion system. *Nat. Commun.* **9**, 1773 (2018).
- Zhang, H. Y. & van Ingen, H. Isotope-labeling strategies for solution NMR studies of macromolecular assemblies. *Curr. Opin. Struct. Biol.* **38**, 75–82 (2016).
- Wiesner, S. & Sprangers, R. Methyl groups as NMR probes for biomolecular interactions. *Curr. Opin. Struct. Biol.* **35**, 60–67 (2015).
- Proudfoot, A., Frank, A. O., Ruggiu, F., Mamo, M. & Lingel, A. Facilitating unambiguous NMR assignments and enabling higher probe density through selective labeling of all methyl containing amino acids. *J. Biomol. NMR* **65**, 15–27 (2016).
- Clark, L. et al. Methyl labeling and TROSY NMR spectroscopy of proteins expressed in the eukaryote *Pichia pastoris*. *J. Biomol. NMR* **62**, 239–245 (2015).
- Suzuki, R. et al. Methyl-selective isotope labeling using α -ketoisovalerate for the yeast *Pichia pastoris* recombinant protein expression system. *J. Biomol. NMR* **71**, 213–223 (2018).
- Kofuku, Y. et al. Deuteration and selective labeling of alanine methyl groups of β_2 -adrenergic receptor expressed in a baculovirus-insect cell expression system. *J. Biomol. NMR* **71**, 185–192 (2018).
- Tugarinov, V., Choy, W. Y., Orekhov, V. Y. & Kay, L. E. Solution NMR-derived global fold of a monomeric 82-kDa enzyme. *Proc. Natl Acad. Sci. USA* **102**, 622–627 (2005).
- Gorman, S. D., Sahu, D., O'Rourke, K. F. & Boehr, D. D. Assigning methyl resonances for protein solution-state NMR studies. *Methods* **148**, 88–99 (2018).
- Kay, L. E., Ikura, M., Tschudin, R. & Bax, A. Three-dimensional triple-resonance NMR spectroscopy of isotopically enriched proteins. *J. Magn. Reson.* **89**, 496–514 (1990).
- Tugarinov, V. & Kay, L. E. Ile, Leu, and Val methyl assignments of the 723-residue malate synthase G using a new labeling strategy and novel NMR methods. *J. Am. Chem. Soc.* **125**, 13868–13878 (2003).
- Sattler, M., Schleucher, J. & Griesinger, C. Heteronuclear multidimensional NMR experiments for the structure determination of proteins in solution employing pulsed field gradients. *Prog. Nucl. Magn. Reson. Spectrosc.* **34**, 93–158 (1999).
- Sprangers, R., Gribun, A., Hwang, P. M., Houry, W. A. & Kay, L. E. Quantitative NMR spectroscopy of supramolecular complexes: Dynamic side pores in ClpP are important for product release. *Proc. Natl Acad. Sci. USA* **102**, 16678–16683 (2005).
- Sprangers, R., Velyvis, A. & Kay, L. E. Solution NMR of supramolecular complexes: providing new insights into function. *Nat. Methods* **4**, 697–703 (2007).
- Gelis, I. et al. Structural basis for signal-sequence recognition by the translocase motor SecA as determined by NMR. *Cell* **131**, 756–769 (2007).
- Xiao, Y., Warner, L. R., Latham, M. P., Ahn, N. G. & Pardi, A. Structure-based assignment of Ile, Leu, and Val methyl groups in the active and inactive forms of the mitogen-activated protein kinase extracellular signal-regulated kinase 2. *Biochemistry* **54**, 4307–4319 (2015).
- Velyvis, A., Schachman, H. K. & Kay, L. E. Assignment of Ile, Leu, and Val methyl correlations in supra-molecular systems: an application to aspartate transcarbamoylase. *J. Am. Chem. Soc.* **131**, 16534–16543 (2009).
- John, M. et al. Sequence-specific and stereospecific assignment of methyl groups using paramagnetic lanthanides. *J. Am. Chem. Soc.* **129**, 13749–13757 (2007).

26. Lescanne, M. et al. Methyl group assignment using pseudocontact shifts with PARAssign. *J. Biomol. NMR* **69**, 183–195 (2017).
27. Venditti, V., Fawzi, N. L. & Clore, G. M. Automated sequence- and stereo-specific assignment of methyl-labeled proteins by paramagnetic relaxation and methyl-methyl nuclear overhauser enhancement spectroscopy. *J. Biomol. NMR* **51**, 319–328 (2011).
28. Xu, Y. Q. et al. Automated assignment in selectively methyl-labeled proteins. *J. Am. Chem. Soc.* **131**, 9480–9481 (2009).
29. Xu, Y. Q. & Matthews, S. MAP-XSII: an improved program for the automatic assignment of methyl resonances in large proteins. *J. Biomol. NMR* **55**, 179–187 (2013).
30. Chao, F.-A., Shi, L., Masterson, L. R. & Veglia, G. FLAMEnGO: A fuzzy logic approach for methyl group assignment using NOESY and paramagnetic relaxation enhancement data. *J. Magn. Reson.* **214**, 103–110 (2012).
31. Chao, F. A. et al. FLAMEnGO 2.0: An enhanced fuzzy logic algorithm for structure-based assignment of methyl group resonances. *J. Magn. Reson.* **245**, 17–23 (2014).
32. Pritisanac, I. et al. Automatic assignment of methyl-NMR spectra of supramolecular machines using graph theory. *J. Am. Chem. Soc.* **139**, 9523–9533 (2017).
33. Monneau, Y. R. et al. Automatic methyl assignment in large proteins by the MAGIC algorithm. *J. Biomol. NMR* **69**, 215–227 (2017).
34. Schmidt, E. & Güntert, P. A new algorithm for reliable and general NMR resonance assignment. *J. Am. Chem. Soc.* **134**, 12817–12829 (2012).
35. Güntert, P. & Buchner, L. Combined automated NOE assignment and structure calculation with CYANA. *J. Biomol. NMR* **62**, 453–471 (2015).
36. Schmidt, E. & Güntert, P. Reliability of exclusively NOESY-based automated resonance assignment and structure determination of proteins. *J. Biomol. NMR* **57**, 193–204 (2013).
37. Würz, J. M. & Güntert, P. Peak picking multidimensional NMR spectra with the contour geometry based algorithm CYPICK. *J. Biomol. NMR* **67**, 63–76 (2017).
38. Shah, D. M. et al. Rapid protein-ligand costructures from sparse NOE data. *J. Med. Chem.* **55**, 10786–10790 (2012).
39. Garrett, D. S. et al. Solution structure of the 30 kDa N-terminal domain of enzyme I of the *Escherichia coli* phosphoenolpyruvate:sugar phosphotransferase system by multidimensional NMR. *Biochemistry* **36**, 2517–2530 (1997).
40. Gardner, K. H., Zhang, X. C., Gehring, K. & Kay, L. E. Solution NMR studies of a 42 kDa *Escherichia coli* maltose binding protein b-cyclodextrin complex: Chemical shift assignments and analysis. *J. Am. Chem. Soc.* **120**, 11738–11748 (1998).
41. Tugarinov, V., Sprangers, R. & Kay, L. E. Probing side-chain dynamics in the proteasome by relaxation violated coherence transfer NMR spectroscopy. *J. Am. Chem. Soc.* **129**, 1743–1750 (2007).
42. Goddard, T. D. & Kneller, D. G. *Sparky 3*. (University of California, 2001)
43. Schmidt, E. et al. Automated solid-state NMR resonance assignment of protein microcrystals and amyloids. *J. Biomol. NMR* **56**, 243–254 (2013).
44. Aeschbacher, T. et al. Automated and assisted RNA resonance assignment using NMR chemical shift statistics. *Nucleic Acids Res.* **41**, e172 (2013).
45. Krähenbühl, B., El Bakkali, I., Schmidt, E., Güntert, P. & Wider, G. Automated NMR resonance assignment strategy for RNA via the phosphodiester backbone based on high-dimensional through-bond APSY experiments. *J. Biomol. NMR* **59**, 87–93 (2014).
46. Schmidt, E. et al. Automated resonance assignment of the 21 kDa stereo-array isotope labeled thiol-disulfide oxidoreductase DsbA. *J. Magn. Reson.* **249**, 88–93 (2014).
47. Lichtenecker, R. J., Coudeville, N., Konrat, R. & Schmid, W. Selective isotope labelling of leucine residues by using α -ketoacid precursor compounds. *ChemBioChem* **14**, 818–821 (2013).
48. Lichtenecker, R. J. et al. Independent valine and leucine isotope labeling in *Escherichia coli* protein overexpression systems. *J. Biomol. NMR* **57**, 205–209 (2013).
49. Gans, P. et al. Stereospecific isotopic labeling of methyl groups for NMR spectroscopic studies of high-molecular-weight proteins. *Angew. Chem. Int. Ed.* **49**, 1958–1962 (2010).
50. Pritisanac, I., Würz, J. M. & Güntert, P. Fully automated assignment of methyl resonances of a 36 kDa protein dimer from sparse NOESY data. *J. Phys.: Conf. Ser.* **1036**, 012008 (2018).
51. Ulrich, E. L. et al. BioMagResBank. *Nucleic Acids Res.* **36**, D402–D408 (2008).
52. Han, B., Liu, Y. F., Ginzinger, S. W. & Wishart, D. S. SHIFTX2: significantly improved protein chemical shift prediction. *J. Biomol. NMR* **50**, 43–57 (2011).
53. Orts, J. et al. NMR-based determination of the 3D structure of the ligand-protein interaction site without protein resonance assignment. *J. Am. Chem. Soc.* **138**, 4393–4400 (2016).
54. Mohanty, B. et al. Determination of ligand binding modes in weak protein-ligand complexes using sparse NMR data. *J. Biomol. NMR* **66**, 195–208 (2016).
55. Lescanne, M. et al. Methyl group reorientation under ligand binding probed by pseudocontact shifts. *J. Biomol. NMR* **71**, 275–285 (2018).
56. Huber, M. et al. A proton-detected 4D solid-state NMR experiment for protein structure determination. *Chemphyschem* **12**, 915–918 (2011).
57. Bartels, C., Güntert, P., Billeter, M. & Wüthrich, K. GARANT—a general algorithm for resonance assignment of multidimensional nuclear magnetic resonance spectra. *J. Comput. Chem.* **18**, 139–149 (1997).
58. Güntert, P., Dötsch, V., Wider, G. & Wüthrich, K. Processing of multidimensional NMR data with the new software PROSA. *J. Biomol. NMR* **2**, 619–629 (1992).
59. Bartels, C., Xia, T. H., Billeter, M., Güntert, P. & Wüthrich, K. The program XEASY for computer-supported NMR spectral analysis of biological macromolecules. *J. Biomol. NMR* **6**, 1–10 (1995).

Acknowledgements

We thank Prof. Andrew Baldwin for help with the MAGMA benchmark data set, and Marta Carneiro, Eiso Ab and Gregg Siegal for providing the HSP90 data set. Financial support by a Eurostars grant of the Swiss Confederation and a Grant-in-Aid for Scientific Research of the Japan Society for the Promotion of Science (JSPS) is gratefully acknowledged.

Author contributions

I.P. and P.G. designed and performed research, and wrote the paper. J.M.W. implemented and performed automated peak picking. T.R.A. assisted in spectral processing, analysis, and data interpretation. All authors contributed to data interpretation and commented on the paper.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at <https://doi.org/10.1038/s41467-019-12837-8>.

Correspondence and requests for materials should be addressed to P.G.

Peer review information *Nature Communications* thanks Benjamin Bardiaux, Martin Billeter and the other, anonymous, reviewers for their contributions to the peer review of this work. Peer review reports are available.

Reprints and permission information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2019