Mollib: A Molecular and NMR Data Analysis Software

Justin L. Lorieau[*]

Department of Chemistry, University of Illinois at Chicago, 845 W Taylor St, Chicago IL 60607

* Corresponding Author: justin@lorieau.com

Abstract

Mollib is a software framework for the analysis of molecular structures, properties and data with an emphasis on data collected by NMR. It uses an open source model and a plugin framework to promote community-driven development of new and enhanced features. Mollib includes tools for the automatic retrieval and caching of protein databank (PDB) structures, the hydrogenation of biomolecules, the analysis of backbone dihedral angles and hydrogen bonds, and the fitting of residual dipolar coupling (RDC) and residual anisotropic chemical shift (RACS) data. In this article, we release version 1.0 of mollib and demonstrate its application to common molecular and NMR data analyses.

Introduction

The analysis and manipulation of biomolecular structures and data is an essential step in the interpretation of molecular information collected by NMR and other biophysical techniques. Many tools are available to analyze the distribution of Ramachandran angles,[1] steric clashes,[2,3] sidechain rotameric states, electrostatics and other features in molecules. In many cases, these tools work independently and are difficult to run together in a comprehensive analysis of a molecular structure. The Molprobity software package is a general structural analysis tool that specializes in crystal structures.[3] It includes a web interface to add hydrogen atoms to molecular structures, conduct an analysis of backbone Ramachandran and sidechain angle outliers and identify steric clashes within molecules. In the analysis of NMR structures, PROCHECK-NMR uses Ramachandran angle and χ_1-angle outliers and NOE violations to evaluate the quality of an NMR structural ensemble.[4] The volume, area dihedral angle reporter (VADAR) additionally includes the characterization of excluded volume, solvent accessible surface area, hydrogen bond energies and steric quality to evaluate structures.[5] The Protein Structure Validation Software (PSVS) suite is another example of a useful web interface that integrates many existing tools, including NMR Protein Recall, Precision and F-measure scores (RPF[6]), PROCHECK,[7] Molprobity,[8] Verify 3D[9] and Prosa II,[10] in validating biomolecular structures.[11] PSVS includes Z-scores based on global quality measures of a structure for Ramachandran angles, steric clashes and other structural features. Likewise, the Resolution by Proxy (ResProx) webserver[12] uses a large number of global quality measures to assign a structure's resolution, based on crystallographic congeners of comparable resolution, and it achieves a high correlation between a structure and its corresponding equivalent resolution.

In the analysis of structural data, more specialized tools are available. For Nuclear Magnetic Resonance (NMR) data, residual dipolar couplings (RDCs) are an example of a high-resolution structural restraint to orient bonds within a molecular alignment frame. Analysis tools for RDC data include DC in NMRPipe,[13] the residual dipolar coupling analysis tool (REDCAT),[14] the prediction of alignment from structure (PALES)[15] and Module 2.[16] These tools determine the order tensor relative to a molecular frame that best matches RDC data, with some tools having useful additional features such as the analysis of errors and dynamics in REDCAT, the visualization of the order tensor with respect to the molecule's frame in Module 2 or the prediction of the steric alignment tensor in PALES. These tools analyze how well RDC data fit to a given structure. However, they work independently of other structural analyses, which must be conducted separately.

In this report, we introduce a new tool, mollib, for the analysis of both molecular structures *and* data with an initial emphasis on data collected by NMR. Mollib presents a simple and unified program and software library with a multitude of analysis tools that can be used independently or together to produce more sophisticated analyses. Mollib is built on an extensible plugin framework that simplifies the use and addition of analysis tools. The plugin framework enables developers to integrate new structural and data analyses that can be either installed separately or as part of the mollib package. Mollib includes extensive documentation for its command-line interface (CLI) and application program interface (API), and its source code is openly available and clearly annotated so that users can easily identify and modify algorithms and contribute to the development of mollib.

Mollib is coded in Python (v2.7 and v3.4-3.6) with performance critical components written in Cython and C. Other Python Structural Biology tools are available that specialize in biomolecular structure visualization (PyMOL), protein dynamics and sequence analysis (Prody, Biopython). Mollib focuses on structural and data analysis. The current version of mollib (v1.0) includes tools to 1) add hydrogen atoms to molecules, 2) measure and characterize distances, angles and dihedrals, 3) list hydrogen bonds and compare these to high-resolution structures, and 4) fit RDC data and residual anisotropic chemical shift data (RACS, also referred to as the residual chemical shift anisotropy) from NMR. All settings are customizable by the user through configuration files. This report focuses on the command-line interface of mollib, which does not require knowledge of programming, with discussions on the implementation details of the analysis algorithms.

Core and Plugin Modules

Version 1.0 of mollib includes the *core* module for processing molecules in the protein databank (PDB) format, a *hydrogens* module to strip and add hydrogens to molecules, an *hbonds* module to identify and classify hydrogen bonds and a *partial alignment* (*pa*) module for fitting RDC and RACS data. See the Supplementary Information for documentation on the various modules and their command-line interfaces and APIs.

Core and Settings. The *core* module (SI manual pg 39) is a Python interface to access molecule properties such as atomic coordinates and masses. Molecules, chains and residues are subclassed Python dictionaries and support indexed accession (ex: molecule['A'][13]['CA']). The molecule object includes functions to calculate and cache molecular properties, and to conduct manipulations such as translations and rotations with Euler angles. The *core* module supports the automatic fetching and caching of structures from the PDB.

The PDB parser is currently coded in Python, and its processing speed compares favorably to existing parsers. Parsing a large, 76MB compressed file, the glutamine synthetase from Mycobacterium tuberculosis (PDB: 1HTQ), with 97,872 atoms in 10 models requires *ca.* 5.5s in single-threaded mode to decompress and parse on a Macbook Pro laptop with an Intel Core i7 processor at 2.7GHz and 16GB of random access memory (RAM). On a 64-bit Linux system with two Intel Xeon CPUs E5-2643 at 3.4GHz and 128GB of RAM, the same operation requires *ca.* 3.8s. On a similar computational platform with an uncompressed PDB file, the computation requires tens of seconds with interpreted languages (BioPython, BioRuby) and 4-5s with compiled languages (hPDB, Rasmol) in single-threaded mode.[17] Compressed, gzipped files are downloaded and cached by default in mollib to reduce bandwidth usage and download times.

In molecules, residues and atoms, the bonding topologies for protein residues and heteratom 'CONECT' annotations are currently supported. Ionizeable groups are supported for Asp, Gln, His, Cys, Tyr and Lys sidechains as well as the α-amino group and terminal carboxylate of each chain. Default pK_as are specified in the settings (SI manual pg 22) and impact the hydrogenation algorithm.

Mollib includes simple functions to load an individual model from an NMR structure or factory methods to load specific models or all models from a PDB file. From the command-line interface, the '--models' option allows the user to specify specific model id numbers for each analysis.

Plugins and Settings. The other components of mollib are integrated as plugins that can be dynamically installed, integrated and removed. Each plugin offers independent analysis functions that can be installed separately from the main mollib distribution. By default, version 1.0 of mollib includes plugins to process and hydrogenate molecules, to measure geometries in molecules, to identify and classify hydrogen bonds and to fit RDC and RACS data.

The *core* module and plugins work together with a settings manager to control the behavior and default values of mollib. Configuration files are simple text files that are either stored in a user's home directory (.mollibrc) or specified on the command line when invoking mollib. The current configuration options can be viewed in the manual (SI manual pg 22-30).

Process and Hydrogens Plugins

All plugins may register preprocessors, processors and postprocessors to manipulate the molecule before, during and after analysis. The *process* plugin (SI manual pg 4) simply allows the preprocessors to operate on the molecule(s), and it gives the user the option to save the processed molecule to a new file.

An example of a preprocessor is the *hydrogens* plugin. The *hydrogens* plugin optionally removes and re-adds hydrogen atoms with optimal geometries to a molecule before analysis and further processing. The *hydrogens* plugin functions analogously to the REDUCE program.[18]

Default topologies are included for protein residues, and the CONECT records for heteroatoms are supported. The *hydrogens* plugin iterates over heavy atoms in the molecule(s) to identify whether a specific atom is sp, sp^2 or sp^3 hybridized as well as the number of hydrogen atoms needed for hydrogenation. Hydrogens are assigned according to the convention described in Markley *et al.,*[19] including prochiral assignments (pro-R and pro-S) and the E/Z designations for planar groups.

Hydrogens are added according to the local geometries of atoms. For example, the first methyl proton will be placed trans to the heaviest atom adjacent to the methyl C-C or C-S group. Hydrogen atoms for ionizeable groups will be placed based on the molecule's specified pH and the default pK_a values in the settings. In cases where multiple degenerate ionizable groups exist, like the two CO groups of an Asp sidechain, the group closest to a hydrogen bond acceptor will be hydrogenated. Ionizeable groups supported include sidechains, α-amino groups and terminal carboxylates. However, the *hydrogens* plugin will not rotate sidechain groups to promote the formation of hydrogen bonds. Unlike REDUCE, the *hydrogens* plugin also does not support the reconfiguration of amide sidechains in Asn and Gln residues, which are occasional misassigned in crystal structures.

Geometry Measurement Plugin

The *measure* plugin (SI manual pg 5) is used to measure basic geometries in molecules. An abbreviated syntax is used to select, calculate and compare multiple distances, angles or dihedral angle values. Listed geometries can be filtered to only include bonded atoms, intra- or interressidue atoms, or intra- or intermolecular atoms. Statistics can be calculated for a group of measurements.

Ramachandran dihedral angles are a special function of the *measure* plugin. The *measure* plugin will list the backbone φ- and ψ-angles for each residue. For the i^{th} residue, φ is calculated from the C_{i-1}-N_i-C^{α}_i-C_i dihedral angle, and ψ is calculated from the N_i-C^{α}_i-C_i-N_{i+1} dihedral angle. In addition to calculating these angles, mollib will report the likelihood of finding a particular set of φ/ψ dihedral angles in high-resolution crystal structures.

Backbone dihedral probabilities (**Fig 1**) are calculated similarly to Molprobity[3] (v4.3). However, dihedral angles in mollib are further grouped by secondary structure classification. Secondary structure assignments are estimated by mollib based on hydrogen bonds and backbone dihedral angles, analogously to the Dictionary of Protein Secondary Structure (DSSP).[1]

Fig 1

Molprobity identifies dihedral angle outliers from one of four probability density maps constructed from 500 high-resolution structures: overall, Gly residues, Pro residues and pre-Pro residues.[3] In mollib, a residue's secondary structure classification is based on hydrogen bonds and backbone dihedral angles, and the residue's dihedral angles are compared to the distribution of dihedral angles for the same secondary structure type in high-resolution structures. The probability of finding a specific set of dihedral angles, as well as the energy penalty, is calculated for each set of backbone dihedral angles. The potentials of mean-force (PMFs) are calculated using a Boltzmann inversion of the probabilities (P) for each set of backbone dihedral angle configurations (Ω).[20]

$$E(\Omega) = -kT \ln P(\Omega) \tag{0}$$

For a given secondary structure type, the energy (in kT) is zero for the highest probability configurations (Ω) in backbone dihedral angles. This operation effectively rescales the highest probability configuration to 100%.

Mollib classifies backbone dihedral angles into 1 of 14 potential energy surfaces (**Fig 1**): α-helical, α-helical (N-term), α-helical (C-term), 3_{10}-helix, π-helix, sheet, sheet (N-term), sheet (C-term), type I and I' turns, type II and II' turns, glycine and no classification. The potential energy contour plots are constructed from *ca.* 11,300 high-resolution structures with over 5 million dihedrals and at least 12,000 dihedral angle pairs per contour map. The high-resolution structures were selected from crystal structures of proteins with at least 50 residues, a resolution between 0.5-1.6Å, an observed R-factor below 0.25, a free R-factor below 0.30 and representative structures with an 80% sequence identity.[20] A comparison of the overall energy contour map between Molprobity and mollib is presented in Figure S1.

We identify dihedral angles for each group of secondary structure type since this distribution is typically much more narrow than the overall distribution used by Molprobity. Residues within an α-helix, for instance, are narrowly clustered around (-62º, -42º). Although the overall Ramachandran map is presented in Fig 1 and Fig S1, it is not used to identify backbone dihedral angle outliers with mollib. This approach more readily identifies outliers for a given secondary structure type, and it does not bias the reported probabilities by the propensities of secondary structure types in the model dataset. Outliers found using this approach indicate either a misassigned secondary structure unit or a residue that does not follow the canonical backbone dihedral angles for a given secondary structure classification. Backbone dihedral outliers do not necessarily indicate an error in the structure, and they likely point to regions of interest with additional dynamics not captured by a single, average state.

Backbone dihedral outliers are easy to identify with mollib's more diverse and stringent grouping of backbone dihedral angles. With ubiquitin's NMR structure (PDB: 2MJB, 76 a.a.), mollib identifies residues G10, K33 and E34 as 'warning' outliers ($E > 3.4kT$, $< 3.3\%$ probability) in backbone dihedral angles and residue G75 as a clear outlier ($E > 5.4kT$, $<0.45\%$ probability). The turn at residues 9-11 and, to an extent, the C-terminus of the α_1-helix at residue 33 are known to be dynamic.[21] The structure at these sites may agree with the motionally-averaged experimental RDC and NOE restraints, yet the average structure likely does not accurately represent the distribution of conformers at these sites. G75 is in the unfolded and dynamic C-terminal tail of ubiquitin,[22] and it visits a large range of backbone dihedral angles. By contrast, MolProbity does not identify backbone dihedral outliers for this structure; all dihedrals angles reside in 'allowed' regions.

In another example with mixed α/β secondary structure, the glutathione synthetase of *E. coli* (PDB: 1GLV, 303 a.a) has 26 warning and 2 outlier backbone dihedral angles, according to MolProbity. Mollib identifies 44 warning and 27 outlier backbone dihedral angles. The 2 outliers identified by MolProbity, P10 and Q315, have high energy penalties of 7.8kT and 12.0kT, respectively, in mollib. Mollib additionally identifies 5 outliers (F22, W137, L180, G181 and L305) with energies above 8.0 kT for their secondary structure classification.

Mollib's assignment of secondary structure units differs in some cases to DSSP, as a more conservative definition of hydrogen bonds is used (see the section on Hydrogen Bonds, below, for further details). Most residues in the NMR structure (PDB: 2MJB[23]) and crystal structure (PDB: 1UBQ[24]) for ubiquitin share the same classifications as DSSP (Table 1), yet differences arise at the termini of secondary structure units and the identification of turns and short helices.

Table 1

Mollib identifies secondary structure units based on contiguous stretches of hydrogen bonded residues with the appropriate dihedral angles. The α_1-helix in ubiquitin (residue 23-34) is correctly identified by both DSSP and mollib. However, residues 38-40 are identified as a 3_{10}-helix by DSSP, whereas mollib identifies this stretch as a type I turn since the 'i' and 'i+3' residues, which share the hydrogen bond, do not have helical dihedrals in the crystal and NMR structures. Likewise, mollib identifies residues 56-59 as a 3_{10}-helix in the crystal structure and a type I turn in the NMR structure. The backbone dihedral angles for these residues are inconsistent with a helix in 2MJB yet they are closer to helical dihedrals in 1UBQ.

An important classification of secondary structure unit that mollib disagrees with DSSP is the π-helix. π-helices are frequently involved in the active site of proteins, and they are also difficult to identify with DSSP and STRIDE.[25] In a clear example, DSSP identifies exclusively α-helical, turn and loop residues in the crystal structure of the LEUTAA (PDB: 2A65, 509 a.a.), a bacterial homolog of a Na^+/Cl^- transporter, yet this structure contains at least 8 π-helices.[25] Mollib identifies 13 π-helices in this structure, or 10 π-helices if helices are counted from contiguous stretches of residues with at least two π-helix hydrogen bonds, as done by Cooley *et al.*[25] Two π-helices are additionally interrupted briefly by α-helical residues (288-290/294-295 and 421-422/426-427), thereby increasing the total count of π-helices from 8 to 10 in mollib.

Hydrogen Bonds Plugin

Hydrogen bonds (SI manual pg 12) are identified based on the identity of acceptor and donor atoms and the geometry of their dipoles. Hydrogen bonds are characterized by at least three parameters (**Fig 2**): 1) the distance between the hydrogen bond donor and acceptor dipole atoms, typically the proton of the donor and the electronegative atom of the acceptor (d_{d1a1}), 2) the angle between the acceptor dipole and the donor hydrogen (θ), and 3) the twist of the acceptor plane (φ).

Fig 2

Mollib can be configured to identify hydrogen bonds for arbitrary hydrogen bond dipoles, including backbone-backbone amide hydrogen bonds, sidechain hydrogen bonds with amine, amide or hydroxyl groups, hydrogen bonds with non-protein molecules and aliphatic hydrogen bonds.

The hydrogen bond PMF energies are calculated using eq (0) with a set of hydrogenated, high-resolution structures—the same structures used to calculate the Ramachandran potentials. The calculation uses the same process as Grishaev and Bax in the Hydrogen Bond Database (HBDB).[20] These potential energy surfaces are highly similar yet mollib includes recent high-resolution structures (up to 2017), and the initial HBDB potential was calculated in 2004. Each hydrogen bond's energy is calculated from 1 of 15 potentials: 9 amide backbone-backbone potentials (α-helical, 3_{10}-helix, π-helix, sheet, type I, type I', type II and type II' turns and isolated) and 5 sidechain amide, amine or hydroxyl potentials.

In mollib v1.0, the energies (**Fig 3**) are calculated for hydrogen bonds with a donor-acceptor distance between 1.5 and 2.5Å, a θ angle between 90° and 180° and a φ angle between -180° and 180°. However, the default θ angle range is set between 105° and 180° to eliminate the overassignment of 3_{10}-helices and to more closely match DSSP assignments. 3_{10}-helices are distinct from the other hydrogen bond groups in our dataset, as these can have donor-acceptor distances above 2.5Å and more acute θ angles (<110°). The user may nevertheless configure the default hydrogen bond geometry ranges.

Fig 3

Mollib accurately identifies nearly all of the experimental ${}^{3h}J_{N\text{-}C'}$-coupling hydrogen bond restraints in the ubiquitin structure refinement (PDB: 2MJB).[23,26] The structure includes 31 ${}^{3h}J_{N\text{-}C'}$-couplings and an additional 6 restraints for a total of 37 hydrogen bonds. Mollib identifies 41 backbone-backbone (bb-bb) and 7 backbone-sidechain (bb-sc) hydrogen bonds (see SI Table S1). The 41 bb-bb hydrogen bonds include all of the restraints from the ubiquitin refinement, except for the P38O…Q41H hydrogen bond.

The experimental hydrogen bonds in the GB1 structure (PDB: 1PGB, 56 a.a.) are also accurately reproduced by mollib. GB1 has 34 bb-bb and 4 bb-sc experimental ${}^{3h}J_{N\text{-}C'}$-coupling restraints.[27] Mollib identifies 46 hydrogen bonds (Table S2), including 33 of the 34 bb-bb experimental hydrogen bonds and all 4 of the bb-sc hydrogen bonds. All hydrogen bonds have favorable probabilities and energies, with the exception of the 1) K28H…A24O, 2) N37H…Y33O, 3) F52H…K4O 6) V54H…I6O bb-bb hydrogen bonds, which are marked as warnings outliers (E > 3.4 kT, <3.3% probability).

Structures optimized with the HBDB potential tend to have fewer hydrogen bond outliers. The recently published ubiquitin NMR structure refined with HBDB (PDB: 2MJB) has 4 warning outliers out of 48 hydrogen bonds whereas the ubiquitin NMR structure refined without HBDB (PDB: 1D3Z) has 5 warning outliers and 2 outliers out of 52 hydrogen bonds. By default, warning outliers are defined as hydrogen bonds with an E > 3.4kT (probability < 3.3%) and outliers have an E > 5.4kT (probability < 0.45%). The energies have been rescaled such that a 100% probability (E = 0 kT) is defined as the most likely configuration for a hydrogen bond in a given classification group.

Partial Alignment Plugin

The *partial alignment* (pa, SI manual pg 17) plugin is used to fit RDC and RACS NMR data. The *pa* plugin includes multiple useful and convenient features not available in similar software packages, including the automatic fetching and fitting of data submitted to the PDB, the hydrogenation of molecules, the fitting of arbitrary CSA tensors, the incorporation of multiple structures, and the inclusion of error analysis and dataset 'fixers.' Mollib uses a simplified data text file format that simply lists the interaction (e.g. '14N-H', '18C', 'A.15H-18HA#' or '14N-C-1') in the first column, the value in the second column and the optional error in the third. Existing datasets in Xplor-NIH[28] and DC can also be used, and a future version of mollib will include additional data formats, including the Biological Magnetic Resonance Data Bank (BMRB) NMRStar format.

To our knowledge, Module 2 is the only publically available software package that currently fits RACS data.[16] Module 2 can only fit a single structure at a time, and backbone ^{15}N and ^{13}C' CSA tensors are supported. Mollib includes static tensors for backbone ^{1}H^{N}, ^{13}C' and ^{15}N nuclei, based on values fit to the RACS dataset on ubiquitin,[29,30] and other tensors can be easily integrated through the settings.

RDC and RACS values are fit to a molecular structure using a Singular Value Decomposition (SVD).[31,32] The RDC between nuclei 'i' and 'j' or the RACS for nucleus 'i' are fit using the A-matrix.

$$\begin{pmatrix} A_{yy}^{(1)} & A_{zz}^{(1)} & A_{xy}^{(1)} & A_{xz}^{(1)} & A_{yz}^{(1)} \\ A_{yy}^{(2)} & A_{zz}^{(2)} & A_{xy}^{(2)} & A_{xz}^{(2)} & A_{yz}^{(2)} \\ . & . & . & . & . \\ . & . & . & . & . \\ A_{yy}^{(N)} & A_{zz}^{(N)} & A_{xy}^{(N)} & A_{xz}^{(N)} & A_{yz}^{(N)} \end{pmatrix} \begin{pmatrix} S_{xx} \\ S_{zz} \\ S_{xy} \\ S_{xz} \\ S_{yz} \end{pmatrix} = \begin{pmatrix} V^{(1)} \\ V^{(2)} \\ . \\ . \\ V^{(N)} \end{pmatrix} \tag{0}$$

The S_{ij} terms are the Saupe matrix components, and the $V^{(m)}$ values are the observed RDCs or RACSs. The A-matrix rows for each interaction 'm' are constructed as follows.[32]

$$A_{ab}^{(m)} = \begin{cases} \Delta_{ab}\cdot\frac{2}{3}\sum\limits_{i=x,y,z}\delta_{ii}^{(m)}(\cos^2\theta_{ai}-\cos^2\theta_{xi})+(1-\Delta_{ab})\cdot\frac{4}{3}\sum\limits_{i=x,y,z}\delta_{ii}^{(m)}\cos\theta_{ai}\cos\theta_{bi} & RACS \\ \Delta_{ab}\cdot\delta_{zz}^{(m)}\cdot(\cos^2\varphi_a-\cos^2\varphi_x)+(1-\Delta_{ab})\cdot\delta_{zz}^{(m)}\cdot 2\cos\varphi_a\cos\varphi_b & RDC \end{cases} \qquad (0)$$

The $\cos\square_{ai}$ and $\cos\varphi_a$ terms are the directional cosines about the 'a' axis between the principal axis system (PAS) and the molecular frame for the RACS and RDC interactions, respectively. The Δ_{ab} terms are the delta functions about the 'a' and 'b' axes. The $\delta_{ii}^{(m)}$ and $\delta_{zz}^{(m)}$ terms are the magnitude of the principal components for the 'i' axis of the RACS tensor or the RDC tensor. Mollib can use pre-calculated values for these, or in the case of RDCs, they can be calculated from gyromagnetic ratios and internuclear distances. Additionally, mollib scales these components by the error ($\sigma^{(m)}$) for each interaction.

$$\begin{aligned} \delta_{ii}^{(m)} &= \frac{\delta_{ii,static}}{\sigma^{(m)}} \quad RACS \\ \delta_{zz}^{(m)} &= \frac{\delta_{zz,static}}{\sigma^{(m)}} \quad RDC \end{aligned} \qquad (0)$$

Experimental errors can be specified for each measurement, or if individual values are not specified, default values for each interaction type are used. With this approach, the predicted RDC and RACS values must also be multiplied by their respective errors, after the SVD of the A-matrix.

For ubiquitin, the static chemical *shift* anisotropy (CSA) values for the backbone ^{1}H^N, ^{13}C' and ^{15}N nuclei are based on ubiquitin datasets in bicelles.[30] The CSA tensors are calculated based on the static ^{1}H-^{15}N dipolar coupling constant (reduced dipolar anisotropy) of 10,823 Hz, corresponding to a motionally averaged bond length of 1.04Å. This value was used to match the fitting of datasets in Cornilescu *et al.* [30,33]

Mollib uses the Z-Y-X convention for CSA tensor rotations from the PAS (**Fig 4**) and the Rose Z-Y-Z Euler convention[34] to describe the orientation of alignment tensors with respect to the molecular frame of the structure. The optimized ^{13}C' reduced anisotropy ($\delta_{C'}$) is -89.1 ppm, the asymmetry (η) is 0.63 and the α_Z angle is 40 degrees. For the ^{15}N tensor, the optimized δ_N is 107.8 ppm, η is 0.16 and the β_Y angle is -20 degrees. For the ^{1}H^N tensor, the optimized δ_H is -5.8 ppm, η is 1.00 and the γ_X angle is -7 degrees. Note that these values are for the chemical shift anisotropy whereas Cornilescu *et al.*[29] reports the chemical *shielding* anisotropy, which is different by a factor of -1.[35] The scatter between the predicted and experimental RACS values is attributable, in part, to site-specific variations in the chemical shift tensor values. These differences are most pronounced for the $^1H^N$ CSA tensors, which vary significantly and may not be accurately described by average values.

Fig 4

In the analysis of RDCs and RACSs together, the reduced anisotropy of dipolar couplings must be doubled in equation (0) since the RDC values themselves are twice as large as their actual values, when measured from the J-coupling. This fact can be verified independently by measuring the span of RDC and RACS values and scaling them by their reduced anisotropies. The factor of 2 does not impact the quality of the SVD fit with RDCs alone. In this case, the fit D_a values themselves will be different by a factor of 2, if the reduced anisotropy of the dipolar coupling is used directly to fit the RDCs.

The sign of RDCs must also be correctly incorporated in the analysis. It is common for ^1H-^{15}N RDCs to be referenced to a positive J-coupling value (*ca.* 93 Hz). To get the correct sign of the anisotropies for the RACS, the correct J-coupling value of *ca.* -93 Hz must be used.

Mollib conveniently includes dataset fixers to easily correct problems with signs in RDC and RACS datasets as well as other problems (SI manual pg 17). Other fixers include an outlier fixer and a 'NH' scale fixer. The outlier fixer identifies outliers in the fit using a Grubbs test, and it reports the fit statistics with and without the outlier points included. The 'NH' scale fixer identifies RDC couplings that have been scaled to match the magnitude of 'NH' RDCs, and it scales these measurements to their original values.

The mollib partial alignment fit to the RDC and RACS data for ubiquitin[30] shows good agreement (**Fig 5**) with the recently published ubiquitin NMR structure[23] (PDB: 2MJB).

Fig 5

The goodness of fit can be evaluated with the Q-factor of the observed and predicted values. The Q-factor is calculated as follows[36]:

$$Q=\sqrt{\sum_{m}\frac{\left(V_{obs}^{(m)}-V_{pred}^{(m)}\right)^{2}}{D_{a,m}^{2}[4+3R_{m}^{2}]/5}} \tag{0}$$

The $D_{a,m}$ and R_m are the magnitude and rhombicity of the dipolar coupling or CSA tensor in the alignment frame [36] for the interaction type of data point 'm.' They adopt a single value for each interaction type based on the reduced anisotropy of the interaction. The Q-factor can be calculated over all RDCs and RACSs or for a subset of interaction types.

The fit can alternatively be calculated from the root mean-squared deviation (RMSD) over the 'N' measurements.

$$RMSD=\sqrt{\sum_{m}^{N}\left(V_{obs}^{(m)}-V_{pred}^{(m)}\right)^{2}/(N-1)} \tag{0}$$

The mollib fits and statistics agree well with previous reports and other software packages (Table 2). The mollib RMSDs for the C', N and ^{H}N RACS are comparable to those reported by Cornilescu *et al.*[30] The discrepancy between these values likely arises from the non-linear regression used in Cornilescu and the SVD fit used by mollib. Likewise, Module 2 uses a Monte-Carlo fitting procedure to fit the HN RDCs and C' and N RACS values. Finally, the reported tensor values are nearly identical to the NMRPipe DC package. The current version of NMRPipe DC only fits RDCs, and the fit ^{1}H-^{15}N RDC RMSD, Q-factor and Saupe tensor values are nearly indistinguishable between mollib and DC.

Table 2

The ubiquitin crystal structure (PDB: 1UBQ) can likewise be fit to the same RDC and RACS dataset, by first hydrogenating the structure in mollib with the '--hydrogenate' option (see SI manual pg 20-21 for examples). The fit Q-factors for the crystal structure are Q_{NH}=16.1%, $Q_{C'}$=27.2%, Q_N=17.9%, Q_H=47.6%, and the overall Q-factor is 30.0%.

Mollib supports the fitting of multiple structures to RDC and RACS data by extending the A-matrix and Saupe matrix in equation (0). The refinement of structural ensembles is also supported by REDCAT.[37] For a single structure, the A-matrix has N x 5 components, for 'N' measurements, and the Saupe vector has 5 x 1 components. For 'M' structures, the A-matrix has N x 5M components and the Saupe vector has 5M x 1 components.

The ubiquitin RDCs and RACSs can be fit to a variety of crystal structures to emulate the conformational distribution of the molecule in solution.[23] Table 3 shows the average Q-factors for ^{1}H-^{15}N RDCs and ^{13}C' RACSs for datasets collected from Pf1, squalamine and two bicelle alignment media. The fits are conducted individually for each crystal structure and in aggregate with all crystal structures. For individual molecules, the Q_{NH}-factors are slightly higher than previously presented by Maltsev *et al.*,[23] which can be attributed to the deviation from ideal H^{N} geometry used in that report. The reduction in Q-factors for the aggregate of structures, calculated over the ensemble, [38] is also consistent with Matlsev *et al.*, as these show a significantly improved fit.

Table 3

The fit statistics reported here are lower since a greater number of degrees of freedom were used in the fits. Matlsev *et al.* used a Monte Carlo procedure with one population for each structure (14 free parameters) and 1 Saupe matrix for each alignment medium (20 free parameters for 4 alignment media). In the currently implementation of mollib, the SVD fits the Saupe matrix for each structure independently, thus producing 5 free parameters for each of the 15 structures. We chose this approach since the Saupe matrix components include the population for each conformer as well as deviations in the α, β and γ Euler angles for different alignments of the structures. Ensembles with similar structures likely fit only a small number of alignment tensors, far fewer than the 15 used here. An SVD with too many alignment tensors will overfit the data and produce artificially small Q-factors and RMSDs, as observed in our ubiquitin ensemble Q-factors. A more rigorous approach would use a non-linear fitting routine with a statistical analysis or a cross-validation procedure to evaluate the minimum number of alignment tensors needed to fit the data. In a future version of mollib, we plan to integrate a non-linear fit and an F-test to evaluate the minimum number of alignment tensors needed to fit an ensemble of structures.

For an ensemble with very different structures, a larger number of alignment tensors is reasonable. An example of an RDC dataset for an ensemble with very different structures is the G8A mutant of the influenza hemagglutinin fusion peptide domain (HAfp-G8A) in dodecylphosphocholine (DPC) micelles.[38] HAfp-G8A adopts a closed, helical-hairpin structure with a 15% population and two open structures (L-shaped and extended) with an 85%

population. The three structures are available in the PDB (PDB: 2LWA, 24 a.a.) and can be fit to the ^{1}H-^{15}N and ^{1}H$^{\alpha}$-^{13}C$^{\alpha}$ RDCs of HAfp-G8A. The Q-factors of the fit are Q_{NH}=2.4% and $Q_{C\alpha H\alpha}$=4.2%, and the overall Q-factor is 3.4%. The very low Q-factors are expected because the same RDCs were used in the refinement of the ensemble structures.[38]

The automatic fetching of structures and datasets can be conveniently used to analyze many NMR structures in the PDB. A table with example fits of PDB structures and RDC datasets is presented in the SI (Table S3) Nearly all of the deposited structures fit well to the deposited RDC data. In most cases, the sign of the deposited NH RDCs must be inverted. In one case, a difference in the assignment labels of RDCs and the atoms in the PDB file produces an artificially large Q_{NH}-factor of 92.3%. A few of the datasets have very low Q-factors, suggesting an overfitting of the RDC data. Some datasets include data from methyls that have been projected onto the C-C or C-S bond vector. These can be fit with the '--project-methyls' option. Methyl ^{1}H-^{13}C RDCs can also be scaled by a user-specified order parameter to account for the motional averaging of the methyl group rotation.

Regression Testing

More than a fifth of the mollib source code consists of regression tests to verify the accuracy and reproducibility of data analyses and reporting. These tests enable developers to make modifications to mollib while ensuring that the results remain consistent between versions. Mollib further incorporates testing tools to ensure compatibility with a wide range of python environments and operating systems.

Mollib includes 3 classes of tests: regression unit tests, doctests and CLI tests. The regression unit tests, in the 'tests' directory, verify that the results of functions are consistent and accurate. The doctests are located within the source code, and they verify the accuracy of the source code documentation examples. Finally, the CLI tests, located in the 'tests/cli' directory, verify the accuracy of commands executed from the command line and those reported in mollib's documentation.

Documentation and Open Source

The mollib source code is open source (GNU Public License v3) and tracked in a git repository currently hosted on github. The distributed nature of git allows users and developers to download the mollib source code and view every change committed to the source code. Citing a specific version of mollib in an article allows readers to load the repository for that version and inspect the specific algorithm used in an analysis.

Mollib includes extensive documentation on the CLI and API in both portable document format (PDF) and hypertext markup language (HTML) format. The documentation source code is also tracked in the same git repository and any version can be retrieved.

Conclusions

We present mollib as a useful molecular and data analysis tool. Mollib simplifies the analysis of molecular structure geometries, hydrogen bonds and partial alignment NMR datasets. It uses a plugin framework to easily extend its functionality, and its open source code promotes contributions from other users. It includes extensive documentation and regression testing, and it implements many useful tools to NMR spectroscopists and other biophysical researchers.

Acknowledgements

This work was supported by the NSF (#1651598).

References

1. Kabsch, W. & Sander, C. Dictionary of protein secondary structure: Pattern recognition of hydrogen bonded and geometrical features. *Biopolymers* **22,** 2577–2637 (1983).
2. Davis, I. W., Murray, L. W., Richardson, J. S. & Richardson, D. C. MOLPROBITY: structure validation and all-atom contact analysis for nucleic acids and their complexes. *Nucleic Acids Res.* **32,** W615-9 (2004).
3. Chen, V. B. *et al.* MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr. D. Biol. Crystallogr.* **66,** 12–21 (2010).
4. Laskowski, R. A., Rullmannn, J. A., MacArthur, M. W., Kaptein, R. & Thornton, J. M. AQUA and PROCHECK-NMR: programs for checking the quality of protein structures solved by NMR. *J. Biomol. NMR* **8,** 477–486 (1996).
5. Willard, L. *et al.* VADAR: A web server for quantitative evaluation of protein structure quality. *Nucleic Acids Res.* **31,** 3316–3319 (2003).
6. Huang, Y. J., Powers, R. & Montelione, G. T. Protein NMR recall, precision, and F-measure scores (RPF scores): Structure quality assessment measures based on information

retrieval statistics. *J. Am. Chem. Soc.* **127,** 1665–1674 (2005).

7. Laskowski, R. A., MacArthur, M. W., Moss, D. S. & Thornton, J. M. PROCHECK: a program to check the stereochemical quality of protein structures. *J. Appl. Crystallogr.* **26,** 283–291 (1993).

8. Davis, I. W. *et al.* MolProbity: all-atom contacts and structure validation for proteins and nucleic acids. *Nucleic Acids Res.* **35,** W375-83 (2007).

9. Lüthy, R., Bowie, J. U. & Eisenberg, D. Assessment of protein models with three-dimensional profiles. *Nature* **356,** 83–85 (1992).

10. Sippl, M. J. Recognition of Errors in 3-Dimensional Structures of Proteins. *Proteins-Structure Funct. Genet.* **17,** 355–362 (1993).

11. Bhattacharya, A., Tejero, R. & Montelione, G. T. Evaluating protein structures determined by structural genomics consortia. *Proteins* **66,** 778–95 (2007).

12. Berjanskii, M., Zhou, J., Liang, Y., Lin, G. & Wishart, D. S. Resolution-by-proxy: A simple measure for assessing and comparing the overall quality of NMR protein structures. *J. Biomol. NMR* **53,** 167–180 (2012).

13. Delaglio, F. *et al.* NMRPipe: A multidimensional spectral processing system based on UNIX pipes. *J. Biomol. NMR* **6,** 277–93 (1995).

14. Valafar, H. & Prestegard, J. H. REDCAT: A residual dipolar coupling analysis tool. *J. Magn. Reson.* **167,** 228–241 (2004).

15. Zweckstetter, M. NMR: prediction of molecular alignment from structure using the PALES software. *Nat. Protoc.* **3,** 679–90 (2008).

16. Dosset, P., Hus, J. C., Marion, D. & Blackledge, M. A novel interactive tool for rigid-body modeling of multi- domain macromolecules using residual dipolar couplings. *J. Biomol. NMR* **20,** 223–231 (2001).

17. Gajda, M. J. hPDB – Haskell library for processing atomic biomolecular structures in protein data bank format. *BMC Res. Notes* **6,** (2013).

18. Word, J. M., Lovell, S. C., Richardson, J. S. & Richardson, D. C. Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation. *J. Mol. Biol.* **285,** 1735–1747 (1999).

19. Markley, J. L. *et al.* Recommendations for the presentation of NMR structures of proteins and nucleic acids. IUPAC-IUBMB-IUPAB Inter-Union Task Group on the Standardization of Data Bases of Protein and Nucleic Acid Structures Determined by NMR Spectroscopy. *J. Biomol. NMR* **12,** 1–23 (1998).

20. Grishaev, A. & Bax, A. An empirical backbone-backbone hydrogen-bonding potential in proteins and its applications to NMR structure refinement and validation. *J. Am. Chem. Soc.* **126,** 7281–92 (2004).

21. Ma, P. *et al.* Observing the overall rocking motion of a protein in a crystal. *Nat. Commun.* **6,** 8361 (2015).

22. Schneider, D. M., Dellwo, M. J. & Wand, A. J. Fast internal main-chain dynamics of human ubiquitin. *Biochemistry* **31,** 3645–3652 (1992).

23. 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. *J. Am. Chem. Soc.* **136,** 3752–5 (2014).

24. Vijay-kumar, S., Bugg, C. E. & Cook, W. J. Structure of ubiquitin refined at 1.8 Å resolution. *J. Mol. Biol.* **194,** 531–544 (1987).

25. Cooley, R. B., Arp, D. J. & Karplus, P. A. Evolutionary Origin of a Secondary Structure: pi-Helices as Cryptic but Widespread Insertional Variations of &alpha;-Helices That Enhance Protein Functionality. *J. Mol. Biol.* **404,** 232–246 (2010).

26. Cordier, F. & Grzesiek, S. Direct Observation of Hydrogen Bonds in Proteins by Interresidue 3hJNC‘Scalar Couplings. *J. Am. Chem. Soc.* **121,** 1601–1602 (1999).

27. Cornilescu, G. *et al.* Correlation between 3h J NC ‘ and Hydrogen Bond Length in Proteins. *J. Am. Chem. Soc.* **121,** 6275–6279 (1999).

28. Schwieters, C. D., Kuszewski, J. J., Tjandra, N. & Clore, G. M. The Xplor-NIH NMR molecular structure determination package. *J. Magn. Reson.* **160,** 65–73 (2003).

29. Cornilescu, G., Marquardt, J. L., Ottiger, M. & Bax, A. Validation of Protein Structure from Anisotropic Carbonyl Chemical Shifts in a Dilute Liquid Crystalline Phase. *J. Am. Chem. Soc.* **120,** 6836–6837 (1998).

30. Cornilescu, G. & Bax, A. Measurement of Proton, Nitrogen, and Carbonyl Chemical Shielding Anisotropies in a Protein Dissolved in a Dilute Liquid Crystalline Phase. *J. Am. Chem. Soc.* **122,** 10143–10154 (2000).

31. Losonczi, J. a, Andrec, M., Fischer, M. W. & Prestegard, J. H. Order matrix analysis of residual dipolar couplings using singular value decomposition. *J. Magn. Reson.* **138,** 334–42 (1999).

32. Liu, Y. & Prestegard, J. H. A device for the measurement of residual chemical shift anisotropy and residual dipolar coupling in soluble and membrane-associated proteins. *J. Biomol. NMR* **47,** 249–258 (2010).

33. Yao, L., Vögeli, B., Ying, J. & Bax, A. NMR determination of amide N-H equilibrium bond length from concerted dipolar coupling measurements. *J. Am. Chem. Soc.* **130,** 16518–20 (2008).

34. Rose, M. E. *Elementary Theory of Angular Momentum*. *Dover* (Dover Publications, Inc., 1957). doi:10.1007/BF02860403

35. Saitô, H., Ando, I. & Ramamoorthy, A. Chemical shift tensor - the heart of NMR: Insights into biological aspects of proteins. *Prog. Nucl. Magn. Reson. Spectrosc.* **57,** 181–228 (2010).

36. Bax, A. Weak alignment offers new NMR opportunities to study protein structure and dynamics. *Protein Sci.* **12,** 1–16 (2003).

37. Schmidt, C., Irausquin, S. J. & Valafar, H. Advances in the REDCAT software package. *BMC Bioinformatics* **14,** 302 (2013).

38. Lorieau, J. L., Louis, J. M., Schwieters, C. D. & Bax, A. pH-triggered, activated-state conformations of the influenza hemagglutinin fusion peptide revealed by NMR. *Proc. Natl. Acad. Sci. U. S. A.* **109,** 19994–9 (2012).

39. Clore, G. M. & Garrett, D. S. R-factor, Free R , and Complete Cross-Validation for Dipolar Coupling Refinement of NMR Structures. *J. Am. Chem. Soc.* **121,** 9008–9012 (1999).

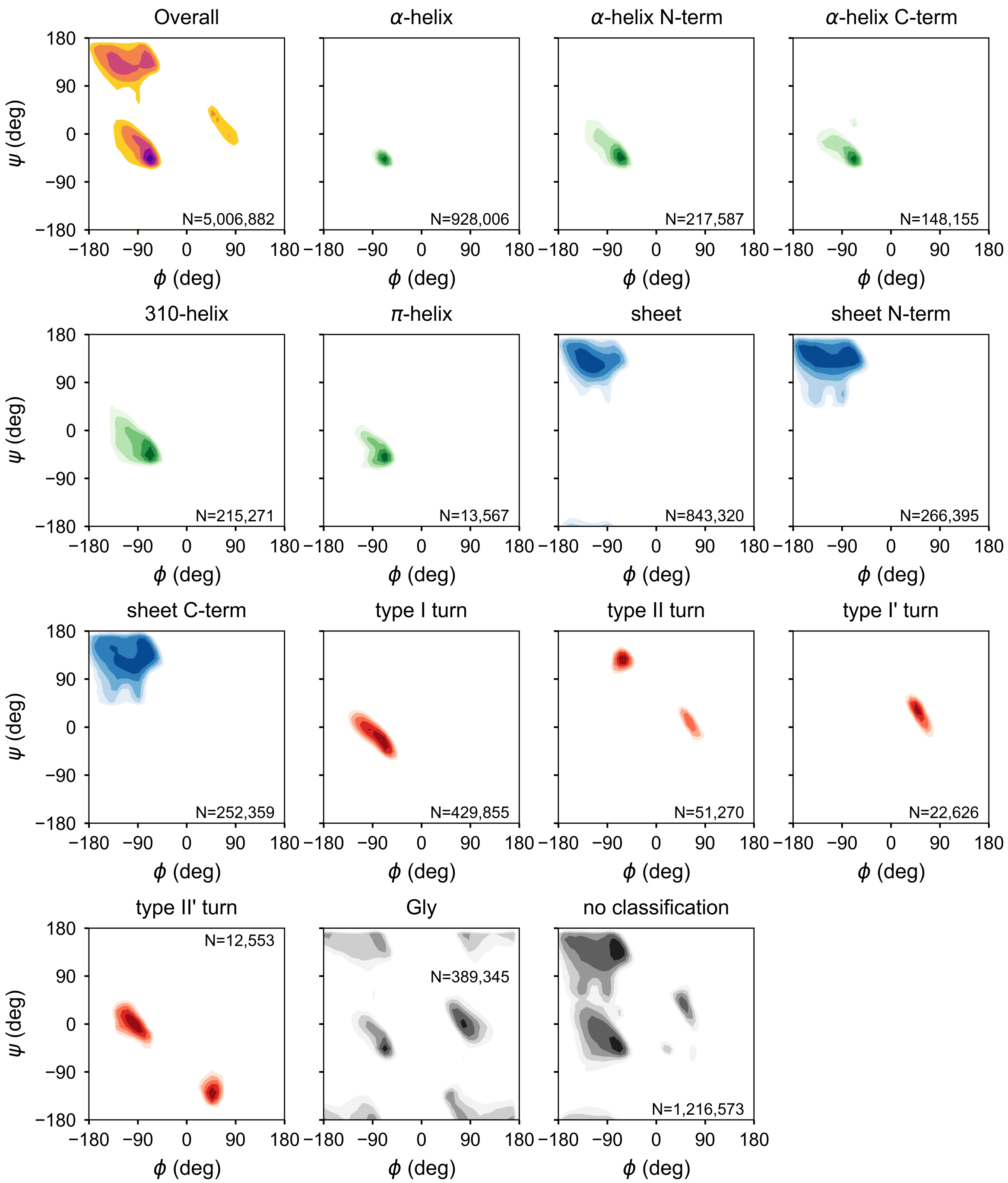

Fig 1. Mollib Ramachandran potential energy contour plots for common secondary structure elements from high-resolution crystal structures in the PDB. Each contour represents one 'kT' level in the energy plotted from eq (0).

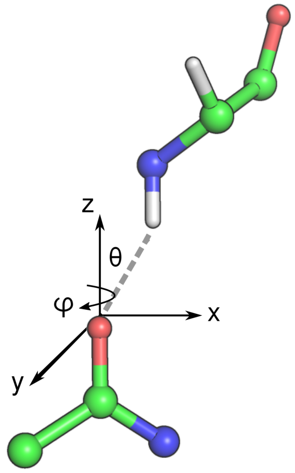

Fig 2. Geometry of a hydrogen bond. A hydrogen bond's geometry is defined by the distance d_{a1d1}, the angle θ and the angle φ. The d_{a1d1} distance represents the length of the vector, shown as a dash grey line, between the first donor atom (H^N) and the first acceptor atom (O). It typically adopts a value of *ca.* 2.0Å for backbone amide hydrogen bonds in proteins. The θ angle represents the angle between the acceptor dipole vector and the d1 donor atom (H^N). The φ angle represents the rotation of the x- and y- axes. The x-axis is defined by the plane including the acceptor dipole (O-C) and the next heaviest atom (N).

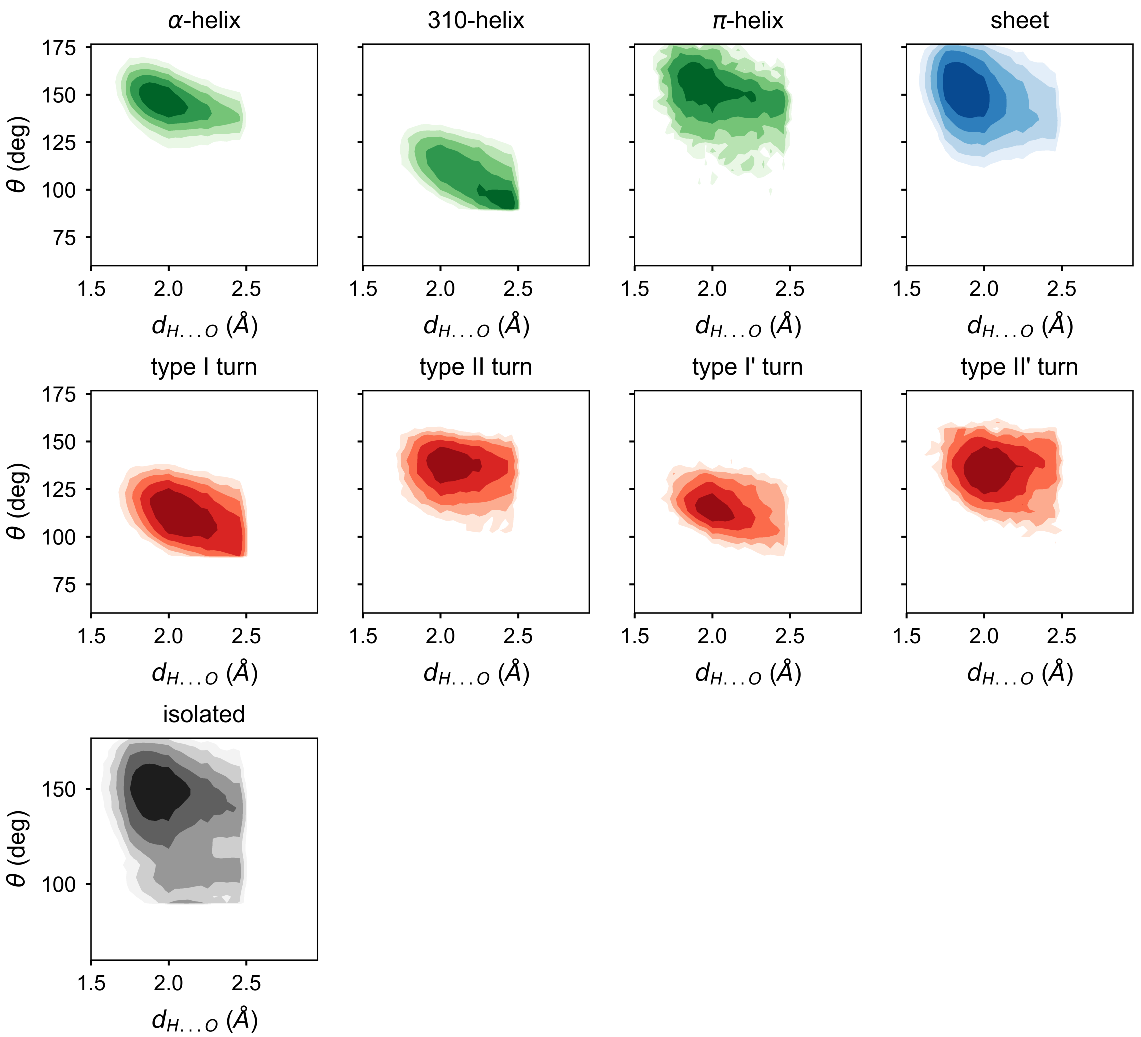

Fig 3. The mollib hydrogen bond potential energy contour plots for common secondary structure elements from high-resolution crystal structures in the PDB. Each contour represents one 'kT' level in the energy plotted from eq (0). The plots show the 2D correlation between the $d_{H...O}$ distance and the θ angle, with the third variable (φ) projected. Hydrogen bonds were detected for $d_{H...O}$ distance ranges between 1.5 and 2.5Å, θ angles between 90º and 180º and φ angles between -180º and 180º.

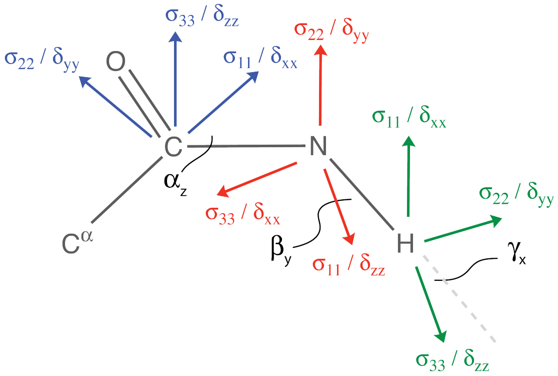

Fig 4. Protein backbone CSA tensor conventions used to fit RACS data in the mollib *partial alignment* plugin. The CSA tensors for the backbone ^{13}C' (blue), ^{15}N (red) and ^{1}H (green) are shown with the Haeberlen convention (δ) and the chemical shielding (σ) convention. Tensor conventions follow those reported in Cornilescu *et al.*[29,30] The ^{13}C' CSA tensor is defined by the axis orthogonal to the peptide plane (zz) and the C'-N vector, and it is rotated about the zz-axis by an angle α_Z. The ^{15}N CSA tensor is defined by the axis orthogonal to the peptide plane (yy) and the N-H vector (zz), and it is rotated about the yy-axis by an angle β_Y. The ^{1}H^{N} CSA tensor is defined by the N-H vector (zz) and the vector orthogonal to the peptide plane (xx), and it is rotated about the xx-axis by an angle γ_X.

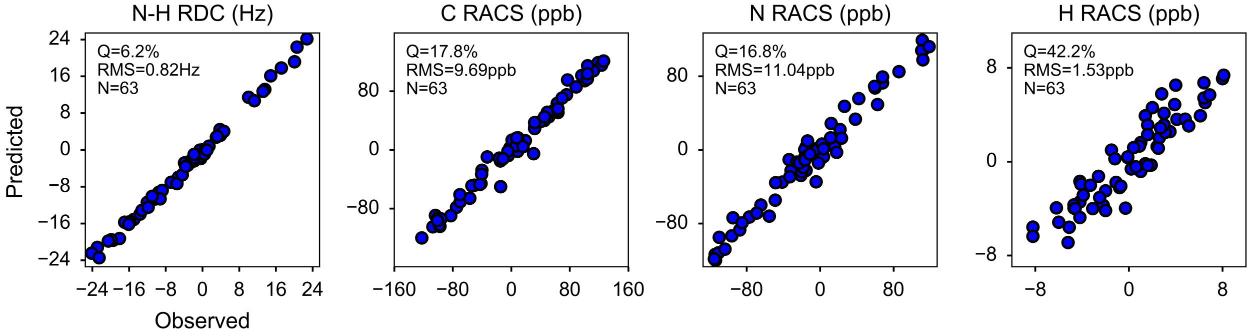

Fig 5. The partial alignment fit of the ubiquitin RDC and RACS data (Cornilescu *et al*.[30]) to the ubiquitin NMR structure (PDB: 2MJB). The observed and predicted RDC values (in Hz) and RACS values (in parts per billion, ppb) are shown.

Table 1: Secondary structure assignment of ubiquitin structures with DSSP and mollib.

Sequence	MQIFVKTLTG KTITLEVEPS DTIENVKAKI QDKEGIPPDQ QRLIFAGKQL
2MJB: DSSP	EEEEE TTS EEEEE TT BHHHHHHHH HHHH GGG EEEEETTEE
2MJB: mollib	EEEEEETT EEEEEE TT HHHHHHHH HHHH TT EEEEETT EE
1UBQ: DSSP	EEEEEETTS EEEEE TT SBHHHHHHHH HHHH GGG EEEEETTEE
1UBQ: mollib	EEEEEETT EEEEEE TT HHHHHHHH HHHH TTT EEEEETT EE
Sequence	EDGRTLSDYN IQKESTLHLV LRLRGG
2MJB: DSSP	TTSBTGGGT TT EEEEE E
2MJB: mollib	ETT TT TTEEEEEE E
1UBQ: DSSP	TTSBTGGGT TT EEEEE E S
1UBQ: mollib	ETT GGGG TTEEEEEE E

Legend: E: β-strand; T: turn; S: bend; B: isolated β-bridge, H: α-helix; G: 3_{10}-helix. Residues in the sequence that have been double underlined represent differences in assignment between DSSP and mollib.

Table 2: Summary of the fit statistics for RDCs and RACSs to the ubiquitin structure [a]

Analysis	RMSD NH (Hz)	RMSD C' (ppb)	RMSD N (ppb)	RMSD H^N (ppb)	Q_{NH} (%)	$Q_{C'}$ (%)	Q_N (%)	Q_H (%)
Mollib	0.82	9.7	11.0	1.5	6.2	17.8	16.8	42.2
Cornilescu et al.[30]	-	9.9	10.5	1.5	-	14 [b]	17 [b]	38 [b]
Module 2 [c]	0.80	14.1	13.9	-	-	-	-	-
NMRPipe[13] DC [d]	0.72	-	-	-	5.5	-	-	-

a. Data were fit against the ubiquitin structure with the PDB accession code 2MJB.[23] The experimental ^{1}H-^{15}N RDCs, ^{13}C' RACSs and ^{15}N RACS are from ubiquitin in bicelles doped with CHAPS (Cornilescu 2000[30]). Values marked with a '-' are not reported.

b. These Q-factors were calculated with the equation $Q = RMS(V_{obs} - V_{pred}) / RMS(V_{obs})$. The Q-factor presented in equation (0) calculates the denominator from a random distribution of vectors.[39]

c. Reported values use the default tensor values in Module 2. The ^{13}C' CSA tensor values used in Module 2 were σ_{zz}=+86.5 ppm, σ_{xx}=-74.7 ppm and σ_{yy}=-11.8 ppm and θ_{tilt} = 38º. The ^{15}N CSA tensor values used in Module 2 were σ_{zz}=+62.8 ppm, σ_{xx}=-108.5 ppm and σ_{yy}=+45.7 ppm and θ_{tilt} = -18º. Module2 currently does not fit H^N CSA tensors or report Q-factors.

d. NMRPipe DC only supports RDCs, and the reported statistics represent the fit with only the NH RDCs. The SVD values of the fit tensor are D_a(HN)=12.9 Hz, Rh=0.616, $S_{xx}=-4.55\cdot10^{-5}$, $S_{yy}=-1.15\cdot10^{-3}$ and $S_{zz}=1.19\cdot10^{-3}$. The corresponding SVD values with mollib for the HN RDCs are D_a(HN)=12.9 Hz, Rh=0.616, $S_{xx}=-4.53\cdot10^{-5}$, $S_{yy}=-1.14\cdot10^{-3}$ and $S_{zz}=1.19\cdot10^{-3}$ with an RMS NH of 0.72 Hz and a Q_{NH}-factor of 5.5%.

Table 3: Summary of the fit statistics for $^{1}H^{15}N$ RDCs and $^{13}C'$ RACSs to ubiquitin crystal structures with a resolution below 1.8Å.

PDB ID	Chain ID	Resolution (Å)	bb RMSD (Å) [a]	Q_{NH} (%) [b]	$Q_{13C'}$ / RMS (% / ppb) [b]
1P3Q	U	1.70	0.639	23.0	27.8 / 11.9
	V	1.70	0.619	31.3	32.8 / 14.3
1UBI	A	1.80	0.319	20.7	25.4 / 11.2
1UBQ	A	1.80	0.325	19.2	28.9 / 12.6
1WRD	B	1.75	0.397	31.5	29.9 / 12.9
2D3G	A	1.70	0.362	24.4	29.7 / 13.1
	B		0.532	24.9	31.8 / 13.7
2ZCC	A	1.40	0.466	26.9	24.9 / 11.1
	B		0.312	24.9	22.9 / 10.1
	C		0.539	26.7	21.8 / 9.3
2ZNV	B	1.60	0.334	22.8	24.3 / 10.7
	C		0.652	36.8	27.9 / 12.1
	E		0.418	24.8	22.1 / 9.8
3BY4	B	1.55	0.334	26.2	28.2 / 12.4
3ONS	A	1.80	0.363	23.4	23.1 / 10.2
All X-ray			0.457	2.0	7.5 / 4.5

a. Calculated relative to the average structure for the backbone heavy atoms of residues 2-69.

b. Calculated for H-N RDCs between Q2-V70. The Q-factors are calculated as an average for 4 datasets: Pf1, squalamine,[23] bicelles[29] and bicelles with CTAB.[30] The SVD was conducted with the '--nofix-sign' option.