Bayesian multi-state multi-condition modeling of a protein structure from X-ray crystallography

Matthew Hancock[a,c,1], **James S Fraser**[a], **Paul D Adams**[a], **and Andrej Sali**[b,1,2]

This manuscript was compiled on June 24, 2024

A model of a protein structure at atomic resolution is key to rationalizing and predicting its biological function. Many such models are computed from a diffraction pattern from X-ray crystallography. Despite the protein crystal containing billions of protein molecules that independently sample the energy landscape during data collection, most models computed from X-ray data depict a single set of atomic coordinates. A model with multiple sets of atomic coordinates (multi-state) may improve the satisfaction of the X-ray data and is a more accurate, precise, and informative depiction of the protein. However, computing a multi-state model is challenging on account of a low data-to-parameter ratio. X-ray datasets collected for the same system under distinct experimental conditions (*eg*, temperature) may provide additional observations, thereby improving the data-to-parameter ratio. Here, we develop, benchmark, and illustrate MultiXray: Bayesian multi-state multi-condition modeling for X-ray crystallography. The input information is J X-ray datasets collected under distinct conditions and a molecular mechanics force field. The model representation is N independent atomic models of the protein structure (states) and the weight of each state under each condition. A Bayesian posterior model density quantifies the match of the model with all X-ray datasets and molecular mechanics. A sample of models is drawn from the posterior model density using molecular dynamics simulations. We benchmark MultiXray on simulated X-ray data and analyze the impact of additional states and conditions on the scoring function and model search. We illustrate MultiXray on temperature-dependent X-ray datasets collected for SARS-CoV-2 M^{pro} and compute multi-state multi-condition models that improve the R^{free} relative to the PDB model by up to 0.05. MultiXray is implemented in our open-source *Integrative Modeling Platform*, relying on integration with *Phenix*, thus making it easily applicable to other systems.

Keyword 1 | Keyword 2 | Keyword 3 | ...

A. A protein crystal is a heterogeneous mix of structural states. X-ray crystallography is an important experimental technique for obtaining structural models of proteins at atomic resolution. In X-ray crystallography, the protein crystal contains between $10^6 - 10^{15}$ copies of the protein Rejto1996-up, Smith2015-bq, Woldeyes2014-bs. Due to the high solvent content of the crystal, protein molecules within the crystal can fluctuate nearly independently throughout data collection and adopt distinct structural states based on the energy landscape DePristo2004-vd, Jensen1997-up, Ringe1986-sa. Despite the resulting structural heterogeneity, a majority of models computed from X-ray datasets describe a single structural state with Gaussian isotropic B-factors Sun2019-yq. Fitting a B-factor to an X-ray dataset convolutes all sources of experimental uncertainty, including the heterogeneous mix of structural states, random thermal motion, and long-range crystallographic disorder Kuzmanic2014-dm, Kuriyan1986-mx. The inability of a B-factor to describe the heterogeneous mix of structural states found within the protein crystal contributes to the inability of protein models to satisfy an X-ray dataset within its theoretical uncertainty Holton2014-fl, Vitkup2002-de. A model that depicts anharmonic conformational substates found in a protein crystal will improve the satisfaction of X-ray data and reveal the structural basis of important biological properties at atomic resolution, such as allosteric networks or hidden cavities for small molecule binding (cryptic pockets) $\mathrm{Van}_d en_B edem2015 - cq, Henzler - Wildman2007 - ej$.

B. Approaches to modeling a heterogenous mix of structural states. There are 2 approaches for computing models that depict multiple structural states from X-ray data. First, conformational substates can be captured by computing single-state models that independently satisfy the X-ray dataset, as is seen in the modeling of Nuclear Magnetic Resonance spectra Schiffer1997-ag. For

Significance Statement

Authors must submit a 120-word maximum statement about the significance of their research paper written at a level understandable to an undergraduate educated scientist outside their field of speciality. The primary goal of the significance statement is to explain the relevance of the work in broad context to a broad readership. The significance statement appears in the paper itself and is required for all research papers.

Author affiliations: [a]Affiliation One; [b]Affiliation Two; [c]Affiliation Three

Please provide details of author contributions here.

Please declare any competing interests here.

[2]To whom correspondence should be addressed. E-mail: salisalilab.org

example, phenix.ensemble_refinement computes an ensemble from snapshots of a molecular simulation restrained by a time-averaged X-ray target function Burnley2012-yr. Such approaches may not find weakly occupied states in a reasonable amount of computation time, as they are dependent on overcoming potentially large barriers in the scoring function. Alternatively, a model can depict conformational substates by introducing additional structural variables Woldeyes2014-bs. All variables are then collectively fit against the X-ray data. The extent and detail of the additional degrees of freedom depend on the available computational power and data quality Wankowicz2024-rh. For example, qFit-3 avoids introducing excessive structural parameters by representing side chains as ensembles of one or more rotameric states Riley2021-ft. Methods that refine multiple fully parameterized atomic models have been limited to systems that diffract to ultra-high resolution Ringe1986-sa, Wilson2000-px, Levin2007-bl, Kuriyan1991-xd, Burling1994-om.

C. Multi-condition crystallography. Often, multiple X-ray datasets are collected for the same system under distinct experimental conditions. One example is multi-temperature crystallography, where data collection is performed at temperatures from cryogenic to near-physiological Thompson2023-pk. Data collection at higher temperatures has been shown to dramatically modify protein dynamics within the protein crystal Frauenfelder1979-zy, Frauenfelder1991-tg, Tilton1992-wj. More recently, models computed from higher temperature datasets have revealed a fuller set of structural states within the protein's energy landscape at atomic resolution Fraser2009-wh, Halle2004-kg, Keedy2014-hz. Comparisons of models of the same system computed at distinct temperatures show similar structural states but shifted thermodynamic equilibrium Ebrahim2021-er, Fraser2009-wh, Du2023-ny. Therefore, multiple X-ray datasets under distinct conditions containing mutual structural information could increase the data-to-parameter ratio and inform a more accurate, precise, and complete multi-state model.

D. Computing a multi-state multi-condition model. Here, we seek to compute a multi-state model from multiple X-ray diffraction patterns collected under distinct conditions (fig:fig1a). We can fit a multi-state model using multiple X-ray datasets without needing ultra-high-resolution X-ray data. We accommodate for the thermodynamic shift of distinct experimental conditions by computing each state's weight under each condition. In other words, we assume the sets of structural states are consistent across all experimental conditions with varied weights, allowing a single multi-state model to be informed by all X-ray datasets and massively boosting the data-to-parameter ratio. The assumption of a consistent set of structural states under all conditions is not limiting because the structural states may include sparsely populated or completely unpopulated states. We formulate a Bayesian posterior model density for the multi-state multi-condition model. We draw a sample from the Bayesian posterior model density using a molecular dynamics algorithm where all structural states are jointly restrained by the satisfaction of all X-ray datasets in addition to molecular mechanics. We benchmarked the method using synthetic X-ray datasets simulated from multi-state SARS-CoV-2 M^{pro} structural models and showed that multiple structural states are needed to satisfy the data and that by including multiple X-ray datasets, all datasets are better satisfied individually. We illustrate our method to compute a multi-state model of the SARS-CoV-2 viral target M^{pro} from multi-temperature X-ray under 6 conditions and show that all datasets are satisfied better by increasing the number of states and X-ray datasets.

Table 1. Comparison of the fitted potential energy surfaces and ab initio benchmark electronic energy calculations

Species	CBS	CV	G3
1. Acetaldehyde	0.0	0.0	0.0
2. Vinyl alcohol	9.1	9.6	13.5
3. Hydroxyethylidene	50.8	51.2	54.0

nomenclature for the TSs refers to the numbered species in the table.

Digital Figures. EPS, high-resolution PDF, and PowerPoint are preferred formats for figures that will be used in the main manuscript. Authors may submit PRC or U3D files for 3D images; these must be accompanied by 2D representations in TIFF, EPS, or high-resolution PDF format. Color images must be in RGB (red, green, blue) mode. Include the font files for any text.

Images must be provided at final size, preferably 1 column width (8.7cm). Figures wider than 1 column should be sized to 11.4cm or 17.8cm wide. Numbers, letters, and symbols should be no smaller than 6 points (2mm) and no larger than 12 points (6mm) after reduction and must be consistent.

Figures and tables should be labelled and referenced in the standard way using the `\label{}` and `\ref{}` commands.

Figure 2 shows an example of how to insert a column-wide figure. To insert a figure wider than one column, please use the `\begin{figure*}...\end{figure*}` environment. Figures wider than one column should be sized to 11.4 cm or 17.8 cm wide. Use `\begin{SCfigure*}...\end{SCfigure*}` for a wide figure with side legends.

Tables. Tables should be included in the main manuscript file and should not be uploaded separately.

Single column equations. Authors may use 1- or 2-column equations in their article, according to their preference.

To allow an equation to span both columns, use the `\begin{figure*}...\end{figure*}` environment mentioned above for figures.

Note that the use of the `widetext` environment for equations is not recommended, and should not be used.

Supporting Information Appendix (SI). Authors should submit SI as a single separate SI Appendix PDF file, combining all text, figures, tables, movie legends, and SI references. SI will be published as provided by the authors; it will not be edited or composed. Additional details can be found in the PNAS Author Center. The PNAS Overleaf SI template can be found here. Refer to the SI Appendix in the manuscript at an appropriate point in the text. Number supporting figures and tables starting with S1, S2, etc.

Authors who place detailed materials and methods in an SI Appendix must provide sufficient detail in the main text methods to enable a reader to follow the logic of the procedures and results and also must reference the SI methods.

Fig. 1. a, modeling can be framed as a model search given some input information. Here, the input information is the CHARMM19 force field parameters and J X-ray datasets collected for the same system under distinct experimental conditions (*eg*, temperature). The representation is the N atomic states containing all the heavy atoms of the system along with the weight matrix that parameterizes the weight of each state under each condition. All states and the weight matrix are scored collectively against each X-ray dataset. Each state is individually scored against the molecular mechanics force field. A sample is drawn from the posterior model density using molecular dynamics. All states are initialized by a starting structure and the force on the atoms is computed from the satisfaction of all X-ray datasets along with the molecular mechanics, and the weights are stochastically sampled.

If a paper is fundamentally a study of a new method or technique, then the methods must be described completely in the main text.

SI Datasets. Supply .xlsx, .csv, .txt, .rtf, or .pdf files. This file type will be published in raw format and will not be edited or composed.

SI Movies. Supply Audio Video Interleave (avi), Quicktime (mov), Windows Media (wmv), animated GIF (gif), or MPEG files. Movie legends should be included in the SI Appendix file. All movies should be submitted at the desired reproduction size and length. Movies should be no more than 10MB in size.

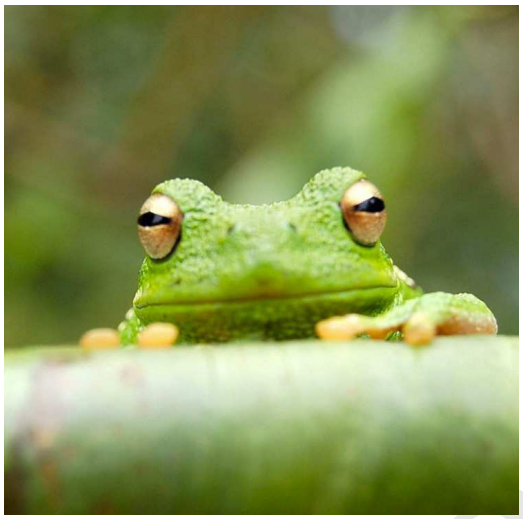

Fig. 2. Placeholder image of a frog with a long example legend to show justification setting.

by

Materials and Methods

Please describe your materials and methods here. This can be more than one paragraph, and may contain subsections and equations as required.

Subsection for Method. Example text for subsection.

ACKNOWLEDGMENTS. Please include your acknowledgments here, set in a single paragraph. Please do not include any acknowledgments in the Supporting Information, or anywhere else in the manuscript.

Fig. 3. This legend would be placed at the side of the figure, rather than below it.

$$\begin{aligned}(x+y)^3 &= (x+y)(x+y)^2 \\ &= (x+y)(x^2+2xy+y^2) \\ &= x^3+3x^2y+3xy^3+x^3.\end{aligned} \quad [1]$$