Supplementary Methods – R-MDD theory and experimental details

**Calculation of measurement time for the case of uniformly sampled experiment.** The total duration of an M-dimensional NMR experiment is defined by three factors: (i) $T_{\text{FID}}$ - time required per single transient of a one-dimensional spectrum (or FID), times the number of transients in a phase cycle used for eliminating spurious signals; (ii) $N_m$ - number of sample points in each of M indirect dimensions; (iii) $N_p=2^{M-1}$ - the number of FID’s required for phase discrimination in the indirect dimensions. Without losing generality of this presentation we assume equal number of points $N_m=N$ in all indirect dimension. Thus, the time in seconds of a conventional experiment with uniformly sampled points can be expressed as following:

$$T^{\text{DFT}}(M, N_m) = N^{M-1} 2^{M-1} T_{\text{FID}}$$  \hspace{1cm} (S1)

**The general MDD model.** Multidimensional decomposition (MDD) model assumes that all essential features of M-dimensional matrix can be described as a sum of small number of tensor products of one-dimensional vectors $1,2$. MDD has been used in different fields as a tool for data analysis and signal processing since the early seventies under various names such as the parallel factor analysis, canonical decomposition, and three-way decomposition. When applied to NMR spectra the MDD can be formulated as follows. Given a matrix $S$ with sizes $N_m$ of its M dimensions $(m=1...M)$ and elements $S_{n_1,n_2,...,n_M}$ , find scalar numbers $\beta a$ and normalized vectors $\beta F_m$, with elements $\beta F_m(n_m)$ $(n_m=1...N_m)$, such that the following norm becomes minimal.

$$| G \bullet [S - \Sigma_\beta (\beta a \beta F^1 \otimes \beta F^2 \otimes \cdots \otimes \beta F^M)]^2 + \lambda \Sigma_\beta (\beta a)^2$$  \hspace{1cm} (S2)

Here, the symbol $\otimes$ denotes tensor product operation; the matrix $S$ corresponds to experimental M-dimensional NMR spectrum in time and/or frequency domain representation. In the case of sparse sampling only a fraction of elements in $S$ is measured and the matrix $G$, which contains elements $g_{n_1,n_2,...,n_M}\in\{0,1\}$, indicates the absence or presence of a particular data point. Accordingly, the symbol $\bullet$ describes element-wise multiplication of matrices. The last term represents a Tikhonov regularization $3$, which is parameterized with the factor $\lambda$ and may be used for improving the convergence of the MDD algorithm $4$. The summation index $\beta$ runs over the number of components used for the decomposition. The range for this index depends on the type of spectrum. For example, for a 3D HNCO spectrum it is roughly equal to the number of protein amide groups.

Each vector $\beta F^m$ is defined by $N_m$ unknown elements. We make an estimate of measurement time needed for MDD reconstruction from the notion that number of measurements should exceed
number of adjustable parameters with a certain redundancy. We also need to make an assumption about maximal number of peaks that have the same position in the directly detected dimension. Altogether, the redundancy and signal overlap give a factor, which is approximately equal to half of the number of peaks in the spectrum. Finally, the measurement time is estimated as:

$$T^{MDD}(M, N_m) = N_c (M-1) N_m T_{FID}$$  \hspace{1cm} (S3)$$

where $N_c$ is number of spectral signals. For example, for HNCO spectrum of ubiquitin $N_c$ is ca 70.

**The Recursive MDD model.** Each of the time domain shapes $\beta F^m$ of length $N_m$ in Eq. S2 is represented as a product of $K_m$ vectors of length $d_m$ so that $N_m = (d_m)^{K_m}$:

$$\beta F^m = \beta V^d \otimes \beta V^2 \otimes \cdots \otimes \beta V^K$$  \hspace{1cm} (S4)$$

This can be interpreted as applying second decomposition to the shapes obtained in the first decomposition, thus the name “recursive”. To rationalize the additional decomposition of each vector $F$ into a set of subvectors $V^k$, (we can omit indices $m$, for simplicity), we shall illustrate the recursive decomposition for $d=2$. The points of exponential decay function $E(t)=\exp(-t)$ sampled at uniformly spaced time points $t=n\Delta t$, $(n=0,\ldots,N)$, constitute vector $E=\{\exp(-\Delta t)^0, \exp(-\Delta t)^1, \ldots \exp(-\Delta t)^{N-1}\}$ of length $N$. The vector can be “folded” into a hypercube with $N$ elements, where each side has length 2. The hypercube can be decomposed into $K=\log_2 N$ vectors $V^k = \{1, \exp(-\Delta t)^k\}$ of length two

$$E=\{1, \exp(-\Delta t)^1\} \otimes \{1, \exp(-\Delta t)^2\} \otimes \{1, \exp(-\Delta t)^4\} \cdots \}$$  \hspace{1cm} (S5)$$

Note that Eq. S4 is a generalization of Eq. S5, when $d$ is any integer number and vector elements are not restricted to integer powers of $\exp(-\Delta t)$. The aim of the second MDD decomposition is to reduce number of unknowns in the least-square minimization of Eq S2. This is achieved by replacing vectors $\beta F^m$ for all $\beta$ in Eq. S2 by the right side of Eq. S4. This merely converts $M$-dimensional decomposition to $(M-1+K_m)$-dimensional one. In case of phase sensitive detection, the complex vector $\beta F^m$ with $2N_m$ unknown elements is defined by much smaller number of real parameters $2(d-1)*\log_d (2N_m)$. Finally, using the same arguments as for Eq. S3, total measurement time needed for robust $R$-MDD reconstruction of the spectrum can be estimated as,

$$T^{RMDD}(M, N_m) = N_c (M-1) (d-1) \log_d (2N_m) T_{FID}$$  \hspace{1cm} (S6)$$

This equation with $N_m=N$ is plotted for $M=2,3,4,5$ as solid curves in Figure 1 of the paper using $d=2$, $T_{FID}=4$ sec, and $N_c=70$ (number of backbone amides in ubiquitin). The curves exhibit logarithmic dependences on number of increments in the indirect dimensions, $N_m$. The exponential dependence $(N_m)^{M-1}$ on the number of dimensions $M$ in Eq. S1 is reduced to linear term $(M-1)$ in
Eq. S6. From the plot it is clear that all of the experimental times needed for the MDD 2D-5D spectral processing are below 5 hours, which is generally the 2D range of standard Fourier spectroscopy.

Eq. S6 provides the following enhancement in time saving relative to eq. S3:

\[ \frac{T_{MDD}(N_m)}{T_{RMDD}(N_m)} = \frac{N_m}{(d-1) \log_d(2N_m)} \]  

(S7)

The ratio increases with larger \( N_m \) (e.g. equals 13 for \( N_m = 100 \)). It does not depend on spectra dimensionality and number of components.

Equations S4 – S7 were derived in the assumption that the recursion, and correspondingly the autoregressive constraint on the time domain line shapes, is applied in all indirect dimensions. However, the constraint can be used in a subset of dimensions. Example is the 3D 15N NOESY experiment, where for reducing number of components and preserving sensitivity, all peaks belonging to one amide group are collected in one MDD component\(^5\). In this case, the recursion can be applied only to the 15\( ^N \)N dimension. Thus, although NOESY type spectra will also profit from R-MDD, the expected timesaving relative to original MDD in this case is moderate.

MDD is a true multidimensional signal processing method, as it analyses entire multidimensional data array simultaneously. An additional potential offered by the method is in its ability to process non-uniform data, when valuable spectrometer time is preferentially spent on obtaining information on signals of interest rather than noise. This insures highest sensitivity for given digital resolution and total measurement time. High sensitivity delivered by the MDD processing have been addressed specifically \(^6\) and demonstrated in practical cases \(^7\). A related technique \(^8\), which is designed for dealing with GFT projections, was also proved successful in identifying weak signals. Likewise, sensitivity of the R-MDD is reflected in demonstrated high accuracy of signal amplitudes, positions, and line shapes as well as ability of the method to deal with signals in wide range of intensities.

**Materials and Methods**

*NMR spectroscopy.* All spectra were recorded using room temperature triple resonance probes with pulse field gradients. The proteins were \(^{13}\)C, \(^{15}\)N labeled (except for barstar): ubiquitin\(^{14}\); barstar-barnase complex\(^{12,13}\); reduced azurin\(^{11}\). The key experimental parameters are listed (Supplementary Table 1).

*Sampling schedules.* Optimal non-linear sampling schedules are devised to match the envelopes of signal coherences in all indirect dimensions for a particular NMR experiment \(^9\). Since \(^{15}\)N chemical
shift evolves in all our experiments during a constant delay, the envelope in this dimension is a constant function. The envelope of the signals in $^{13}\text{C}$ dimension is the exponential decay with transverse relaxation time $T_2$ (Supplementary Table 1). The convolution of the two envelopes produces a two-dimensional probability density function in two acquisition dimensions $t_1$ and $t_2$. The sampling was randomly generated in accordance with the density function. For example, in the sampling schedule optimized for the 3D HNCA of azurin (Supplementary Figure 1) in addition to the $T_2$-decay, the sampling density is modulated by cosine function due to $J(^{13}\text{C}^{\alpha},^{13}\text{C}^{\beta})$-coupling of 35 Hz.

**Data processing and reconstruction** for all spectra were done using the same procedure. It is exemplified below for the azurin case. Initially time domain data along the directly detected dimension, $t_3$ of the reference spectrum were multiplied with a squared cosine-bell window function, zero-filled to 2K points, and Fourier transformed with nmrPipe package. The “truncated” data set was derived from the “reference” by truncation after initial 40 ($t_1$), 40 ($t_2$) points. The “sparse” data set was extracted from the “reference” according to the sampling schedule (Supplementary Figure 1) containing 1600 points ($t_1,t_2$). A complete time domain signal containing $2^9=512$ ($t_1$), $2^6=64$ ($t_2$) complex points was reconstructed from the sparse data set using the R-MDD procedure (Eqs. S2, S4) implemented in a homebuilt software mddNMR. Note that the final size in $^{13}\text{C}^{\alpha}$ dimension ($t_1$) in the reconstruction is 25% larger than that in the reference spectrum. This shows that R-MDD does not only fill gaps between the measured points, but also can effectively extrapolate time domain data. Application of the recursion (Eq. S4) to dimensions $t_1$ and $t_2$ but not to $t_3$ resulted finally in $9+6+1 = 16$ – dimensional decomposition in Eq. S2. The R-MDD calculation and reconstruction were performed individually for 48 overlapped 3D segments covering the range from 5.75 to 10.75 ppm ($w_3$-direct $^1\text{H}$), each corresponding to 32 frequency-domain points. The segmentation doesn’t affect the algorithm convergence and quality of reconstruction. It was done only to save computational time by parallel processing, since segments can be processed independently. To account for possible incomplete reconstruction of peak tails on the segments borders (in $w_3$) only the central 16 point parts were merged to obtain the whole spectrum reconstruction. The number of the MDD components for each of the 48 segments was estimated as following: the number of $^1\text{H}^N$-$^{15}\text{N}$ correlations in the corresponding $^1\text{H}$ strip of the $^{15}\text{N}$ HSQC spectrum was multiplied by 4 to account for the intra- and inter-residue duplets in the HNCA spectrum; the number was further increased by 30% to account for presence of possible minor peaks and some of the large noise features or spectral artefacts. After the reconstruction, which produced a regular 3D data set in the format of nmrPipe, the time domain signal was
multiplied with a squared cosine-bell window function, zero-filled to double size and Fourier transformed. This gave digital resolution of 0.029 ppm (5.17 Hz) for $^{13}\text{C}$ dimension. The peaks positions in the reference spectrum are determined using ‘pkFindROI’ subroutine of the nmrPipe package $^{10}$. The exact peak positions in all peak lists were refined by three-point interpolation. In the analysis the parameters of duplets were determined as average of two singlets. To check stability of the algorithm convergence we repeated the R-MDD calculation with the same experimental input but different initial approximation for the shapes $\theta V^K$ (Eq. S4). The result was essentially the same as for the first run described in the main text: the all-signal r.m.s.d. to reference was 0.0062 ppm in the CA dimension and the correlation of intensities was 0.998. These values for the difference between the two solutions were: 0.0036 ppm and 0.9985, respectively.

References