AssignSLPGUI Tutorial

Robert V. Williams, V. Paduchuri, James H. Prestegard

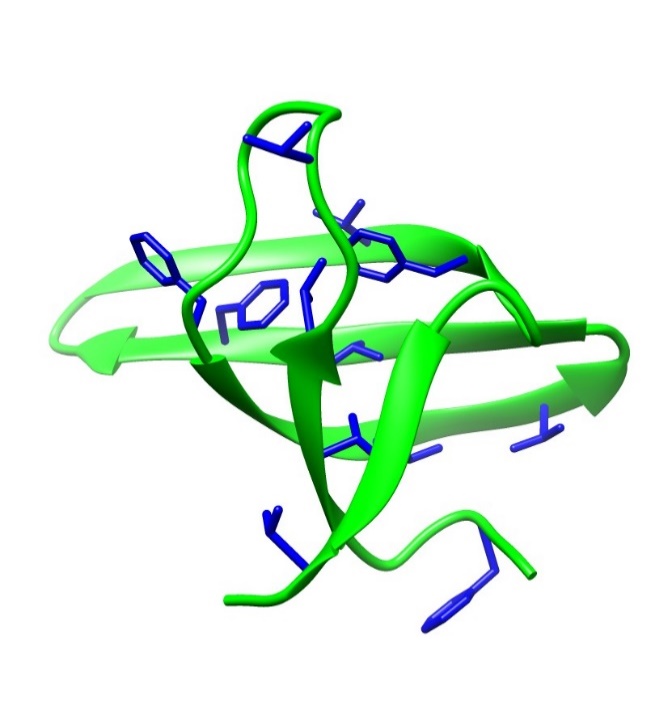
*Updated for v2.4.9 on 9/9/2024 J. Prestegard (jpresteg@uga.edu)*

**Overview**

AssignSLPGUI is a software package for NMR resonance assignment of sparsely labeled proteins. Sparse labeling is typically done by adding a single (or a few) types of isotopically labeled amino acid to a protein expression medium. It is particularly useful for large proteins and proteins expressed in cell cultures that require supplementation with amino acids in addition to more basic metabolic substrates. The package focuses on assignment of simple 2D (15N-1H or 13C-1H)spectra (HSQC or HETCOR) that are often used in subsequent protein-protein or ligand-protein association studies. It assumes that a good structural model for the protein already exists, from experimental or computational studies. This updated version incorporates a graphical user interface (GUI) and builds upon previous work by Gordon Chalmers. It is built on a MATLAB platform and may be run either on systems with a licensed version of MATLAB or on systems having a compiled version and the free MATLAB runtime library. If you use this software, please cite:

Williams, RV; Rogals, MJ; Eletsky, A; Huang, C; Morris, LC; Moremen, KW; Prestegard, JH; (2022). “AssignSLP\_GUI, a software tool exploiting AI for NMR resonance assignment of sparsely labeled proteins”, *Journal of Magnetic Resonance***, 345**:1-9.

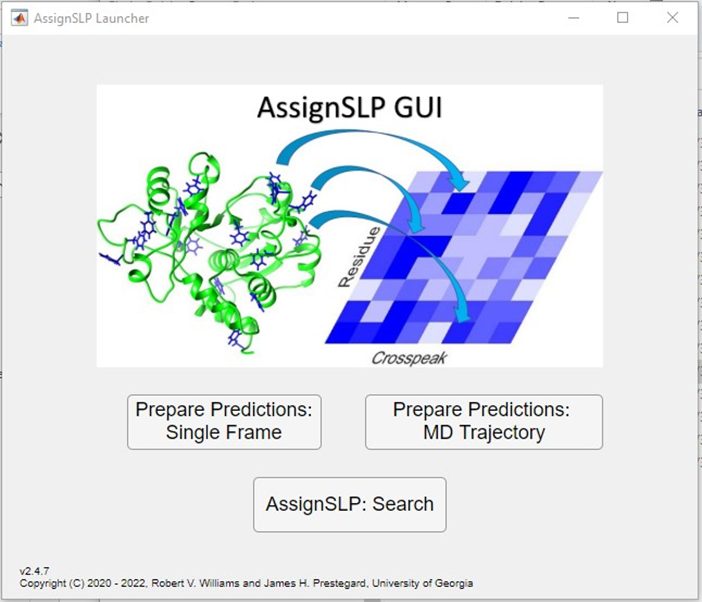
1. **Introduction**

This tutorial will walk you through running AssignSLPGUI, using it to generate a predicted data set, running a genetic algorithm search and reaching a resonance assignment, initially on a small example protein and then on a larger one. The small protein is from a photosynthetic bacteria’s electron transport chain and has both NMR and X-ray structures (Northeast Structural Genomics Consortium target SgR42, PDB entries 3C4S (X-ray) and 2JZ2(NMR)). It has complete sets of NOE, RDC, and chemical shift data, along with assignments (BMRB entry 15604). We will use only a subset of 1H, 15N data (PHE and VAL data) as an illustration of an application to a sparsely labeled protein. The necessary data are contained in an “…experiment\_input” workbook in the “examples” folder supplied on installation.

There is a second example that is a little more challenging, the N-terminal domain from CEACAM1, a cell-surface signaling and adhesion molecule. It has been labeled with 13C in valine and alanine methyl groups. In addition to chemical shift and NOE data it makes use of pseudo-contact shift (PCS) and paramagnetic relaxation enhancement (PRE) data. The tutorial also includes descriptive information on more advanced applications including “Validation Mode” which is important in assigning confidence levels to assignments and “Trajectory Mode” which takes account of conformational averaging of experimental data.

1. **Launching the GUI versions**

The procedure for launching the GUI is slightly different for source and complied versions of AssignSLPGUI. Source versions require opening MATLAB prior to executing the AssignSLPlauncer.mlapp. Compiled versions open directly from a system dependent command. See the readme files supplied in the installation package for explicit instructions. In either case the following launch screen with the AssignSLPGUI logo will eventually appear. Follow the steps provided below.

**

1. **Preparing the Prediction input file from an X-Ray crystal structure (PDB 3C4S)**
2. Click the “Prepare Predictions: Single Frame” button in the launcher window.
3. The prediction file preparation tool will appear.

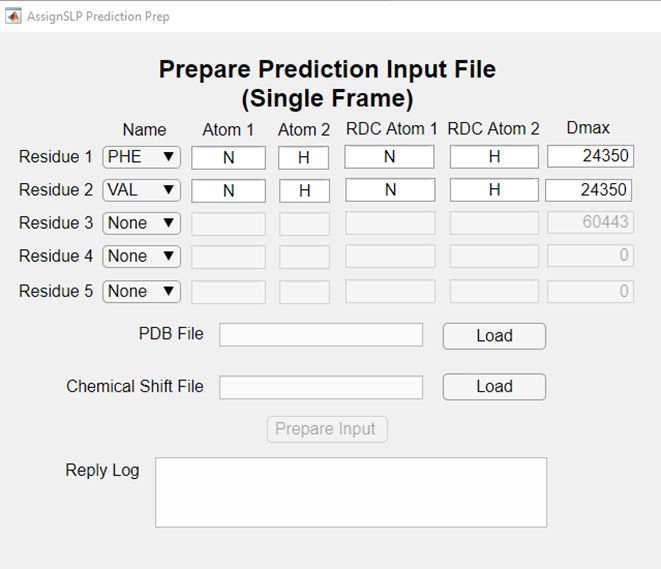
Graphical user interface

Description automatically generated

1. The left side of this window allows the user to define up to 5 spin-pair types for prediction based on amino acid type and atom names. The default values are more appropriate for the CEACAM example; lets begin with the 3C4S example.
2. Look at the HSQC sheet of the “3C4S\_experiment\_input.xlsx” workbook to see the type of residue labeling that exists. You will see that we have chemical shift data for PHE and VAL 15N-1H spin pairs for which we will want to predict crosspeak assignments. Looking at the other sheets you will see that, in addition to the 15N-1H chemical shift data, NOESY and RDC data are available. Other sheets enforce constraints coming from other data. For example, mutational data may assign a specific crosspeak to a specific site, or separate preparations using different labeled amino acids may allow crosspeak separation by amino acid type (constraint sheet). TOCSY data, or data from a preparation using a uniformly labeled amino acid, may assign a set of 13C-1H crosspeaks to the same residue (pairing sheet). Let’s begin by identifying residue type 1 on the prediction window. The order of entry should be the same as in the experiment workbook
   1. Change the dropdown menu to PHE.
   2. In the atom 1 field, type “N”
   3. In the atom 2 field, type “H”
   4. In the “RDC Atom 1” field, type “N”
   5. In the “RDC Atom 2” field, type “H”
   6. In the “Dmax” field, type 24350

*Note: for this example, all spin pairs are of the same type and the value used for Dmax is not important.*

1. Repeat this process for residue type 2
   1. Change dropdown to VAL
   2. In the atom 1 field, type “N”
   3. In the atom 2 field, type “H”
   4. In the “RDC Atom 1” field, type “N”
   5. In the “RDC Atom 2” field, type “H”
   6. In the “Dmax” field, type 24350
2. Make sure the remaining residue lines are inactive by setting the dropdown menus to “None”.
3. The residue fields should now match the screenshot below.



1. The right side of the prepare predictions screen (item 2 above) has an “Options” tab, with several adjustable parameters. For this example, only the “The NOE Distance Cutoff” will be used. The default value is 4 Angstroms. We’ll leave it here for this example.
2. The “Paramagnetic Species” field is empty by default. This example doesn’t use paramagnetic data and we’ll leave this blank. Since the species field is blank, the value of “PRE minimum distance, r0” is ignored. The “Magnetic field strength, B0” and “Temperature” fields will also be ignored.
3. Load a PDB file by clicking the “Load” button next to the PDB file field. A file browser window will appear. Navigate to the Example Files directory and choose 3c4s\_addH.pdb.
   1. NOTE – this file was prepared by fetching 3C4S.pdb from the PDB and adding hydrogen atoms using Chimera.

Graphical user interface

Description automatically generated

Click here to load PDB

File name appears after loading

1. Select a chemical shift prediction file by clicking the second “Load” button. In the file browser window navigate to the Example Files directory and select “3c4s.cs”.
   1. NOTE – This is an output file from the ShiftX2 webserver, saved in NMR-STAR format. Any number of shift predictors can be used as long as the output is in NMR-STAR format. Note that predicted shifts are used for NOE contacts as well, so it is important that the prediction contains all protons.

Graphical user interface

Description automatically generated

Click here to generate the predictions

Click here to load chemical shift file

File name appears after loading

1. Now that both PDB and Chemical Shift files have been loaded, the “Prepare Input” button will become enabled. Click it to prepare predicted HSQC and NOE peak lists and coordinates for RDC back-calculation.
2. After a few seconds, the right side of the interface will update with tabs to select tables showing the predictions. The HSQC peak list is visible by default. The RDC and NOE information can be seen by clicking the appropriate tab above the table. Note that the “PREs” and “PCS” tabs will remain empty.

DataTypes

Table

Description automatically generated

1. Save an excel workbook containing the predicted data by clicking the “Save Excel Workbook” button. This will open a file browser dialog where you can choose a location and name the output file.

Close the prediction prep tool window.

1. **Running the genetic algorithm search**
2. Return to the launcher window and click the “AssignSLP: Search” button. The search GUI will appear in a new window.

Graphical user interface, application

Description automatically generated

1. Load experimental data by clicking the “Load” button to the right of “Experimental Data File”. A file browser window will appear. Navigate to the “examples” directory and select the Excel workbook named “3C4S\_experiment\_input.xlsx”.

Graphical user interface, text, application, email

Description automatically generated

Click Here!

1. The Experimental Data File textbox will update with the filename and the Reply Log will display relevant information, indicating the data was loaded successfully.
   1. *WARNING: After selecting the file, the GUI window may appear to vanish. The window has just moved behind other windows that are open.*
   2. See the document on “Preparing the Experimental Workbook” for more information about the structure of the experimental input.

Graphical user interface, text, application, email

Description automatically generated

c

c

1. Load predicted data by clicking the “Load” button to the right of “Predicted Data File”. Another file browser window will appear. Select the Excel workbook prepared in the previous section.
   1. NOTE – The GUI window may appear to vanish again, but it has simply moved behind other windows.
2. The “Predicted Data File” textbox will update with the filename of the Excel workbook. The Reply Log will update again.

Graphical user interface, text, application, email

Description automatically generated

c

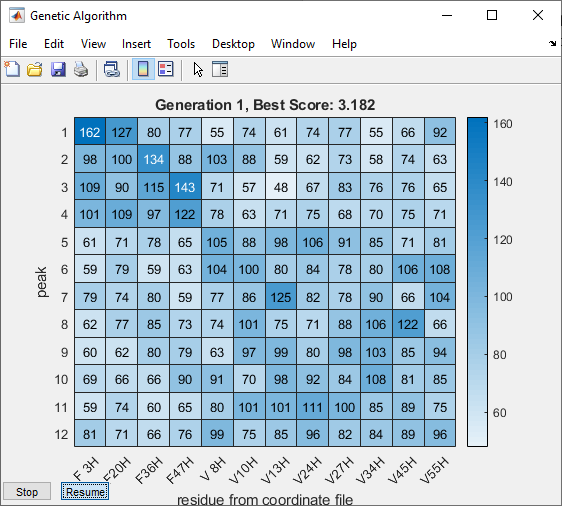
c

c

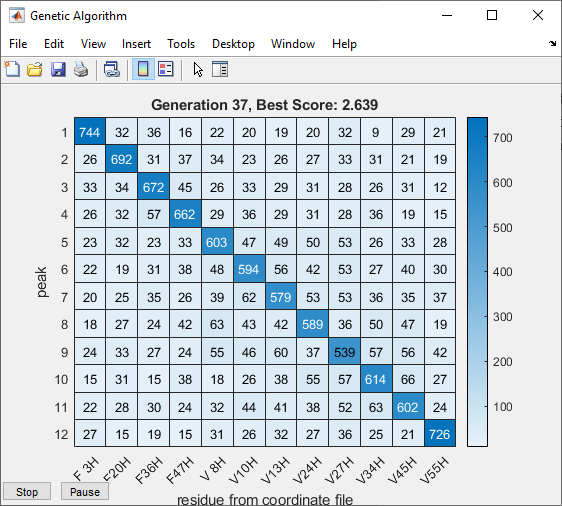
This button is now enabled

*Note: predicted and experiment workbooks can be loaded in either order.*

1. Now that both experimental and predicted data files have been loaded, the “Go!” button should become active. Clicking this button will start the genetic algorithm search. By default, the genetic algorithm (GA) is running with 1000 individuals that pair HSQC crosspeaks with specific labeled sites in our protein. At each generation they undergo mutations or crossovers (actually segment inversions) to generate new individuals with different pairings. These individuals are scored based on a match of predicted and experimental data; a fraction of individuals having the lowest scores is saved and the remainder undergo another round of mutation and crossover. This continues until no significant change in score occurs or the process reaches a set maximum number of generations (parameters set in the options menu; see section 5.4 below). Genetic algorithms are not infallible. So, you can set the # of replicate GA searches. By default the search is repeated 5 times and the best result across all 5 searches is saved. In our example the same assignment will be found all 5 times, so this can be set to 1, but in other cases, looking at the scores of repetitions can be informative. Click “Go!” now.
2. Select a file location to save the output folder. This folder contains temporary MATLAB files useful for debugging as well as subsequent analysis.
3. After a few seconds, another window will appear displaying a heatmap. This heatmap reflects the assignments made by the current population of individuals. Each box in the heatmap shows the number of times individuals assign a particular crosspeak to a particular residue. For example, crosspeak 1 is assigned to residue F3 by 162 of the 1000 individuals in the figure below. This will update periodically as the search algorithm runs.



1. After about 30 generations the algorithm will converge and stop running. In this example, the correct assignment is already known and experiment datasets were arranged in the correct order. The converged assignment appears along the diagonal.



1. If the number of repetitions is greater than 1, a second search will begin and the heatmap will revert to a random distribution. This process will repeat until the specified number of replicate searches is completed. It is important to ascertain that the same minimum solution is found multiple times, as the search can find local as opposed to global minimum assignments.
2. The main GUI window will now update with the lowest scoring result across all replicate GA searches. The reply log will display the total run time. On the right panel of the search window, the assignment table will populate with the most optimal assignment from the search.
3. Having completed a search, now click the “Analyze Results” button on the search window. A new window will appear:

Graphical user interface, chart, line chart

Description automatically generated

1. This window gives more detailed information on the scores from the final generation of the genetic algorithm search. The graph on the left shows “Plot Total” by default. These are total scores for individuals ranked lowest to highest. Other buttons allow you to look at the contribution from different data types. The table below presents the same data numerically. The right half of the screen shows additional information for RDCs and/or PCSs from a single assignment. Selecting a row (another individual) in the scores table (bottom left) will cause the RDC/PCS information to update for that assignment. At this point many individuals carry the same assignments, so you may have to move down in the list to see some changes.
2. Clicking “OK” will close this window.
3. It may be useful to explore the effect of changing some variables using this small protein example. For example, uncheck some data type (RDCs?) on the search window. How many assignments are correct using just shifts and NOEs? Or, change the weighting of a data type (for example, decrease RDC error estimates in the experiment workbook). What effect does this have?
4. **PCS and PRE data:**

Paramagnetic Relaxation Enhancements (PREs) and Pseudo Contact Shifts (PCSs) provide useful long-distance constraints that nicely complement NOEs, chemical shifts and RDCs. The example file we will use is from the cell-surface adhesion/signaling molecule, CEACAM1. This protein is best known for its use as a marker for certain types of cancer and as a target for therapeutic intervention in these diseases. It is highly glycosylated and best expressed in mammalian cells, where sparse labeling proves especially advantageous. The example illustrates the use of 13C-methyl labels (valine and alanine in this case), as well as the use of PCSs and PREs. Because of the more extensive data set and increased degeneracy of data on a larger protein, adjusting search parameters and determining confidence levels for assignments becomes more important.

The geometric dependence of PCSs and RDCs are the same. So, the algorithms used in the search routine are identical. It is just that the atom pairs for PCS calculations are the protons of HSQC atoms (Atom 2 in the Prepare Predictions window) and the ion at the center of a large anisotropic magnetic susceptibility tensor, as opposed to the pairs specified as RDC atoms in the Prepare Predictions window. The coordinates for the PCS atom pairs are put out in a separate PCS sheet; coordinates for the pair of RDC atoms remain in an RDC sheet. PCS predictions are triggered on finding the symbol for the paramagnetic ion in the “Paramagnetic Species” window (Tb in our example). In our example, no RDC data are included, so the RDC sheet created will be ignored.

The equivalent of a Dmax for RDCs (PCSmax) is calculated internally using the B0 and Temp values entered in the “Paramagnetic Species” window. With values of 21T and 298K PCSmax is 5.9e+07. However, the size is inconsequential in this example, since it is a product of order parameters and Dmax (PCSmax) that determines predicted values, and calculated order parameters will compensate for any variation in choice of PCSmax. There is, however, an important case where the PCSmax value does matter, namely the case where field induced RDCs and PCS are collected on the same paramagnetic sample. Combining the data is advantageous, as five unknowns must be determined in any order tensor calculation, and relatively large data sets are needed to make this determination reliable. The easiest way to combine the data is to move PCS data to the RDC sheets and enter the PCSmax values where the RDCmax values would normally occur. Dmax values must correspond to the RDC atoms entered. For a directly bonded 13C-1H pair, a value of minus 60,200 is appropriate; for a methyl 13C-1H pair, RDCs are affected by methyl rotation. It is best to use a C-C-Methyl pair (CA-CB for alanine) and adjust Dmax for the difference in bond length as well as rotation. We find Dmax of +39,000 to be appropriate.

Calculation of predicted PRE values is also triggered by the appearance of a symbol for a paramagnetic ion in the “Paramagnetic Species” window. PREs are calculated assuming exponential decay with acquisition time in proportion to 2.3(R0/R)6, where R0 is an estimate of the distance at which 10% of the initial peak intensity remains and R is the distance between the proton of an isotopically labeled site and the paramagnetic ion. An experimental workbook, pdb files and predicted chemical shift are provided for CEACAM. Many of the subsequent steps in running ASSIGN\_SLP are similar to the previous example and they will be only briefly described:

1. From the launch window select “Prepare Predictions”. The Predictions window may default to an appropriate set of parameters. If not, fill in the residue types as ALA and VAL, with two lines for VAL, as there are two methyl groups; fill in the atom types as CB, HB for alanine and CG1, HG1 and CG2, HG2 for valine. You can leave the RDC atoms at their defaults, or change them to another pair of interest, but predictions will only be used if experimental RDCs are provided. Load the pdb and chemical shift files from the examples directory. Since this example uses paramagnetic data enter the species type as TB (refer to the pdb file for verification). For enter an R0 value of 15Å. For PCSs enter a field strength (B0) and a temperature. Values are only important if PCSs are to be compared to field induced RDCs, but the program will look for them anyway. Press “Prepare Predictions” and save the workbook.
2. From the launch window select “AssignSLP Search”. On the search window load experiment and prediction workbooks as before. But, before executing the search it will be useful to adjust some search parameters.
3. Along the top of the window there is an “options” menu, with a selection for “Genetic Algorithm”. Click this button.
4. The “AssignSLPGUI: Options” window will appear.

Graphical user interface

Description automatically generated

The “Scoring Function Settings” are used only to weight the contribution of different types of data to the total score; most score contributions are divided by the errors entered to weight their contribution. NOEs are treated differently, since it is difficult to estimate errors in data represented by a multiple peak vector. Instead, scores are multiplied by a ratio of the average NOE intensity over all vectors divided by the intensity of the vector with the least data (assumed to be mostly noise). Adjusting the NOE factor upward weights NOEs more heavily. However, the default value, calculated from a typical input, is a good starting point for both 3C4S and CEACAM1.

1. “GA Search Settings” affect how the genetic algorithm runs and are more important for CEACAM. The meanings of Max Generations and Population size are obvious. Generation Limit is essentially a number of iterations for which a minimal change in score has occurred. The default numbers work well for the 3C4S example, but for CEACAM1 it is best to adjust Max Generations to 300 and Generation Limit to 50. Once adjusted, press “OK”; you will return to the Search window.
2. In the CEACAM example it is OK to leave the # of replicate GA searches at 5. Press “Go!” and watch the score display at the top of the heatmap window. Note that the list of crosspeaks is no longer in the order of assignments, so solutions converge to a somewhat random set of boxes. Make a note of the score at the end of each repetition. If the same low value (and set of most populated boxes) is not found more than once, it may be wise to raise the number of repetitions.
3. Once completed, you can activate the “Analysis” routine as you did for the 3c4s example.
4. **Validation Mode (on search window)**

Under typical use, a multiple repetition GA search will converge upon a single assignment, but this is no guarantee that the assignment is correct. For this reason, we developed a feature called “Validation Mode” which allows the user to assess the confidence of the assignment for each HSQC peak.

1. Clicking the “Validation Mode” button will start a Monte Carlo style process that repeats the GA search many times (200 by default). This procedure can take many hours to run on a single processor, so it is best executed on a multiple processor computer.
2. Separate entries are made in “Errors for Monte Carlo” on the options page; these are used in the Validation mode to generate new datasets. It is important to keep these latter error estimates realistic (about half the default search parameters), and possibly somewhat below actual estimated errors. They determine the extent to which new experimental data sets generated by Gaussian sampling deviate from the experimental values. Ideally, they should deviate from true values; deviation from experimental values will cause some excessive deviation, hence the use of smaller error estimates.
3. The “# of Rounds” under “Validation Mode” is the number of times a new data set is generated and used in a genetic algorithm search. A setting of 200 will make a fractional assignment among individuals of 0.95 for a particular crosspeak correlate with a 95% confidence level. More rounds are required if you wish to trust fractional assignments lower than this (~250 for 80% and above).
4. “Parallel Processes” determines the number of CPUs used in validation mode. The time to run in validation mode can be cut significantly by using multiple processors. Most PCs and laptops will have at least 2 processors. Laboratory workstations may have 24; university systems can have hundreds. On a Linux system the command, “nproc”, will show you the number available

**7. Prepare Predictions: Trajectory Mode (Launch Window)**

**Table

Description automatically generated**

1. This mode is also time consuming and may be appropriate only for more advanced studies of systems exhibiting significant internal motion. The program takes in an MD trajectory and aligns the frames so that overall and internal motion can be treated separately. To reduce computation times, while retaining acceptable accuracy trajectories should be acquired (or reduced to) one frame each 4 ps. An effective isotropic correlation time for molecular tumbling is entered and the trajectory is broken into segments of 3 times this value to allow treatment of internal motion. NOEs are treated at the spin-pair level, but more accurately than the 1/r6 approximation used for single frame analysis. The correlation time dependence of NOEs treated properly and a “complete” relaxation matrix is used to account for indirect NOE effects. The NOE treatment used is described more completely in a paper by Chalmers et al. in the Journal of Magnetic Resonance 265 (2016) 1–9.
2. RDCs are also corrected by using the trajectory to calculate an order parameter and effective bond vector. The effective bond vectors replace the bond vectors used in a single frame calculation and the experimental RDCs are divided at each trial assignment by the order parameter to mimic what would be seen if no internal motions existed**.** Oder parameters entered in the prediction workbook should be >0.1 to prevent excessively large RDC predictions**.** At this point no averaging is done for PCSs and PREs; the first frame of the trajectory is used as a representative rigid structure.
3. Chemical shifts are predicted externally to this program and lists are uploaded. So, the user is responsible for averaging shifts. There are software packages that take in multiple model pdb files and return an ensemble averaged chemical shift list (PPM, for example; <http://spin.ccic.osu.edu/index.php/ppm>). Note that most software packages are trained on solution data and inherently include some averaging, particularly for residues that frequently occur in less ordered regions. Caution: Check that shifts for all spins are present, including those involved in NOEs.

**8. Accelerated MD**

1. Unfortunately, with computational facilities currently available to most users, it is not practical to acquire cMD trajectories longer than a few microseconds. This is far too short to properly sample all conformations accessed by proteins. Accelerated Molecular Dynamics (aMD) trajectories provides one solution to this problem (Pierce, et al. Journal of Chemical Theory and Computation, 2012, 8: 2997-3002). aMD trajectories can be used in ASSIGN\_SLP, but not directly, as they include high energy structures due to the boost energies used to facilitate crossing barriers between conformational states. To allow use of these trajectories, frames can be clustered, energies reweighted and representative structures from the most probable structures selected (Rogals et al., ACS Chemical Biology, 2022, 17:3527-3534. These can be used as starting points for short conventional MD runs (~100ns), which can be strung together in proportion to the probabilities assigned to starting structures.

**9. Acknowledgements**

This work was supported by NIH grant R01 GM134335. Chemical shift predictions were calculated using the ShiftX2 webserver (www.shiftx2.ca): Beomsoo Han, Yifeng Liu, Simon Ginzinger, and David Wishart (2011) 50:43-57. Assign\_SLP uses functions included in the MDtoolbox package for MATLAB (https://github.com/ymatsunaga/mdtoolbox).Y. Matsunaga, and Y. Sugita, J. Chem. Phys. (2018) 148:241731. We thank Professor Dan Hall, Professor Jaxk Reeves and Huimin Hu, of the University of Georgia Statistical Consulting Center, for contributions underlying the code used in the Validation module. We thank Monique Rogals and Alex Eletsky for acquiring the CEACAM1 data, and we thank Varshith Paduchuri for resolving a number of code issues and implementing the advanced trajectory mode described above.