









# SPyCi-PDB: A modular command-line interface for back-calculating experimental datatypes of protein structures.

Zi Hao Liu <sup>1,2</sup>✉, Oufan Zhang<sup>3,4</sup>, João M. C. Teixeira <sup>5</sup>, Jie Li <sup>3,4</sup>,  
Teresa Head-Gordon <sup>3,4,6,7</sup>, and Julie D. Forman-Kay <sup>1,2</sup>✉

**1** Molecular Medicine Program, Hospital for Sick Children, Toronto, Ontario, Canada **2** Department of Biochemistry, University of Toronto, Toronto, Ontario, Canada **3** Pitzer Center for Theoretical Chemistry, University of California, Berkeley, California, United States of America **4** Department of Chemistry, University of California, Berkeley, California, United States of America **5** Department of Biomedical Sciences, University of Padua, Padova, Italy **6** Department of Chemical and Biomolecular Engineering, University of California, Berkeley, California, United States of America **7** Department of Bioengineering, University of California, Berkeley, California, United States of America ✉ Corresponding author

DOI: [10.21105/joss.04861](https://doi.org/10.21105/joss.04861)

## Software

- [Review](#) 
- [Repository](#) 
- [Archive](#) 

Editor: [Antonia Mey](#)  

## Reviewers:

- [@dotsdl](#)
- [@lohedges](#)
- [@JenkeScheen](#)
- [@sulstice](#)

Submitted: 21 September 2022

Published: 10 May 2023

## License

Authors of papers retain copyright and release the work under a Creative Commons Attribution 4.0 International License ([CC BY 4.0](https://creativecommons.org/licenses/by/4.0/)).

## Summary

Structural determination of proteins has been a central scientific focus since the early 1960s (Dill et al., 2008) with technological advances facilitating experimental structures of stable, folded proteins by nuclear magnetic resonance (NMR) spectroscopy (Kanelis et al., 2001), X-ray crystallography (Smyth, 2000), and cryo-electron microscopy (Malhotra et al., 2019), as well as the recent computational prediction of structures (Baek et al., 2021; Jumper et al., 2021). Modeling intrinsically disordered proteins (IDPs) and intrinsically disordered regions (IDRs), however, remains challenging due to their highly dynamic nature and low propensity to form low energy folded structures (Mittag & Forman-Kay, 2007).

Currently, approaches to model IDPs/IDRs generally start with initial pools of structures that sample potentially accessible conformations and then utilize experimental data to narrow the pool. One method to generate initial conformational ensembles of IDPs/IDRs uses sampling techniques such as in TraDES (Feldman & Hogue, 2000, 2001), Flexible-meccano (Ozenne et al., 2012), FastFloppyTail (Ferrie & Petersson, 2020), IDPConformerGenerator (Teixeira et al., 2022), and others (Estañá et al., 2019), that rely on the torsion angle distributions found in high-resolution folded protein structures deposited in the RCSB Protein Data Bank (Berman, 2000). Another more computationally expensive approach generates conformational ensembles using molecular dynamics (MD) simulations with different force-fields (Robustelli et al., 2018; Salvi et al., 2016).

After generating the initial pool of structures, back-calculations to experimental data and reweighting using Monte-Carlo (Krzeminski et al., 2012) or Bayesian statistics (Bottaro et al., 2020; Brookes & Head-Gordon, 2016; Lincoff et al., 2020) can be performed to define structural ensembles that better match solution NMR, small-angle X-ray scattering (SAXS), single molecule fluorescence (SMF), and other experimentally obtained data from these IDPs/IDRs. An emerging method to generate conformations of IDPs/IDRs uses machine learning generative models based on ensembles generated from sampling or MD techniques as training data and reinforces learning with experimental data (Zhang et al., 2022). Both of these general approaches rely on back-calculation of “experimental observables” from coordinates of conformers within the ensembles, a task that is increasingly complex due to the various models for interpretation of experimental data and the numerous tools available.

Here we present **SPyCi-PDB**, designed to facilitate and streamline this back-calculation stage

by acting as a platform for internal back-calculator functions as well as published third-party software, utilizing PDB structures of disordered protein conformations. One goal of **SPyCi-PDB** is to minimize the existing issues with different data-formats from software and scripts within the IDP/IDR research community and improve accessibility to researchers with a range of computational expertise. In this release, **SPyCi-PDB** can back-calculate NMR chemical shift (CS), paramagnetic resonance enhancement (PRE), nuclear Overhauser effect (NOE), 3J-HNHA coupling (JC), and residual dipolar coupling (RDC) data; hydrodynamic radius (Rh) data from NMR, light scattering, or size exclusion chromatography; SAXS; and single-molecule fluorescence resonance energy transfer (smFRET) values from all-atom PDB structures of IDP/IDR conformations.

## Statement of Need

As new software packages and *in silico* methodologies emerge to better model IDP/IDR structures, back-calculations to multiple experimental datatypes are required to quantitatively assess the conformers generated. However, interpretation of solution data, as a simple calculation from the sum of sampled conformations within IDP/IDR ensembles is fraught with pitfalls. For example, commonly used approaches for back-calculating NOE and PRE data for dynamic protein systems treat only the distance and do not incorporate the contribution of dynamics of the vector connecting the interacting points, potentially leading to underestimations of the potential range of distances sampled (Brookes & Head-Gordon, 2016; Krzeminski et al., 2012; Lincoff et al., 2020). In addition, even for stable systems, back-calculation is not trivial, with even state-of-the-art back-calculators of chemical shifts, such as in UCBSHift (Li et al., 2020), leading to errors that can be large relative to the expected deviation of experimental values. Given the rapidly developing nature of different software tools to perform back-calculations, **SPyCi-PDB** should assist by providing a user-friendly, all-in-one package to reduce time and confusion in this back-calculation step as well as open opportunities for future collaborations and integration of new experimental datatypes. Furthermore, **SPyCi-PDB** aims to unify different input and output data formats from different experimental datatypes to increase productivity and accelerate research. As stated in the documentation hosted by ReadTheDocs, input formats are conventional comma-delimited tables (e.g. .CSV, .TXT), while the output format is human-readable .JSON files that can be easily manipulated using Python or other software based on the user's ultimate needs. **SPyCi-PDB** was also developed to integrate into the IDPConformerGenerator platform (Teixeira et al., 2022).

Ultimately, given the complicated and dynamic exchanging nature of IDPs, new back-calculators are needed to be developed to address the current challenges in interpretation. By creating a tool with modularity and best practices, we aim to encourage the researcher community to contribute towards this platform to further the goal of improved modelling of IDPs and IDRs.

## Implementation

As `spycipdb` is written completely in Python, it is compatible with any platform able to execute Python ( $\geq 3.8$ ,  $< 4.0$ ). However, certain third-party extensions to perform back-calculations (SAXS and RDC) have only been tested on 64-bit Ubuntu 18.04.X LTS and 20.04.X LTS, as well as WSL 2.0 on 64-bit Windows 11.

In the production version 0.3.5, four out of eight modules of **SPyCi-PDB**'s back-calculators (`pre`, `noe`, `jc`, `smfret`) use internal mathematical equations and PDB structure processing algorithms from IDPConformerGenerator libraries (Teixeira et al., 2022). The `pre` (1) and `noe` (2) module calculates scalar distances between pairs of atoms according to the pairs derived from the experimental template. It utilizes an algorithm that matches atom names of each residue with allowance for multiple assignments for `noe`. The `jc` (3) module uses the Karplus curve, a simple cosine function, to back-calculate the desired J-couplings according to residue

number as provided by the experimental template file (Pérez et al., 2001). Finally, the smfret (4) module takes into consideration residue pairs and a scale factor to adjust for dye size from the experimental setup to back-calculate distances between two alpha-Carbon (CA) atoms (Lincoff et al., 2020). The aforementioned equations are as follows:

$$\sqrt{\delta x^2 + \delta y^2 + \delta z^2} \quad (1)$$

$$\sqrt[6]{\left(\frac{\delta x^2 + \delta y^2 + \delta z^2}{N}\right)^3} \quad (2)$$

$$\cos\left(\varphi - \frac{\pi}{3}\right) \quad (3)$$

$$\frac{1}{1 + \left(\frac{D \cdot \sqrt{\frac{|R_1 - R_2| + r}{R_1 - R_2}}}{S}\right)^6} \quad (4)$$

Where  $\delta x$ ,  $\delta y$ ,  $\delta z$  are the Cartesian differences between two atoms of interest (1, 2),  $N$  represents the number of combinations for NOE atom pairs (2),  $\varphi$  is the Phi torsion angle of interest (3),  $D$  is the scalar distance between the residues of interest with  $R_1$  and  $R_2$  being the vector Cartesian coordinates for the residues and  $S$  being the scale factor according to experimental information.

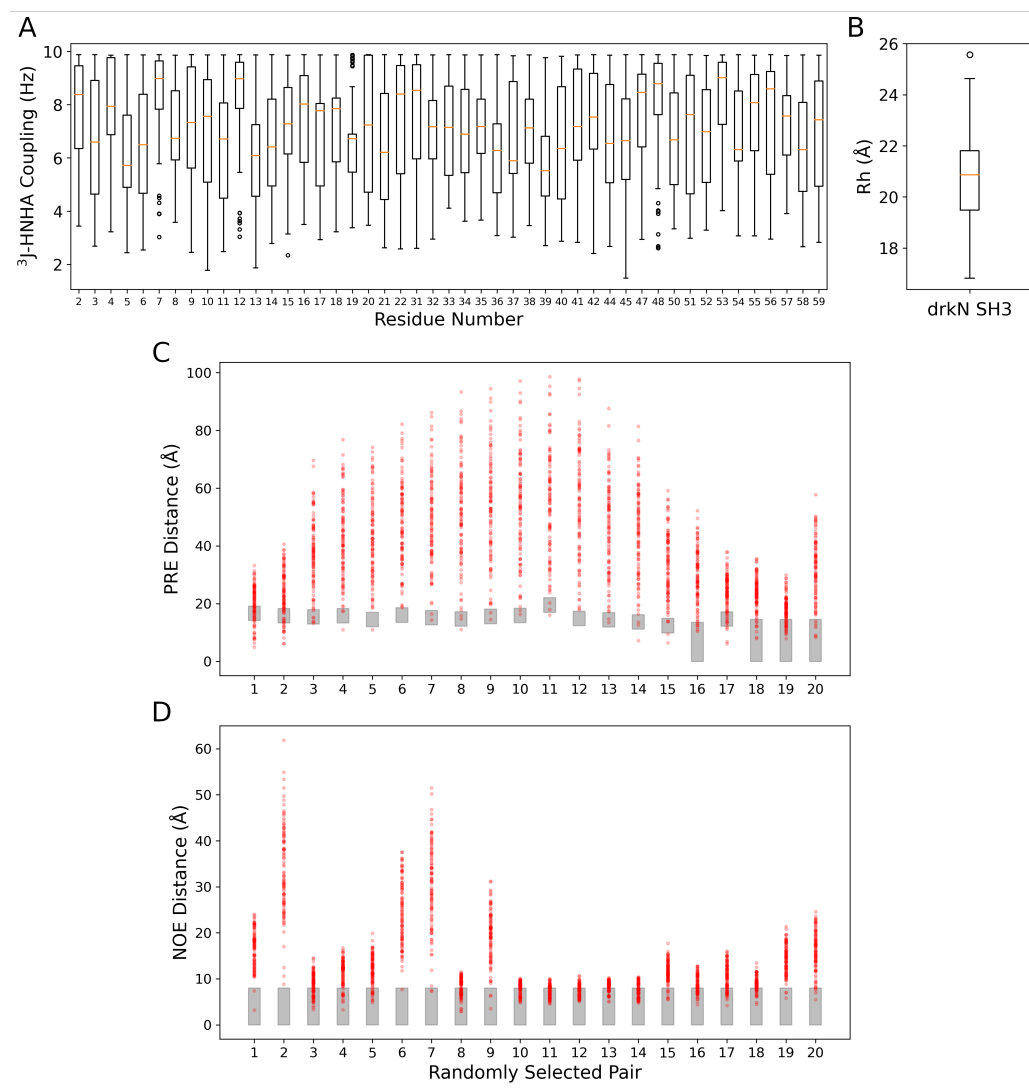
The remaining 4 modules (cs, saxs, rh, rdc) call upon third-party academic software: UCBSHift, a machine learning algorithm that uses structural alignment for experimental chemical shift replication and employs a random forest regression on curated data to most accurately predict protein chemical shifts (Li et al., 2020); CRY SOL v3, an updated version of the well-established SAXS back-calculator from ATSAS that can now evaluate the hydration shell by populating the protein structure with dummy water (Franke et al., 2017); HullRad, to calculate hydrodynamic radius (Rh) by using a convex hull model to estimate the hydrodynamic properties of a macromolecule (Fleming & Fleming, 2018); and PALES, using the steric obstruction model to derive dipolar coupling (RDC) information from the average orientation of the 3D coordinates (Zweckstetter & Bax, 2000). Thorough testing of each module has been performed to ensure smooth installation and troubleshooting, as well as retaining or providing multiprocessing capabilities that may not have been implemented in their standalone forms. When choosing third-party software, we prioritized those written in Python for ease of integration.

We plan to integrate alternative methods to calculate experimental datatypes internally, such as using a parameterizable fluorescence lifetime and the Förster distance, as used in the Naudi-Fabra et al. study of describing intrinsically disordered proteins using smFRET, NMR, and SAXS (Naudi-Fabra et al., 2021). Future additions to the SPyCi-PDB interface suite are welcome and easy to perform given its modular design.

Detailed installation/troubleshooting instructions, real-world usage examples, and input/output formats are provided both in the project's documentation hosted on ReadTheDocs (<https://spyci-pdb.readthedocs.io/en/stable/>) and within the modules through the `--help` argument. Plots of sample outputs from the jc, rh, pre, and noe modules using the example structures and data in the repository are shown in Figure 1.

Comparing the back-calculated PRE and NOE distance values to the experimental observables, the default Euclidean distance interpretation yields some values agreeing with the experimental range (Figure 1C, 1D). With a greater sample size, we would likely capture more back-calculated data agreeing with the experimental ranges. The internal plotting features in the noe, pre, rh, and jc modules of SPyCi-PDB is useful for users to gauge the quality of the initial pool before downstream reweighting. Furthermore, with the integration of different back-calculation

methods, these plots will provide the user with a useful comparison between back-calculation philosophies.



**Figure 1:** Plots of distributions of back-calculated experimental data of 100 structures of the unfolded state of the Drk N-terminal SH3 domain (drkN SH3) generated using IDPConformerGenerator (Teixeira et al., 2022). Panel (A) shows back-calculated  $^3\text{J}$ -HNHA couplings in Hz based on the Karplus equation with A, B, and C constants from Lincoff et al. (Lincoff et al., 2020). Only residues with experimental data to compare will generate a back-calculated J coupling value. Panel (B) shows the distribution of back-calculated Rh values in Angstroms using HullRad (Fleming & Fleming, 2018). Panels (C) and (D) show twenty randomly selected pairs of back-calculated PRE and NOE distances, respectively. The ranges of experimental values are represented as grey boxes while back-calculated values for each conformer are shown as red dots.

## Acknowledgements

T.H.-G. and J.D.F.-K. acknowledge funding from the National Institute of Health under Grant 2R01GM127627-05. J.D.F.-K. also acknowledges support from the Natural Sciences and Engineering Research Council of Canada (2016-06718) and from the Canada Research Chairs Program.

The motivation behind this project was to create a modular-yet-standalone software package to back-calculate experimental datatypes for conformers generated by the IDPConformerGenerator (Teixeira et al., 2022) platform.

## References

- Baek, M., DiMaio, F., Anishchenko, I., Dauparas, J., Ovchinnikov, S., Lee, G. R., Wang, J., Cong, Q., Kinch, L. N., Schaeffer, R. D., Millán, C., Park, H., Adams, C., Glassman, C. R., DeGiovanni, A., Pereira, J. H., Rodrigues, A. V., Dijk, A. A. van, Ebrecht, A. C., ... Baker, D. (2021). Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, 373(6557), 871–876. <https://doi.org/10.1126/science.abj8754>
- Berman, H. M. (2000). The protein data bank. *Nucleic Acids Research*, 28(1), 235–242. <https://doi.org/10.1093/nar/28.1.235>
- Bottaro, S., Bengtsen, T., & Lindorff-Larsen, K. (2020). Integrating molecular simulation and experimental data: A bayesian/maximum entropy reweighting approach. In *Methods in molecular biology* (pp. 219–240). Springer US. [https://doi.org/10.1007/978-1-0716-0270-6\\_15](https://doi.org/10.1007/978-1-0716-0270-6_15)
- Brookes, D. H., & Head-Gordon, T. (2016). Experimental inferential structure determination of ensembles for intrinsically disordered proteins. *Journal of the American Chemical Society*, 138(13), 4530–4538. <https://doi.org/10.1021/jacs.6b00351>
- Dill, K. A., Ozkan, S. B., Shell, M. S., & Weikl, T. R. (2008). The protein folding problem. *Annual Review of Biophysics*, 37(1), 289–316. <https://doi.org/10.1146/annurev.biophys.37.092707.153558>
- Estaña, A., Sibille, N., Delaforge, E., Vaisset, M., Cortés, J., & Bernadó, P. (2019). Realistic ensemble models of intrinsically disordered proteins using a structure-encoding coil database. *Structure*, 27(2), 381–391.e2. <https://doi.org/10.1016/j.str.2018.10.016>
- Feldman, H. J., & Hogue, C. W. V. (2000). A fast method to sample real protein conformational space. *Proteins: Structure, Function, and Genetics*, 39(2), 112–131. [https://doi.org/10.1002/\(sici\)1097-0134\(20000501\)39:2%3C112::aid-prot2%3E3.0.co;2-b](https://doi.org/10.1002/(sici)1097-0134(20000501)39:2%3C112::aid-prot2%3E3.0.co;2-b)
- Feldman, H. J., & Hogue, C. W. V. (2001). Probabilistic sampling of protein conformations: New hope for brute force? *Proteins: Structure, Function, and Genetics*, 46(1), 8–23. <https://doi.org/10.1002/prot.1163>
- Ferrie, J. J., & Petersson, E. J. (2020). A unified de novo approach for predicting the structures of ordered and disordered proteins. *The Journal of Physical Chemistry B*, 124(27), 5538–5548. <https://doi.org/10.1021/acs.jpcc.0c02924>
- Fleming, P. J., & Fleming, K. G. (2018). HullRad: Fast calculations of folded and disordered protein and nucleic acid hydrodynamic properties. *Biophysical Journal*, 114(4), 856–869. <https://doi.org/10.1016/j.bpj.2018.01.002>
- Franke, D., Petoukhov, M. V., Konarev, P. V., Panjkovich, A., Tuukkanen, A., Mertens, H. D. T., Kikhney, A. G., Hajizadeh, N. R., Franklin, J. M., Jeffries, C. M., & Svergun, D. I. (2017). ATSAS 2.8: A comprehensive data analysis suite for small-angle scattering from macromolecular solutions. *Journal of Applied Crystallography*, 50(4), 1212–1225. <https://doi.org/10.1107/s1600576717007786>
- Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Židek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., ... Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873), 583–589. <https://doi.org/10.1038/s41586-021-03819-2>

- Kanelis, V., Forman-Kay, J. D., & Kay, L. E. (2001). Multidimensional NMR methods for protein structure determination. *IUBMB Life (International Union of Biochemistry and Molecular Biology: Life)*, 52(6), 291–302. <https://doi.org/10.1080/152165401317291147>
- Krzeminski, M., Marsh, J. A., Neale, C., Choy, W.-Y., & Forman-Kay, J. D. (2012). Characterization of disordered proteins with ENSEMBLE. *Bioinformatics*, 29(3), 398–399. <https://doi.org/10.1093/bioinformatics/bts701>
- Li, J., Bennett, K. C., Liu, Y., Martin, M. V., & Head-Gordon, T. (2020). Accurate prediction of chemical shifts for aqueous protein structure on “real world” data. *Chemical Science*, 11(12), 3180–3191. <https://doi.org/10.1039/c9sc06561j>
- Lincoff, J., Haghightalari, M., Krzeminski, M., Teixeira, J. M. C., Gomes, G.-N. W., Gradinaru, C. C., Forman-Kay, J. D., & Head-Gordon, T. (2020). Extended experimental inferential structure determination method in determining the structural ensembles of disordered protein states. *Communications Chemistry*, 3(1). <https://doi.org/10.1038/s42004-020-0323-0>
- Malhotra, S., Träger, S., Peraro, M. D., & Topf, M. (2019). Modelling structures in cryo-EM maps. *Current Opinion in Structural Biology*, 58, 105–114. <https://doi.org/10.1016/j.sbi.2019.05.024>
- Mittag, T., & Forman-Kay, J. D. (2007). Atomic-level characterization of disordered protein ensembles. *Current Opinion in Structural Biology*, 17(1), 3–14. <https://doi.org/10.1016/j.sbi.2007.01.009>
- Naudi-Fabra, S., Tengo, M., Jensen, M. R., Blackledge, M., & Milles, S. (2021). Quantitative description of intrinsically disordered proteins using single-molecule FRET, NMR, and SAXS. *Journal of the American Chemical Society*, 143(48), 20109–20121. <https://doi.org/10.1021/jacs.1c06264>
- Ozenne, V., Bauer, F., Salmon, L., Huang, J.-r., Jensen, M. R., Segard, S., Bernado, P., Charavay, C., & Blackledge, M. (2012). Flexible-meccano: A tool for the generation of explicit ensemble descriptions of intrinsically disordered proteins and their associated experimental observables. *Bioinformatics*, 28(11), 1463–1470. <https://doi.org/10.1093/bioinformatics/bts172>
- Pérez, C., Löhr, F., Rüterjans, H., & Schmidt, J. M. (2001). Self-consistent karplus parameterization of 3J couplings depending on the polypeptide side-chain torsion  $\chi_1$ . *Journal of the American Chemical Society*, 123(29), 7081–7093. <https://doi.org/10.1021/ja003724j>
- Robustelli, P., Piana, S., & Shaw, D. E. (2018). Developing a molecular dynamics force field for both folded and disordered protein states. *Proceedings of the National Academy of Sciences*, 115(21). <https://doi.org/10.1073/pnas.1800690115>
- Salvi, N., Abyzov, A., & Blackledge, M. (2016). Multi-timescale dynamics in intrinsically disordered proteins from NMR relaxation and molecular simulation. *The Journal of Physical Chemistry Letters*, 7(13), 2483–2489. <https://doi.org/10.1021/acs.jpcllett.6b00885>
- Smyth, M. S. (2000). X ray crystallography. *Molecular Pathology*, 53(1), 8–14. <https://doi.org/10.1136/mp.53.1.8>
- Teixeira, J. M. C., Liu, Z. H., Namini, A., Li, J., Vernon, R. M., Krzeminski, M., Shamandy, A. A., Zhang, O., Haghightalari, M., Yu, L., Head-Gordon, T., & Forman-Kay, J. D. (2022). IDPConformerGenerator: A flexible software suite for sampling the conformational space of disordered protein states. *The Journal of Physical Chemistry A*. <https://doi.org/10.1021/acs.jpca.2c03726>
- Zhang, O., Haghightalari, M., Li, J., Teixeira, J. M. C., Namini, A., Liu, Z.-H., Forman-Kay, J. D., & Head-Gordon, T. (2022). Learning to evolve structural ensembles of unfolded and disordered proteins using experimental solution data. arXiv. <https://doi.org/10.48550/ARXIV.2206.12667>



Zweckstetter, M., & Bax, A. (2000). 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(15), 3791–3792. <https://doi.org/10.1021/ja0000908>