Supplementary Information

Noise peak filtering in multi-dimensional NMR spectra using convolutional neural networks

Naohiro Kobayashi[1,*], Yoshikazu Hattori[2], Takashi Nagata[3,4], Shoko Shinya[1], Peter Güntert[5,6,7], Chojiro Kojima[8] and Toshimichi Fujiwara[1*]

[1]Institute for Protein Research, Osaka University, Yamadaoka 3-2, Suita, Osaka, 565-0871, Japan
[2]Faculty of Pharmaceutical Sciences, Tokushima Bunri University, Tokushima, 770-8514, Japan
[3]Institute of Advanced Energy, Kyoto University, Kyoto, 611-0011, Japan
[4]Graduate School of Energy Science, Kyoto University, Kyoto, 611-0011, Japan
[5]Institute of Biophysical Chemistry, Goethe-University, 60438 Frankfurt am Main, Germany
[6]Department of Chemistry, Tokyo Metropolitan University, 192-0397, Japan
[7]Laboratory of Physical Chemistry, ETH Zürich, 8093 Zürich, Switzerland
[8]College of Engineering Science, Yokohama National University, Yokohama, 240-0801, Japan

*To whom correspondence should be addressed.

Details of methods described in main text

Preparation of training data

All the spectroscopic data used for training neural networks were acquired using a Bruker AVANCE III instrument (600 MHz, equipped with QCI cryo-probe ^{1}H/^{13}C/^{15}N/^{31}P). The sample, uniformly labeled ubiquitin (100% ^{15}N/^{13}C), for the NMR experiments was expressed in *E. coli* with M9 medium and purified using standard methods. The spectroscopic data sets include 2D ^{1}H-^{15}N HSQC and ^{1}H-^{13}C HSQC for aliphatic and aromatic regions (constant time evolution method), 3D HNCO, CBCA(CO)NH, HNCACB, HBHA(CO)NH and HCCH-TOCSY for the aliphatic region, and ^{15}N-edited NOESY and ^{13}C-edited NOESY for the aliphatic region. Additional spectrum data sets for training were also generated using identical FID data sets by applying slightly different window functions, phase miss-matches, base line corrections, and bulk water signal subtraction.

Peak lists for the above spectra were prepared using a combination of manual and automated picking (as discussed below); altogether more than 60,000 peaks were identified. Noise peaks were eliminated using the HSQC-masked filter method (as discussed below) and manually by visual inspection in MagRO-NMRView, an updated version of the NMR analysis tool named KUJIRA (Kobayashi *et al.*, 2007). Finally, ~2,800 noise and ~2,800 real signals were selected for training.

To prepare the training image data, 2D sub-matrices were extracted from the spectroscopic data based on the detected peak positions with a width of 0.4 ppm for the ^{1}H dimension, 2.0 ppm for the ^{13}C dimension, and 4.0 ppm for the ^{15}N dimension. For the NOESY experiments, the width for the ^{1}H indirect dimension was set at 0.6 ppm. The sub-matrix data were prepared as *xy*-images for the 2D spectra, and as *xy*- and transposed *zy*-images for the 3D spectra. The data intensity in the sub-matrix was normalized by the absolute intensity at the detected peak position and then scaled into a 0–255 grayscale, with values of 0–127 for negative and 128–255 for positive intensities. The absolute values of data points with absolute intensities below a manually optimized threshold value were set to zero (127 in grayscale). The data were further smoothed by standard bilinear interpolation into a 40 × 40 matrix, as shown in Fig. S1.

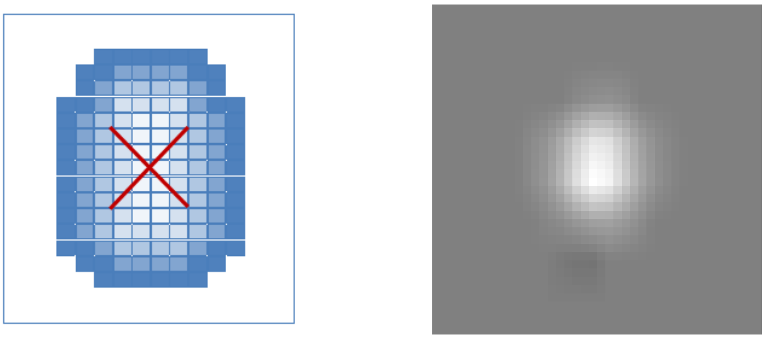

Fig. S1 Schematic representation of extracted 2D sub-matrix (left) and grayscale image generated using methods described in this section.

The obtained grayscale images were rotated by 90°, 180°, and 270°, mirrored horizontally and vertically, and their signs were changed to extend the data to ~58,000 images. These were randomly renumbered, labeled with noise:0 or signal:1 according to visual inspection, and 20% of the data were extracted as test data. Fig. S2 shows some of the images for the convolutional neural network (CNN) training.

Detail of the CNN noise filtering system and training

All CNN training and test calculations were performed on a standard PC with an Intel Core i7 6700K (over-clocked 4.7 GHz) CPU, nVIDIA GTX1080Ti (11GB) GPU, DDR4 64GB memory, m.2 SSD (PCIe) 256GB, and 32GB RAM-disk. Ubuntu16.04LTS 64bit was used as the operating system, with

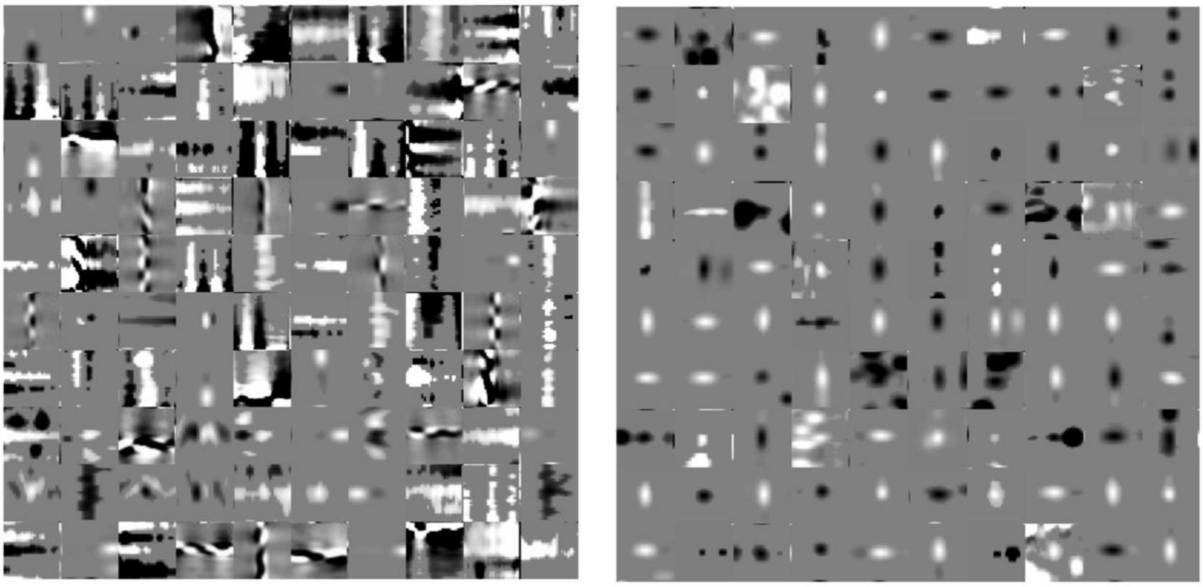

Fig. S2 Some of the training images for CNN noise filtration: 100 randomly selected images for noise (left) and signal (right).

The CNN were built using the Cognitive Neural Network Tool Kit (developed by Microsoft) version 2.0, as shown in Fig. 1 in the main text. The networks were constructed from two sets, each consisting of a convolutional layer and a max-pooling layer. The input layer was set to 1600 single-channel features corresponding to the 40 × 40 grayscale data matrices. The convolutional layers 1 and 2 comprised 16 and 32 local linear filters with 3 × 3 pixels, and one data point stride with automated zero-padding. For the max-pooling layers 1 and 2, a 2 × 2 window was used with a two data point stride, and dropout was used with a rate of 0.5 for each epoch in max-pooling layer 2. The dense layer was fully connected to the output tensor from max-pooling layer 2, which had 400 dimensions, followed by the final output with two units using the cross-entropy with soft-max function. The rectified linear unit (ReLU), as an activation function, was applied to the convolutional layers and the dense layer. Stochastic gradient descent (SGD) was used to accelerate CNN training for the number of epochs 32 with the mini-batch size as 32. The learning rate was 0.1 for the first 16 epochs and 0.01 for the remaining ones. The momentum per mini-batch was set to 0.0 for the first 10 epochs and 0.3 for the remaining ones. More details of the CNN techniques using CNTK are available in a web tutorial:
https://github.com/Microsoft/CNTK/wiki/Tutorial2#going-deep-convolutional-neural-networks-cnns

. The neural network parameters for training (mainly filter pixel size, learning rate, and momentum rate) were optimized manually to obtain a final accuracy with the test data of 98.5%. The soft-max function was used to export the results from two units of output layers as 0.0~1.0 float data.

Small three-layered neural networks (ZNN) were designed to validate the suspiciousness of the identified peaks. Networks consisting of a five-unit input layer, 32-unit hidden layer, and two-unit output layer were constructed using SGD, mini-batch methods and cross-entropy with soft-max function as error function with mini-batch size of 8 for 16 epochs by CNTK. The ReLU function and dropout with a rate of 0.5 were used for the hidden layer. The learning rate for each mini-batch was set 0.1 for the first 8 epochs and 0.01 for the remaining ones. About 12,000 peaks (20% for testing) were used to obtain an accuracy of ~95%. The input layer was evaluated in terms of five features, F_{znn} [n] (n = 0...4). The F_{znn} [0] is derived as shown below from the value F_{level} which has been statistically scaled intensity for each peaks calculated by following equation:

$$F_{\mathrm{level}} = \frac{||I_x| - t_x| - m_x}{r_x \sigma_x}$$

where I_x is the peak intensity of spectrum x beyond the threshold value t_x. The mean value m_x and the standard deviation σ_x of the peak intensities were calculated for all peaks excluding diagonal peaks such as in 3D HCCH-TOCSY, ^{15}N-edited NOESY and ^{13}C-edited NOESY, then recalculate them by the peaks excluding the strong peaks with intensity 2.0 times higher than the standard deviation. The weight factor r_x for the standard deviation is given by

$$r_x = \begin{cases} 2.0 & if\ N_x/M_x > 0.5 \\ N_x/M_x & if\ 0.1 \le N_x/M_x \le 0.5 \\ 0.1 & if\ N_x/M_x < 0.1 \end{cases}$$

where N_x is the number of expected peaks estimated from the spectrum x and protein sequence and M_x the number of actually detected peaks. The F_{level} is a kind of Z-value scaled based on the distribution of peak intensities where the weight factor can soften the weakness of signals if many more than the expected number of peaks are identified. The scaled F_{znn}[0] is obtained using scaled F_{level}:

$$sF_{\mathrm{level}} = 255 \times w_f \times \left(F_{\mathrm{level}} + b_f\right)$$

$$F_{\mathrm{znn}}[0] = \begin{cases} 0 & if\ sF_{\mathrm{level}} < 0 \\ sF_{\mathrm{level}} & if\ 0 \le sF_{\mathrm{level}} \le 255 \\ 255 & sF_{\mathrm{level}} > 255 \end{cases}$$

where w_f and b_f are a weight and bias for F_{level} value scaling, respectively (default is 0.2 and 2.0).

The input units F_{znn}[1] and F_{znn}[2] are corresponding to the judgment results performed by CNN, a simple comparison of two-unit output layer of CNN O_{cnn}[0] and O_{cnn}[1]. The ZNN input, F_{znn}[1], can be calculate from CNN output for xy-image;

$$F_{\mathrm{znn}}[1] = \begin{cases} 0 & if\ O_{cnn}[0] \ge O_{cnn}[1]\ (noise) \\ 255 & if\ O_{cnn}[0] < O_{cnn}[1]\ (signal) \end{cases}$$

F_{znn}[2] can be obtained similarly from output of CNN for zy-image. For 2D spectra, F_{znn}[2] is set to F_{znn}[1].

The input units 3 and 4 are derived from the absolute values of the difference between output values from two-unit from output layer of CNN. The ZNN input, F_{znn}[3], can be calculated from CNN output for xy-image:

$$F_{\mathrm{znn}}[3] = 255\ \times \max\left(O_{\mathrm{cnn}}[p]\right) \quad \text{with}\ p = 0\ \text{or}\ 1$$

where max(x) gives maximal element in the provided vector x. F_{znn}[4] can be obtained similarly from output of CNN for zy-image. For 2D spectra, F_{znn}[4] is set to F_{znn}[3]. On the output layer soft-max function was used to export $O_{\mathrm{znn}}[p] \in [0,1]$ (p = 0: noise, p = 1: signal) as the peak/noise probability. The entire system of CNN noise filtering is shown in Fig. S3.

In the final stage of the CNN filtering system, the user can export a filtered peak list based on the results from ZNN with cut-off value as shown in Fig. S4. As a result, the peaks are labeled as described in the flow-chart of Fig. S4. (If the user enables the option "strong" in the GUI, the peaks labeled as "M" are eliminated)

For CNN filtration of the actual data, the standard peak width for each dimension was estimated to optimize the image window size by checking the peak width at half-height for the top 50% highest peaks. The following sections explain the filtering pathway of fully automated peak identification and noise filtration for 2D and 3D spectra.

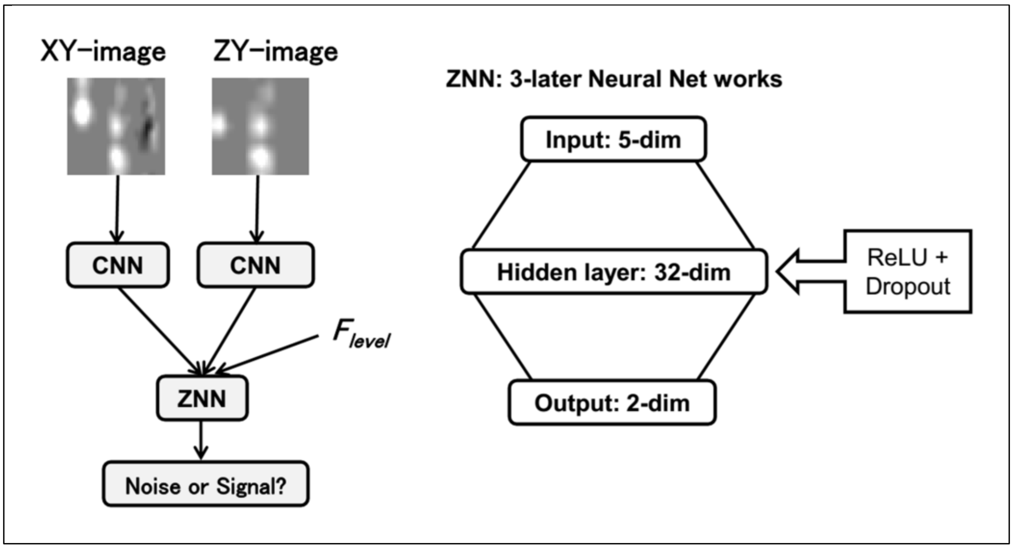

Fig. S3 Overview of the CNN filter system (left). Two CNN units are used for image recognition in the extracted *xy*- and *zy*-images giving two units of output $O_{\mathrm{cnn}}[0]$ and $O_{\mathrm{cnn}}[1]$. The right panel represents the construction of the 3-layer neural network, ZNN.

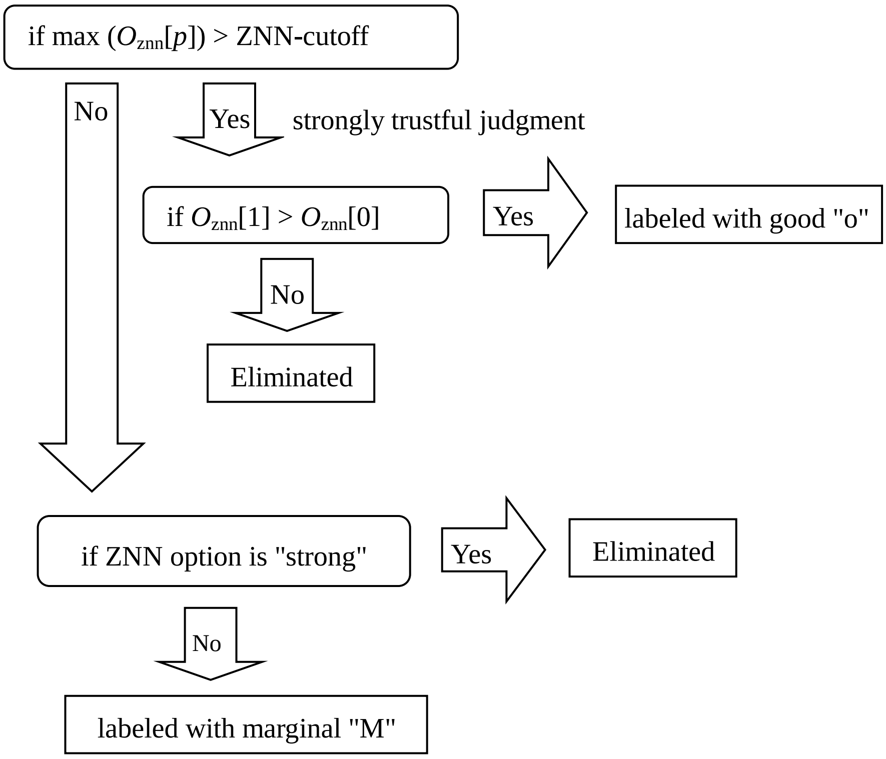
if max ($O_{\mathrm{znn}}[p]$) > ZNN-cutoff
No
Yes
strongly trustful judgment
if $O_{\mathrm{znn}}[1] > O_{\mathrm{znn}}[0]$
Yes
labeled with good "o"
No
Eliminated
if ZNN option is "strong"
Yes
Eliminated
No
labeled with marginal "M"

Fig. S4 Flow chart of final decision routine. $O_{\mathrm{znn}}[0]$ and $O_{\mathrm{znn}}[1]$ are values from the output layer of ZNN. By the maximal value in $O_{\mathrm{znn}}[p]$ the final peaks are labeled or eliminated depending on the value as a reliability of this routine with the ZNN-cutoff value (default 0.6).

Automated determination for spectrum data threshold stage

The automated spectrum data threshold was determined by randomly selecting 1,000 points of spectrum data. The spectrum data threshold was set to 6 times the median of their absolute intensities. Only for this estimation, the data points within 1.0 ppm from the water signal for the *x*-axis of 2D ^{1}H–^{13}C HSQC for aliphatic, 3D HCCH-TOCSY for aliphatic, and ^{13}C-edited NOESY for aliphatic were ignored. In the "deep" option mode, half the threshold will be used.

Automated peak picking stage

Automated peak identification is based on searching for local maxima in the spectrum. A data point is a local maximum if its (absolute) intensity is higher than for all its neighbors (8 in 2D, 26 in 3D). Then quadratic interpolation is used to approximate the peak maximum for each dimension using three surrounding data points [$i-1$], [i] and [$i+1$], as shown in Fig. S5. This method is simple and quick, therefore many spectrum viewers have used this function for peak detection. All the peak lists have the file name labeled with *_auto.xpk.

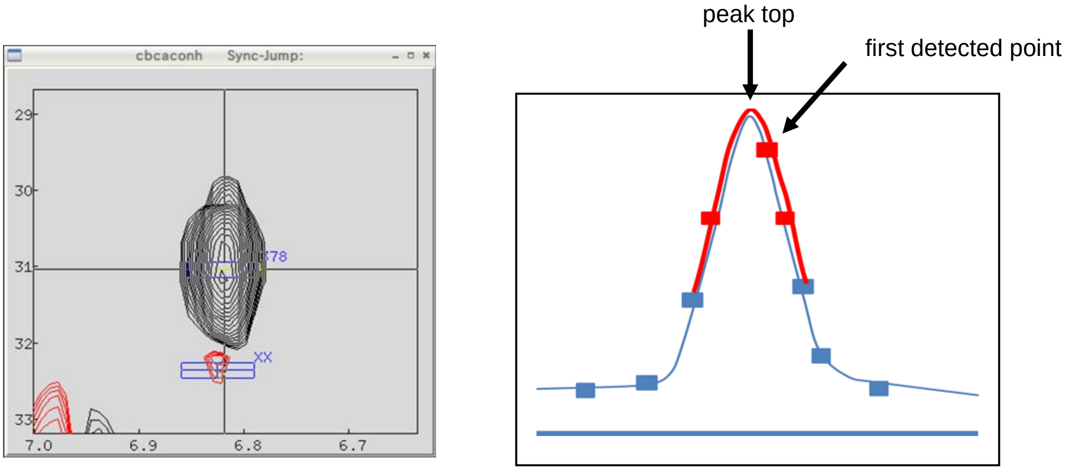

Fig. S5 2D plane of a 3D CBCA(CO)NH spectrum (left), and schematic representation of the estimation of the peak position (right) using quadratic interpolation algorithm with three data points (red).

HSQC-masked filter stage

Two-dimensional HSQC spectra are normally acquired to confirm the peak positions in 3D spectra, just like a projection image. Therefore 2D HSQC peak lists can be useful for roughly filtering noise peaks. For the benchmark, 2D peak lists were used with a tolerance of 0.04 ppm for ^{1}H, and 0.30 ppm for ^{13}C and ^{15}N dimensions. For ^{13}C-edited NOESY-type experiments only, the signals near the water signal in the indirect dimension (*y*-axis) are filtered because they are considered to originate from water exchanging with the target molecule (with 0.10 ppm tolerance). The user can set the filter tolerance values for the *x*- and *z*-axis and also set the position of the water signal (see instructions of the tool). If the user wants to apply custom filter data, the *_mask.xpk file in the job directory will have higher priority for using this filter process. The filtered peaks lists are labeled with *_filt.xpk.

CNN → ZNN filter stage

This stage using CNN filter as shown in Fig 1 of main text followed by ZNN as shown in the Fig S3. The final decision routine as shown in Fig. S4 will export the peaks lists labeled with *_cnn.xpk FLYA module will convert the peak lists using the priority order of file label; auto < filt < cnn. The GUI has options which can allow the user to select CNN and ZNN neural network models and ZNN cut-off value (default is 0.6).

Benchmarks

Benchmark 1 is for a small protein for which time-domain data are available from the BioMagResBank (BMRB) as bmr16647 (Ramelot *et al.*, to be published) were obtained. The 2D and 3D spectroscopic data sets selected (ten spectra) for this benchmark are shown in Table 1 and S1. Spectra were obtained from the time-domain data with NMR-Pipe 2010 (Delaglio *et al.*, 1995) using macro files provided along with archived data, without any modification except for conversion of the NMR-Pipe format to NMR-View (Johnson *et al.*, 1994) or Sparky UCSF format. The protein is an SH3 domain named CpR74A with 74 residues that was uniformly labeled with ^{13}C and ^{15}N. The corresponding NMR structure coordinates are available as PDB-ID 2KRS from wwPDB. The spectroscopic data were acquired using a 600 MHz Varian INOVA instrument for 2D ^{1}H-^{13}C HSQC for aromatics (constant time), and 3D HNCO, CBCA(CO)NH, HNCACB, and HCCH-TOCSY for aliphatics. 2D ^{1}H-^{15}N HSQC, ^{1}H-^{13}C HSQC for aliphatics, 3D ^{15}N-edited NOESY and ^{13}C-edited NOESY for aliphatics and aromatics were obtained using an 850 MHz Bruker Avance III instrument. The assignment completeness levels obtained by the authors are as follows:

	Total	^{1}H	^{13}C	^{15}N
All	90.5% (774/855)	90.5% (399/441)	90.9% (299/329)	89.4% (76/85)
Backbone	91.2% (395/433)	90.5% (133/147)	91.5% (195/213)	91.8% (67/73)
Aliphatic	90.1% (439/487)	90.5% (266/294)	90.6% (164/181)	75.0% (9/12)
Aromatic	69.2% (54/78)	69.2% (27/39)	68.4% (26/38)	100.0% (1/1)

(for more details about counting policies, see our web page for bmr16647: http://bmrbdep.pdbj.org)

Table S1. Benchmark results of bmr16647 for fully automated peak identification.

spectrum name	auto[b]	filter[c]	CNN[d]	manual[e]	FP[f]	FN[g]	Recall[h]	Precision[i]	F-value[j]
^{1}H-^{13}C HSQC AR[a]	136	--	23	21	2	0	100.0	73.4	95.2
HNCO	1,152	891	122	91	25	2	98.0	79.5	87.8
^{13}C NOESY AR[a]	960	320	170	157	10	1	99.4	94.1	96.7

See the other spectrum data in Table 1 of main text. a: AL and AR mean aliphatic and aromatic resonance regions, respectively. b: number of peaks obtained by quick peak identification with automated threshold optimization, as described in this document. c: number of peaks after HSQC-masked filtering only for 3D peak lists. d: number of peaks after CNN noise filtration applied to peak lists (named *auto.xpk) for 2D and 3D peak lists (named *_filt.xpk). e: number of peaks optimized by manual operation by experienced NMR scientist. f: number of false positive peaks identified as real peaks that were noise. Tolerances for chemical shift positions of peaks are 0.03 ppm for ^{1}H and 0.4 ppm for ^{15}N and ^{13}C atoms. g: number of false negative real peaks eliminated by CNN noise filtration. h: Recall values (%) were calculated as TP/(TP + FN). TP (true positive) is the number of correctly identified peaks. i: The precision value was calculated as TP/(TP + FP). j: F-values (%) were calculated as (2 × Recall × Precision)/(Recall + Precision). Not including diagonal peaks for NOESY and TOCSY type spectra for the F-value calculation.

The table S2 shows the results of peak identification using manually optimized threshold for each spectrum. The scores for Recall and F-value were worse than those determined by automatic manner. There are the weak but real peaks in the peak tables were failed to be identified. A too large number of FP peaks will lead to wrong assignments even though FLYA and CYANA are tolerant against noise peaks.

Table S2. Benchmark results of bmr16647 for automated peak identification with manually optimized threshold values.

spectrum name	auto[b]	filter[c]	manual[e]	FP[f]	FN[g]	Recall[h]	F-value[j]	Threshold ratio[k]
^{1}H–^{15}N HSQC	79	--	173	1	1	98.7	98.7	0.07
^{1}H–^{13}C HSQC AL[a]	452	--	815	189	9	96.7	72.7	0.15
^{1}H–^{13}C HSQC AR[a]	104	--	83	58	0	100	61.3	0.26
HNCO	266	197	91	82	14	89.1	70.6	0.22
CBCA(CO)NH	286	243	136	110	6	95.7	69.6	0.15
HNCACB	388	347	236	123	13	94.5	76.7	0.54
HCCH-TOCSY AL[a]	8,860	1,702	651	1,089	32	94.2	52.1	0.77
^{15}N NOESY	2,880	1,472	734	178	9	99.3	87.9	0.37
^{13}C NOESY AL[a]	12,951	2,997	1,636	1,106	194	90.7	74.4	0.31
^{13}C NOESY AR[a]	960	526	157	250	3	98.9	68.6	1.31

a-j: same as indicated in Table S1. k: the ratio of threshold value for each spectra (auto / manual).

Threshold values for the spectra were determined automatically; the total number of peaks automatically detected was 46,532, including bulk water noise signals. The HSQC-masked filter eliminated a large number of noise signals, giving a total of 15,865 peaks for the 3D spectra. Filt_Robot with the CNN noise filter system further eliminated noise peaks to give 4,754 peaks for all spectra. Figs. S6 and S7 show the peaks in the spectrum window in MagRO-NMRView.

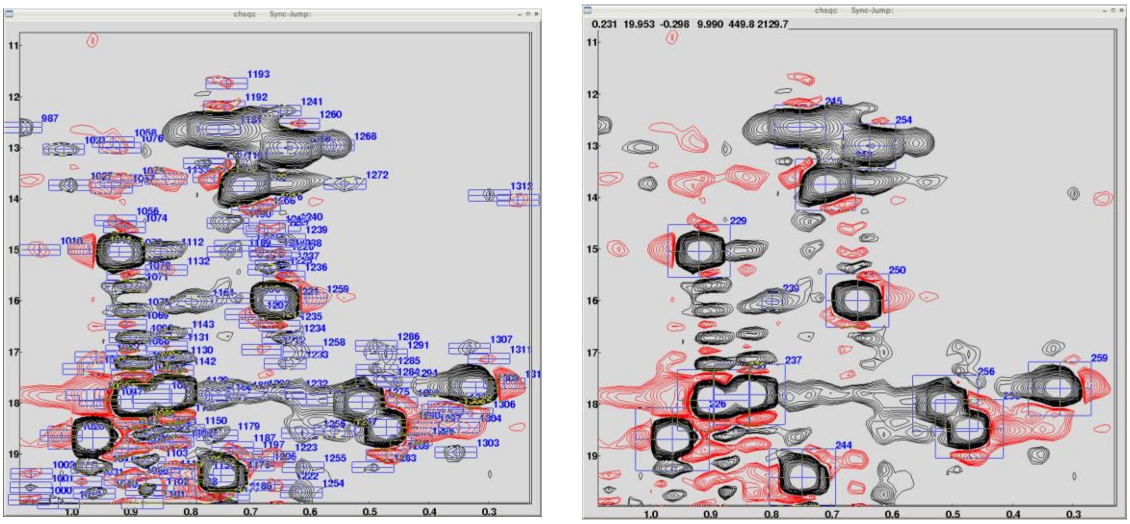

Fig. S6 2D ^{1}H-^{13}C HSQC for aliphatic spectrum derived from bmr16647. Fully automated peak picking before (left) and after CNN noise filtration (right).

Benchmark 2 was performed for a protein with 147 amino acid residues. The 2D and 3D spectroscopic data sets (eleven spectra) for these benchmark are shown in Table S3. Spectra were processed in the same way as for benchmark 1. The protein is a mutant of a human nuclear lamin domain (Lamin-G465D) which was expressed and purified uniformly labeled with stable isotopes ^{13}C and ^{15}N by standard methods as mentioned above for ubiquitin. NMR and X-ray structure coordinates of the corresponding wildtype protein are available as PDB-ID 1IFR and 1IVT from wwPDB.

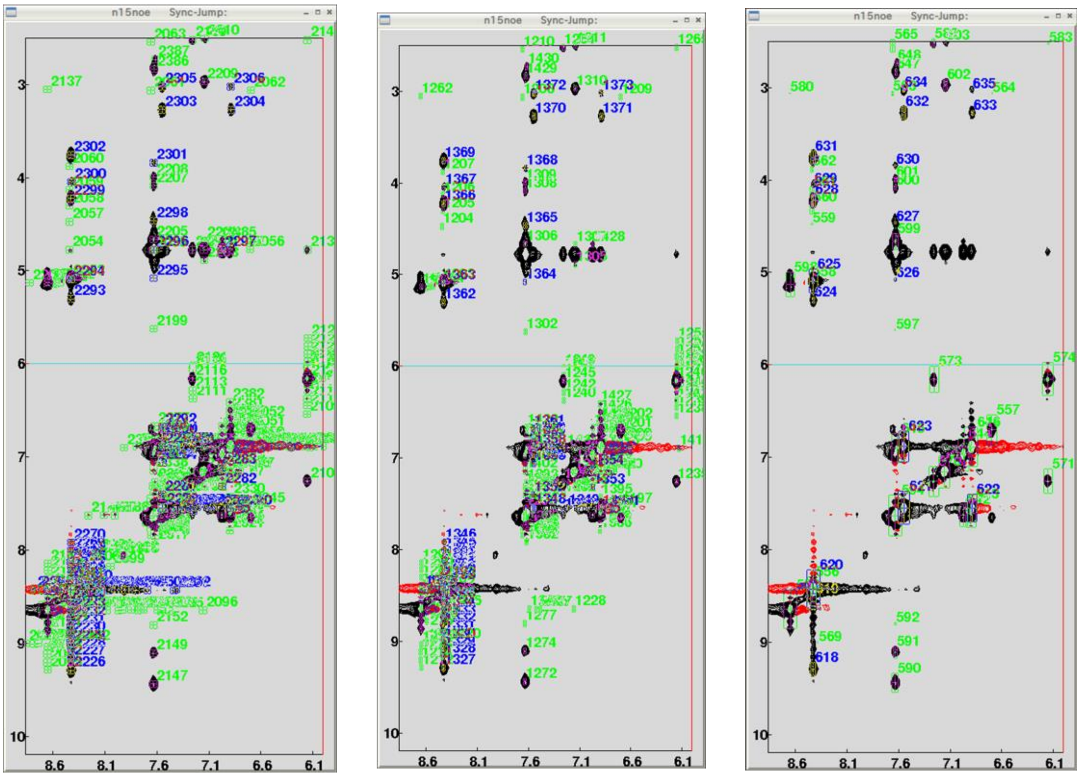

Fig. S7 2D slice of 3D ^{15}N-edited NOESY derived from bmr16647. Left panel shows peak boxes identified using fully automated method. Middle panel shows peaks after filtering with HSQC-masked filter. Right panel shows peak boxes after CNN noise filtering. Note that peaks derived from chemical exchange with water were automatically eliminated on the *y*-axis (*x*-axis:HN, *y*-axis:^{1}H NOE, 2D-slice on a certain *z*-axis point: ^{15}N). The blue boxes are peaks just on the *z*-plane while green boxes on the close *z*-slices (3-data points).

The spectroscopic data were acquired using a 800 MHz Bruker Avance III instrument for 2D ^{1}H-^{15}N HSQC and 3D HNCO, CBCA(CO)NH, HNCACB, and HCCH-TOCSY for aliphatics and aromatics. 2D ^{1}H-^{13}C HSQC for aliphatics and aromatics, 3D ^{15}N-edited NOESY and ^{13}C-edited NOESY for aliphatics and aromatics were obtained using an 950 MHz Bruker Avance III instrument (sample concentration was ~0.5 mM of protein in 30 mM potassium phosphate, pH 6.3 and 3 mM dithiothreitol, measured at 298 K). All the NMR machines are equipped with QCI cryo-probe ^{1}H/^{13}C/^{15}N/^{31}P.

Table S3. Benchmark results of Lamin-G465D for fully automated peak identification.

spectrum name	auto[b]	filter[c,k]	CNN[d,l]	manual[e]	FP[f]	FN[g]	Recall[h]	Recall[i]	F-value[j]
^{1}H-^{13}C HSQC AR[a]	120	83	41	43	43	0	100.0	48.2	65.0
HNCO	2,753	1,932	175	152	24	1	99.3	86.3	92.4
HCCH-TOCSY AR[a]	1,503	218	64	64	2	2	94.3	96.4	96.4
^{13}C NOESY AR[a]	3,439	459	284	312	4	17	94.3	96.4	96.4

See the other spectrum data in Table 1 of main text. a-j: same as indicated in Table S1. k: Filtered positive signals only and slightly wider tolerance for HSQC-masked filter (tolerances *x*-axis: 0.04, *z*-axis: 0.5 and water: 0.05 ppm). l: "strong" option was not applied except for 2D spectra. The Recall was bad for 2D ^{1}H-^{13}C HSQC for aromatic because of a large number of artifacts derived from the aliphatic signals (manual peak list did not contain them).

Threshold values for the spectra were determined automatically; the total number of peaks automatically detected was 108,957, including bulk water noise signals.

The HSQC-masked filter eliminated a large number of noise signals, giving a total of 33,676 peaks for the 3D spectra. Filt_Robot with the CNN noise filter system further eliminated noise peaks to give 10,436 peaks for all spectra. Figs. S8, S9 and S10 show the peaks in the spectrum window in MagRO-NMRView.

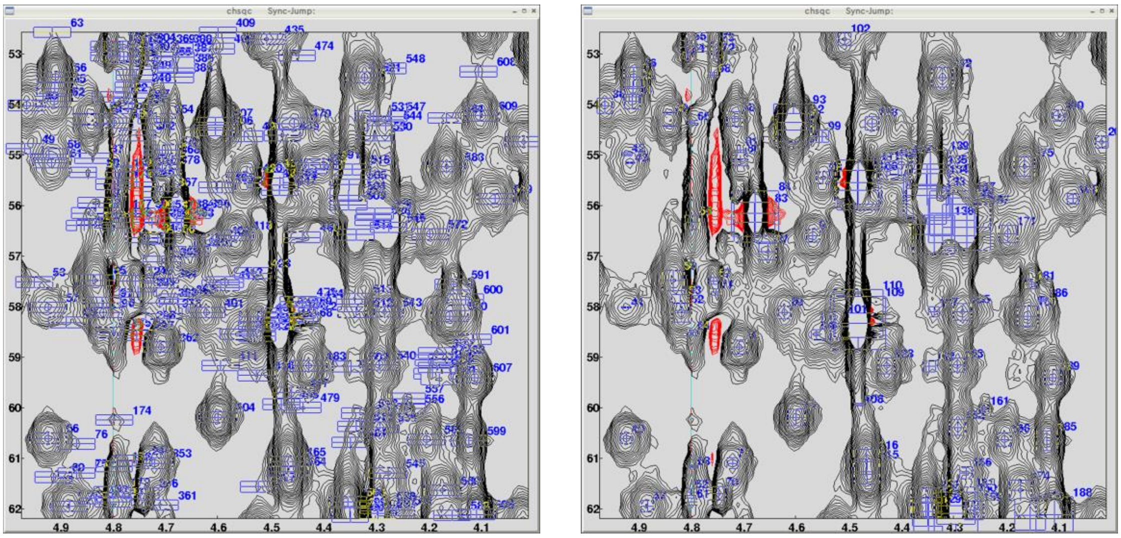

Fig. S8 2D ^{1}H-^{13}C HSQC for aliphatic spectrum derived from Lamin-G465D. Fully automated peak picking before (left) and after CNN noise filtration (right).

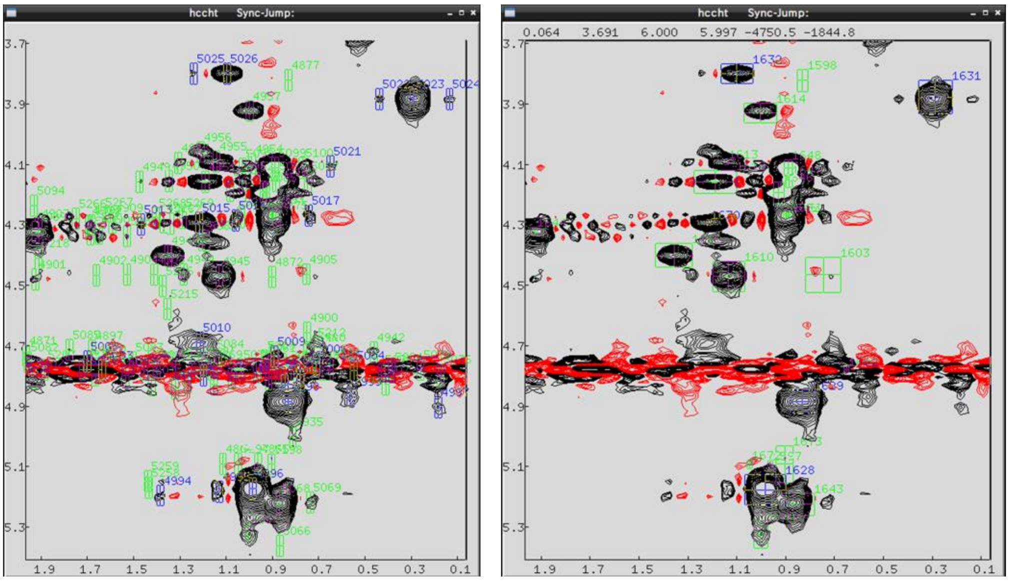

Fig. S9 3D HCCH-TOCSY for aliphatic spectrum derived from Lamin-G465D extracted on 21.36 ppm of *z*-axis (^{13}C). After HSQC-masked filter (left) and after CNN noise filtration (right). A large number of noise peaks from bulk water were eliminated by CNN-filter.

The filtered peak lists were converted with Filt_Robot to XEASY peak list format. Using the filtered peak lists, FLYA calculations including three stages of structure calculations (Schmidt and Güntert, 2012; Schmidt and Güntert, 2013) were performed with CYANA version 3.98 for fully automatic signal assignment and structure determination. The chemical shift tolerances were set to 0.03 ppm for ^{1}H, 0.4 ppm for ^{13}C, and 0.4 ppm for ^{15}N atoms.

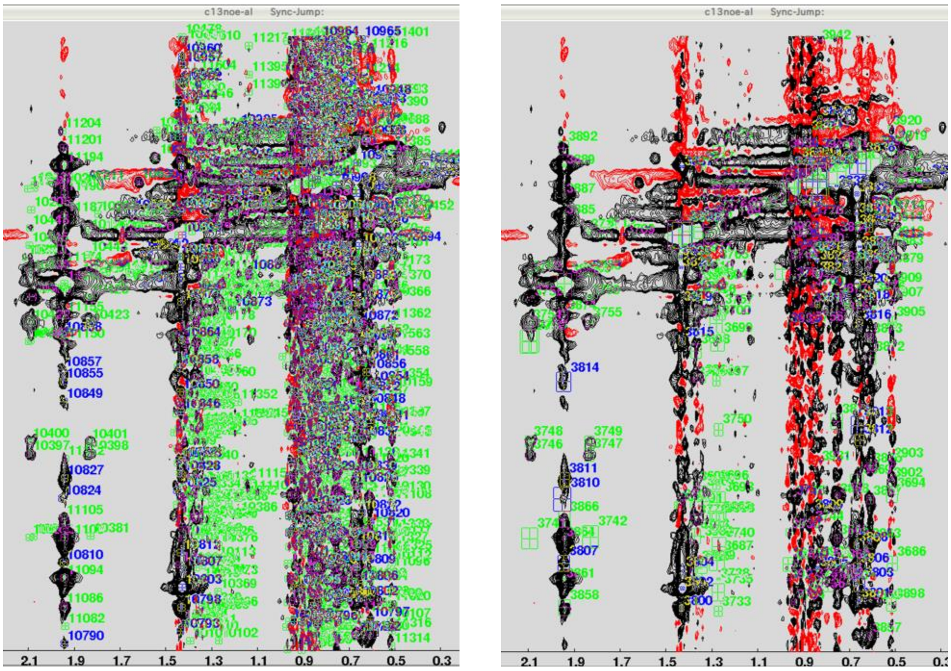

Fig. S10 3D ^{13}C-edited NOESY for aliphatic spectrum derived from Lamin-G465D extracted on 21.02 ppm of *z*-axis (^{13}C). HSQC-masked filter (left) and after CNN noise filtration (right). The remarkable point is there are seriously large number of noise peaks on the methyl region but CNN-filter can correctly take real peaks.

The calculation was performed using a 20-core Xeon E5-4627 (2.6-3.0 GHz) CPU, and took 60~90 min. The completeness and accuracy of the FLYA resonance assignments (labeled as strong by FLYA) were as follows:

Assigned signals	**Completeness**	**Accuracy**
[Benchmark 1: bmr16647]		
Backbone (HN-^{15}N)	98.5% (66/67)	100.0% (66/66)
Aliphatic side-chains	95.0% (452/476)	98.7% (446/452)
Aromatic side-chains	100.0% (32/32)	96.9% (31/32)
[Benchmark 2: Lamin-G465D]		
Backbone (HN-^{15}N)	92.8% (129/139)	98.4% (127/129)
Aliphatic side-chains	93.2% (928/996)	97.8% (908/928)
Aromatic side-chains	92.4% (61/66)	95.1% (58/61)

For the above assessment of the backbone signals, the pairing of HN-^{15}N signals was examined. The chemical shift error tolerances were set at 0.05, 0.40, and 0.40 ppm for ^{1}H, ^{15}N, and ^{13}C, respectively. The better results were selected for the assessments of swappable atoms, for example Ser-HB2/HB3 and Val-CG1/CG2.

After the FLYA calculation, Filt_Robot imported the chemical shift assignments that were labeled as strong (i.e. reliable) by FLYA to produce a TALOS+ (Shen *et al.*, 2009) input file. A TALOS+ calculation was performed to obtain restraints for the backbone dihedral angles (ϕ/ψ) with a minimal width of 10º. Using Filt_Robot, the files required for the CYANA calculation (Güntert and Buchner, 2015) were prepared using the CNN-filtered peak lists for 3D ^{15}N-edited NOESY and ^{13}C-edited NOESY for aliphatics and aromatics, and the TALOS+ dihedral angle restraints.

To obtain the structure, we ran a CYANA calculation on the same computer as was used for the FLYA calculations. The proline residue having cis-configuration can be predicted from the chemical shits of Cβ and Cγ (Schubert *et al*., 2002), the strongly probable one is automatically labeled with cPRO in the CYANA sequence file. The chemical shift tolerances were set to 0.03 ppm, 0.4 ppm, and 0.4 ppm for ^{1}H, ^{13}C, and ^{15}N atoms, respectively. One-hundred structures were calculated by CYANA, including seven cycles and a final stage using simulated annealing with 10,000 dynamics steps to give 20 final structural models.

The ordered regions in the structure ensemble of bmr16647 were automatically identified and overlaid using FitRobot (Kobayashi, 2014). The RMSD to the mean coordinates for the backbone atoms (N, C$^{\alpha}$ and C' at residues 1-65) was 0.40 Å, whereas the backbone RMSD for the original NMR structure with PDB-ID 2KRS corresponding to the NMR data of bmr16647 was 0.4 Å (residues 2–61), as shown in Fig. S11.

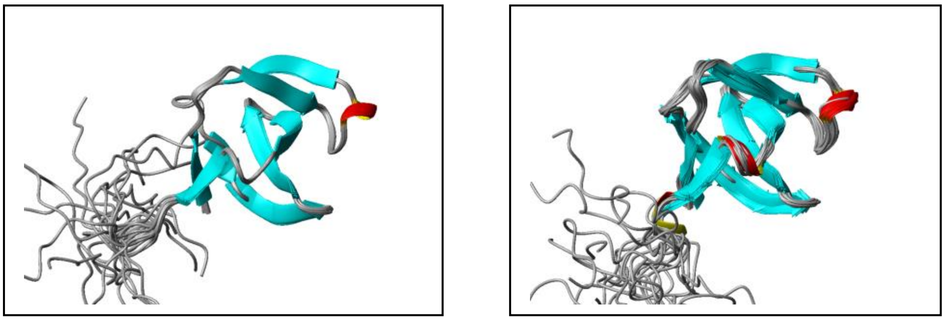

Fig. S11 Ribbon representations of overlaid models (bmr16647) determined using our fully automated method (left), and the original NMR structure determination 2KRS (right).

The averaged RMSD bias of all determined NMR structure ensemble against mean coordinates of 2KRS was 1.01 ± 0.04 Å. The FLYA, TALOS+ and CYANA calculations were similarly performed using the peak lists as shown in Table S2, revealing high completeness (95.1% for all atoms) and accuracy (97.2% for all atoms) of chemical shift assignments. However the CYANA structure was showing the averaged RMSD bias 1.80 ± 0.02 Å indicating that the NOE peak lists were improved by the noise elimination with the CNN filter. The major reason of this discrepancy was the large number of FP and FN peaks in NOE peak lists identified even with manually optimized threshold value. The number of missing and artifacts appearing on NOE dimension in 3D NOESY spectra was obviously beyond the error tolerance of automated NOE assignment by CYANA (Buchner and Güntert, 2015). In spite of the fact, the results support that the CNN filter with relatively lower threshold can be suitable for automated chemical shit assignments and structure calculation by FLYA and CYANA.

The structure ensemble of Lamin-G465D calculated by CYANA was overlaid on secondary structure regions (436-437, 441-453, 464-473, 477-482, 489-487, 495-507, 511-543) as shown in Fig. S12, then compared with the structures determined by X-ray crystallography (PDB-ID 1IFR) and manually refined NMR data to bet RMSD bias for the backbone atoms (N, C$^{\alpha}$ and C') on the secondary structure regions using MolMol (Koradi *et al*., 1996). The averaged RMSD bias for X-ray and NMR structures were 0.97 ± 0.04 Å and 1.10 ± 0.03 Å, respectively, indicating that the mutation of Gly-465 to Asp may not have led to a large structural change. Interestingly the old NMR structure published by Krimm et al., (Krimm *et al*., 2002) showed a larger RMSD bias, 2.02 ±0.02 Å. This may indicate how difficult it was to correctly determine the solution structure of human Lamin in the time when the original structure was published, and demonstrates how the new strategy mentioned in this report will strongly assist automated NMR spectrum analysis.

Again very high Recall values are a key point for the automation, i.e. the filter system has a high capability not to eliminate real peaks. Without the spectrum data for connecting the information from backbone assignments to side-chain signals such as 3D H(CCO)NH, C(CO)NH and HBHA(CO)NH, the assignment process would be extremely difficult even for manual operation. In the benchmark 2, there are also many noise peaks derived from the sharp diagonal peaks.

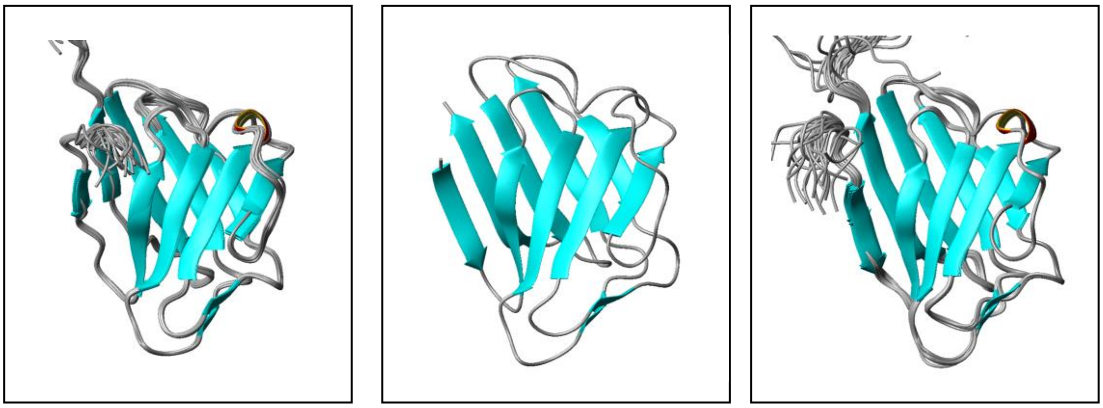

Fig. S12 Ribbon representations of overlaid models (Lamin-G465D) determined using our fully automated method (left), X-ray structure (middle) and the NMR structure determination by manual operation (right). Disordered regions (residues 401-425) are omitted.

All of the benchmark data are available from our web-site, including FID data, NMR-Pipe scripts, and processed spectrum data sets in NMR-View and Sparky formats. Also the demo package including the calculated FLYA, TALOS+ and CYANA data for reproducing the capability of CNN filter system. The virtual-box image (ova file, ~2.6GB) is available to reconstruct the whole Ubuntu virtual machine including CNTK and OpenMPI for getting ready to try the demo data quickly.

Summary

In conclusion, the fully automated manner of peak identification from limited spectrum data sets is strongly expedited by the CNN noise filter to get correct NMR structures quickly. This system may work with other peak pickers such as AUTOPSY (Koradi *et al*., 1998), PICKY (Alipanahi *et al*., 2009), WavPeak (Liu *et al*., 2012), CV-Peak picker (Klukowski *et al*., 2015), INFOS (Smith, 2017) and CYPICK (Würz and Güntert, 2017) for the initial stage of peak identification if the user can convert the peak lists and spectrum data into NMR-View format. CNN is a well matured technology which has been widely used for image recognition. For example one of the most remarkable applications of CNN could be found in magnetic resonance imaging (MRI). Very recently CNN was applied to filtering artifacts in MRI data (Gurbani *et al*., 2018). Although the high robustness of our system, we would recommend the user to carefully inspect the final peak lists on spectrum viewer as there must be a few percent of FP and FN peaks in especially the case for the study relying on limited spectrum data sets such as HSQC perturbation. Especially the peak separation and alignment of HSQC projection are critical on each stage of the automated strategy. Nevertheless the system can eliminate more than 99% of the noise peaks estimated from row peak lists, indicating that the methods described in this report will reduce the tediousness for handling NMR peaks and is suitable for the automated NMR analysis in structural genomics.

References

Alipanahi,B. *et al*. (2009) PICKY: a novel SVD-based NMR spectra peak picking method. *Bioinformatics,* 25, 268–275.

Buchner,L. and Güntert,P. (2015) Systematic evaluation of combined automated NOE assignment and structure calculation with CYANA. *J. Biomol. NMR.*, 62, 81-95.

Delaglio,F. *et al*. (1995) NMRPipe: a multidimensional spectral processing system based on UNIX pipes. *J. Biomol. NMR*, 6, 277–293.

Güntert P. and Buchner L. (2015) Combined automated NOE assignment and structure calculation with CYANA. *J. Biomol. NMR*, 62, 453–471.

Gurbani,S.S. *et al*. (2018) A convolutional neural network to filter artifacts in spectroscopic MRI. *Magn. Reson. Med.,* in print.

Johnson,B.A. and Blevins,R.A. (1994) NMR View: A computer program for the visualization and analysis of NMR data. *J. Biomol. NMR,* 4, 603–614.

Klukowski,P. *et al*. (2015) Computer vision-based automated peak picking applied to protein NMR spectra. *Bioinformatics,* 31, 2981–2988.

Kobayashi,N. *et al*. (2007) KUJIRA, a package of integrated modules for systematic and interactive analysis of NMR data directed to high-throughput NMR structure studies. *J. Biomol. NMR.,* 39, 31-52.

Kobayashi,N. (2014) A robust method for quantitative identification of ordered cores in an ensemble of biomolecular structures by non-linear multi-dimensional scaling using inter-atomic distance variance matrix. *J. Biomol. NMR*, 58, 61–67.

Koradi,R., *et al*. (1996) MOLMOL: a program for display and analysis of macromolecular structures. *J. Mol. Graph.,* 14, 51–55.

Koradi,R. *et al*. (1996) Automated peak picking and peak integration in macromolecular NMR spectra using AUTOPSY. *J. Magn. Reson.*, 135, 288–297.

Krimm,I. *et al*. (2002) The Ig-like structure of the C-terminal domain of lamin A/C, mutated in muscular dystrophies, cardiomyopathy, and partial lipodystrophy. *Structure*, 10, 811-823.

Liu,Z. *et al*. (2012) WaVPeak: picking NMR peaks through wavelet-based smoothing and volume-based filtering." *Bioinformatics,* 28, 914–920 (2012).

Schmidt,E. and Güntert,P. (2012) A new algorithm for reliable and general NMR resonance assignment. *J. Am. Chem. Soc.*, 134, 12817–12829.

Schmidt,E. and Güntert,P. (2013) Reliability of exclusively NOESY-based automated resonance assignment and structure determination of proteins. *J. Biomol. NMR,* 57, 193–204.

Schubert,M. *et al*. (2002) A software tool for the prediction of Xaa-Pro peptide bond conformations in proteins based on C-13 chemical shift statistics. *J. Biomol. NMR,* 24, 149–154.

Shen,Y. *et al*. (2009) TALOS+: A hybrid method for predicting protein backbone torsion angles from NMR chemical shifts. *J. Biomol. NMR*, 44, 213–223.

Smith,A.A. (2017) INFOS: spectrum fitting software for NMR analysis. *J. Biomol. NMR* 67, 77–94.

Würz,J.M. and Güntert,P. (2017) Peak picking multidimensional NMR spectra with the contour geometry based algorithm CYPICK. *J. Biomol. NMR,* 67, 63–76.

Acknowledgements

We thank Dr. H. Tamaki and Profs. Y. Miyanoiri and T. Sugiki at Osaka University for fruitful discussions about deep learning and peak identification methods, and Mr. T. Iwata for preparing the web-site for the tool and demo data, and Profs. S. Sakuraba at Kyoto University and A. Kinjo at Osaka University, and T. Ikeya at Tokyo Metropolitan University, Mr. T. Kengo, Drs. T. Yamazaki, T. Kasai, and T. Kigawa at RIKEN Yokohama Institute, T. Sakamoto and G. Kawai at Chiba Institute of Technology, Y. Muto at Musashino University and Y. Yoshimura and S.I. Tate at Hiroshima University and T. Ikegami at Yokohama City University for various discussions and suggestions. We are also grateful to BioMagResBank (BMRB) for providing NMR spectroscopic data. The authors also especially thank the Microsoft Corp. for distributing CNTK with a MIT license.