NMR Protein Structure Determination via Local Conformational Mapping

Mohammad ElGamacy, Michael Riss[†], Hongbo Zhu, Vincent Truffault and Murray Coles*

Dept. of Protein Evolution, Max-Planck-Institute for Developmental Biology, 72076 Tübingen, Germany

[†] Dept. of Informatics, Technical University of Munich, 85748 Garching, Germany

* Corresponding author Murray Coles:
Department of Protein Evolution
Max-Planck-Institute for Developmental Biology
Max-Planck-Ring 5
72076 Tübingen
Germany
Tel. +49 7071 601 451
Fax +49 7071 601 349
murray.coles@tuebingen.mpg.de

Running title: An R-factor in NMR structure determination

ABSTRACT

We present a generalized method for mapping protein population landscapes by NMR spectroscopy via direct comparison of experimental spectra to expectation spectra spanning an arbitrary conformational space. Signal decomposition of the experimental spectrum directly yields the relative populations of local conformational microstates with an attached R-factor. In this way, interpretation of the spectra in terms of peak picking or assignment is completely avoided. We demonstrate that our method extracts sufficient information from a single 3D NOESY experiment to perform initial model building, refinement and validation, thus offering a complete *de novo* structure determination protocol.

INTRODUCTION

Internal motions of protein are key determinates in their function. Understanding these processes requires accurate descriptions of protein structures and their conformational diversity. NMR spectroscopy is uniquely placed to address these issues, offering atomic-resolution data on samples in native-like physical states. Time averaging of NMR parameters has long been exploited to localise and characterize the timescales of internal dynamics [1]. However, the data is also ensemble-averaged over all molecules in the sample volume and should thus provide information on the nature and population of underlying conformational microstates. Such data is vital to understanding protein function at the molecular level, but are not readily accessible to conventional NMR structure determination methods [2-5].

Here we query the propensities of individual microstates by means of spectral decomposition. Systematic back-calculation of expectation spectra across a conformational space allows reconstruction of the experimental spectra. In NMR spectroscopy, the richest source of structural data are NOESY spectra, which report on inter-proton distances within a detection limit of 5 to 6 Å. Due to this short spatial range, a large fraction of NOESY intensity can be explained within short, linear sequence fragments. Each such fragment thus represents a sub-space that could be searched systematically to provide detailed information on local dihedral angles and their distributions. Moreover, comparison of the back-calculated and experimental data would provide a quantitative quality measure: an NMR R-factor.

A difficulty in realising this approach lies in the nature of NOESY data itself. The information content of NOESY spectra is very unevenly distributed across the observed intensities, thus small, informative peaks can be overwhelmed by inaccuracies in back-calculation and spectral artifacts. For this reason, quantitative comparison of back-calculated and experimental spectra has been far less applicable in NMR structure determination than equivalent measures used in crystallography [6-8]. Here we show that these obstacles can be largely averted in the 3D CNH-NOESY experiment [9]. This is an implementation of a ^{13}C-HSQC-NOESY-^{15}N-HSQC where the indirect proton dimension

has been omitted, thus displaying contacts to backbone amide protons in a well-resolved ^{13}C dimension. It exploits the higher dispersion and more homogeneous effective linewidths of the heteronuclei, while suppressing water-exchange cross-peaks and obviating the need for stereospecific proton assignments. Crucially, it intrinsically lacks large, uninformative diagonal peaks and the associated baseline and truncation artifacts. Combined, these represent decisive advantages in the accuracy of back-calculation.

In this work we demonstrate that a single 3D CNH-NOESY spectrum contains sufficient information to define population maps of local dihedral sub-spaces. Analytical decomposition expresses the experimental spectra as a linear combination of elements of a features set of back-calculated spectra. In this way, both the reliance on a knowledge-base and the interpretation of spectra in terms of peak or assignment lists can be eliminated. This conformational mapping provides highly detailed data for model building and refinement, with progress monitored by a quantitative R-factor. We validate this method against human Ubiquitin (hUb), widely considered the gold standard for NMR-based protein structure determination [5,10-12]. We further demonstrate the generality of the method by solving the structures of four example proteins.

RESULTS

An R-factor from CNH-NOESY data

We have adapted existing routines to back-calculate CNH-NOESY spectra, obtaining 1D ^{13}C strips for each backbone amide proton. These are compared directly to equivalent strips extracted from the experimental 3D matrix. An R-factor expressing the discrepancy between experimental and expectation spectra is readily calculated as the fractional root mean squared residual (see Materials and Methods). This R-factor is analogous to its crystallographic counterpart, except that it is calculated on a per-residue basis. Our back-calculation routines very accurately reproduce the intensities and line-shapes of experimental CNH-NOESY data collected for hUb (Fig. 1A). The back-calculated spectra are also highly sensitive to backbone and sidechain dihedral angles (Fig. 1B and Supplementary Fig. S1), a prerequisite for conformational mapping.

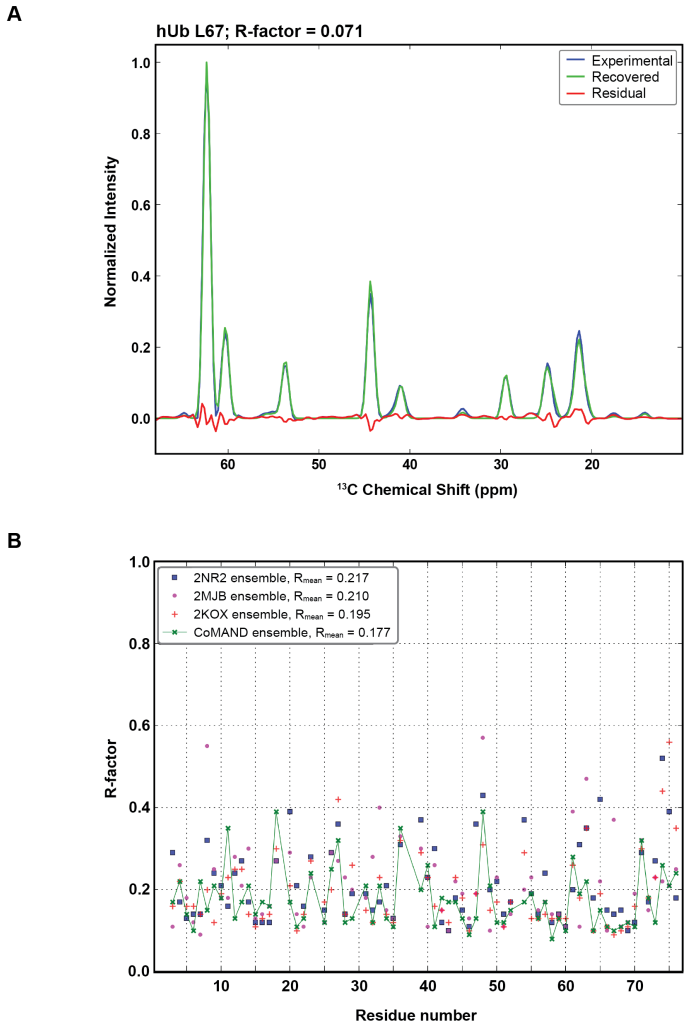

Figure 1. NOESY back-calculation in the CoMAND method. **A)** An example comparison demonstrating the quality of back-calculation of CNH-NOESY spectra. The experimental spectrum for L67 in human ubiquitin (hUb) is in blue and the recovered spectrum back-calculated by averaging across the 640 models of the 2KOX structure ensemble [11] is in green. The residual signal is in red (R-factor = 0.071). **B)** R-factors plotted across the sequence for three literature ensembles. These ensembles have been compiled to emphasise different aspects of the hUb structure: 2KOX to elucidate internal motions [11], 2NR2 via a minimal under-restraining, minimal over-restraining procedure [5] and 2MJB to represent a static average pose [12]. The average structures for all three ensembles are very similar and differences in R-factors are therefore attributable to the different representations of conformational diversity (see Supplemental Figure S1 for specific examples). These are compared to the CoMAND ensemble (green line).

NOESY intensities are time and ensemble averages over the conformational microstates sampled during the measurement. For this reason, R-factors improve with an accurate and comprehensive description of the ensemble. We demonstrate this for hUb using three reference ensembles that have been compiled according to different metrics. Two have been compiled to elucidate internal motions: the 2KOX ensemble using a large set of residual dipolar coupling (RDC) data [11], and the 2NR2 ensemble according to minimum under-restraining, minimum over-restraining criteria and including S_{NH}^2 order parameters derived from relaxation data [5]. The third, 2MJB, provides a static control ensemble close to the average structure [12]. High R-factors are obtained where these ensembles either over- or under-estimate the conformational diversity. Moreover, the use of a residue-wise R-factor in evaluating the ensemble can localize such diversity (Fig. 1B).

Mapping local conformational spaces

For the CNH-NOESY, the vast majority of cross-peak intensity can be explained by intra-residue contacts and those to the immediately preceding residue. The R-factor for residue *i* is thus strongly dependent on conformation in a shifted Ramachandran space defined by the backbone dihedral angles ψ_{i-1} (here denoted υ_i) and φ_i. This is extended by including the relevant sidechain rotamers up to χ_{1_i} and χ_{2_i}, representing a periodic space within a dipeptide fragment that can be searched exhaustively (Fig 2A). The back-calculated spectra for these systematically sampled conformers constitute a features set that can be used to decompose the experimental spectra (Fig 2B). Here we characterize the solution ensemble as a linear combination of elements of the features set, weighted by their respective populations. Calculating these weights is analogous to parts-based representation of complex spectral mixtures often encountered in other fields of spectroscopy [13].

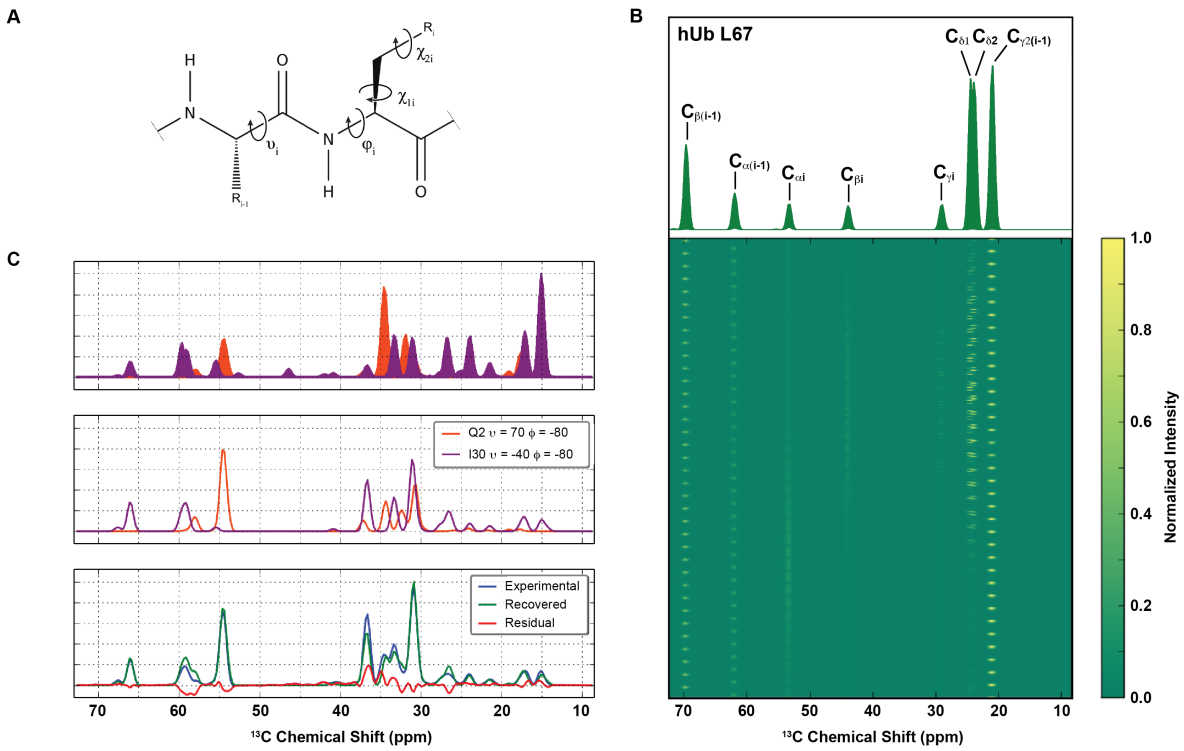

Figure 2. Conformational mapping in the CoMAND method. **A)** The definition of the conformational search space. Note that in the shifted Ramachandran space, υ_i is equivalent to $\psi_{(i-1)}$. **B)** The features matrix for L67 in hUb shown as a stacked plot where each row is a spectrum back-calculated via systematic conformational sampling, resulting in periodic intensity patterns. The order of sampling, from fastest to slowest, is χ_2, χ_1, υ, φ with 10° steps for backbone and 120° steps for sidechain angles. The intensity of each peak in the spectrum displays a different dependency on the dihedrals, underlining the power of the data to discriminate individual conformations. The projection of this plot – i.e. all members of the features set overlaid - is shown above with individual peaks assigned. **C)** Decomposing overlapped spectra. The top panel shows all members of the concatenated features set for Q2 (orange) and I30 (purple) in hUb. Two-component factorisation successfully decomposes the completely overlapped experimental spectra, yielding the correct conformations of the respective residues (middle panel).

Decomposition of the experimental spectrum can be framed as a positive matrix factorisation problem [14]. The features matrix is represented by W comprising back-calculated spectra for l conformers, resolved to m points along the ^{13}C dimension. A solution can thus be found for a vector of weights h in order to reconstruct the observed experimental spectrum V:

$$V \approx WH, \qquad V \in \mathbb{R}_{\geq 0}^{m \times n}, W \in \mathbb{R}_{\geq 0}^{m \times l}, H \in \mathbb{R}_{\geq 0}^{l \times n}$$

Where n=1 if one spectrum is considered per residue.

To map the energetic landscape of the conformational space, the yielded estimates of the population weights can in turn be expressed via a Boltzmann factor relative to a reference conformer (Materials and Methods). Examples of these conformational maps for hUb are shown in Figure 3.

An inherent question in factorization methods is the uniqueness of the solution. Here uniqueness is limited by cross-peaks that cannot be explained within the dipeptide space, but overlap with peak positions in the features set. The larger the fraction of such contamination, the more difficult finding a unique solution becomes. Given the low intensity of potentially contaminating peaks and the high dispersion of the ^{13}C dimension, the extent of contamination is usually minor. The extreme case of contamination is the coincidence of ^{15}N-HSQC positions for two or more residues, resulting in overlap of the experimental strips. Such a situation can be solved by concatenation of the features sets of the overlapped residues and solving in a multiple-dipeptide space. Figure 2C shows a typical example of this situation, where conformational maps have been obtained for two overlapped residues.

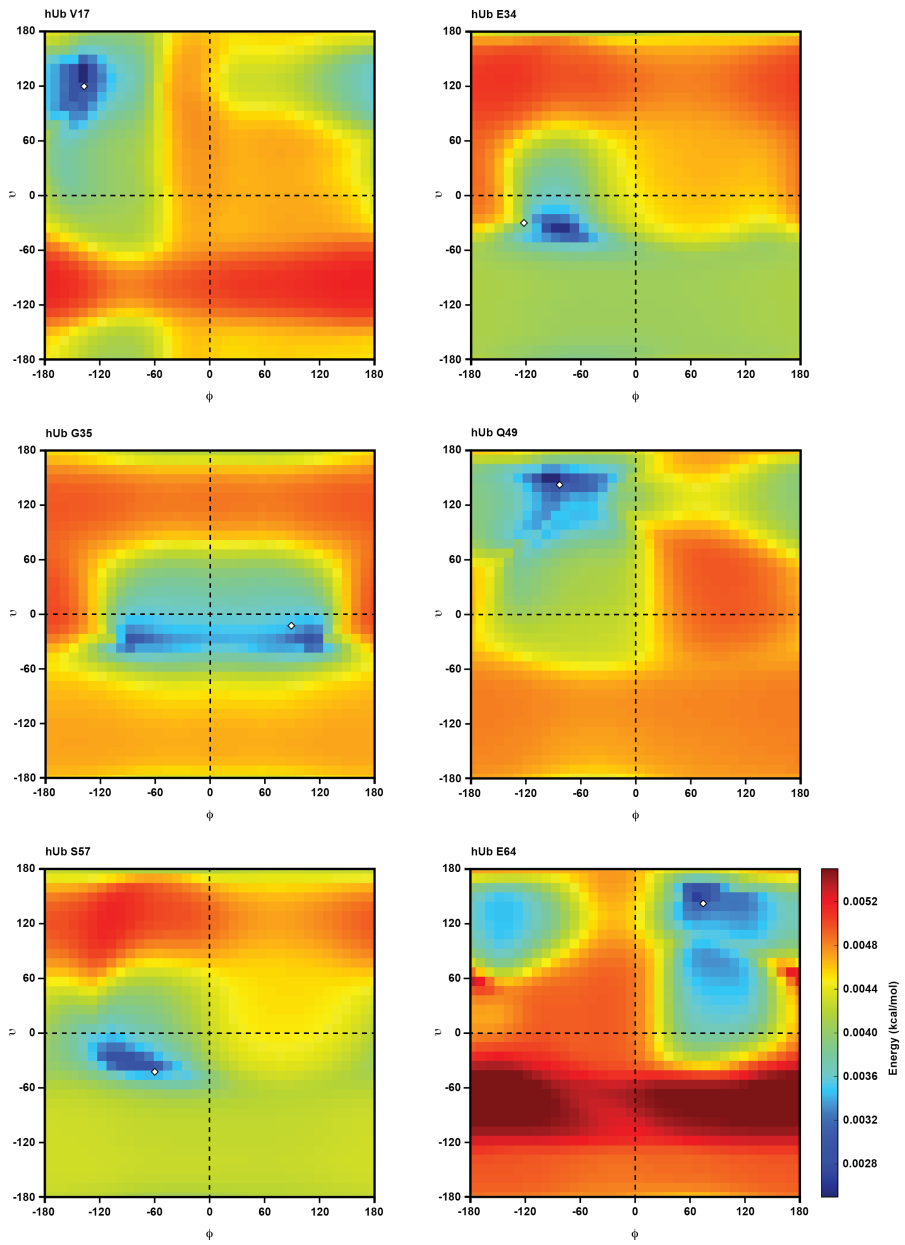

Figure 3. Conformation maps from spectral decomposition. Conformational maps are shown for six examples residues representing different secondary structural contexts in hUb. They are expressed as heat maps of conformer free energy change, relative to a two-conformer global minimum reference state. For non-glycine residues, a two-dimensional (υ, φ) slice through the full three- or four-dimensional map at the minimum χ_1/χ_2 position is shown. The map for G35 displays typical pseudo-symmetry about the φ=0 axis due to the achiral nature of glycine residues. In each map, the minima agree very well with the corresponding crystallographic conformations (1UBQ; white diamonds).

182 **Structure determination with conformational maps: the CoMAND method**

183 The conformational maps obtained from spectral decomposition provide rich
184 information for structure determination. At the simplest level, global minima can
185 provide local torsion angles sufficient for model building. These initial dihedrals
186 constitute an agnostic starting point, as they are derived directly from the data without
187 recourse to heuristics or conformational databases. A unique feature of this method is
188 the deployment of R-factors as an objective convergence test that captures both local
189 and long-range contacts. The latter can be isolated by examining the difference between
190 R-factors obtained for a linear peptide fragment and those from the full, folded model.
191 We term this measure the fold factor (F), and it should be negative if the model explains
192 long-range contacts well. Figure 4 shows that the average fold factor (F_{mean}) is a sensitive
193 overall measure of correct folding, while the sequence profile can localize misfolded or
194 poorly defined regions. Owing to this independent measure of convergence, any routine
195 can be used to build initial models. Here we employ either a Rosetta-based protocol
196 (Materials and Methods) or a purpose-designed molecular dynamics routine: Simulated
197 Annealing Replica *Seilschaft* (SARS) for initial model building (Materials and Methods).

198 In order to construct high-resolution ensembles with accurate description of the
199 underlying microstates, the experimental information contained in the conformational
200 maps can be encoded in two possible representations. The first is a temporal equilibrium
201 ensemble derived from molecular dynamics simulations. Here the conformer probability
202 distribution is imposed via a grid-based dihedral energy correction term (CMAP [15]; Fig. 3
203 and Materials and Methods). These override the standard force field CMAPs with
204 bespoke ones on a per-residue basis, providing an experimentally augmented ensemble
205 representation. The other is to aggregate a set of frames from a generalized ensemble
206 using the standard force field, based on an R-factor selection criterion, providing a wider
207 coverage of the phase space.

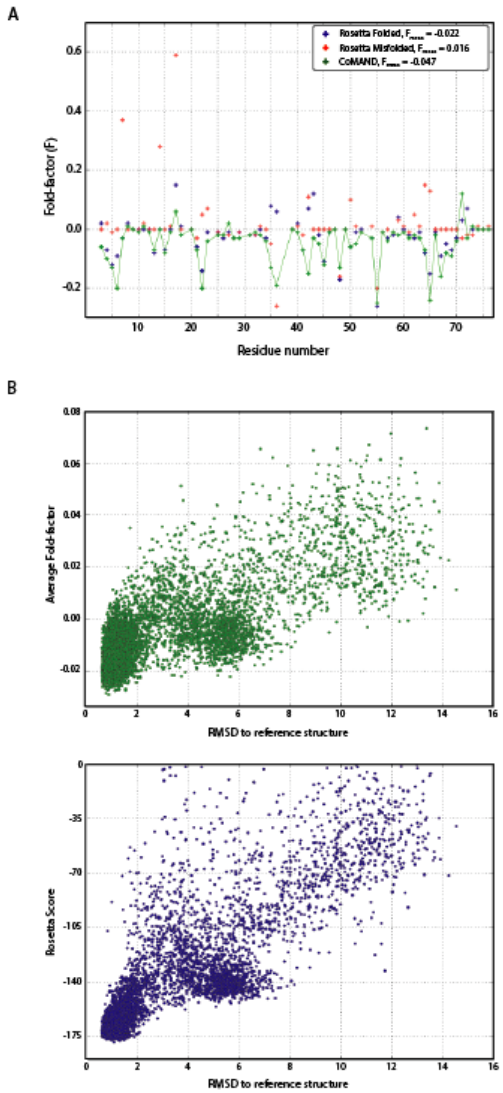

Figure 4. The R-factor as objective target function in model building. **A)** Fold factors plotted across the sequence for hUb. Fold factors (F) are calculated on a per-residue basis as the difference between the R-factor calculated for a linear peptide and that calculated for the full structure. This isolates the component of the R-factor not explained by local contacts. Negative values indicate residues in well-folded environments. Values are shown for an initial folded model from Rosetta runs plus a Rosetta structure misfolded by a strand swap in the N-terminal β-hairpin. These are compared to a representative of the final CoMAND ensemble (green line). **B)** Comparison of the average fold factor (F_{mean}) versus the Rosetta score as selection criteria for well-folded hUb models. Both measures are plotted against the RMSD to the reference structure (1UBQ) for the same set of 7215 Rosetta decoys with sub-zero score. Structures with low F_{mean} are consistently close to the reference structure.

220 We name this method of *de novo* structure determination CoMAND (for
221 Conformational Mapping by Analytical NOESY Decomposition). In addition to hUb, we
222 present examples for four structure determination projects from our Institute. U3Sfl
223 (125 amino-acids) is a protein designed as a chimera of sub-domain sized fragments
224 (Patino, S. T. & Höcker, B. *Personal communication*) KH-S1 (170 amino-acids) is a fusion
225 construct of the KH and S1 domains of *E. coli* exosomal polynucleotide phosphorylase
226 (Audin, M. H. & Sprangers, R. *Personal communication*). MlbQ is a protein implicated in
227 self-resistance to endogenous lantibiotics in actinomycetes [16]. The final example, polb4,
228 is a protein designed to reconstruct the polymerase beta N-terminal domain using two
229 unrelated peptide fragments and is presented here as a *de novo* structure
230 determination.
231 For all five proteins, we first built starting models. For U3Sfl, KH-S1 and hUb, we
232 extracted backbone dihedral angles from the factorization minima and used these for
233 fragment picking in a Rosetta protocol (Materials and Methods). For MlbQ and polb4 we
234 applied SARS, supplementing the CHARMM36 forcefield with bespoke conformational
235 maps, starting from completely extended chains. For MlbQ, folding was accelerated by
236 the addition of 10 unambiguous NOE distance restraints. We used the average R-factor
237 across the full length of the protein as a criterion for selecting models, choosing a single
238 Rosetta decoy or a single frame from the SARS runs. These models were very similar to
239 respective reference structures (Fig. 5). For U3Sfl this was a structure we had previously
240 solved by manual analysis (RMSD over backbone atoms 1.98 Å). For KH-S1, crystal
241 structures are available for homologues of the individual domains (4AM3; 1.48 Å and
242 4NNG; 1.67 Å). For MlbQ, this was the published solution structure (2MVO; 1.92 Å). As
243 no structure has previously been solved for polb4, we used the design target as a
244 reference structure (RMSD over backbone atoms 1.48 Å).

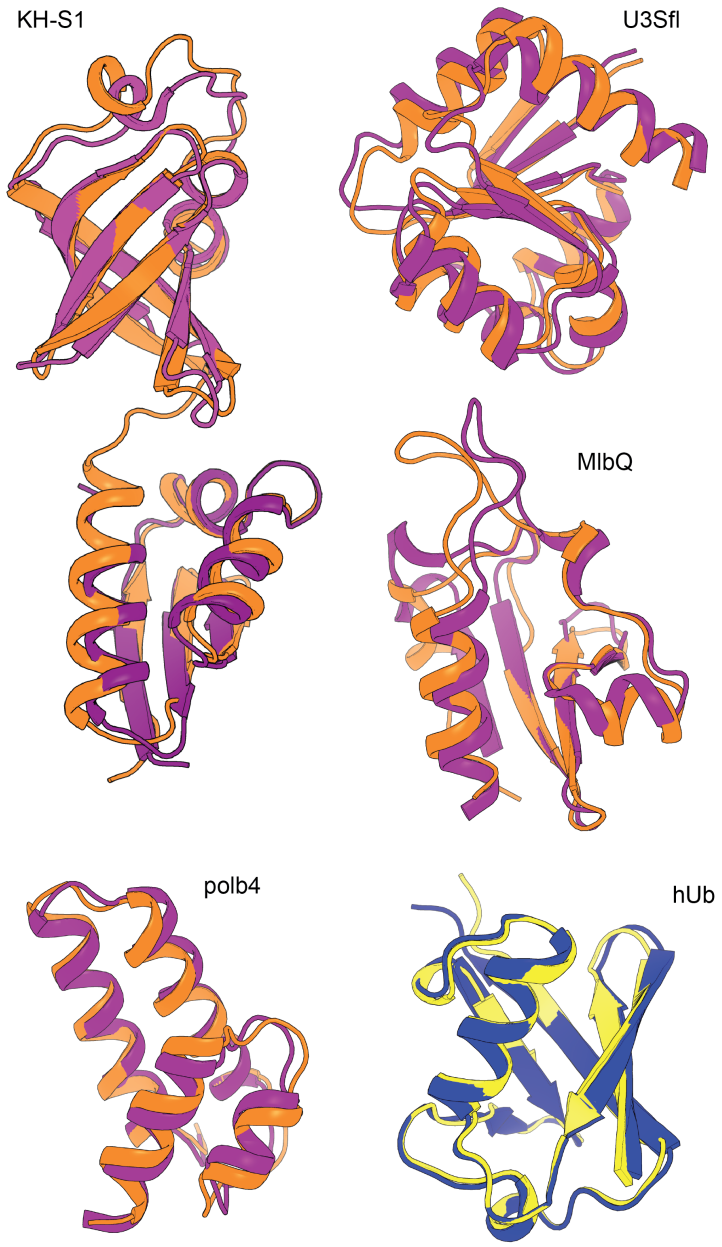

Figure 5. The CoMAND structure gallery. Initial models (orange) are shown superimposed on their respective reference structures. For U3Sfl and MlbQ the reference structures are previously solved solution structures. For KH-S1, the KH domain reference structure is 4AM3 (light purple) and the S1 domain reference structure is 4NNG (dark purple). The *de novo* structure determined for polb4 is compared to the design target. A single model from the refined CoMAND ensemble for hUb is shown in yellow and the reference structure (1UBQ) in blue (RMSD over backbone atoms 0.49 Å).

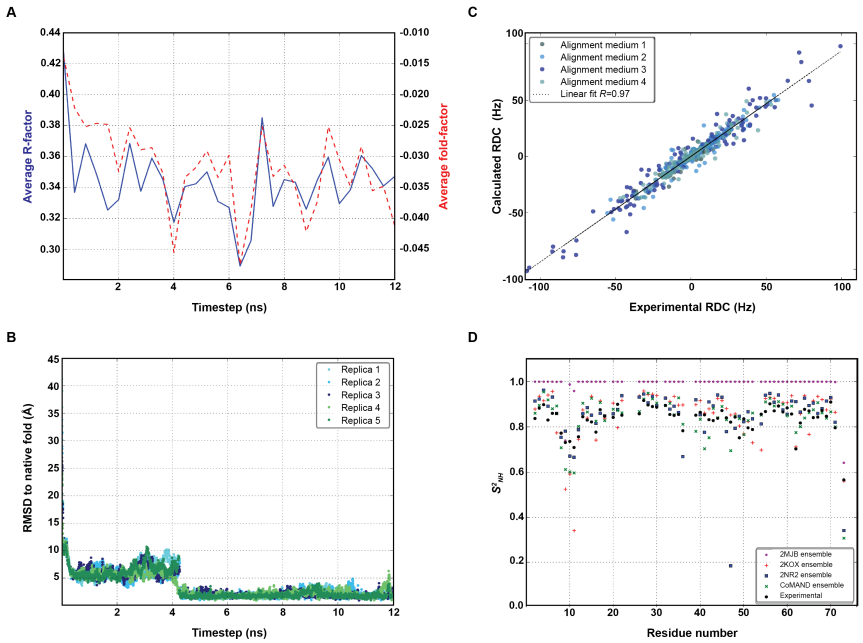

Figure 6. Validation of CoMAND ensembles. **A)** Evolution of the average sequence R-factor and fold factor along the first replica of the SARS folding trajectory of polb4. The lowest R-factor structure (time point of 6.4 ns) was chosen as a low-resolution model (See also Supplemental Movie S1) **B)** The RMSD from the native fold along the trajectory of the same SARS folding simulation. Traces for all five replicates are shown, illustrating the convergence of the protocol. **C)** The CoMAND ensemble independently reproduces NMR observables. The correlation between residual RDC values back-calculated from the CoMAND ensemble and experimental values in four different alignment media is shown [12]. The Q-factor expressing the agreement between prediction and experiment for this data set is 0.24. Similarly good agreement is obtained between back-calculated and experiment $^3J_{HNH\alpha}$ coupling constants (correlation coefficient = 0.95; Supplementary Fig. S2). RDC values report on the orientation of various bond vectors to an external molecular alignment medium and are thus sensitive to both local and global structure. $^3J_{HNH\alpha}$ coupling constants report on local ϕ angles. Neither parameter was used in compiling the CoMAND ensemble. **D)** Calculated S^2_{NH} order parameter values across the sequence of hUb using the first 20 models of the 2MJB, 2KOX, 2NR2 and CoMAND ensembles. The CoMAND ensemble best reproduces experimental values derived from NMR relaxation analysis [29] (correlation coefficient=0.82). Correlations for the reference ensembles are shown in Supplementary Fig. S2).

For refinement we conducted unrestrained molecular dynamics simulations in explicit solvent for microsecond timescales, seeded by the starting model. This was followed by a frame-picking procedure that employs a greedy optimizer to minimize the average R-factor across the ensemble. Given the wealth of structural and dynamics data available for hUb, we compiled such a refined ensemble and compared it to the reference ensembles (2KOX, 2MJB and 2NR2). The resulting ensemble of 20 conformers shows better correlation to experimental NH order parameters than the literature ensembles. It is also comparable in predicting experimental scalar couplings and RDCs to ensembles that have been built on one or more of these observables plus thousands of NOE restraints (Figure 6 and Supplement). This demonstrates the depth of the structural information captured when NOESY spectra are analyzed holistically.

DISCUSSION

NOESY spectra can be seen as an encoding of a proton-proton contact map with an approximate upper distance limit of 5 Å. If correctly decoded as a set of distance restraints, this information is sufficient to solve the structure with high accuracy and precision. However, crowded spectra and the consequent spectral overlap mean that the encoding is ambiguous. Spectral editing, for example via additional frequency dimensions, can only partially alleviate this problem, often at considerable cost in experiment time [17]. The consequences of this ambiguity are not only that individual cross-peaks cannot be uniquely assigned – i.e. attributed to a specific proton-proton contact – but also that cross-peaks may comprise significant intensity from several contacts. Conventional NMR structure determination protocols interpret NOESY spectra through peak-picking, assignment and conversion into distance restraints under a paradigm of one peak; one assignment; one restraint. Even automated routines that specifically consider ambiguities will resolve to a single effective restraint per picked peak. This represents a compromise that is not justified by the underlying nature of the data, affecting either the accuracy or precision of distance estimates. In contrast, the

CoMAND method makes no interpretation of cross-peaks, and thus stands outside this conventional assignment paradigm.

Given the ambiguity of NOESY data, the incorporation of unambiguous data from other sources is advantageous in NMR structure determination protocols. Particularly useful are data that define local dihedral angles (e.g. scalar couplings), as these are poorly defined by imprecise NOE-based distance estimates. Backbone dihedral angle predictions derived from chemical shift heuristics using the program TALOS are very widely used [18]. For example, the CS-Rosetta approach exploits this data to build structural models within the Rosetta framework [19]. In contrast to previous methods, we show that direct signal decomposition can yield backbone dihedrals unambiguously and without heuristics. Moreover, we demonstrate that the NOESY data can be leveraged to map the underlying conformational landscape in a systematic fashion. A further key advantage over existing methods is that the whole process of structure determination, including resonance assignment, model building, refinement and conformational mixture elucidation can be objectively assessed by the R-factor as a single metric.

In recent years development of analysis methods in solution NMR of proteins has been driven by the need to make automation more reliable, while using less data and extending the range to larger proteins and more difficult cases, such as membrane proteins. CoMAND contributes to this effort in that it leverages a small set of spectra on a single sample into a high-resolution structure and is therefore applicable where protein concentration or stability are limiting. As the method involves minimum user intervention after the resonance assignment stage, it is also intrinsically suited to automation. However, the most unique feature of the method lies in the power to obtain accurate descriptions of protein conformational ensembles. We therefore anticipate that the method can be applied to studying ligand binding or allosteric processes, promising to elucidate subtle conformational changes in an unprecedented level of detail.

MATERIALS AND METHODS

NMR Spectroscopy

Backbone and sidechain assignments for *de novo* structure determinations were obtained using standard triple resonance experiments. For human Ubiquitin literature values were used. Slight correction of ^{13}C shifts against the respective CNH-NOESY spectra was necessary to account for calibration differences between spectrometers and spectrum types. 3D CNH-NOESY spectra were acquired at 800 MHz on a Bruker AvanceIII spectrometer equipped with room temperature probehead. Indirect ^{13}C dimensions were typically acquired with ~100 time increments and processed with linear prediction and zero filling to 256 data points. The ^{13}C sweep width was set to cover aliphatic carbon resonances; i.e. ~10-73 ppm, resulting in a resolution of ~30 Hz per point. At this resolution, $^{1}J_{CC}$ couplings are unresolved and the spectra were run in non-constant time mode. Broadband ^{13}C pulses were used to excite aromatic resonances and these were folded into the aliphatic window without phase inversion.

CNH-NOESY spectra were analysed by extracting one-dimensional ^{13}C sub-spectra chosen from a search area centred on assigned ^{15}N-HSQC positions (typically 1-3 points in each dimension). As these sub-spectra contain only cross-peaks to a specific amide proton, choosing the strip with highest integral maximises the signal-to-noise. Residues with overlapping search areas were examined separately. In most cases strips with acceptable separation of signals could be obtained. Where this was not possible the residues were flagged as overlapped and a joint strip constructed by summing those at the estimated maxima of the respective components. A set of strips well separated from assigned HSQC positions were averaged to define a global noise level for the spectrum.

360 **NOESY back-calculation**

361 In order to back-calculate 3D CNH-NOESY spectra we modified the program
362 SPIRIT [20] by porting it to C++ and extending it to accommodate any combination of
363 proton and heteronuclear dimensions. We name this program SHINE, for Simulation of
364 Hetero-Indirect NOESY Experiments. The calculations are based on a full relaxation
365 matrix and thus account for spin diffusion in static structures. Internal motion of the
366 protein is treated by ensemble averaging over *n* contributing microstates, effectively
367 applying an *n*-state jump model of motion, where the life-time of a microstate is
368 assumed to be long compared to the interconversion time. This does not account for
369 true time-averaged phenomena, such as motion-mediated spin-diffusion, which are
370 treated as negligible for the current application. Inputs for the calculations are a
371 chemical shift list, a test structure and a set of simulation parameters. The latter are
372 largely spectral details, such as spectrometer frequencies and sweep widths, which are
373 extracted automatically from the corresponding experimental files (Bruker format), but
374 also include an estimate of a global molecular correlation time. Resonances are
375 modelled as gaussians, with ^{13}C linewidths assigned on a class basis, taking into account
376 unresolved $^{1}J_{CC}$ couplings. Here it should be noted that the short acquisition times in an
377 indirect ^{13}C dimension (<10 ms on an 800 MHz spectrometer) mean that effective
378 lineshapes are largely governed by apodisation of the time domain. They are thus
379 considerably more homogeneous than for a proton dimension.

380 The computational demand of NOESY back-calculation depends on the number
381 of protons in the relaxation network. For this reason we employ sub-structures
382 containing <150 atoms. These can be linear peptides or fragments of a folded structure.
383 Linear peptides are typically tri- or penta-peptides where the test residue is in the
384 second last position. Fragments are compiled at the residue level; residues are included
385 in the sub-structure if they contain a proton within a given radius (typically 5 Å) of a
386 target residue proton. The output is a one-dimensional strip displaying contacts to a
387 single backbone amide proton. In this mode, back-calculation typically takes less than 20

388 ms per conformer on a single processor core. The program also outputs a list of peak
389 intensities that can be used to build multi-dimensional spectra suitable for viewing in
390 SPARKY [21], with annotation of individual cross peaks.

391 Calculation of features sets is performed for each residue for which an
392 experimental spectrum is available. The starting structures are linear peptide fragments
393 extracted from an arbitrary structural model. In the current work these were tripeptides
394 centred on the test residue. For back-calculation of contacts to the amide proton of
395 residue i this peptide is modified through a set of torsion angles in a shifted
396 Ramachandran space: angles ψ_{i-1} (here denoted υ_i) and φ_i and up to two sidechain χ
397 angles. The backbone angles were searched at 10° granularity, while sidechain angles
398 sample all staggered rotamers. This results in a maximum of 11664 conformers. No
399 checks for steric clashes are applied to test if a conformer is physically reasonable and
400 bond lengths and angles remain constant throughout. An exception is for proline
401 residues where the search space is restricted to a physically realistic range in φ_i. As
402 proline residues lack an amide proton, their features sets are compiled by back-
403 calculation of the amide proton of the previous residue, which is the most sensitive
404 reporter on the υ angle of proline The set of back-calculated spectra for each residue are
405 stored as a single file. For residues flagged as overlapped, the features set files are
406 concatenated to create a joint set.

407 **Calculation of R-factors**

408 We define the R-factor as the relative RMS residual between an experimental
409 spectrum and the expectation spectrum back-calculated from a structural model. The
410 use of RMS is in analogy to the quality factor (Q-factor) calculated for residual dipolar
411 coupling data [22]. The use of RMS tends to emphasize large outliers relative to the R-
412 factor used in crystallography, which averages absolute differences between
413 experimental and back-calculated structure factors. It also provides a convenient
414 definition of the optimum scaling factor for the back-calculated spectrum s_{calc}, which
415 can be calculated as:

$$s_{calc} = \frac{\boldsymbol{v}_{exp} \cdot \boldsymbol{v}_{calc}}{\|\boldsymbol{v}_{calc}\|^2}$$

where $\boldsymbol{v}_{exp}$ and $\boldsymbol{v}_{calc}$ are the experimental and back-calculated spectral intensities vectors, respectively.

The R-factor is then calculated as:

$$\boldsymbol{R} = \frac{\sqrt{\sum_i^n (\boldsymbol{v}_{exp_i} - \boldsymbol{s}_{calc}\boldsymbol{v}_{calc_i})^2}}{\sqrt{\sum_i^n \boldsymbol{v}_{exp_i}^2}}$$

The theoretical range of the metric is from 0 to 1, however the maximum value can only be reached if there is no correspondence between peaks in the experimental and expectation spectra, in which case the scaling factor, s_{calc}, will approach zero. The practical minimum value is limited by the RMS noise on a per-residue basis.

Calculation of fold factors

Back-calculation for a structure can be carried out either on a linear peptide or as a fragment of the full structure. The R-factor can therefore be calculated on either basis. Calculation for peptides cannot explain any peaks outside the linear context, whereas all peaks should be explained in a fragment. The difference between fragment and peptide R-factors therefore reports on the fraction of cross-peak intensity that can only be explained in the folded structure. Here we report this difference as a per-residue fold factor, F:

$$F_i = R_i^{full} - R_i^{pentapeptide}$$

where R_i^{full} is based on a fragment from the full model and $R_i^{pentapeptide}$ is the penapeptide based R-factor for residue i. As R-factors should decrease as more of the cross-peak intensity is explained, the fold factor will be consistently negative for well-

437 folded structures. High positive values are indications of misfolding, while continuous
438 stretches of values close to zero should only be seen for unstructured regions.

439

440

441 **Factorisation and shifted space-CMAP construction**

442

443 The dipeptide conformational space was sampled according to the following
444 granularity: $\Delta\upsilon = 10°, \Delta\varphi = 10°, \Delta\chi_1 = 120°, \Delta\chi_2 = 120°$. The experimental vector $\boldsymbol{v} \in$
445 $\mathbb{R}^m$ consisted of $m = 256$ data points of the acquired CNH-NOESY strip (i.e. NOE
446 intensities *vs.* ^{13}C chemical shift), while $W \in \mathbb{R}_{\geq 0}^{m \times l}$ contained all of the back-calculated
447 spectra of the l conformers sampled. With the aim of solving for the positive factors
448 vector $\boldsymbol{h}$ that weights each column of W to best explain $\boldsymbol{v}$.

449

450 The principal solution can be defined as follows:

451

452 $$\min_{h \geq 0} \|\boldsymbol{v} - W\boldsymbol{h}\|^2$$

453

454 $\boldsymbol{h}$ can be also derived directly once the Moore-Penrose pseudo-inverse of the back-
455 calculated spectra matrix, W^+, is computed as:

456

457 $$\boldsymbol{h} = \sqrt{\boldsymbol{v}^2 W^{+2}}$$

458

459 The uniqueness of the solution in positive matrix factorization is limited by the
460 ranks of the component matrices with the upper bound of $\min(m, n)$ [23]. Here, as
461 $rank(W) \gg rank(\boldsymbol{v}) = rank(\boldsymbol{h}) = 1$, the limiting factor is the rank of $\boldsymbol{v}$. Thus a 2-
462 conformer block-wise factorisation was sought, being closest to this limit. The 2-
463 conformer solution is also computationally tractable for handling the fine degree of
464 conformational sampling described above. This solution offers a recovered spectrum

465 that has a higher or equal weight against any single-conformer solution. The former two-
466 component weight h_{ref} was used as highest propensity reference state for estimating
467 the relative normalised propensities of every other available conformer h_i. A Boltzmann
468 factor can be directly used to estimate the energy of every conformer according to:
469

470 $$\frac{h_{ref}}{h_i} = e^{\frac{\Delta G_{ref \to i}}{kT}}$$

471

472 The above procedure was performed across the (υ, φ) planes at every (χ_1, χ_2)
473 combination. And lowest R-factor yielding plane was the one embedded into the
474 CHARMM36 force field to generate MD ensembles with experimentally derived
475 backbone dihedrals energy surfaces.
476

477 **Model building using Rosetta**

478 The Rosetta software package (version 3.6) [24] was used to build structural models
479 using backbone dihedral restraints derived from conformational mapping. First, a
480 Rosetta dihedral angle constraint file (.cst) was compiled. For each residue position *i*, a
481 MultiConstratin field was written to comprise both dihedral angles υ_i and φ_i. When
482 multiple dihedral angles were possible for one residue position, an AmbigousConstratint
483 field was used to include all possibilities. The Rosetta fragment picking program
484 *fragment_picker* [25] was used to select 3mer and 9mer fragments satisfying the dihedral
485 restraints from the PDB database. The DihedralConstraintsScore weight used was 500
486 and the minimum allowed 100. The SecondarySimilarity (weight 150, minimum 1.5) and
487 RamaScore (weight 150, minimum 1.5) both used psipred. FragmentCrmsd was not
488 used. Other *fragment_picker* parameters were defaults. The fragment database
489 vall.jul19.2011 was searched and 200 3mer and 200 9mer fragments were picked for
490 each position in the protein. The Rosetta *ab initio* folding program *AbinitioRelax* was
491 then used to fold the target proteins using these picked fragments (flag file: –ex1, -ex2 -
492 use_input_sc, -flip_HNQ, -no_optH false, -silent_gz 1). Typically, 20,000 decoys were

generated for each target protein.

SARS simulations

The SARS framework was designed to provide robust and efficient conformational sampling without relying on any bioinformatic data, with the only input being CNH-NOESY-acquired dihedral preferences encoded as residue-wise biasing potentials into a standard atomistic force field. The sampling acceleration scheme is based on the assumption that the native structure lies at the global minimum of the potential energy surface that is led to by successively deeper local minima. The algorithm initiates multiple replicas from the same fully extended peptide chain starting system. It then combines alternating rounds of simulated annealing between two temperature baths with conjugate gradient minimisation. Each such round constitutes a search step in a collective swarming behaviour that guides the configuration exchange between replicas whenever a new minimum is reached. In this way, all of the high-energy replicas follow the lead of the lowest energy one. The implementation details and convergence properties of the SARS method will be detailed in a separate publication.

Molecular dynamics of human ubiquitin

An initial low-resolution model generated by ROSETTA from the CNH-NOESY-acquired dihedral restraints was taken as the input coordinates for molecular dynamics simulations. The standard trajectories were acquired from 10 independent replicas conducted using the standard CHARMM36 force field [26]. In contrast, the guided trajectories were acquired from 10 independent replicas where the systems were built using an augmented CHARMM36 with bespoke CMAP potentials. The CMAP potentials were directly constructed on a per-residue basis, in the shifted Ramachandran space (υ_i,

522 φ_i), from the energy maps as described above. The energy maps where scaled by an
523 arbitrary factor depending on restraint level required, and the standard CMAPs were
524 adopted wherever a bespoke one was not available. These residues with unmodified
525 cross-terms were M1, Q2, G10, P19, E24, I30, P37, P38, G53 and G76.
526
527 All of the simulations were performed in explicit TIP3P water as solvent,
528 containing 9226 solvent molecules each, and neutralised by 0.15 M Sodium Chloride in a
529 cubic periodic unit cell. Energy minimisation was performed through 5000 steps of
530 conjugate gradient minimisation, which was followed by 30 ns of NPT equilibration. The
531 time step was set to 2 fs and a Langevin Piston was set to 1 Atm at oscillation period of
532 200 fs and damping period of 50 fs and a temperature of 298 K. A Langevin thermostat
533 was accordingly set with damping coefficient of 1 ps^{-1}. A nonbonded interactions
534 distance cutoff was set to 12.0 Å at a switching distance of 10.0 Å with all nonbonded
535 force and pair list evaluations were performed every timestep, and long-range
536 electrostatics were computed using the Smooth Particle Mesh Ewald method [27] as
537 implemented in the NAMD engine [28]. Data was collected from the ensuing NVT
538 trajectories, dumping coordinates every 5 ps for analysis. Frame picking was done from
539 the 10 production trajectories of 30 ns each based on the standard force field that would
540 represent a steady state canonical ensemble.
541
542 **Ensemble building**
543
544 To compile the final CoMAND ensemble for hUb, we performed frame picking
545 from an equilibrium ensemble, such that the compiled conformers belong to microstates
546 of minimal free energy and maximal entropy at the target temperature. This should
547 provide more physically realistic final models compared to those collected from
548 constrained tempering schemes with unrealistic Hamiltonians. To pick frames from the
549 production trajectories we applied a greedy algorithm aimed at minimising the average
550 R-factor for all residues where experimental CNH-NOESY strips were available. For

consistency, residues lacking experimental strips (M1, Q2, G10, P19, E24, I30, P37, P38, G53, G76) were excluded from the following comparisons with other datasets.

Validation versus NMR observables

Expectation NMR observables were calculated for the CoMAND and reference hUb ensembles. For uniformity, only the first 20 conformers from each ensemble were considered for comparisons.

For backbone amide order parameters, frames were aligned into a singular molecular frame of reference that best fits backbone atoms of residues 1 through 70. The order parameter, S_{NH}^2, was directly computed according to the method described by Nederveen and Bonvin [29] through the following equation:

$$S_{NH}^2 = \frac{1}{2}\left(3\sum_{i=1}^{3}\sum_{j=1}^{3}\langle\mu_i\mu_j\rangle^2 - 1\right)$$

Where μ_i is the normalised internuclear NH bond vector in the molecular frame of reference and $\langle\cdot\rangle$ represents the ensemble-averaged value.

Expectation ${}^3J_{H^N H^\alpha}$ coupling constants were calculated according to the following Karplus function:

$${}^3J_{H^N H^\alpha} = 6.98\cos^2\theta - 1.38\cos\theta + 1.72$$

Where θ is the $H^N - N - C^\alpha - H^\alpha$ dihedral angle. These were compared to literature experimental values [30].

Deviations from experimental residual dipolar couplings and an overall Q-factor for the CoMAND ensemble were calculated using the program PALES [31] using 996 backbone coupling published for 2MJB across four alignment media [12].

Accession Numbers

Coordinates for the CoMAND ensemble for human ubiquitin have been submitted to the protein data bank.

Acknowledgements

We thank Birte Höcker and Remco Sprangers and their co-workers for providing samples for U3Sfl and KH-S1. Programming of the SHINE package was supported by Professors Alois Knoll and Horst Kessler at the Technical University, Munich. We thank Andre Noll (MPI for Developmental Biology, Tübingen) for helpful discussions.

MG designed and implemented spectral decomposition routines and the SARS protocol, performed molecular dynamics calculations and analysed their output. MC implemented back-calculation routines. MR programmed SHINE. HZ performed Rosetta calculations and analysis. VT and MC acquired and analysed NMR spectra. MC conceived and directed the project. The manuscript was written by MC and MG and edited by all the authors.

This work was supported by institutional funds of the Max Planck Society. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

601 **REFERENCES**

602

603 1 Lange, O. F. *et al.* Recognition dynamics up to microseconds revealed from an
604 RDC-derived ubiquitin ensemble in solution. *Science* 320, 1471-1475 (2008).
605 2 Bonvin, A. M. & Brunger, A. T. Do NOE distances contain enough information to
606 assess the relative populations of multi-conformer structures? *J Biomol NMR* 7,
607 72-76 (1996).
608 3 Burgi, R., Pitera, J. & van Gunsteren, W. F. Assessing the effect of conformational
609 averaging on the measured values of observables. *J Biomol NMR* **19**, 305-320
610 (2001).
611 4 Lindorff-Larsen, K., Best, R. B., Depristo, M. A., Dobson, C. M. & Vendruscolo, M.
612 Simultaneous determination of protein structure and dynamics. *Nature* **433**, 128-
613 132 (2005).
614 5 Richter, B., Gsponer, J., Várnai, P., Salvatella, X. & Vendruscolo, M. The MUMO
615 (minimal under-restraining minimal over-restraining) method for the
616 determination of native state ensembles of proteins. *Journal of Biomolecular*
617 *NMR* **37** (2007).
618 6 Brunger, A., Clore, G., Gronenborn, A., Saffrich, R. & Nilges, M. Assessing the
619 quality of solution nuclear magnetic resonance structures by complete cross-
620 validation. *Science* **261**, 328-331 (1993).
621 7 Spronk, C. A. E. M., Nabuurs, S. B., Krieger, E., Vriend, G. & Vuister, G. W.
622 Validation of protein structures derived by NMR spectroscopy. *Progress in*
623 *Nuclear Magnetic Resonance Spectroscopy* **45**, 315-337 (2004).
624 8 Huang, Y. J., Rosato, A., Singh, G. & Montelione, G. T. RPF: a quality assessment
625 tool for protein NMR structures. *Nucleic Acids Research* **40**, W542-W546 (2012).
626 9 Diercks, T., Coles, M. & Kessler, H. An efficient strategy for assignment of cross-
627 peaks in 3D heteronuclear NOESY experiments. *Journal of Biomolecular NMR* **15**,
628 177-180 (1999).

10 Lindorff-Larsen, K., Maragakis, P., Piana, S. & Shaw, D. E. Picosecond to Millisecond Structural Dynamics in Human Ubiquitin. *The Journal of Physical Chemistry B* **120**, 8313-8320 (2016).

11 Fenwick, R. B. *et al.* Weak Long-Range Correlated Motions in a Surface Patch of Ubiquitin Involved in Molecular Recognition. *Journal of the American Chemical Society* **133** (2011).

12 Maltsev, A. S., Grishaev, A., Roche, J., Zasloff, M. & Bax, A. Improved Cross Validation of a Static Ubiquitin Structure Derived from High Precision Residual Dipolar Couplings Measured in a Drug-Based Liquid Crystalline Phase. *Journal of the American Chemical Society* **136**, 3752-3755 (2014).

13 Xu, Q., Sachs, J. R., Wang, T.-C. & Schaefer, W. H. Quantification and Identification of Components in Solution Mixtures from 1D Proton NMR Spectra Using Singular Value Decomposition. *Analytical Chemistry* **78**, 7175-7185 (2006).

14 Paatero, P. & Tapper, U. Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics* **5**, 111-126 (1994).

15 Mackerell, A. D., Feig, M. & Brooks, C. L. Extending the treatment of backbone energetics in protein force fields: Limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations. *Journal of Computational Chemistry* **25**, 1400-1415 (2004).

16 Pozzi, R. *et al.* Distinct mechanisms contribute to immunity in the lantibiotic NAI-107 producer strain Microbispora ATCC PTA-5024. *Environmental Microbiology* **18**, 118-132 (2016).

17 Güntert, P. Automated NMR protein structure calculation. *Progress in Nuclear Magnetic Resonance Spectroscopy* **43**, 105-125 (2003).

18 Shen, Y., Delaglio, F., Cornilescu, G. & Bax, A. TALOS+: a hybrid method for predicting protein backbone torsion angles from NMR chemical shifts. *Journal of Biomolecular NMR* **44**, 213-223 (2009).

19 Shen, Y. *et al.* Consistent blind protein structure generation from NMR chemical shift data. *Proceedings of the National Academy of Sciences* **105** (2008).

20 Zhu, L., Dyson, H. J. & Wright, P. E. A NOESY-HSQC simulation program, SPIRIT. *Journal of Biomolecular NMR* **11**, 17-29 (1998).

21 SPARKY 3 (University of California, San Fransisco).

22 Cornilescu, G., Marquardt, J. L., Ottiger, M. & Bax, A. Validation of Protein Structure from Anisotropic Carbonyl Chemical Shifts in a Dilute Liquid Crystalline Phase. *Journal of the American Chemical Society* **120** (1998).

23 Wang, Y. X. & Zhang, Y. J. Nonnegative Matrix Factorization: A Comprehensive Review. *IEEE Transactions on Knowledge and Data Engineering* **25**, 1336-1353 (2013).

24 Das, R. & Baker, D. Macromolecular modeling with rosetta. *Annual review of biochemistry* **77**, 363-382 (2008).

25 Gront, D., Kulp, D. W., Vernon, R. M., Strauss, C. E. & Baker, D. Generalized fragment picking in Rosetta: design, protocols and applications. *PloS one* **6**, e23294 (2011).

26 Huang, J. & MacKerell, A. D. CHARMM36 all-atom additive protein force field: Validation based on comparison to NMR data. *Journal of Computational Chemistry* **34**, 2135-2145 (2013).

27 Essmann, U. *et al.* A smooth particle mesh Ewald method. *The Journal of Chemical Physics* **103**, 8577-8593 (1995).

28 Phillips, J. C. *et al.* Scalable molecular dynamics with NAMD. *Journal of Computational Chemistry* **26**, 1781-1802 (2005).

29 Nederveen, A. J. & Bonvin, A. M. J. J. NMR Relaxation and Internal Dynamics of Ubiquitin from a 0.2 µs MD Simulation. *Journal of Chemical Theory and Computation* **1**, 363-374 (2005).

30 Wang, A. C. & Bax, A. Determination of the backbone dihedral angles phi in human ubiquitin from reparametrized empirical Karplus equations. *Journal of the American Chemical Society* **118**, 2483-2494 (1996).

31 Zweckstetter, M. & Bax, A. Prediction of sterically induced alignment in a dilute liquid crystalline phase: Aid to protein structure determination by NMR. *Journal of the American Chemical Society* **122**, 3791-3792 (2000).