Title
Exploring the relationship between conformational heterogeneity and ligand binding

Author
Wankowicz, Stephanie Anne

Publication Date
2023

Peer reviewed|Thesis/dissertation

Exploring the relationship between conformational heterogeneity and ligand binding

by
Stephanie A Wankowicz

DISSERTATION
Submitted in partial satisfaction of the requirements for degree of
DOCTOR OF PHILOSOPHY

in

Biophysics

in the

GRADUATE DIVISION
of the
UNIVERSITY OF CALIFORNIA, SAN FRANCISCO

Approved:

DocuSigned by:
James Fraser
430BBB9A04D24A3...

James Fraser
Chair

DocuSigned by:
Andrej Sali
[illegible]

Andrej Sali

DocuSigned by:
Yifan Cheng
[illegible]

Yifan Cheng

DocuSigned by:
Michael Keiser
4DF1BD06D670465...

Michael Keiser

Committee Members

Acknowledgments

This thesis is dedicated to the many family members I have lost during graduate school Joanne Anderson, Kathy Allston, Eleanor Topping, Mike Mullane, and Dave Mullane. Their memories constantly remind me that science is nothing if we do not care and support each other.

To my advisor, James Fraser, I am grateful for your support and mentorship throughout my Ph.D. You have shown me the importance of fighting for what is most important in life and science, and your delicate balance of supporting your family, leading your lab, and advocating for changes in scientific education and communication is inspiring. I am happy knowing my graduate career is just the beginning of our working relationship.

I also want to thank my previous advisors, Eliezer Van Allen, Joaquim Bellmunt, and Paul Kostecki. Paul was the first person who suggested I become a researcher. Those words have never left me. Joaquim was an enthusiastic mentor who encouraged me to always go a step further with my research. I would not be where I am today without the platform and freedom he provided. Eli was a phenomenal mentor who helped me during one of the darkest moments of my career. I am grateful to have had the chance to watch him build his lab with intention both scientifically and operationally. I am grateful for his mentorship and friendship.

I further want to thank my other lab mates and collaborators at UMass Amherst, Dana-Farber Cancer Institute, Broad Institute, and UCSF who have supported and guided me

throughout my career. I especially want to thank David Liu, Brendan Reardon, Meng Xiao He, Jihye Park, Galen Correy, Iris Young, Robbie Diaz, Willow Coyote-Maestas, Andrej Sali, Yifan Cheng, Michael Keiser, and Nicole Flowers.

I would also like to thank the San Francisco running community. I feel so blessed to have met so many amazing people through many groups including November Project, SFRC, West Valley, and Trail Thursdays. Having goals and friends outside of science was a very nice breath of fresh air. I would also like to thank the many friends I have made through climbing as well as Erin Thompson and Jess Bolton, who I was lucky enough to 'ride' through the pandemic with. Finally, I want to thank my many other friends who always had an open ear including Hannah Moverman and Kayla Calabro.

My family's support has been wonderful during my entire education and crazy career path including graduate school. My parents, Denis and Jodie, and siblings, Dylan and Candace, always had faith in me and always reminded me of success when I was at my lowest. I also want to thank my in-laws, Cindy, Nick, Chris, Gabby, and Allison who also were huge supports.

Finally, I want to thank my husband, Alex, for his constant support and encouragement on this twisted path of my career and life. Your unwavering love and support have been my anchor during the good times and the bad. I am forever grateful for your love.

Contributions

Chapter 1

Riley, B.T., Wankowicz, S.A., de Oliveira, S.H., van Zundert, G.C., Hogan, D.W., Fraser, J.S., Keedy, D.A. and van den Bedem, H., 2021. qFit 3: Protein and ligand multiconformer modeling for X-ray crystallographic and single-particle cryo-EM density maps. Protein Science, 30(1), pp.270-285.

Chapter 2

Wankowicz, S.A., de Oliveira, S.H., Hogan, D.W., van den Bedem, H. and Fraser, J.S., 2022. Ligand binding remodels protein side-chain conformational heterogeneity. Elife, 11, p.e74114.

Chapter 3

Wankowicz, S.A., Fraser, J.S., 2023. Making sense of the chaos: uncovering the mechanisms of conformational entropy.

Exploring the relationship between conformational heterogeneity and ligand binding

Stephanie Anne Wankowicz

Abstract

Protein folding converts a disordered polymer to a globular structure, reducing many conformational degrees of freedom and incurring a significant conformational entropy penalty. However, residual conformational entropy is retained in a protein's folded native state. Subtle changes between positions within the native state, mostly from sidechains, alters residual conformational entropy, leading to differences in binding affinity and allosteric communication. While NMR has provided measurements of conformational entropy, these measurements do not provide information on where this entropy is coming from, such as if this is coming from a sidechain moving harmonically or anharmonically. However, we can take advantage of the fact that X-ray crystallography and CryoEM experimental data capture the conformational ensemble allowing us to measure the motion of residues and their atomistic structure. This provides an unparalleled platform to answer how, where, and why conformational entropy.

The first chapter of this thesis presents the improvements to the qFit algorithm. This algorithm allows for the automated modeling of multiple conformations per residue across a protein for high resolution X-ray crystallography or cryo-EM. We present algorithm improvements including the ability to run the program on a laptop. This algorithm was the basis for much of the future work of my thesis.

The second chapter contains the findings of the relationship between conformational heterogeneity and ligand binding. Using qFit, we identified the changes in conformational heterogeneity between matched bound and unbound high resolution X-ray structures. We identified a reciprocal relationship upon ligand binding where as binding site residues become more rigid, distant residues become more flexible, indicating an entropic compensation.

The third chapter contains my future outlook on the questions and techniques to probe conformational entropy mechanism. This chapter includes how to integrate new modeling techniques to understand how different motions of residues lead to differences in conformational entropy.

Table of Contents

List of Figures

qFit 3: Protein and ligand multiconformer modeling for X-ray crystallographic and single-particle cryo-EM density maps.

Ligand binding remodels protein side-chain conformational heterogeneity.

Chapter I

qFit 3: Protein and ligand multiconformer modeling for X-ray crystallographic and single-particle cryo-EM density maps.

Contributing authors

Blake T. Riley[1], Stephanie A. Wankowicz[2,3], Saulo H. P. de Oliveira[4], Gydo C. P. van Zundert[5], Daniel W. Hogan[2], James S. Fraser[2], Daniel A. Keedy[1,6,7], Henry van den Bedem[2,8]

[1]Structural Biology Initiative, CUNY Advanced Science Research Center, New York, NY 10031

[2]Department of Bioengineering and Therapeutic Sciences, University of California San Francisco, San Francisco, CA, USA

[3]Biophysics Graduate Program, University of California San Francisco, San Francisco, CA, USA

[4]Frontier Medicines Corporation, South San Francisco, CA 94080

[5]Schrödinger, New York, NY 10036

[6]Department of Chemistry and Biochemistry, City College of New York, New York, NY 10031

[7]Ph.D. Programs in Biochemistry, Biology, and Chemistry, The Graduate Center – City University of New York, New York, NY 10016

[8]Atomwise, Inc., San Francisco, CA 94103

Preface

The bulk of this chapter appears as Riley *et al.* preprinted in *bioRxiv* in 2021, and a version of which was ultimately published in *Protein Science* later the same year.

Abstract

New X-ray crystallography and cryo-electron microscopy (cryo-EM) approaches yield vast amounts of structural data from dynamic proteins and their complexes. Modeling the full conformational ensemble can provide important biological insights, but identifying and modeling an internally consistent set of alternate conformations remains a formidable challenge. qFit efficiently automates this process by generating a parsimonious multiconformer model. We refactored qFit from a distributed application into software that runs efficiently on a small server, desktop, or laptop. We describe the new qFit 3 software and provide some examples. qFit 3 is open-source under the MIT license, and is available at https://github.com/ExcitedStates/qfit-3.0.

Introduction

Conformational dynamics play an essential role in many aspects of protein function, including ligand binding, allostery, and enzyme turnover [1,2]. In each of these processes, the protein does not adopt a single conformation, but rather a conformational ensemble including a number of low-energy states. This ensemble can then be redistributed or reshaped by small-molecule binding, post-translational modifications, or other perturbations, thereby controlling biological function. To fully understand the fundamental interplay between protein conformational heterogeneity and function, it is

necessary to develop experimental and computational techniques to reveal alternative protein conformations in atomic detail.

X-ray crystallography is a powerful tool for addressing this need. Because individual protein molecules in the crystal lattice sample different conformations, there is a growing appreciation that crystallographic electron density maps contain a wealth of information about sparsely populated, alternative protein conformations[3]. Moreover, crystallography is undergoing an experimental renaissance: new tools are emerging with the potential to bias conformational distributions in crystals and gain new mechanistic insights into the links between protein dynamics and function.

For example, crystallographic datasets collected across multiple temperatures — as opposed to at a single cryogenic temperature — often reveal ensembles with more conformational diversity[4–8], including at dynamic enzyme active sites [9]. High-throughput crystallographic protein:ligand screening can identify otherwise undetectable low-occupancy ligand-bound protein states[7,10,11]. And time-resolved diffraction experiments, triggered by a variety of stimuli[12–15], can offer detailed windows into how protein conformational ensembles evolve in real time. Time-resolved experiments are becoming more accessible as serial microcrystallography experiments can take place not only at X-ray free-electron lasers, but also at third-generation synchrotrons with microfocus beamlines [16]. Serial microcrystallography can also help reveal alternative protein states by dissecting distinct crystal polymorphs within the microcrystal population [17]. These advances, coupled with an ever-growing level of automation and faster X-ray detectors

[18], are yielding larger amounts of data that highlight the need for automated (rather than manual) computational methods for modeling alternative conformations and their correlations in electron density maps.

In parallel to the renaissance for X-ray crystallography, cryo-electron microscopy (cryo-EM) is in the midst of a "resolution revolution" [19]. Recently, cryo-EM structures of apoferritin at "atomic resolution" (1.2–1.25 Å) [20,21] demonstrated how far this method has come in recent years. Similar to electron density maps from X-ray crystallography, Coulomb potential maps from cryo-EM reveal evidence for alternative protein states, which in this case are sampled by individual protein molecules on the microscopy grid. Unfortunately, so far no methods exist for unbiased and automatic modeling of correlated alternative conformations in cryo-EM maps. Additionally, many cryo-EM structures feature large protein complexes with thousands of amino acids, posing a significant challenge to traditional model building approaches. Efficient, automated algorithms [22] could meet this challenge for cryo-EM.

There is thus a clear need for computational model-building methods that better explain X-ray and cryo-EM data by incorporating alternative conformations. Protein conformational heterogeneity can be represented using various approaches, including B-factors, multi-copy ensembles, or multiconformer models [1]. First, B-factors are present for every atom in the Protein Data Bank (PDB) [23] file format. Theoretically, B-factors represent the harmonic, thermal displacement of each atom about its mean position, either isotropically with one parameter or anisotropically with six parameters [24].

However, in practice, B-factors often absorb uncertainty in a more general sense about each atom's position, and are insufficient representations of anharmonic motions such as transitions between side-chain rotamers [25]. Second, multi-copy ensemble models consist of some number (>1) of full, independent copies of the protein with distinct xyz coordinates and B-factors that collectively explain the experimental data [26]. Ensemble models can successfully describe discrete conformational heterogeneity such as rotamer transitions -- but they unnecessarily inflate the number of model parameters for those regions of the protein with essentially a single, unique conformation [27]. Finally, multiconformer models lie somewhere in the middle in terms of model complexity. A multiconformer model represents local, anharmonic features in the data with a small number (2–5) of discrete conformations, but represents regions of the protein that show little to no evidence of flexibility with a single conformation. These conformations are assigned labels ("alternative locations" or "altlocs"), such as A, B, etc., with corresponding occupancies in the PDB format on a per-atom basis. Groups of atoms whose alternate positions are correlated (side chains, stretches of contiguous backbone, collective exchange across an active site, etc.) are assigned the same label and occupancy. When constructed in a parsimonious manner, multiconformer models can limit a model's complexity while maximizing its explanatory power.

To efficiently generate parsimonious multiconformer models for protein X-ray crystal structures, we previously introduced the software package qFit [28]. Besides providing mechanistic insights, for example by revealing hidden protein contact signaling networks [29] and allosteric pathways [7], multiconformer qFit models have also established

that the conformational ensemble at room temperature is not dominated by radiation damage [30], and that the effect of crystal dehydration on the conformational ensemble is similar to that of cryocooling [31]. We recently introduced multiconformer treatment of ligands in complex with proteins in a standalone version, *qFit-ligand* [32]. However, previous versions of qFit were computationally demanding (requiring a high-performance computing cluster), and were restricted to density maps from X-ray crystallography only, among other limitations.

Here we report a new, refactored version of qFit, which we call qFit 3, with several key improvements. qFit 3 operates on maps from either X-ray crystallography or cryo-EM. It combines multiconformer modeling of proteins and of ligands complexed with proteins (from *qFit-ligand*) in a single software package written in Python. The software distribution includes a script to refine the multiconformer model generated by qFit with Phenix [33]. Importantly, we reduced the runtime by two orders of magnitude. qFit 3 typically runs for a ~300 residue protein in several hours on a laptop, making it significantly more accessible to users.

Overall, qFit 3 reveals hidden alternative conformations in protein structures in a rapid, automated, and unbiased manner. This new software will allow a broader array of users to explore conformational heterogeneity in their systems of interest. It will also smooth the path toward integrating new and exciting types of structural biology data, including series of datasets related by temperature, ligands, or time, as well as biologically important and/or large protein systems from X-ray free electron lasers (XFELs) or cryo-

EM. qFit 3 will thus empower novel studies of the relationship between protein dynamics and biological function.

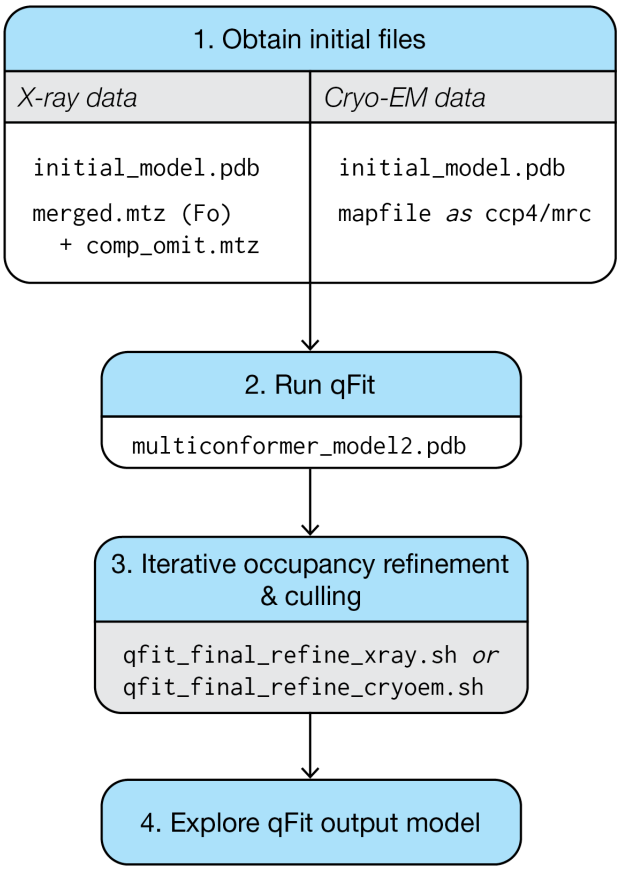
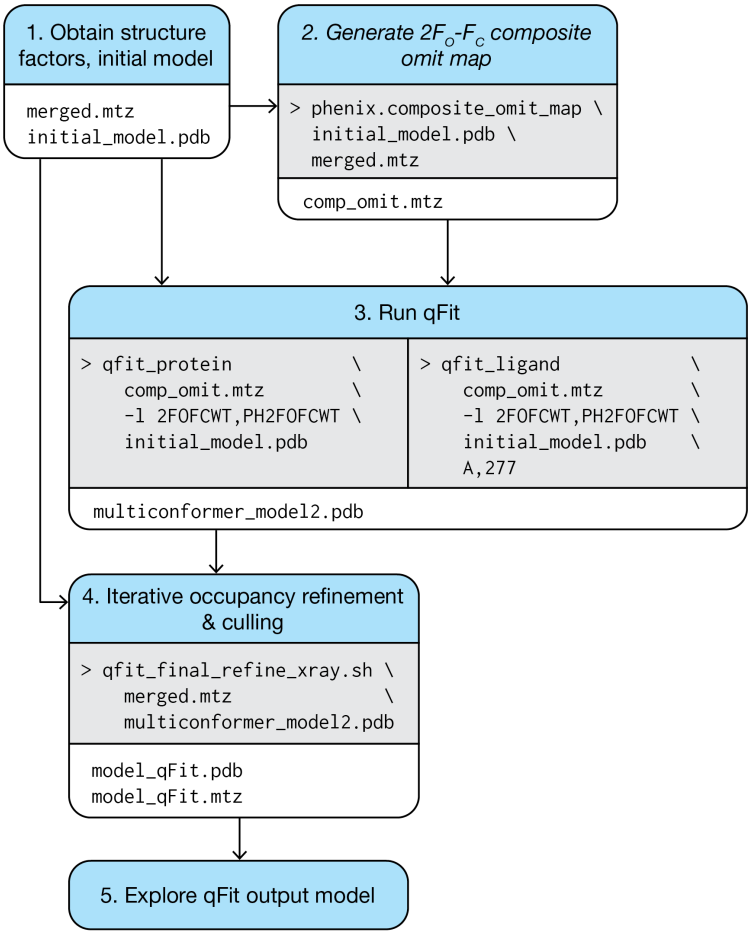


Figure 1.1| Usage flowchart for qFit 3 for either protein or ligand inputs and for either X-ray or cryo-EM data. (1) qFit requires an initial model and map information. In the case of X-ray diffraction data, qFit will require both the structure factors and a high-quality, unbiased map, such as a composite omit map. (2) With these files, qFit will generate a parsimonious model (multiconformer_model2.pdb) containing the fewest number of sampled conformers that explain the experimental data. (3) This intermediate/preliminary model should proceed through an iterative procedure to refine the occupancies of conformers in the model, and cull those conformers that have <9% occupancy. (4) The resulting model can then be used to explore conformational diversity.



Supplementary Figure 1.1| A flowchart for typical use of qFit with X-ray data.

Results

qFit was completely refactored in the Python programming language and released as open-source software; see Methods and the GitHub repository (https://github.com/ExcitedStates/qfit-3.0) for more details. A typical qFit 3 workflow is illustrated in **Figure 1.1 and Supplementary Figure 1.1**. qFit 3 takes as minimal input a starting model and either a real-space map in the MRC/CCP4 format or map coefficients in the MTZ format. For X-ray crystallography, the preferred map is a composite omit map to minimize model bias, which can be readily generated with Phenix. For cryo-EM, the input is a real-space map together with the resolution of the

data and a flag to use electron scattering factors for generating synthetic densities. qFit 3 relies on a sample-and-select procedure based on constrained optimization to identify alternative conformations of proteins and their ligands. To ensure optimal model selection and prevent overfitting, qFit 3 evaluates increasing model complexities, selecting the model with the lowest Bayesian Information Criterion (BIC). qFit 3 now also provides all functionality to model ligand alternate conformations, previously available separately in *qFit-ligand*. A distinctly important new feature is qFit 3's capability to model alternate conformations into cryo-EM maps. Numerous additional options and details are described in the Methods section and can be found in the qFit 3 GitHub repository. Here, we demonstrate typical use cases of qFit for protein systems and their ligands. All analyses in this section used default parameters, unless otherwise stated.

We first carried out qFit 3 modeling on a previously deposited cryogenic X-ray structure of a protein tyrosine phosphatase, PTPN18 (PDB ID: 2OC3) [34]. While the deposited model includes ten residues with alternate conformers, a difference density map shows unmodeled positive density over 3σ around Phe30 and Gln34 **(Figure 1.2)**. qFit 3 models suggest that an alternate conformer for Phe30 and an ensemble of three side-chain conformers for Gln34 better fit the density, and reduce nearby difference density peaks (**Figure 1.2A)**. Running on a quad-core processor, qFit sampled and selected alternative conformations for this 290-residue protein in 12.75 hours.

The default algorithm of qFit 3 changed slightly compared to earlier versions. Previously, each amino acid in turn was truncated at the Cβ atom and refined anisotropically. This had two advantages: 1) it generally positioned the Cβ atom at the peak *average* density of potential alternate conformations, and 2) the anisotropy of the atomic displacement parameter provided guidance for backbone motions. Although this earlier version often better captured subtle backbone movements, it led to significantly increased computational expense and complexity [35]. Nonetheless, the present version of qFit can be made to mimic the behavior of the earlier algorithm on a single residue by providing an alternative input. A thoroughly-tested room-temperature structure of the peptidyl-prolyl cis-trans isomerase CypA (PDB ID: 3K0N) displays multiple conformers for Phe113 [9]. Starting from a single conformer (**Figure 1.2**), we truncated Phe113 at Cβ, refined the structure anisotropically, calculated a composite omit map, and used this as input to qFit 3. This pre-processing enabled qFit to recapitulate the alternative conformations observed in the published model (**Figure 1.2**). With Phe113 in place, qFit 3 ran in 460 min over the other 161 residues. This computationally expensive pre-processing procedure is provided as an option, and improved backbone modeling will be a focus of future development.

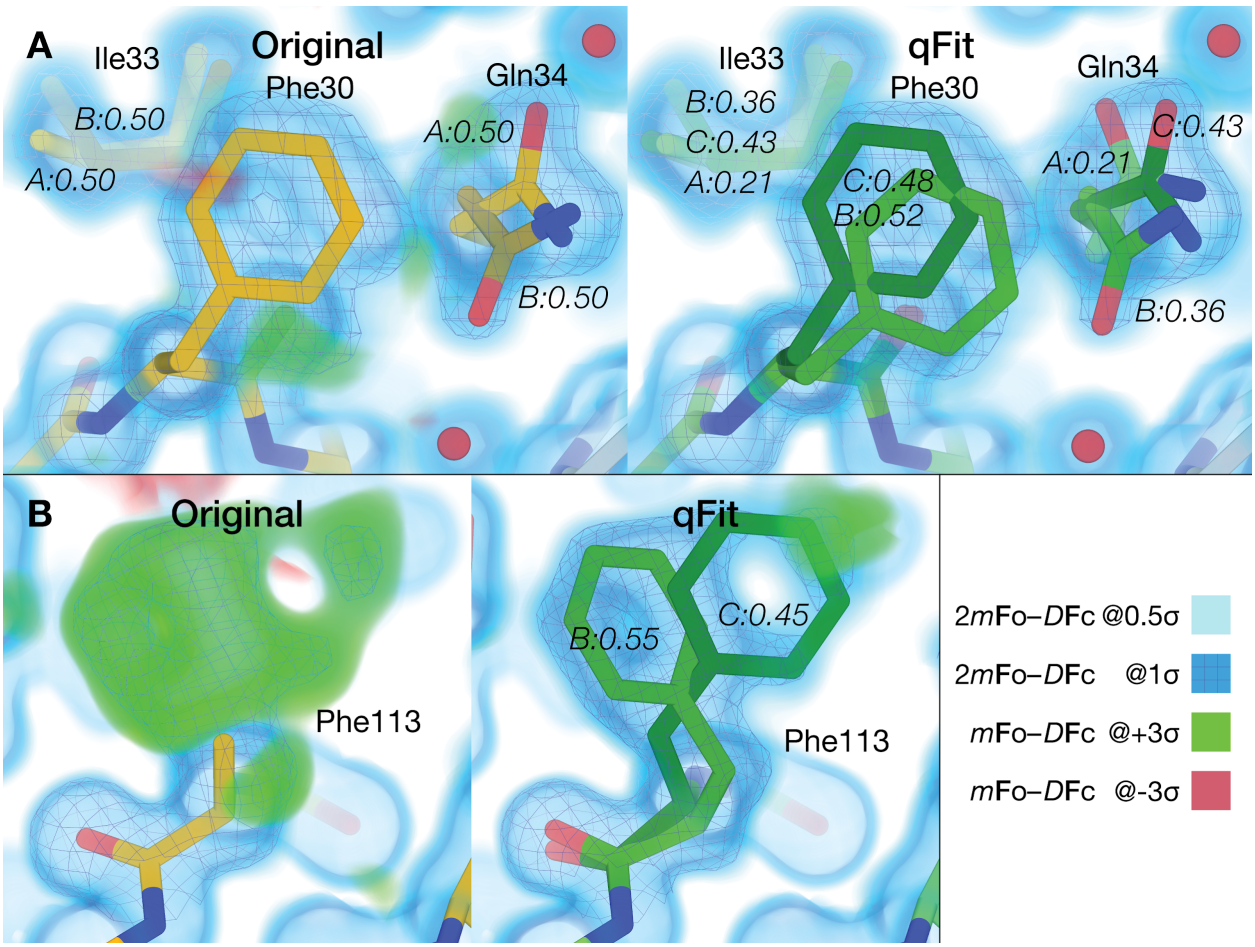
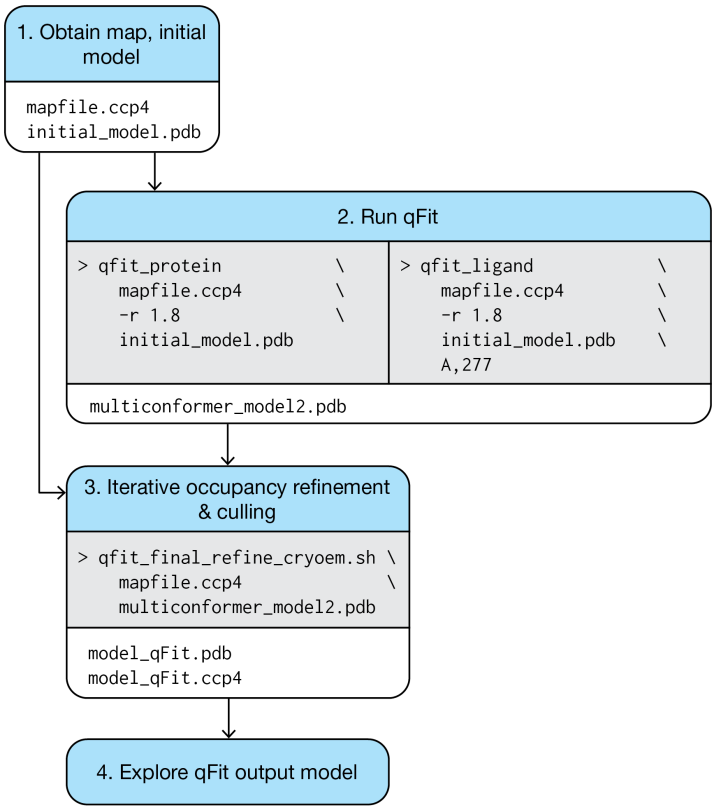


Figure 1.2| qFit 3 recapitulates deposited alternate conformations in X-ray crystallography density maps, and suggests additional conformations to explain unmodeled density. *(A) Left*: PTPN18 (PDB ID: 2oc3) displays regions of unmodeled density near Phe30 and Gln34 in the deposited *m*Fo-*D*Fc difference density map at +3σ (green cloud). These are visible in a 2*m*Fo-*D*Fc composite omit density map contoured at 1σ (blue mesh), which is clarified by a low-density 0.5σ contour (blue cloud). *Right*: qFit 3 adds extra conformers to model these residues. Gln34 is modeled by three conformers (corresponding to the rotamers **mm**110, **mt**0, **mt**0 [25]); Phe30 is also modeled by two conformers (both in the "favored" **t**80 rotamer space). The distance between Phe30 and Gln34 doesn't lead to steric hindrance between any of the conformers of either residue. Note that qFit 3 sets the minimum number of conformers in Ile33 to three (because of Gln34) to ensure backbone consistency; Phe30 is part of another backbone segment. (B) *Left*: Following the methodology in qFit 2 [35], Phe113 was truncated at Cβ and refined. Both the composite omit map and the difference map indicated the presence of at least two conformers for this residue. *Right:* qFit 3 sampled and selected two conformers of Phe113 (matching the two known ones) to explain the density in the composite omit map.

qFit 3, for the first time, also accepts cryo-EM density maps as input. We have adopted the simplified scattering factor calculation of averaging the contributions of all atoms to calculate synthetic maps, as is used in real-space refinement in Phenix (**Supplementary Figure 1.2)**[36]. As an example application of this new functionality, we ran qFit 3 on two ultra-high-resolution cryo-EM structures: β3 GABA receptor [21] (1.2 Å resolution) and apoferritin [20] (1.7 Å resolution). qFit 3 was run on both chain A and the entire structure for both examples. Chain A of apoferritin (176 residues) had a runtime of 112 minutes using four cores.



Supplementary Figure 1.2| A flowchart for typical use of qFit with cryo-EM data.

For these examples, qFit 3 captured both previously modeled and newly modeled alternative conformations (**Figure 1.3**). Within chain A, there were originally 19 residues with modeled alternative conformers. qFit 3 successfully identified alternate conformations for 16 (84.2%) of these residues and suggested 66 additional residues with alternative conformations. In **Figure 1.3,** we demonstrate the ability of qFit 3 to recapitulate alternative conformers in Ser124. In **Figure 1.3**, we demonstrate the ability of qFit 3 to detect a new alternative rotamer for Gln14 (pt0 and mm-40 [25], RMSF 1.16 Å).

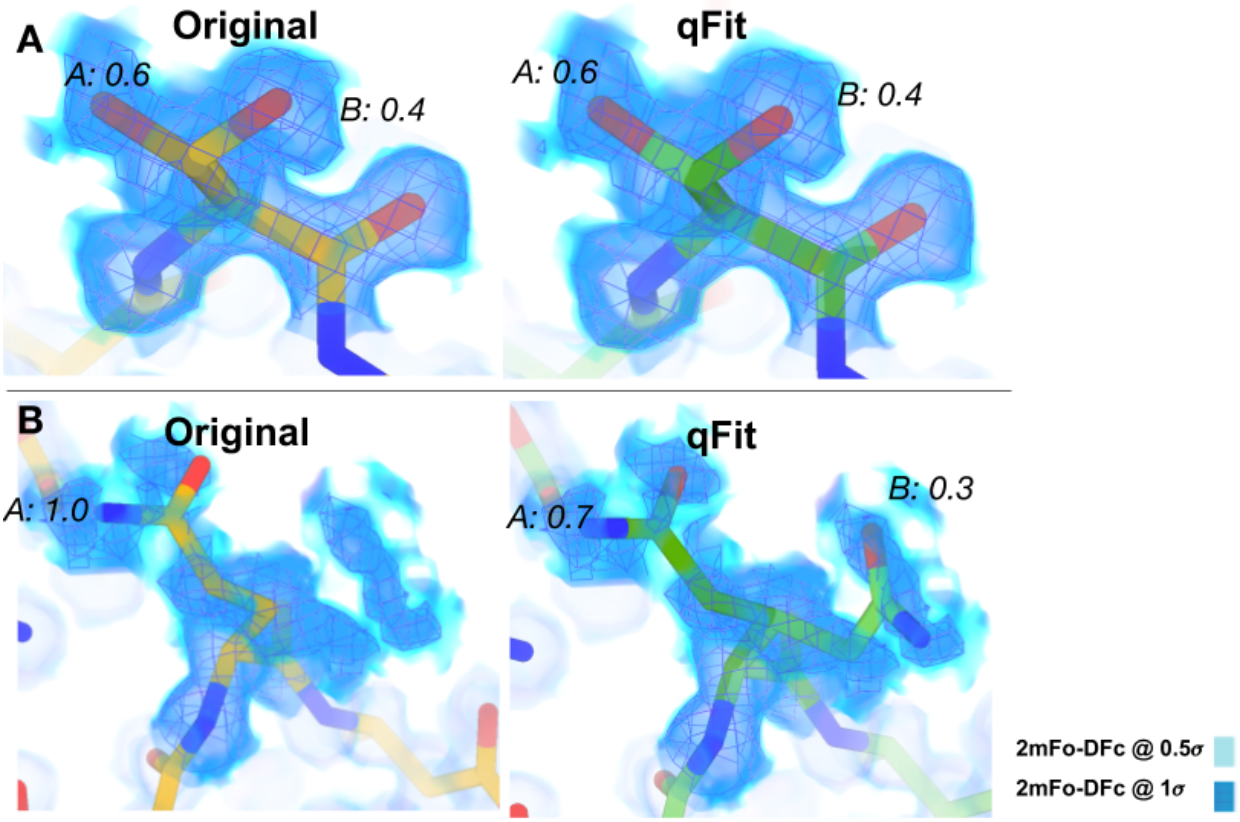


Figure 1.3| qFit 3 recapitulates deposited alternate conformations in cryo-EM density maps, and suggests alternate conformations to explain noisy data. (A) *Left*: Deposited alternative conformations for Ser113 in a high-resolution published cryo-EM structure of apoferritin (PDB ID: 6v21). These are visible in a 2*m*Fo-*D*Fc composite omit density map contoured at 1σ (dark blue cloud) and at 0.5σ (light blue cloud and blue mesh). *Right*: qFit 3 and subsequent refinement successfully modeled identical alternative conformations. Occupancies are indicated in italics. *(B) Left*: Deposited single conformation for Gln14 in the same structure of apoferritin.*Right*: qFit 3 and subsequent refinement identifies the original conformer, plus an alternative conformer (**mt** and **tt** rotamers [25]).

Additionally, qFit 3 can determine alternative conformations of ligands [32]. Distinct ligand conformations can play an important role in determining binding affinities, activity, and disassociation from the protein. Visualizing ligand alternate conformations can help determine the role of entropy in binding affinity, or help guide lead optimization in drug discovery [37]. *qFit-ligand* takes a model, map, and information about the position of the ligand of interest (chain and residue number). The output is a set of conformations of

the ligand. In **Figure 4**, we show two examples of ligands taking on multiple conformations to two different proteins, CDK2[38] and Human Leukotriene A4 Hydrolase[39].

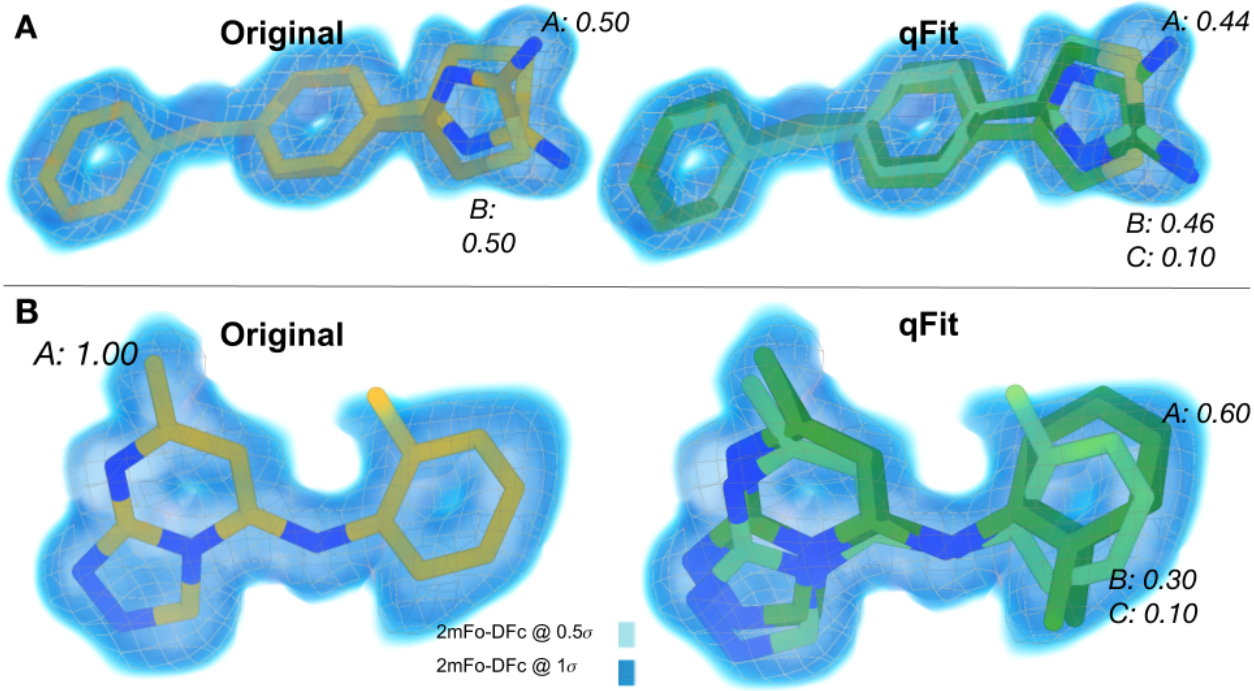


Figure 1.4| qFit 3 generates occupancy-weighted multiconformer models for bound ligands. (A) *Left:* Deposited alternative conformations of thiazolylpyrimidine, an inhibitor of CDK2, in a co-crystal structure (PDB ID: 5hq5). The 2*m*Fo-*D*Fc composite omit density map is contoured at 1σ (dark blue cloud) and at 0.5σ (light blue cloud and grey mesh). Occupancies of alternative conformations are labeled in italics. *Right: qFit-ligand* successfully identifies both deposited alternative thiazolylpyrimidine conformations, as well as an additional, similar conformer. *(B) Left:* Deposited conformation of 4-(4-benzylphenyl)thiazol-2-amine, an epoxide hydrolase selective inhibitor, co-crystallized with human Leukotriene A4 Hydrolase (PDB ID: 4l2l)[39]. *Right: qFit-ligand* models both the deposited 4-(4-benzylphenyl)thiazol-2-amine conformation and suggests two additional conformations that, unlike the deposited conformation, fit entirely within the 1σ density contour.

Discussion

qFit 3 is a significantly faster implementation of the qFit algorithm that can now run on commodity computer hardware like a laptop. It is open-source and freely available, with simple installation instructions. qFit 3's speed enables application of the qFit approach to series of multiple datasets generated by new high-throughput methods in crystallography; to large, increasingly high-resolution cryo-EM structures with many thousands of amino acids; and to many more structural bioinformatics studies that focus on conformational heterogeneity.

Although qFit 3 can be run in an automated fashion on large (numbers of) structures, the user should apply caution in interpreting its multiconformer models. False positives can occur when qFit 3 selects spurious alternative protein conformations based on density that corresponds to other atoms such as water molecules. False negatives can occur when qFit 3 fails to sample backbone conformational space sufficiently. Development of qFit is ongoing and the user community is invited to contribute to the open-source project at https://github.com/ExcitedStates/qfit-3.0.

To improve qFit further, we envision several new developments. For example, qFit's backbone sampling methodology has ample room for improvement. Currently in qFit, each amino acid's backbone is translated along the principal axes of the anisotropic ellipsoid of the Cß atom (or O for Gly), while closure of the backbone is maintained by torsion-based nullspace inverse kinematics, thus positioning it to accommodate suitable alternative side-chain rotamers (Methods). Although this current backbone sampling is powerful for capturing small-scale motions, it is limited in its ability to capture larger

ones (**Figure 2B**). A suite of backbone sampling methods in qFit, ranging from backrubs [40] and helix "shear" [41,42] to inverse-kinematics-based loop modeling [43], would be able to overcome this limitation. These new methods will allow qFit to model alternative conformations that are related to each other by larger, biologically relevant motions, as with loops in protein tyrosine phosphatase 1B (PTP1B) [7] and helices in isocyanide hydratase (ICH) [15]. A related challenge is that hierarchical alternative conformations — such as alternative loop or helix backbone positions that each have alternative side-chain rotamers — are not supported in the existing PDB format. It may be possible to use additional restraints to bypass this limitation, as with refinement of the multi-state models from PanDDA [44], which are conceptually related but distinct from the multiconformer models from qFit. Alternatively, the new PDBx/mmCIF format that was recently adopted by the PDB could be used to explicitly define hierarchical relationships between alternative conformations.

Another important direction is improving ligand models, and correlating protein alternate conformations with alternate ligand binding modes. Currently, qFit lacks chemical knowledge of ligand atoms such as hybridisation and protonation. Incorporating this knowledge, for example with the help of sophisticated force fields that work in tandem with crystallography maps [45], will greatly improve ligand model quality and help determine the precise interactions between protein and ligand.

Finally, the problem of compositional heterogeneity must be addressed. Some of the alternative conformations in the protein may be in response to the ordering of other

components in the unit cell (heteroatoms such as ligands, crystallographic additives, and solvent). While multi-dataset approaches, such as PanDDA [11], may increase confidence in modeling partially occupied ligands and crystal additives, addressing the problem of partially occupied solvent may be bootstrapped by using stereotypical interactions in a solvated rotamer library [46]. Solving this problem will also help to better define the border between proteins or ligands and bulk solvent [47], which is likely to be key to reducing the "R-factor gap in crystallography" [48].

X-ray crystallography and cryo-electron microscopy remain the dominant experimental techniques to obtain structural information for proteins and their complexes with other macromolecules or with ligands, like therapeutic chemical compounds. New, emerging experimental techniques in X-ray crystallography and ever-increasing resolution limits in cryo-EM can reveal an ensemble of protein and ligand conformations that can provide insights into molecular mechanisms and function. qFit 3 automates interpreting an ensemble from X-ray or cryo-EM density maps, and generates an unbiased, internally consistent, parsimonious model of conformational heterogeneity. We refactored qFit with a specific focus on efficiency and ease-of-use, so that it effortlessly installs and runs on a standard laptop to facilitate advanced interpretation of experimental structural biology data.

Methods

qFit algorithm

qFit samples a large number of conformers and uses a deterministic approach to select a small ensemble of these conformers that parsimoniously explains local density. The method starts from an initial single-conformer model and generates candidate conformers for each residue/ligand in the initial structure. It evaluates all possible combinations of these conformers to determine an optimal ensemble. A final relabeling step ensures that conformers of different residues/ligands have consistent altloc labels. For all analyses in this manuscript, default parameters were used unless otherwise stated. **Figure 1.5** provides a graphical overview of both the *qFit-protein* and *qFit-ligand* algorithms, the two main command-line utilities of the qFit 3 package for automatic multiconformer modeling of proteins or ligands.

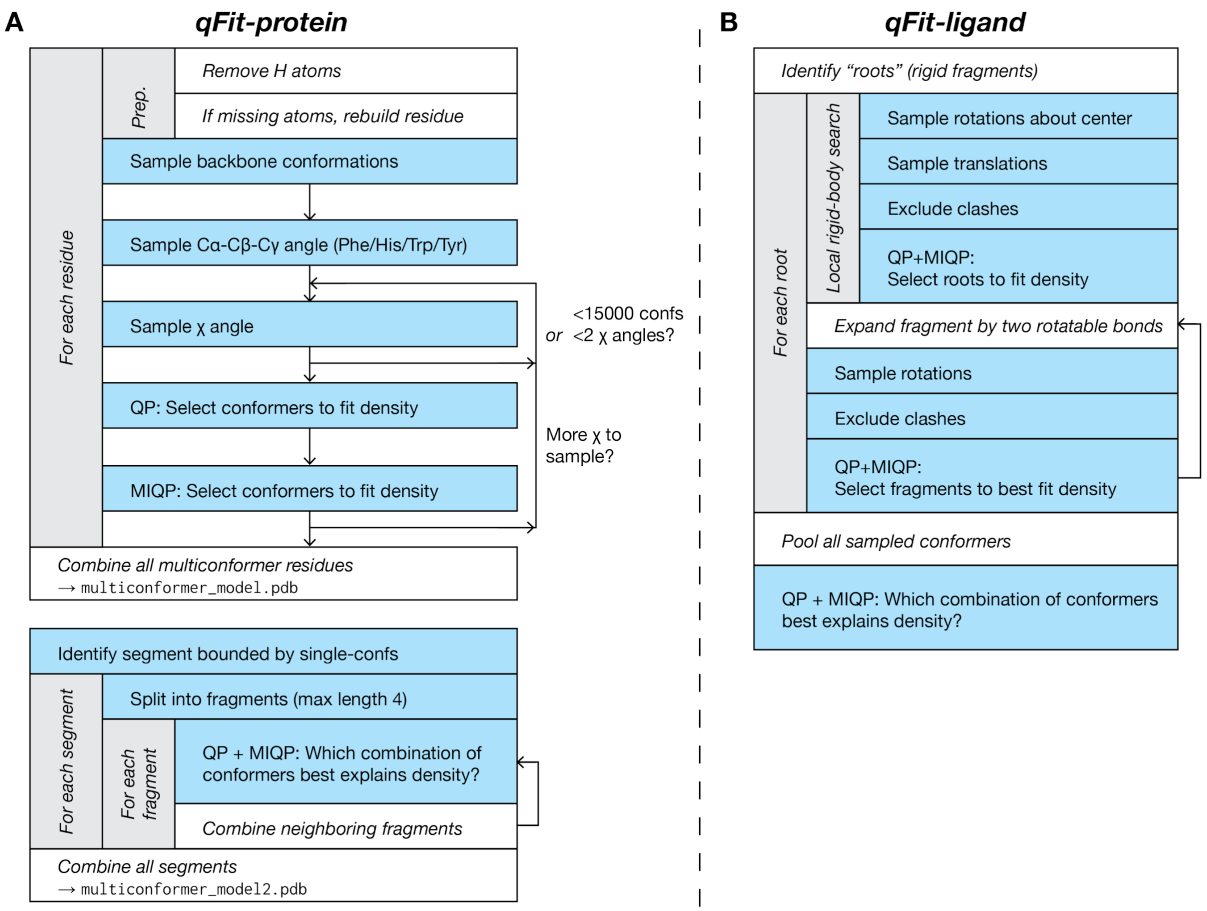


Figure 1.5| A flowchart of the sample-and-select protocols for (A) *qFit-protein*, and (B) *qFit-ligand*.

The qFit 3 protocol accepts input density maps or map coefficients in several commonly accepted crystallographic or cryo-EM file formats (MTZ, CCP4). For best performance, we recommend the use of a composite omit map for crystallographic densities [49]. All runs of qFit 3 on crystal structures described in this manuscript used an input composite omit map generated with the *phenix.composite_omit_map* command from the Phenix software suite [33]. Refinement was carried out on each partial model (*omit-type=refine*) and default parameters were used for this calculation. qFit 3 also expects a PDB file containing the structure of interest as input. Hydrogens are automatically removed to provide uniform treatment of input models. Note that during the final refinement stage,

hydrogens will be (re-)added (see Final refinement script). For analyses described in this manuscript, we removed all alternate conformers (except for altloc A) using the *phenix.pdbtools* executable and used the resulting single-conformer input structure as input for all subsequent modeling.

Map treatment

qFit 3 converts the input maps to absolute scale following the protocol described in reference [50]. The software creates a lookup table corresponding to the theoretical spatial density value distribution for each atomic element for radial shells spaced at 0.01 Å. The mask radius for this calculation is resolution-dependent (default radius = 0.5 Å + resolution/3). qFit 3 indirectly avoids clashes during sampling by means of a real-space density subtraction. It uses all atoms whose conformations are not being sampled to calculate a density map to perform this real-space subtraction. This prevents undesirable modeling into density from neighboring residues/side chains. The mask radius and an option to use excluded volume for clash detection instead, as detailed in [32], can be determined via the command line. Different sets of scattering factors are used for electron density maps from X-ray crystallography vs. Coulomb potential maps from cryo-EM. For convenience, we refer to both types of maps as “density maps” in this paper.

Conformational sampling for residues

qFit 3 exhaustively samples residue conformations in three stages: backbone sampling, Cα-Cß-Cγ bond angle sampling (for certain residues), and side-chain sampling (**Figure**

5A). These are all enabled by default, but can be individually disabled via command-line options.

Backbone sampling

qFit 3 samples backbone conformations by means of a nullspace inverse kinematics algorithm [28,35,43]. Backbone sampling for each residue extends to neighboring residues, two on each side. Backbone sampling is not performed if a residue lacks two neighbors on both sides (e.g., close to terminal residues). The Cß atom of the residue of interest (or O atom for Gly) is moved in the direction of the major and minor axes of its thermal ellipsoid. By default, three amplitudes for this sampling are used (0.1 + σ, 0.2 + σ, 0.3 + σ), where σ is randomly selected in the interval [-0.125, 0.125].

The amplitude scaling factor and the maximum value of σ can be defined at input. In total, three amplitudes times six directions = 18 positions for the Cß (O in case of glycine) are tested. The five-residue fragment is then deformed using nullspace inverse kinematics and dihedral angle degrees of freedom. The input conformation is also added to the ensemble, leading to 19 backbone conformations after backbone sampling. Peptide flips [35] are not yet implemented in qFit 3.

Cα-Cß-Cγ bond angle sampling

For amino acids with large planar aromatic groups (Phe, Tyr, Trp, His), qFit samples around the Cα-Cß-Cγ bond angle of the 19 backbone conformations resulting from the previous sampling step. For each conformation, we sample the Cα-Cß-Cγ bond angles

as follows: [θ - 7.5º, θ - 3.75º, θ, θ + 3.75º, θ + 7.5º]. Both the range and the step of the bond angle sampling can be adjusted via command line. This step expands the number of sampled conformations to 95 for the large planar aromatic residues.

Side-chain sampling

Side-chain sampling in qFit 3 is performed by iteratively rotating around the χ angles of ideal rotamers. The protocol begins by rotating around χ_1. For each of the (19 or 95) backbone conformations, we rotate around each of the rotamers for the target residue in the penultimate rotamer library [25]. For each rotamer, we explore a sampling window using a rotamer neighborhood of [-60º, +60º] at 10º intervals. Both the sampling window and the step size can be defined via command-line options. For the default parameters, at most 19*5*(8+1)*13 = 11,115 conformations are generated (with either Phe, Tyr, Trp, or His), which provides a balance between performance and accuracy. From this set, we remove conformations that lack support from the subtracted density map (voxel with minimum density intensity < 0.3 e^{-1} $Å^{-3}$), conformations that contain self-collisions (based on hard spheres), and conformations that are redundant (using an all-atom RMSD threshold of 0.01 Å). These exclusion strategies can be adjusted via command-line options. For protein and ligand atoms, B-factor sampling is also a non-default option.

Once the backbone and χ_1 sampling is complete, the protocol initiates a selection step based on our optimization strategy (see Optimization protocol for more details). We select all atoms starting from the backbone up to the atoms involved in the χ angle

being sampled (χ_1 in this first iteration). The remaining atoms are rendered inactive, and their density contribution is not taken into account during optimization. Up to five conformers can be selected at each iteration, which then serve as the basis for sampling of subsequent χ angles.

From the second iteration onwards, we sample up to two χ angles simultaneously (also defined at command line). After sampling χ_i we exclude unsupported, clashing, and redundant conformers (as outlined above) and use this filtered conformer ensemble to sample around χ_{i+1}. In the worst case scenario (Arg), χ_i leads to 5*(34+1)*13 = 2,275 conformers and up to 2,275*(34+1)*13 = 1,035,125 conformations are produced for χ_{i+1}. In practice, this number of conformations is never produced owing to redundancy. We limit the number of conformations that can be used during optimization to 15,000 for computational efficiency and memory (RAM) constraints. If sampling two χ angles in a single iteration leads to more than 15,000 conformers, we reduce sampling to a single χ for that iteration. Side-chain sampling concludes when all χ angles have been sampled.

Conformational sampling for ligands: Ligand sampling in qFit 3 is performed in two steps: a local rigid body search followed by an iterative step which samples the degrees of freedom about the flexible areas of the ligand [32] (**Figure 5B).** For the local search, we identify all possible roots, i.e. rigid fragments of atoms. Rigid fragments are defined as a set of connected atoms that do not contain a rotatable bond. We sample conformations starting from each possible root. Around the center of each ligand root, we test 100 possible rotations, by sampling rotation space in intervals of [0°, 10°]. For each rotation,

we enumerate possible translations for x, y, and z coordinates in the interval [-0.2 Å, 0.2 Å] at 0.1 Å increments. The local search leads to 100(rotations)*125(translations) = 12,500 conformers. We then exclude conformers that do not have support from the density (voxel with minimum density intensity < 0.3 e^{-1} $Å^{-3}$) and conformations that are redundant, using an all-atom RMSD threshold cutoff of 0.01 Å. Additionally, conformers with internal (ligand) or external clashes (receptor) are removed using a spatial hashing algorithm, which efficiently converts the 3D coordinates to a 1D hash table to determine if the sampled portion of the ligand occupies the same spatial coordinates as any other part of the ligand and/or receptor. After this exclusion step, remaining conformations are used as input for the optimization routine (see below), which selects up to five conformers of each root to best represent the local density.

Still treating each root independently, we take the root fragments selected by the local rigid body search and "expand" each fragment to the full ligand, by iteratively sampling around rotatable bonds. The protocol follows a rotatable bond hierarchy from the root to the extremities of the molecule. For each rotatable bond, we sample all angles in a [0^o, 360^o] interval at 10^o increments. Two rotatable bonds are sampled at a time, leading to 5*36*36 = 6,480 conformations per iteration. At each iteration, we exclude conformers that do not have support from the density (voxel with minimum density intensity < 0.3 e^{-1} $Å^{-3}$), those with an all-atom RMSD of <0.01 Å, or that contain internal or external clashes. After exclusion, qFit uses the optimization routine to select up to five conformers to be used for the next iteration. After all rotatable bonds have been sampled, up to five conformers can be output for each root. One final optimization step

is used to select up to five consensus conformers from the pool of conformers produced across all roots.

Optimization Protocol: We frame the problem of selecting a subset of conformers that best represents local density as an optimization problem. Each conformer has an occupancy ω_i associated with it. The vector of all occupancies $\boldsymbol{\omega}^T$ contains the variables for the optimization, with the extra constraints that ω_i are non-negative and their sum lies in the unit interval. We optimize real-space residuals, calculated from the observed density (ρ^{obs}) against the occupancy-weighted sum of the calculated densities (ρ_i^{calc}) for all conformers. We can formulate this problem as constrained quadratic optimization:

$$\begin{aligned} \min_{\omega} \quad & ||\rho^{obs} - \sum_i \omega_i \rho_i^{calc}||_2 \\ \text{subject to} \quad & 0 \leq \sum_i \omega_i \leq 1 \\ & \omega_i \geq 0 \quad \text{for } i = 1, \ldots, N. \end{aligned}$$

Residuals are calculated over all voxels within (0.5 + resolution / 3) Å from any active atoms across all input conformers. To prevent overfitting conformers with arbitrarily small occupancies, we require a threshold constraint on the occupancies, turning the problem into a mixed-integer quadratic program (MIQP):

$$\begin{aligned} \min_{\omega} \quad & ||\rho^{obs} - \sum_i \omega_i \rho_i^{calc}||_2 \\ \text{subject to} \quad & z_i t_{d_{min}} \leq \omega_i \leq z_i, \; z_i \in \{0,1\}^N \\ & 0 \leq \sum_i \omega_i \leq 1, \text{ for } i = 1, \ldots, N. \end{aligned}$$

Note that this ensures that the number of conformers selected is at most $1/t_{d_{min}}$. The optimal threshold parameter is determined using a penalized-likelihood criteria (see below). An MIQP is NP-hard, thus applying an MIQP solver directly to the conformers output from our sampling step is computationally inefficient [28,35]. Applying a QP solver to the thousands of conformers output from our sampling routine, and then selecting the QP-fitted conformers with non-zero occupancy as input for MIQP, allows for near-optimal solutions to be calculated within a tractable time. Our protocol uses cvxopt (https://cvxopt.org/) and a proprietary, freely available implementation of the IBM ILOG CPLEX Optimization Studio (Python API, version 12.10) to solve QP and MIQP programs.

Achieving parsimony by means of the Bayesian Information Criterion (BIC)

To prevent overfitting and to ensure optimal model selection, we use the Bayesian Information Criterion to decide on model complexity. For every optimization call in qFit, we iteratively test increasing values of the threshold parameter $t_{d_{min}}$ and determine if the gain of information justifies the use of a more complex model. We fit iteratively, allowing the maximum number of conformers to vary from 1 up to 5 conformers ranked according to real-space correlation. For each iteration, we use our combined QP/MIQP routine to optimize the real-space residual sum of squares (RSS). We calculate the BIC for each level of complexity according to the following formula:

$$\mathrm{BIC} = n \ln(\mathrm{RSS}/n) + k \ln(n),$$

where n is the number of voxels in our resolution-dependent mask (see previous section for details) and $k = 4n_{active\ atoms}/t_{d_{min}}$ is the number of parameters in the model.

Each active atom has four parameters: x, y, z, and B-factor. Note that the occupancies are variables and not parameters. The factor $1/t_{d_{min}}$ is a proxy for model complexity and imposes a limit on the maximum number of conformations. We select the number of conformers that minimizes the BIC.

Parallelization

qFit 3 can be run individually for a single residue or ligand of interest, or in parallel across a whole protein using Python's *multiprocessing* module to spawn embarrassingly parallel subprocesses that run qFit across all residues in a target protein.

Validation metrics

For each residue/ligand modeled by qFit 3, we output several validation metrics, which include the BIC and the related Akaike information criterion $\mathrm{AIC} = 2k + n\ln(\mathrm{RSS})$ with n and k as above. qFit 3 also reports a confidence interval for the real-space cross-correlation of the proposed conformers. The confidence interval is calculated from the Fisher z-score of the real-space cross-correlation r [51]:

$$z = 0.5\ln((1+r)/(1-r))$$

Note that the z-score is approximately normally distributed with a standard deviation of $\sigma = (n-3)^{-\frac{1}{2}}$, where n is the number of voxels in our resolution-dependent mask around the set of conformers being assessed. qFit 3 reports the 95% confidence interval $z \pm 1.96\sigma$ for the cross-correlation. Overlapping intervals suggest that the gain in

cross-correlation is statistically not significant; we cannot reject the null hypothesis that the cross-correlations are the same at 95% confidence.

These auxiliary validation metrics are not used to filter results, but provide a guideline for balancing gain of information vs. model complexity.

Building an internally consistent structural model

In the procedure above, residues are modeled independently, i.e., without taking into account multiconformer models for neighboring residues. This leads to two modeling inconsistencies. First, consecutive residues may have different occupancies for each altloc, or even a different number of alternate conformations. Second, alternate conformers of (not necessarily consecutive) side chains in a spatial neighborhood can clash owing to inconsistent assignment of altloc identifiers. To resolve these two inconsistencies, we execute two routines: qFit-segment, which addresses the problem of inconsistency along the backbone, and qFit-relabel, which resolves clashing alternate conformers between neighboring residues by reassigning altloc labels.

The qFit-segment routine starts by identifying all segments along the backbone for which all residues have at least two backbone conformers. To mark the start and end points of such backbone segments, we identify residues for which either (a) a single conformer was output, or (b) where the backbone Cα and O atoms of that residue's conformers do not deviate by more than 0.05 Å. A segment is then delimited by these single-backbone-conformer residues. To create consistent segments, we proceed

iteratively. We break the segments in fragments of up to 4 residues (adjustable via the command line). We enumerate all possible combinations of conformers for the fragment, which at worst case leads to $5^4 = 625$ possible conformers. We use our optimization strategy (QP/MIQP iteratively, using the BIC) to select up to five conformers per fragment based on optimal fit to the experimental map (not based on covalent geometry). To ensure consistency with the PDB file format and compatibility with refinement software, we duplicate conformers for some residues within a fragment as needed to ensure that all consecutive residues have the same number of backbone conformers. Once all 4-residue fragments have been modeled in this fashion, we proceed to enumerate all possible combinations of such length-4 fragments. This leads to fragments of at most length 16, and, again, at worst case $5^4 = 625$ possible conformers. We continue to iterate in this fashion, enumerating all possible combinations and solving/modeling, until the segment is completed. The output of the qFit-segment routine is segments, each with up to five conformers, for which the backbone is consistent, i.e., for which all atoms for each conformer have the same label and occupancy.

Next, qFit-relabel relies on simulated annealing (SA) optimization of a Lennard-Jones potential to reassign altloc labels. We calculate the pairwise Lennard-Jones potential across every atom of all conformers output by qFit. Parameters for the Lennard-Jones calculation were taken from the Amber ff99SB forcefield [52]. The procedure selects five segments at random (a segment can include a single residue in this case) and randomly shuffles their labels. We then assess the change in the Lennard-Jones potential and

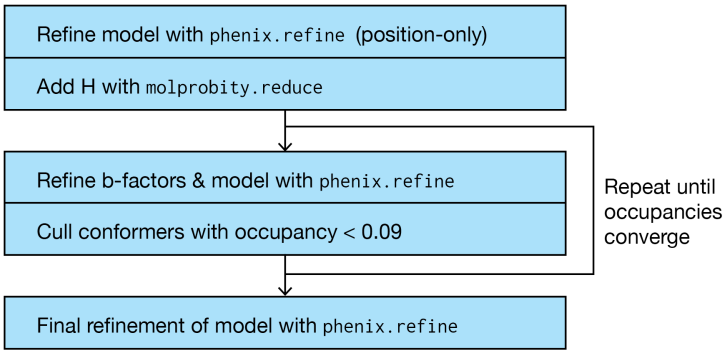
either accept or reject this move. The probability of accepting an unfavorable move is defined as:

$$P = \exp(-\Delta LJ/Temperature)$$

The temperature begins at 273 (arbitrary units), and is decreased by 10% every 10,000 perturbations. By default, 100,000 perturbations are sampled during relabeling. Benchmarking suggests that this value is sufficient for the scoring function to converge (data not shown). The output of the relabeling routine is a multiconformer model with up to five conformers per residue, in which backbones are consistent and in which alternate conformers for side chains are not clashing.

Final refinement script

We performed iterative refinement on the qFit multiconfomer models using version 1.18 of the Phenix software suite[33] to normalize the initially distorted covalent geometry, to ensure that the output models are properly fit into density (**Supplementary Figure 1.3)**, and to remove any unnecessary conformers.



Supplementary Figure 1.3| A flowchart for the recommended final refinement procedure. This was used for all structures modeled by qFit in this paper, and is contained in both qfit_final_refine_xray.sh and qfit_final_refine_cryoem.sh.

For X-ray crystallography structures, this iterative refinement protocol uses the *phenix.refine* executable (script name: qfit_final_refine_xray.sh). The initial round of refinement is done without hydrogens and uses the strategy=*individual_sites. We then (re-)add hydrogens to the model [53]. The next rounds of refinement use the following parameters: strategy=*individual_sites *individual_adp *occupancies, number_of_macro_cycles=5. At each iteration, we remove all conformers for which the occupancy fell below a cutoff of 0.09. This iterative cycle continues for as long as atoms are being removed due to this occupancy cutoff criterion. We then perform one last refinement round.

For cryo-EM structures, we use a similar refinement protocol as described above, but using *phenix.real_space_refine* [36] (script name: qfit_final_refine_cryoem.sh). All rounds of real-space refinement use the default parameters.

High-performance and cloud computing

qFit is capable of scaling from single laptops to large high-performance computing clusters. The following instructions enable qFit on Amazon's AWS, and should readily generalize to other cloud providers and RPM-based Linux distributions.

We describe configurations at two different scales: a single instance and an autoscaling cluster with a free master instance.

Launch an instance that will be used to execute qFit. AWS's c5.9xlarge instance has an appropriate number of cores and amount of memory for most proteins.

The following Bash script, reproduced from docs/aws_deploy.sh in the qFit repository, installs qFit and its dependencies within a conda environment:

```
#!/usr/bin/env bash

# Tested on Amazon Linux 2, but should work on most RPM-based Linux distros

# install Anaconda RPM GPG keys
sudo rpm --import https://repo.anaconda.com/pkgs/misc/gpgkeys/anaconda.asc

# add Anaconda repository
cat <<EOF | sudo tee /etc/yum.repos.d/conda.repo
[conda]
name=Conda
baseurl=https://repo.anaconda.com/pkgs/misc/rpmrepo/conda
enabled=1
gpgcheck=1
gpgkey=https://repo.anaconda.com/pkgs/misc/gpgkeys/anaconda.asc
EOF

sudo yum -y install conda
sudo yum -y install git gcc

source /opt/conda/etc/profile.d/conda.sh
```

```
conda create -y --name qfit
conda activate qfit

conda install -y -c anaconda mkl
conda install -y -c anaconda -c ibmdecisionoptimization cvxopt cplex

git clone https://github.com/ExcitedStates/qfit-3.0.git
cd qfit-3.0/

# Optionally, uncomment the following line to set a specific version of qFit
#git checkout v3.2.0
pip install .
```

Consider creating an image of the instance at this point to avoid executing the above script each time an instance is launched from a base instance.

After installation, it is necessary to execute source

/opt/conda/etc/profile.d/conda.sh

to set up conda within your Bash shell then activate the conda environment by executing

conda activate qfit.

Using the example described it qFit's README.md, alternative conformers for all residues in 3K0N can be calculated by executing

qfit_protein 3K0N.mtz -l 2FOFCWT,PH2FOFCWT 3K0N.pdb -p 36

for 3K0N.mtz and 3K0N.pdb in the current working directory, utilizing up to 36 cores.

Autoscaling cluster

Additionally, ParallelCluster can be used to create an autoscaling cluster to maximize efficiency of cloud resources.

Autoscaling cluster

Cluster creation and configuration

ParallelCluster is a suite of officially supported open-source tools used to create an autoscaling cluster on AWS.

After installation, pcluster configure provides a setup assistant to configure a cluster. A series of prompts guides the user through selection of region, scheduler, operating system, minimum and maximum size, master and compute instance type, and network configuration. These instructions assume selection of Slurm as the scheduler and Amazon Linux 2 as the operating system.

The following [cluster] section of the configuration file (saved at ~/.parallelcluster/config on Linux) represents reasonable settings:

```
[cluster default]
key_name = ###redacted###
base_os = alinux2
scheduler = slurm
master_instance_type = t2.micro
cluster_type = ondemand
compute_instance_type = c5.9xlarge
max_queue_size = 10
maintain_initial_size = false
vpc_settings = default
post_install = https://raw.githubusercontent.com/ExcitedStates/qfit-3.0/master/docs/aws_deploy.sh
```

This cluster will always run a t2.micro master instance, the first 750 hours per month of which are free, and a variable number of c5.9xlarge compute instances. While the scheduler's queue is empty and all jobs have finished, no compute instance will be running; when a job is submitted, a new compute instance will be launched so long as the total number would not exceed max_queue_size. New instances will download and execute the file at post_install URL, installing and configuring qFit.

For reduced costs in exchange for risk of job termination, cluster_type can be set to spot instead of ondemand. Spot pricing and risk of interruption are variable and depend on instance type, which should be considered when selecting compute instance type.

The cluster can be created with the command pcluster create default, accessed via SSH with pcluster ssh default and deleted with pcluster delete default.

References

1. van den Bedem H, Fraser JS (2015) Integrative, dynamic structural biology at atomic resolution--it's about time. Nat. Methods 12:307–318.

2. Aviram HY, Pirchi M, Mazal H, Barak Y, Riven I, Haran G (2018) Direct observation of ultrafast large-scale dynamics of an enzyme under turnover conditions. Proc. Natl. Acad. Sci. U. S. A. 115:3243–3248.

3. Lang PT, Ng H-L, Fraser JS, Corn JE, Echols N, Sales M, Holton JM, Alber T (2010) Automated electron-density sampling reveals widespread conformational polymorphism in proteins. Protein Sci. 19:1420–1431.

4. Fraser JS, van den Bedem H, Samelson AJ, Lang PT, Holton JM, Echols N, Alber T (2011) Accessing protein conformational ensembles using room-temperature X-ray crystallography. Proc. Natl. Acad. Sci. U. S. A. 108:16247–16252.

5. Keedy DA, Van Den Bedem H, Sivak DA, Petsko GA (2014) Crystal cryocooling distorts conformational heterogeneity in a model Michaelis complex of DHFR. Structure [Internet]. Available from: https://www.sciencedirect.com/science/article/pii/S0969212614001403

6. Keedy DA, Kenner LR, Warkentin M, Woldeyes RA, Hopkins JB, Thompson MC, Brewster AS, Van Benschoten AH, Baxter EL, Uervirojnangkoorn M, et al. (2015) Mapping the conformational landscape of a dynamic enzyme by multitemperature and XFEL crystallography. Elife [Internet] 4. Available from: http://dx.doi.org/10.7554/eLife.07574

7. Keedy DA, Hill ZB, Biel JT, Kang E, Rettenmaier TJ (2018) An expanded allosteric network in PTP1B by multitemperature crystallography, fragment screening, and covalent tethering. Elife [Internet]. Available from: https://elifesciences.org/articles/36307

8. Doukov T, Herschlag D, Yabukarski F (2020) A Robust Method for Collecting X-ray Diffraction Data from Protein Crystals across Physiological Temperatures. bioRxiv [Internet]:2020.03.17.995852. Available from: https://www.biorxiv.org/content/10.1101/2020.03.17.995852v1

9. Fraser JS, Clarkson MW, Degnan SC, Erion R, Kern D, Alber T (2009) Hidden alternative structures of proline isomerase essential for catalysis. Nature 462:669–673.

10. Pearce NM, Bradley AR, Krojer T, Marsden BD, Deane CM, von Delft F (2017) Partial-occupancy binders identified by the Pan-Dataset Density Analysis method offer new chemical opportunities and reveal cryptic binding sites. Struct Dyn 4:032104.

11. Pearce NM, Krojer T, Bradley AR, Collins P, Nowak RP, Talon R, Marsden BD, Kelm S, Shi J, Deane CM, et al. (2017) A multi-crystal method for extracting obscured crystallographic states from conventionally uninterpretable electron density. Nat. Commun. 8:15123.

12. Tenboer J, Basu S, Zatsepin N, Pande K, Milathianaki D, Frank M, Hunter M, Boutet S, Williams GJ, Koglin JE, et al. (2014) Time-resolved serial crystallography captures high-resolution intermediates of photoactive yellow protein. Science 346:1242–1246.

13. Hekstra DR, White KI, Socolich MA, Henning RW, Šrajer V, Ranganathan R (2016) Electric-field-stimulated protein mechanics. Nature 540:400–405.

14. Thompson MC, Barad BA, Wolff AM, Sun Cho H, Schotte F, Schwarz DMC, Anfinrud P, Fraser JS (2019) Temperature-jump solution X-ray scattering reveals distinct motions in a dynamic enzyme. Nat. Chem. 11:1058–1066.

15. Dasgupta M, Budday D, de Oliveira SHP, Madzelan P, Marchany-Rivera D, Seravalli J, Hayes B, Sierra RG, Boutet S, Hunter MS, et al. (2019) Mix-and-inject XFEL crystallography reveals gated conformational dynamics during enzyme catalysis. Proc. Natl. Acad. Sci. U. S. A. 116:25634–25640.

16. Ebrahim A, Moreno-Chicano T, Appleby MV, Chaplin AK, Beale JH, Sherrell DA, Duyvesteyn HME, Owada S, Tono K, Sugimoto H, et al. (2019) Dose-resolved serial synchrotron and XFEL structures of radiation-sensitive metalloproteins. IUCrJ 6:543–551.

17. Ebrahim A, Appleby MV, Axford D, Beale J, Moreno-Chicano T, Sherrell DA, Strange RW, Hough MA, Owen RL (2019) Resolving polymorphs and radiation-driven effects in microcrystals using fixed-target serial synchrotron crystallography. Acta Crystallogr D Struct Biol 75:151–159.

18. Casanas A, Warshamanage R, Finke AD, Panepucci E, Olieric V, Nöll A, Tampé R, Brandstetter S, Förster A, Mueller M, et al. (2016) EIGER detector: application in macromolecular crystallography. Acta Crystallogr D Struct Biol 72:1036–1048.

19. Callaway E (2020) Revolutionary cryo-EM is taking over structural biology. Nature

578:201.

20. Yip KM, Fischer N, Paknia E, Chari A, Stark H (2020) Breaking the next Cryo-EM resolution barrier – Atomic resolution determination of proteins! bioRxiv [Internet]:2020.05.21.106740. Available from: https://www.biorxiv.org/content/10.1101/2020.05.21.106740v1

21. Chirgadze DY, Murshudov G, Aricescu AR, Scheres S (2020) Single-particle cryo-EM at atomic resolution. BioRxiv [Internet]. Available from: https://www.biorxiv.org/content/10.1101/2020.05.22.110189v1.abstract

22. Li P-N, de Oliveira SHP, Wakatsuki S, van den Bedem H (2020) Sequence-guided protein structure determination using graph convolutional and recurrent networks. arXiv [cs.LG] [Internet]. Available from: http://arxiv.org/abs/2007.06847

23. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) The Protein Data Bank. Nucleic Acids Res. 28:235–242.

24. Merritt EA (1999) Expanding the model: anisotropic displacement parameters in protein structure refinement. Acta Crystallogr. D Biol. Crystallogr. 55:1109–1117.

25. Lovell SC, Word JM, Richardson JS, Richardson DC (2000) The penultimate rotamer library. Proteins 40:389–408.

26. Burnley BT, Afonine PV, Adams PD, Gros P (2012) Modelling dynamics in protein crystal structures by ensemble refinement. Elife 1:e00311.

27. Babcock NS, Keedy DA, Fraser JS, Sivak DA (2018) Model selection for biological

crystallography. bioRxiv [Internet]. Available from: https://www.biorxiv.org/content/10.1101/448795v1.abstract

28. van den Bedem H, Dhanik A, Latombe JC, Deacon AM (2009) Modeling discrete heterogeneity in X-ray diffraction data by fitting multi-conformers. Acta Crystallogr. D Biol. Crystallogr. 65:1107–1117.

29. Brock JS, Hamberg M, Balagunaseelan N, Goodman M, Morgenstern R, Strandback E, Samuelsson B, Rinaldo-Matthis A, Haeggström JZ (2016) A dynamic Asp-Arg interaction is essential for catalysis in microsomal prostaglandin E2 synthase. Proc. Natl. Acad. Sci. U. S. A. 113:972–977.

30. Russi S, González A, Kenner LR, Keedy DA, Fraser JS, van den Bedem H (2017) Conformational variation of proteins at room temperature is not dominated by radiation damage. J. Synchrotron Radiat. 24:73–82.

31. Atakisi H, Moreau DW, Thorne RE (2018) Effects of protein-crystal hydration and temperature on side-chain conformational heterogeneity in monoclinic lysozyme crystals. Acta Crystallogr D Struct Biol 74:264–278.

32. van Zundert GCP, Hudson BM, de Oliveira SHP, Keedy DA, Fonseca R, Heliou A, Suresh P, Borrelli K, Day T, Fraser JS, et al. (2018) qFit-ligand Reveals Widespread Conformational Heterogeneity of Drug-Like Molecules in X-Ray Electron Density Maps. J. Med. Chem. 61:11183–11198.

33. Liebschner D, Afonine PV, Baker ML, Bunkóczi G, Chen VB, Croll TI, Hintze B, Hung LW, Jain S, McCoy AJ, et al. (2019) Macromolecular structure determination

using X-rays, neutrons and electrons: recent developments in Phenix. Acta Crystallogr D Struct Biol 75:861–877.

34. Barr AJ, Ugochukwu E, Lee WH, King ONF, Filippakopoulos P, Alfano I, Savitsky P, Burgess-Brown NA, Müller S, Knapp S (2009) Large-scale structural analysis of the classical human protein tyrosine phosphatome. Cell 136:352–363.

35. Keedy DA, Fraser JS (2015) Exposing hidden alternative backbone conformations in X-ray crystallography using qFit. PLoS Comput. Biol. [Internet]. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4624436/

36. Afonine PV, Poon BK, Read RJ, Sobolev OV, Terwilliger TC, Urzhumtsev A, Adams PD (2018) Real-space refinement in PHENIX for cryo-EM and crystallography. Acta Crystallogr D Struct Biol 74:531–544.

37. Su H, Zou Y, Chen G, Dou H, Xie H, Yuan X, Zhang X, Zhang N, Li M, Xu Y (2020) Exploration of Fragment Binding Poses Leading to Efficient Discovery of Highly Potent and Orally Effective Inhibitors of FABP4 for Anti-inflammation. J. Med. Chem. 63:4090–4106.

38. Bank RPD 4EK8. Available from: https://www.rcsb.org/structure/4ek8

39. Stsiapanava A, Olsson U, Wan M, Kleinschmidt T, Rutishauser D, Zubarev RA, Samuelsson B, Rinaldo-Matthis A, Haeggström JZ (2014) Binding of Pro-Gly-Pro at the active site of leukotriene A4 hydrolase/aminopeptidase and development of an epoxide hydrolase selective inhibitor. Proc. Natl. Acad. Sci. U. S. A. 111:4227–4232.

40. Davis IW, Arendall WB 3rd, Richardson DC, Richardson JS (2006) The backrub

motion: how protein backbone shrugs when a sidechain dances. Structure 14:265–274.

41. Smith CA, Kortemme T (2008) Backrub-like backbone simulation recapitulates natural protein conformational variability and improves mutant side-chain prediction. J. Mol. Biol. 380:742–756.

42. Hallen MA, Keedy DA (2013) Dead-end elimination with perturbations (DEEPer): A provable protein design algorithm with continuous sidechain and backbone flexibility. : Structure, Function, and … [Internet]. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/prot.24150

43. van den Bedem H, Lotan I, Latombe JC (2005) Real-space protein-model completion: an inverse-kinematics approach. Section D: Biological … [Internet]. Available from: https://scripts.iucr.org/cgi-bin/paper?wd5022

44. Pearce NM, Krojer T, von Delft F (2017) Proper modelling of ligand binding requires an ensemble of bound and unbound states. Acta Crystallogr D Struct Biol 73:256–266.

45. van Zundert GCP, Moriarty NW, Sobolev OV, Adams PD, Borrelli KW (2020) Macromolecular refinement of X-ray and cryo-electron microscopy structures with Phenix / OPLS3e for improved structure and ligand quality. bioRxiv [Internet]:2020.07.10.198093. Available from: https://www.biorxiv.org/content/10.1101/2020.07.10.198093v2

46. Jiang L, Kuhlman B, Kortemme T, Baker D (2005) A “solvated rotamer” approach to modeling water-mediated hydrogen bonds at protein--protein interfaces. Proteins: Struct. Funct. Bioinf. 58:893–904.

47. Liebschner D, Afonine PV, Moriarty NW, Poon BK, Sobolev OV, Terwilliger TC, Adams PD (2017) Polder maps: improving OMIT maps by excluding bulk solvent. Acta Crystallogr D Struct Biol 73:148–157.

48. Holton JM, Classen S, Frankel KA, Tainer JA (2014) The R-factor gap in macromolecular crystallography: an untapped potential for insights on accurate structures. FEBS J. 281:4046–4060.

49. Terwilliger TC, Grosse-Kunstleve RW, Afonine PV, Moriarty NW, Adams PD, Read RJ, Zwart PH, Hung L-W (2008) Iterative-build OMIT maps: map improvement by iterative model building and refinement without model bias. Acta Crystallogr. D Biol. Crystallogr. 64:515–524.

50. Lang PT, Holton JM, Fraser JS, Alber T (2014) Protein structural ensembles are revealed by redefining X-ray electron density noise. Proc. Natl. Acad. Sci. U. S. A. 111:237–242.

51. Volkmann N (2009) Confidence intervals for fitting of atomic models into low-resolution densities. Acta Crystallogr. D Biol. Crystallogr. 65:679–689.

52. Hornak V, Abel R, Okur A, Strockbine B, Roitberg A, Simmerling C (2006) Comparison of multiple AMBER force fields and development of improved protein backbone parameters. Proteins: Struct. Funct. Bioinf. 65:712–725.

53. Word JM, Lovell SC, Richardson JS, Richardson DC (1999) Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation. J. Mol. Biol. 285:1735–1747.

Chapter II

Ligand binding remodels protein side chain conformational heterogeneity

Stephanie A. Wankowicz[1,2], Saulo de Oliveira[3], Daniel Hogan[1], Henry van den Bedem[1,3], James S. Fraser[1,*]

1) Department of Bioengineering and Therapeutic Sciences, University of California, San Francisco, San Francisco, CA 94158, USA.
2) Biophysics Graduate Program, University of California San Francisco, San Francisco, CA 94158, USA.
3) Atomwise Inc. San Francisco, CA 94103 USA

Preface

The bulk of this chapter appears as Wankowicz *et al.* preprinted in *bioRxiv* in 2022, and a version of which was ultimately published in *eLife* later the same year.

Abstract

While protein conformational heterogeneity plays an important role in many aspects of biological function, including ligand binding, its impact has been difficult to quantify. Macromolecular X-ray diffraction is commonly interpreted with a static structure, but it can provide information on both the anharmonic and harmonic contributions to conformational heterogeneity. Here, through multiconformer modeling of time- and space-averaged electron density, we measure conformational heterogeneity of 743 stringently matched pairs of crystallographic datasets that reflect unbound/apo and ligand-bound/holo states. When comparing the conformational heterogeneity of side chains, we observe that when binding site residues become more rigid upon ligand binding, distant residues tend to become more flexible, especially in non-solvent exposed regions. Among ligand properties, we observe increased protein flexibility as the number of hydrogen bonds decrease and relative hydrophobicity increases. Across a series of 13 inhibitor bound structures of CDK2, we find that conformational heterogeneity is correlated with inhibitor features and identify how conformational changes propagate differences in conformational heterogeneity away from the binding site. Collectively, our findings agree with models emerging from NMR studies suggesting that residual side chain entropy can modulate affinity and point to the need to integrate both static conformational changes and conformational heterogeneity in models of ligand binding.

Introduction

Ligand binding is essential for many protein functions, including enzyme catalysis, receptor activation, and drug response [1]. Ligand binding reshapes the protein conformational ensemble between the ligand-bound (holo) and unbound (apo) states, stabilizing some conformations and destabilizing others [2]. Despite the dynamic nature of proteins, when comparing structures, often only static conformational changes are considered. However, differences due to ligand binding can range from large, collective movements, such as a loop closure over the binding pocket, to small, local fluctuations of side chains [3]. Differences in binding affinity and specificity are most often attributed to the enthalpic portion of binding free energy, including visualized interactions between the receptor and ligand. On the other hand, conformational heterogeneity, especially side chain fluctuations, can also contribute energetically to the binding affinity by modulating entropy [4,5]. While the individual fluctuation of residues are small, they can add up to significantly contribute to the entropic portion of binding free energy. Previous work examining a diverse set of protein complexes calculated that protein conformational entropy can contribute between -2 (favoring) and 4 (disfavoring) kcal/mol to binding free energy [6,7]. A holistic understanding of the origins of binding would ideally explore both enthalpic and entropic energetic contributions to binding affinity [8].

Side chain conformational heterogeneity, including jumps between and variation within rotameric conformations, measured by Nuclear Magnetic Resonance (NMR) relaxation studies has been linked to entropy [6,9]. In principle, complementary information could be accessed by other structural methods. Structural information from X-ray crystallography or Cryo-electron microscopy (CryoEM), typically produces a single set of structural

coordinates. However, the underlying density maps are created from thousands-to-millions of protein molecules, and averaged in both time and space through the crystal lattice or electron microscope particle stack [10,11]. When averaged in a single density map, conformational heterogeneity across these copies can manifest as “anharmonic disorder”, which can be modeled using multiple alternative conformations, or “harmonic disorder”, which can be modeled by B-factors/atomic displacement parameters (Figure 2.1). Molecular dynamics experiments have demonstrated that if alternative conformations are not modeled correctly and consistently, then B-factors take on values that are not representative of the underlying conformational heterogeneity [12,13]. Moreover, B-factors incorporate many effects, including the biases and restraints of the refinement programs, modeling errors, crystal lattice defects, and occupancy changes of atoms. Therefore, consistently modeling X-ray structures as multiconformer models, with alternative side chain and backbone conformations, along with B-factors, may better complement the view emerging from NMR and improve our understanding of the energetics of binding [14].

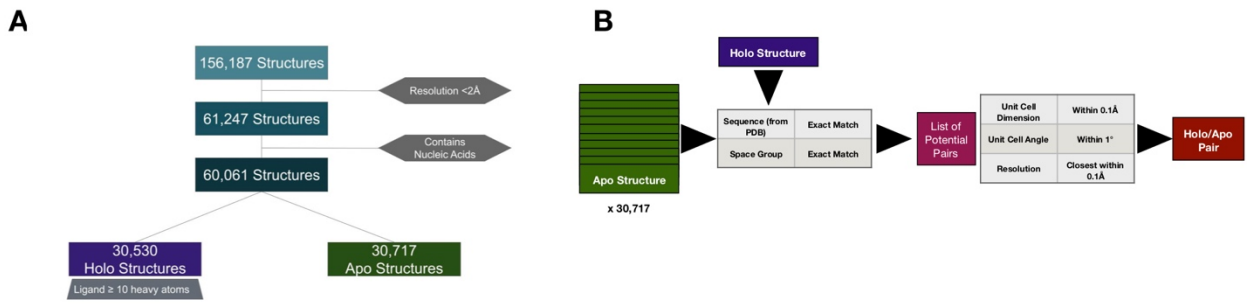
Here, we examine how protein side chain conformational heterogeneity changes upon ligand binding by assembling a large, high-quality dataset of matched holo and apo X-ray crystallography structures. To integrate both harmonic and anharmonic disorder, we use a consistent multiconformer modeling procedure, qFit [15] and crystallographic order parameters [16]. We test the hypothesis that ligand binding narrows the conformational ensemble, resulting in a decrease in heterogeneity of side chains in the holo structure compared with the apo structure. Our analysis reveals complex patterns of

conformational heterogeneity that vary between and within proteins upon ligand binding. Specifically, in proteins where binding site residues become more rigid upon ligand binding, distant residues tend to become less rigid. This observation suggests that both natural and artificial ligands can modulate the natural composition of the protein conformational heterogeneity across the entire receptor to modulate the free energy of binding.

Results

Assembling the dataset

To assess the differences in conformational heterogeneity upon ligand binding, we identified high quality, high resolution (2Å resolution or better) X-ray crystallography datasets from the PDB [17]. We classified structures as holo if they had a ligand with 10 or more heavy atoms, excluding common crystallographic additives (**Supplementary Figure 2.1**). Structures without ligands, excluding common crystallographic additives, were classified as apo (**Supplementary Figure 2.1**). We identified holo/apo matched pairs by requiring the same sequence and near-isomorphous crystallographic parameters. Furthermore, we required the resolution difference between holo and apo pairs to be 0.1 Å or less, selecting representative apo structures to minimize the difference in resolution (**Supplementary Figure 2.1**). This stringently matched ligand holo-apo dataset contained 1,205 pairs. We also used identical selection criteria to create a control dataset of 293 apo-apo pairs, taken from the set of holo/apo pairs.



Supplementary Figure 2.1| The dataset selection process. (A) To select holo/apo matched pairs, we first categorized the PDB structures into holo or apo structures, removing structures with a resolution worse than 2Å, not resolved using X-ray crystallography, and those that include nucleic acids. Holo structures (n=30,530) were required to have a ligand, not including common crystallographic additives, with 10 or more heavy atoms. All others were classified as apo (n=30,171). (B) For every holo structure, we compared it to the 30,717 apo structures first matching for exact sequence and space group and controlling for similar unit cell dimensions (within 0.1 Å) and angles (within 1 degree). Finally, we selected the structures paired for resolution within 0.1 Å.

Re-refining and qFit modeling of apo/holo pairs

To minimize biases resulting from different model refinement protocols, we re-refined all structures using the deposited structure factors and *phenix.refine* [18]. The majority of structures in our re-refined dataset had less than 2% of residues modeled with alternative conformations, likely reflecting undermodeling of conformational heterogeneity represented in the PDB, based on prior literature [19]**.** To more consistently assess conformational heterogeneity, we rebuilt all structures using qFit, an automated multiconformer modeling algorithm (Keedy et al., 2015; Riley et al., 2021) with subsequent refinement using *phenix.refine* [18]. While qFit has biases, running all models through a consistent protocol will avoid manual biases that could creep into the holo or apo structures specifically. Additionally, by re-building each model as a multiconformer model, we were able to better distinguish the contributions of harmonic and anharmonic conformational heterogeneity across the structure (Figure 2.1). All models went through

additional quality control, removing structures that resulted in large increases in R-free at each refinement step, high clashscores, or large RMSD between the pairs (**Methods, Supplementary Figure 2.2**). This procedure resulted in 743 pairs. Due to apo datasets serving as the reference state for multiple ligand bound structures, our dataset consists of 743 unique holo structures and 432 unique apo structures.

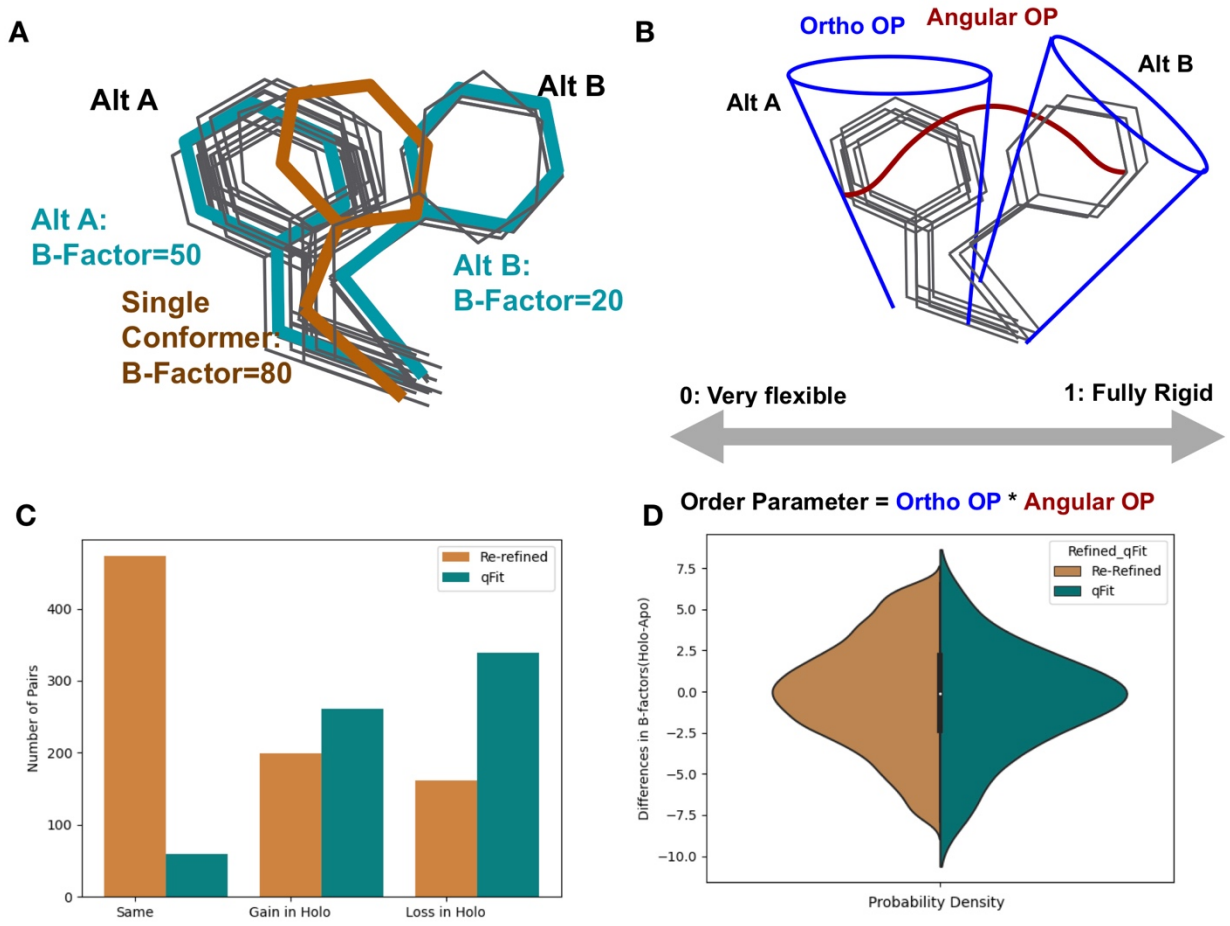
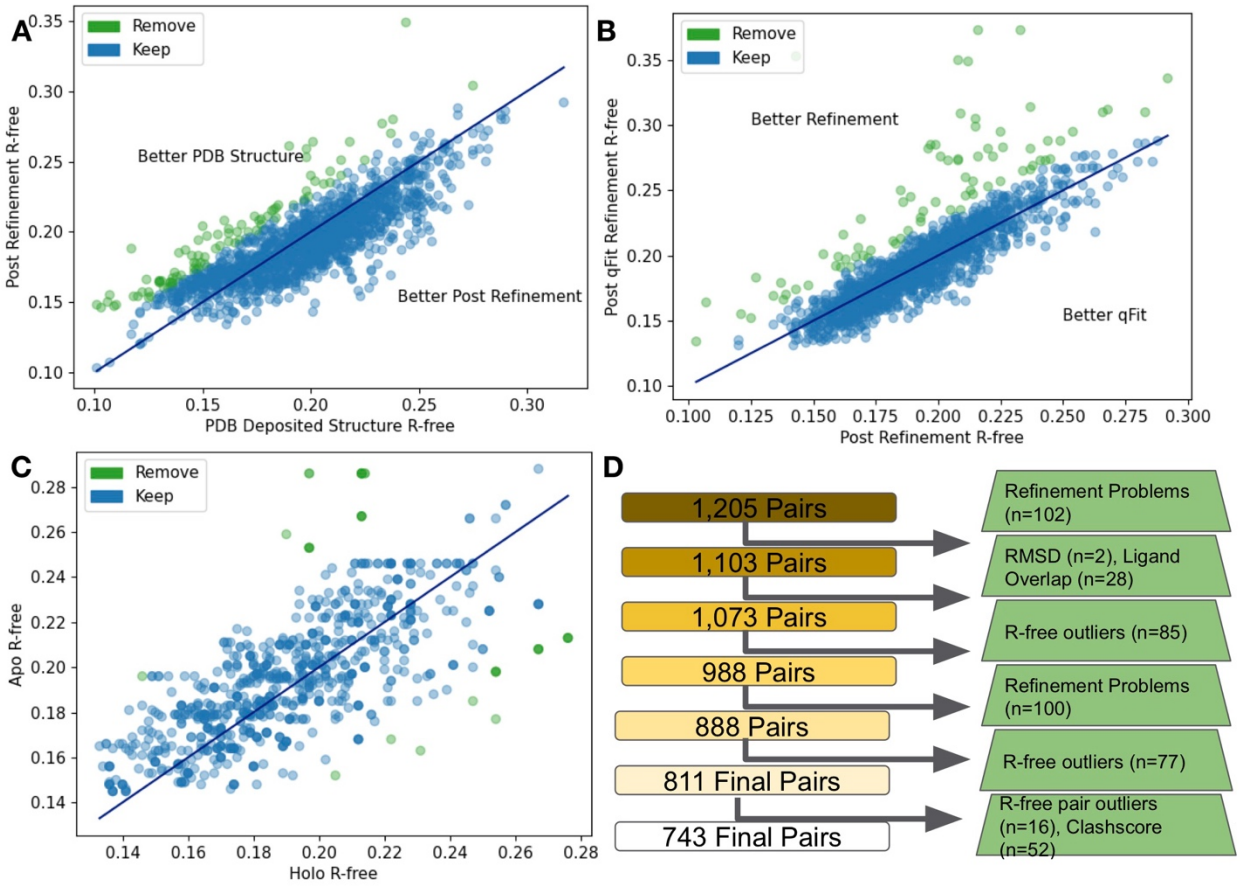


Figure 2.1| Representing structural data as multiconformer models. (A) The grey outlines represent snapshots of the true underlying ensemble of the phenylalanine residue. The orange stick represents the residue modeled as a single conformer. The teal sticks represent the residue modeled as alternative conformers. The single conformer accounts for all heterogeneity in the B-factor, increasing the B-factor and reducing our ability to determine harmonic versus anharmonic motion. When a residue is modeled using alternative conformers, this heterogeneity is divided between harmonic heterogeneity, captured by the B-factors of each alternative conformation and the anharmonic heterogeneity, captured by spread in coordinates between the alternative conformations. (B) To quantify the conformational heterogeneity of each

residue, we used multi-conformer order parameters [16], which are the products of the ortho order parameter, which captures the harmonic or B-factor portion of each conformation and the angular order parameter, which captures the anharmonic portion or the displacement between alternative conformers. These are multiplied to produce the final order parameter (Methods). (C) The change in the number of alternative conformers (holo-apo) in binding site residues. In the re-refined dataset (orange), the majority structures have the same number of alternative conformers in the binding site, with the second most popular category gaining alternative conformers in the holo structure. In the qfit dataset (teal), the majority of structures lose an alternative conformer in the holo structure, with the second most common category being gaining an alternative conformer. (D) The differences in B-factors (holo-apo) in the re-refined (orange) and qFit (teal) datasets. Overall, there was no significant difference in B-factors between holo and apo structures in both the re-refined and qFit datasets.

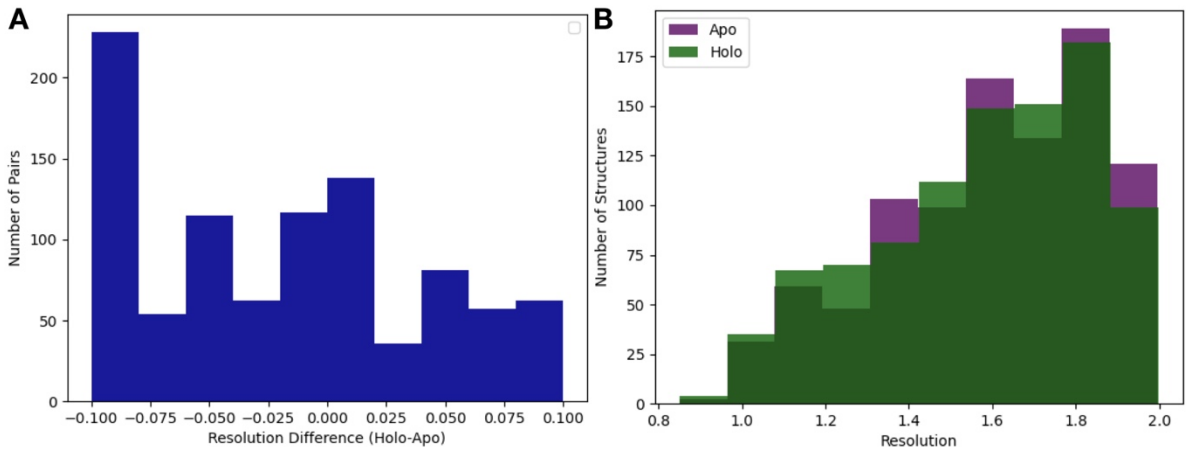


Supplementary Figure 2.2| Quality control of multiconformer models. (A) The differences in R-free values between the PDB deposited structures and after re-refinement. 85 structures were removed (green) as their R-free increased by more than 2.5%. (B) The difference in R-free statistics between the re-refined structures and the qFit structures. 77 structures were removed (green) as their R-free increased by more than 2.5%. (C) The difference in R-free statistics in qFit structures between the holo and apo structure. 16 pairs were removed (green) as ther R-free statistics differed by 5% or more between the pairs. (D) Flowchart representing our quality control process, with removed structures in green boxes.

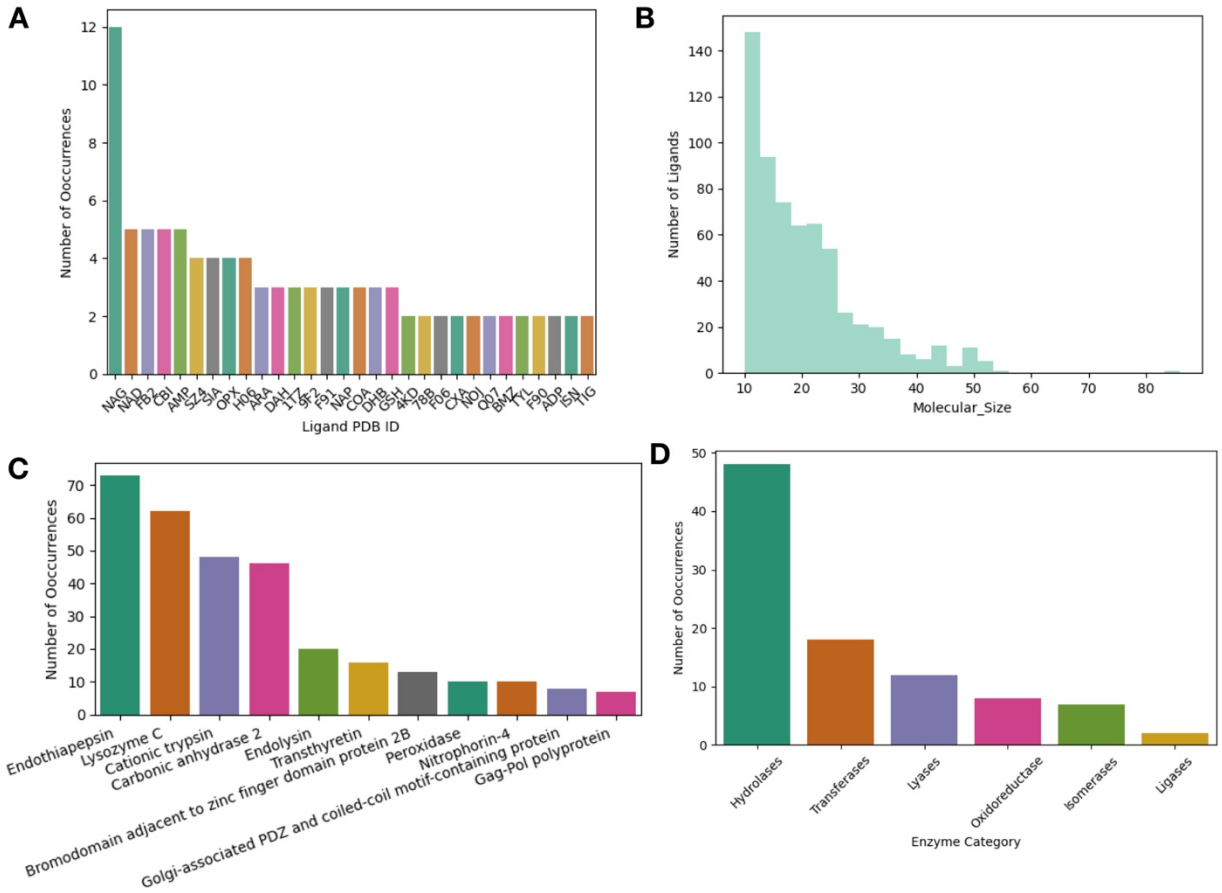
Properties of the apo/holo pairs

The median resolution across our dataset was 1.6Å with a small trend towards improved (higher) resolution in the apo structure (0.01Å, median improvement (holo-apo); $p=3.8x10^{-20}$, Wilcoxon signed rank test; **Supplementary Figure 2.3**). Across the dataset, 546 unique ligands were present in the structures, with 134 of these (e.g. NAG,

AMP, etc) appearing in multiple structures (**Supplementary Figure 2.4**). The median number of ligand heavy atoms was 19, with only 10 very large ligands (>50 heavy atoms, e.g. Atazanavir; **Supplementary Figure 2.4**). The proteins in the dataset represent 315 unique Uniprot IDs, with a bias towards enzymes that have been used for model systems for structural biology, including: Endothiopepsin (n=73 pairs), Lysozyme (n=62 pairs), Trypsin (n=48 pairs), and Carbonic Anhydrase 2 (n=46 pairs; **Supplementary Figure 2.4**).



Supplementary Figure 2.3| Resolution difference in apo/holo pairs. (A) Resolution difference between pairs (holo-apo). The median pairwise difference was 0.01Å, with slightly better resolution in the apo structures, and the standard deviation was 0.06Å. (B) The distribution of resolution (median=1.6Å) of the apo (n=432) and holo (n=743) dataset. The median apo resolution was 1.58Å and the median holo resolution was 1.58Å.



Supplementary Figure 2.4| Ligand statistics of holo structures. (A)The top 30 ligands in our dataset by PDB chemical ID. NAG (2-acetamido-2-deoxy-beta-D-glucopyranose) and H06 ((E)-4-((2-nicotinoylhydrazono)methyl) benzimidamide) where the most frequent ligands in our dataset. (B) The distribution of the number of heavy atoms of a ligand of interest. The median number of heavy atoms was 19. There were only 10 very large ligands (>50 heavy atoms, e.g. Atazanavir). (C) The most common proteins in our dataset. Eleven proteins in our dataset were included in 6 or more pairs. This included our most common proteins including: Endothiopepsin (n=73 pairs), Lysozyme (n=62 pairs), Trypsin (n=48 pairs), and Carbonic Anhydrase 2 (n=46 pairs). (D) The distribution of enzymes (n=95) based on their Enzyme Commission Number.

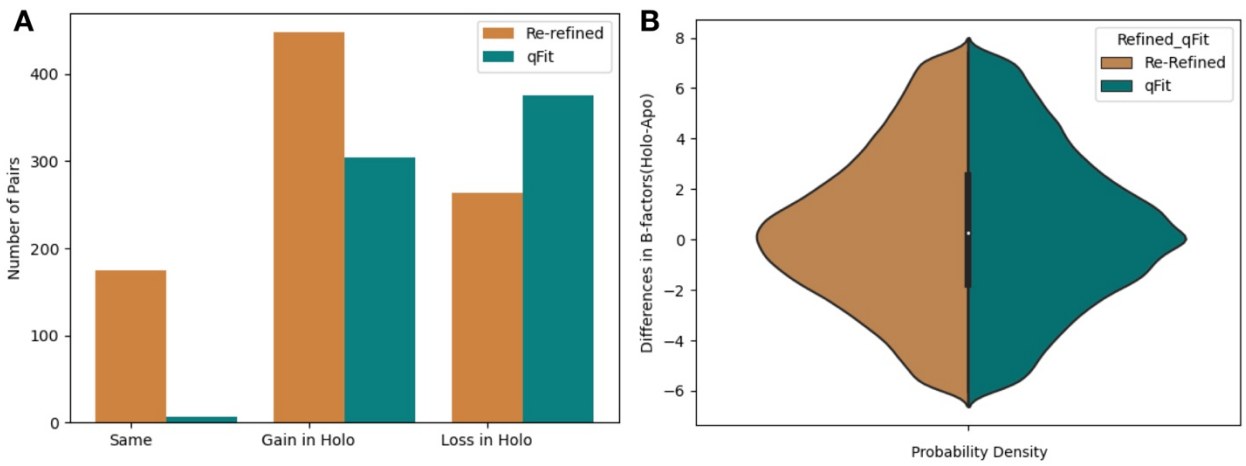
Conformational Heterogeneity across the Re-refined and qFit dataset

To determine the differences in conformational heterogeneity upon ligand binding in both the re-refined and qFit models, we assessed four commonly used metrics: the number of alternative conformers, B-factors (atomic displacement parameter), root-mean-square fluctuations (RMSF), and rotamer changes.

Number of Alternative Conformations

Alternative conformations were modeled at low frequency in the re-refined dataset compared to the qFit modeled structures (1.7% vs. 47.8% of residues). In the re-refined dataset, there is a bias to increased modeling of alternative conformations in the holo dataset (50.5% gain vs. 29.8% loss), whereas more even representation was observed in the qFit dataset (44.3% gain vs. 54.8% loss; **Supplementary Figure 2.5**). These results suggest that the trend of increased side chain conformational heterogeneity in PDB deposited structures may have its origin in human bias with more careful human attention to careful model building of binding site residues in holo structures.

We next focused our analysis on binding site residues, defined as any residue with a heavy atom within 5Å of any ligand heavy atom. In the re-refined dataset, 23.9% of the matched pairs had a gain in alternative conformations in the holo model compared to only 19.3% losing an alternative conformer in the holo model, suggesting, counter-intuitively, that ligand binding increases local side chain mobility (**Figure 2.1**). However, in the qFit dataset, holo models tend to lose alternative conformations in the binding site residues (39.7% gain vs. 51.5% loss; **Figure 2.1**).

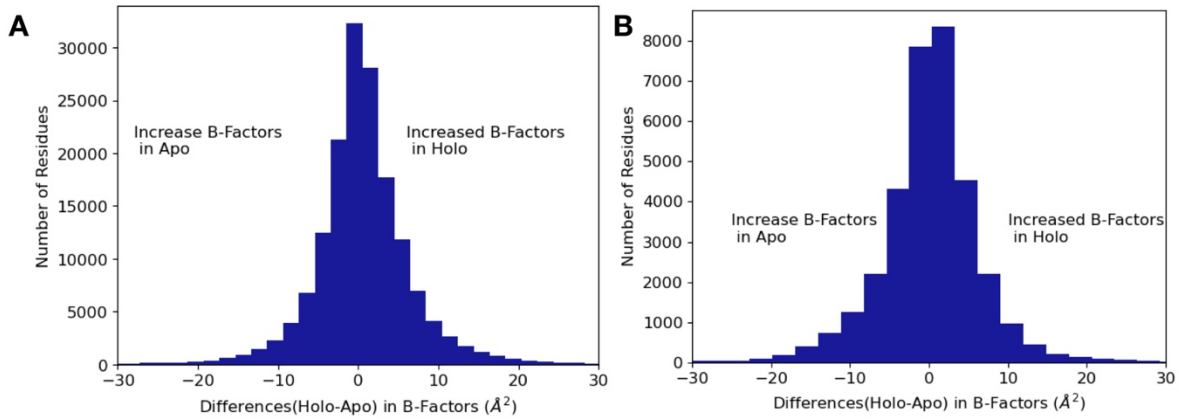


Supplementary Figure 2.5| Alternative Conformers and B-factors. (A) The change in the number of alternative conformers (holo-apo) across all residues. In the re-refined dataset (orange), the majority models have a gain of the number of alternative conformers in the holo, with the second most common category being a loss of alternative conformers. In the qfit dataset (teal), the majority of structures lose an alternative conformer in the holo model, with the second most common category being gaining an alternative conformer. (B) The difference in B-factors across all residues. There was a slight increase in B-factors in holo models in both the re-refined and the qFit datasets.

B-factors

Next we explored the harmonic contribution to conformational heterogeneity as modeled by B-factors on a pairwise, residue by residue basis. Across all residues in the re-refined dataset, B-factors were slightly higher in holo models (0.31$Å^2$, median difference (holo-apo); p=4.4x10^{-208}, Wilcoxon signed rank test; **Supplementary Figure 2.5**). In the qFit dataset, similar to the re-refined structures, holo residues had slightly higher B-factors (0.34$Å^2$, median difference (holo-apo); p=5.6x10^{-264}, Wilcoxon signed rank test; **Supplementary Figure 2.6**). Of note, the B-factors in the qFit dataset are slightly smaller than the re-refined dataset (13.41$Å^2$ vs. 13.94$Å^2$, average B-factors) reflecting the tendency for alternative conformation effects to be modeled as increased B-factors. When examining the binding site residues, there was no significant difference

in B-factors between the holo and apo models in both the re-refined (0.01Å^2, median difference in B-factors; p=0.34, Wilcoxon signed rank test; **Figure 1D**) and qFit datasets (0.06Å^2, median difference in B-factors; p=0.7, Wilcoxon signed rank test; **Figure 1D, Figure 1- Figure Supplement 6B**). The lack of change in B-factors close to ligands between the holo and apo models indicate that changes between the holo and apo B-factors are driven by signals distant from the binding site.



Supplementary Figure 2.6| B-factor differences between apo and holo. (A) The difference in B-factors between holo and apo pairs. The range of the difference in B-factors was -199.8Å^2 to 197.0Å^2, here we remove the most 10% extreme values, which are due to poor density in loop regions leading to high B-factors for those individual residues. Across all residues, on average B-factors were higher in holo structures compared to apo (0.34Å^2, median difference (holo-apo); p=4.4x10^{-208}, Wilcoxon signed rank test). (B) In binding site residues, B-factors were on average the same between holo and apo residues (0.06Å^2, median difference in B-factors; p=0.7, Wilcoxon signed rank test).

Conformational differences incorporating alternative conformations

Because of the low number of alternative conformers in the re-refined dataset, we only explored the anharmonic differences for side chains between the holo and apo models in the qFit dataset. First, to determine the extent of conformational change of alternative conformations, we compared the rotameric distribution of side chains. Side chain rotamer changes between apo and holo structures have been reported to be very

prevalent in single conformer models, with 90% of binding sites having at least one residue changing rotamers upon ligand binding [20]. To accommodate multiconformer models, we assigned all conformations to distinct rotamers using *phenix.rotalyze*. We classified each residue as having “no change” in rotamers if the set of rotamer assignments matched for the holo and apo residue. In binding sites, we also observed that “no change” was the most common outcome for residues (78.6%; **Figure 2.2**). In the second largest category, “distinct”, the holo and apo residue shared no rotamer assignments (15.5% of residues; **Figure 2.2**).

A more complicated situation occurs when some, but not all, of the rotamer assignments are shared across apo and holo residue. We classified 2.6% of residues as “remodeled - holo loss” (**Figure 2.2**) if distinct, additional rotameric conformations were populated in the apo residue only and 3.8% of residues as “remodeled - holo gain” (**Figure 2.2**) if distinct, additional rotameric conformations were populated in the holo residue only. These results suggest a counterintuitive interpretation of binding site residues increasing their conformational heterogeneity upon ligand binding. However, a major potential confounder is that holo structures reflect an ensemble average of two compositional states (apo and holo) with alternative conformations representing the apo state at reduced occupancy, which we examined by subsetting the ligands based on relative B-factors (see below). A potential for a third category of remodeling, where both apo and holo residues share at least one conformation and each have at least one additional conformation, did not occur in our dataset.

Across apo-holo pairs there was a large range of the percentages of binding site residues with the same rotamer classification in the pairs (23.2% to 100.0%), indicating that side chain remodeling can be quite variable (**Figure 2.2**). We found 11% of binding sites had all residues classified as “same” between pairs, consistent with a previous study that used single conformer models [20]. As an example of such a “pre-organized” binding site is Galectin-3 bound to thiodigalactoside (PDB: 5NFC, 4JC1; **Figure 2.2**). In contrast, 67% of binding site residues have a rotamer status difference in transthyretin (PDB: 1CZR, 3CFN; **Figure 2.2**), including a rotamer change in Leu101 to avoid a clash with the ligand.

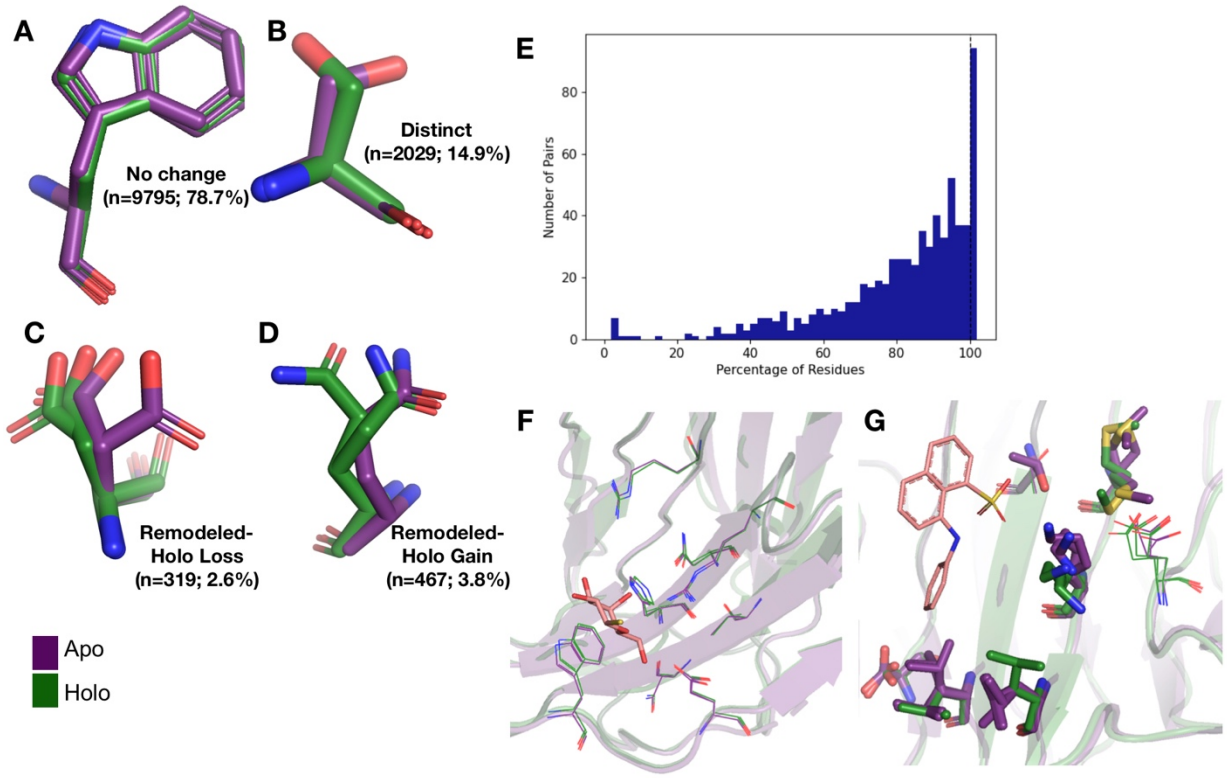
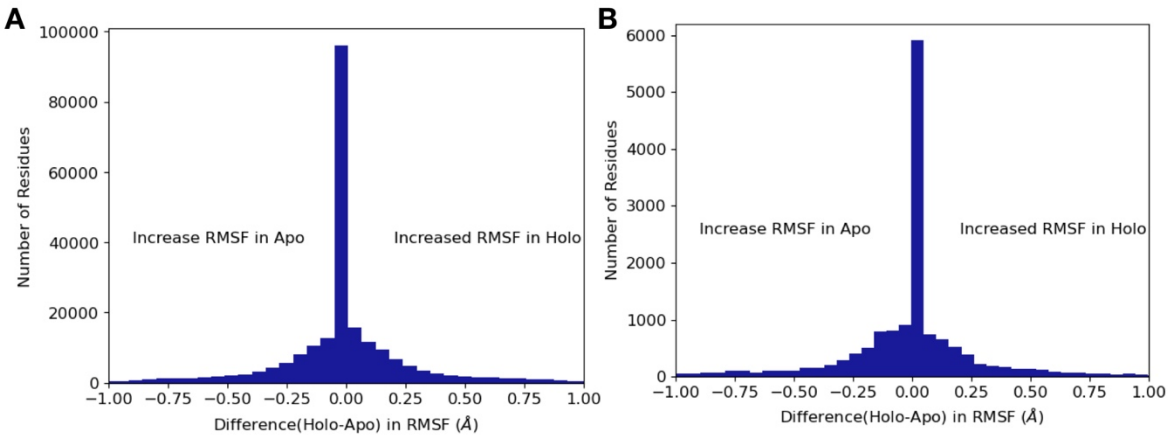


Figure 2.2| Rotamer changes between apo and holo pairs. Examples of rotamer changes between apo (purple) and holo (green) binding site residues. (A) Example residues for: 'no change' in rotamer status, accounting for 78.7% of binding site residues; (B) "distinct" rotamers, accounting for 14.9% of binding site residues; (C) "remodeled- holo loss", accounting for 2.6% of binding site residues; and (D) "remodeled- holo gain", accounting for 3.8% of binding site residues. (E) The percentage of residues in the binding site that have the same rotamer status in the holo and apo structures. The black line highlights the 11% of pairs that had the same rotamer status for all binding site residues. (F) Paired galectin-3 apo (purple; PDB: 5NFC) and holo (green; PDB: 4JC1, ligand: thiodigalactoside) multiconformer models with no changes in rotamer status in any binding site residues. (G) Paired transthyretin apo (purple; PDB: 1ZCR) and holo (green; PDB: 3CFN, ligand: 1-anilino-8-naphthalene) multiconformer models with 6 out of 9 residues with remodeled or different rotamer status in the binding site residues. Residues with rotamer changes are shown as sticks. Residues with the no change in rotamer status are shown as lines.

To compare the magnitude of fluctuations between alternative conformations, we calculated RMSF for all residues. This analysis suggested that, on average, apo residues have slightly greater conformational heterogeneity than holo residues (-0.006Å, mean difference of RMSF(holo-apo); $p=3.7x10^{-8}$, Wilcoxon signed rank test;

Supplementary Figure 2.7). This trend was somewhat stronger in binding site residues (-0.02 Å, mean difference of RMSF(holo-apo); $p=4.5x10^{-29}$, Wilcoxon signed rank test; **Supplementary Figure 2.7**). Our RMSF results suggest that, on average, there is a slight decrease in heterogeneity upon ligand binding and that this reduction is most prevalent at residues distant from the binding site.



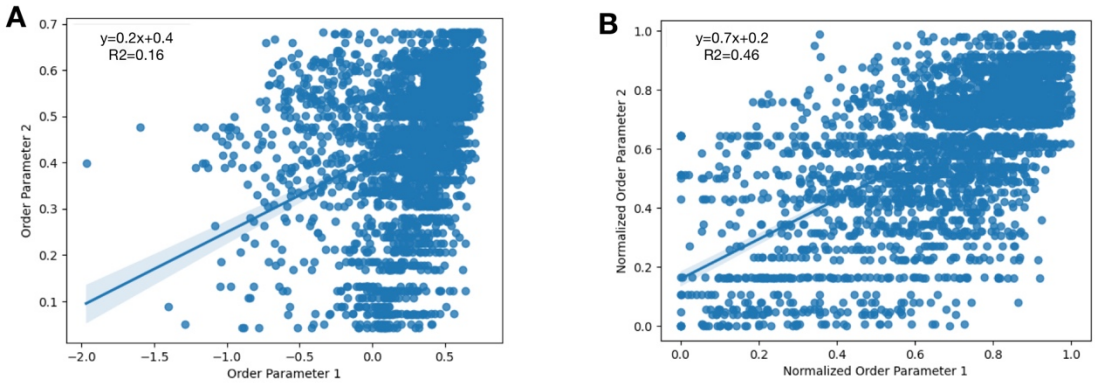
Supplementary Figure 2.7| RMSF differences between apo and holo. *(A) Across all residues, apo residues had a higher RMSF compared to holo residues (0.17*Å vs. *0.16*Å, mean RMSF; *-0.006, mean difference: $p=4.5x10^{-29}$,Wilcoxon signed rank test). (B) Within binding site residues, apo residues also had a higher RMSF, compared to holo residues (0.17*Å vs. *0.15*Å, mean RMSF*; -0.02, mean difference; $p=3.7x10^{-8}$, Wilcoxon signed rank test).*

Collectively, these results do not conform to a simple model. There is a large amount of variability in the response across datasets and the median responses reveal only small biases. Nonetheless, considering those average responses, upon binding a ligand, the RMSF analysis suggests decreases in heterogeneity at the binding site, whereas the rotamer comparison has a slight bias to increased heterogeneity at the binding site, and B-factors only change at distant sites. One interpretation is that heterogeneity is reduced in binding site residues by a small number of anharmonic conformational

changes, as observed by the RMSF reduction, paired with an increase in harmonic fluctuations far away, as observed by an increase in the B-factors. However, it is difficult to interpret these changes separately, as conformational heterogeneity is a combination of both harmonic and anharmonic motion and there is potential degeneracy in modeling alternative conformations, even with qFit [21]. Therefore, we moved to using an integrated measurement of order parameters that can account for these complications [16].

Order parameters integrate both harmonic and anharmonic conformational heterogeneity

To integrate the anharmonic fluctuations between alternative conformers with the harmonic fluctuations modeled by B-factors [12], we used a crystallographic order parameter (**Figure 2.1**) [16]. Order parameters allow us to capture the conformational entropy both within and between side chain rotamer wells. While order parameters are traditionally used in NMR or molecular dynamic simulations, they can be calculated for multiconformer X-ray models and, in some cases, show reasonable agreement with solution measures [16]. We focused on the order parameters of the first torsion angle (χ1) of every sidechain for all residues except for glycine and proline. Order parameters are measured on a scale of 0 to 1, with 1 representing a fully rigid residue and 0 representing a fully flexible residue. Below, we analyze the differences in normalized order parameters between paired residues (**Methods, Supplementary Figure 2.8**).



Supplementary Figure 2.8| Order parameter normalization. To normalize the order parameters across all structures, we looked at 31 lysozyme structures and compared their order parameters. We randomly selected 3 PDBs as our 'control' order parameters (PDBs: 1VAT, 4NHS, 5LIO). (A) For every residue, we plotted the initial order parameters of our control PDBs compared to all of the other PDBs in our dataset. We obtained a slope of 0.2 with an R^2 of 0.16. (B) After applying our equation that accounts for average B-factor and resolution (**Methods**) we re-plotted the normalized order parameters. Here we obtained a slope of 0.7 and an R^2 of 0.46.

As an additional control, we compared our apo/holo dataset to a dataset of apo/apo pairs. In examining the differences in order parameters, both in the holo/apo pairs and the apo/apo pairs, there are no large differences in conformational heterogeneity, as indicated by a median difference in order parameters of approximately 0. However, in the holo/apo pairs there is a much wider range of order parameter differences, indicating that ligand binding impacts conformational heterogeneity beyond experimental variability ($p=3.4x10^{-17}$, individual Mann-Whitney U test; **Figure 2.3**).

Next, to examine whether different regions of the protein were driving this higher variability, we compared the differences in order parameters among binding site residues, within 5Å of any ligand heavy atom, compared to a control dataset which matched the number of, type and solvent exposure within the protein for each binding site residue. In binding site residues, the holo structures had a slightly, but significantly,

increased order parameters, suggesting reduced conformational heterogeneity compared to the control dataset (0.034 vs. 0, median difference (holo-apo) order parameter; $p=3.4x10^{-7}$, individual Mann-Whitney U test; **Figure 2.3**). While there is a larger range of responses, this indicates that, in general, binding site residues become more rigid upon ligand binding.

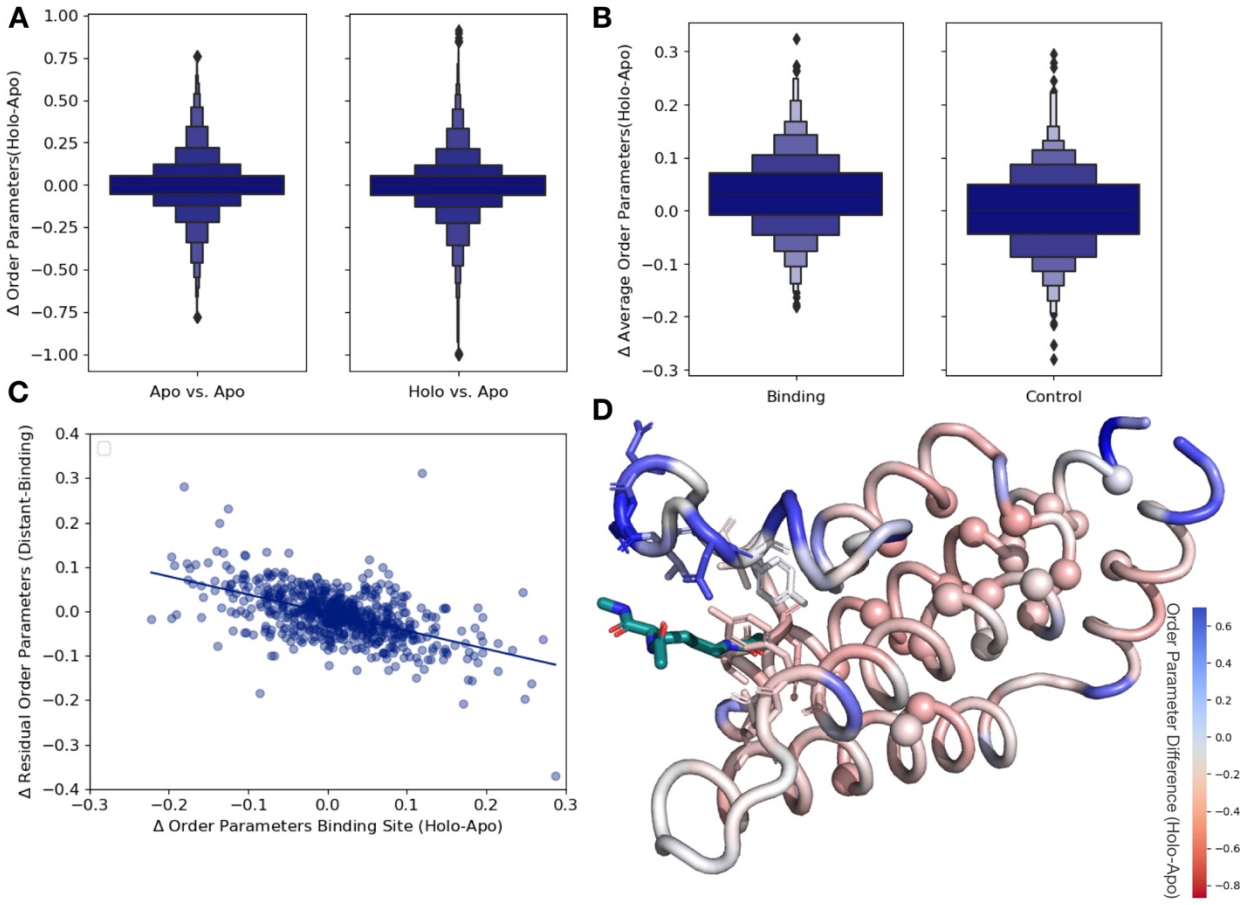


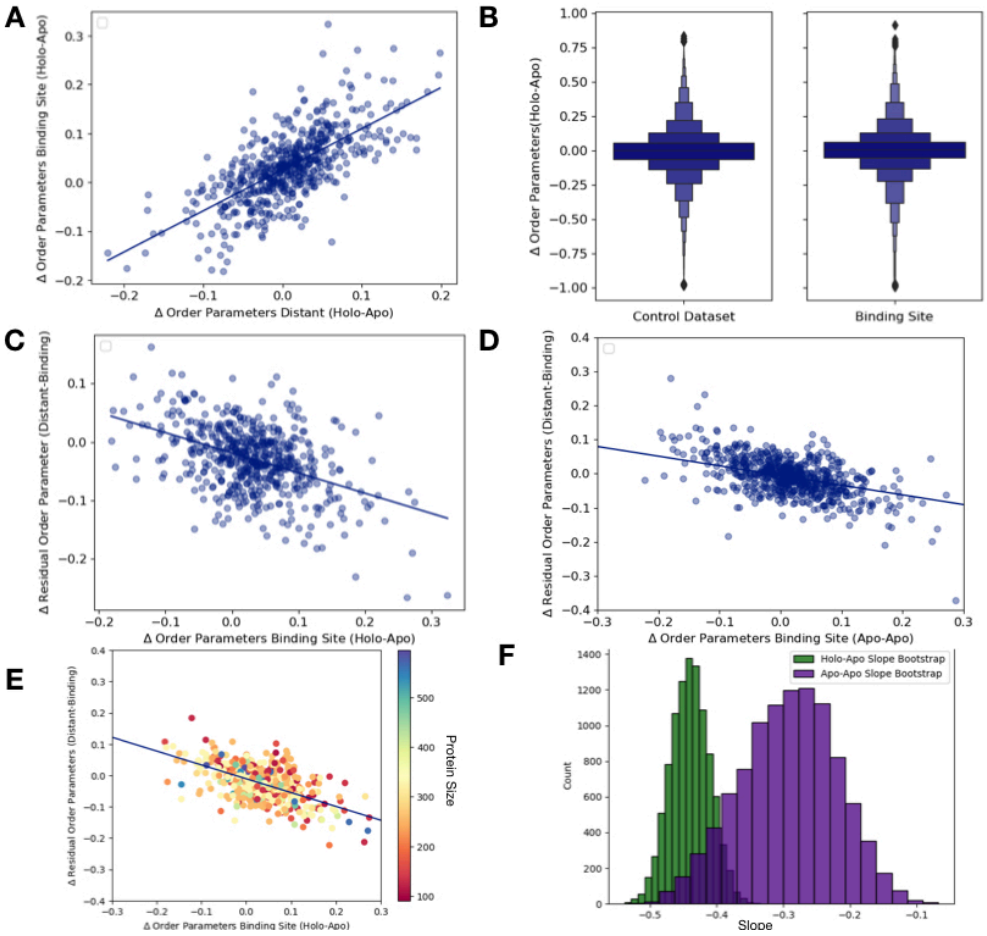
Figure 2.3| Ligand binding alters conformational heterogeneity patterns. (A) Across all residues, the distribution of order parameter changes is much wider in Holo-Apo pairs compared to Apo-Apo pairs ($p=3.4x10^{-17}$, individual Mann-Whitney U test), however there is no median difference in order parameters upon ligand binding (median difference: 0 for both) indicating that ligands have varying impacts across different proteins. (B) The distribution of the average differences of order parameters in binding site residues compared to the average differences in a control dataset made up of the same number, type, and solvent exposure of amino acids. Comparing the holo/apo structures, on average binding site residues got more rigid upon binding. The median difference in order parameters was 0.03 for the binding site residues, compared to 0 for the control dataset ($p=3.4x10^{-7}$, individual Mann-Whitney U test). (C) The relationship of the difference in order parameters between the holo and apo residues in binding site residues versus the residual order parameter in distant, non-solvent exposed residues. We observed a negative trend (slope=-0.44) indicated that structures that had a loss of heterogeneity in the binding site (right on the x-axis) had a relative gain in heterogeneity in residues distant from the binding site that were not solvent exposed (top on the y-axis). (D) We explore this trend on a structure of human ATAD2 bromodomain (PDB: 5A5N). Residues are colored by the differences between the average binding site order parameter minus the order parameter for each residue. Blue residues are less dynamic than the average binding site residue and red residues are more dynamic than the average binding site residue. Binding site residues are represented by sticks and

distant, non-solvent exposed alpha carbons are represented by spheres. The ligand ((2S)-2,6-diacetamido-N-methylhexanamide) is colored in teal.

Spatial distribution of conformational heterogeneity changes

Based on the large range of order parameter differences we observed across the protein, along with the decrease in heterogeneity localized to binding site residues, we next explored the relationship between changes in heterogeneity in binding site residues and the rest of the protein. The difference in order parameters between the holo and apo models were correlated in both the binding site and distant residues (**Supplementary Figure 2.9**), indicating that ligand binding generally caused global changes to flexibility. Given the average rigidification of the binding site residues (Figure 2.3**, Supplementary Figure 2.9**)**,** these results predict a general trend of decreased conformational heterogeneity in the ligand binding site would be associated with a relative increase in conformational heterogeneity at distant sites in the protein. This pattern suggests that the residual change in heterogeneity (the difference between the average order parameter of the distant residues and the average order parameter of the binding site residues) should be inversely related to the change in the binding site residues: more rigidified binding sites will have more flexible than expected distant sites, and vice versa. Therefore, we explored the relationship between binding site residues and distant residues, defined as those more than 10Å away from any heavy atom in the ligand. Indeed, on a protein-by-protein basis, the relationship between binding site residues and residual changes at distant sites follows this trend (**Supplementary Figure 2.9**)**.** Consistent with studies suggesting significant residual conformational heterogeneity in folded buried residues [22] and the potential for those buried residues to

change heterogeneity upon ligand binding [23], this trend is even stronger in residues that were more than 10Å away from any heavy atom in the ligand and less than 20% solvent exposed (slope=-0.44, r^2=0.46; p=5.1x10^{-50}, two-sided t-test; Figure 2.3). This indicates that proteins that lose conformational heterogeneity in the binding site are associated with a relative increase in conformational heterogeneity in distant, non-solvent exposed residues.



Supplementary Figure 2.9| Conformational heterogeneity analysis. (A) The relationship between the average order parameter in distant, non-solvent exposed residues versus the average order parameters in binding site residues (n=743, slope=0.79, r^2=0.65; p=6.5x10^{-89}, two-sided t-test). (B) We compare the difference in order parameters in each binding site residues of Holo-Apo pairs compared to a control dataset made up of the same number, type, and solvent exposure of amino acids. Comparing the holo/apo structures, on average binding site residues got more rigid upon binding. The median difference in order parameters was 0.03 for the binding site residues, compared to 0 for the control dataset (p=3.4x10^{-7}, individual Mann-Whitney U test).(C) The relationship between the residual order parameters in all distant residues versus binding site residue order parameters (n=743, slope=-0.34, r^2=0.17; p=4.6x10^{-28}, two-sided t-test). (D) The relationship between the residual order parameters in distant, non-solvent exposed residues versus binding site residues in the apo and apo control dataset residues (n=283, slope=-0.28, r^2=0.20; p=1.8x10^{-34}, two-sided t-test). (D) The relationship between the residual order parameters in distant, non-solvent exposed residues versus binding site residues in the holo dataset residues colored by size of protein. (F) The bootstrap analysis of the overlap of the slope of distant, average order parameters of non-solvent exposed residue versus average order parameters of binding site residue between holo-apo (green) and apo-apo (purple). While there was some overlap the mean slope of holo-apo (-0.44) was more than two standard deviations

away from the mean slope of the apo-apo (-0.28). Comparing the two bootstrap distributions using a z-test, the z-value was -191.26 with a p-value of 0.0.

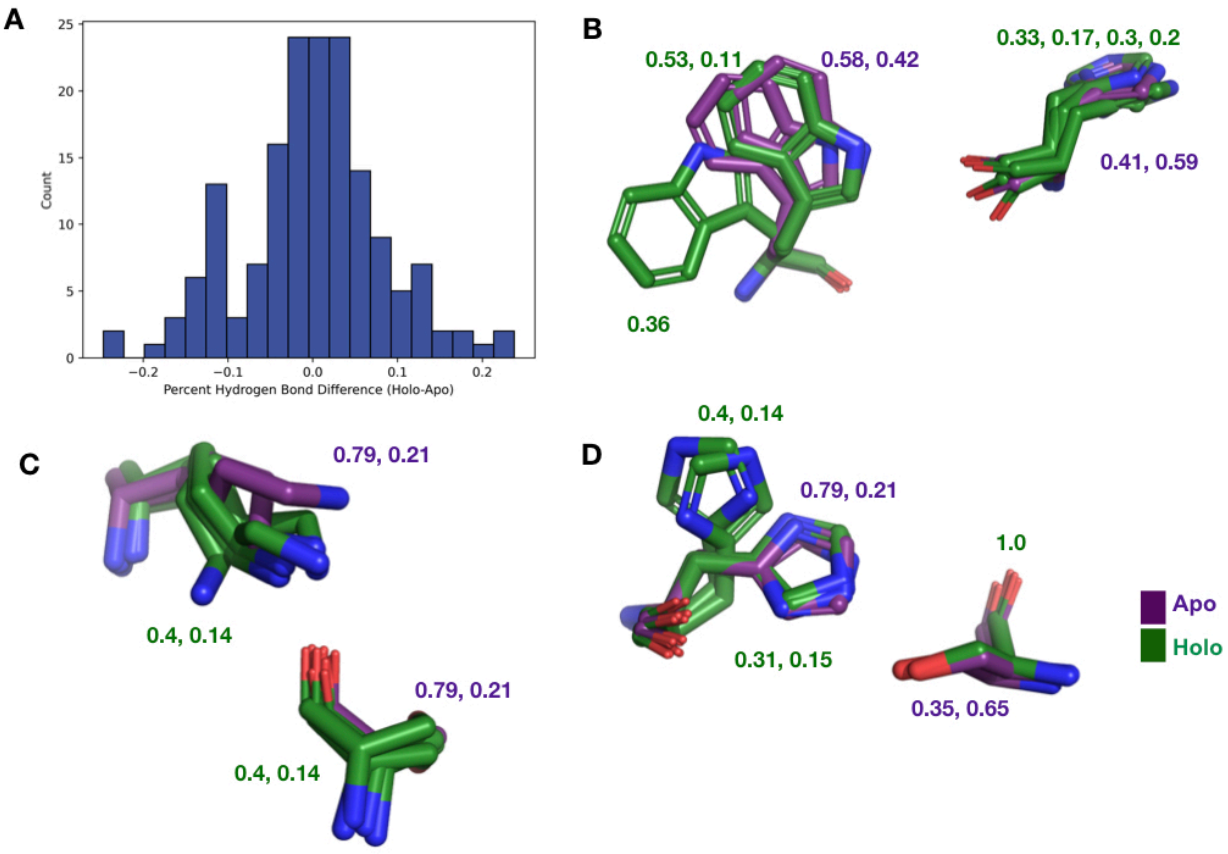
There are three likely origins of this effect. First, this may reflect a feature of the distribution of order parameters around the mean value within each protein. Second, this may reflect a topological feature of protein packing, whereby packing optimization of certain areas of a protein decreases the optimization of other parts of the protein [24]. Third, this may reflect the stabilization of certain conformations in a ligand bound protein. As a control for these effects, we compared the residual order parameter differences between the buried, non-solvent exposed residues and the binding site residues in apo-apo pairs. Globally the trends were similar, but weaker in both correlation and magnitude (slope=-0.28, r^2=0.20; p=1.8x10^{-34}, two-sided t-test; **Supplementary Figure 2.9**). Therefore, we interpret the trend we observe as mainly based on protein topology, specifically that proteins have areas where there are less efficiently packed alternative conformers, likely to enable entropic compensation across the protein during various functions, including ligand binding. We interpret that the stronger signal we observed in the holo-apo dataset is due to the ligand perturbation, which is also reflected in the median rigidification of binding site residues (Figure 2.3). We hypothesize that we are observing this innate protein property being used, specifically optimizing the binding site residues to bind a ligand, while decreasing the optimization elsewhere in the protein.

As an example to visualize this trend, we mapped the change in order parameters onto the structure of the human ATAD2 bromodomain (PDB ID:5A5N). In ATAD2, the binding

site residues rigidify upon ligand binding whereas the majority of distant residues are more heterogeneous compared to the binding site residues (Figure 2.3). Specifically, this difference is greatest between binding residues and non-solvent exposed residues, as previously observed in lysozyme (**Supplementary Figure 2.9** [6,23]). However, as in the global analysis, the ATAD2 example demonstrates there is a large range of changes in binding site order parameters, consistent with NMR examples that show a heterogeneous response both close to and distant from ligands (Caro, Valentine, and Wand 2021),

Hydrogen bond patterns change upon ligand binding

We next investigated changes in protein side chain hydrogen bonds upon ligand binding. Here we applied HBplus[25] to identify hydrogen bonds for each side chain alternative conformation (Methods). We examined the occupancy weighted hydrogen bonds in binding site residues, using a hydrogen bond cutoff of 3.2Å. Overall, we observed the creation of 0.06 hydrogen bonds per residue in holo binding sites (**Supplementary Figure 2.10**), which translates to 10% of structures gaining one full hydrogen bond in the holo structure. This is likely indicative of stable binding sites in holo structures. This follows a trend observed previously where upon ligand binding, hydrogen bonds to water molecules decrease, but hydrogen bonds to other protein atoms increase[20].



Supplementary Figure 2.10| Hydrogen bonding patterns. We examined the difference in hydrogen bonds across all binding site residues. (A) The percentage difference in hydrogen bonds between holo and apo structures in binding site residues. (B) We observe W118 having a much different conformation in altB in apo structure breaking the hydrogen bond with H122. (C) K64 in the apo structure is unable to make any hydrogen bonds with S133 due to wandering nitrogen in the last chi angle of K64. (D) alt A and B in H97 of the apo structure have a much different conformation from H97 in the holo structure.

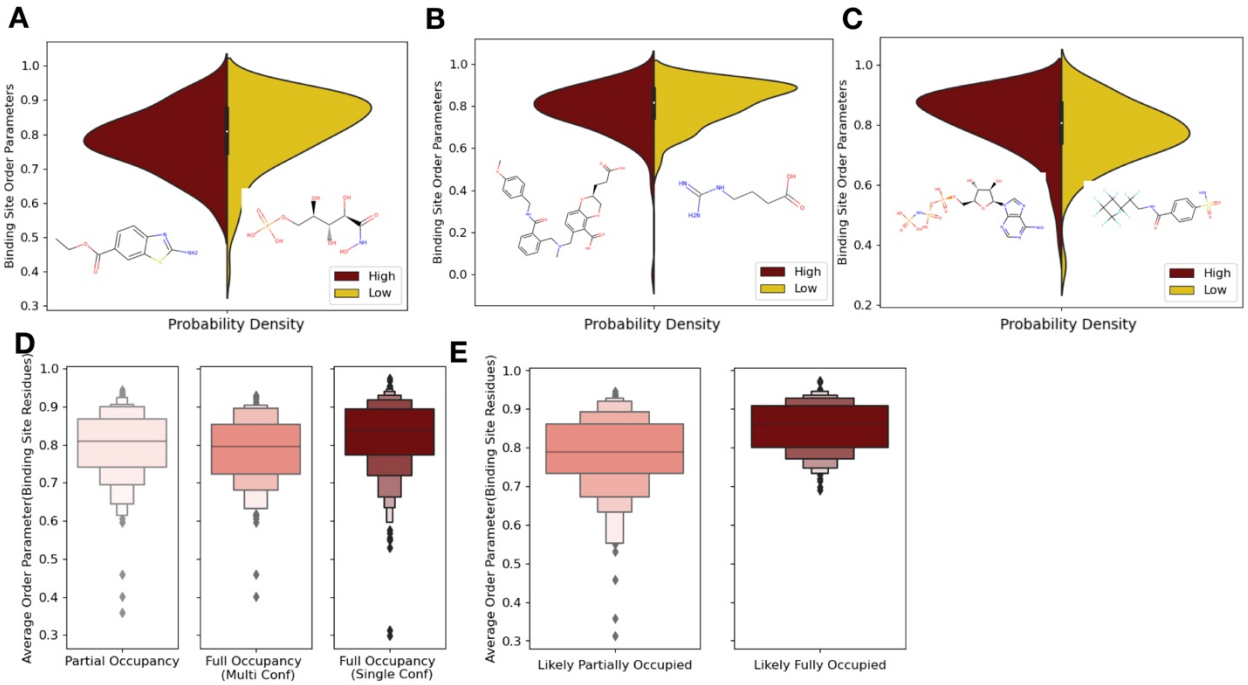
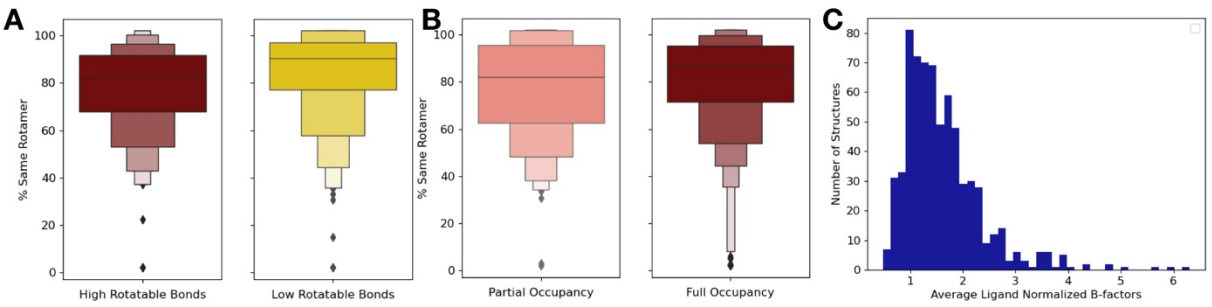


Figure 2.4| Ligand properties impact binding site order parameters**.** (A) Ligands with higher logP value (maroon), indicative of more greasy or hydrophobic ligands, versus ligands with a lower logP value (gold), had lower in order parameters in the binding site residues (0.78 vs. 0.84, median order parameter; $p=7.5x10^{-6}$, independent Mann Whitney U test) [Example ligands: low logP: 5-phospho-d-arabinohyroamic acid, high logP: ethyl 2-amino-1,3-benzothiazole-6-carboxylate]. (B) Ligands with relatively higher molecular weight (maroon) had higher order parameters compared to those with lower molecular weight(gold; 0.79 vs. 0.83, median order parameter; p=0.0001, independent Mann Whitney U test). [Example ligands: High number of heavy atoms: (2S)-2-(3-hydroxy-3-oxopropyl)-6-[[[2-[(4-methoxyphenyl)methylcarbamoyl]phenyl] methyl-methyl-amino]methyl]-2,3-dihydro-1,4-benzodioxine-5-carboxylic acid, low number of heavy atoms: 4-carbamimidamidobutanoic acid]. (C)Ligands with relatively higher hydrogen bonds per heavy atom (maroon) had higher order parameters compared to those with lower molecular weight (gold; 0.84 vs. 0.79, median order parameter; $p=5.9x10^{-5}$, independent Mann Whitney U test) [example ligands: low hydrogen bond: 4-sulfamoyl-N-(2,2,3,3,4,4,5,5,6,6,6-undecafluorohexyl) benzamide, high hydrogen bond: phosphoaminophosphonic acid-adenylate ester]. (D) Binding site order parameters were lower in ligands with partial occupancy (light pink; 0.79, median order parameter) and mutliconformer ligands adding to full occupancy (salmon; 0.80, median order parameter), compared to single conformer ligands with full occupancy (dark red; 0.83, median order parameter; $p=4.9x10^{-8}$, independent Mann Whitney U test). (E) In fully occupied ligands, ligands in the top quartile of ligand B-factors, controlled for by the mean alpha carbon B-factor, had lower binding site order parameters (salmon; 0.79, median order parameter) compared to ligands in the bottom quartile (dark red; 0.85, median order parameter; $p=1.6x10^{-11}$, independent Mann Whitney U test).



Supplementary Figure 2.11| Conformational heterogeneity and ligand properties. (A) We explored if the top and bottom quartiles of rotatable bond ligands were associated with an increase or decrease of rotamer changes, as defined as the percentage of close residues with the same rotamer in the holo and apo structure. The ligands in the top quartile of rotatable bonds had less rotamers that were the same between holo and apo structures versus ligands in the bottom quartile of rotatable bonds (80% vs. 88%, median same percentage of rotamers, p=0.001, independent Mann Whitney U test). (B) There was no significant difference in the percentage of the same rotamers between partially occupied and fully occupied ligands (80% vs. 85%, median percentage of the same rotamer; p=0.11, independent Mann Whitney U test). (C) In fully occupied ligands, the median B-factor was 24.8, with a range of 5.5 to 99.3.

Ligand properties influence conformational heterogeneity

Next we investigated how ligand properties impact the conformational heterogeneity of binding site residues. For ligand properties dictated by the size of the ligand (number of rotatable bonds and number of hydrogen bonds) we normalized these metrics by the molecular weight of the ligand. For each property, we compared the highest and lowest quartiles by both the absolute order parameters of the holo structure and the order parameters differences between holo and apo pairs. No significant associations existed when comparing the differences between holo and apo order parameters, but the characteristics of the holo binding site and the rotamer changes were correlated with ligand properties in several cases.

We hypothesized that ligand properties associated with increase ligand dynamics, including more rotatable bonds, higher lipophilicity (logP), fewer hydrogen bonds, and more heavy atoms would be associated with increased conformational heterogeneity (an increase in absolute order parameters or a smaller difference between the apo and holo order parameters; [26]. While molecules with fewer rotatable bonds (lower quartile: <2 (n=134) vs. upper quartile: >6 (n=134)) were indeed associated with more rigid binding sites (lower quartile: 0.83 vs. upper quartile: 0.81, individual Mann Whitney U test), this was not significant. However, higher numbers of rotatable bonds were associated with a lower number of same rotamers between the apo and holo binding site residues (88% vs. 80%, percentage same rotamer; $p=6.0x10^{-6}$, individual Mann Whitney U test (**Supplementary Figure 2.11**). Increased lipophilicity (logP, upper quartile: <0.04 (n=134) vs. lower quartile: >2.69 (n=134)), was significantly associated with a more flexible binding site (0.79 vs. 0.84, median order parameters; $p=7.5x10^{-6}$, individual Mann Whitney U test; **Figure 2.4**). Previous studies have indicated that increased lipophilicity generates more nonspecific binding interactions [27]. Larger compounds (upper quartile: >26 heavy atoms (n=134) vs. lower quartile: <13 heavy atoms (n=134)) are also associated with more flexible binding sites (0.83 vs. 0.79, median order parameter; p=0.0001, individual Mann Whitney U test; **Figure 2.4**). Large compounds, thus a larger ligand surface area, are associated with more nonspecific binding interactions, which is compatible with increased protein conformational heterogeneity. Finally, more total hydrogen bonds per heavy atom(upper quartile: >0.47 (n=134) vs. lower quartile: <0.25 (n=134)) are associated with more rigid binding sites (0.84 vs. 0.79, median order parameter; $p=5.9x10^{-5}$, individual Mann Whitney U test;

Figure 2.4). This trend holds even when examining hydrogen bond donors or acceptors separately.

From these results, an intuitive general picture emerges where more specific, directional interactions, such as hydrogen bonds[28], are more likely to lock the corresponding protein residue in place, thus creating more rigid binding site residues[29]. Whereas the more non-specific interactions are correlated with more flexible binding site residues. There is also a wide range of deviation from this general picture, likely reflecting that natural and artificial optimization of ligands is based on free energy, not any specific thermodynamic component or interaction type. These trends emphasize the need to monitor both the impacts of ligands on specific interactions with the protein along with conformational heterogeneity of the protein. Additionally, these results suggest that specific interactions can be tuned to rigidify a binding site. Paired with our findings of the relationship between order parameters in binding site and distant residues, ligand impacts are likely propagated throughout the protein. Ligands with more specific interactions, thus a less flexible binding site, will likely have a corresponding increase in conformational heterogeneity distant from the binding site.

Reduced ligand occupancy and conformational heterogeneity

One potential confounder for quantifying the change in conformational heterogeneity of binding site residues is that the ligands may not be fully occupied in the crystal. There were 193 structures with ligands with alternative conformations or partially occupied ligands in our datasets (**Figure 2.4**). Of these 193, 125 ligands had less than full occupancy, whereas 68 had alternative conformations that amounted to full occupancy.

The vast majority of ligands (n=425) were modeled originally with full occupancy. Fully occupied ligands were associated with more rigid binding sites than partially occupied ligands or ligands with alternative conformers (0.84 vs. 0.79, mean order parameters of binding site residues; $p=2.9x10^{-7}$, individual Mann Whitney U test; **Figure 2.4**). There was no difference observed between the partially occupied ligands and ligands with alternative conformers (p=0.15, individual Mann Whitney U test). We also explored if partially occupied ligands were associated with more rotamer changes between holo and apo pairs, but no significant difference existed (80% vs. 85%, median percentage of the same rotamer; **Supplementary Figure 2.11**).

While the scattering contributions of B-factor and occupancy changes are subtle (but distinct), most models likely include true occupancy changes as elevated B-factors. We observed a wide range of average ligand B-factors and, as expected, a lack of correlation between the ligand B-factors and ligand occupancy [30–32]. As a proxy for likely partially occupied ligands, we normalized the ligand B-factor by the mean C-alpha B-factor to identify ligands with higher B-factors than expected (**Supplementary Figure 2.11**). We examined the outer two quartiles of the normalized ligand B-Factors (>0.016 vs. <0.005, median normalized B-factor). In these "likely partially occupied" ligands, we observed greater conformational heterogeneity (0.86 vs 0.80, mean order parameter; $p=1.6x10^{-11}$; individual Mann Whitney U test, **Figure 2.4**). In structures with modeled partially occupied ligands and likely partially occupied ligands, we learned that binding site residues tend to have more apparent conformational heterogeneity, likely due a combination of compositional and conformational heterogeneity.

Conformational heterogeneity for multiple ligands to CDK2

To better understand our findings in the context of multiple, diverse ligands binding to a single receptor, we examined Cyclin-dependent Kinase 2 (CDK2), a cyclin kinase family that regulates the G1 to S transition in the cell cycle. Our dataset contains 13 protein-inhibitor complexes, including both type I and type II inhibitors, all of which share the same apo model (PDB ID: 1PW2). We hierarchically clustered the residues and ligands by difference in order parameters between the holo and apo models, identifying three distinct clusters of residues. The first cluster (blue, **Figure 2.5**), consisting of 13 residues, are rigidified upon ligand binding. This cluster included residues scattered throughout both the N- and C-lobes of CDK2 that rigidified upon ligand binding. This dispersed pattern is similar to the trend of rigidification that is observed by NMR in PKA upon substrate binding, suggesting that changes in conformational dynamics in kinases systems are structurally dispersed as a function of ligand state [33]. Two notable residues in this cluster, Glu127 and Val18, contact the inhibitors. Upon ligand binding, Val18 transitions from multiple conformers to a single conformation. Glu127 has a similar conformation in the apo and type II structures of two distinct alternative side chain rotamers, whereas in the type I inhibitor structure, the alternative conformers cluster around the same rotamer (**Supplementary Figure 2.12**).

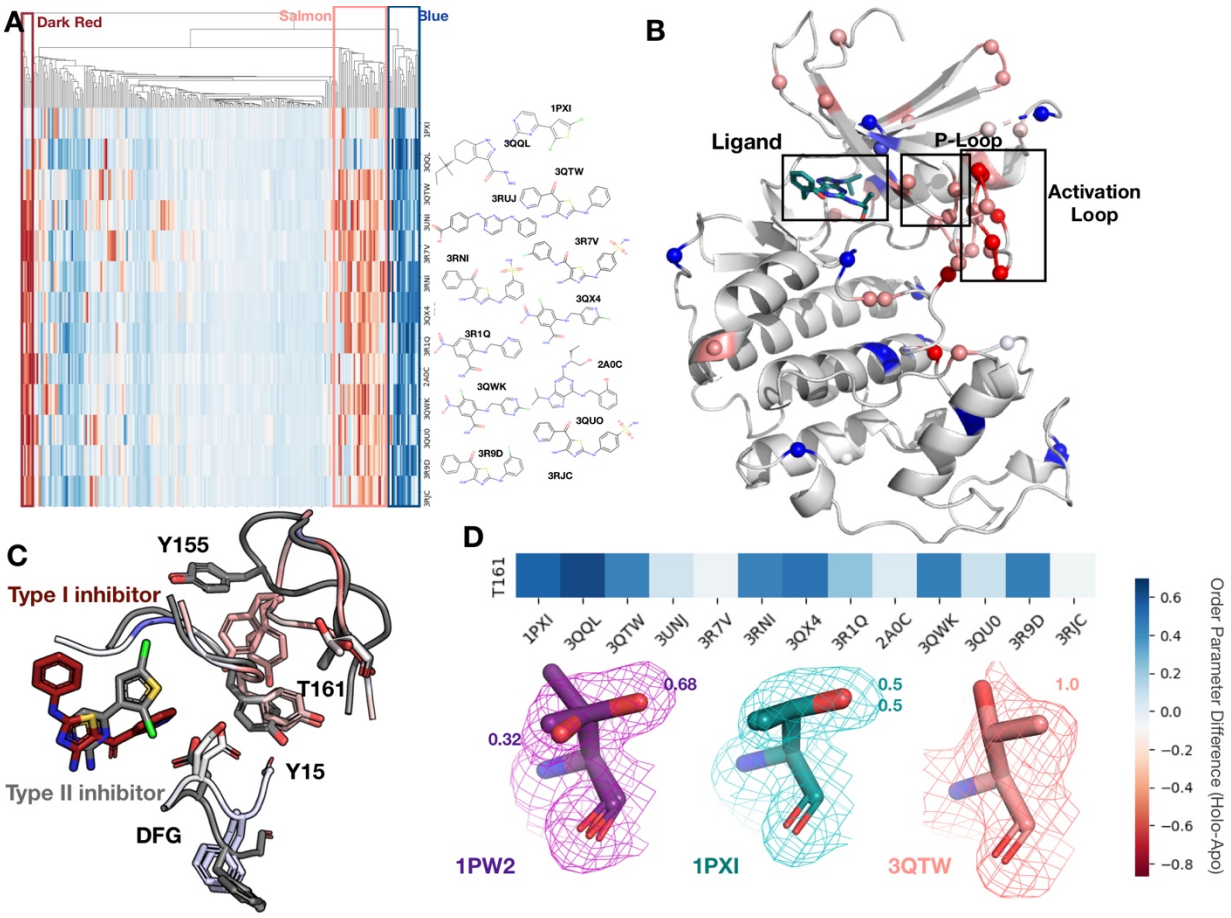
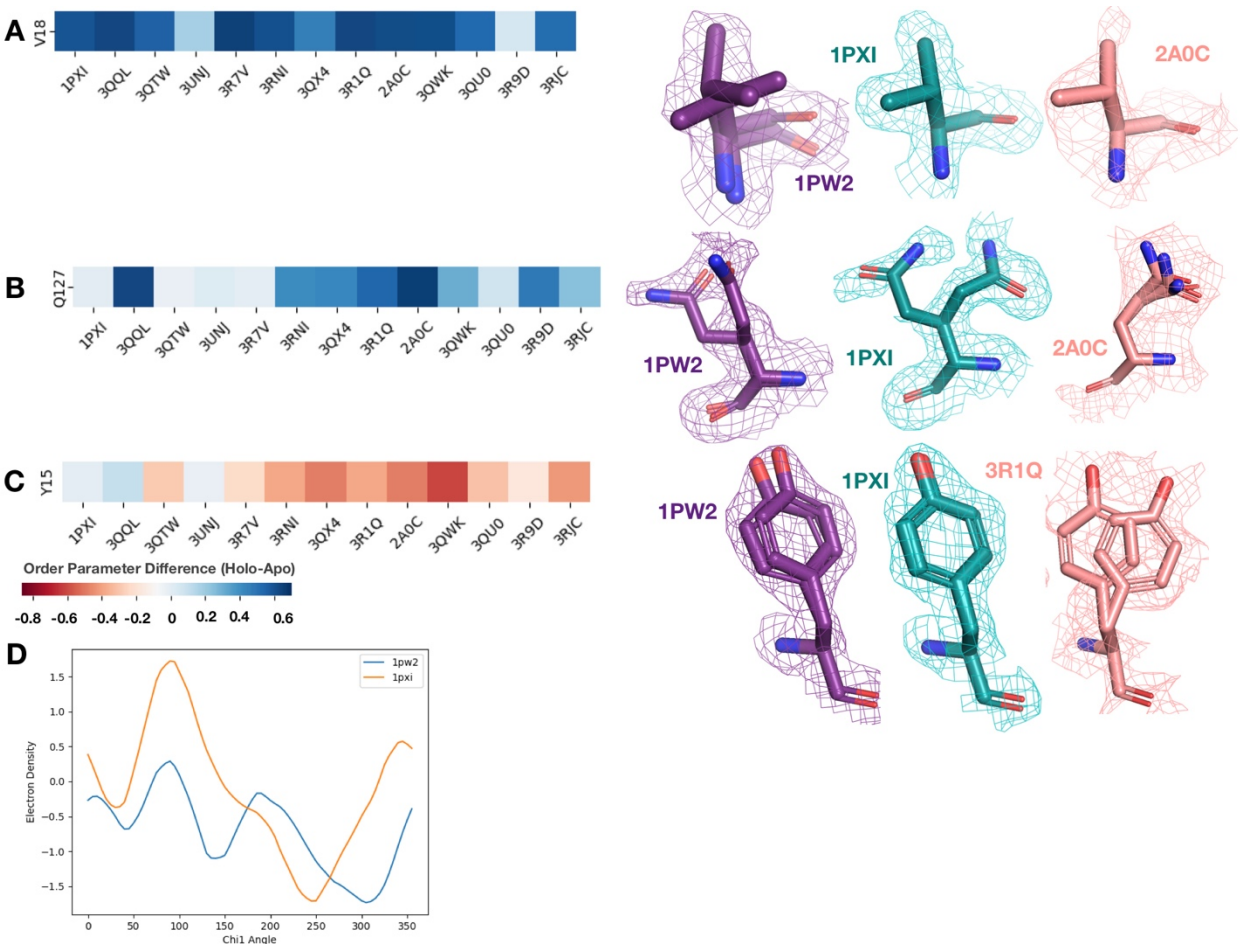


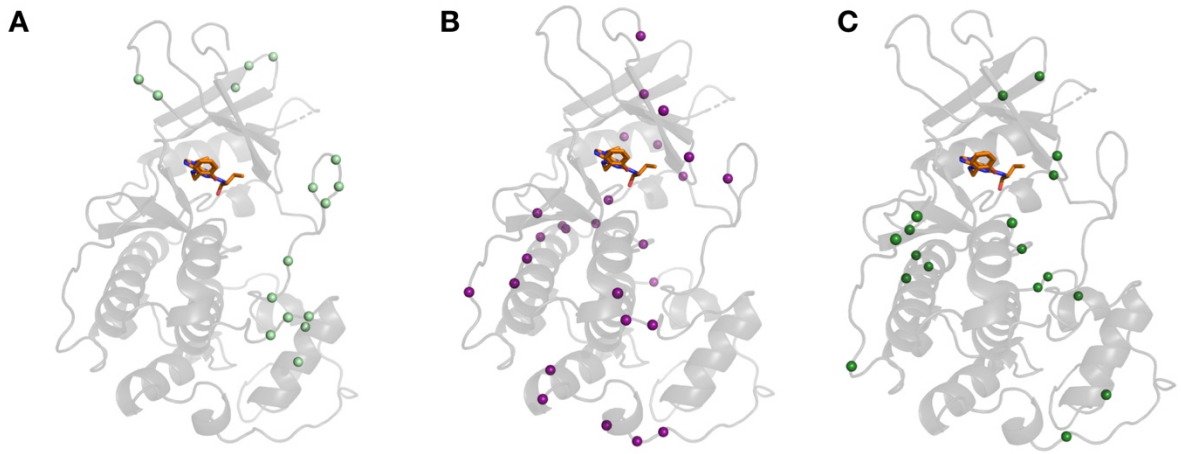
Figure 2.5| Conformational change and heterogeneity in CDK2. (A) The clustermap of all residues in the 13 CDK2 protein/ligand pairs. Red values indicate a negative difference (holo-apo) in order parameters, indicating that the holo structures have more heterogeneity compared to the apo. Blue values indicate positive differences, indicating that the apo structures have more heterogeneity compared to the holo. We highlighted three important clusters, the left red cluster, middle salmon cluster, and right blue cluster. (B) A representative structure (PDB: 3QTW) is shown with each residue colored by the difference in order parameter, corresponding to the same coloring scheme as the clustermap. The three distinct clusters (dark red, salmon, blue) are shown in spheres. (C) Many of the key differences between type I inhibitor (PDB: 3QTW) and type II inhibitor (PDB: 1PXI) are located in the DFG motif, p-loop, and activation loop. The type II inhibitor structure is colored in grey and the type I inhibitor is colored as the difference in order parameters between the type I inhibitor and type II inhibitor structures. Red signifies a more dynamic region in the type I inhibitor structure, blue signifies a less dynamic region in the type I inhibitor structure. Changes in the DFG motif, propagates changes, both structural and in dynamics, in the p-loop (highlighted by Tyr15), which progates even larger changes in the activation loop between the two inhibitors, including changes in conformation of Thr161, the phosphorylation site of CDK2. (D)Threonine 161, the phosphorylation site for CDK2. We looked at the supporting density for specific residues between the apo (PDB: 1PW2, purple), type II (PDB: 1PXI, teal), and type I

(PDB: 3QTW, salmon) inhibitors. 2Fo-Fc electron density is shown at 1 sigma. Occupancies of the alternative conformers are labeled with the corresponding color. The apo structure has multiple conformations, whereas the type I model only has one, and the type II model has two very similar conformations, but these are in different rotamer states compared to the apo.



Supplementary Figure 2.12| CDK2 density in key residues. We looked at the difference in order parameters (holo-apo) and the supporting density for specific residues between the apo (PDB: 1PW2, purple), type II (PDB: 1PXI, 3QQL, teal), and type I (PDB: 2A0C, 3QTW, 3R1Q, salmon) inhibitors. All density is shown at 1 sigma. (A) Valine 18, one of the ligand contacts for both the type I and type II inhibitors. Across all holo structures this residue becomes more rigid, including losing an alternative conformer and changing rotamers in the holo structure. This residue is also a part of the blue cluster in the heatmap. (B) Glutamine 127, one of the ligand contacts for both type I and type II inhibitors. This residue has two very different alternative conformers in the apo structure. In the type II inhibitor structure, there are again two very different alternative conformers whereas in the type I inhibitor structure, there are three very similar alternative conformers. This residue is also a part of the blue cluster in the heatmap. (C) Tyrosine 15 in the P-loop has varying differences in order parameters. In the type II inhibitor, this tyrosine gets more rigid, along with the rest of the p-loop, however in the type I inhibitor structures, this tyrosine along with the rest of the p-loop becomes more dynamic. (D) Ringer analysis to detect alternative conformations in electron density maps. Ringer detected two peaks for 1PW2, indicating two alternative conformers, whereas only one peak was detected for 1PXI, indicating only one conformation.

The second cluster (salmon, **Figure 2.5**), consists of 14 residues that increase flexibility upon ligand binding. The majority of these residues connect the p-loop and the activation loop (**Figure 2.5**). The electron density is very weak for many of these residues in most of the holo structures, driving their modeling in multiple conformations and elevated B-factors (**Supplementary Figure 2.12**). We also observed that many of these residues had sidechain to sidechain hydrogen bonds that were lost upon ligand binding (**Supplementary Figure 2.13**). The third cluster (dark red, **Figure 2.5**) is comprised of five residues that became more flexible in all, but two holo datasets, which are the only type II inhibitors in the dataset. These were all located on the activation loop of the kinase (**Figure 2.5**). As type II inhibitors, the two molecules [PDB: 1PXI (ligand: CK1) and PDB: 3QQL (ligand: X03)] bind the DFG out conformation present in the apo dataset (PDB: 1PW2) and do not have as drastic of a rigidifying effect as the type I inhibitors. Notably, these two inhibitors were also smaller than the type I inhibitors and the reduced contacts may also drive some of this effect. We also observed that the hydrogen bonds gained in the holo structure are inhibitor specific (**Supplementary Figure 2.13**).



Supplementary Figure 2.13| Hydrogen bond differences in CDK2. We examined the difference in hydrogen bonds across CDK2 structures. (A) Hydrogen bonds broken in the majority of holo structures located in loop regions, especially present in the activation loop. (B) Hydrogen bonds formed upon ligand binding was unique to inhibitors as observed in 3qtw(B, purple) and 2a0c (C, green).

The multiconformer models also provide a structural rationale for these changes. The differences in DFG conformation change the contacts with the P-loop, which allow for greater side chain flexibility in the “up” form compatible with type I inhibitors. The interface between the P-loop and the activation loop is weaker and residues such as Tyr155 adopt multiple conformations. At the base of the activation loop, Thr161, a critical phosphorylation site, changes conformation, with a rigidifying effect common to both type I and II inhibitors (Figure 2.5**, Supplementary Figure 2.13**). The conformation of Thr161 found in the type II inhibitors overlap, with one of the conformations populated in the multiconformer apo model. In contrast, the type I inhibitors adopt a distinct conformation. This case study highlights how modeling information present in the density can reveal changes beyond those in single conformer structures.

Discussion

By creating a large dataset of stringent matched pairs of apo and holo multiconformer models, we identified a pattern of conformational heterogeneity consistent with smaller scale studies of individual proteins [6]. We observed that individual proteins greatly varied in amount and direction of change of conformational heterogeneity, as observed in previous studies (Caro et al., 2017). In general, we found that binding site residues tend to become more rigid upon ligand binding. But similar to the entire protein, there was a large range of effects, including many sites becoming more flexible when bound to a ligand. The trends suggest that disorder-order transitions between binding site residues and distant residues are common and potentially a selected property of many proteins [33]. Specifically, our data suggests that some of the entropy lost by the rigidification incurred by binding site residues upon ligand binding can be compensated with an increase in disorder in distant residues. This finding generalizes the phenomenon has been observed in single protein analyses with NMR and MD simulation [23,34,35]. Both theoretical and experimental analyses suggest that the relationship between local packing optimization and small voids that permit alternative conformations will be key to predictably mapping this relationship [6,24]. Using temperature or pressure as perturbations during X-ray data collection can help to further map the connection between packing “quality” and side chain conformational heterogeneity in greater detail [6]. While NMR order parameter studies only take into account movement that is shorter than the tumbling time for the protein (Hoffmann et al., 2021; Gangé et al., 1998), our results are insensitive to timescale. In addition, it is quite likely that our use of cyro-cooled structures causes an underestimate of the heterogeneity occurring in these

datasets [36], and may potentially bias our results by locking in certain populations of the protein ensemble. This effect may particularly impact areas of the protein and surrounding solvent that go from a preorganized, low energy state to a more dynamic state as observed in Galectin-3 and Barnase[6,37]. This study can also serve as a template to investigate other perturbations including mutations, pressure, or temperature.

We observed a complex interplay between conformational change and dynamics in our analysis of 13 inhibitor-bound datasets of the kinase CDK2, in the same crystal form and space group. The ability to explore one protein with multiple ligands highlights the utility of crystal systems amenable to isomorphous soaking or co-crystallization [38]. We identified differences in conformational heterogeneity between type I and type II inhibitors that can be classified along with well-known changes, such as differences in the DFG motif. Tuning distal site dynamics may be a viable strategy for modulating the affinity of kinase inhibitors and affect the pattern of protein-protein interactions on distal surfaces, which is of critical importance in CDK inhibitor development [39,40].

We note that our work is not sensitive to many facets of the complex changes associated with ligand binding [1]. Our stringent resolution matching criterion may also render us blind to the most severe effects on conformational heterogeneity, whereby ligand binding causes a more widespread change leading to a loss or gain of diffraction power. In addition, water molecules play an important role in ligand binding, both in the release of ordered water molecules contributing to binding through entropy and in

forming specific interactions [41,42]. Additionally, ligand conformational heterogeneity has been highlighted by several recent studies [30,43,44]. Another caveat in our analysis is the limitations of qFit modeling for modeling extensive backbone heterogeneity into weak electron density. Ensemble modeling methods, which leverage molecular dynamics for sampling and use a different model representation may be more appropriate for examining these systems [45,46]. Future work, integrating the conformational heterogeneity of the protein, ligand, and water molecules will create better predictions and explanations of the energetics of binding. In addition, this would allow us to interpret the impact of specific interactions and alterations on both the entropy and enthalpy of all components of the system.

Our study, as well as previous NMR studies [7,9], only leverage a limited set of side chain dihedral angles. However, comparisons with molecular dynamics simulations suggest that small sets of side chain dihedrals alone may be representative of the overall changes in dynamics of the system [44748]. What is the thermodynamic impact of restricting side chain conformational heterogeneity? Protein folding studies and theory indicate that restricting the rotamer of even a single side chain can incur an entropic penalty of binding of ~0.5kcal/mol [49]. While we observe many such restrictions in binding sites due to specific interactions with ligands, our data point to corresponding changes away from the binding site that help balance this cost. Overall, the median increase in rigidity we observe in binding site residues (0.03 order parameter increase) would create an energetic penalty of approximately ~0.1-0.5kcal/mol, based off of the entropy meter calculated in Caro et al 2017 [6,7], with outliers having even larger

thermodynamic consequences. Given the constraints of maintaining a folded conformational ensemble upon ligand binding, it is likely that ligand binding generally acts to restrict degrees of freedom locally and that protein topological constraints favor increased motion in distal regions [24]. This overall effect likely manifests because optimizing affinity is desirable for medicinal chemistry and for the selective pressures experienced by many proteins. Such optimization is insensitive as to whether the free energy is driven enthalpically or entropically. However, given the attention paid to designing and optimizing enthalpic interactions, there is likely unleveraged potential in optimizing the entropic component as well. As more sophisticated models of conformational heterogeneity are created and validated [50] the strategy of rationally tuning conformational heterogeneity to improve binding affinity may be an attainable design strategy.

Methods

Dataset

Our dataset was compiled using a snapshot of the PDB[51] in September 2019, containing 156,187 structures. We then selected structures that had a resolution better or equal to 2Å (n= 64,557). We also excluded any structure that contained nucleic acids (n=2,280) or covalently bound ligands (n=1,030). We identified holo structures(n=30,530), defined as those that contained at least one ligand, defined as any HETATM residue with 10 or more heavy atoms, excluding common crystallographic additives.

To create apo/holo pairs, we took each holo structure and compared it to each potential apo structure (n=30,717), defined as structures without a ligand bound. A pair was defined according to the following criteria:

-same space group

-exact sequence or exact sequence after removing the first or last five base pairs of either structure

-a resolution difference between the two structures less than 0.1Å

-dimensions of unit cells do not differ by more than 1Å

-angles of the unit cells do not differ by more than 1 degree

This gave us 15,214 pairs. We then subsetted this list down to provide only one apo structure per holo structure, based on the smallest resolution difference. This produced a final pair set of 1,205 with 1,143 unique structures.

We also created a pairset with 458 unique apo/apo pairs using the same criteria as the apo/holo pairset.

Refinement

We re-refined all structures using phenix.refine (https://www.phenix-online.org/documentation/ reference/refinement.html). This was done using phenix version 1.17.1-3660. We performed anisotropic refinement on all pairs where both PDBs had a resolution better than 1.5Å. All other refinement was run isotropically. Refinement used the following parameters:

- Refine strategy: individual sites + individual adp + occupancies

- Number of macro cycles: 8
- NQH flips: True
- Optimize xyz weight: True
- Optimize adp weight: True
- Hydrogen refine: Riding

We removed 102 structures because of incompatibility with our re-refinement pipeline due to breaks in the protein chain or ligand incompatibility. We removed 88 structures where the R-free increased by >2.5% compared to the value reported in the PDB header (**Supplementary Figure 2.2**).

Running qFit

qFit-3.0[15] (version 3.2.0) was run using a composite omit map and the re-refined structure on the default parameters(https://github.com/ExcitedStates/qfit-3.0/). We ran qFit on Amazon Web Services (AWS). We used an auto scaling cluster of images controlled by the scheduler via ParallelCluser. Please see the qFit github for a script that will install qFit on AWS's default OS image, using conda to install its dependencies.

After qFit, we re-ran refinement as suggested by qFit-3.0. Briefly, this involves three rounds of refinement. The first refines coordinates only, the second goes through a cyclical round of refinement until the majority of low occupancy conformers are removed, and the third refinement polishes the structure, including hydrogens. The script used for post qFit refinement can be found here:

https://github.com/ExcitedStates/qfit-3.0/blob/master/scripts/post/qfit_final_refine_xray.sh. We removed 100 structures because of incompatibility with refinement after qFit rebuilding.

Quality Control

From our original dataset (n=1,205 pairs), we removed 28 apo structures that had a crystallographic additive or amino acid in the binding site that partially overlaid with the holo structure. We set a minimum ligand occupancy threshold of 0.15, which did not remove any pairs from our dataset. Chains were renamed according to the difference in distance between the two chains. We also re-numbered each chain based on the apo structure. We then superimposed the two structures using the pymol align function. We measured the alpha carbon root mean squared difference (RMSD) between the two structures as well as the difference in just binding site residues. Structures were removed if the mean RMSD of the entire structure was greater than 1Å or if the mean RMSD in the binding site residues were greater than 0.5Å. We removed two pairs based on these RMSD criteria.

We also assessed the difference in R-free values for each refinement step (before/after qFit). If the post refinement R-free value was 2.5% larger than the pre refinement R-value, then the structure was removed (n=85, 77 structures removed; **Supplement Figure 2.2**) . Additionally, we compared the final R-free values between apo and holo pairs, removing pairs with R-free values with more than a 5% difference (n=16 pairs removed; **Supplement Figure 2.2**). We ran the clashscore function out of Molprobity[52] to identify severe clashes in our dataset. We removed any structures with a clashscore

greater than 15, removing 52 structures. After filtering out pairs that failed our quality checks, our dataset contained 743 matched apo/holo pairs.

Alternative Conformations

Side chains were considered alternative conformers if there was at least one atom that was modeled with an alternative conformer. Our re-refinement procedure changes the occupancy, coordinates, and B-factors of these conformations, but it does not add or delete conformations.

B-factors

B-factors were assessed on a residue basis by averaging the B-factors of all heavy atoms for each residue. For residues with multiple conformations, we took the mean B-factor for all heavy atoms in each side chain, weighted by occupancy. For structures modeled anisotropically, we used the isotropic equivalent B-factor from phenix.

Root Mean Squared Fluctuation (RMSF)

RMSF was chosen over root mean squared deviation as many alternative conformers were predicted to have the same occupancy, thus making it difficult to define which was the main conformer. RMSF was measured for each residue based on all side chain heavy atoms. RMSF finds the geometric center of each atom in all alternative conformers. It then takes the distance between the geometric mean of each conformer's side chain heavy atoms and the overall geometric center. It then takes the squared mean of all of those distances, weighted by occupancy.

Order Parameters

Order parameters were measured for each residue (except proline and glycine) by taking into account both the angle of alternative conformers (s2angle), by measuring the chi1 angle, and the B-factors of alpha or beta carbons along with an attached hydrogen(s2ortho)[16]. To account for differences in B-factors as resolution changes, we investigated the correlation between order parameters in 32 apo lysozyme structures ranging in resolution from 1.1 to 2Å. We optimized the s2ortho parameter by looking for the normalization that would provide a slope closest to one and have the smallest root mean squared error (**Supplement Figure 2.1**). We normalized the s2ortho portion using the following equation:

s2orthonormalized=s2ortho*Bfactoralpha carbon/10(resolution)

The final order parameter reported in the paper is:

s2calc = s2orthonormalized*s2ang

Rotamer Analysis

Rotamers were determined using phenix.rotalyze[52] with manually relaxing the outlier criteria to 0.1%. Each alternative conformation has its own rotamer state. Rotamers were compared on a residue by residue basis between the holo and apo, taking into account each rotamer state for each alternative conformation. Residues were classified as “no change” if rotamers matched across the apo and holo residue, “distinct” if the matched residue shared no rotamer assignments. Residues were classified as “remodeled- holo loss” if distinct, additional rotameric conformations were populated in the apo residue only, and “remodeled - holo gain” if distinct, additional rotameric conformations were populated in the holo residue only.

Hydrogen Bond Analysis

To assess for the changes in hydrogen bonding across all pairs in our study, we applied HBplus[25] to every multiconfomer structure. HBplus identifies hydrogen bonds when the distance between the hydrogen and acceptor are less than 3.2 Å, with a maximum angle of 90 degrees. Since HBplus, nor any other software program we could identify, looks at hydrogen bonds in reference to alternative conformers, we split up each multiconformer PDB by alternative conformation. For example, the altA PDB contained all atoms that had an alternative conformer A as well as all atoms with no alternative conformation.

We then examined all of the hydrogen bonds for each PDB in binding site residues. We only considered hydrogen bonds between side chains or between side chains and the main chain. Hydrogen bonds were weighted based on the lowest occupancy of the acceptor or donor atom. We then controlled for the number of residues in the binding site.

Solvent Exposed Surface Area

We calculated the relative accessible surface area (RASA) using Define Secondary Structure of Proteins (DSSP)[53] with the Tien et al [54] definition of Max accessible surface area (MaxASA). Residues with a RASA of ≥20% were considered solvent exposed[55].

Ligand Analysis

We obtained the ligand properties using RDkit (version 2021.03.2) by importing SDF files of each ligand in our dataset. To account from the multiple hypothesis testing, we applied a Bonferroni correction, with an alpha of 0.05, as we were testing 10 hypotheses leaving us with a corrected significance value of 0.005.

Occupancy of the ligands were taken directly from the PDB file and correspond with the ligand occupancy from the deposited structure. Ligand B-factors were normalized using the mean alpha carbon B-factor of all residues in the structure.

If there were multiple ligands of interest in a structure, we looked at the properties of the ligand and surrounding protein residues in chain A or in the lowest alphabetical chain.

Protein Type Analysis

Protein names and enzyme names were extracted from Uniprot[56]. Names and properties were connected using PDB IDs.

Ringer Analysis

Protein Residues of interest were put through names and enzyme names were extracted from Uniprot[56]. Names and properties were connected using PDB IDs.

Statistics

Paired Wilcoxen test was used for all matched properties (comparing holo v. apo matched residues or structures). Individual Mann-Whitney U test was used for all non-match properties, including ligand properties. Two-sided t-test was used to compare the significance of the slopes.

Code

Code can be found in the following repositories:

-Dataset selection: https://github.com/stephaniewankowicz/PDB_selection_pipeline

-Refinement/qFit pipeline: https://github.com/stephaniewankowicz/refinement_qFit

-Analysis/Figures: https://github.com/fraser-lab/Apo_Holo_Analysis

-qFit: https://github.com/ExcitedStates/qfit-3.0.

References

1. Mobley, D. L. & Dill, K. A. Binding of small-molecule ligands to proteins: ‘what you see’ is not always ‘what you get’. Structure 17, 489–498 (2009).

2. Boehr, D. D., Nussinov, R. & Wright, P. E. The role of dynamic conformational ensembles in biomolecular recognition. Nat. Chem. Biol. 5, 789–796 (2009).

3. Gutteridge, A. & Thornton, J. Conformational changes observed in enzyme crystal structures upon substrate binding. J. Mol. Biol. 346, 21–28 (2005).

4. Wand, A. J. & Sharp, K. A. Measuring Entropy in Molecular Recognition by Proteins. Annu. Rev. Biophys. 47, 41–61 (2018).

5. Tzeng, S.-R. & Kalodimos, C. G. Protein activity regulation by conformational entropy. Nature 488, 236–240 (2012).

6. Caro, J. A., Valentine, K. G. & Wand, A. J. Structural origins of protein conformational entropy. bioRxiv 2021.02.12.430981 (2021) doi:10.1101/2021.02.12.430981.

7. Caro, J. A. et al. Entropy in molecular recognition by proteins. Proc. Natl. Acad. Sci. U. S. A. 114, 6563–6568 (2017).

8. Zhou, H.-X. & Gilson, M. K. Theory of free energy and entropy in noncovalent binding. Chem. Rev. 109, 4092–4107 (2009).

9. Frederick, K. K., Marlow, M. S., Valentine, K. G. & Wand, A. J. Conformational entropy in molecular recognition by proteins. Nature 448, 325–329 (2007).

10. Woldeyes, R. A., Sivak, D. A. & Fraser, J. S. E pluribus unum, no more: from one crystal, many conformations. Curr. Opin. Struct. Biol. 28, 56–62 (2014).

11. Cheng, Y., Grigorieff, N., Penczek, P. A. & Walz, T. A primer to single-particle cryo-electron microscopy. Cell 161, 438–449 (2015).

12. Kuzmanic, A., Pannu, N. S. & Zagrovic, B. X-ray refinement significantly underestimates the level of microscopic heterogeneity in biomolecular crystals. Nat. Commun. 5, 3220 (2014).

13. Kuriyan, J., Petsko, G. A., Levy, R. M. & Karplus, M. Effect of anisotropy and anharmonicity on protein crystallographic refinement. An evaluation by molecular dynamics. J. Mol. Biol. 190, 227–254 (1986).

14. van den Bedem, H. & Fraser, J. S. Integrative, dynamic structural biology at atomic resolution--it's about time. Nat. Methods 12, 307–318 (2015).

15. Riley, B. T. et al. qFit 3: Protein and ligand multiconformer modeling for X-ray crystallographic and single-particle cryo-EM density maps. Protein Sci. 30, 270–285 (2021).

16. Fenwick, R. B., van den Bedem, H., Fraser, J. S. & Wright, P. E. Integrated description of protein dynamics from room-temperature X-ray crystallography and NMR. Proc. Natl. Acad. Sci. U. S. A. 111, E445–54 (2014).

17. Berman, H. M. et al. The Protein Data Bank. Nucleic Acids Res. 28, 235–242 (2000).

18. Liebschner, D. et al. Macromolecular structure determination using X-rays, neutrons and electrons: recent developments in Phenix. Acta Crystallogr D Struct Biol 75, 861–877 (2019).

19. Lang, P. T. et al. Automated electron-density sampling reveals widespread conformational polymorphism in proteins. Protein Sci. 19, 1420–1431 (2010).

20. Gaudreault, F., Chartier, M. & Najmanovich, R. Side-chain rotamer changes upon ligand binding: common, crucial, correlate with entropy and rearrange hydrogen bonding. Bioinformatics 28, i423–i430 (2012).

21. van den Bedem, H., Dhanik, A., Latombe, J. C. & Deacon, A. M. Modeling discrete heterogeneity in X-ray diffraction data by fitting multi-conformers. Acta Crystallogr. D Biol. Crystallogr. 65, 1107–1117 (2009).

22. Wong, K.-B. & Daggett, V. Barstar Has a Highly Dynamic Hydrophobic Core: Evidence from Molecular Dynamics Simulations and Nuclear Magnetic Resonance Relaxation Data†. Biochemistry vol. 37 11182–11192 Preprint at https://doi.org/10.1021/bi980552i (1998).

23. Moorman, V. R., Valentine, K. G. & Wand, A. J. The dynamical response of hen egg white lysozyme to the binding of a carbohydrate ligand. Protein Sci. 21, 1066–1073 (2012).

24. Bromberg, S. & Dill, K. A. Side-chain entropy and packing in proteins. Protein Sci. 3, 997–1009 (1994).

25. McDonald, I. K. & Thornton, J. M. Satisfying hydrogen bonding potential in proteins. J. Mol. Biol. 238, 777–793 (1994).

26. Wicker, J. G. P. & Cooper, R. I. Will it crystallise? Predicting crystallinity of molecular materials. CrystEngComm vol. 17 1927–1934 Preprint at https://doi.org/10.1039/c4ce01912a (2015).

27. Olsson, T. S. G., Williams, M. A., Pitt, W. R. & Ladbury, J. E. The thermodynamics of protein-ligand interaction and solvation: insights for ligand design. J. Mol. Biol. 384, 1002–1017 (2008).

28. Bissantz, C., Kuhn, B. & Stahl, M. A medicinal chemist's guide to molecular interactions. J. Med. Chem. 53, 5061–5084 (2010).

29. Majewski, M., Ruiz-Carmona, S. & Barril, X. An investigation of structural stability in protein-ligand complexes reveals the balance between order and disorder. Communications Chemistry vol. 2 Preprint at https://doi.org/10.1038/s42004-019-0205-5 (2019).

30.van Zundert, G. C. P. et al. qFit-ligand Reveals Widespread Conformational Heterogeneity of Drug-Like Molecules in X-Ray Electron Density Maps. J. Med. Chem. 61, 11183–11198 (2018).

31. Bhat, T. N. Correlation between occupancy and temperature factors of solvent molecules in crystal structures of proteins. Acta Crystallogr. A 45 (Pt 1), 145–146 (1989).

32. Carugo, O. Correlation between occupancy and B factor of water molecules in protein crystal structures. Protein Eng. 12, 1021–1024 (1999).

33. Kim, J. et al. A dynamic hydrophobic core orchestrates allostery in protein kinases. Sci Adv 3, e1600663 (2017).

34. Wang, Y. et al. Globally correlated conformational entropy underlies positive and negative cooperativity in a kinase's enzymatic cycle. Nature Communications vol. 10 Preprint at https://doi.org/10.1038/s41467-019-08655-7 (2019).

35. Gohlke, H., Kuhn, L. A. & Case, D. A. Change in protein flexibility upon complex formation: analysis of Ras-Raf using molecular dynamics and a molecular framework approach. Proteins 56, 322–337 (2004).

36. Fraser, J. S. et al. Accessing protein conformational ensembles using room-temperature X-ray crystallography. Proc. Natl. Acad. Sci. U. S. A. 108, 16247–16252 (2011).

37. Diehl, C. et al. Protein flexibility and conformational entropy in ligand design targeting the carbohydrate recognition domain of galectin-3. J. Am. Chem. Soc. 132, 14577–14589 (2010).

38. Steuber, H. et al. Expect the Unexpected or Caveat for Drug Designers: Multiple Structure Determinations Using Aldose Reductase Crystals Treated under Varying Soaking and Co-crystallisation Conditions. Journal of Molecular Biology vol. 363 174–187 Preprint at https://doi.org/10.1016/j.jmb.2006.08.011 (2006).

39. Jhaveri, K. et al. The evolution of cyclin dependent kinase inhibitors in the treatment of cancer. Expert Rev. Anticancer Ther. (2021) doi:10.1080/14737140.2021.1944109.

40. Wood, D. J. & Endicott, J. A. Structural insights into the functional diversity of the CDK-cyclin family. Open Biol. 8, (2018).

41. Breiten, B. et al. Water networks contribute to enthalpy/entropy compensation in protein-ligand binding. J. Am. Chem. Soc. 135, 15579–15584 (2013).

42. Verteramo, M. L. et al. Interplay between Conformational Entropy and Solvation Entropy in Protein–Ligand Binding. Journal of the American Chemical Society vol. 141 2012–2026 Preprint at https://doi.org/10.1021/jacs.8b11099 (2019).

43. Jain, A. N. et al. XGen: Real-Space Fitting of Complex Ligand Conformational Ensembles to X-ray Electron Density Maps. J. Med. Chem. 63, 10509–10528 (2020).

44. Caldararu, O., Ekberg, V., Logan, D. T., Oksanen, E. & Ryde, U. Exploring ligand dynamics in protein crystal structures with ensemble refinement. Acta Crystallogr D Struct Biol 77, 1099–1115 (2021).

45. Burnley, B. T., Afonine, P. V., Adams, P. D. & Gros, P. Modelling dynamics in protein crystal structures by ensemble refinement. Elife 1, e00311 (2012).

46. Eshun-Wilson, L. et al. Effects of α-tubulin acetylation on microtubule structure and stability. Proc. Natl. Acad. Sci. U. S. A. 116, 10366–10371 (2019).

47. Kasinath, V., Sharp, K. A. & Wand, A. J. Microscopic insights into the NMR relaxation-based protein conformational entropy meter. J. Am. Chem. Soc. 135, 15092–15100 (2013).

48. Chatfield, D. C. & Wong, S. E. Methyl Motional Parameters in Crystalline l-Alanine: Molecular Dynamics Simulation and NMR. The Journal of Physical Chemistry B vol. 104 11342–11348 Preprint at https://doi.org/10.1021/jp0018089 (2000).

49. Doig, A. J. & Sternberg, M. J. Side-chain conformational entropy in protein folding. Protein Sci. 4, 2247–2251 (1995).

50. Rosenbaum, D. et al. Inferring a Continuous Distribution of Atom Coordinates from Cryo-EM Images using VAEs. (2021).

51. Berman, H. Protein Data Bank Project at Rutgers University. Preprint at https://doi.org/10.2172/805813 (2002).

52. Williams, C. J. et al. MolProbity: More and better reference data for improved all-atom structure validation. Protein Sci. 27, 293–315 (2018).

53. Kabsch, W. & Sander, C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22, 2577–2637 (1983).

54. Tien, M. Z., Meyer, A. G., Sydykova, D. K., Spielman, S. J. & Wilke, C. O. Maximum allowed solvent accessibilites of residues in proteins. PLoS One 8, e80635 (2013).

55. Wu, W., Wang, Z., Cong, P. & Li, T. Accurate prediction of protein relative solvent accessibility using a balanced model. BioData Min. 10, 1 (2017).

56. UniProt Consortium. UniProt: a hub for protein information. Nucleic Acids Res. 43, D204–12 (2015).

Chapter 3

Making sense of the chaos: uncovering the mechanisms of conformational entropy

Stephanie A. Wankowicz[1,2], James S. Fraser[1]

1) Department of Bioengineering and Therapeutic Sciences, University of California, San Francisco, San Francisco, CA 94158, USA.

2) Biophysics Graduate Program, University of California San Francisco, San Francisco, CA 94158, USA.

Preface

The bulk of this chapter will appear as Wankowicz & Fraser in 2023.

Abstract

Protein folding converts a disordered polymer to a globular structure, reducing many conformational degrees of freedom and incurring a significant conformational entropy penalty. However, residual entropy is retained through the motion of the protein between different related conformations that define its native state, often referred to as the conformational ensemble. Subtle changes in protein motion, mostly from sidechains, can alter this residual conformational entropy, leading to differences in binding affinity and allosteric communication. While NMR has provided measurements of conformational entropy, these measurements do not provide information on where this entropy is coming from, such as if this is coming from a sidechain moving harmonically or anharmonically. Extracting this information from molecular simulations is currently impossible as the timescale of protein motion is beyond the timescale of molecular simulations. However, we can take advantage of the fact that X-ray crystallography and CryoEM experimental data capture the conformational ensemble allowing us to measure the motion of residues and their atomistic structure. This provides an unparalleled platform to answer how, where, and why conformational entropy. A mechanistic understanding of conformational entropy will help fill in the gaps on an often-forgotten dimension of biological control and function, leading to improved explanations of allostery and binding events.

The Problem of Entropy and Binding

During protein folding, proteins go from a disordered state, where the polymer access an almost infinite number of states, to its native state, where the access to the majority of states is restricted by steric clashes, especially in backbone atoms[1]. As a result, the protein loses a significant amount of entropy, as defined by the Boltzmann entropy equation [$S=k_B \ln(W)$][2]. However, sidechains can still have a large number of semi-independent conformations, accounting for a significant amount of remaining conformational entropy within the native state. The transition from an unfolded to folded protein involves a reduction of states, which is observed by narrowing a free energy landscape (**Figure 3.1**). Even though the free energy landscape narrows upon folding, there is still considerable width, representing the conformational ensemble of a folded protein.

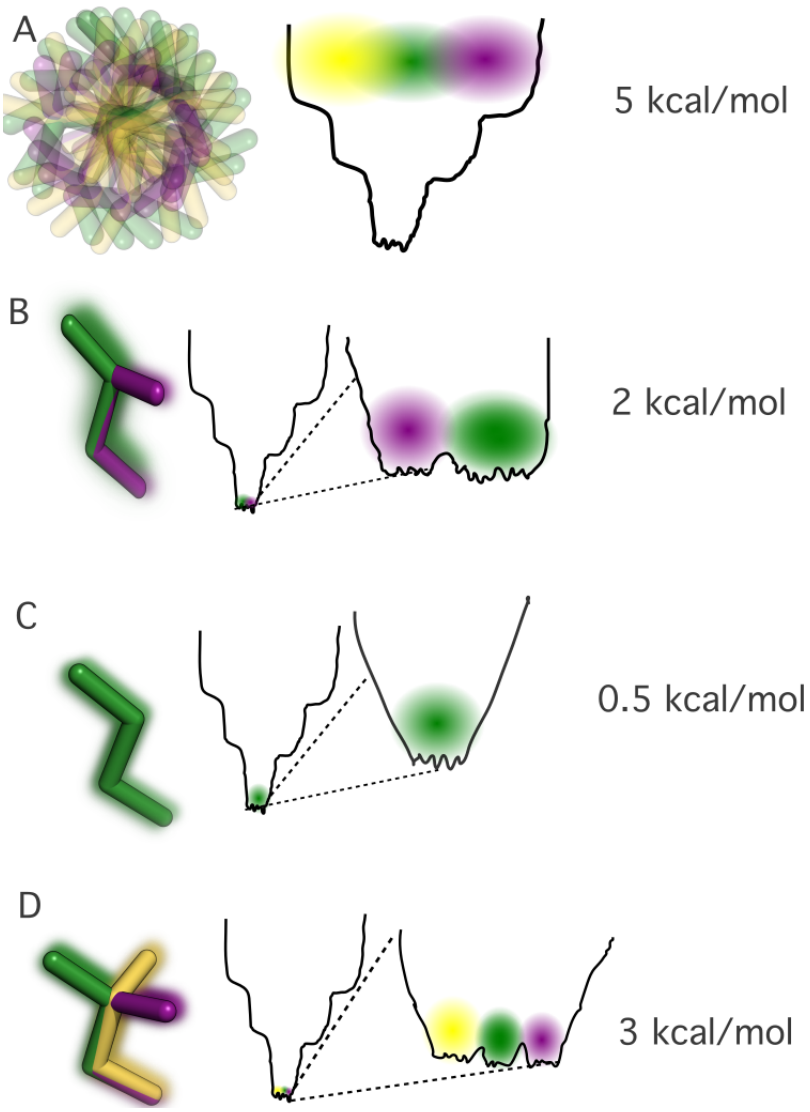


Figure 3.1| Conformation and energy landscape of serine residue. A. Serine residue in an unfolded protein can be in any orientation. It's location on an energy landscape is towards the top and can traverse the entire landscape. B. Upon folding, the serine residue can only be in two conformations, significantly reducing its conformational entropy.

To illustrate the concept of sidechain conformational entropy, consider a single serine sidechain in a polypeptide. In the unfolded state, the serine sidechain can equally access all three rotamer conformations, and the backbone is only restricted linearly (**Figure 3.1**). Upon folding, the serine backbone is reduced to only one conformation, with some harmonic motion, due to steric clashes. However, the serine sidechain may

still have access to multiple conformations, as this small, slightly polar amino acid can happily interact with many neighboring residues. It is likely due to steric clashes with surrounding amino acids that the sidechain will lose access to one or more conformations, leading to a reduction, but not the elimination, of conformational entropy in the folded state. In a free energy landscape, the loss of a potential serine conformation is observed as the loss of that energy well, making it much less likely for that conformation to exist. However, the sidechain contains conformational entropy through being able to anharmonically move between conformations, represented as the movement between energy wells in the free energy landscape. The sidechains can also move harmonically about each confirmation, represented by the width of free energy wells. Further, if the two remaining conformations are not equally likely, this will increase the conformational entropy due to the uncertainty of position. Entropy, as defined by the Boltzmann equation, increases with more microstates. If two conformations have equal probability, you only need to have two microstates to described their distribtuion. However, if the two conformations are unequal in their likelihood, you would need more than two microstates to describe their distribution, increasing the entropy in the system.

Perturbation to proteins, including macromolecular binding and mutations, alters the protein's free energy landscape by manipulating energy well depth and width within the folded protein state. From our serine example, ligand binding may lead to the elimination of one of the two remaining sidechain conformations (**Figure 1**). This would represent a perturbation that significantly decreases a well's depth, making it difficult for the serine to access one of its conformations. This is structurally observed as a conformational change[4], but is also related to a reduction in conformational entropy.

Alternatively, ligand binding may cause subangstrom movement of ligand-interacting sidechains towards the ligand, resulting in a void[3] around our serine sidechain, increasing the harmonic motion of the sidechains and increasing conformational entropy (**Figure 1**). This creation of voids or pockets of areas where sidechains can increase their harmonic or anharmonic motion provides an energy reservoir with the native fold. On a free energy landscape, this would result in an alteration of the width of the free energy basin[5]. With the widening of a free energy basin, the overall structure may look similar, but the average subangstrom movement of sidechains increases, leading to an increase in conformational entropy. This energy reservoir may counteract entropy reduction elsewhere or provide energy for other protein functions, including macromolecular binding, signaling, macromolecular machines, and catalysis.

As stated above, a perturbation can have unique effects on different areas of the protein. This is based on the uneven redistribution of sidechain motions upon perturbation. This redistribution can result in different energetic binding properties but can also result in allosteric communication. Going back to our serine example, ligand binding likely alters the motion of each serine conformation differently. It may be possible that only one serine sidechain conformation increases in motion, leading to a directional propagation of motion of neighboring sidechain[6], which will further be impacted by residue type[7]. It is critical to evaluate how, why, and where changes in conformational entropy arise within a protein, as this can determine its functional effects[8].

The interplay between conformational entropy, structure, and function is complex. Over 40 years ago, Cooper & Dryden theoretically postulated that protein thermal fluctuations

could provide energy for macromolecular binding and other functions[5]. We have only recently been able to experimentally measure this, highlighting the ensemble nature of allostery and the ability of proteins to contain a plurality of allosteric mechanisms[9]. However, to complete Cooper & Dryden's theory, we must understand how the many potential atomistic motions of proteins lead to differences in conformational entropy and protein function.

Ways of Measuring Ensembles

Nuclear magnetic resonance (NMR) relaxation techniques have measured conformational entropy and correlated it with protein functions[1011]. Conformational entropy is measured using order parameters, which provide a site-specific measurement of the degree of motion of the NH or methyl groups on the picosecond-nanosecond timescale[12,13]. NMR order parameters have been quantitatively linked with conformational entropy, demonstrating changes in methyl order parameters correlating with changes in experimentally measured protein conformational entropy. These studies showed that the motion of sidechain atoms, but not backbone atoms, holds significant entropy. However, these entropy estimations degenerate in relation to the structural model and ignore motions that occur on slower timescales. Order parameters lack information about the directions and extent of these motions, dampening our understanding of the mechanism of conformational entropy. Molecular dynamics (MD) can provide insight into protein motions, but their timescale is computationally limited. Further, small shifts in populations or the motion of single atoms can be overshadowed by the large dimensionality of most systems. Connecting residue motion to the quantitative effects of conformational entropy is critical.

Ensemble-based structural techniques, X-ray crystallography and CryoEM, capture protein conformations from millions to trillions of molecules. Models are usually created by placing atoms in their mean atomic positions and assigning a B-factor to capture motion. However, the underlying data contains information on conformations in all the captured molecules. Due to the low signal-to-noise ratio, decoding all these conformations is challenging. Nevertheless, new computational techniques can model more conformations, leading to an atomistic understanding of protein conformation entropy. The modeling of the protein conformational ensemble from X-ray or CryoEM allows for the connection of the atomic properties of residues, including positions and movements, to free energy landscapes to elucidate the macroscopic behavior of proteins.

Recently, we demonstrated how to use multiconformer modeling to extract quantitive information on conformational entropy from thousands of cryogenic X-ray structures. This allows us to extract information on the directions and types of motions that increase or decrease conformational entropy. By examining over 700 paired bound and unbound structures, we demonstrated that distant residues tend to become more flexible when binding site residues become more rigid upon ligand binding, distant residues tend to become more flexible[14]. This indicated that entropic compensation might be a widespread phenomenon.

Increasingly model building algorithms for X-ray crystallography and Cryo-EM take advantage of the ensemble nature of the data. The output of these models can help provide an atomistic explanation of the mechanisms of conformational entropy, which is essential to understand the fundamentals of allostery. This explanation is key to

understanding how protein function changes upon perturbation or in different environments. Here, we present how these new algorithms may be applied to answering some of the open questions of the mechanisms of conformational entropy.

Examples of how entropy influences binding

Protein conformational entropy correlates with binding entropy[15,16]. However, what determines how and where conformational changes are upon binding is still unknown, limiting our mechanistic explanation. Both protein and ligand properties have demonstrated different effects on conformational entropy, highlighting the complexity of untangling the impact of conformational entropy. Further, how conformational entropy interplays with binding enthalpy still needs to be discovered. Examining these problems with models that can connect residue locations and motion to conformational entropy has the potential to answer many of these outstanding questions.

Ligand and protein alterations can impact conformational entropy changes upon binding. In PDZ domains, found in proteins with diverse functions, the truncation of an alpha helix 3 (α3), 6 angstroms away from the binding site, reduces peptide binding affinity by 21 fold[17,18]. α3 truncation does not change the structure of PDZ but increases sidechain flexibility, indicating that binding affinity difference is due to conformational entropy changes. PDZ domain conformational entropy and binding affinity may also depend on other domains[17], indicating that proteins can tune the conformational entropy of distant domains, manipulating their binding affinity. In Calmodulin, peptides can increase or decrease calmodulin's conformational entropy[16], indicating that substrate properties also impact conformational entropy.

From these experiments, is unclear how the truncation of α3 or the binding of peptides causes an increase in conformational entropy. While we know that neither of these events leads to a complete unfolding of the protein, how sidechains increase their conformational entropy is unknown. The truncation of α3 may lead to increased volume around many residues within their existing conformations due to the lack of α3 packing the rest of the protein. It is also possible that the lack of α3 leads to the new conformation of many sidechains, potentially disrupting hydrogen bonds and salt bridges, leading to a lack of stability within the binding site.

These hypotheses could be probed using qFit, an automated and parsimonious multiconformer modeling software[19,20]. Multiconformer modeling places all or part of a residue into multiple positions with corresponding occupancies, as supported by the electron density data. We can then use these models to calculate quantitive backbone and sidechain conformational heterogeneity for every residue, similar to NMR order parameters[21]. Multiconformer models simultaneously provide information on the residues' position and their conformational entropy. Further, we can use these models to create contact networks between residues to hypothesize how proteins can transfer their conformational entropy to perform their functions[22].

Conformational entropy may also drive allosteric communication during macromolecular binding. In kinases, allosteric binding cooperativity occurs between the nucleotide and substrate binding site, with PKA demonstrating positive cooperation[23] and SRC demonstrating negative cooperation[24]. In PKA and SRC, conformational entropy is

thought to drive this cooperation. In both cases, inhibitors with similar binding affinities and similar overall structures have drastically different changes in NMR order parameters, correlating with differences in substrate binding affinities. The inhibitor and substrate binding sites are connected through the kinase's hydrophobic center, where structural features, including R and C-spine, are located[25]. If conformational entropy drives the cooperativity effects and goes through the hydrophobic center, how do changes in conformational entropy interact with these structural features? We could investigate these questions using multiconformer modeling, allowing us to observe how specific interactions with ligands or kinase structural features interplay with conformational entropy changes. Given that there are 100s of kinase structures, we could apply Xtrapol8 to extract positions and occupancies of rare conformational by comparing an excited and base dataset[26]. Xtrapol8 creates weighted difference maps to visualize rare conformations unbiasedly. It then creates optimally weighted extrapolated structure factors and allows for real or reciprocal space refinement of the excited state. Finally, it provides estimates of the occupancies of the excited states. This method produces results similar to the map-deconvolution algorithm, PANDDA, which identifies and models low occupancy ligands and ligand fragments[26,27]. This modeling may identify the residues' motion in the kinase's core and how these motions interact with the R or C spine. It could also investigate differences in contact networks, which may provide clues as to where and why conformational entropy is changing[22].

Additionally, given the large number of kinase structures avaliable, we could also create protein pseduoensembles of PKA or SRC[28]. Pseudoensembles have uncovered

conformational clusters, highlighting how ligands perturb the conformational landscape and tend to bind or select different major protein conformations[29–32]. Further analysis of the motion within each state can still be probed further using multiconformer modeling. We can combine pseudoensembles and multiconformer modeling, using qFit, to assess for cooperative structural effects between the major conformational changes and the conformational entropy of the protein. These methods would improve our estimates of residue populations, leading to improved mechanistic hypotheses. Combining these modeling methods may uncover the interplay between major and minor changes in protein conformations[33].

Human thymidylate synthase also displays positive cooperativity in the dual binding of dUMP, its native nucleotide substrate[34]. Positive cooperativity is driven by a reduction in sidechain conformational entropy during the first dUMP binding event, allowing the second dUMP to bind with no entropic cost. This represents about 10kcal/mol energetic driver for the second binding effect. This energetic driver is incredibly specific as the binding of thymidylate synthase's product, which only differs from dUMP by one methyl group, does not induce this change in entropy[35]. This specificity was also observed in peptides binding to calmodulin[16]. However, what is driving the difference in conformational entropy upon extremely similar substrates is unclear. Different specific interactions likely lead to differences in volume elsewhere in the protein. However, NMR cannot capture how specific interactions lead to these volume differences. It could be possible that a residue with multiple conformations in the unbound state can now only be in one conformation due to clashing with the ligand and can then trap cascading

residues in a more confined space, reducing conformational entropy. Nevertheless, that clash may not exist with a different substrate, allowing many residues to keep the same conformational entropy. By applying multiconformer modeling to these structures, we may uncover how the different interactions of two highly similar ligands produce drastic differences in conformational entropy.

Entropy-driven cooperativity is also frequently observed in DNA-binding proteins. Protein-DNA complexes are often dynamic, potentially due to their large search space[36] or to compensate for the cost of burying polar interfaces. There is evidence from multiple proteins that metal ion binding at a distant site changes DNA binding affinity by reducing its conformational entropy [37–40]. This is likely due to more specific interactions between multiple residues and the metal ion, but we need ensemble-based models to confirm this hypothesis.

Conformational entropy may also impact larger conformational changes. Upon phosphorylation in CheY, a chemotaxis response regulator, in binding site residues decreased flexibility, but flexibility, specifically, fast motion, increased along its allosteric pathways[41]. The increase in fast motion flexibility correlated with a decrease in larger conformational changes. However, how this flexibility gets propagated and why small fast motion disrupts the ability for larger conformational changes still needs to be discovered. The large conformational change may need more correlated motion, which may be too energetically unfavorable when residues are moving quickly between two or more conformations. It is also possible that one or two residues sterically block the

ability for the large conformation to occur.

As X-ray crystallography can detect motions across timescales, we can capture both types of motions observed in CheY. To tease out the interplay of these motions, the extensible-component hierarchical TLS (ECHT) B-factor model models atomic disorder on multiple levels to more accurately capture the different length scales of motion[42,43]. This modeling can be applied to time-averaged ensemble refinement, which uses MD, restrained by a time-averaged agreement with the X-ray structure factors, to generate multiple models[44]. ECHT-based refinement may show how changes in conformational entropy are necessary for the conformational switch in CheY.

Examples of how entropy influences catalysis

Enzyme catalysis requires precise positions of catalytic residues while maintaining enough flexibility to allow the reaction to proceed. Subangstrom changes in the position or motions of residues can completely erase the catalytic efficacy of an enzyme. Human histidine triad nucleotide-binding (hHINT) proteins catalyze nucleotide phosphoramidase and acyl-phosphatase reactions. In hHINT1, when a surface glutamine 13 angstroms from the active site is mutated to alanine, it significantly impacts rate constants. However, the structure had no changes in positions in water structures or residues between this residue and the active site. However, protein residues increased their motion between the surface alanine and active site. In another ‘hint ’at the specificity of conformational entropy, the arginine residue in the same location in hHINT2, which is highly homologous to hHINT1, is mutated to alanine, there is no change in rate

constants[45]. In ketosteroid isomerase, room temperature X-ray structures and functional studies emphasized the importance of the probability of different positions of residues[29], which can only be identified through multiconformer models. These models may be able to capture the elusive transition-state structures. While there has been significant progress in capturing these rare states using time-resolved cryo-EM and X-ray crystallography[46–49], it may be possible to extract this information from multiconformer models, especially with room temperature or multitemperature models[50].

Examples of how entropy influences Molecular Machines

Many large protein complexes drive essential protein functions by acting as biomolecular machines. These complexes must harness chemical and thermodynamic energy to drive DNA replication, RNA transcription, and protein synthesis. While some molecular machines use ATP or GTP to complete their functions, others do not, indicating they harness their limited motions to complete these arduous functions. Either way, entropic reservoirs likely help some complexes complete their functions. For example, ribosomes undergo large conformational changes during translation, driven by GTP hydrolysis. When GTP hydrolysis occurs, it causes spontaneous thermal fluctuations of the ribosome, likely increasing conformational entropy[51]. Single-molecule experiments demonstrated that this motion is more coordinated both through pre-existing structural elements in the ribosomes but also by tRNA[52]. This helps with tRNA binding, allowing the entropic motion of the ribosome to overcome the enthalpic penalty of binding tRNA, enabling faster transitions between ribosome states[53]. This phenomenon is likely not unique to the ribosome but has been impossible to probe

using NMR. With high-resolution and time-resolved cryo-EM, we can observe how binding events change the conformational entropy of large molecular machines and how these machines likely evolved to take advantage of their entropy reservoir.

Examples of how mutations influence entropy

Mutations also impact conformational entropy. This may help explain why mutations distant from binding sites impact the affinity of ligands or proteins. Single residue mutations can also change protein conformational entropy with varying impacts. In adenylate kinase[8], alanine to glycine mutation in a solvent-exposed residue far from the binding site changed the relative substrate binding affinity by ~0.5kcal/mol. A similar mutation in a different domain changed the turnover rate, not the binding affinity highlighting how location is critical to the impact conformational entropy can have. In DNA-binding proteins, mutations outside the DNA-binding interface can change the conformational and binding entropy[40]. Further, in a designed protein based on streptococcal protein G domain β1, the leucine to valine mutation prevents the local unfolding of a helix[54]. However, the same mutation in another residue does not result in a change in flexibility. Why two similar mutations impact protein dynamics drastically differently is still unclear. All of these mutations are hypothesized to cause local unfolding. Nevertheless, it is unclear what local unfolding looks like. For example, local unfolding could cause an increase in the motions of all residues while retaining the same rotamer status, or it could create a void where one or two residues now flip between two or more rotamer well. Uncovering which of these potential situations are true will help identify how local unfolding leads to drastically different effects. Modeling

the mutation as multiconformers may uncover how the positions of these residues lead to differences in motion. It is also possible to model protein motions by describing bonds rather than atomic positions[55]. This allows a non-linear representation of the protein conformational space, highlighting correlated, subtle conformational entropy changes throughout the structure[56]. We could also measure the packing entropy, which determines how much space the protein can move locally, using the algorithm PACKMAN[57]. This may uncover how residue neighborhoods lead to packing and conformational entropy differences.

Mutations can also impact protein-protein binding. The Spike protein of SARS-CoV2 is constantly acquiring mutations to increase its binding affinity to ACE2. While most mutations have increased the enthalpy, a recent mutation on the receptor binding domain, N501Y, had more favorable binding with increasing temperature[58]. MD simulations showed that N501Y increases the motion of sidechain residues in both the RBD and ACE2[59], without impacting the binding surfaces between the two proteins. It was predicted that an increase in conformational entropy drives the favorable binding of these mutations. The impact of conformational entropy becomes even more complex when exploring potential other perturbations of proteins, including post-translational modifications. On top of mutations impacting conformational entropy in the RBD, glycans on ACE2 impact the binding affinity of the SARS-CoV2 Spike protein due to reducing the entropy of the glycans[60]. However, as glycan flexibility depends on their environment[61], different glycans may have different entropic impacts, making this a tricky thing to predict.

Membrane proteins

Membrane proteins are the cell's sensors to the outside world. These large proteins bind small ligands outside cells and facilitate different functions inside the cell, sometimes transmitting that signal over 60 angstroms. Conformational entropy changes ligand binding likely determines this allosteric communiciation[62]. In neurotensin receptor 1, a prototypical peptide-binding GPCR, orthosteric agonists and antagonists rigidify NST1 to different degrees, potentially leading to their pharmacological difference. This highlights a potential role for conformational entropy in GPCR ligand discrimination. As other examples have shown, designing a ligand for a rigid protein state may result in decreased free energy compared to designing a ligand for a more flexible state of the protein.

Solvent and Ligand Entropy

Solvent entropy is also intertwined with protein conformational entropy. While solvent entropy is often estimated by measuring solvent-accessible surface area differences, this is a crude estimation. Additionally, there is evidence that solvent networks can allosterically communicate with distant parts of the protein.

Ligand and solvent interactions change the distribution of conformational entropy[63,64]. Many ligands can bind in multiple orientations, with different poses impacting the protein differently[65]. Multiple ligand poses and their relative occupancies can sometimes be difficult to detect. However, programs such as qFit Ligand can identify conformationally

averaged ligand poses in X-ray and CryoEM structures[66]. If a ligand can bind in multiple poses, the ligand's entropy is not reduced to zero from the theoretical high entropy state within solvent, potentially increasing the entropy of binding. Additionally, the different poses of the ligand may have unique impacts on the protein conformational entropy.

Solvent molecules also contribute significantly to the system's entropy [64]. It has been suggested that solvent motion can also act as an allosteric pathway, allowing for the transmission of signals throughout a protein[63]. Moreover, while there is likely a significant interplay between the solvent's entropy and the protein's conformational entropy, protein conformational entropy is not tied to solvent motion[67].

Intrinsically Disordered Regions

Intrinsically disordered regions (IDRs) contain a large amount of conformational entropy, as they do not have a defined native or folded state. However, many IDRs can become ordered in certain scenarios. Further, different IDRs have differing amounts of conformational entropy depending on the transient and weak intramolecular interactions they have. This conformational entropy can be significantly reduced through interactions with the rest of the protein, including tethering, or becoming ordered upon interaction with another macromolecule.

Future Directions

While our ability to model conformational heterogeneity is improving, there are still significant gaps in translating these methods into biological insights. First, while we

presented multiple ways to model conformational entropy from X-ray or cryo-EM data, it is unclear how these motions translate into energetics. To provide a complete picture of the mechanisms of conformational entropy, we must be able to relate the observables in structural studies with the experimentally measured entropy. Additionally, most measurements only capture the first chi angle, which only partially accounts for conformational entropy.

The modeling discussed here was focused on cryogenically cooled X-ray crystallography. However, cooling can restrict protein motion[68,69]. Room-temperature X-ray crystallography detects higher-energy protein conformations[70,71], with X-ray free-electron lasers reducing room temperature related radiation damage[72]. Further, time-resolved and temperature jump experiments detect conformational states that change over time[49,73]. The growing ease of detecting the position of hydrogens using neutron crystallography will lead to better estimates of binding affinity or turnover rates due to our ability to detect critical hydrogens[74,75]. Developing computational tools to enable new biological discoveries is increasingly vital.

Computational methods are also constantly improving. We see the most necessary gains in developing methods to estimate the independence or correlation of motions on different timescales and between residues. In cryo-EM, flexible refinement can identify different conformations of proteins, but these different conformations are often large changes[76]. We encourage the creation of mixed ensemble-based and multiconformer models, which may identify the connection between conformational entropy and large

conformational changes. Methods for sampling conformations or positions of residues can also be improved. Creating probabilistic diffusion models could also be used to identify different residue or loop conformations[77–79].

Finally, we need validation metrics and comparisons between different experimental methods. How do order parameters compare between X-ray or cryoEM and NMR? Are differences due to restrictions in the crystals or the time restraints in NMR? It may also be possible to use MD simulations to help determine how low-probability events look in diffraction patterns or electron density[80,81]. We also need to establish new data depositions methods to enable data sharing. While there has been significant progress with fragment depositions[82], it still needs to be determined how we can track or deposit structure re-modeled as multiconformer or ensemble models[83].

Open lines of inquiry

This review aims to represent the many roles conformational entropy plays in biological function and improved modeling many answers how conformational entropy impacts function. Beyond the examples above, other open lines of inquiry may be viewed in the light of conformational entropy.

Many examples above explore how proteins can regulate function or react to new perturbations. It is crucial to assess the contributions of conformational entropy and its interaction with enthalpic interactions[84]. One possible way to make proteins more sensitive to their environment is to detect new "sensors," such as ligands, which adapt

the protein to perform certain functions. However, using traditional allostery models would involve creating or modifying the binding site to allow ligands to bind and then creating a specific path to their effectors. However, if allosteric regulation can occur due to changes in the dynamics of the protein, this provides an easier way for new functions to evolve, allowing the protein to respond to new stimuli. Further, pluripotent allostery, where ligands depend on external factors, such as metabolic and proteomic concentration, may act through slight changes in conformational entropy, providing another way to tune protein actions[85]. These ideas may bring to light why certain proteins have different functions in different cell types or cellular compartments.

The link between the sequence and conformational entropy is still unknown. Alphafold, Rosettafold, and diffusion-based models have provided insight into single conformer structures, but conformational entropy cannot be extracted from these predictions[86–8878]. While algorithms are being developed to coax structure prediction software into modeling multiple states, it is unclear how best to implement or interpret these findings[89,90]. Further, significant improvements in sidechain placement are critical to integrating the structure prediction with conformational entropy. Uncovering this link would provide an enormous platform for hypothesis generation and prediction of the impacts of perturbations. Machine learning may also help us uncover the dependencies of conformational entropy. Using an autoencoder, we may be able to use the latent space to discover additional dependencies of conformational entropy. Further, this could be used to determine how the unlimited number of perturbations may impact conformational entropy.

Integrating the impact of conformational entropy in ligand design will likely increase our ability to predict the binding and impact of molecules. Conformational entropy has likely been used to optimize drug-like molecules without us consciously realizing it. However, the above methods may allow for a much more data-driven perspective for ligand design. Conformational entropy should be exploited in protein design. Molecular dynamics and NMR have helped manipulate the underlying conformational landscape of proteins[91]. It is possible to design proteins that can switch between multiple conformations with multiconformer modeling and measure how residues conformations change. This may help overcome the difficulty of designing catalytically efficient proteins[92]. Considering conformational entropy will also likely help design a stable structure, as these fluctuations help stabilize structures[93].

Conclusions

Obtaining atomistic and mechanistic explanations of conformational entropy will help move the concept of entropy in macromolecular interactions from theory to experimentally testable and measurable. The methods and approaches discussed in this review will expand our toolbox of experimentally measurable protein interactions from mostly focused on only enthalpically driven to entropically and enthalpically driven. Many examples of the impact of conformational entropy upon ligand binding have demonstrated that the free energy of binding decreases when the protein must become more rigid to bind its ligand. When thinking about ligand design, a rigid, stable protein is usually what ligands are designed for. However, it may be more energetically favorable

to design for a more flexible state of the protein to reduce entropy loss upon binding.

This shift of thinking may also help us understand protein evolution. Many proteins have likely evolved to absorb energetic losses upon perturbation, which also helps them evolve new functions. The manipulation of entropy through macromolecular binding is a subtle but powerful way to regulate biological function. Measuring and visualizing entropy will allow for the explanation of biological phenomena rooted more solidly in energetic theory.

References

1. Ramachandran, G. N., Ramakrishnan, C. & Sasisekharan, V. Stereochemistry of polypeptide chain configurations. *Journal of Molecular Biology* vol. 7 95–99 Preprint at https://doi.org/10.1016/s0022-2836(63)80023-6 (1963).

2. Makhatadze, G. I. & Privalov, P. L. On the entropy of protein folding. *Protein Sci.* **5**, 507–510 (1996).

3. Liang, J. & Dill, K. A. Are proteins well-packed? *Biophys. J.* **81**, 751–766 (2001).

4. Monod, J., Wyman, J. & Changeux, J. P. ON THE NATURE OF ALLOSTERIC TRANSITIONS: A PLAUSIBLE MODEL. *J. Mol. Biol.* **12**, 88–118 (1965).

5. Cooper, A. & Dryden, D. T. Allostery without conformational change. A plausible model. *Eur. Biophys. J.* **11**, 103–109 (1984).

6. Hilser, V. J., Dowdy, D., Oas, T. G. & Freire, E. The structural distribution of cooperative interactions in proteins: analysis of the native state ensemble. *Proc. Natl. Acad. Sci. U. S. A.* **95**, 9903–9908 (1998).

7. Shapovalov, M. V. & Dunbrack, R. L., Jr. A smoothed backbone-dependent rotamer library for proteins derived from adaptive kernel density estimates and regressions. *Structure* **19**, 844–858 (2011).

8. Saavedra, H. G., Wrabl, J. O., Anderson, J. A., Li, J. & Hilser, V. J. Dynamic allostery can drive cold adaptation in enzymes. *Nature* **558**, 324–328 (2018).

9. Motlagh, H. N., Wrabl, J. O., Li, J. & Hilser, V. J. The ensemble nature of allostery. *Nature* **508**, 331–339 (2014).

10. Wand, A. J. The dark energy of proteins comes to light: conformational entropy and its role in protein function revealed by NMR relaxation. *Curr. Opin. Struct. Biol.* **23**, 75–

81 (2013).

11. Igumenova, T. I., Frederick, K. K. & Wand, A. J. Characterization of the fast dynamics of protein amino acid side chains using NMR relaxation in solution. *Chem. Rev.* **106**, 1672–1699 (2006).

12. Lipari, G. & Szabo, A. Model-free approach to the interpretation of nuclear magnetic resonance relaxation in macromolecules. 1. Theory and range of validity. *Journal of the American Chemical Society* vol. 104 4546–4559 Preprint at https://doi.org/10.1021/ja00381a009 (1982).

13. Lipari, G. & Szabo, A. Model-free approach to the interpretation of nuclear magnetic resonance relaxation in macromolecules. 2. Analysis of experimental results. *Journal of the American Chemical Society* vol. 104 4559–4570 Preprint at https://doi.org/10.1021/ja00381a010 (1982).

14. Wankowicz, S. A., de Oliveira, S. H., Hogan, D. W., van den Bedem, H. & Fraser, J. S. Ligand binding remodels protein side-chain conformational heterogeneity. *Elife* **11**, (2022).

15. Caro, J. A. *et al.* Entropy in molecular recognition by proteins. *Proc. Natl. Acad. Sci. U. S. A.* **114**, 6563–6568 (2017).

16. Frederick, K. K., Marlow, M. S., Valentine, K. G. & Wand, A. J. Conformational entropy in molecular recognition by proteins. *Nature* **448**, 325–329 (2007).

17. Laursen, L., Kliche, J., Gianni, S. & Jemth, P. Supertertiary protein structure affects an allosteric network. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 24294–24304 (2020).

18. Petit, C. M., Zhang, J., Sapienza, P. J., Fuentes, E. J. & Lee, A. L. Hidden dynamic allostery in a PDZ domain. *Proceedings of the National Academy of Sciences* vol. 106

18249–18254 Preprint at https://doi.org/10.1073/pnas.0904492106 (2009).

19. Riley, B. T. *et al.* qFit 3: Protein and ligand multiconformer modeling for X-ray crystallographic and single-particle cryo-EM density maps. *Protein Sci.* **30**, 270–285 (2021).

20. Keedy, D. A., Fraser, J. S. & van den Bedem, H. Exposing Hidden Alternative Backbone Conformations in X-ray Crystallography Using qFit. *PLoS Comput. Biol.* **11**, e1004507 (2015).

21. Fenwick, R. B., van den Bedem, H., Fraser, J. S. & Wright, P. E. Integrated description of protein dynamics from room-temperature X-ray crystallography and NMR. *Proc. Natl. Acad. Sci. U. S. A.* **111**, E445–54 (2014).

22. van den Bedem, H., Bhabha, G., Yang, K., Wright, P. E. & Fraser, J. S. Automated identification of functional dynamic contact networks from X-ray crystallography. *Nat. Methods* **10**, 896–902 (2013).

23. Olivieri, C. *et al.* ATP-competitive inhibitors modulate the substrate binding cooperativity of a kinase by altering its conformational entropy. *Sci Adv* **8**, eabo0696 (2022).

24. Pucheta-Martínez, E. *et al.* An Allosteric Cross-Talk Between the Activation Loop and the ATP Binding Site Regulates the Activation of Src Kinase. *Sci. Rep.* **6**, 24235 (2016).

25. McClendon, C. L., Kornev, A. P., Gilson, M. K. & Taylor, S. S. Dynamic architecture of a protein kinase. *Proc. Natl. Acad. Sci. U. S. A.* **111**, E4623–31 (2014).

26. De Zitter, E., Coquelle, N., Oeser, P., Barends, T. R. M. & Colletier, J.-P. Xtrapol8 enables automatic elucidation of low-occupancy intermediate-states in crystallographic

studies. *Commun Biol* **5**, 640 (2022).

27. Pearce, N. M. *et al.* A multi-crystal method for extracting obscured crystallographic states from conventionally uninterpretable electron density. *Nat. Commun.* **8**, 15123 (2017).

28. Best, R. B., Lindorff-Larsen, K., DePristo, M. A. & Vendruscolo, M. Relation between native ensembles and experimental structures of proteins. *Proc. Natl. Acad. Sci. U. S. A.* **103**, 10901–10906 (2006).

29. Yabukarski, F. *et al.* Ensemble-function relationships to dissect mechanisms of enzyme catalysis. *Sci Adv* **8**, eabn7738 (2022).

30. Merski, M., Fischer, M., Balius, T. E., Eidam, O. & Shoichet, B. K. Homologous ligands accommodated by discrete conformations of a buried cavity. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 5039–5044 (2015).

31. Stachowski, T. R. & Fischer, M. Large-Scale Ligand Perturbations of the Protein Conformational Landscape Reveal State-Specific Interaction Hotspots. *J. Med. Chem.* **65**, 13692–13704 (2022).

32. Modi, V. & Dunbrack, R. L., Jr. Defining a new nomenclature for the structures of active and inactive kinases. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 6818–6827 (2019).

33. Fenley, A. T., Muddana, H. S. & Gilson, M. K. Entropy-enthalpy transduction caused by conformational shifts can obscure the forces driving protein-ligand binding. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 20006–20011 (2012).

34. Bonin, J. P. *et al.* Positive Cooperativity in Substrate Binding by Human Thymidylate Synthase. *Biophys. J.* **120**, 4137 (2021).

35. Bonin, J. P., Sapienza, P. J. & Lee, A. L. Dynamic allostery in substrate binding by

human thymidylate synthase. *Elife* **11**, (2022).

36. Kalodimos, C. G. *et al.* Structure and flexibility adaptation in nonspecific and specific protein-DNA complexes. *Science* **305**, 386–389 (2004).

37. Eicken, C. *et al.* A metal-ligand-mediated intersubunit allosteric switch in related SmtB/ArsR zinc sensor proteins. *J. Mol. Biol.* **333**, 683–695 (2003).

38. Arunkumar, A. I., Campanello, G. C. & Giedroc, D. P. Solution structure of a paradigm ArsR family zinc sensor in the DNA-bound state. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 18177–18182 (2009).

39. Capdevila, D. A., Braymer, J. J., Edmonds, K. A., Wu, H. & Giedroc, D. P. Entropy redistribution controls allostery in a metalloregulatory protein. *Proc. Natl. Acad. Sci. U. S. A.* **114**, 4424–4429 (2017).

40. Tzeng, S.-R. & Kalodimos, C. G. Protein activity regulation by conformational entropy. *Nature* **488**, 236–240 (2012).

41. McDonald, L. R., Whitley, M. J., Boyer, J. A. & Lee, A. L. Colocalization of fast and slow timescale dynamics in the allosteric signaling protein CheY. *J. Mol. Biol.* **425**, 2372–2381 (2013).

42. Pearce, N. M. & Gros, P. A method for intuitively extracting macromolecular dynamics from structural disorder. *Nat. Commun.* **12**, 5493 (2021).

43. Ploscariu, N., Burnley, T., Gros, P. & Pearce, N. M. Improving sampling of crystallographic disorder in ensemble refinement. *Acta Crystallogr D Struct Biol* **77**, 1357–1364 (2021).

44. Burnley, B. T., Tom Burnley, B., Afonine, P. V., Adams, P. D. & Gros, P. Modelling dynamics in protein crystal structures by ensemble refinement. *eLife* vol. 1 Preprint at

https://doi.org/10.7554/elife.00311 (2012).

45. Strom, A. *et al.* Dynamic Long-Range Interactions Influence Substrate Binding and Catalysis by Human Histidine Triad Nucleotide-Binding Proteins (HINTs), Key Regulators of Multiple Cellular Processes and Activators of Antiviral ProTides. *Biochemistry* **61**, 2648–2661 (2022).

46. Carbone, C. E. *et al.* Time-resolved cryo-EM visualizes ribosomal translocation with EF-G and GTP. *Nat. Commun.* **12**, 7236 (2021).

47. Kaledhonkar, S. *et al.* Late steps in bacterial translation initiation visualized using time-resolved cryo-EM. *Nature* **570**, 400–404 (2019).

48. Mäeots, M.-E. *et al.* Modular microfluidics enables kinetic insight from time-resolved cryo-EM. *Nat. Commun.* **11**, 3465 (2020).

49. Kupitz, C. *et al.* Serial time-resolved crystallography of photosystem II using a femtosecond X-ray laser. *Nature* **513**, 261–265 (2014).

50. Yabukarski, F. *et al.* Assessment of enzyme active site positioning and tests of catalytic mechanisms through X-ray-derived conformational ensembles. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 33204–33215 (2020).

51. Petrychenko, V. *et al.* Structural mechanism of GTPase-powered ribosome-tRNA movement. *Nat. Commun.* **12**, 5933 (2021).

52. Fei, J. *et al.* Allosteric collaboration between elongation factor G and the ribosomal L1 stalk directs tRNA movements during translation. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 15702–15707 (2009).

53. Ray, K. K. *et al.* Entropic control of the free energy landscape of an archetypal biomolecular machine. *bioRxiv* 2022.10.03.510626 (2022)

doi:10.1101/2022.10.03.510626.

54. Han, K.-L., Zhang, X. & Yang, M.-J. *Protein Conformational Dynamics*. (Springer Science & Business Media, 2014).

55. Ginn, H. M. Vagabond: bond-based parametrization reduces overfitting for refinement of proteins. *Acta Crystallogr D Struct Biol* **77**, 424–437 (2021).

56. Ginn, H. M. Torsion angles to map and visualize the conformational space of a protein. Preprint at https://doi.org/10.1101/2022.08.04.502807.

57. Khade, P. M. & Jernigan, R. L. Entropies Derived from the Packing Geometries within a Single Protein Structure. *ACS Omega* **7**, 20719–20730 (2022).

58. Prévost, J. *et al.* Impact of temperature on the affinity of SARS-CoV-2 Spike glycoprotein for host ACE2. *J. Biol. Chem.* **297**, 101151 (2021).

59. Vergara, N. G., Gatchel, M. & Abrams, C. F. Entropic overcompensation of the N501Y mutation on SARS-CoV-2 S binding to ACE2. *bioRxiv* (2022) doi:10.1101/2022.08.30.505841.

60. Mugnai, M. L., Shin, S. & Thirumalai, D. Reduction in RBD Binding Affinity to Glycosylated ACE2 is Entropic in Origin. Preprint at https://doi.org/10.1101/2022.10.12.511994.

61. Casalino, L. *et al.* Beyond Shielding: The Roles of Glycans in the SARS-CoV-2 Spike Protein. *ACS Cent Sci* **6**, 1722–1734 (2020).

62. Reinhart, G. D., Hartleip, S. B. & Symcox, M. M. Role of coupling entropy in establishing the nature and magnitude of allosteric response. *Proc. Natl. Acad. Sci. U. S. A.* **86**, 4032–4036 (1989).

63. Capdevila, D. A. *et al.* Functional Role of Solvent Entropy and Conformational

Entropy of Metal Binding in a Dynamically Driven Allosteric System. *J. Am. Chem. Soc.* **140**, 9108–9119 (2018).

64. Gorman, S. D., Winston, D. S., Sahu, D. & Boehr, D. D. Different Solvent and Conformational Entropy Contributions to the Allosteric Activation and Inhibition Mechanisms of Yeast Chorismate Mutase. *Biochemistry* **59**, 2528–2540 (2020).

65. Bruning, J. B. *et al.* Coupling of receptor conformation and ligand orientation determine graded activity. *Nat. Chem. Biol.* **6**, 837–843 (2010).

66. van Zundert, G. C. P. *et al.* qFit-ligand Reveals Widespread Conformational Heterogeneity of Drug-Like Molecules in X-Ray Electron Density Maps. *J. Med. Chem.* **61**, 11183–11198 (2018).

67. Marques, B. S. *et al.* Protein conformational entropy is not slaved to water. *Sci. Rep.* **10**, 1–8 (2020).

68. Halle, B. Biomolecular cryocrystallography: structural changes during flash-cooling. *Proc. Natl. Acad. Sci. U. S. A.* **101**, 4793–4798 (2004).

69. Lang, P. T., Holton, J. M., Fraser, J. S. & Alber, T. Protein structural ensembles are revealed by redefining X-ray electron density noise. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 237–242 (2014).

70. Fraser, J. S. *et al.* Hidden alternative structures of proline isomerase essential for catalysis. *Nature* **462**, 669–673 (2009).

71. Keedy, D. A. *et al.* Mapping the conformational landscape of a dynamic enzyme by multitemperature and XFEL crystallography. *Elife* **4**, (2015).

72. Young, I. D. *et al.* Structure of photosystem II and substrate binding at room temperature. *Nature* **540**, 453–457 (2016).

73. Thompson, M. C. *et al.* Temperature-Jump Solution X-ray Scattering Reveals Distinct Motions in a Dynamic Enzyme. Preprint at https://doi.org/10.1101/476432.

74. Helliwell, J. R. Relating protein crystal structure to ligand-binding thermodynamics. *Acta Crystallogr. Sect. F Struct. Biol. Cryst. Commun.* **78**, 403–407 (2022).

75. Correy, G. J. *et al.* The mechanisms of catalysis and ligand binding for the SARS-CoV-2 NSP3 macrodomain from neutron and x-ray diffraction at room temperature. *Sci Adv* **8**, eabo5083 (2022).

76. Levy, A., Wetzstein, G., Martel, J., Poitevin, F. & Zhong, E. D. Amortized Inference for Heterogeneous Reconstruction in Cryo-EM. (2022) doi:10.48550/arXiv.2210.07387.

77. Anand, N. & Achim, T. Protein Structure and Sequence Generation with Equivariant Denoising Diffusion Probabilistic Models. (2022) doi:10.48550/arXiv.2205.15019.

78. Ingraham, J. *et al.* Illuminating protein space with a programmable generative model. *bioRxiv* 2022.12.01.518682 (2022) doi:10.1101/2022.12.01.518682.

79. Watson, J. L. *et al.* Broadly applicable and accurate protein design by integrating structure prediction networks and diffusion generative models. *bioRxiv* 2022.12.09.519842 (2022) doi:10.1101/2022.12.09.519842.

80. Liu, N., Mikhailovskii, O., Skrynnikov, N. R. & Xue, Y. Simulating diffraction photographs based on molecular dynamics trajectories of a protein crystal: a new option to examine structure-solving strategies in protein crystallography. *IUCrJ* vol. 10 Preprint at https://doi.org/10.1107/s2052252522011198 (2023).

81. Vögele, M. *et al.* Systematic Analysis of Biomolecular Conformational Ensembles with PENSA. (2022) doi:10.48550/arXiv.2212.02714.

82. Weiss, M. S. *et al.* Of problems and opportunities-How to treat and how to not treat

crystallographic fragment screening data. *Protein Sci.* **31**, e4391 (2022).

83. Miller, M. D. & Phillips, G. N., Jr. Moving beyond static snapshots: Protein dynamics and the Protein Data Bank. *J. Biol. Chem.* **296**, 100749 (2021).

84. Foulkes-Murzycki, J. E., Rosi, C., Kurt Yilmaz, N., Shafer, R. W. & Schiffer, C. A. Cooperative effects of drug-resistance mutations in the flap region of HIV-1 protease. *ACS Chem. Biol.* **8**, 513–518 (2013).

85. Byun, J. A. *et al.* Allosteric pluripotency as revealed by protein kinase A. *Sci Adv* **6**, eabb1250 (2020).

86. Lane, T. J. Protein structure prediction has reached the single-structure frontier. *Nat. Methods* (2023) doi:10.1038/s41592-022-01760-4.

87. Baek, M. *et al.* Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **373**, 871–876 (2021).

88. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).

89. Del Alamo, D., Sala, D., Mchaourab, H. S. & Meiler, J. Sampling alternative conformational states of transporters and receptors with AlphaFold2. *Elife* **11**, (2022).

90. Robertson, A. J., Courtney, J. M., Shen, Y., Ying, J. & Bax, A. Concordance of X-ray and AlphaFold2 Models of SARS-CoV-2 Main Protease with Residual Dipolar Couplings Measured in Solution. *J. Am. Chem. Soc.* **143**, 19306–19310 (2021).

91. Damry, A. M., Mayer, M. M., Broom, A., Goto, N. K. & Chica, R. A. Origin of Conformational Dynamics in a Globular Protein. Preprint at https://doi.org/10.1101/724286.

92. Kuhlman, B. & Bradley, P. Advances in protein structure prediction and design.

Nature Reviews Molecular Cell Biology vol. 20 681–697 Preprint at https://doi.org/10.1038/s41580-019-0163-x (2019).

93. Norn, C. *et al.* Protein sequence design by conformational landscape optimization. *Proc. Natl. Acad. Sci. U. S. A.* **118**, (2021).

94. Wang, Z., Abeysinghe, T., Finer-Moore, J. S., Stroud, R. M. & Kohen, A. A remote mutation affects the hydride transfer by disrupting concerted protein motions in thymidylate synthase. *J. Am. Chem. Soc.* **134**, 17722–17730 (2012).

95. Lumry, R. & Rajender, S. Enthalpy-entropy compensation phenomena in water solutions of proteins and small molecules: a ubiquitous property of water. *Biopolymers* **9**, 1125–1227 (1970).

96. Bonomi, M., Pellarin, R. & Vendruscolo, M. Simultaneous Determination of Protein Structure and Dynamics Using Cryo-Electron Microscopy. *Biophys. J.* **114**, 1604–1613 (2018).

97. Caldararu, O., Kumar, R., Oksanen, E., Logan, D. T. & Ryde, U. Are crystallographic B-factors suitable for calculating protein conformational entropy? *Phys. Chem. Chem. Phys.* **21**, 18149–18160 (2019).

Publishing Agreement

It is the policy of the University to encourage open access and broad distribution of all theses, dissertations, and manuscripts. The Graduate Division will facilitate the distribution of UCSF theses, dissertations, and manuscripts to the UCSF Library for open access and distribution. UCSF will make such theses, dissertations, and manuscripts accessible to the public and will take reasonable steps to preserve these works in perpetuity.

I hereby grant the non-exclusive, perpetual right to The Regents of the University of California to reproduce, publicly display, distribute, preserve, and publish copies of my thesis, dissertation, or manuscript in any form or media, now existing or later derived, including access online for teaching, research, and public service purposes.

DocuSigned by:
Stephanie Wankowicz
0FE2F6F52A5549F...
Author Signature

3/6/2023
Date