Mapping and modeling X-ray diffuse scattering from protein crystals

by

Andrew Holland Van Benschoten

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Biophysics

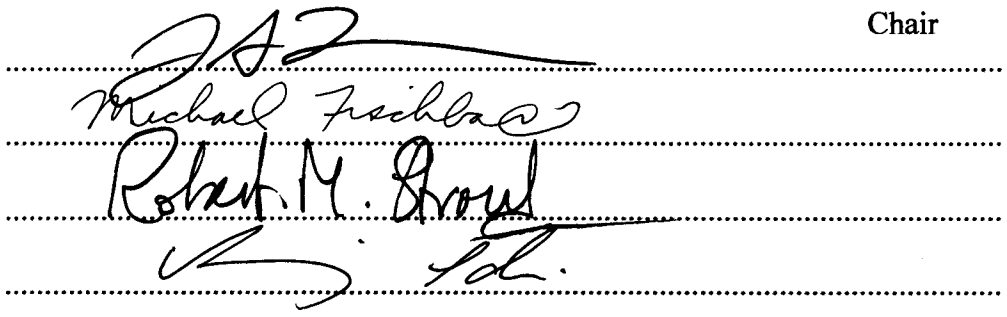
in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, SAN FRANCISCO

Approved:



Chair

Committee in Charge

UMI Number: 3714309

All rights reserved

INFORMATION TO ALL USERS
The quality of this reproduction is dependent upon the quality of the copy submitted.

In the unlikely event that the author did not send a complete manuscript
and there are missing pages, these will be noted. Also, if material had to be removed,
a note will indicate the deletion.

UMI 3714309

Published by ProQuest LLC (2015). Copyright in the Dissertation held by the Author.

Microform Edition © ProQuest LLC.
All rights reserved. This work is protected against
unauthorized copying under Title 17, United States Code

ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 - 1346

Copyright 2015
by
Andrew Van Benschoten

To Sarah and Abigail, the loves of my life.

Acknowledgements

My time in graduate school would not have been possible without the love and support of my family. Primary thanks go to my parents Dave and Greta Van Benschoten, who instilled in me the love of learning through 18 years of home education and beyond. My brothers Matthew, Mark and John deserve recognition for putting up with my nerdy pursuits and being the best classmates (as well as siblings) possible. The impact of my many relatives cannot be understated. Thanks to my grandparents Dr. Henry and Jackie Holland for demonstrating that it's possible to combine ethics with intellect, as well as to my great-uncle Jack Crawford, who is easily the best example of a lifelong learner that I know. I also want to acknowledge my aunt Dr. Roseanne Sension for enjoyable science talks over the holidays and the constant reminder that crystallography is the "easiest subject in biophysics".

My deepest gratitude must be expressed to my undergraduate research advisor Cathy Drennan and graduate mentor Christine Piro for introducing me to X-ray diffraction and helping lay the foundation of my scientific career. Thanks to professors Iain Cheeseman, Hazel Sieve, Tyler Jacks and Terry Orr-Weaver for fascinating undergraduate lectures and guidance on how to further my biology studies. Surviving MIT is only possible with a solid group of friends to lean on, vent to and live life with. I am forever indebted to Josh Suresh, Jeremiah Noordhoek, Jack Bourbonnais, Dave & Abby Reens, Paul Morales, Zac & Ruth Nelson, Chuky Mbagwu and Wes McDougal for keeping me sane and (relatively) well adjusted.

Success in graduate school depends on the investment and mentorship of older graduate students, staff and faculty. Thanks to David Paquette for single-handedly corralling a class of

rowdy first-year students and making sure we learned the fundamentals of biophysics. The personal guidance of James Kraemer and David Booth was similarly invaluable. Special thanks to James Partridge, who holds the unique distinction of being the only teacher I've had for both my undergraduate and graduate careers. Chris Waddling was instrumental in teaching me the principles of data processing. Thanks to Gigi Knudsen for mass spectroscopy guidance and Kyoshi Egami for providing unfettered access to the Superose 6 column and some absurdly powerful electron microscopes.

Though we've all gone our separate ways, I'll always remember the Mission Bay Tower Crew of Susan Chen, D'Juan Farmer, Jack Yue, Kyle Barlow, Liron Noiman, Ben Spangler and the camaraderie we shared. My Christian community has been a bedrock for me through the last few years, particularly Arjun Thounoujam (who also spent untold hours helping me build Python libraries), John George, Wilbur Hu and Kyle Gong. San Diego 2013 will live on forever. Thanks to Kristen Herberg, my beer buddy and Frisbee partner, and Robert Bajada for frequent outings to the shooting range. To my RSF Community Group (Moses Ike, Brian Sun, Laura Richardson, Laura Mead, Tony Chan, Gabe & Jenni Sudario, Emily Tang, Kylie Veverka and many others) I can't begin to express my gratitude. You helped me through my toughest times, introduced me to my wife and are a continual source of love and support as I navigate fatherhood.

I can't imagine a better place to pursue X-ray crystallography than the Bay Area. I am deeply indebted to Ana Gonzalez and Lisa Dunn at SSRL for giving me far more experimental time on beamlines 11-1 and 12-2 than my proposal score deserved. James Holton and George Meigs at ALS played a major role in developing the experimental methods behind diffuse scattering data

collection. The folks in Paul Adams' PHENIX group were a tremendous resource and displayed never-ending patience as I took my first steps in the world of programming. Special thanks to Aaron Brewster and Nick Sauter, who debugged countless issues within DIALS-LUNUS.

I've been blessed to experience some of the finest collaboration in science. Mike Wall's extensive contributions range from forming the foundation of 3D diffuse scattering analysis two decades ago to patiently putting up with three years of never-ending questions. Pavel Afonine and Alexandre Urzhumtsev worked tirelessly on how to best analyze TLS refinement and produced some of the most beautiful matrix algebra I've seen. Special thanks goes to Brianna Williams for providing me with my first opportunity to solve a new protein structure. The pinnacle of scientific collaboration occurred for me at the 2014 Gordon Research Conference on Methods in Diffraction Analysis, where Elspeth Garman, John Kuryian, David Case, John Spence and many others challenged my scientific thinking while suggesting innumerable avenues of future study. Finally, I want to individually thank each of the members of my thesis committee: Bob Stroud for deepening my understanding of (as well as passion for) the mathematics underlying crystallography, Andrej Sali for assisting with my transformation into a computational scientist and Michael Fischbach for keeping me grounded and focused on the big picture.

My best memories of UCSF have been as a member of the Fraser Lab, from Snapchat sessions to Canada Day parties to our quest to occupy every building in Mission Bay at least once. Rahel Woldeyes, David Mavor and I were Jaime's first crop of students and passed through the fires of grad student life together. Special thanks to Lillian Kenner for playing devil's advocate in all

areas of life. Daniel Keedy, Lin Liu and Mike Thompson have been perfect examples of what postdocs should be: knowledgeable, approachable and always willing to lend a hand. Daniel gets extra points for being an expert “pun-smith”. The dynamic research/baking duo of Justin Biel and Benjamin Barad has been fantastic to work with and help ensure that the future of the lab remains bright.

I especially want to thank my Ph.D advisor, Dr. James Fraser. His influence on my scientific identity cannot be understated, as he taught me how to ask the big questions, write effective scientific literature and how to collaborate with everyone under the sun. His presentation about room-temperature crystallography during my graduate interview weekend (and the offer that I could work on similar topics if I joined his lab) is what ultimately tipped the scales in favor of UCSF. I’m extremely thankful for his introduction to computational biology, Python and data visualization, as it stoked my desire to become a data scientist. Though many professors view their role as complete when a student graduates, Jaime works tirelessly to make sure that each of his advisees can pursue their vocational passion. I’m grateful for his assistance as I leave UCSF for Silicon Valley and can only hope that every graduate student feels this supported in their post-Ph.D pursuits.

My most important acknowledgements go to my beautiful wife Sarah and lovely daughter Abigail. I love you two ladies more than you can imagine. Sarah, I never would have made it this far without your support: early on you made it clear that you believed in me and would stand by me in both the good and the bad. You planned our big “pre-qualifying exam date” to help me stay sane the day before my presentation, put up with numerous late-night beam line trips and

gave me a safe place to revel in my successes and vent my frustrations. Abby, you may not be old enough to read these words, but you've helped your daddy understand what's most important in life. I'm thrilled to watch you grow into a young woman and share with you all the cool things this world has to offer.

Abstract

Understanding the physical basis of enzyme dynamics is a major challenge in biology. Although modeling the motion of individual atoms is straightforward, combining these movements into descriptions of macromolecular function proves more difficult. X-ray crystallography produces atomic-level visualizations of an ensemble of countless molecules; however, current methods capture only the average protein conformation and thus cannot completely describe the underlying dynamics. A parallel source of information, diffuse scattering, is present in diffraction images and directly reports on correlated atomic motions.

I created experimental and computational tools to measure macromolecular diffuse scattering and compare it against hypotheses of correlated motion. The first tool, *phenix.diffuse*, calculates diffuse scattering patterns from known structural ensembles. I applied this software to the refinement technique *Translation-Libration-Screw* and solved a pre-existing degeneracy within the predicted motion of glycerophosphodiesterase GpdQ. Surprisingly, I also uncovered a fundamental flaw in the implementation of TLS refinement in structural biology software, revealing unphysical motions to be present in nearly 25% of all known macromolecular structures.

Next, I developed the comprehensive pipeline *DIALS-LUNUS* for the measurement of macromolecular diffuse scattering. This system was applied to crystals of the proline isomerase cyclophilin A (CypA) and trypsin, ultimately producing high-resolution diffuse maps of both proteins. These maps were compared to several distinct models of motion that were previously indistinguishable to crystallographic techniques. By comparing the experimental data to each

predicted diffuse scattering pattern, I was able to successfully identify the most probable mechanism of motion. Ultimately, these studies provide a new avenue of exploration in the pursuit of understanding molecules as dynamic entities.

Table of Contents

List of Tables

Chapter 2

Chapter 3

Chapter 4

List of Figures

Chapter 2

Chapter 3

Chapter 4

Chapter 1

Introduction

Any good introductory biology course will doubtless cover the fascinating topic of enzyme selectivity. Despite the astronomical number of molecules in a cell, binding partners are able to uniquely identify each other and collaborate on critical chemical reactions. Perhaps more importantly, these seemingly improbable events happen frequently enough to sustain the processes of life. The 20th century saw two competing theories attempt to explain this paradox. The first, the "lock-and-key" hypothesis, was proposed by chemist Emil Fischer in 1890 and defined both enzyme and substrate as uniquely shaped rigid bodies. In the same way that a door is opened by only a single key (out of many possibilities), biological selectivity is achieved through perfect shape complementarity. In stark contrast stood the induced fit mechanism (Koshland, 1958), in which weak initial interactions between binding partners triggered structural rearrangements, resulting in a more ideal enzyme-substrate fit. Though both lines of reasoning make intuitive sense, substantial evidence has emerged over the past few decades that support the induced-fit hypothesis (Lakowitz and Weber, 1973; Frauenfelder and Petsko, 1980; James *et.al*, 2003; Fraser *et.al*, 2009). Indeed, we now know that molecules flex, bend and twist in order to accomplish functions as diverse as enzymatic turnover, signaling regulation and protein-protein interactions (Woldeyes *et.al*, 2014).

This shift in understanding molecules as conformational ensembles rather than unbending chunks of carbon has also taken place in the field of X-ray crystallography. Crystallography is by nature an ensemble experiment: many copies of the molecule come together to form a rigid lattice. If we assume the molecule adopts an "average" structure at each lattice point, our system can be perfectly described by Bragg's Law of X-ray scattering (Bragg & Bragg, 1913). According to Bragg's law, diffraction events occur in well-defined, repeating regions of reciprocal space known as "Bragg spots" and can be modeled using simple mathematics. If we measure enough Bragg spots, apply a Fourier transform and perform a bit of electron density modeling, we'll end up with a snapshot of this "average" protein structure. This trick of simplifying our crystalline system to the "average" protein structure enables much easier parameter fitting and provided great benefit to early crystallographers, who were forced to perform the intimidating task of solving protein structures by hand. However, it was soon realized that a single average conformation was inadequate to capture the rich dynamics present within the crystal. In 1979, Frauenfelder *et.al* introduced the notion of atomic B-factors: Gaussian modulating terms that captured the spread in atomic electron density. Including these parameters provided a more accurate fit to observed data; differences in B-factor values also provided evidence for localized molecular flexibility. In a similar fashion, mathematical models such as the Translation-Libration-Screw approximation ("TLS", Schomaker & Trueblood, 1968) provided implicit descriptions of flexibility present in the macromolecule. As data collection and analysis methods improved, explicitly modeling part (or all) of the conformational states into the "average" structural data became possible (Gros *et.al*, 1990; Rader and Agard, 1997). A powerful example of the usefulness of defining the explicit ensemble to was produced by Fraser *et.al* (2009), in which structural refinement of high-resolution human proline isomerase

cyclophilin A data revealed two distinct protein conformations that defined a mechanism for catalytic turnover. As this mechanism was validated in parallel through Nuclear Magnetic Resonance analysis and mutational studies, modern crystallography is now capable of providing a detailed glimpse into the inner dynamics of biological macromolecules.

However, this structural understanding of enzyme dynamics is fundamentally at odds with the "average" protein model provided by Bragg data, as there may be several orthogonal models of correlated atomic displacement that fit the cumulative density equally well. To break this degeneracy we must look beyond the Bragg spots to the realm of X-ray diffuse scattering, which occurs throughout reciprocal space and results from correlated variation in the electron density distributions (Phillips *et al.*, 1980; Chacko & Phillips, 1992; Faure *et al.*, 1994; Clarage & Phillips, 1997; Mizuguchi *et al.*, 1994). This variation breaks from the theoretical "perfect" crystal lattice, leading to diffraction outside of the regions predicted by Bragg's law. Incorporating diffuse scattering analysis into crystallographic investigations would help create rich pictures of correlated molecular motion. Indeed, mathematical descriptions of the molecular imperfections underpinning various diffuse scattering patterns have been formalized (Guinier, 1963; Amoros & Amoros, 1964) and diffuse scattering analysis has previously shed light on small molecule motion (Welberry & Butler, 1994; Estermann & Steurer, 1998; Michels-Clark *et al.*, 2013).

Unfortunately, real-world limitations have historically prevented the application of diffuse scattering analysis to more complex targets. The measured signal is weak, often a thousand-fold less intense than the corresponding Bragg peaks. As a result, crystals must undergo long periods

of X-ray exposure for these features to be revealex, frequently causing significant radiation damage to the lattice. Additionally, the increased parametric complexity of macromolecular motion significantly complecates the modeling of diffuse scattering beyond the smallest atomic systems. Despite these challenges, several notable attempts have been made to analyze the diffuse signal arising from protein crystals. Phillips and co-workers produced initial investigations of the global motions of tropomyosin in the crystalline lattice (Phillips *et.al*, 1980) and demonstrated that the observed diffuse features could be roughly reproduced if the short and long filament arms were assumed to fluctuate over distances of 5 Å and 8 Å, respectively. Doucet & Benoit (1987) later proposed that sharp diffuse streaks observed along the Bragg planes in lysozyme could be explained by molecular vibrations coupled across multiple unit cells along the *a* and *c* lattice axes. Comparison of experimental lysozyme features to simulated diffuse intensities similarly revealed nearest-neighbor interactions across the unit cell boundaries (Clarage *et.al*, 1992), as well as rough agreement with normal-mode calculations of the protein structure (Faure *et.al*, 1994). In each of these instances, however, experimental diffuse data was limited to a single diffraction frame. Recognizing the need for a more complete picture of diffuse scattering, Wall *et.al* (1997) introduced the first computational framework for the measurement of diffuse signal across the entirety of reciprocal space. This framework was applied to diffraction data from *Staphylococcal* nuclease crystals, revealing that 3D diffuse scattering features were reproducible and displayed crystallographic symmetry. Though the computational elements for obtaining complete pictures of macromolecular diffuse scattering were now available, inherent limitations within Charge-Coupled Device (CCD) detectors (present at nearly every synchrotron beamline worldwide) continued to prohibit diffuse intensity measurement. These cameras were highly noisy and possessed significant point-spread functions, causing weak

diffuse intensities to wash out in the diffraction background. Thus, in the absence of technically challenging detector modification (Wall, 1996), hardware limitations proved to be yet another obstacle in the pursuit of routine diffuse scattering analysis.

However, the recently developed pixel-array detector (Gruner *et.al*, 2002) has overcome these technical challenges and provides perhaps the final component for developing diffuse scattering into a standard crystallographic technique. These detectors provide exquisite single-photon diffraction detail across a low-noise background, providing an ideal setting for the measurement of weak signal. When coupled with 21st-century room-temperature diffraction methods (Fraser *et.al*, 2011), it becomes possible to extract complete Bragg and diffuse datasets from single protein crystals with little to no radiation damage. A significant parallel development has been the convergence of computational crystallographic tools into well-defined open-source packages. Perhaps most notable is the Computational Crystallography Toolbox and the Python-based Hierarchical ENvironment for Integrated Xtallography (PHENIX; Adams *et.al*, 2010), which provides the groundwork for assimilating pre-existing diffuse measurement programs into standard crystallographic refinement routines. Building on the work of others before me, I have developed computational and experimental tools to capture diffuse scattering intensities and fit the resulting data to models of protein motion. In Chapter 2 I outline my mathematical investigation of TLS refinement, long thought to be quantifiable through diffuse scattering analysis (Tickle & Moss, 1999), and reveal unexpected issues in the computational implementations of TLS within all current refinement software packages. In Chapter 3 I introduce *phenix.diffuse*, a tool capable of simulating X-ray diffuse scattering from structural ensembles and thus making the critical connection between experimental data and generalized

models of motion. Finally, in Chapter 4 I present DIALS-LUNUS, an experimental pipeline for quantifying diffuse signal from diffraction frames. Using this tool I created two new three-dimensional diffuse maps from the proteins cyclophilin A and trypsin. These maps were then utilized to select distinct models of motion for each molecule. My work provides the foundation for using every bit of available diffraction information to further our understanding of biological molecules as dynamic structural ensembles.

1.1 References

Adams, P. D., Afonine, P. V., Bunkoczi, G., Chen, V. B., Davis, I. W., Echols, N., Headd, J. J., Hung, L. W., Kapral, G. J., Grosse-Kunstleve, R. W., McCoy, A. J., Moriarty, N. W., Oeffner, R., Read, R. J., Richardson, D. C., Richardson, J. S., Terwilliger, T. C. & Zwart, P. H. (2010). *Acta crystallographica. Section D, Biological crystallography* **66**, 213-221.

Amorós, J. L. & Amorós, M. (1968). *Molecular crystals: their transforms and diffuse scattering* (Vol. 6). New York: Wiley.

Bragg, W. H., & Bragg, W.L. (1913) *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character,* 428-438.

Chacko, S., Phillips, G.N. Jr. (1992) *Biophysical Journal* **61**, 1256

Clarage, J. B., Clarage, M. S., Phillips, W. C., Sweet, R. M. & Caspar, D. L. (1992). *Proteins* **12**, 145-157.

Clarage, J. B., Phillips, G.N. Jr (1997) *Methods in Enzymology* **277**, 407-432

Doucet, J. & Benoit, J. P. (1987). *Nature* **325**, 643-646.

Estermann, M. A. & Steurer, W. (1998). *Phase Transitions* **67**, 165-195.

Faure, P., Micu, A., Perahia, D., Doucet, J., Smith, J. C., & Benoit, J. P. (1994). *Nature Structural & Molecular Biology **1***, 124-128.

Fraser, J. S., Clarkson, M. W., Degnan, S. C., Erion, R., Kern, D. & Alber, T. (2009). *Nature* **462**, 669-673.

Fraser, J. S., van den Bedem, H., Samelson, A. J., Lang, P.T., Holton, J. M., Echols, N., Alber, T. (2011) *Proceedings of the National Academy of Sciences* **108**, 16247-16252

Frauenfelder, H. & Petsko, G.A. (1980). *Biophysics Journal* **32**, 465

Frauenfelder, H., Petsko, G.A., and Tsernoglou, D. (1979) *Nature* **280**, 558-563.

Gros, P., van Gunsteren, W. F., & Hol, W. G. (1990). *Science* **249**, 1149-1152.

Gruner, S. M., Tate, M. W. & Eikenberry, E.F. (2002) *Review of Scientific Instruments* **73**, 2815-2842

Guinier, A. (1963). *X-ray diffraction in crystals, imperfect crystals, and amorphous bodies*. Courier Dover Publications.

James, L.C., Roversi, P., & Tawfik, D.S. (2003) *Science* **299**, 1362-1367.

Koshland, D.E. (1958). *Proceedings of the National Academy of Sciences* **44**, 98-104

Lakowitz, J.R & Weber, G. (1973). *Biochemistry* **12**, 4171

Michels-Clark, T., Lynch, V., Hoffmann, C., Hauser, J., Weber, T., Harrison, R. & Burgi, H. (2013). *Journal of Applied Crystallography* **46**, 1616-1625.

Mizuguchi, K., Kidera, A. & Go, N. (1994) *Proteins* **18**, 34-48

Phillips, G.N Jr, Fillers, J.P. & Cohen, C. (1980) *Biophysical Journal* **32**, 485-502

Rader, S. D. & Agard, D. A. (1997). *Protein science* **6**, 1375-1386.

Schomaker, V., & Trueblood, K. N. (1968). *Acta Crystallographica Section B: Structural Crystallography and Crystal Chemistry* **24**, 63-76.

Tickle, I. J., & Moss, D. S. (1999). *IUCr99 Computing School, London, United Kingdom* [Online *http://people.cryst.bbk.ac.uk/~tickle/iucr99/iucrcs99.html*]
Wall, M. E. (1996). PhD thesis, Princeton University.

Wall, M. E., Ealick, S. E. & Gruner, S. M. (1997). *Proceedings of the National Academy of Sciences of the United States of America* **94**, 6180-6184

Welberry, T.R & Butler, B.D. (1994) *Journal of Applied Crystallography* **27**, 205-231

Woldeyes, R. A., Sivak, D. A., & Fraser, J. S. (2014). *Current opinion in structural biology* **28**, 56-62.

Chapter 2

From deep TLS validation to ensembles of atomic models built from elemental motions

2.1 Abstract

The *Translation Libration Screw* model (Schomaker & Trueblood, 1968) describes concerted motions of atomic groups. This type of refinement can increase the agreement between calculated and experimental diffraction data and potentially shed light on molecular mechanisms. Three matrices *T*, *L* and *S* describe the combination of the atomic vibrations and librations. Because these matrices describe physical motions, their elements must satisfy several conditions described in this article. Refining the *T*, *L* and *S* matrix elements as independent parameters without taking these conditions into account may result in matrices that do not represent real-space movement. We describe a mathematical framework and the computational tools to analyze *TLS* matrices, resulting in either explicit decomposition into descriptions of the underlying motions or the detection of broken conditions. As a first application we present an algorithm to generate structural ensembles that are consistent with given *TLS* parameters. All methods are implemented as part of the *Phenix* project.

2.2 Introduction

Independent and concerted molecular motions

Crystallographic models that are used to fit diffraction data (X-ray, neutron, or electron diffraction) describe each atom by its central position $\mathbf{r_0}$ and parameters modeling displacement from the central position. Small-magnitude structure disorder, particularly thermal motion, is described by the so-called Debye-Waller factor that reflects the probability of an atom moving from its central position by certain amount. A contribution of each atom to a structure factor corresponding to integer indices (h,k,l) is multiplied by this factor that in a harmonic approximation can be presented as

$$\exp\left(-2\pi^2 \mathbf{h}^{\tau} O^{-1} U_{Cart} \left(O^{-1}\right)^{\tau} \mathbf{h}\right) \quad . \qquad (1)$$

(see for example, Grosse-Kunstleve & Adams, 2002, and references therein). Here O is the orthogonalization matrix for the given crystal, **h** is the column vector of integer indices (h,k,l), t signifies matrix or vector transposition, and U_{Cart} is an atomic displacement parameter, ADP. (In Grosse-Kunstleve & Adams (2002) the orthogonalization matrix is noted as A; here this letter is reserved for the matrix in development of U_{Cart}, following Tickle & Moss, 1999). The symmetric positive definite matrix U_{Cart} is defined by the average atomic shifts (and their correlations) along the coordinate axes. The matrix U_{Cart} varies between atoms and is diagonal (with equal elements) for atoms that are assumed to be moving isotropically.

For each atom, U_{Cart} accumulates several contributions arising from difference sources, including overall crystal anisotropy U_{cryst}, various kinds of concerted motions U_{group} and independent displacement of individual atoms U_{local} (see for example Dunitz & White, 1973; Prince & Finger, 1973; Johnson, 1980; Sheriff & Hendrickson, 1987; Murshudov *et al*., 1999; Winn *et al*., 2001).

A description of the concerted molecular motion of an atomic group by means of the *TLS* formalism has been introduced by Cruickshank (1956) and Schomaker & Trueblood (1968) and developed further in a number of works such as Johnson (1970), Scheringer (1973), Howlin *et al*. (1989, 1993), Kuriyan & Weis (1991), Schomaker & Trueblood (1998), Tickle & Moss (1999), Murshudov *et al*. (1999), Winn *et al*. (2001, 2003) and Painter & Merritt (2005, 2006a, 2006b). This type of motion is of special interest for several reasons. First, it may characterize the anisotropic mobility of atomic groups and thus give insight into molecular mechanism. Second, describing only the concerted motion can simplify the crystallographic model and reduce the number of parameters needed to model the data while simultaneously providing a more physically realistic description of atomic displacements.

A common misconception of TLS parametrization is that its sole merit is providing an economical way to account for anisotropic atomic motions at low resolution. There are a number of examples of using the *TLS* formalism to analyze flexibility of the molecules relevant to their mechanisms (see for example Kuriyan & Weis, 1991; Harris *et al*., 1992; Sali *et al*., 1992 ; Wilson & Brunger, 2000; Raaijmakers *et al*., 2001; Yousef *et al*., 2002; Papiz & Prince, 2003;

Chaudhry *et al*., 2004). Thus physically meaningful *TLS* matrices can provide useful structural information. Moreover, this information can be used to build ensembles of atomic models that explicitly represent corresponding conformations. In turn, this allows for a better description of diffraction data, particularly in the case of diffuse X-ray scattering (Van Benschoten *et al*., 2015). Thus, *TLS* parameterization can be useful regardless of the resolution of the available diffraction data.

TLS model

Since displacement of a rigid group of atoms is a composition of translation and rotation (see for example Goldstein, 1950), Schomaker & Trueblood (1968) presented the matrices $U_{group,n}$ for a concerted motion of a group of atoms $n = 1, 2, \ldots N$ as a sum

$$U_{group,n} = T + A_n L A_n^{\tau} + A_n S + S^{\tau} A_n^{\tau} \qquad (2)$$

Anti-symmetric matrices A_n are functions of the Cartesian coordinates (x_n, y_n, z_n) of atom n

$$A_n = \begin{pmatrix} 0 & z_n & -y_n \\ -z_n & 0 & x_n \\ y_n & -x_n & 0 \end{pmatrix} . \qquad (3)$$

Matrix S and symmetric matrices T and L are common to all atoms within each rigid group. L describes librations (oscillating rotations) around three principal rotation axes mutually orthogonal to each other. T describes apparent translations of the atomic group (the term 'vibrations' might actually be more appropriate for random translations around a central position). S describes screw motions, *i.e.* the combination of librations and vibrations. We use the term 'apparent translation' because matrix T may have an additional contribution from librations as discussed in Section 2.

Thus, explicit information about atomic motions can be encoded into *TLS* matrices making them inexplicit descriptors of such motions. These two ways of describing atomic group motions have their merits. Explicit description allows for a straightforward interpretation and analysis of the motions, while using *TLS* formalism provides an easier framework for calculating structure factors. As they arise from a specific combination of motions, elements of *TLS* matrices must obey certain conditions. However, current refinement programs treat elements of the *TLS* matrices as independent variables with a constraint on the trace of the matrix S (as discussed

below) and postrefinement enforcement of the resulting $U_{group,n}$ to be non-negative definite (Winn *et al*., 2001). Previously Zucker *et al*. (2010) analyzed all PDB entries containing *TLS* descriptions and suggested tools to validate the *TLS* parameters. However, this was done from the different viewpoint of smoothness of the ADP for neighboring groups. Failure to enforce all conditions on the individual components of $U_{group,n}$, *i.e.* on the *TLS* matrices, may result in these matrices being physically nonsensical and thus invalidating the *TLS* model. With methods and tools presented in this manuscript, we analyzed all structures from the PDB (Bernstein *et al*., 1977; Berman *et al*., 2000; about 105,000 entries, 25,000 of which contain *TLS* models with a total of 200,000 sets of matrices). A third of these sets contains *T* or *L* matrices that are non positive semidefinite and another third (Table 2.1) cannot describe libration-vibration correlated motions due to other reasons discussed in Sections 2 – 5. Some of these errors are trivial to fix, *e.g.* correcting marginally negative eigenvalues of *T* and *L* or modifying the trace of *S* (examples are given in Section 6 and in Table 2.1).

On the physical meaning and use of TLS

By definition, *TLS* is a physical description of the anisotropic harmonic motion of atomic groups in the crystal (Schomaker & Trueblood, 1968) and has no relations to the diffraction data quality (resolution, for instance). Consequently, it is incorrect to consider the *TLS* matrices to be merely a tool to describe the uncertainties in atomic positions in the context of a reduced number of refinement parameters. The physical meaning of the individual *T*, *L* and *S* matrices is as important as the condition imposed on $U_{group,n}$ to be positive definite. Nonsensical individual *TLS* matrices invalidate the underlying physical model. *B* values need to be positive, occupancies must range between 0 and 1 and atomic coordinates should define model geometry in accordance with chemical knowledge. Similarly, *TLS* matrices need to make sense according to the model they describe. Efforts to constrain *TLS* parameters to keep them physically meaningful have been discussed previously (Winn *et al*., 2001; Painter & Merritt, 2006a).

Because *TLS* matrices are physically meaningful, the corresponding composite motions can be extracted from these matrices and in turn help to analyze molecular mobility. However, while calculating *TLS* matrices from corresponding libration and vibration parameters is rather straightforward (Section 2), the inverse procedure is less trivial. As discussed previously (Johnson, 1970; Scheringer, 1973; Tickle & Moss, 1999) the problem itself is poorly posed since

the same set of diffraction data (and consequently the same set of the *TLS* matrices) may correspond to different motions of the contributing atoms or atomic groups. Moreover, there are computational difficulties if all the conditions on the matrices have not been considered (Sections 3-5). There exist several programs that aim to analyze *TLS* matrices (Howlin *et al*., 1993; Painter & Merritt, 2005); however, we could not find in the literature a comprehensive mathematical protocol for describing these procedures.

The set of *TLS* matrices that correspond to physically possible combinations of motions is smaller than the set of all possible *TLS* matrices. Since for any function its minimum on a larger set of parameters may be deeper than that for its subset, this means that refinement with versus without imposing conditions on *TLS* matrices may result in higher *R*-factors. Since *TLS* modeling is an approximation to the true molecular motions and also strongly depends on the assignment of TLS groups, higher *R*-factors as result of using *TLS* may be indicative of this model being counterproductive in that particular case.

Summary of presented work

In our work described in this article we address the following points.

- We describe the algorithm (Fig. 2.1) that interprets the *TLS* matrices in terms of parameters of the corresponding motions. This includes the direction of the principal axes of vibration and libration, the corresponding root-mean-square displacements, the position of the libration axes, as well as the correlations between vibration and libration displacements.
- We present a complete list of conditions that must be fulfilled to make the aforementioned *TLS* decomposition possible; this includes widely known conditions (e.g. *T* and *L* must be positive semidefinite) as well as a number of less trivial conditions that to the best of our knowledge have not been previously discussed.
- We describe the calculation protocols in a ready-to-program style so that they can be implemented in existing or future software. Most of calculations described in the manuscript are straightforward; less trivial expressions and proofs can be found in Appendix A as well as in the review Urzhumtsev *et al*. (2013) below referred to as UAA-2013.
- We implemented the described algorithms in the open source Computational Crystallography Toolbox (*cctbx*; Grosse-Kunstleve *et al*., 2002). Also, we made two end-

user applications available in *Phenix* suite (Adams *et al*., 2010).: *phenix.tls_analysis* for analysis and validation of refined *TLS* matrices and their underlying motions, and *phenix.tls_as_xyz* for generating ensembles of structures consistent with *TLS* matrices.

- We applied the program to all PDB entries containing *TLS* matrices. We discovered that majority of these matrices are nonsensical. For a number of cases marginal modification of the *TLS* matrices can correct the errors.
- We used the program to generate a structural ensemble for the calculation of X-ray diffuse scattering (Van Benschoten *et al*., 2015).

2.3 Calculating *TLS* matrices from elemental motions

This section provides a step-by-step protocol for the calculation of *TLS* matrices from the parameters of the composite vibrations and librations. Inverting this scheme provides the method of extracting libration-vibration parameters from the *TLS* matrices.

Constructing TLS matrices from the parameters of the libration and vibration

The matrices in equation (2) depend on the basis in which the atomic coordinates are given. We use the index in square brackets to indicate which basis is used. Let the atoms be given in some basis denoted [M]; for example it may be the basis corresponding to the model deposited into PDB. Even if a rigid group is involved in several simultaneous motions (supposing that the amplitudes of these motions are relatively small and the motions are harmonic), the total motion can be described by a libration around three axes $\boldsymbol{l}_x$, $\boldsymbol{l}_y$, $\boldsymbol{l}_z$ that are mutually orthogonal and by a vibration along three other mutually orthogonal axes, $\mathbf{v}_x$, $\mathbf{v}_y$, $\mathbf{v}_z$. These triplets of axes form the other two bases, [L] and [V].

In equation (2) the matrix T is a sum of several components. In the absence of librations (that is, matrices L and S are zero) it is equal to the contribution V arising from pure vibrations. In the basis [V] this matrix is diagonal

$$V_{[V]} = \begin{pmatrix} \langle t_x^2 \rangle & 0 & 0 \\ 0 & \langle t_y^2 \rangle & 0 \\ 0 & 0 & \langle t_z^2 \rangle \end{pmatrix}$$

(4)

Here $\langle t_x^2 \rangle, \langle t_y^2 \rangle, \langle t_z^2 \rangle$ are the corresponding squared root-mean-square deviations (*rmsd*) along the principal vibration axes $\mathbf{v_x}$, $\mathbf{v_y}$, $\mathbf{v_z}$ and are expressed in Å^2. If there are librations the matrix L is always diagonal in the basis [L]:

$$L_{[L]} = \begin{pmatrix} \langle d_x^2 \rangle & 0 & 0 \\ 0 & \langle d_y^2 \rangle & 0 \\ 0 & 0 & \langle d_z^2 \rangle \end{pmatrix}$$

(5)

Here $\langle d_x^2 \rangle, \langle d_y^2 \rangle, \langle d_z^2 \rangle$ are the squared *rmsd*s of the vibration angles expressed in squared radians; for small deviations they are numerically equal to the squared *rmsd*s of points at a unit distance from the corresponding axes.

In reality the principal vibration and libration axes are not parallel to each other; practically, it is convenient to express the matrices in a common basis. Basis [L] is more convenient for this since in this basis the elements of S (see below) are easily expressed through geometric parameters of librations. Matrix V in this basis is no longer diagonal but is instead equal to

$$V_{[L]} = R_{VL}^{\tau} V_{[V]} R_{VL} \quad .$$

(6)

Here R_{VL} is the transition matrix that describes the rotation that superposes the vectors $\mathbf{v_x}$, $\mathbf{v_y}$, $\mathbf{v_z}$ with the vectors $\boldsymbol{l_x}$, $\boldsymbol{l_y}$, $\boldsymbol{l_z}$ (Appendix A). Frequently, vibration and libration motions are not independent but instead are correlated to form screw rotations. It is convenient to characterize screw rotations by the parameters s_x, s_y, s_z such as for a screw rotation by d_z radians around an axis parallel to $\boldsymbol{l_z}$ each atom is shifted by s_z Ångstroms along this axis; a similar definition is

used for two other parameters. Such a correlation generates an additional contribution $C_{[L]}$ to the T matrix that arises from screw motions

$$T_{C[L]} = V_{[L]} + C_{[L]} = V_{[L]} + \begin{pmatrix} s_x^2\left\langle d_x^2\right\rangle & 0 & 0 \\ 0 & s_y^2\left\langle d_y^2\right\rangle & 0 \\ 0 & 0 & s_z^2\left\langle d_z^2\right\rangle \end{pmatrix} \tag{7}$$

and also results in a non-zero S matrix

$$S_{[L]} = \begin{pmatrix} s_x\left\langle d_x^2\right\rangle & 0 & 0 \\ 0 & s_y\left\langle d_y^2\right\rangle & 0 \\ 0 & 0 & s_z\left\langle d_z^2\right\rangle \end{pmatrix} \tag{8}$$

Finally, the principal libration axes do not necessarily pass through the origin, or even have a common point (*i.e.* may not intersect). If they pass through the points $\mathbf{w}_{[L]}^{lx} = \left(w_x^{lx}, w_y^{lx}, w_z^{lx}\right)$, $\mathbf{w}_{[L]}^{ly} = \left(w_x^{ly}, w_y^{ly}, w_z^{ly}\right)$, $\mathbf{w}_{[L]}^{lz} = \left(w_x^{lz}, w_y^{lz}, w_z^{lz}\right)$, respectively, this generates an additional component to the T matrix

$$T_{[L]} = T_{C[L]} + D_{W[L]} \tag{9}$$

where

$$D_{W[L]} = \tag{10}$$

$$\begin{pmatrix} \left(w_z^j\right)^2\left\langle d_y^2\right\rangle + \left(w_y^k\right)^2\left\langle d_z^2\right\rangle & -w_x^k w_y^k\left\langle d_z^2\right\rangle & -w_x^j w_z^j\left\langle d_y^2\right\rangle \\ -w_x^k w_y^k\left\langle d_z^2\right\rangle & \left(w_z^i\right)^2\left\langle d_x^2\right\rangle + \left(w_x^k\right)^2\left\langle d_z^2\right\rangle & -w_y^i w_z^i\left\langle d_x^2\right\rangle \\ -w_x^j w_z^j\left\langle d_y^2\right\rangle & -w_y^i w_z^i\left\langle d_x^2\right\rangle & \left(w_y^i\right)^2\left\langle d_x^2\right\rangle + \left(w_x^j\right)^2\left\langle d_y^2\right\rangle \end{pmatrix}$$

Taking into account both the screw motion and the position of the libration axes, the matrix S becomes

$$S_{[L]} = \begin{pmatrix} s_x \langle d_x^2 \rangle & w_z^i \langle d_x^2 \rangle & -w_y^i \langle d_x^2 \rangle \\ -w_z^j \langle d_y^2 \rangle & s_y \langle d_y^2 \rangle & w_x^j \langle d_y^2 \rangle \\ w_y^k \langle d_z^2 \rangle & -w_x^k \langle d_z^2 \rangle & s_z \langle d_z^2 \rangle \end{pmatrix} \quad (11)$$

(if the axes pass through the origin this matrix is the diagonal matrix (8)). Finally, the matrices in the original basis [M] where they are reported together with the atomic coordinates, are obtained from $L_{[L]}$ (5), $T_{[L]}$ (9), $S_{[L]}$ (11) as

$$\begin{aligned} L_{[M]} &= R_{ML} L_{[L]} R_{ML}^{\tau} \\ T_{[M]} &= R_{ML} T_{[L]} R_{ML}^{\tau} \\ S_{[M]} &= R_{ML} S_{[L]} R_{ML}^{\tau} \end{aligned} \quad (12)$$

Here R_{ML} is the transition matrix from the basis [M] to the basis [L] (Appendix A).

Molecular basis and center of reaction

The *TLS* matrices also depend on the choice of the origin. Clearly, the coordinates of the position of the libration axes change as function of the origin. Usually the origin is taken to be the center of mass of the atomic group or in the point where the mean atomic displacements are similar in magnitude to each other due to librations around each of the principal axes. This second point is called the center of diffusion (Brenner, 1967) or the center of reaction (Tickle & Moss, 1999). Choosing the origin at the center of reaction minimizes the trace of *T* and makes *S* symmetric (Brennen, 1967; Tickle & Moss, 1999; UAA-2013). Shifting from one origin to another changes *T* and *S* but does not change *L* and does not modify the algorithm of the search for the composite motions. In what follows we consider the matrices to be in their original basis (for example as they are defined in the PDB).

Calculating elemental motions from *TLS* matrices: libration axes

This section provides a step-by-step explanation of the inverse problem, *i.e.* calculating the vibration and libration axes and the corresponding *rmsd*s, the position of the libration axes, and the parameters describing the correlations between librations and vibrations from given *TLS* matrices.

Diagonalization of the L matrix ([L] basis; step A)

Suppose that we know the elements of the matrices (12) in the basis [M]. By construction, the matrices *T* and *L* should be positive semidefinite (Appendix B) and symmetric, $T_{[M]xy} = T_{[M]yx}$, $T_{[M]xz} = T_{[M]zx}$, $T_{[M]yz} = T_{[M]zy}$ and $L_{[M]xy} = L_{[M]yx}$, $L_{[M]xz} = L_{[M]zx}$, $L_{[M]yz} = L_{[M]zy}$. These properties remain such for any rotation of the coordinate system, *i.e.* in any Cartesian basis; this is important for further analysis of the *T* matrices.

We start the procedure from the matrix $L_{[M]}$, which depends only on the libration parameters. The principal libration axes correspond to its three mutually orthogonal eigenvectors. First we search for the corresponding eigenvalues $0 \le \lambda_1 \le \lambda_2 \le \lambda_3$, which must be non-negative (see expression (5); eigenvalues do not change with the coordinate system). Let $\boldsymbol{l}_1$, $\boldsymbol{l}_2$, $\boldsymbol{l}_3$ be the corresponding normalized eigenvectors from which we construct a new basis [L] as

$$\boldsymbol{l}_x = \pm\boldsymbol{l}_1,\ \boldsymbol{l}_y = \boldsymbol{l}_2,\ \boldsymbol{l}_z = \boldsymbol{l}_3 \qquad (13)$$

The appropriate sign for $\boldsymbol{l}_x$ is chosen so that the vectors in (13) form a right-hand triad; for example one can take $\boldsymbol{l}_x = \boldsymbol{l}_y \times \boldsymbol{l}_z$ that guarantees such the condition. The *TLS* matrices in the [L] basis are

$$\begin{aligned} L_{[L]} &= R_{ML}^{\tau} L_{[M]} R_{ML} \\ T_{[L]} &= R_{ML}^{\tau} T_{[M]} R_{ML} \\ S_{[L]} &= R_{ML}^{\tau} S_{[M]} R_{ML} \end{aligned} \qquad (14)$$

where R_{ML} is the transition matrix from the basis [M] into basis [L] (Appendix A). In this new basis matrix $L_{[L]}$ is diagonal with the elements $L_{[L]xx} = \lambda_1$, $L_{[L]yy} = \lambda_2$, $L_{[L]zz} = \lambda_3$, giving the estimates $\langle d_x^2 \rangle = L_{[L]xx}$, $\langle d_y^2 \rangle = L_{[L]yy}$, $\langle d_z^2 \rangle = L_{[L]zz}$ of the squared libration amplitudes around the

three principal libration axes.

Position of the libration axes in the [L] basis (step B)

In the basis [L] the libration axes are parallel to the coordinate axes but do not necessarily coincide with them. Let them pass through some points $\mathbf{w}^{lx}$, $\mathbf{w}^{ly}$, $\mathbf{w}^{lz}$, respectively, that we are looking for. Using equation (11) we calculate the coordinates of these points as

$$
\begin{aligned}
w^{lx}_{[L]y} &= -\frac{S_{[L]xz}}{L_{[L]xx}},\ w^{lx}_{[L]z} = \frac{S_{[L]xy}}{L_{[L]xx}},\\
w^{ly}_{[L]x} &= \frac{S_{[L]yz}}{L_{[L]yy}},\ w^{ly}_{[L]z} = -\frac{S_{[L]yx}}{L_{[L]yy}},\\
w^{lz}_{[L]x} &= -\frac{S_{[L]zy}}{L_{[L]zz}},\ w^{lz}_{[L]y} = \frac{S_{[L]zx}}{L_{[L]zz}}
\end{aligned}
\qquad (15)
$$

A zero value for any denominator in (15) means that there is no rotation around the corresponding axis; in this case the two corresponding numerator values must also be equal to zero and thus assign zero values to the corresponding coordinates in (15); otherwise the input matrices are incompatible and the procedure must stop (Appendix B). The *x*-component of $\mathbf{w}^{lx}$, *y*-component of $\mathbf{w}^{ly}$ and *z*-component of $\mathbf{w}^{lz}$ in the basis [L] can be any values. For presentation purposes it might be useful to assign them as

$$
\begin{aligned}
\mathbf{w}^{lx}_{[L]x} &= \tfrac{1}{2}\left(\mathbf{w}^{ly}_{[L]x} + \mathbf{w}^{lz}_{[L]x}\right),\ \mathbf{w}^{ly}_{[L]y} = \tfrac{1}{2}\left(\mathbf{w}^{lx}_{[L]y} + \mathbf{w}^{lz}_{[L]y}\right),\\
\mathbf{w}^{lz}_{[L]z} &= \tfrac{1}{2}\left(\mathbf{w}^{lx}_{[L]z} + \mathbf{w}^{lz}_{[L]z}\right)
\end{aligned}
\qquad (16)
$$

that will position each of these points in the middle of the two other axes.

Knowing the positions (15) – (16) of the libration axes and elements of $L_{[L]}$ we can calculate the contribution $D_{W[L]}$ (10) from an apparent translation due to the displacement of the libration axes from the origin. Then inverting (9) we can calculate the residual matrix $T_{C[L]}$ after removal of this contribution:

$$T_{C[L]} = T_{[L]} - D_{W[L]}$$

(17)

Matrix (17) must be positive semidefinite (Appendix B) as it is a sum (7) of two positive semidefinite matrices. Matrices $S_{[L]}$ and $L_{[L]}$ are not modified at this step.

Calculating elemental motions from *TLS* matrices: screw components (step C)

Correlation between libration and vibration and a choice of the diagonal elements of S

Next we use the matrices $L_{[L]}$ and $S_{[L]}$ to determine the screw parameters s_x, s_y, s_z, remove the screw contribution from the $T_{C[L]}$ matrix using equations (7) and (17), and finally extract the matrix $V_{[L]}$ for uncorrelated vibrations. However, there is an ambiguity in the definition of $S_{[L]}$ which is apparent from the observation that the matrices $U_{concerted,n}$ of individual atoms will not change if the same number t is added or removed simultaneously from all three diagonal elements of $S_{[L]}$. This is usually known as indetermination of the trace of this matrix. The choice of this number has been discussed by Schomaker & Trueblood (1968). A current practice (Section 6.1 provides an illustration) is to choose t such that it minimizes the trace (rather its absolute value) of the resulting matrix

$$S_C(t) = S_{[L]} - tI$$

(18)

(where I is a unit matrix), *i.e.* minimizing vibration-libration correlation (UAA-2013). The unconditioned minimization

$$\min_t \left| S_C(t) \right| = \min_t \left| \left(S_{[L]xx} + S_{[L]yy} + S_{[L]zz} \right) - 3t \right| = 0$$

(19)

gives

$$t_0 = \tfrac{1}{3}\left(S_{[L]xx} + S_{[L]yy} + S_{[L]zz} \right) = \tfrac{1}{3} tr S_{[L]}$$

(20)

However, this value may lead to physically unrealistic matrices for which libration-vibration

decomposition is impossible. Intuitively, if the elements of matrix S and the corresponding values s_x, s_y, s_z are too large, the matrix V in (7) may be not positive definite for a given $T_{C[L]}$. The next sections describe a procedure that defines the constraints on the diagonal elements of matrix S when using (18).

Cauchy-Schwarz conditions

After removing $D_{W[L]}$ (eq. (17)), the set of matrices $T_{C[L]}, L_{[L]}$ and the matrix $S_{[L]}$ with the removed off-diagonal elements (reducing the matrix (11) to the form (8)) correspond to a combination of vibrations with screw rotations around the axes crossing the origin. The diagonal elements of these matrices must satisfy the Cauchy-Schwarz inequality (Appendix A)

$$S_{[L],xx}^2 \le T_{C[L]xx} L_{[L]xx}; \quad S_{[L],yy}^2 \le T_{C[L]yy} L_{[L]yy}; \quad S_{[L],zz}^2 \le T_{C[L]zz} L_{[L]zz} \tag{21}$$

that in turn defines the conditions (Appendices A and B):

$$\begin{aligned} \left(S_{[L]xx} - t\right)^2 &\le T_{C[L]xx} L_{[L]xx} \\ \left(S_{[L]yy} - t\right)^2 &\le T_{C[L]yy} L_{[L]yy} \\ \left(S_{[L]zz} - t\right)^2 &\le T_{C[L]zz} L_{[L]zz} \end{aligned} \tag{22}$$

or

$$t_{\min,C} \le t \le t_{\max,C} \tag{23}$$

with

$$\begin{aligned} t_{\min,C} &= \max\left\{S_{[L]xx} - r_x; S_{[L]yy} - r_y; S_{[L]zz} - r_z\right\} \\ t_{\max,C} &= \min\left\{S_{[L]xx} + r_x; S_{[L]yy} + r_y; S_{[L]zz} + r_z\right\} \end{aligned} \tag{24}$$

$$r_x = \left(\tfrac{1}{2} T_{C[L]xx} L_{[L]xx}\right)^{1/2}, \; r_y = \left(\tfrac{1}{2} T_{C[L]yy} L_{[L]yy}\right)^{1/2}, \; r_z = \left(\tfrac{1}{2} T_{C[L]zz} L_{[L]zz}\right)^{1/2}$$

In particular, this unambiguously defines the t value if one of the diagonal elements of the matrix $L_{[L]}$ is zero so that the trace of $S_{[L]}$ cannot be changed or assigned arbitrarily (see section 4.4).

Positive semidefinition of the V matrix

The last condition to check is that the matrix V is positive semidefinite. Let us suppose that all diagonal elements of the matrix $L_{[L]}$ are different from zero; section 4.4 considers the alternative case. From equations (5), (7), (8) and (18) we find the expression for the screw contribution

$$C_{[L]}(t) = \begin{pmatrix} S^2_{C,xx} L^{-1}_{[L]xx} & 0 & 0 \\ 0 & S^2_{C,yy} L^{-1}_{[L]yy} & 0 \\ 0 & 0 & S^2_{C,zz} L^{-1}_{[L]zz} \end{pmatrix} = \tag{25}$$

$$= \begin{pmatrix} \left(S_{[L]xx} - t\right)^2 L^{-1}_{[L]xx} & 0 & 0 \\ 0 & \left(S_{[L]yy} - t\right)^2 L^{-1}_{[L]yy} & 0 \\ 0 & 0 & \left(S_{[L]zz} - t\right)^2 L^{-1}_{[L]zz} \end{pmatrix}$$

to be subtracted from matrix (17) as

$$V_{[L]} = T_{C[L]} - C_{[L]}(t) \quad .$$

(26)

Matrix $V_{[L]}$ is positive semidefinite along with

$$V_\Lambda = \begin{pmatrix} v_{XX} & v_{XY} & v_{XZ} \\ v_{YX} & v_{YY} & v_{YZ} \\ v_{ZX} & v_{ZY} & v_{ZZ} \end{pmatrix} = \Lambda^\tau V_{[L]} \Lambda = \Lambda^\tau T_{C[L]} \Lambda - \Lambda^\tau C_{[L]}(t) \Lambda = T_\Lambda - C_\Lambda(t)$$

(27)

where

$$\Lambda = \Lambda^\tau = \begin{pmatrix} L^{1/2}_{[L]xx} & 0 & 0 \\ 0 & L^{1/2}_{[L]yy} & 0 \\ 0 & 0 & L^{1/2}_{[L]zz} \end{pmatrix}$$

(28)

$$T_\Lambda = \begin{pmatrix} T_{C[L]xx}L_{[L]xx} & T_{C[L]xy}L_{[L]xx}^{1/2}L_{[L]yy}^{1/2} & T_{C[L]xz}L_{[L]xx}^{1/2}L_{[L]zz}^{1/2} \\ T_{C[L]yx}L_{[L]xx}^{1/2}L_{[L]yy}^{1/2} & T_{C[L]yy}L_{[L]yy} & T_{C[L]yz}L_{[L]yy}^{1/2}L_{[L]zz}^{1/2} \\ T_{C[L]zx}L_{[L]xx}^{1/2}L_{[L]zz}^{1/2} & T_{C[L]zy}L_{[L]yy}^{1/2}L_{[L]zz}^{1/2} & T_{C[L]zz}L_{[L]zz} \end{pmatrix}$$

(29)

$$C_\Lambda(t) = \begin{pmatrix} (S_{[L]xx}-t)^2 & 0 & 0 \\ 0 & (S_{[L]yy}-t)^2 & 0 \\ 0 & 0 & (S_{[L]zz}-t)^2 \end{pmatrix}$$

(30)

Necessarily, all diagonal terms of (30) cannot be larger than the maximal eigenvalue $\tau_{\max}$ of matrix (29) giving a necessary condition (Appendix B)

$$t_{\min,\tau} = \max\{S_{[L]xx}; S_{[L]yy}; S_{[L]zz}\} - \tau_{\max}^{1/2} \le t \le \min\{S_{[L]xx}; S_{[L]yy}; S_{[L]zz}\} + \tau_{\max}^{1/2} = t_{\max,\tau}. \quad (31)$$

Another condition (all diagonal terms of (30) are not larger than the minimum eigenvalue $\tau_{\min}$ of (29)) is sufficient but not necessary.

Matrix V_Λ is positive semidefinite if and only if all three of its real eigenvalues are non-negative (some of them may coincide with each other). They are the roots of the cubic characteristic equation

$$\nu^3 + a_S\nu^2 + b_S\nu + c_S = 0$$

(32)

with the coefficients

$$a_S(t) = -trV_\Lambda, \quad (33)$$

$$b_S(t) = \det\begin{pmatrix} v_{XX} & v_{XY} \\ v_{YX} & v_{YY} \end{pmatrix} + \det\begin{pmatrix} v_{YY} & v_{YZ} \\ v_{ZY} & v_{ZZ} \end{pmatrix} + \det\begin{pmatrix} v_{ZZ} & v_{ZX} \\ v_{XZ} & v_{XX} \end{pmatrix}$$

(34)

$$c_S(t) = -\det V_\Lambda$$

(35)

The roots of (32) are positive if and only if three inequalities below hold simultaneously:

$$a_S(t) \le 0 \quad , \quad b_S(t) \ge 0 \quad , \quad c_S(t) \le 0 \tag{36}$$

where the left parts are polynomials of order 2, 4 and 6 of the parameter t, all with the unit highest-order coefficient (Appendix A). The first condition in (36) defines the interval for t values (Appendix B):

$$t_{\min,a} = t_0 - t_a \le t \le t_{\max,a} = t_0 + t_a \tag{37}$$

with

$$t_a = \left[t_0^2 + \tfrac{1}{3} trT_\Lambda - \tfrac{1}{3}\left(S^2_{[L]xx} + S^2_{[L]yy} + S^2_{[L]zz} \right) \right]^{1/2} \tag{38}$$

We failed to find analytical expressions corresponding to the two other inequalities. As a result, the following numerical procedure is suggested to find the best t value that is physically acceptable:

a) Calculate t_0 value (equation (20));

b) Calculate the interval $(t_{\min}, t_{\max})$ for allowed t values as intersection of intervals (23), (31) and (39); $t_{\min} = \max\{t_{\min,C}, t_{\min,\tau}, t_{\min,a}\}$, $t_{\max} = \min\{t_{\max,C}, t_{\max,\tau}, t_{\max,a}\}$; if $t_{\min} > t_{\max}$ the problem has no solution and the procedure stops (Appendix B);

c) If $t_{\min} = t_{\max}$ we check the conditions $b_S(t_{\min}) \ge 0$, $c_S(t_{\min}) \le 0$, or the condition that V_Λ is positive semidefinite; if the conditions are satisfied we assign $t_S = t_{\min}$ otherwise the problem has no solution and the procedure stops (Appendix B);

d) If $t_{\min} < t_{\max}$ we search numerically, in a fine grid, for the point t_S in the interval $(t_{\min}, t_{\max})$ and closest to t_0 such that $b_S(t_S) \ge 0$, $c_S(t_S) \le 0$; if for any point of this interval at least one of these inequalities is wrong, then the procedure stops (Appendix B);

e) We accept the value obtained at the step *c* or *d* as the final t_S.

Singular sets of rotation

When one of the $L_{[L]xx}, L_{[L]yy}, L_{[L]zz}$ values is zero (that is there is no rotation around the corresponding axis), straightforward use of the standard procedure including (25) becomes impossible. However, in this case the t_S value must be equal to $S_{C,xx}$, $S_{C,yy}$ or $S_{C,zz}$, corresponding to the axes with no rotation, making the corresponding diagonal element in (25) equal to zero and turning the corresponding inequality in (24) into an equality. For example, if $L_{[L]xx} = 0$ then $t_S = S_{L,xx}$ resulting in $C_{[L]xx} = 0$. We simply need to check two other conditions in (21) and the condition that the residual matrix is positive semidefinite (for example by calculating (36)). If t_S does not satisfy these conditions, the problem has no solution (Appendix B).

Screw parameters

For the $t = t_S$ determined above we calculate the matrix $S_C(t_S)$ (18). From this matrix we obtain the screw parameters $s_x = S_{C,xx} L^{-1}_{[L]xx}$, $s_y = S_{C,yy} L^{-1}_{[L]yy}$, $s_z = S_{C,zz} L^{-1}_{[L]zz}$ for the rotation axes currently aligned with the coordinate axes of the basis [L]. If one of the $L_{[L]xx}, L_{[L]yy}, L_{[L]zz}$ values is equal to zero, the corresponding diagonal element of S_C must also be equal to zero (otherwise the matrices are inconsistent with each other and the procedure stops (Appendix B) and we assign the corresponding screw parameter, s_x, s_y or s_z to be zero.

Calculating elemental motions from *TLS* matrices: vibration components (step D)

Matrix V and vibration parameters in [L] basis

For the known t_S, matrix $C_{[L]}(t_S)$ and then $V_{[L]}$ are calculated as (25) - (26). The values of parameters of the independent vibrations are calculated from the $V_{[L]}$ matrix similarly to those for the independent librations, as we obtain them from $L_{[M]}$. First, we calculate the three eigenvalues $0 \le \mu_1 \le \mu_2 \le \mu_3$ of matrix $V_{[L]}$ (Appendix B; in practice, all of them are strictly

positive). Then we identify three corresponding unit eigenvectors $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3$ that are orthogonal to each other and assign

$$\mathbf{v}_x = \pm\mathbf{v}_1, \mathbf{v}_y = \mathbf{v}_2, \mathbf{v}_z = \mathbf{v}_3 \quad (39)$$

(the sign for $\mathbf{v_x}$ is taken so that the vectors (39) form a right-hand triad). We remind that these axes define the basis [V] in which matrix $V_{[V]}$ (6) is diagonal with elements $V_{[V]xx} = \mu_1$, $V_{[V]yy} = \mu_2$, $V_{[V]zz} = \mu_3$ defining the last missing parameters, namely the values of the squared *rmsd*s along these axes: $\langle t_x^2 \rangle = V_{[V]xx}$, $\langle t_y^2 \rangle = V_{[V]yy}$, $\langle t_z^2 \rangle = V_{[V]zz}$.

Vibration and libration axes in [M] basis

The libration and vibration amplitudes and screw parameters are independent of the choice of the basis, and the direction of the libration axes is known in the principal [M] basis. However the directions of the uncorrelated translations $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3$ that were calculated in section 4 and the points $\mathbf{w}_{[L]}^{lx}$, $\mathbf{w}_{[L]}^{ly}$, $\mathbf{w}_{[L]}^{lz}$ belonging to the libration axes (section 3.2) are now known in the [L] basis.

To obtain the coordinates $\left(w_{[M]x}^{lx}, w_{[M]y}^{lx}, w_{[M]z}^{lx}\right)$, $\left(w_{[M]x}^{ly}, w_{[M]y}^{ly}, w_{[M]z}^{ly}\right)$, $\left(w_{[M]x}^{lz}, w_{[M]y}^{lz}, w_{[M]z}^{lz}\right)$ of these points in the [M] basis we apply the transformation

$$\mathbf{w}_{[M]}^{lx} = R_{ML}\mathbf{w}_{[L]}^{lx} \quad , \quad \mathbf{w}_{[M]}^{ly} = R_{ML}\mathbf{w}_{[L]}^{ly} \quad , \quad \mathbf{w}_{[M]}^{lz} = R_{ML}\mathbf{w}_{[L]}^{lz} \quad (40)$$

Similarly, the vectors defining the direction of the axes $\mathbf{v}_x, \mathbf{v}_y, \mathbf{v}_z$ in the basis [M] can be obtained as

$$\mathbf{v}_{[M]x} = R_{ML}\mathbf{v}_{[L]x} \quad , \quad \mathbf{v}_{[M]y} = R_{ML}\mathbf{v}_{[L]y} \quad , \quad \mathbf{v}_{[M]z} = R_{ML}\mathbf{v}_{[L]z} \quad (41)$$

This step finalizes extracting the parameters of the motions that correspond to the given set of *TLS* matrices. Section 6 provides some examples of this procedure applied to models deposited in the PDB. Section 7 describes an example of when knowledge of the motion

parameters extracted from the *TLS* matrices is necessary to explicitly simulate the ensemble of corresponding structures in order to simulate diffuse scattering.

2.4 Examples of the *TLS* matrix analysis

As pointed out in the Introduction, there are numerous examples of fruitful application of the *TLS* formalism to structural studies. The goal of this section is to illustrate the algorithm described above, describe possible traps and discuss further developments.

Survey of available TLS matrices in the PDB

We have analyzed the *TLS* matrices available in the PDB. From the overall 106,761 entries (as of March 2015), 25,904 use TLS modeling. More than 20,000 of these systems have several *TLS* groups, resulting in a total of 203,261 sets of *TLS* matrices (Fig 2.5a) with the largest number of groups per entry being 283 (PDB code 3u8m). About a third of these sets have negative eigenvalues for the deposited *T* or *L* matrices. Some of these values are only slightly negative (Figs. 2.5b and 2.5c) and can be considered rounding errors, while the worst values are as small as -0.28 radians^2 for *L* and -20.72 Å^2 for *T*. For 11,412 *T* matrices and 138 *L* matrices all three eigenvalues are negative.

Another third of the *TLS* groups cannot be interpreted by elemental motions due to other reasons described in sections 3-4 (Table 2.1).

After an initial screen to find the positive *T* and *L* matrices, we then ran a search for the elemental motions in two modes. First, we tried to decompose the *TLS* matrices as taken directly from the PDB files. As expected, the average value of *trS* is $3\cdot10^{-5}$ Å , (*i.e.* practically zero) and the corresponding *rms* is $s = 10^{-2}$ Å. About 120,000 *S* matrices have $|\ trS\ | < 10^{-4}$ Å. The number of the matrices with $|\ trS\ |$ larger than 1*s*, 3 *s*, 10 *s* and 20 *s* Å is only 3772, 486, 31 and 3, respectively. We then applied the aforementioned algorithm with the optimal choice of the value t_S to be subtracted from the diagonal *S* elements in each case.

Table 2.1 shows the results of both runs and illustrates that we can fix the problems found in 6,500 of the *TLS* sets (corresponding to about 500 PDB entries) by a correction of the diagonal elements of the *S* matrix as described above. The Table takes into account possible rounding

errors by correcting slightly negative eigenvalues (those closer in value to 0 than 10^{-5} of the default units Å^2, radians^2, Å·radians for *T*, *L*, *S*, respectively). For example, when running the algorithm in the *S*-optimizing mode the program can formally calculate the *V* matrix for about 70,000 sets. For 2296 cases this matrix has negative eigenvalues (Fig. 2.5d) while in 2294 of them these eigenvalues are closer to 0 than 10^{-5} Å^2; for such matrices the program makes automatic corrections and continues the process.

It is important to note that even if the parameters of the elemental motions can be formally extracted from the *TLS* matrices, this does not guarantee that they make physical sense. Clearly, the vibration amplitudes on the order of 20 Å^2 cannot represent harmonic vibrations (Fig. 2.5d). Similarly, the linear rotation approximation contained in *TLS* theory is valid only up to approximately 0.1 radians; much larger values can be found in the PDB (Fig. 2.5b). This is also true for the screw parameters. The products $s_x d_x, s_y d_y, s_z d_z$ show the mean shifts along the screw axes due to librations around these axes; the values found in the PDB approaching 3Å seem to be too large to describe harmonic motions.

For a more detailed analysis we selected several entries from the PDB. For each structure, we applied a standard *TLS* refinement protocol as implemented in *phenix.refine* (Afonine *et al*., 2012) including automatic determination of the *TLS* groups. During refinement, 20 matrix elements were refined independently; 6 for *T*, 6 for *L* and 8 for *S*; the three diagonal elements of *S* have been constrained such that the trace of the matrix is equal to 0. The procedure described above (Sections 3-5) was then applied to all sets of obtained *TLS* matrices.

We remind a reader that the elements of the *L* and *S* matrices are expressed in radians^2 and Å·radians while in the PDB files they are kept in degrees^2 and in Å·degrees, respectively; this makes their mantissa smaller and thus more convenient for archiving.

Synaptotagmin

The crystals of synaptotagmin III (PDB code 1dqv) contain two copies of the molecule in the asymmetric unit. The structure after re-refinement by *phenix.refine* without *TLS* modeling has R_{work} = 0.200 and R_{free} = 0.231 at a resolution of 2.5 Å. Performing *TLS* refinement with each molecule taken as a single *TLS* group reduced *R*-factors to R_{work} = 0.177 and R_{free} = 0.211 indicating this additional modeling significantly improves agreement with the experimental data. Table 2.2 shows the two sets of matrices and Table 2.3 contains the corresponding motion

parameters extracted using our approach. For the two groups both vibrations and librations are practically isotropic and are of the same order of magnitude. Fig. 2.2a shows the principal axes of these motions.

Calmodulin

The structure of calmodulin (PDB code 1exr) was determined previously at a resolution of 1.0 Å. We use this example to illustrate possible problems that could be solved by a minimal correction of the *TLS* values. For re-refinement with *phenix.refine* the model was automatically split into four *TLS* groups. For the first group, one of the eigenvalues of the matrix *L* was equal to $-0.23 \cdot 10^{-4}$ radian^2. If we consider this value to be zero (in this case the zero value must be also assigned to off-diagonal elements of the first row of the matrix *S*), all composite motions can be extracted routinely with only two libration axes. Corresponding modifications of the resulting matrices $U_{group,n}$ (2) can be compensated for by respective modification of the individual contributions $U_{local,n}$. This keeps the total ADP parameters $U_{Cart,n}$ unchanged, which maintains the previously calculated structure factors and *R*-factors. An accurate separation of total atomic displacement parameter values into contributions from several sources (for example, Murshudov *et al*., 1999; Winn *et al*., 2001, 2003; Afonine *et al*., 2013) is a separate ongoing project (Afonine & Urzhumtsev, 2007).

For the second *TLS* group, the refined *TLS* matrix elements contained one degenerate libration. The procedure described in sections 3-5 was successfully applied. Note that this procedure modified the diagonal elements of the matrix *S*, removing an appropriate value of the parameter t_S (Section 4.4) and making *trS* non-zero.

For the third group, the screw parameters were extremely large, leading to the procedure's inability to find a positive semidefinite $V_{[L]}$ (27). All three eigenvalues of the matrix *L* were extremely small (0.0, $0.08 \cdot 10^{-4}$ and $0.35 \cdot 10^{-4}$ radians^2), resulting in high computational instability. If we replace matrix *L* (and respectively *S*) by zero matrices, defining all librations to be absent, the vibration parameters can easily be found from *T*. In fact, this *TLS* group is a helix held at both ends by large domains which leads to the expectation of a pure vibration motion.

Finally, for the fourth group one of the diagonal elements of the matrix *T* was marginally negative. Increasing all diagonal elements of the matrix *T* by 0.002 Å^2 makes this matrix positive

definite (this corresponds to B = 0.16 Å^2). As discussed above, this adjustment can be compensated for by removing the equivalent amount from individual atomic contributions $U_{local,n}$ (assuming such subtraction keeps the individual atomic contributions positive). This group vibrates in a plane (Fig. 2.2b) and the principal vibration axis of group 3 (the helix) is parallel to this plane, leading to the plausible hypothesis that groups 3 and 4 at least partially move together or slide one with respect to another.

To check the influence of the manual modification on the *TLS* matrices, we recalculated the *R*-factors before and after performing these changes without updating the individual atomic contributions $U_{local,n}$. For all modifications described above, including the ensemble of modifications applied together, the *R*-factors varied only in the 4th significant digit.

This example demonstrates that current refinement procedures may result in *TLS* matrices that do not satisfy the previously mentioned conditions. However, small changes to refined *TLS* matrices may be sufficient correction. This highlights the need to use appropriate restraints or constraints on refinable parameters within the *TLS* model.

Initiation translation factor 2 (IF2)

The structure of IF2 (PDB code 4b3x) has recently been solved in one of our laboratories (Simonetti *et al*., 2013) with R_{work} = 0.180 and R_{free} = 0.219 at a resolution of 1.95 Å. *A posteriori TLS* refinement was done with two groups: the first group included the N-terminal and the following long helix, and the second included the rest of structure. The re-refined model had better R_{work} = 0.176 and especially R_{free}= 0.203. In this example the *TLS* matrices from the first group were not directly interpretable because the residual matrix $V_{[L]}$ was not positive semidefinite (the minimal eigenvalue was -0.05). Similarly to the last *TLS* group in calmodulin, we artificially added 0.06 Å^2 to all diagonal elements of the matrix *T*, corresponding to roughly 5 Å^2 (the same amount has been removed from the residual atomic *B* values, thus leaving the *R*-factors unchanged). This correction allowed for interpretation of the *TLS* matrices in terms of elemental motions. We note that for the first *TLS* group one of the rotations was degenerate and that the assignment *trS* = 0 would make this matrix incompatible with *L*. Table 2.3 shows that vibrations of this group are essentially anisotropic. Fig. 2.2c also shows that the libration axes for

this group pass quite far away from the molecule, which makes the corresponding rotation similar to a translation. Additionally, we believe that the large s_z value indicates that the matrix *S* is not well defined to be physically significant in this case. The matrices for the second group were interpreted and revealed isotropic vibrations and librations.

Finally, we tried to apply the same procedure after choosing the *TLS* groups manually as residues 1-50 (N terminal), 51-69 (helix), 70-333 (G domain) and 343-363 (connector to the C domain absent in this structure). As before, the matrices were physically interpretable for the G domain. For the groups 2 and 4, after an adjustment similar to those discussed above (a slight increasing the diagonal *T* elements with decreasing the residual *B* values of the individual atoms), we obtained a pure vibration for the helix (as for the calmodulin case) and a libration around a single axis for the terminal group. In contrast, we failed to find reasonably small corrections for the matrices of the first group that would make them physically interpretable.

This case exemplifies a situation when the current *TLS* refinement protocol resulted in matrices (a few of which are physically nonsensical) that significantly reduced the *R*-factors without providing physically meaningful refined *TLS* parameters for one of the groups. This again highlights the need to improve *TLS* refinement algorithms.

2.5 Interpreting *TLS* matrices with a structural ensemble

Generation of an explicit set of atomic models with a variability consistent with TLS

Some structural problems may explicitly require a set of models that describe a given mobility, *e.g.* corresponding to the *TLS* matrices for harmonic motion (here we exclude larger-scale anharmonic motions for which other techniques such as molecular dynamic trajectories have traditionally been used, as originated by McCammon *et al*., 1977). An example of such a problem is described in Van Benschoten *et al*. (2015), and briefly presented in Section 7.4, in which X-ray diffuse scattering data were compared with calculated data corresponding to different types of molecular motion.

As soon as a combination of vibrations and librations is extracted from the *TLS* matrices, we can explicitly build a corresponding set of models. If a model deposited in the PDB contains *TLS* matrices, the matrices can be decomposed as described above. A decomposition of this

motion into three vibrations and three librations provides the atomic shifts underlying the total displacement.

It is generally more convenient to generate each group of atomic shifts in its own basis: shifts $\Delta^{V}_{[V]}\mathbf{r}_n$ due to vibration in the [V] basis and shifts $\Delta^{L}_{[L]}\mathbf{r}_n$ due to libration in the [L] basis. Here we are working in a linear approximation such that rotation angles are on the order of 0.1 radian. For each particular set of generated shifts, they are transformed into the [M] basis as $\Delta^{V}_{[M]}\mathbf{r}_n$ and $\Delta^{L}_{[M]}\mathbf{r}_n$ and their sum

$$\Delta_{[M]}\mathbf{r}_n = \Delta^{L}_{[M]}\mathbf{r}_n + \Delta^{V}_{[M]}\mathbf{r}_n \quad (42)$$

is applied to the corresponding atoms. Details of model generation are discussed in the next sections. This procedure is repeated independently multiple times, leading to structural models distributed according to the *TLS* matrices.

Calculation of the model shift due to libration

Let us suppose that we know the following axes in the basis [M]: the three mutually orthogonal axes $\boldsymbol{l}_\mathbf{x}$, $\boldsymbol{l}_\mathbf{y}$, $\boldsymbol{l}_\mathbf{z}$ for independent libration as well as the coordinates of the points $\mathbf{w}^{lx}_{[M]}$, $\mathbf{w}^{ly}_{[M]}$, $\mathbf{w}^{lz}_{[M]}$ that belong to each of them. We recalculate the coordinates of these points and the coordinates $\left(x_{[M]n}, y_{[M]n}, z_{[M]n}\right)$, n = 1, 2, …, N, of all atoms $\mathbf{r}_{[M]n}$ of the group into the [L] basis as

$$\mathbf{r}_{[L]n} = R^{-1}_{ML}\mathbf{r}_{[M]n} = R^{\tau}_{ML}\mathbf{r}_{[M]n} \quad (43)$$

(similar relations are for the points $\mathbf{w}^{lx}_{[L]}$, $\mathbf{w}^{ly}_{[L]}$, $\mathbf{w}^{lz}_{[L]}$). We remind the reader that the squared libration amplitudes $\left\langle d_x^2 \right\rangle = L_{[L]xx} = \lambda_1$, $\left\langle d_y^2 \right\rangle = L_{[L]yy} = \lambda_2$, $\left\langle d_z^2 \right\rangle = L_{[L]zz} = \lambda_3$ (Section 3.2) and the screw parameters s_x, s_y, s_z (Section 4.5) are independent of the basis.

For an atom at a distance R = 1 Å from the rotation axis, the probability of the shifts

d_x, d_y, d_z , which are numerically equal to the rotation angle in radians, are equal to:

axis parallel to $\boldsymbol{l_x}$: $P(d_x) = \sqrt{2\pi\lambda_1}\,\exp\left(-d_x^2 / 2\lambda_1\right)$

axis parallel to $\boldsymbol{l_y}$: $P(d_y) = \sqrt{2\pi\lambda_2}\,\exp\left(-d_y^2 / 2\lambda_2\right)$

(44)

axis parallel to $\boldsymbol{l_z}$: $P(d_z) = \sqrt{2\pi\lambda_3}\,\exp\left(-d_z^2 / 2\lambda_3\right)$

If one of the eigenvalues is equal to 0 then the corresponding *d* is equal to 0 with unit probability. The particular values of d_{x0}, d_{y0}, d_{z0} are obtained using a random number generator with an underlying normal distribution (44).

For each of the axes $\boldsymbol{l_x}$, $\boldsymbol{l_y}$, $\boldsymbol{l_z}$ for each atom *n* described by the vector $\mathbf{r}_n$ we calculate the coordinates, in the [L] basis, of its shifts $\Delta^{lx}_{[L]}\mathbf{r}_n$, $\Delta^{ly}_{[L]}\mathbf{r}_n$, $\Delta^{lz}_{[L]}\mathbf{r}_n$ due to the corresponding rotations by d_{x0}, d_{y0}, d_{z0} (Appendix A). The overall shift due to libration around the three axes is the sum

$$\Delta^{L}_{[L]}\mathbf{r}_n = \Delta^{lx}_{[L]}\mathbf{r}_n + \Delta^{ly}_{[L]}\mathbf{r}_n + \Delta^{lz}_{[L]}\mathbf{r}_n$$

(45)

It changes from one atom of the group to another and must be calculated for all atoms of the group with the same (d_{x0}, d_{y0}, d_{z0}) values for a particular instance of the three rotations.

To transform the atomic shift (45) from the [L] basis into the initial [M] basis, we invert equation (43):

$$\Delta^{L}_{[M]}\mathbf{r}_n = R_{ML}\Delta^{L}_{[L]}\mathbf{r}_n$$

(46)

Calculation of the model shift due to vibration

In the harmonic approximation, the independent vibration shifts t_x, t_y, t_z expressed in the [V] basis are distributed accordingly to the probability laws:

$$P(t_x) = \sqrt{2\pi V_{[V]xx}}\,\exp\left(-t_x^2 / 2V_{[V]xx}\right) = \sqrt{2\pi\mu_1}\,\exp\left(-t_x^2 / 2\mu_1\right)$$

$$P(t_y) = \sqrt{2\pi V_{[V]yy}}\,\exp\left(-t_y^2 / 2V_{[V]yy}\right) = \sqrt{2\pi\mu_2}\,\exp\left(-t_y^2 / 2\mu_2\right)$$

(47)

$$P(t_z) = \sqrt{2\pi V_{[V]zz}}\,\exp\left(-t_z^2 / 2V_{[V]zz}\right) = \sqrt{2\pi\mu_3}\,\exp\left(-t_z^2 / 2\mu_3\right)$$

Using a random number generator, for each model we obtain particular values of t_{x0}, t_{y0}, t_{z0} using (47). If one of the eigenvalues *m* is equal to zero, the zero value is assigned to the corresponding shift. The overall translational shift, common to all atoms of the rigid group, is equal to

$$\Delta^V_{[V]}\mathbf{r}_n = t_{x0}\mathbf{v}_x + t_{y0}\mathbf{v}_y + t_{z0}\mathbf{v}_z. \qquad (48)$$

In order to obtain this shift in the [M] basis we calculate, similarly to equation (46),

$$\Delta^V_{[M]}\mathbf{r}_n = R_{MV}\Delta^V_{[V]}\mathbf{r}_n$$

(49)

Validation and application to GpdQ

We generated the ensembles produced by alternative *TLS* refinements of the glycerophosphodiesterase GpdQ (Jackson *et al.*, 2007). GpdQ is found in *Enterobacter aerogenes* and contributes to the homeostasis of the cell membrane by hydrolyzing the 3'-5' phosphodiester bond in glycerophosphodiesters. Each dimer contains three distinct domains per monomer: an a/b sandwich fold containing the active site, a domain-swapped active site cap and a novel dimerization domain comprised of dual-stranded antiparallel b-sheets connected by a small b-sheet. Due to the high global *B*-factors and presence of diffuse signal (Fig. 2.3), Jackson *et al*. (2007) performed three separate *TLS* refinements to model the crystalline disorder: *Entire molecule*, *Monomer* and *Sub-domain*. All *TLS* refinement attempts improved the R_{free} values when compared to the standard isotropic *B*-factor refinement; however, there was no significant difference among the final R_{free} values from the various *TLS* runs. We hypothesized that the diffuse scattering produced by each *TLS* motion would contain significant differences, as diffuse signal is a direct result of correlated motion. The notion that *TLS* refinement produces unique diffuse signal has been suggested previously (Tickle & Moss, 1999). Physical ensembles of the

TLS motion, rather than a mathematical description, were required to generate 3D diffuse scattering maps from *phenix.diffuse*. Visual inspection confirmed that the ensembles produced by *phenix.tls_as_xyz* replicated the anisotropic motion predicted by *TLS* thermal ellipsoids (Fig. 2.4). Additionally, we calculated the structure factors predicted by the original *TLS* refinement *Entire molecule* and compared them to F_{model} values (for example as defined in Afonine *et al*., 2012) produced by various *phenix.tls_as_xyz* ensemble sizes. The structure factors converged to a global correlation value of 0.965, demonstrating that *phenix.tls_as_xyz* ensembles accurately represent the motions predicted by *TLS* refinement. Physical representation of the underlying motion also revealed that, while two of the *TLS* refinements produced motion with small variances (a necessity within *TLS* theory), using each functional region as a *TLS* group produced fluctuations that are clearly non-physical (Fig. 2.3). Thus, viewing *TLS* refinement in the form of a structural ensemble is a valuable check of the validity of the results, as matrix elements that satisfy the previously described conditions may still produce motions that are clearly implausible.

2.6 Discussion

While our previous review on the subject (UAA-2013) describes the computational details of obtaining the *TLS* matrices from a known set of vibration and libration parameters (including the position of the axes and correlation of these motions), the current work focuses on the opposite problem of extracting these parameters from a given set of *TLS* matrices. The problem is not as simple as it may at first seem.

This difficulty arises because current structure refinement programs vary the matrix elements as independent parameters and often ignore critical constraints on real-space motions. A second difficulty is that identical motions may be represented by different vibration-libration combinations. As a consequence, there is no one-to-one relationship between these parameters and the set of *TLS* matrices. In particular, the traditional way of choosing the matrix *S* so that its trace is equal to zero may result in a mutually inconsistent combination of *TLS* matrices.

This manuscript describes the constraints that should be used to validate a given set of *T*, *L* and *S* matrices. Beyond the well-known conditions of non-negativity for the eigenvalues of *T* and *L*, we also discuss the conditions that relate the matrices, a crucial step in ensuring that the

results of *TLS* refinement correspond to physically possible combinations of librations and vibrations.

The survey of PDB entries with *TLS* matrices available revealed that roughly 85% of deposited models contain physically nonsensical *TLS* matrices (Table 2.1). This highlights two urgent needs. The first step must be changing existing refinement programs so that they apply appropriate restraints or constraints on refinable parameters of the *TLS* model. This must be followed by the implementation and use of comprehensive validation of *TLS* refinement results.

The utility of our presented algorithm is two-fold: it validates *TLS* matrices to confirm that they make physical sense and interprets *TLS* matrices in terms of the elemental motions that they describe. The information about atomic group motions conveyed by the TLS model can be used to analyze possible molecular mechanisms (as illustrated previously). Descriptions of TLS motion may also be used to generate an ensemble of molecular conformations, from which the predicted diffuse scattering signal can be calculated (Van Benschoten *et al*., 2015).

TLS matrix representation is a convenient way of encoding concerted motions into a form that is suitable for the calculation of structure factors and, in turn, structure refinement. There are two drawbacks to the standard implementation of this method. First, *TLS* matrices cannot be readily interpreted in terms of underlying motions but rather require additional processing in order for this information to be extracted. Secondly, direct refinement of the *TLS* matrix elements may result in refined matrices that do not make physical sense. To address these two drawbacks we propose using the set of vibration and libration parameters as refinable variables (an ongoing project for the authors) and reporting them in the PDB files. Indeed, using actual motion descriptors as refinement variables will allow for more effective application of physical constraints and in turn guarantee that refined values make physical sense. This will also simplify the analysis of refinement results, as they will be readily available for interpretation. Finally, this strategy would reduce data overfitting from physically nonsensical atomic models.

The current procedures for analysis and validation of *TLS* parameters, as well as the algorithm for generating a set of models from given libration and vibration parameters, are implemented in the *Phenix* suite and are called *phenix.tls_analysis* and *phenix.tls_as_xyz*, respectively; they are available starting with version dev-1890.

2.5 Tables

Table 2.1) Number of PDB entries with at least one physical conditions on *TLS* matrices broken. The statistics is shown for the matrices in the PDB (25904 entries with the *TLS* matrices from the total number of 106,761 entries, as of March 2015) with the default condition $trS = 0$ (upper line) and with the optimal choice of the diagonal S elements whenever possible as described in Sections 3-4 (bottom line). The conditions are, from left to right: matrices T and L are positive semidefinite ($T \geq 0$ & $L \geq 0$); an absence of libration around one of the axes requires the corresponding elements of the S matrix to be equal to 0 (s=0 & w=0); matrix T is positive semidefinite after the contribution due to the displacement of libration axes is removed ($T_C \geq 0$); elements of the S matrix are limited by the corresponding elements of the T and L matrices accordingly to the Cauchy conditions ($S \leq TL$); residual V matrix is positive semidefinite ($V \geq 0$). The column ($V \geq 0$) includes all conditions from Sections 4.3 - 4.4. When one of the conditions is broken further conditions were not checked.

Mode	Total PDB number	Total *TLS* number	Conditions broken					Total *TLS* broken	Total *TLS* OK	Total PDB broken
			$T \geq 0$ & $L \geq 0$	s=0 & w=0	$T_C \geq 0$	$S \leq TL$	$V \geq 0$			
$t_S = 0$	25904	203261	71362	3104	52254	n/a	10492	137212	66049	22707
best t_S	25904	203261	71362	3104	52255	133	3776	130630	72631	22201

Table 2.2) Examples of the *TLS* matrices. The matrix elements extracted from the PDB files after refinement (section 6).

PDB code	chain, residue number	T (Å^2)	L (degree^2)	S (Å·degree)
1dqv	A1 – A97	0.1777 0.0090 -0.0044 0.0090 0.1306 0.0019 -0.0044 0.0019 0.1372	1.4462 -0.0160 -0.2656 -0.0160 1.2556 0.4713 -0.2656 0.4713 0.8689	0.0467 -0.0523 0.0566 0.1010 0.0032 -0.0164 0.0090 0.0188 0.0560
	B1 – B97	0.1777 0.0090 -0.0044 0.0090 0.1306 0.0019 -0.0044 0.0019 0.1372	1.4462 -0.0160 -0.2656 -0.0160 1.2556 0.4713 -0.2656 0.4713 0.8689	0.0467 -0.0523 0.0566 0.1010 0.0032 -0.0164 0.0090 0.0188 0.0560
1exr	A2 – A30	0.0899 0.0040 -0.0004 0.0040 0.1333 0.0058 -0.0004 0.0058 0.0728	1.3491 -0.3760 -0.3971 -0.3760 0.6103 -0.3389 -0.3971 -0.3389 0.3698	-0.0249 -0.3537 -0.0874 0.1275 0.0783 -0.0144 0.0183 0.0542 -0.0103
	A31 - A74	0.0925 0.0037 0.0041 0.0037 0.0673 0.0062 0.0041 0.0062 0.1119	0.3464 0.3638 0.2923 0.3638 0.3283 0.1212 0.2923 0.1212 0.3799	-0.0220 -0.0419 -0.0793 -0.0061 0.0018 0.1161 -0.0041 -0.0385 -0.0009
	A75 – A84	0.2433 0.0144 0.0917 0.0144 0.2867 0.1720 0.0917 0.1720 0.1749	0.0736 0.0171 0.0565 0.0171 0.0068 -0.0203 0.0565 -0.0203 0.0336	0.4357 0.1151 0.2346 -0.2521 -0.3549 -0.2041 -0.3793 -0.1499 0.0111
	A85 – A147	0.0747 -0.0110 0.0066 -0.0110 0.1384 0.0062 0.0066 0.0062 0.0673	0.6097 -0.0786 -0.1864 -0.0786 0.6474 -0.6233 -0.1864 -0.6233 0.9637	0.0180 0.1466 0.0378 0.0155 -0.0872 -0.0542 -0.0440 0.1022 -0.0852
4b3x	A1 – A65	0.4663 0.0991 -0.0764 0.0991 0.5443 -0.0321 -0.0764 -0.0321 0.5001	0.4738 0.0063 0.2318 0.0063 0.2120 -0.0584 0.2318 -0.0584 0.1312	0.0391 -0.0307 -0.4316 0.0587 0.1786 -0.2003 0.3665 0.4293 0.0403
	A66 - A363	0.1649 -0.0259 0.0184 -0.0259 0.1422 0.0055 0.0184 0.0055 0.2028	0.8808 -0.0912 -0.1736 -0.0912 0.9522 0.0972 -0.1736 0.0972 1.6563	-0.0345 0.0102 -0.0661 0.1159 -0.0222 0.0999 0.0424 -0.1330 -0.0237

Table 2.3) Examples of parameters of the elemental motions found from the decomposition of the *TLS* matrices. The parameters are given in the units used in this article allowing an easy estimation of the corresponding atomic displacements. Direction of the libration and rotation axes is not given.

PDB code	chain, residue number	*T:* t_x, t_y, t_z (Å)	*L:* d_x, d_y, d_z (rad)	*S:* s_x, s_y, s_z (Å)	*trS*
1dqv	A1-A97	.3455 .3671 .4172	.01239 .02044 .02273	1.343 1.137 -1.319	0
	B1-B97	.3634 .3885 .4166	.01608 .01753 .03069	0.679 -1.177 0.200	0
1exr	A2-A30	.1944 .2663 .2870	.00000 .01602 .02182	0.000 2.951 3.408	>0
	A31-A74	.2110 .2939 .3068	.00000 .00860 .01637	0.000 -18.14 -5.028	<0
	A75-A84	.1692 .4906 .6598	.00000 .00000 .00000	0.000 0.000 0.000	0
	A85-A147	.0002 .2270 .3078	.00553 .01418 .02109	20.83 0.800 -1.672	≈0
4b3x	A1-A65	.0994 .6064 .7116	.00000 .00825 .01343	0.000 2.718 -11.05	<0
	A66-A363	.3306 .4102 .4413	.01568 .01720 .02283	3.164 -2.276 -0.197	0

2.8 Figures

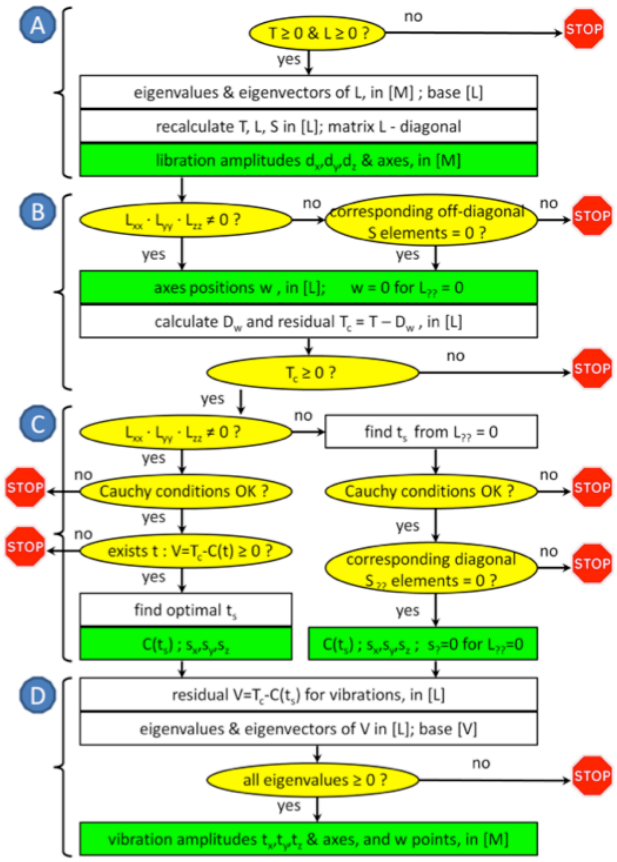


Figure 2.1) General flowchart of the *TLS* decomposition into libration and vibration composite motions. Yellow ellipses are for conditions to the verified. Green rectangles are for the output parameters of the composite motions. Letters A-D indicate different steps of the procedure as described in the text.

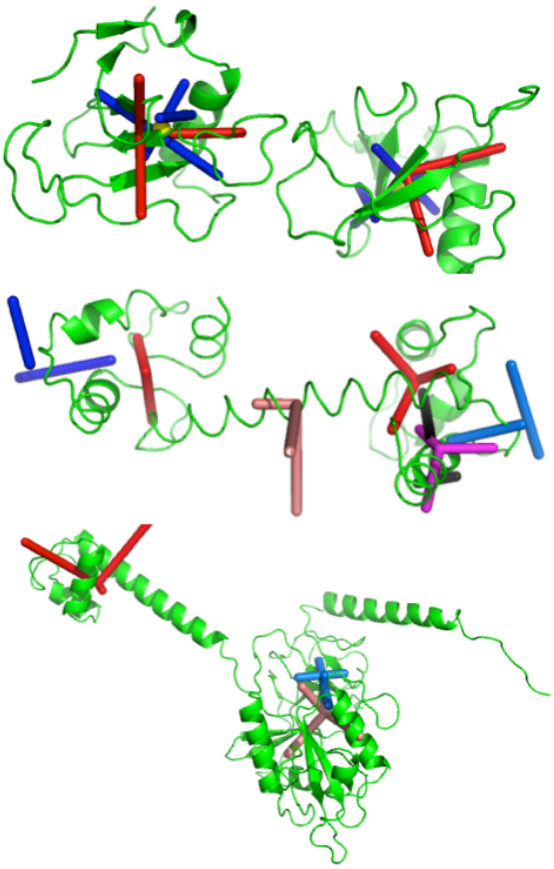


Figure 2.2) Examples of the vibration-libration ensembles. Red / salmon / magenta sticks indicate the principal vibration axes with the origin in the centre of the group; blue / marine / black sticks are for the libration axes. Yellow spheres for the 1dqv model show the reaction centers. a) 1dqv model. b) 1exr model; note pure vibrations for the group 3 (the helix) and absence of one of libration axes for the groups 1 and 2. c) 4b3x model. Libration axes for the first group are not shown being too far from the molecule.

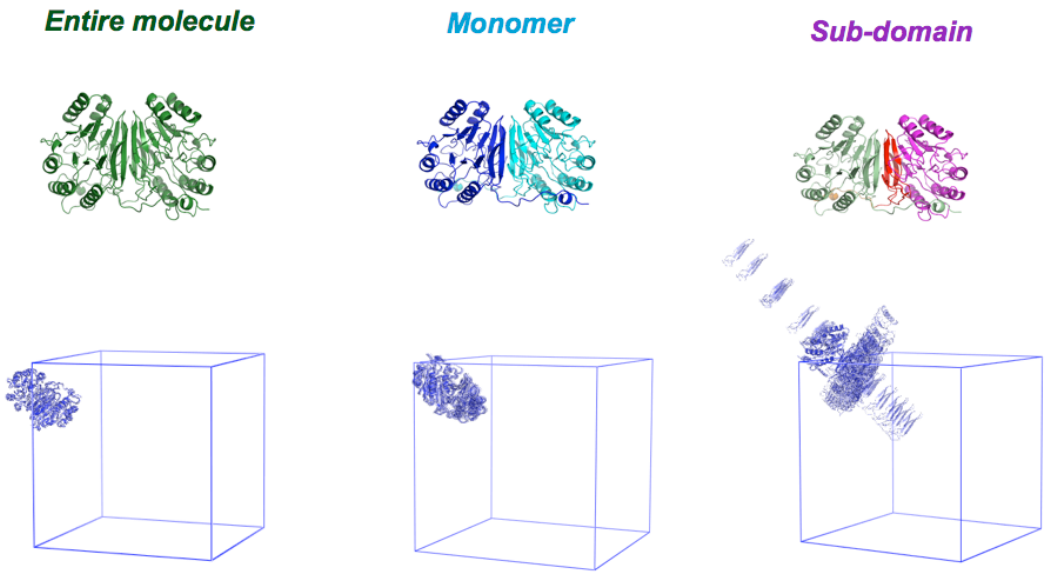


Figure 2.3) GpdQ TLS ensembles. The GpdQ *TLS* groups are projected onto the protein structure. The corresponding ensembles produced by *phenix.tls_as_xyz* are shown below. Each *TLS* PDB ensemble is shown as a single asymmetric unit outlined by the unit cell. An increase in overall motion is apparent going from left to right. The 20 member ensemble is shown for visual simplicity.

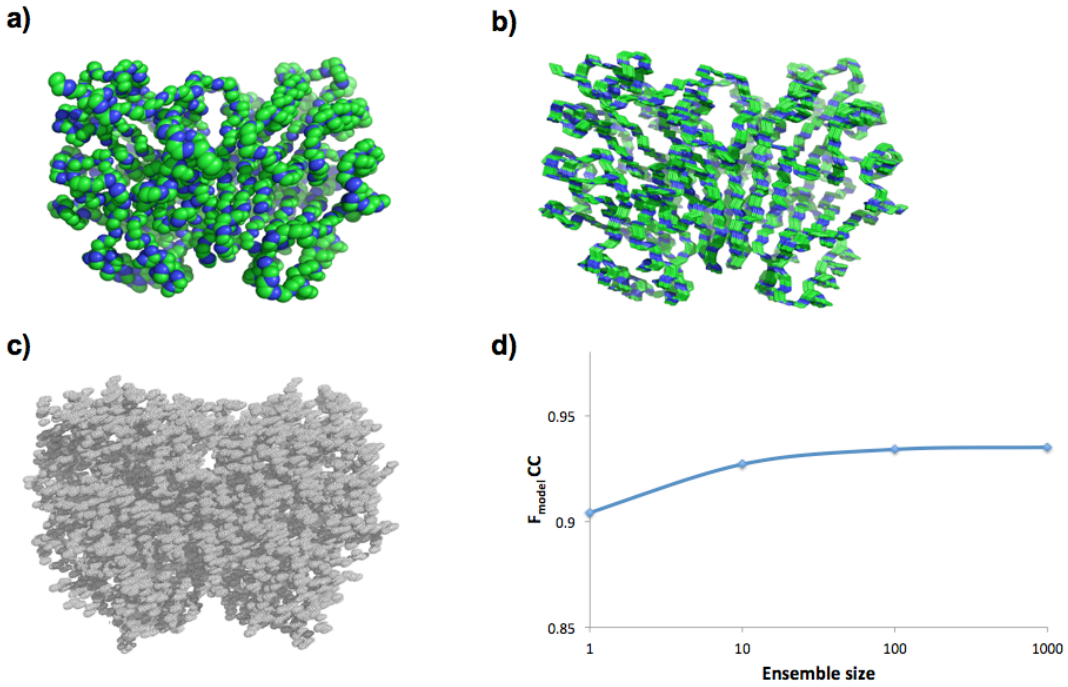


Figure 2.4) ***phenix.tls_as_xyz*** **ensembles replicate** ***TLS*** **anisotropic motion.** a) GpdQ backbone with thermal ellipsoid representation of "entire molecule" *TLS* anisotropic *B*-factors. b) *phenix.tls_as_xyz* ensemble backbones produced from "entire molecule" *TLS* refinement. c) Complete electron density predicted by "entire molecule" *TLS* refinement. d) Global correlation coefficient between experimental structure factor amplitudes F_{obs} of the original GpdQ 'entire motion' refinement and *phenix.tls_as_xyz* ensembles of various sizes. Convergence values plateau at 0.935.

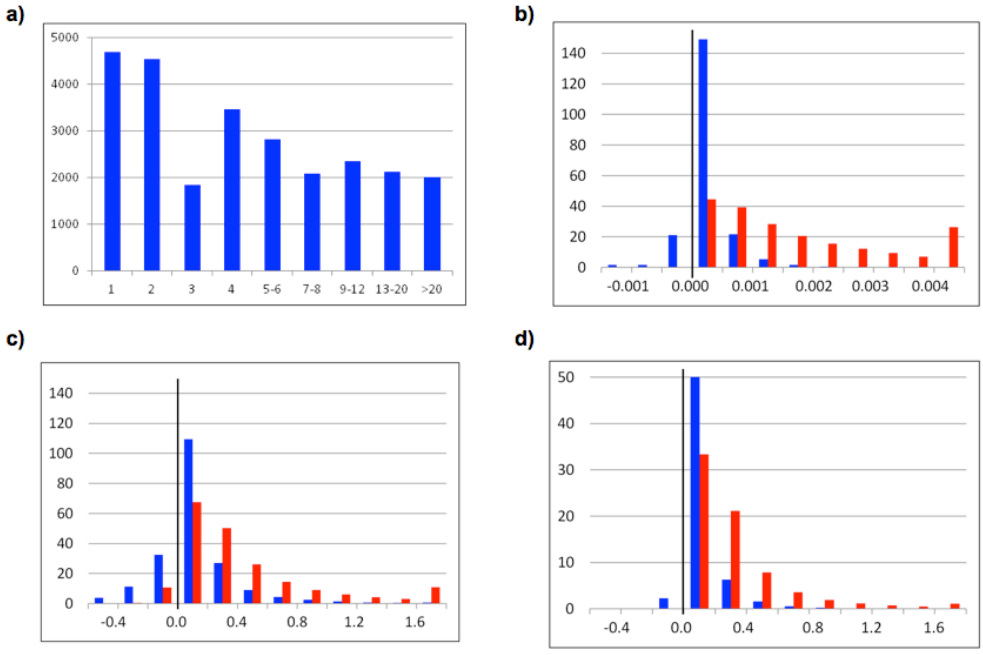


Figure 2.5) The number of PDB entries (in thousands) as a function of various parameters. The blue histogram in (b-d) is for the minimum eigenvalue and the red histogram is for the maximum eigenvalue. The leftmost and rightmost bins include all the cases with the values respectively less than or greater than the limits given at the axis. The eigenvalues are given in radians^2 for *L* and in Å^2 for *T* and *L*. The total number of the *TLS* groups is 203,261 for the diagrams (a-c) and about 70,000 for the diagram d) when the matrix *V* could be calculated. a) The number of the *TLS* groups per entry; the largest is 283. b) Distribution of eigenvalues of the matrix *L*; minimum eigenvalue varies from -0.285 to 0.164, maximum eigenvalue varies from -0.001 to 0.409. c) Distribution of eigenvalues of the matrix *T*; minimum eigenvalue varies from -20.716 to 6.852, maximum eigenvalue varies from -1.551 to 28.676. d) Distribution of eigenvalues of the matrix *V* (*S* matrix optimized as described in the article); minimum eigenvalue varies from -0.001 to 2.815, maximum eigenvalue varies from 0 to 5.950.

2.9 References

Adams, P.D., Afonine, P.V., Bunkóczi, G., Chen, V.B., Davis, I.W., Echols, N., Headd, J.J., Hung, L.-W., Kapral, G.J., Grosse-Kunstleve, R.W., McCoy, A.J., Moriarty, N.W., Oeffner, R., Read, R.J., Richardson, D.C., Richardson, J.S., Terwilliger, T.C. & Zwart, P.H. (2010). *Acta Cryst.,* D**66**, 213-221.

Afonine, P., Urzhumtsev, A. (2007). *CCP4 Newsletter on Protein Crystallography,* **45**, http://www.ccp4.ac.uk/newsletters/newsletter45/articles/

Afonine, P.V., Echols, N., Grosse-Kunstleve, R.W., Headd, J.J., Moriarty, N.W., Mustyakimov, M., Terwilliger, T., Urzhumtsev, A., Zwart, P.H., Adams, P.D. (2012). *Acta Cryst.,* D**68,** 352-367.

Afonine, P.V., Grosse-Kunstleve, R.W., Adams, P.D., Urzhumtsev, A. (2013). *Acta Cryst.,* D**69***,* 625-634

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N. & Bourne, P.E. (2000). *Nucleic Acids Research*. **28**, 235-242.

Bernstein, F.C., Koetzle, T.F., Williams, G.J., Meyer, E.F., Jr., Brice, M.D., Rodgers, J.R., Kennard, O. Shimanouchi, T. & Tasumi, M. (1977). *J.Mol.Biol*., **112**, 535-542.

Brenner, H. (1967). *J.Colloid Interface Chem.*, **23**, 407-435.

Chaudhry, C., Horwich, A.L., Brunger, A.T. & Adams, P.D. (2004). *J.Molec.Biol.*, **342**, 229-245.

Cruickshank, D.W.J. (1956). *Acta Cryst.* **9**, 754-756.

Dauter, Z., Murshudov & Wilson, K.S. (2011). *International Tables for Crysallography vol.F, eds. Arnold E., Himmel D.M., Rossmann M.G*., Wiley & Sons, Chichester, 485-498.

Dunitz, J.D. & White, D.N.J. (1973). *Acta Cryst.*, **A**29, 93-94.

Goldstein, H. (1950). *Classical Mechanics*, Cambridge, Massachusetts: Addison-Wesley.

Grosse-Kunstleve, R.W. & Adams, P.D. (2002). *J.Appl. Cryst.*, **35**, 477-480.

Grosse-Kunstleve, R. W., Sauter, N.K., Moriarty, N.W. & Adams, P. D. (2002). *J. Appl. Cryst.,* **35**, 126-136.

Harris, G.W., Pickersgill, R.W., Howlin, B. & Moss, D.S. (1992). *Acta Cryst.*, B**48**, 67-75.

Howlin, B., Moss, D.S. & Harris, G.W. (1989). *Acta Cryst.*, A**45**, 851-861.

Howlin, B., Butler, S.A., Moss, D.S., Harris, G.W. & Driessen, H.P.C. (1993). *J.Appl. Cryst.*, **26**,

622-626.

Jackson, C. J., Carr, P. D., Liu, J. W., Watt, S. J., Beck, J. L. & Ollis, D. L. (2007). *J. Mol. Biol.*, **367**, 1047-1062.

Johnson, C.K. (1970). In *Crystallographic Computing*, ed. F.R.Ahmed, Munksgaard, Copenhagen, 220-226.

Johnson, C.K. (1980). In *Computing in Crystallography*, eds. R.Diamond, S.Ramaseshan, K.Venkatesan, Indian Academy of Sciences, Bangalor, India, 14.01-14.19.

Kuriyan, J. & Weis, W.I. (1991). *Proc.Natl.Acad.Sci. USA*, **88**, 2773-2777.

McCammon, J.A., Gelin, B.R. & Karplus, M. (1977). *Nature*, **267**, 585-590.

Murshudov, G.N., Vagin, A.A., Lebedev, A., Wilson, K.S. & Dodson, E.J. (1999). *Acta Cryst.*, D**55**, 247-255.

Painter, J. & Merritt, E.A. (2005). *Acta Cryst.*, D**61**, 465-471.

Painter, J. & Merritt, E.A. (2006a). *Acta Cryst.*, D**62**, 439-450.

Painter, J. & Merritt, E.A. (2006b). *J.Appl. Cryst.*, **39**, 109-111.

Papiz, M.Z. & Prince, S.M. (2003). *J.Molec.Biol.*, **326**, 1523-1538.

Prince, E. & Finger, L.M. (1972). *Acta Cryst.,* B**29**, 179-183.

Raaijmakers, H., Toro, I., Birkenbihl, R., Kemper, B. & Suck, D. (2001). *J.Molec.Biol.*, **308**, 311-323.

Sali, A., Veerapandian, B., Cooper, J.B., Moss, D.S., Hofmann, T. & Blundell, T.L. (1992). *Proteins Struct. Funct. Genet.*, **12**, 158-170.

Scheringer, C. (1973). *Acta Cryst.* A**29**, 554-570.

Schomaker, V. & Trueblood, K.N. (1968). *Acta Cryst.* B**24**, 63-76.

Schomaker, V. & Trueblood, K.N. (1998). *Acta Cryst.* B**54**, 507-514.

Sheriff, S. & Hendrickson, W. A. (1987). *Acta Cryst*., A**43**, 118–121

Tickle, I. & Moss, D.S. (1999). Notes from IUCr Cryst.Computing School, *http://public-1.cryst.bbk.ac.uk/~tickle/iucr99/iucrcs99.htm.*

Urzhumtsev, A., Adams, P.D. & Afonine, P. (2013). *Crystallography reviews,* **19**, 230-270.

Van Benschoten, A.H., Afonine, P.V., Terwilliger, T.C., Wall, M.E., Jackson, C.J., Sauter, N.K., Adams, P.D., Urzhumtsev, A. & Fraser, J.S. (2015). *Acta Cryst.*, D**71**

Williams, B.B., Van Benschoten, A. H., Cimermancic, P., Donia, M. S., Zimmermann, M., Taketani, M., Ishihara, A., Kashyap, P. C., Fraser, J.S. & Fischbach, M. A. (2014) *Cell Host & Microbe,* **16**, 495-503

Wilson, M.A. & Brunger, A.T. (2000). *J.Molec.Biol.*, **301**, 1237-1256.

Winn, M.D., Isupov, M.N. & Murshudov, G.N. (2001). *Acta Cryst.*, D**57,** 122-133.

Winn, M.D., Murshudov, G.N. & Papiz, M.Z. (2003). *Methods in Enzymology.*, **374**, 300-321.

Yousef, M.S., Fabiola, F., Gattis, J.L., Somasundaram, T. & Chapman, M.S. (2002). *Acta Cryst.*, D**58**, 2009-2017.

Zucker, F., Champ, P.C. & Merritt, E.A. (2010). *Acta Cryst.*, D**66,** 889-900.

2.10 Appendix A. Technical details of the algorithm

Definition of the transition matrices

Let us have three mutually orthogonal unit vectors $\boldsymbol{l}_x$, $\boldsymbol{l}_y$, $\boldsymbol{l}_z$ described respectively by their coordinates $\left((l_x)_{[M]x},(l_x)_{[M]y},(l_x)_{[M]z}\right)$, $\left((l_y)_{[M]x},(l_y)_{[M]y},(l_y)_{[M]z}\right)$, $\left((l_z)_{[M]x},(l_z)_{[M]y},(l_z)_{[M]z}\right)$ in the Cartesian basis [M]. These vectors can be considered as a new basis [L]. The coordinates of a vector **r** in [L] and [M] are expressed through each other using the transition matrix R_{ML} as

$$\begin{pmatrix} x_{[M]} \\ y_{[M]} \\ z_{[M]} \end{pmatrix} = R_{ML} \begin{pmatrix} x_{[L]} \\ y_{[L]} \\ z_{[L]} \end{pmatrix} = \begin{pmatrix} (l_x)_{[M]x} & (l_y)_{[M]x} & (l_z)_{[M]x} \\ (l_x)_{[M]y} & (l_y)_{[M]y} & (l_z)_{[M]y} \\ (l_x)_{[M]z} & (l_y)_{[M]z} & (l_z)_{[M]z} \end{pmatrix} \begin{pmatrix} x_{[L]} \\ y_{[L]} \\ z_{[L]} \end{pmatrix}$$

(50)

Transition matrices for other pairs of basis, for example from [V] to [L] (section 2.1), [M] to [V] and *vice versa* (Section 7.3) are defined in a similar way.

Cauchy conditions on the elements of the TLS matrices

Let d_x, d_y, d_z and u_x, u_y, u_z be random displacements due to rotations and translations, respectively. Since $S_{xx} = \langle d_x u_x \rangle$, $V_{xx} = \langle u_x u_x \rangle$, $L_{xx} = \langle d_x d_x \rangle$ (Schomaker & Trueblood (1968); see also equations (8.5) - (8.7) in UAA-2013) , it follows from the Cauchy inequality that

$$S_{xx}^2 \le T_{xx} L_{xx}$$

(51)

In the basis [L] with $S_{[L]} = S_C(t_S)$ (eq. (18)), condition (51) becomes

$$\left(S_{[L]xx} - t\right)^2 \le T_{C[L]xx} L_{[L]xx}$$

(52)

Similarly we obtain two other conditions

$$\left(S_{[L]yy} - t\right)^2 \le T_{C[L]yy} L_{[L]yy}, \qquad \left(S_{[L]zz} - t\right)^2 \le T_{C[L]zz} L_{[L]zz}$$

(53)

Polynomials for the coefficients of the characteristic equation

If t_{xx}, t_{xy}, t_{xz} etc are respective elements of the matrix T_Λ (29), the coefficients (36) of the characteristic equation as functions of the parameter *t* are:

$$a_S(t) = \left[\left(t - S_{[L]xx}\right)^2 - t_{xx}\right] + \left[\left(t - S_{[L]yy}\right)^2 - t_{yy}\right] + \left[\left(t - S_{[L]zz}\right)^2 - t_{zz}\right]$$

(54)

$$b_S(t) = \left[\left(t - S_{[L]xx}\right)^2 - t_{xx}\right]\left[\left(t - S_{[L]yy}\right)^2 - t_{yy}\right] + \left[\left(t - S_{[L]yy}\right)^2 - t_{yy}\right]\left[\left(t - S_{[L]zz}\right)^2 - t_{zz}\right] +$$
$$+ \left[\left(t - S_{[L]zz}\right)^2 - t_{zz}\right]\left[\left(t - S_{[L]xx}\right)^2 - t_{xx}\right] - \left[t_{xy}^2 + t_{yz}^2 + t_{zx}^2\right] \qquad (55)$$

$$c_S(t) = \left[\left(t - S_{[L]xx}\right)^2 - t_{xx}\right]\left[\left(t - S_{[L]yy}\right)^2 - t_{yy}\right]\left[\left(t - S_{[L]zz}\right)^2 - t_{zz}\right] -$$

(56)

$$- t_{yz}^2\left[\left(t - S_{[L]xx}\right)^2 - t_{xx}\right] - t_{xz}^2\left[\left(t - S_{[L]yy}\right)^2 - t_{yy}\right] - t_{xy}^2\left[\left(t - S_{[L]zz}\right)^2 - t_{zz}\right] - 2t_{xy}t_{yz}t_{xz}$$

Explicit expression for the atomic shifts due to rotations with given parameters

Let $\left(x_{[L]}, y_{[L]}, z_{[L]}\right)$ be Cartesian coordinates of a point $\mathbf{r}$ in the basis [L]. For a rotation around the axis parallel to $\boldsymbol{l}_z$ and crossing the point $\mathbf{w}_{[L]}^{lz} = \left(w_{[L]x}^{lz}, w_{[L]y}^{lz}, w_{[L]z}^{lz}\right)$, we recalculate first the coordinates of the vector $\mathbf{r} - \mathbf{w}_{[L]}^{lz}$ with respect to the rotation axis

$$x_{[A]} = x_{[L]} - w_x^{lz}; y_{[A]} = y_{[L]} - w_y^{lz}; z_{[A]} = z_{[L]},$$

(57)

If $\mathbf{r}'$ stands for the position of the same point after rotation by angle d_{z0} around this axis, the coordinates of $\mathbf{r}' - \mathbf{w}_{[L]}^{lz}$, the point with respect to the axis, are

$$\left(x_{[A]}\cos d_{z0} - y_{[A]}\sin d_{z0}\right), \left(x_{[A]}\sin d_{z0} + y_{[A]}\cos d_{z0}\right), \left(z_{[A]} + s_z d_{z0}\right).$$

(58)

This gives the atomic shift

$$\Delta_{[L]}^{lz}\mathbf{r} = \mathbf{r}' - \mathbf{r} = \left(\mathbf{r}' - \mathbf{w}_{[L]}^{lz}\right) - \left(\mathbf{r} - \mathbf{w}_{[L]}^{lz}\right) =$$

(59)

$$= \left[\left(x_{[L]} - w_x^{lz}\right)\left(\cos d_{z0} - 1\right) - \left(y_{[L]} - w_y^{lz}\right)\sin d_{z0}\right]\boldsymbol{l}_x +$$

$$+\left[\left(x_{[L]}-w_x^{lz}\right)\sin d_{z0}+\left(y_{[L]}-w_y^{lz}\right)\left(\cos d_{z0}-1\right)\right]\boldsymbol{l}_y+s_z d_{z0}\boldsymbol{l}_z$$

There are similar expressions for the shift due to rotations around two other axes:

$$\Delta^{lx}_{[L]}\mathbf{r}=\left[\left(y_{[L]}-w_y^{lx}\right)\left(\cos d_{x0}-1\right)-\left(z_{[L]}-w_z^{lx}\right)\sin d_{x0}\right]\boldsymbol{l}_y+$$

(60)

$$+\left[\left(y_{[L]}-w_y^{lx}\right)\sin d_{x0}+\left(z_{[L]}-w_z^{lx}\right)\left(\cos d_{x0}-1\right)\right]\boldsymbol{l}_z+s_x d_{x0}\boldsymbol{l}_x$$

$$\Delta^{ly}_{[L]}\mathbf{r}=\left[\left(z_{[L]}-w_z^{ly}\right)\left(\cos d_{y0}-1\right)-\left(x_{[L]}-w_x^{ly}\right)\sin d_{y0}\right]\boldsymbol{l}_z+$$

(61)

$$+\left[\left(z_{[L]}-w_z^{ly}\right)\sin d_{y0}+\left(x_{[L]}-w_x^{ly}\right)\left(\cos d_{y0}-1\right)\right]\boldsymbol{l}_x+s_y d_{y0}\boldsymbol{l}_y$$

2.11 Appendix B. List of abnormal situations requiring procedural interruption

This appendix summarizes the situations when the described algorithm breaks. Each condition below starts from the corresponding program message and then refers to the main text and to Figure 2.1. To analyze the PDB content, the program can be run in a special regime when at step C we assign $t_S = 0$ (this corresponds to the current default constraint *trS* = 0), *i.e.* when the matrix *S* is taken without any correction. In this regime we calculate directly the matrices *C* and $V_{[L]}$ and check the conditions (k-m).

Step A: basis [L]; determination of the libration axes and amplitudes

a) "Input matrix L[M] is not positive semidefinite". Section 3.1.

b) "Input matrix T[M] is not positive semidefinite". Section 3.1.

*Step B: determination of the points **w** at the libration axes*

c) "Non-zero off-diagonal S[L] and zero L[L] elements". Section 3.2, eq. (15).

d) "Matrix T_C[L] is not positive semidefinite". Section 3.2, eq. (17).

Step C: determination of the screw parameters

left branch (librations around all three axes)

e) “Empty (tmin_c , tmax_c) interval”. Section 4.2, eq. (23). $t_{\min,C} > t_{\max,C}$.

f) “Empty (tmin_t , tmax_t) interval”. Section 4.3, eq. (31). $t_{\min,\tau} > t_{\max,\tau}$.

g) “Negative argument when estimating tmin_a”. Section 4.3, eq. (38).

h) “Intersection of the intervals for t_S is empty”. Section 4.3, step b. $t_{\min} > t_{\max}$

i) “t_min = t_max giving non positive semidefinite V_lambda”. Section 4.3, step c.

j) “Interval (t_min , t_max) has no t value giving positive semidefinite V”. Section 4.3, step d.

right branch (no libration around at least one of the axes)

k) “Cauchy-Schwarz conditions are wrong for the found t_S”. Eq. (22) with t_S calculated in section 4.4.

l) “Non-zero diagonal S[L] element for a zero L[L] element”. Section 4.4.

Step D: determination of the vibration parameters

m) “Matrix V[L] is not positive semidefinite”. Section 5.1.

Extra checks at step C when some conditions may fail due to rounding errors:

- When calculating square roots (24), the arguments are non negative by previous conditions (a) and (d) since the diagonal elements of a positive semidefinite matrix are non negative.
- When calculating square roots (28), the arguments are non negative by previous condition (a).
- When calculating square roots (31), the argument $\tau_{\max}$ is non negative since the eigenvalues of $T_{C[L]}$ are also non negative by previous condition (d).

Chapter 3

Predicting X-ray Diffuse Scattering from Translation Libration Screw Structural Ensembles

3.1 Abstract

Identifying the intramolecular motions of proteins and nucleic acids is a major challenge in macromolecular X-ray crystallography. Because Bragg diffraction describes the average positional distribution of crystalline atoms with imperfect precision, the resulting electron density can be compatible with multiple models of motion. Diffuse X-ray scattering can reduce this degeneracy by directly reporting on correlated atomic displacements. Although recent technological advances are increasing the potential to accurately measure diffuse scattering, computational modeling and validation tools are still needed to quantify the agreement between experimental data and different parameterizations of crystalline disorder. A new tool, *phenix.diffuse*, addresses this need by employing Guinier's equation to calculate diffuse scattering from Protein Data Bank (PDB)-formatted structural ensembles. As an example case, *phenix.diffuse* is applied to Translation-Libration-Screw (TLS) refinement, which models rigid body displacement for segments of the macromolecule. To enable calculation of diffuse scattering from TLS refined structures, *phenix.tls_models* builds multi-model PDB files that sample the underlying T, L and S tensors. In the glycerophosphodiesterase GpdQ, alternative TLS group partitioning and different motional correlations between groups yield markedly dissimilar diffuse scattering maps with distinct implications for molecular mechanism and allostery. These methods demonstrate how, in principle, X-ray diffuse scattering could extend macromolecular structural refinement, validation, and analysis.

3.2 Introduction

Protein flexibility is essential for enzymatic turnover, signaling regulation and protein-protein interactions (Fraser & Jackson, 2011). The motions enabling these functions span length-scales from a few angstroms to many nanometers and include transitions between side chain rotamers (Fraser *et al.*, 2009), loop openings and closings (Qin *et al.*, 1998; Williams *et al.*, 2014) and rigid-body subunit rotations (Korostelev & Noller, 2007). Multiple crystal structures are routinely compared to identify these motions and to derive hypotheses about the role of correlated motions in executing protein function. However, if only a single crystal form is available, evidence of concerted motion must be extracted from the spread in the electron density.

Extracting this information is possible because protein conformational heterogeneity across unit cells in space and within unit cells during the X-ray exposure time leads to an ensemble-averaged electron density map. Atomic vibrations are commonly fit with individual B-factors, which describe the electron density distribution as a continuous isotropic Gaussian envelope around a central location and predominantly encompass disorder from thermal motion. Discrete conformational heterogeneity and crystal packing defects can be described as ensembles of structural models with partial occupancy (Burnley *et al.*, 2013; Rader & Agard, 1997; Gros *et al.*, 1990; van den Bedem *et al.*, 2009; Levin *et al.*, 2007; Wall *et al.*, 1997). If high-resolution diffraction data are available, anisotropic directionality can be added to B-factors by modeling a Gaussian distribution along each real-space axis, yielding an ellipsoid that shows the predominant direction of the electron density.

However, the large number of parameters required for anisotropic B-factor refinement renders it inaccessible for most macromolecular diffraction experiments. Translation-Libration-Screw (TLS) modeling, introduced by Schomaker and Trueblood (1968), can describe concerted, rigid-body displacement for groups of atoms (for comprehensive review see Urzhumtsev *et al.*, 2013). In TLS refinement, the target protein is segmented into independent rigid bodies that undergo small translations ("vibrations") and rotations ("librations"). The anisotropic displacement of TLS refinement can be fully described with twenty parameters per rigid body, each of which can potentially contain many atoms. This small number of parameters compares favorably to the six parameters per atom demanded by individual anisotropic B-factor refinement and allows grouped anisotropic B-factors to be modeled at mid to low-resolution ranges. TLS refinement often leads to better agreement between observed and calculated structure factors, as measured by decreasing R_{free} values. The potential for improved statistics when relatively few observations are available has positioned TLS as a general refinement technique: 22% of the structures deposited in the Protein Data Bank (PDB; Bernstein *et al.*, 1977; Berman *et al*, 2000) employ TLS refinement in some form. TLS refinement is a component of many major structural refinement programs such as *Refmac* (Murshudov *et al*, 1997; Winn *et al.*, 2001), BUSTER-TNT (Bricogne *et al*, 1993, 2011) and *phenix.refine* (Afonine *et al*, 2012). These programs can select TLS groups automatically, based on biochemical intuition, or with the assistance of external web servers (Painter & Merritt, 2006a; Painter & Merritt, 2006b).

TLS refinement naturally suggests concerted structural motions, which can be assigned biological significance and subsequently tested with additional experiments. Visualization programs such as TLSViewer (Painter & Merritt, 2005) can convert the T, L and S tensors into a

description of domain-scale mechanical motions and molecular graphics programs, such as Chimera (Pettersen *et.al*, 2004), Coot (Emsley and Cowtan, 2004) or PyMol (DeLano, 2002), can be used to visualize the resulting anisotropic ellipsoids. For example, TLS refinement of the large multi-protein complex GroEL revealed subunit tilting that may play a role in transmitting conformational changes upon GroES or nucleotide binding (Chaudhry *et al.*, 2004) **(Figure 3.1a-b)**. Similarly, TLS modeling of the ribosome structure implied a "ratcheting" rotation of the 50S and 30S subunits around the peptidyl transferase center during tRNA translocation (Korostelev & Noller, 2007).

A potential complication of TLS refinement is that there is no information regarding correlations between groups; thus, many different rigid body arrangements can result in equivalent improvement of refinement statistics (Moore, 2009; Tickle & Moss, 1999). The inability to discriminate among alternate TLS models stems from the exclusive usage of Bragg diffraction data in model refinement. Because Bragg data reports on electron density averaged across all unit cells, there may be several models of correlated structural displacement that fit the density equally well. Thus, TLS refinement might improve the modeled electron density but incorrectly describe the correlated motion that occurs in the crystal during the diffraction experiment. Drawing on additional sources of information such as patterns of steric clashes (van den Bedem *et al.*, 2013), NMR spectroscopy (Ruschak & Kay, 2012), or mutational analysis (Fraser *et al.*, 2009) can be used to distinguish competing models of correlated motion between non-bonded atoms.

An additional, yet rarely used, data source that can discriminate between these models is X-ray diffuse scattering from protein crystals, which results from correlated variation in the electron density distributions (Phillips *et al.*, 1980; Chacko & Phillips, 1992; Faure *et al.*, 1994; Clarage & Phillips, 1997; Mizuguchi *et al.*, 1994). This variation breaks from the hypothetical "perfect" crystal lattice, leading to diffraction outside of the regions of reciprocal space predicted by Bragg's law. The theoretical relationship between conformational heterogeneity within unit cells and diffuse scattering has been available for decades (Guinier, 1963; Amorós & Amorós, 1968) and small-molecule crystallographers have used diffuse scattering data in refinement and model validation (Estermann & Steurer, 1998; Michels-Clark *et al.*, 2013, Welberry & Butler, 1994).

The potential of macromolecular diffuse scattering to break the degeneracy within refinement methods such as TLS, including information about the location and length scale of macromolecular disorder, has long been recognized (Thune & Badger, 1995; Perez *et al.*, 1996; Hery *et al.*, 1998; Tickle & Moss, 1999). Diffuse scattering maps predicted by models of motion can be calculated using either an all-atom covariance matrix or the equation $I(q)_{diffuse} = N.(\langle|f_n(q)|^2\rangle - |\langle f_n(q)\rangle|^2)$ (often called Guinier's equation, where q is the scattering vector, n is the complex structure factor of the n-th protein conformation and N is the number of unit cells in the crystal) (Micu & Smith, 1994; Lindner & Smith, 2012). The covariance matrix describes correlated displacements between every pair of atoms, whereas Guinier's equation models diffuse scattering from an ensemble of structure factors. Calculation of the covariance matrix has been used to model crystalline normal modes and TLS parameterization (Riccardi *et al.*, 2010). It is also possible to explicitly estimate each matrix element from molecular dynamics trajectories (Meinhold & Smith, 2007). The size of the covariance matrix scales as the square of the number

of atoms, making full matrix calculations expensive to compute for large systems. This poses a significant challenge to quantitative diffuse scattering analysis. For these reasons, a straightforward method that calculates diffuse scatter from discrete multi-model PDB files may be preferable.

To meet this need, we developed *phenix.diffuse*, a new tool within the *Phenix* software suite (Adams *et al.*, 2010), which uses Guinier's equation to calculate diffuse scattering from multi-model (ensemble) PDB files. Thus, *phenix.diffuse* can be applied to any motional model represented as an explicit ensemble of related structures. As a first application we have simulated the diffuse scattering produced by alternative TLS refinements of the glycerophosphodiesterase GpdQ (Jackson *et al.*, 2007). GpdQ is found in *Enterobacter aerogenes* and contributes to the homeostasis of the cell membrane by hydrolyzing the 3'-5' phosphodiester bond in glycerophosphodiesters. Each chain of the dimeric enzyme contains three distinct structural elements: an α/β sandwich fold containing the active site, a domain-swapped active site cap and a novel dimerization domain comprised of dual-stranded antiparallel β-sheets connected by a small β-sheet. Though the catalytic mechanism of GpdQ is similar to other metallo-phosphoesterases, some substrates are too large to pass through the active site entrance as it is modeled in the crystal structure. Protein dynamics must therefore play a role in substrate entry and product release. Normal mode analysis of the GpdQ hexamer suggested high mobility in the cap domain and a breathing motion centered on the catalytic and dimerization domains (Jackson *et al.*, 2007). Due to the high global B-factors and presence of diffuse signal in the diffraction images **(Figure 3.1c)**, Jackson and colleagues performed three separate TLS refinements to model the crystalline disorder. All three TLS refinements improved the R_{free} values when

compared to the standard isotropic B-factor refinement; however, there was no significant difference among the final R_{free} values from the refinements initiated with distinct TLS groupings. In contrast, our results reveal significant differences between the diffuse intensities predicted by the motion from each TLS refinement, highlighting the possible usefulness of diffuse scattering in optimizing structure refinement.

3.3 Methods

GpdQ refinement. Based on the original refinement strategy of Jackson *et al.* (2007), we performed three different TLS refinements on the zinc-bound structure of GpdQ (PDB ID: 2DXN): *"Entire molecule",* one TLS group for all residues; *"Monomer"*, one TLS group for each of the two individual chains; and *"Sub-domain"*, one TLS group for each of the α/β sandwich domain (residues 1-196), the "dimerization" domain (residues 197-255) and the "cap" domain (residues 257-271) of each chain. The pre-TLS refinement R_{work} and R_{free} were 19.1% and 23.1%, respectively. After defining the TLS groups, each structure was re-refined for 5 macrocycles in *Phenix.refine*. The strategy included refinement of the individual coordinates and isotropic B-factors, water picking and refinement of TLS parameters for defined TLS groups. Both the X-ray/atomic displacement parameters and X-ray/stereochemistry weights were optimized (Afonine *et al*, 2011). The final R_{work}, R_{free} values for each refinement were as follows: *"Entire molecule"* (14.6%, 18.9%), *"Monomer"* (14.9%, 19.0%), *"Sub-domain"* (14.9%, 19.3%), suggesting approximately equal agreement with the Bragg data **(Figure 3.1d)**.

In TLS refinement, the eigenvalues of the T and L matrices describe the variance of the motional displacement along each orthogonal real-space axis. To avoid an unphysical description of TLS

motion *(Urzhumtsev et al. accompanying manuscript)*, we inspected the eigenvalues of each TLS refinement to ensure non-negative eigenvalues for the T and L matrices **(Table 3.1)**. Although solvent is expected to contribute significantly to experimental diffuse scattering, we removed water molecules after refinement. This step, along with removing bulk solvent from the starting structure, ensures that all subsequent diffuse scattering simulations only reflect correlated motions implicit in the TLS refinement.

***phenix.tls_models* and TLS ensemble generation.** We used *phenix.tls_models* (*Urzhumtsev et al., accompanying manuscript*) to convert the TLS matrices to a structural ensemble. *phenix.tls_models* receives as input a structure with TLS header information, separates the molecule into individual TLS groups and randomly samples the real-space distribution for each group based on mathematical decomposition of the T, L and S matrices. The trace of the matrix S is set to 0 during these calculations. The sampled PDB files are then either re-assembled into a multi-model PDB ensemble or output with no further changes **(Figure 3.2)**. To ensure adequate sampling of the underlying Gaussian distributions, we generated ensembles of different sizes and monitored the convergence of the global correlation coefficient between diffuse maps in which spherically-symmetric sources of diffuse scattering have been removed ("anisotropic maps": **Table 3.2**). These maps offer an improved comparison relative to the raw diffuse signal because they correct for the resolution dependency of diffuse scattering, which would otherwise lead to overestimation of inter-map correlation. We determined that an ensemble size of 1000 models was sufficient for effective sampling of each TLS refinement. The extent of the motions predicted by the "sub-domain" refinement **(Figure 3.3)** is quite surprising and likely results from a lack of chemical restraints within the TLS refinement implementation in Phenix. While

subdividing the "monomer" TLS refinement into smaller components might intuitively produce similar refinement statistics, the tensors between all three groups are substantially different and thus describe dissimilar motions.

phenix.diffuse. *Phenix.diffuse* implements Guinier's description of diffuse scattering (Guinier, 1963) **(Figure 3.4a)** . Diffuse scattering is calculated entirely from a series of unit cell "snapshots" contained in a multi-model PDB ensemble and assumes no motional correlation between crystal unit cells. This simplification ignores sources of disorder spanning multiple unit cells, which can contribute to experimentally measured diffuse scattering (Clarage *et al.*, 1992; Doucet & Benoit, 1987; Wall *et al.*, 1997). *phenix.diffuse* can model these large-scale effects through the analysis of a "supercell" containing multiple unit cell copies as implemented in several recent MD simulations of small proteins (Janowski *et al.*, 2013; Kuzmanic *et al.*, 2014). Guinier's equation can be applied to arbitrarily sized crystalline regions; thus, a system of multiple unit cells allows for analysis of motions that occur between and across unit cells. In line with previous diffuse scattering simulations (Wall *et al.*, 2014), our program calculates structure factors for each ensemble member at the Bragg lattice positions, from which each term in Guinier's equation is determined.

GpdQ TLS diffuse scattering simulation. We simulated the diffuse scattering of each of the GpdQ TLS ensembles to 3.0Å resolution. Unless otherwise stated, all TLS groups within a given refinement were assumed to move independently of one another. Since the diffraction data for GpdQ PDB entry 2DXN extends to 2.9Å, our simulation should be sufficient for future comparisons with experimental maps. As the resulting diffuse scattering data is identical in

format to descriptions of Bragg X-ray reflections, *phenix.reflection_statistics* was used to perform all statistical analyses. All reported correlation values are global Pearson correlation coefficients calculated between the described two sets of diffuse intensities. As previously mentioned (and described in Wall *et al.* (1997)), spherically symmetric sources of diffuse scattering contribute significantly to the observed intensity. In order to remove these confounding effects, we used the LUNUS software package (Wall, 2009) to subtract from each point the average radial diffuse intensity (**Figure 3.5).**

GpdQ diffraction image processing and radial averaging. Diffraction images used to determine the GpdQ Bragg structure were collected at the Advanced Photon Source (Lemont, IL) under cryogenic temperatures with 0.25^{o} oscillation wedges (Jackson *et al.*, 2006). Subsequent processing was performed using LUNUS (Wall, 2009). Pixels correlating to the beamstop shadow and CCD detector panels were removed with the LUNUS *punchim* and *thrshim* routines. Solid-angle normalization and beam polarization were corrected using *polarim* and *normim*. Mode filtering was applied as previously described (Wall *et al.*, 1997). The radial intensity profile was calculated from a single image using the *avgrim* function, which calculates radial intensities on a per-pixel scale. The radial profile for the experimental GpdQ data was scaled by a factor of 1000x to better facilitate qualitative comparisons to the simulations.

3.4 Diffuse scattering is dependent on TLS grouping.

The raw diffuse intensity predicted by the motions described from each TLS refinement strategy rises as a function of the number of TLS groups **(Figure 3.6)**. The "entire molecule" and "monomer" maps show a similar range of intensity values: $0\text{-}4.52\text{x}10^{6}$ and $0\text{-}8.34\text{x}10^{6}$

respectively. The "subdomain" map displays a much wider dynamic range (0-4.71x10^8) **(Figure 3.3c)**. This trend likely results from an increase in the amplitude of TLS motion, particularly within the dimerization region of the "subdomain" model. **(Figure 3.3)**. However, "sub-domain" map intensities greater than 1x10^7 are limited to a resolution range of 11Å and lower. The "entire molecule" and "monomer" maps also possess "primary diffuse shell" regions surrounding the origin, though they only extend out to a resolution range of 30Å. This region will be particularly difficult to measure experimentally given the presence of a beamstop, which blocks access to signal around F_{000} (Lang *et al.*, 2014). Each diffuse map has a dip in radial intensity between the primary diffuse shell before the diffuse intensity rises in a second shell **(Figure 3.7a).** In contrast to the "sub-domain" map, the strongest diffuse intensities for the "entire molecule" and "monomer" maps occur within this secondary shell. The width between the primary and secondary diffuse shells decreases as the number of TLS groups increases, due to an expansion in the primary diffuse shell radius. As X-ray detectors can easily measure intensities in the regions of reciprocal space occupied by the secondary shell, a significant fraction of the diffuse scattering predicted by TLS refinement can potentially be compared to experimental data.

To determine if the different TLS groupings yielded distinct diffuse scattering predictions, we calculated global Pearson correlation coefficients between each refinement's anisotropic signal. The anisotropic comparison revealed little similarity between maps (CC range from 0.031 to 0.312) **(Figure 3.6)**. Comparing the correlation values across resolution bins reveals the anisotropic diffuse signal correlations remain consistently poor across scattering vector length **(Figure 3.7b)**. The large discrepancy between the maps calculated with different TLS models contrasts with the high similarity of experimental maps of anisotropic diffuse signal from

different crystals of *Staphylococcal* nuclease (CC = 0.93) (Wall *et al.*, 1997). This result suggests the experimentally measured diffuse signal will be sufficiently precise to distinguish between TLS-related diffuse scattering models (Wall *et al.*, 2014). However, other sources of disorder will need to be accounted for before models of TLS motion can be effectively compared to experimental data.

3.5 Correlations between TLS groups can be detected by diffuse scattering.

Although TLS refinement makes no assumptions regarding motion between groups, diffuse scattering can test whether correlated rigid body fluctuations do, in fact, exist. To illustrate this concept, we simultaneously sampled the motions along the translation and libration eigenvectors to produce "parallel" and "antiparallel" correlated motions for the "monomer" GpdQ TLS refinement **(Figure 3.8).** For the "parallel" model, the correlated motion consists of sampling along all translation and libration eigenvectors in step sizes of $\sigma/2$, where σ is obtained from the underlying Gaussian distribution in each direction, for a total of 10 steps (-2.5σ to 2.5σ). Simply reversing the direction of sampling for the chain B translation eigenvectors created the "antiparallel" motion. In contrast to the simulation in **Figure 3.6a**, which assumed no correlation between TLS groups, here we have introduced correlated motion between GpdQ monomers. Next, we simulated the diffuse scattering produced by the "parallel" and "antiparallel" correlated motions. Both raw maps display strong secondary shell characteristics in combination with a weak primary shell of diffuse scattering **(Figure 3.8c).** A diffuse intensity difference map **(Figure 3.8d)** shows that discrepancies between the raw maps occur across the entirety of reciprocal space. Comparing the anisotropic diffuse intensity correlation across resolution bins reveals a general decreasing trend as the scattering vector length increases

(Figure 3.8e). In contrast to the previous TLS simulations, the correlation values are highest at low resolution. The low global Pearson correlation coefficient (0.375) demonstrates there are quantitative differences between the two maps. However, these inter-group correlation differences will be slightly more difficult to detect than changes between specific TLS models, where the correlation coefficients range from 0.031 to 0.312.

3.6 TLS models yield unique radial profiles of diffuse intensity.

We calculated the radial diffuse intensity profile for a GpdQ diffraction frame and for the three TLS refinements **(Figure 3.9)**. Although radial averaging removes the rich directional information present in diffuse scattering, this simplification has been successfully used to assess agreement between distinct diffuse maps (Meinhold & Smith, 2005; Meinhold & Smith, 2007). For the experimental GpdQ map, a peak at 8.5Å and a shoulder at 6Å are observed. None of these features are observed in the raw TLS radial profiles, except for a local maximum at 4.5Å and shoulder at 4Å for the "monomer" refinement. Rather, the dominating feature for each TLS simulation is the secondary diffuse scattering shell, which varies between maps in both width and maximum radial value. This result is not surprising, as the experimental diffuse scattering from GpdQ reflects a much broader group of correlated motions than simply TLS-related movement within the macromolecule. For example, disordered solvent is expected to significantly contribute to experimental diffuse measurements (Wall *et al.*, 1997). As solvent molecules were not modeled in our TLS ensembles, this is a likely source of the discrepancy between GpdQ experiment and simulation. The liquid-like motions (LLM) model, in which atoms interact only with nearest neighbors to produce a gelatinous crystalline environment, can also be used to explain the diffuse scattering intensity. Comparing the diffuse maps of

Staphylococcal nuclease (Wall *et al.*, 1997), pig insulin (Caspar *et al.*, 1988) and hen egg-white lysozyme (Clarage *et al.*, 1992) to LLM models maximized correlations across distances of 6-10Å. Thus, a more thorough analysis involving several models of disorder must be applied to GpdQ to improve the fit to the experimental diffuse data.

3.7 Distinct patterns of diffuse signal can be calculated at Non-Bragg Indices

While *phenix.diffuse* currently calculates the diffuse signal under Bragg peaks, diffuse scattering occurs throughout the entirety of reciprocal space. To more completely sample reciprocal space between the Bragg spots, we increased the unit cell boundaries. Expanding the unit cell in real space allows for finer sampling of the underlying Fourier transform **(Figure 3.10)**. The resulting structure factors can be re-scaled to the original lattice points, leading to fractional *hkl* sampling. These fractional values are then assigned to the nearest integer *hkl* index and averaged, leading to a single diffuse intensity value associated with each Bragg peak. Although it is clearly possible to output a map consisting of these fractional values and thereby produce a more accurate picture of diffuse scattering, we chose the integer values because diffuse scattering processing techniques commonly calculate the average diffuse intensity across pixels within a 1x1x1 voxel around each Bragg point (Wall, 1996). This average value is then assigned to the *hkl* index, leading to the same 1:1 correlation between lattice points and diffuse intensity values. Although it is tempting to use this method in our current analysis, the unit cell expansion method does not maintain the expected crystallographic symmetry for any crystal system with a screw axis. Introducing vacuum into our structure factor calculations will satisfy other symmetry operations, but as GpdQ possesses a screw axis we are currently unable to more finely sample its predicted diffuse scattering. Therefore, we can use this method to compare data between simulated models

of motion, but not between simulated models and experimental data. More advanced simulation methods will need to incorporate screw axes, either by defining a new supercell for simulation or directly calculating structure factors at fractional *hkl* indices. Cognizant of these limitations, we calculated the diffuse scattering of each of the GpdQ TLS ensembles to 3.0Å resolution in a P1 cell, with a sub-sampling of 4x4x4 around each Bragg lattice point (**Figure 3.10c).** These calculations confirm that each TLS motion produces distinct patterns of diffuse signal throughout reciprocal space.

3.8 Discussion

Accurate modeling of conformational dynamics is important for understanding macromolecular function. Although many models may fit the existing data equally well, they can often suggest different correlated motions. Our results indicate that comparisons to experimental diffuse scattering can break the degeneracy between different TLS refinements, as different selections of rigid bodies (along with their associated correlations) can produce markedly different diffuse patterns. For example, alternative correlations between TLS groups have equivalent average electron density, but result in unique diffuse scattering predictions. More generally, any model proposed through TLS refinement should agree with the experimental diffuse data, as this data directly reflects the existing protein disorder (Moore, 2009).

Despite this synergy between TLS refinement and diffuse scattering, there are many potential complications when applying TLS X-ray refinement to model protein dynamics. As the T and L matrices describe independent translations and librations, these motions must be physically sensible. Our review of protein structures deposited in the Protein Data Bank indicates that 88%

of refinements employing TLS (24% of the total PDB) do not satisfy this physical requirement (Urzhumtsev *et al.*, complementary paper). We hypothesize that this discrepancy arises due to a lack of restraints applied to refined TLS parameters to ensure their physical plausibility. Even if this criterion is met, current TLS refinement methods still do not impose chemical restraints between TLS groups, which can lead to displacements that are chemically unreasonable. Our TLS refinement of the GpdQ subdomain is one such example, as it produces rigid body displacements that extend across the entirety of the unit cell **(Figure 3.3c)**. Thus, validation checks of TLS refinement (such as those implemented in *phenix.tls_models*) are critical, as is employing TLS refinement within a broader framework of restraints. Alternative techniques, such as the Phase Integrated Method (PIM), which derives anisotropic B factors from low-frequency normal modes (Chen *et al.*, 2010), may significantly improve the biochemical accuracy of modeling efforts. In PIM, the fit between model and experiment is significantly improved by calculating normal modes in the context of the asymmetric unit rather than individual molecules (Lu & Ma, 2013).

Numerous sources of crystalline disorder combine to produce observed diffuse intensity patterns. Perhaps the most critical step in diffuse signal analysis is the determination of the relative contribution from each source; phenix.diffuse represents an important step towards performing such an investigation. Many causes of disorder can be described in terms of structural ensembles; thus, our tool enables the diffuse scattering produced by each source to be calculated. As experimental diffuse intensity is simply the sum of its independent components, optimizing the relative weights of the hypothesized sources of disorder to best fit the observed diffuse scattering may provide a feasible method of comprehensive diffuse scattering analysis.

With the increasing availability of modeling tools, the lack of high-quality three-dimensional datasets is now a key bottleneck in diffuse scattering analysis. One challenge in data collection is that long X-ray exposures can be required to reveal diffuse features. This can lead to blooming around saturated Bragg spots in diffraction images collected using commercially-available charge-coupled device (CCD) area detectors (Gruner *et al.*, 2002) Blooming can artificially increase pixel values between the Bragg spots, where the diffuse intensity is measured (Glover *et al.*, 1991). Although CCD detectors can be configured to eliminate spot blooming at the cost of decreasing dynamic range (Wall, 1996; Wall *et al.*, 1997), this configuration is not available in commercial detectors. The development of pixel array detectors, which possess much higher dynamic ranges as well as very small point-spread functions, has opened the door to more accurate measurement of diffuse signal. Additionally, methods for processing diffuse scattering data from raw image frames to complete reciprocal space map are under active development (Wall *et al.*, 2014). Because acoustic scattering is maximized at Bragg peaks (Glover *et.al*, 1991), diffuse signal will be most straightforward to measure in intervening regions. These methods will be applied to new datasets of simultaneous Bragg and diffuse scattering data. Instead of being included in the background corrections in estimating Bragg peak intensities, these diffuse intensities will increase the data available for refinement, enable more accurate quantification of interatomic distances (Kuzmanic *et al.*, 2011), and allow for the simultaneous refinement of multiple coupled protein motions (Wilson, 2013).

3.9 Tables

Table 3.1) Eigenvalues of GpdQ TLS refinement matrices. a) "Entire molecule" **b)** "Monomer" **c)** "Sub-domain". It is important to note that, for the "Sub-domain" refinement, L#5 and L#6 have negative eigenvalues. Due to their low value, however, these eigenvalues were set to zero for subsequent calculations.

a)

T	L
0.854	1.405
0.258	0.717
0.338	0.172

b)

T Monomer A	L Monomer A	T Monomer B	L Monomer B
0.873	1.843	0.850	1.896
0.236	0.021	0.192	1.103
0.327	0.822	0.329	0.500

c)

T Sandwich (A)	L Sandwich (A)	T Dimerization (A)	L Dimerization (A)	T Cap (A)	L Cap (A)
0.917	0.005	0.942	1.420	0.902	0.154
0.247	1.396	0.313	0.181	0.475	0.062
0.367	0.957	0.375	0.057	0.323	0.005

T Sandwich (B)	L Sandwich (B)	T Dimerization (B)	L Dimerization (B)	T Cap (A)	L Cap (A)
0.940	0.216	0.897	1.07	0.938	0.155
0.170	1.265	0.267	0.059	0.638	**-0.003**
0.399	0.838	0.368	**-0.001**	0.477	0.031

Table 3.2) Multi-model ensembles are necessary for adequate random sampling of TLS motions. Two ensembles independently sampling the underlying TLS distributions were used to generate anisotropic diffuse scattering maps. Global CC values between the two maps are shown. These simulations were conducted in triplicate, producing the CC standard deviation shown in parentheses. All maps were simulated to 3 Angstrom resolution.

	10	***50***	***100***	***500***	***1000***
Entire Molecule	0.886 (0.027)	0.956 (0.019)	0.988 (0.005)	0.996 (0.002)	0.999 (0.000)
Monomer	0.809 (0.087)	0.924 (0.008)	0.952 (0.005)	0.992 (0.002)	0.997 (0.001)
Sub-domain	0.944 (0.012)	0.984 (0.005)	0.992 (0.001)	0.999 (0.000)	0.999 (0.000)

3.10 Figures

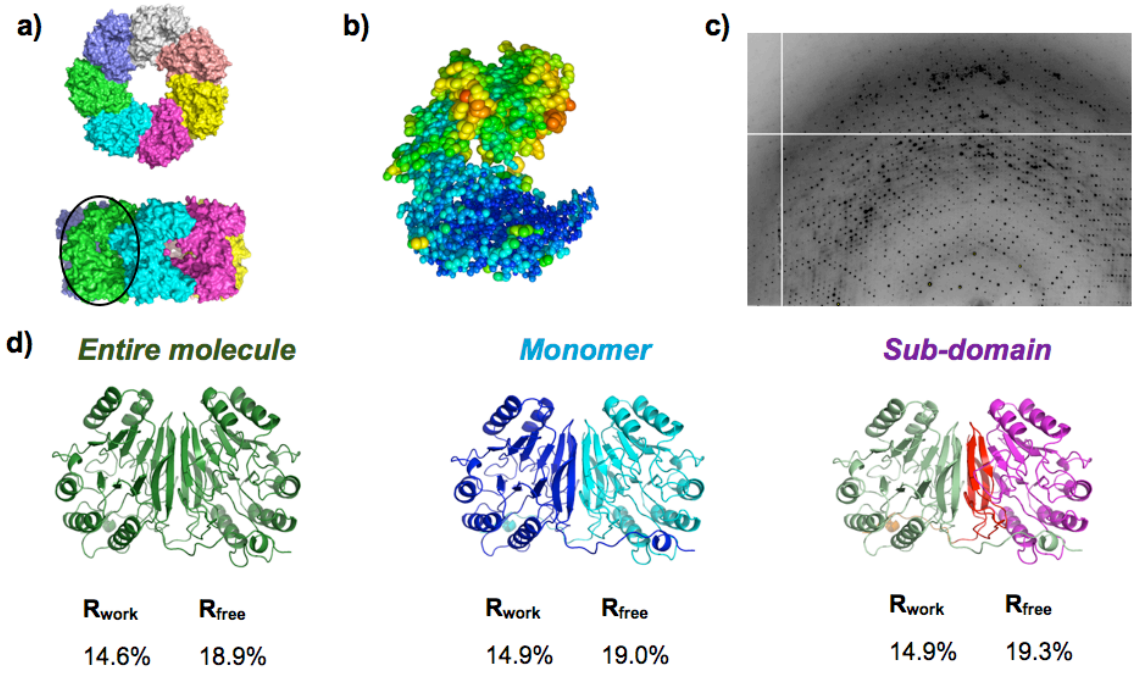


Figure 3.1) TLS refinement suggests macromolecular motions linked to function a) Top and side view of GroEL. Each color denotes a unique chain. **b)** TLS refinement of GroEL subunits reveals a "tilting" motion around the center of the subunit. **c)** GpdQ diffraction image showing significant diffuse scattering features. **d)** Refinement of GpdQ fails to produce substantial changes in R_{work} and R_{free} values between alternate TLS groups. TLS refinement significantly improves the overall R_{free} (23.1% pre-TLS).

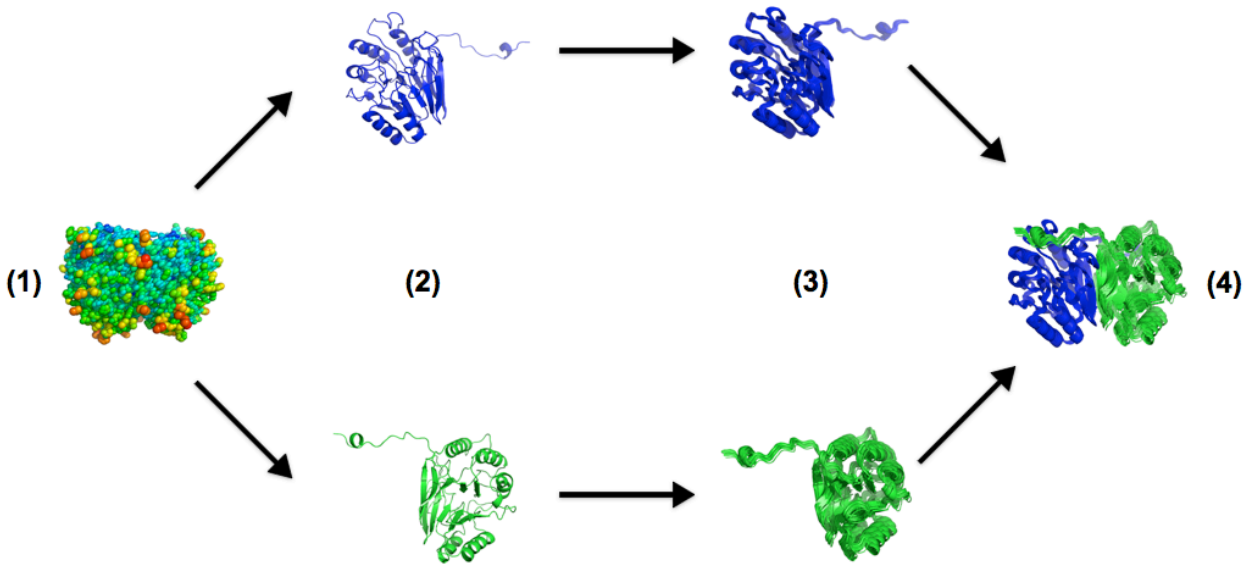


Figure 3.2) Overview of *Phenix.tls_as_xyz* The input PDB **(1)** is broken down into its constituent TLS groups **(2)** and TLS ensembles are generated for each group independently **(3)**. These groups are then re-assembled into the complete protein structure on a model-by-model basis **(4)**.

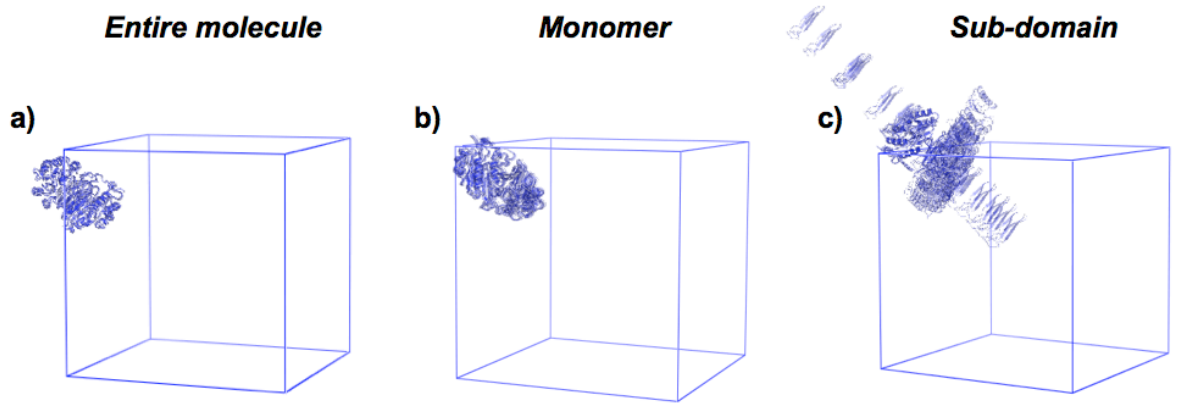


Figure 3.3) Structural ensembles of GpdQ TLS motions. Each TLS PDB ensemble is shown as a single asymmetric unit outlined by the unit cell. An increase in overall motion is apparent going from left to right. The 20 member ensemble is shown for visual simplicity. It is important to note that the chemically unreasonable motion produced by the sub-domain TLS refinement is not immediately apparent from the T and L eigenvalues presented in Supplemental Table 1, highlighting the need for the more thorough matrix analysis presented in our accompanying paper.

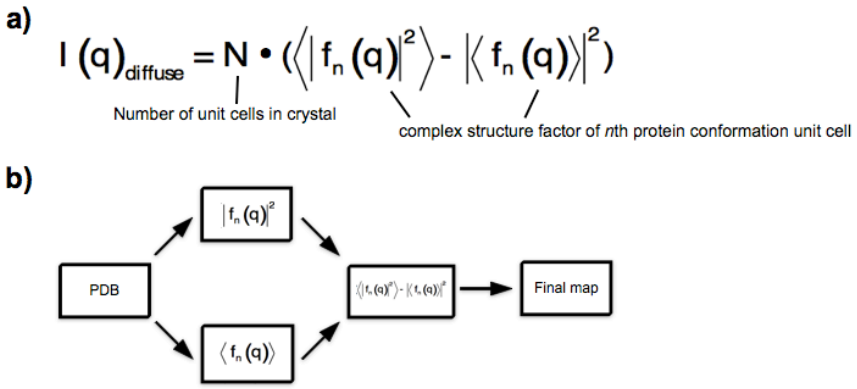


Figure 3.4) Overview of *Phenix.diffuse* a) The general form of Guinier's equation, The motion to be analyzed is captured in a series of "snapshots" defined by the the multi-model PDB. **b)** The general program flow. Each term in Guinier's equation is calculated separately from the structural ensembles and then combined to obtain the final map.

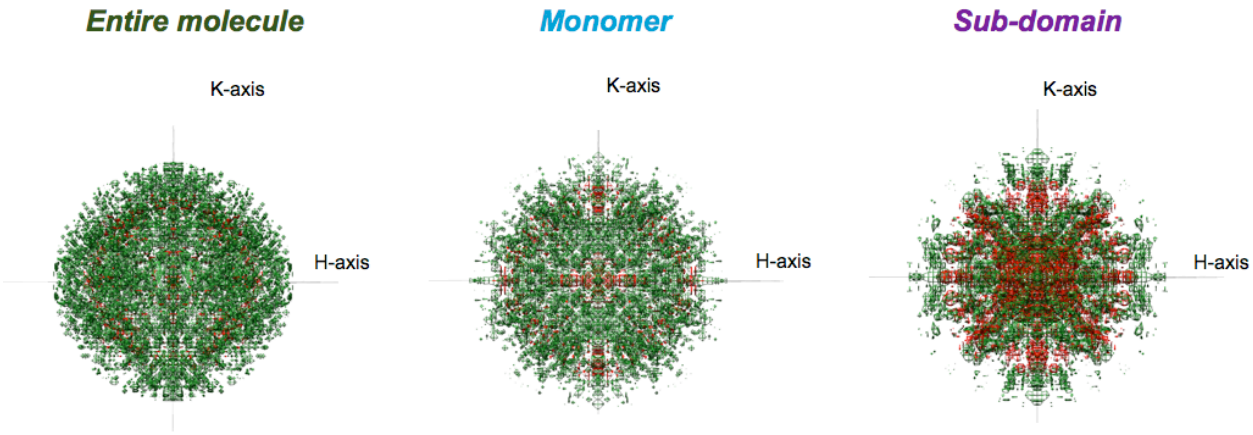
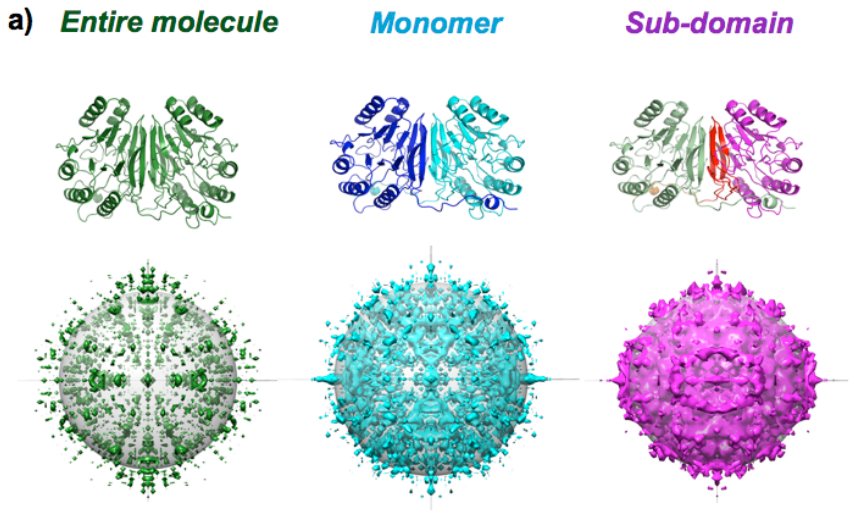


Figure 3.5) Anisotropic diffuse scattering maps. Positive and negative anisotropic density is shown as green and red mesh, respectively. Absolute threshold levels shown for the positive and negative signals are equivalent. The maps are shown to their full 3 Angstrom resolution limit.



b)

	Entire molecule	Monomer	Sub-domain
Entire molecule	1	0.300	0.031
Monomer		1	0.312
Sub-domain			1

Figure 3.6) Differing TLS groups produce unique diffuse scattering. a) The GpdQ TLS groups projected onto the structure, along with the calculated diffuse scattering (looking down the L axis; grey sphere denotes 4Å resolution). The "Monomer" and "Sub-domain" maps are shown at equivalent density thresholds, while "Entire molecule" is set at 60% of the density threshold. No correlation is assumed between TLS rigid body groups. **b)** Pearson correlation coefficients between anisotropic maps.

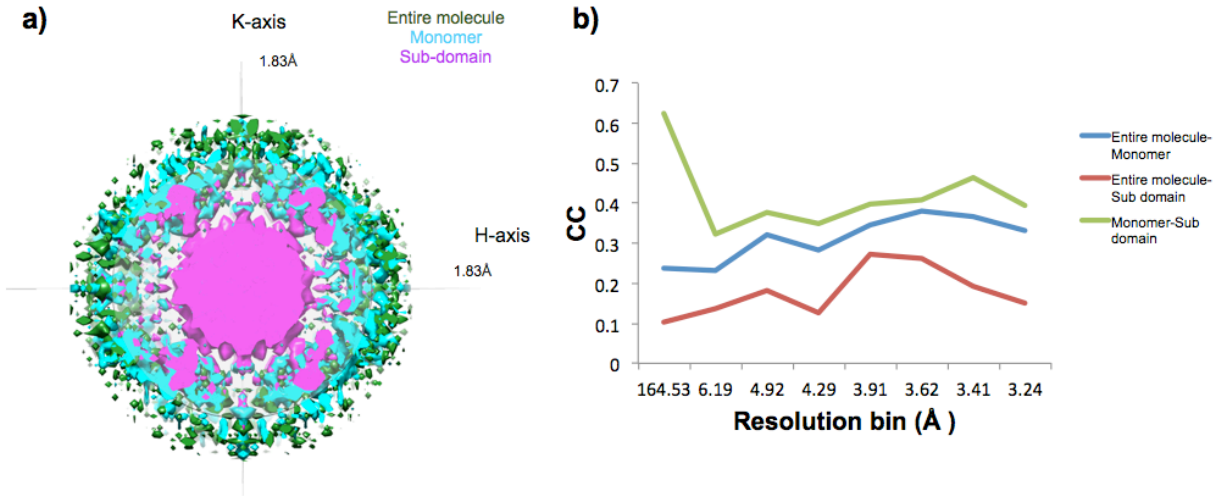


Figure 3.7) Comparison of simulated GpdQ TLS diffuse scattering maps. a) Cross-section of simulated TLS diffuse scattering maps. Primary and secondary diffuse intensity shells, separated by a gap, can be observed in each model. As the number of TLS groups increase, the intensity shells grow closer, predominantly due to an expansion in primary intensity shell size. **b)** Pearson correlation values between each set of maps across resolution bins.

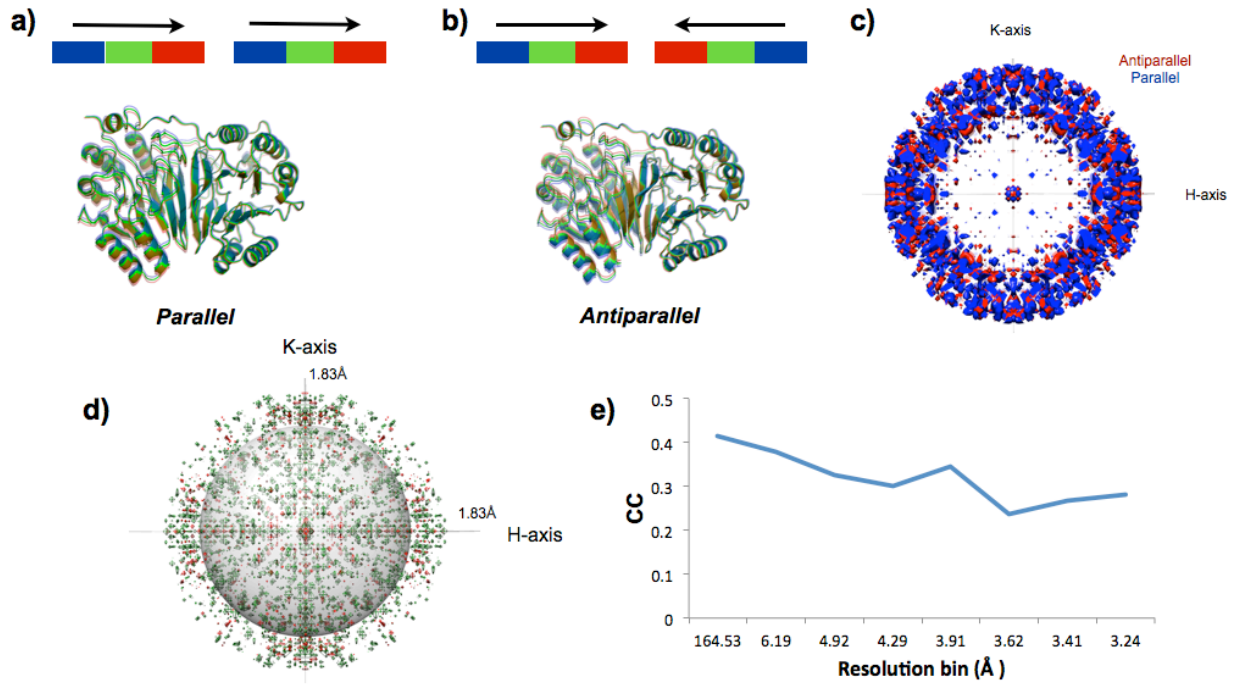


Figure 3.8) Different correlations between TLS groups produce unique diffuse scattering. Parallel **(a)** and antiparallel **(b)** TLS motions in GpdQ chains result in measurable differences between diffuse scattering patterns (CC = 0.375). Color bars indicate the directionality of the TLS motions; each color represents a unique molecular position. **c)** A map cutaway reveals strong secondary shell features with a small primary diffuse shell (looking down the L axis; grey sphere denotes 4Å resolution). **d)** Intensity differences between raw "anti" and "parallel" diffuse maps (green: positive, red: negative) highlights the qualitative changes caused by alternative TLS group correlations. **e)** Correlation values across anisotropic map resolution bins reveal highest correlation occurs between the maps at low resolution and decreases as a function of scattering vector length.

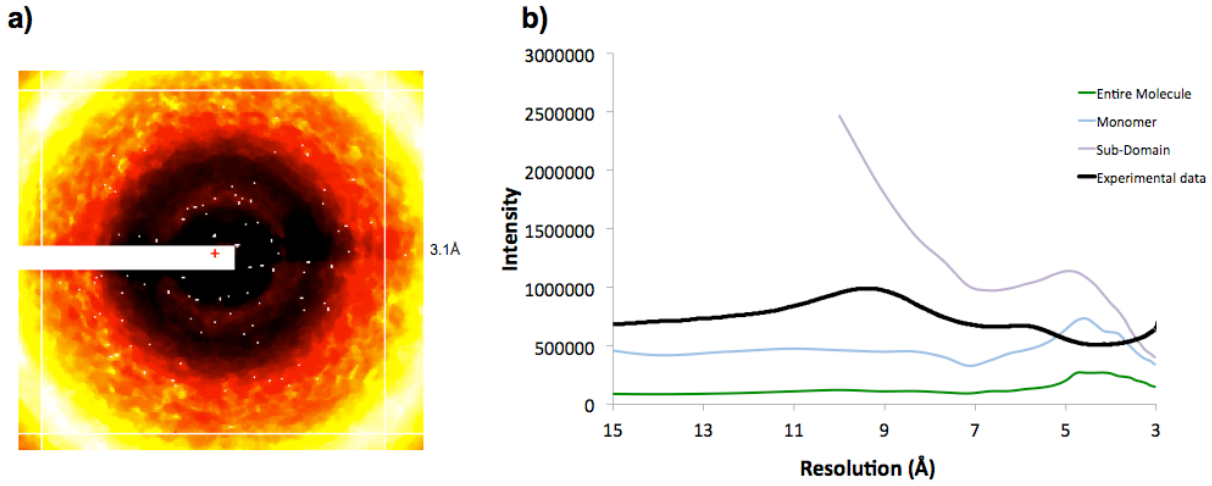
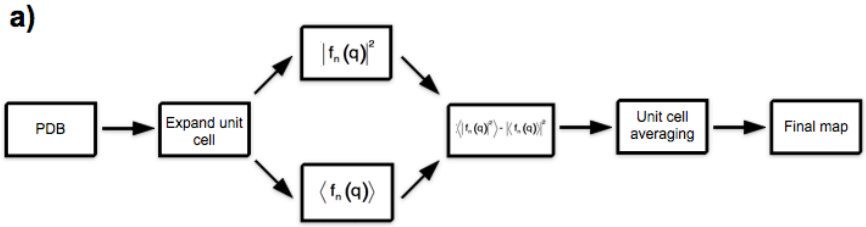
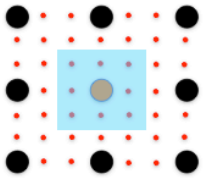


Figure 3.9) TLS models yield unique radial profiles of diffuse intensity. a) Mode filtered GpdQ diffraction image used for radial intensity calculation. The white regions correspond to pixels thrown out due to detector panel and beamstop artifacts, as well as Bragg scattering contamination. **b)** Radial diffuse intensity profiles for experimental and simulated GpdQ data. Resolution data below 15Å (roughly corresponding to the primary diffuse shell) were removed for more accurate visual comparison. The "Sub-domain" map exceeds the limits of the Y-axis at lower than 10Å resolution.

a)



b)



c)

	Entire molecule	*Monomer*	*Sub-domain*
Entire molecule	1	0.389	0.057
Monomer		1	0.113
Sub-domain			1

Figure 3.10) Unit cell expansion allows for reciprocal space subsampling a) The input PDB's unit cell is expanded to create the desired unit cell sampling, each term in Guinier's equation is calculated separately and then the second term is subtracted from the first to obtain the diffuse intensity. The "pseudo-unit cells" are then averaged across, producing the final diffuse scattering map. **b)** Unit cell expansion allowing for 3x subsampling of reciprocal space. True/"pseudo" Bragg peaks are shown in black/orange and red, respectively. The intensity values of the eight pseudo peaks and one orange peak in the blue box are averaged and the resulting value is assigned to the orange peak's Bragg index. **c)** Pearson correlation coefficients between maps.

3.11 References

Adams, P. D., Afonine, P. V., Bunkoczi, G., Chen, V. B., Davis, I. W., Echols, N., Headd, J. J., Hung, L. W., Kapral, G. J., Grosse-Kunstleve, R. W., McCoy, A. J., Moriarty, N. W., Oeffner, R., Read, R. J., Richardson, D. C., Richardson, J. S., Terwilliger, T. C. & Zwart, P. H. (2010). *Acta crystallographica. Section D, Biological crystallography* **66**, 213-221.

Afonine, P. V., Grosse-Kunstleve, R. W., Echols, N., Headd, J. J., Moriarty, N. W., Mustyakimov, M., Terwilliger, T. C., Urzhumtsev , A., Zwart, P. H. & Adams, P. D. (2012). *Acta Crystallographica Section D: Biological Crystallography* **68**, 352-367.

Amorós, J. L. & Amorós, M. (1968). *Molecular crystals: their transforms and diffuse scattering* (Vol. 6). New York: Wiley.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I.N. & Bourne, P. E. (2000). *Nucleic acids research*, *28*(1), 235-242.

Bernstein, F. C., Koetzle, T. F., Williams, G. J., Meyer, E. F., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). *European Journal of Biochemistry* **80**, 319-324.

Bricogne, G. (1993). *Acta crystallographica. Section D, Biological crystallography* **49**, 37-60.

Bricogne, G., Blanc, E., Brandl, M., Flensburg, C., Keller, P., Paciorek, W. Roversi, P., Sharff, A., Smart, O.S, Vonrhein, C. & Womack, T. O. (2011). *Cambridge, United Kingdom: Global Phasing Ltd*.

Burnley, B. T., Afonine, P. V., Adams, P. D. & Gros, P. (2012). *Elife* **1**.

Chacko, S. & Phillips, G.N. Jr. (1992) *Biophysical Journal* **61**, 1256-1266

Chaudhry, C., Horwich, A. L., Brunger, A. T. & Adams, P. D. (2004). *Journal of molecular biology* **342**, 229-245.

Chen, X., Wang, Q., Ni, F. & Ma, J. (2010). *Proceedings of the National Academy of Sciences*, 107,11352-11357.

Clarage, J. B., Clarage, M. S., Phillips, W. C., Sweet, R. M. & Caspar, D. L. (1992). *Proteins* **12**, 145-157.

Clarage, J. B., Phillips, G.N. Jr (1997) *Methods in Enzymology* **277**, 407-432

DeLano, W. L. (2002). The PyMOL molecular graphics system.

Doucet, J. & Benoit, J. P. (1987). *Nature* **325**, 643-646.

Emsley, P., & Cowtan, K. (2004). Coot: model-building tools for molecular graphics. *Acta Crystallographica Section D: Biological Crystallography*, *60*(12), 2126-2132.

Estermann, M. A. & Steurer, W. (1998). *Phase Transitions* **67**, 165-195.

Fraser, J. S., Clarkson, M. W., Degnan, S. C., Erion, R., Kern, D. & Alber, T. (2009). *Nature* **462**, 669-673.

Fraser, J. S., & Jackson, C. J. (2011) *Cellular and Molecular Life Sciences*, *68*(11), 1829-1841.

Faure, P., Micu, A., Perahia, D., Doucet, J., Smith, J.C., & Benoit, J. P. (1994) *Nature* **1**, 124-128

Glover, I. D., Harris, G. W., Helliwell, J. R., & Moss, D. S. (1991). *Acta Crystallographica Section B: Structural Science* **47**, 960-968.

Gros, P., van Gunsteren, W. F., & Hol, W. G. (1990). *Science* **249**, 1149-1152.

Gruner, S. M., Tate, M. W. & Eikenberry, E.F. (2002) *Review of Scientific Instruments* **73**, 2815-2842

Guinier, A. (1963). *X-ray diffraction in crystals, imperfect crystals, and amorphous bodies*. Courier Dover Publications.

Hery, S., Genest, D. & Smith, J.C (1998) *Journal of Molecular Biology* **279**, 303-319

Jackson, C. J., Carr, P. D., Kim, H. K., Liu, J. W., & Ollis, D. L. (2006). *Acta Crystallographica Section F* **62**(7), 659-661.

Jackson, C. J., Carr, P. D., Liu, J. W., Watt, S. J., Beck, J. L. & Ollis, D. L. (2007). *Journal of molecular biology* **367**, 1047-1062.

Janowski, P. A., Cerutti, D. S., Holton, J. & Case, D. A. (2013). *Journal of the American Chemical Society* **135**, 7938-7948.

Korostelev, A. & Noller, H. F. (2007). *Journal of molecular biology* **373**, 1058-1070.

Kendall, M. G., & Stuart, A. (1958) *The Advanced Theory of Statistics.* C. Griffin & C., London.

Kuzmanic, A., Kruschel, D., van Gunsteren, W. F., Pannu, N. S. & Zagrovic, B. (2011). *Journal of molecular biology* **411**, 286-297.

Kuzmanic, A., Pannu, N. S. & Zagrovic, B. (2014). *Nature communications* **5**, 3220.

Lang, P. T., Holton, J. M., Fraser, J. S. & Alber, T. (2014). *Proceedings of the National Academy of Sciences of the United States of America* **111**, 237-242.

Levin, E. J., Kondrashov, D. A., Wesenberg, G. E. & Phillips G. N., Jr. (2007). *Structure* **15**, 1040-1052.

Linder, B. & Smith, J.C. (2012) *Computer Physics Communications* **183**, 1491-1501

Lu, M., & Ma, J. (2013). *Journal of Molecular Biology*, **425**, 1082-1098.

Meinhold, L. & Smith, J.C. (2005). *Physical Review Letters* **95**, 218103

Meinhold, L. & Smith, J. C. (2007). *Proteins* **66**, 941-953.

Michels-Clark, T., Lynch, V., Hoffmann, C., Hauser, J., Weber, T., Harrison, R. & Burgi, H. (2013). *Journal of Applied Crystallography* **46**, 1616-1625.

Micu, A. M. & Smith, J.C. (1994) *Computer Physics Communications* **91** 331-338

Mizuguchi, K., Kidera, A. & Go, N. (1994) *Proteins* **18**, 34-48

Moore, P. B. (2009). *Structure* **17**, 1307-1315.

Murshudov, G. N., Vagin, A. A., & Dodson, E. J. (1997). *Acta Crystallographica Section D: Biological Crystallography* **53**, 240-255.

Painter, J. & Merritt, E. A. (2005). *Acta crystallographica. Section D: Biological crystallography* **61**, 465-471.

a) Painter, J., & Merritt, E. A. (2006a). *Acta Crystallographica Section D: Biological Crystallography* **62**, 439-450.

b) Painter, J. & Merritt, E.A. (2006b) *Journal of Applied Crystallography* **39**, 109-111

Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., & Ferrin, T. E. (2004). *Journal of computational chemistry* **25**, 1605-1612.

Perez, J., Faure, P. & Benoit, J.P. (1996) *Acta crystallographica. Section D, Biological Crystallography* **52**, 722-729

Phillips, G.N Jr, Fillers, J.P. & Cohen, C. (1980) *Biophysical Journal* **32**, 485-502

Qin, B. Y., Bewley, M. C., Creamer, L. K., Baker, H. M., Baker, E. N. & Jameson, G. B. (1998). *Biochemistry* **37**, 14014-14023.

Rader, S. D. & Agard, D. A. (1997). *Protein science* **6**, 1375-1386.

Riccardi, D., Cui, Q. & Phillips, G. N., Jr. (2010). *Biophysical journal* **99**, 2616-2625.

Ruschak, A. M. & Kay, L. E. (2012). *Proceedings of the National Academy of Sciences of the United States of America* **109**, E3454-3462.

Schomaker, V. & Trueblood, K. N. (1968). *Acta Crystallographica Section B* **24**, 63-76.

Tickle, I. J., & Moss, D. S. (1999). *IUCr99 Computing School, London, United Kingdom* [Online *http://people.cryst.bbk.ac.uk/~tickle/iucr99/iucrcs99.html*]

Thüne, T., & Badger, J. (1995) *Progress in biophysics and molecular biology*, *63*(3), 251-276.

Urzhumtsev, A., Afonine, P. V., Van Benschoten, A. H., Fraser, J. S. & Adams, P. D. (2014) *Acta Crystallographica Section D: Biological Crystallography*, submitted.

Urzhumtsev, A., Afonine, P. V. & Adams, P. D. (2013). *Crystallography Review* **19**, 230-270.

van den Bedem, H., Dhanik, A., Latombe, J. C., & Deacon, A. M. (2009). *Acta Crystallographica Section D: Biological Crystallography*,*65*(10), 1107-1117.

van den Bedem, H., Bhabha, G., Yang, K., Wright, P. E. & Fraser, J. S. (2013). *Nature methods* **10**, 896-902.

Wall, M. E. (1996). PhD thesis*,* Princeton University.

Wall, M. E. (2009). *Methods in molecular biology* **544**, 269-279.

Wall, M. E., Adams, P. D., Fraser, J. S. & Sauter, N. K. (2014). *Structure* **22**, 182-184.

Wall, M.E., Clarage J.B. & Phillips, G.N. Jr (1997) *Structure* **5**, 1599-1612

Wall, M. E., Ealick, S. E. & Gruner, S. M. (1997). *Proceedings of the National Academy of Sciences of the United States of America* **94**, 6180-6184.

Wall, M.E., Van Benschoten, A.H., Sauter, N.K, Adams, P.D, Fraser, J.S & Terwilliger, T.C. (2014) *Proceedings of the National Academy of Sciences of the United States of* America **111**, 17887-17892

Welberry, T.R & Butler, B.D. (1994) *Journal of Applied Crystallography* **27**, 205-231

Williams, B.B., Van Benschoten, A. H., Cimermancic, P., Donia, M. S., Zimmermann, M., Taketani, M., Ishihara, A., Kashyap, P. C., Fraser, J.S. & Fischbach, M. A. (2014) *Cell Host & Microbe,* **16**, 495-503

Wilson, M. A. (2013). *Nature methods* **10**, 835-837.

Winn, M. D., Isupov, M. N. & Murshudov, G. N. (2001). *Acta crystallographica. Section D, Biological crystallography* **57**, 122-133.

Chapter 4

Mapping and modeling X-ray diffuse scattering from protein crystals

4.1 Abstract

Correlated atomic motions underlie macromolecular functionality. While it is challenging to uncover these movements through traditional biophysical experiments, X-ray diffuse scattering directly reports on these linked atomic motions. However, measuring and analyzing diffuse scattering signal has been historically challenging. Here, we introduce *LUNUS-DIALS*, a comprehensive experimental and computational pipeline for creating three-dimensional maps of diffuse scattering. This tool was applied to diffraction datasets collected from Cyclophilin A and trypsin crystals, resulting in high-resolution reciprocal space diffuse scattering maps from both proteins. We used these maps to break molecular motional degeneracy resulting from alternative TLS refinement strategies and select the TLS grouping most consistent with the diffuse data. We also applied the liquid-like motions model to CypA, revealing a high degree of global correlation with implications for the scientific understanding of crystalline lattices. These results represent a significant technical advancement in macromolecular diffuse analysis and direct evidence of diffuse scattering's ability to uncover correlated motions.

4.2 Introduction

Structural dynamics play a critical role in enzymatic reactions (Williams *et.al*, 2014), protein-protein interactions and signaling cascades (Woldeyes *et.al*, 2014). One of the most useful tools for uncovering these motions is X-ray crystallography, which provides atomic-level detail of a molecule's inner workings. Crystallography is by definition an ensemble experiment: unit cell heterogeneity results from multiple conformational states, thermal motion and crystalline defects.

Although crystallographers have historically modeled proteins as single structural snapshots, parameters have been gradually introduced to account for both atomic and molecular motion. Examples include thermal B-factors (Frauenfelder *et.al*, 1979) and Translation-Libration-Screw structural refinement (Schomaker & Trueblood, 1968). Newly-emergent methods such as room-temperature crystallography (Fraser *et.al*, 2011) and the X-ray Free Electron Laser (Keedy *et.al*, 2015) provide further avenues for understanding the motion present in these once (supposedly) static environments.

However, there remains a fundamental limit in the amount of dynamics information present within modern crystallography experiments. These analyses focus on the signal at Bragg peaks: repetitive regions of reciprocal space containing significant X-ray scattering intensity. Crystallographic theory dictates that the signal present in these spots is a combination of all diffraction events occurring in the crystal (Bragg & Bragg, 1913); thus, the electron density maps produced from this data is an average of the electron density across all unit cells. This in turn can lead to degeneracy when attempting to model the correlation between atomic motions. Additional sources of information, such as patterns of steric clashes (van den Bedem *et al.*, 2013), NMR spectroscopy (Ruschak & Kay, 2012), or mutational analysis (Fraser *et al.*, 2009) can often be used to distinguish competing models of correlated motion between non-bonded atoms. However, this need for external experiments further emphasizes the limitations of current crystallographic analysis in capturing protein motions.

A parallel data source capable of directly measuring atomic correlation is X-ray diffuse scattering (Phillips *et al.*, 1980; Chacko & Phillips, 1992; Faure *et al.*, 1994; Clarage & Phillips,

1997; Mizuguchi *et al.*, 1994). Deviations away from a perfect crystal (thermal motion, crystal defects, static disorder) violate the repetitive structure of the lattice and lead to diffraction outside of the regions of reciprocal space predicted by Bragg's Law. The relationship between unit cell heterogeneity and diffuse signal has been available for decades (Guinier, 1963; Amorós & Amorós, 1968). As a result, the ability of diffuse scattering to break the degeneracy between multiple models of motion has long been recognized and multiple attempts have been made over the past few decades to quantify and model this signal (Thune & Badger, 1995; Perez *et al.*, 1996; Hery *et al.*, 1998; Tickle & Moss, 1999). In every case, analysis consists of calculating diffuse intensity patterns produced by some hypothetical motion and comparing to experimental data. Multiple strategies currently exist for calculating diffuse scattering, including all-atom covariance matrices (Riccardi *et.al*, 2010) and our recently-published tool *phenix.diffuse* (Van Benschoten *et.al*, 2015), which uses Guinier's equation to calculate reciprocal space diffuse scattering map from multi-model PDB files.

More challenging has been the collection and measurement of experimental diffuse intensities. Diffuse scattering is 1000x weaker than Bragg signal and thus requires overexposure of the X-ray image. Unfortunately, overexposure leads to spot blooming and the washing out of diffuse signal in the region immediately surrounding Bragg peaks. Furthermore, diffuse data must be collected at room temperature, as cryogenic conditions will "freeze out" many of the motions contributing to the observed signal (Fraser *et.al*, 2011). However, room-temperature data collection greatly increases crystal radiation sensitivity and leads to faster crystal decay. Because of these technical challenges, experimental diffuse scattering data has historically been limited to single diffraction frames. Recently, Wall *et.al* (1997) developed a computational framework for

measuring diffuse scattering across the entirety of reciprocal space. Through mode filtering, sharp Bragg intensities are removed and the subsequent frames, which only contain diffuse scattering signal, can be scaled together and integrated into a complete diffuse map. Though this process was originally applied to specialized CCD detectors (Wall, 1996), the recent commercial development of pixel-array detectors, which possess tight point-spread functions and single-photon sensitivity, has opened up the realm of experimental diffuse scattering measurement to the general scientific community. Thus, the computational and experimental framework is now in place for diffuse scattering to become a routine component of structural analysis.

Despite this ability to measure diffuse scattering through mode filtering, several additional features are necessary for any comprehensive diffuse analysis software package. The processed images must be assembled into a three-dimensional lattice, symmetry operators must be applied and the resulting data written out in a standardized file format. Though diffuse signal is of primary interest, it is also critical to have the corresponding electron density map produced from the Bragg data. Not only does the Bragg signal provide a frame of reference for the raw diffuse intensities, but the electron density map supplies a crystal-specific context for subsequent modeling attempts. If this context is not considered, subsequent models of motion may fail to capture unique crystalline nuances. Additionally, spherically symmetric sources of diffuse signal such as disordered solvent and scattering angle effects contribute significantly to the observed intensity. These sources must be removed in order to facilitate more accurate motional analysis (Wall *et.al*, 1997). Here we introduce a new program, *DIALS-LUNUS*, that achieves all of these goals through the combination of pre-existing software and new computational routines **(Figure 4.1)**. The end result is a comprehensive diffuse scattering analysis pipeline capable of converting

raw datasets into corresponding maps of Bragg and diffuse diffraction, ultimately providing new insights into protein dynamics.

We used this program to construct diffuse datasets of the human isomerase cyclophilin A (CypA) and the serine protease trypsin. These datasets triple of the amount of three-dimensional macromolecular diffuse scattering maps available to the scientific community, providing a critical advancement for the field (Wall *et.al*, 2014). We then compared these maps to various models of correlated motion predicted from TLS structure refinement, as well as the Liquid-Like Motion (LLM) lattice model (Clarage *et.al*, 1992). Our results demonstrate that observed diffuse features can be accurately modeled and assist in the selection of competing hypotheses of motion. Long-term, diffuse X-ray scattering can guide structure refinement as well as further our understanding of the enzyme as a dynamic entity.

4.3 Methods

Protein purification and crystallization. Trypsin crystals were obtained according to the method of Liebschner *et.al* (2013). Lyophilized bovine pancreas trypsin was purchased from Sigma-Aldrich (T1005) and dissolved at a concentration of 30 mg/mL into 30mM HEPES pH 7.0, 5 mg/mL benzamidine and 3mM $CaCl_2$. Crystals were obtained from a solution of 200mM Ammonium sulfate, 100mM Na cacodylate pH 6.5, 20% PEG 8000 and 15% glycerol. CypA was purified and crystallized as previously described (Fraser *et.al*, 2009). Briefly, the protein was concentrated to 60 mg/mL in 20mM HEPES pH 7.5, 100mM NaCl and 500mM TCEP. Trays were set with a precipitant solution of 100mM HEPES pH 7.5, 22% PEG 3350 and 5mM TCEP. Both crystal forms were obtained using the hanging-drop method.

Crystallographic data collection. Diffraction data was collected at the Stanford Synchrotron Radiation Lightsource (Menlo Park, CA). Each dataset was collected from a single crystal at 273K. To prevent dehydration, crystals were coated in a thin film of paratone. For CypA, a single set of 0.5 degree oscillation images were collected and used for both Bragg and diffuse data processing. A total of 360 images were collected across a 180 degree phi rotation. Trypsin data was gathered through an "interleaved" process of collecting an image still for the diffuse signal, followed by a one degree oscillation for the Bragg data. This process was repeated for a total of 135 degrees.

Bragg data processing. All Bragg diffraction data was processed using XDS and XSCALE within the *xia2* software package (Winter, 2009). Molecular replacement solutions were found using the Phaser-MR package within the *Phenix* software suite (Adams *et.al*, 2010). The PDB search models were 4I8G (Trypsin) and 2CPL (CypA). Initial structural refinement was performed using *phenix.refine* (Afonine *et.al*, 2011). The strategy included refinement of individual atomic coordinates and water picking. Both the X-ray/atomic displacement parameters and X-ray/stereochemistry weights were optimized. Isotropic B-factos were chosen for the initial structures to allow for non-negligible R-factor optimization by subsequent TLS refinement strategies. All structures were refined for a total of 5 macrocycles. Statistics for the "initial" Bragg structures are shown in **Table 4.1**.

Diffuse data processing. Pixels corresponding to the beamstop and image edges were removed using the *punchim* and *windim* routines within the LUNUS software package (Wall, 1997). This was performed to avoid contaminant scattering and edge effects. Beam polarization was corrected using *polarim*; polarization values were determined by analyzing the azimuthal profile

of sample diffraction images. A solid-angle normalization (*normim*) correction was also applied. Finally, mode-filtering was used to remove Bragg peaks from diffraction images and leave only the diffuse background intensities. This was accomplished using the LUNUS routine *modeim*, with the mask and bin sizes set to 20 and 1, respectively. A radial scattering vector intensity profile was then calculated and used to scale diffuse frames across the entire dataset. The stills were then assembled into a 3D lattice using novel DIALS routines produced from the Computational Crystallography Toolbox (*cctbx.sourceforge.net*). Finally, a combination of LUNUS and *Phenix* programs were used to remove the spherically-symmetric scattering sources, symmetrize the experimental data and calculate correlation values both globally and across resolution bins.

Bragg model building

CypA TLS refinement. Three independent TLS refinements were performed on CypA. *Whole molecule* denotes selection of the entire molecule as a single TLS group. *Phenix* signifies identification of the TLS groups with the aid of *phenix.find_tls_groups*. A total of 8 TLS groups were chosen by this method: residues 2-14, 15-41, 42-64, 65-84, 85-122, 123-135, 136-145 and 146-165. *TLSMD* describes selection of the TLS groups through the TLS Motion Determination web server (Painter & Merritt, 2006a; Painter & Merritt, 2006b). Again, 8 TLS groups were identified: residues 2-15, 16-55, 56-80, 81-85, 86-91, 92-124, 125-143 and 144-165. All TLS refinement was performed within *phenix.refine* through 5 macrocycles. Aside from the inclusion of TLS refinement, these macrocycles were identical to the initial structure refinement described above. CypA diffuse maps were simulated to a resolution of 1.4 angstroms using *phenix.diffuse*.

Trypsin TLS refinement. The previously-described *whole molecule, Phenix* and *TLSMD* TLS refinement strategies were similarly applied to trypsin. Phenix automation selected 7 TLS groups: residues 16-54, 55-103, 104-123, 124-140, 141-155, 156-225 and 226-245. The TLSMD web server selected the following 9 groups: 16-52, 53-98, 99-115, 116-144, 145-171, 172-220, 221-224, 225-237 and 238-245. Diffuse maps were generated to a resolution of 1.25 angstroms using *phenix.diffuse*.

4.4 Experimental diffuse maps possess crystallographic symmetry

The symmetrized diffuse scattering maps are shown in **Figure 4.2**. The CypA dataset is 98% complete to a resolution of 1.4 Angstroms, while trypsin possesses 95% completeness to 1.25 Angstroms. Two unique statistics were defined to quantify the level of crystallographic symmetry in each anisotropic map. In order to evaluate the presence of Friedel symmetry, we averaged diffuse values across Friedel pairs to create a symmetrized map *I_Friedel* and calculated the Pearson Correlation Coefficient (PCC) between the symmetrized and unsymmetrized data to obtain the statistic $CC_{Friedel}$. For CypA and trypsin, $CC_{Friedel} = 0.90$ and 0.95 respectively, demonstrating that Friedel symmetry is conserved across diffuse intensities. In order to analyze the presence of space group symmetry in the diffuse maps, we averaged P222-related reflections (since both crystals displayed P 21 21 21 symmetry) to produce the map *I_p222*. This map was then compared to the unsymmetrized diffuse map to obtain a correlation value CC_{Sym}. High levels of symmetry are observed for both CypA ($CC_{Sym} = 0.70$) and trypsin ($CC_{Sym} = 0.69$). Thus, our data demonstrates that crystallographic symmetry is preserved in diffuse signal across multiple systems. It is also important to note that while the trypsin data was collected using the common "interleaved" diffuse data collection strategy (where still images are

captured in between standard one degree oscillations), the CypA dataset consisted entirely of oscillation frames, a procedure more frequently utilized in standard crystallography. As the CypA map displays equivalent levels of symmetry to the trypsin map, this suggests that diffuse signal is robust (to a degree) with respect to data collection method.

4.5 Diffuse signal can select between equivalent models of motion

CypA. In order to understand the molecular motions taking place in the CypA crystal, we compared the diffuse map to three alternative TLS refinements of the Bragg structure: *phenix*, *tlsmd* and *whole molecule* **(Figure 4.3a).** Although all three refinements predict different motions, the R-free values between the structures are nearly identical, again highlighting the limitations of Bragg structure refinement. All three maps show little similarity with each other: the global correlation coefficient between the maps range between 0.07 and 0.22 **(Figure 4.3b)**. Both the *phenix* and *TLSMD* motions display poor global fit to experimental data (PCC = 0.03 and 0.04 respectively), while the *whole molecule* TLS motion displays a better fit (0.14). An analysis of the correlation by resolution **(Figure 4.5)** shows that the *whole molecule* motion correlates best at low resolution (above 3A), while the *phenix* and *tlsmd* models remain consistently poor across resolution bins. This result prompted further investigation of the *whole molecule* map at the subsampled resolution range of 3A and below. Applying this cutoff to the data reveals that the map correlation peaks in the range between 3.56-3.37A, where PCC = 0.373. In this range the experimental map correlates less well with *phenix* (0.005) and *tlsmd* (0.157), demonstrating that comparing specific regions of reciprocal space may serve as a viable method of boosting the ability to best match motional hypothesis to experiment. As the experimental diffuse intensity is strongest within this resolution range (the “solvent ring” found

between 5.0 and 3.3 Angstroms), these results provide a solid foundation for further investigations int the correlated motions present within CypA.

Trypsin. As with CypA, our three trypsin TLS refinement strategies yielded roughly equivalent R_{free} values (16.6-16.7%; **Figure 4.4a**). Anisotropic map comparison reveals that although the *Whole molecule* motion is dissimilar to both the *Phenix* and *TLSMD* predictions (PCC = 0.03 and 0.05, respectively), the *Phenix* and *TLSMD* refinements show a significant degree of similarity with a global PCC of 0.515 **(Figure 4.4b)**. However, we believe that this occurs because the two refinement strategies selected similar TLS groups (as described in the Methods section). Comparison of the simulated maps to the experimental data again reveals that the *Whole molecule* strategy (PCC = 0.08) fits slightly better than either *Phenix* (PCC = 0.02) or *TLSMD* (PCC = 0.02), although to a lower degree than in the case of CypA. Map correlation values across resolution bin shows that all three simulations agree best with experimental data between 2.56-2.03 Angstroms. However, the low overall correlation between all three models of motion and the data suggest that additional refinement strategies not considered here may better explain the observed diffuse scattering. Alternatively, further improvements in data quality might lead to better agreement between model and experiment.

4.6 Liquid-like motions accurately model global diffuse features

In order to better explain the observed diffuse signal within CypA, we applied a liquid-like motions (LLM) model of molecular motion (Clarage *et.al*, 1992). The LLM framework describes the crystal lattice as a soft, elastic environment in which individual molecules undergo internal fluctuations and interact most strongly with nearest neighbors. These fluctuations decay exponentially according to a correlation length γ and standard deviation σ. Previous results

demonstrate that the LLM model is highly successful in fitting experimental diffuse scattering data (Wall, 1997; Wall, 1998). For CypA, parameter optimization (where γ = 6.0 Angstroms and σ = 0.53 Angstroms) yielded an overall correlation of 0.525 to the experimental data. This is roughly in line with Wall *et.al*'s previous LLM fit to Staph nuclease diffuse maps, where PCC = 0.595. Subdividing the model into resolution bins reveals that the highest correlation between data and experiment occurs in the range 2.39-2.09A, where PCC = 0.665 **(Figure 4.6)**.

4.7 Discussion

Diffuse scattering has long been viewed as a valuable source of information on protein dynamics, but experimental and computational challenges have hampered extensive investigation. Our results represent a significant step towards moving X-ray diffuse scattering analysis into the mainstream of structural biology. DIALS-LUNUS makes possible the analysis of diffuse data collected on a wide range of available detectors, with no need for technical modifications or the development of complex computational scripts. Furthermore, our comparison of CypA and trypsin maps to various refinement models demonstrate that diffuse signal can serve to distinguish between degenerate models of motion and potentially guide refinement efforts.

It is not surprising that the *Whole molecule* TLS refinement provides the best agreement with experimental data. This group selection roughly approximates many crystalline imperfections present in the crystal, including vibrational phonons and lattice shear. Indeed, treating the entire molecule as a single TLS group is a common strategy for modeling global crystalline disorder during macromolecular refinement (Burnley *et.al*, 2012). The results of Doucet and Benoit (1987) revealed that correlated short-range rigid movements take place between molecules

within the crystalline unit cell, further supporting this observation. However, Clarage *et.al* (1992) proposed that these rigid unit cell motions might only supply ~10% of the observable diffuse scattering, which could explain the observed correlation of 0.143 in the case of CypA.

The agreement of the liquid-like motions model with experimental diffuse data across multiple systems warrants further consideration. Whereas most crystallographic refinement programs consider the crystal lattice to be a rigid entity (with a bit of thermal noise), the fit between LLM and diffuse data strongly suggests that this assumption should be revisited. This model provides a new paradigm for understanding and modeling intra-lattice effects: a gelatinous macromolecular environment in which molecules soften and interact with each other across defined correlation distances. Indeed, changing the description of the crystalline lattice to this softer network could provide significant improvements in both Bragg and diffuse structural refinement.

Because our optimal correlation between model and map is only 0.525 for the CypA LLM, it is clear that there is still much room for improvement in modeling the crystalline disorder contributing to diffuse scattering. Future efforts will need to take into consideration the myriad of intramolecular (loop openings, side chain flips), intermolecular (phonons, lattice contacts) and crystal-specific (mosaicity, microlattices) effects. However, this work clearly demonstrates the value of simultaneously refining Bragg and diffuse data in crystallographic analysis. Rather than being thrown out as noise, diffuse intensities can provide a valuable set of experimental restraints and unlock new insight into protein dynamics. Further work will need to be performed in order for this goal to be completely realized; however, the computational and experimental routines presented here will greatly simplify this process.

4.8 Tables

Table 4.1) Initial Bragg data refinement statistics

	CypA	Trypsin	GpdQ
Resolution range	38.66-1.4	23.29-1.25	45.56-2.9
Space group	P 21 21 21	P 21 21 21	P 21 3
Unit cell	42.91, 52.44, 89.12	54.81, 58.51, 67.42	164.27, 164.27, 164.27
Completeness (%)	98	95	100
R_{work} (%)	17.88	15.90	19.67
R_{free} (%)	19.50	17.41	23.37
RMS (bonds)	0.007	0.013	0.007
RMS (angles)	1.16	1.61	1.33
Ramachandran favored	97	98	94
Ramachandran allowed	3	2	5
Ramachandran outliers	0	0	1
Clashscore	0.79	2.59	10.89
Average B-factor	21.42	14.57	68.73

4.9 Figures

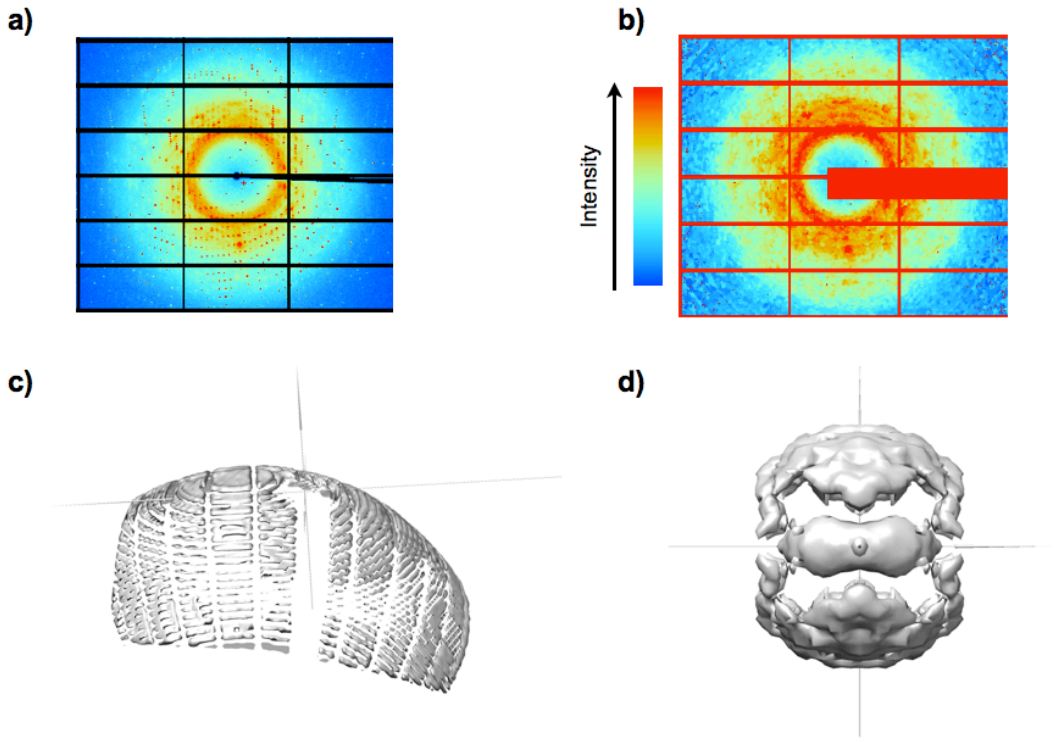
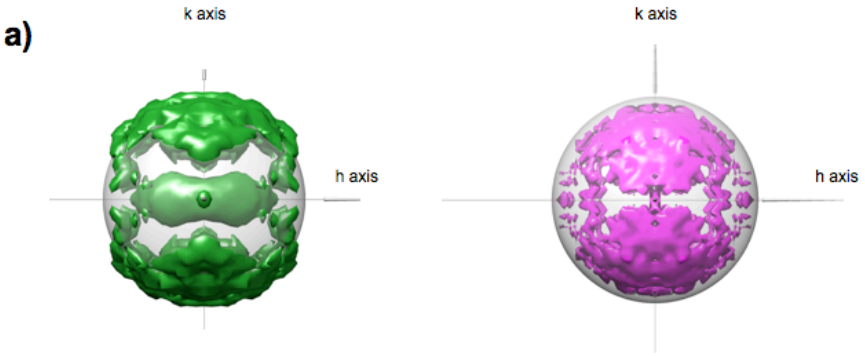


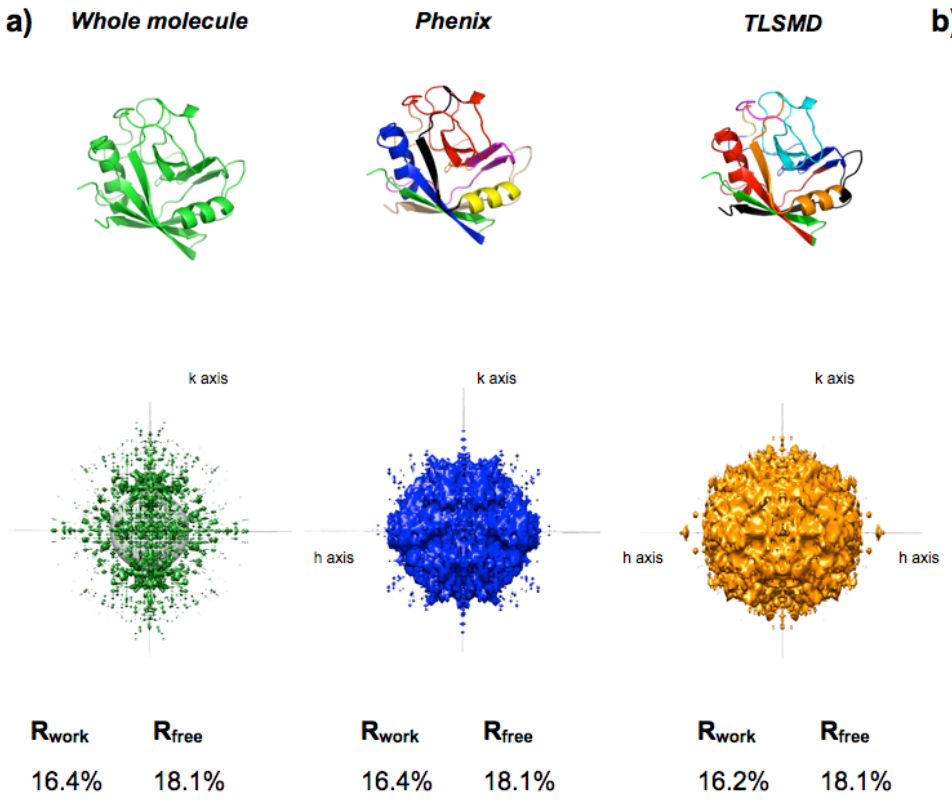
Figure 4.1) Overview of DIALS-LUNUS Raw CypA diffraction images **(a)** are thresholded and mode filtered **(b)** to remove Bragg peaks and contaminant scattering. These processed images are integrated in parallel to produce individual slices of the diffuse scattering across reciprocal space **(c)**. All slices are combined and symmetrized into a comprehensive three-dimensional map.



b)

Map	$CC_{Friedel}$	CC_{Sym}
CypA	0.90	0.70
Trypsin	0.95	0.69

Figure 4.2) Experimental diffuse scattering maps a) Isotropic diffuse scattering maps after application of symmetry operators. Grey reference spheres denote a resolution of 2.0A. **b)** Crystallographic symmetry statistics for anisotropic maps.

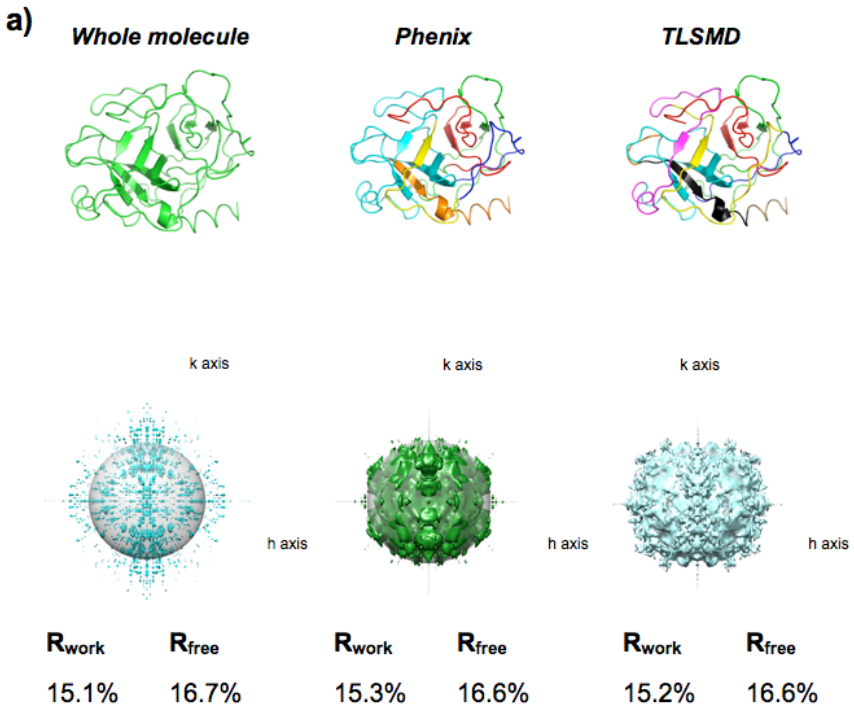


b)

	Whole molecule	Phenix	TLSMD
Whole molecule	1	0.116	0.066
Phenix		1	0.220
TLSMD			1

Figure 4.3) Different CypA TLS groups produce unique diffuse scattering. a) The CypA TLS groups projected onto the structure, along with the calculated diffuse scattering (looking down the L axis; grey sphere denotes 4Å resolution). All maps are shown at an equivalent intensity isosurface. **b)** Pearson correlation values between anisotropic maps.

a)



b)

	Whole molecule	*Phenix*	*TLSMD*
Whole molecule	1	0.026	0.048
Phenix		1	0.515
TLSMD			1

Figure 4.4) Different Trypsin TLS groups produce unique diffuse scattering. a) TLS groups for each refinement strategy projected onto the CypA structure, along with the calculated diffuse scattering for each predicted TLS motion (looking down the L axis; grey sphere denotes 2Å resolution). *Whole molecule* is shown at 25% of the intensity as *phenix* and *TLSMD*. **b)** Pearson correlation values between anisotropic maps.

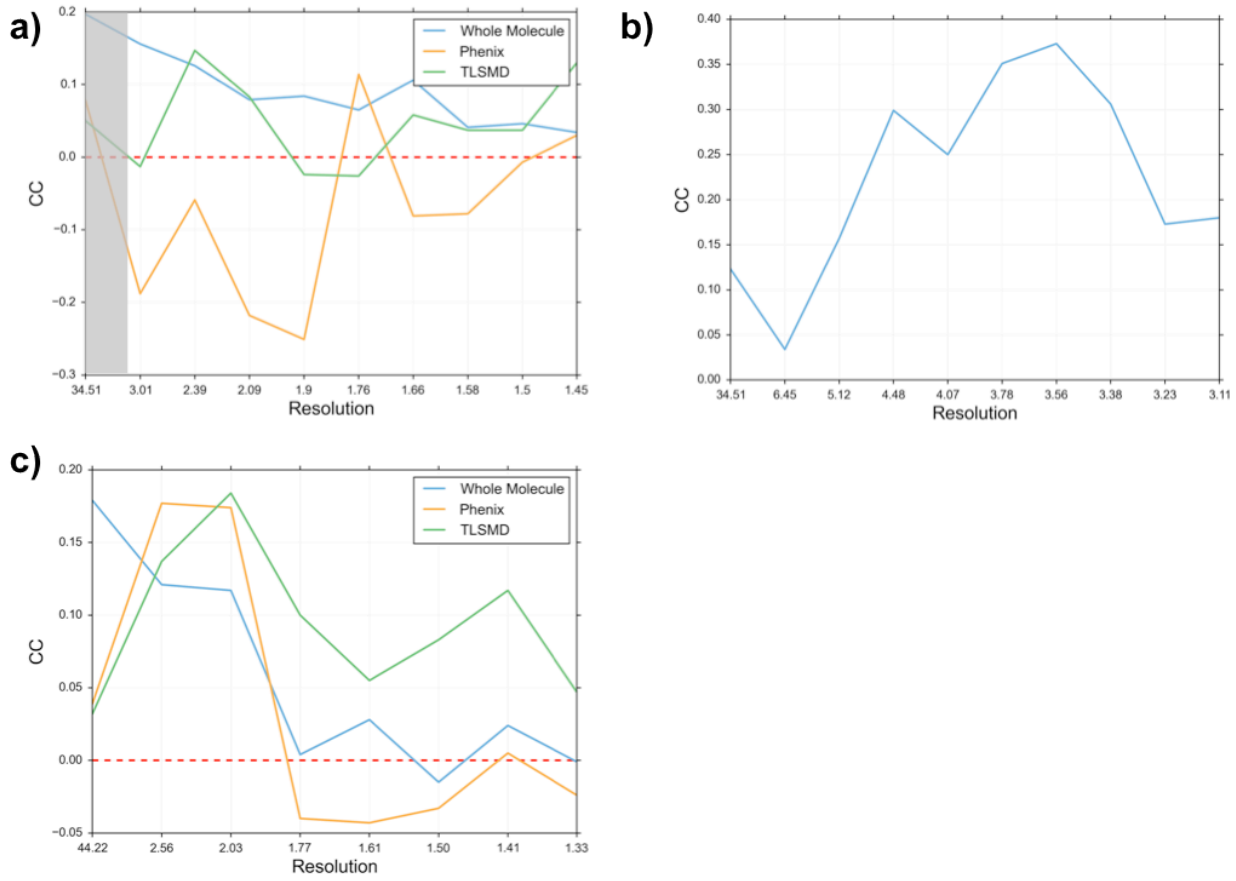


Figure 4.5) Diffuse scattering can distinguish between TLS motions. a) Pearson correlation coefficient values between CypA TLS refinements and experimental diffuse data as a function of resolution bin. **b)** PCC between *Whole Molecule* TLS refinement and experimental map in the highlighted sub-range from a) **c)** Pearson correlation coefficient values between CypA TLS refinements and experimental diffuse data as a function of resolution bin.

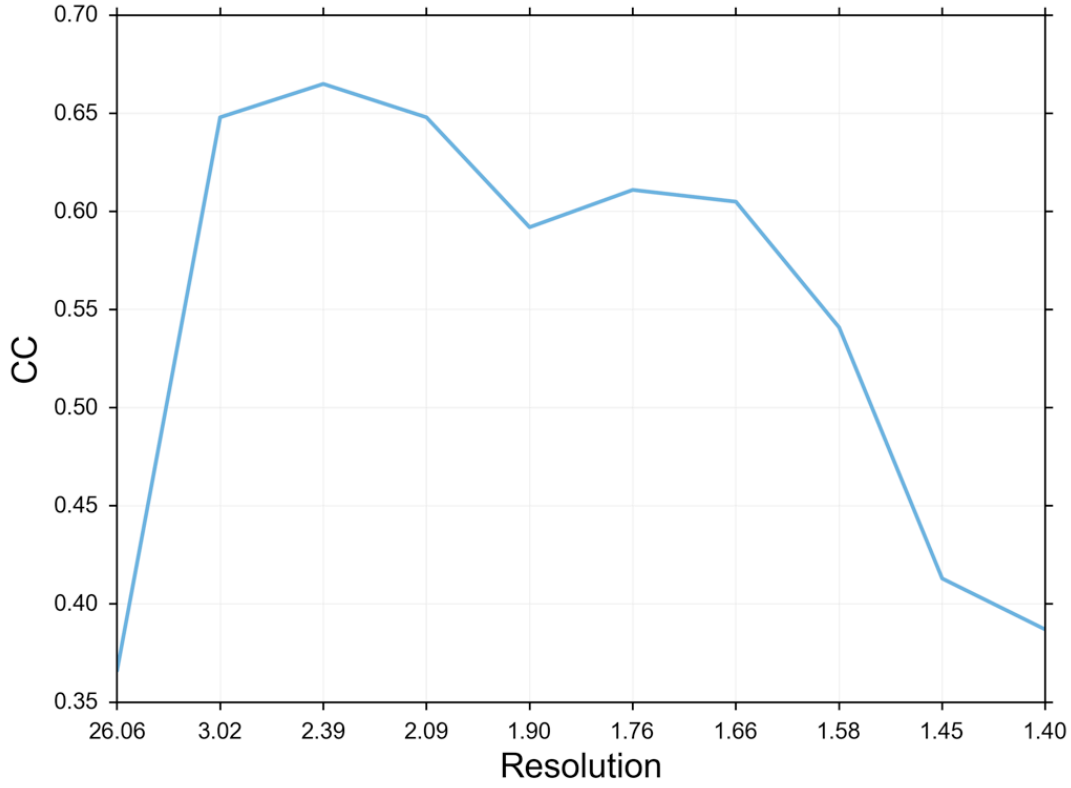


Figure 4.6) CypA Liquid-Like Motions agree with experimental data. Global correlation between CypA experimental map and LLM best-fit, where $\gamma = 6.0$A and $\sigma = 0.53$A.

4.10 References

Adams, P. D., Afonine, P. V., Bunkoczi, G., Chen, V. B., Davis, I. W., Echols, N., Headd, J. J., Hung, L. W., Kapral, G. J., Grosse-Kunstleve, R. W., McCoy, A. J., Moriarty, N. W., Oeffner, R., Read, R. J., Richardson, D. C., Richardson, J. S., Terwilliger, T. C. & Zwart, P. H. (2010). *Acta crystallographica. Section D, Biological crystallography* **66**, 213-221.

Afonine, P. V., Grosse-Kunstleve, R. W., Echols, N., Headd, J. J., Moriarty, N. W., Mustyakimov, M., Terwilliger, T. C., Urzhumtsev , A., Zwart, P. H. & Adams, P. D. (2012). *Acta Crystallographica Section D: Biological Crystallography* **68**, 352-367.

Amorós, J. L. & Amorós, M. (1968). *Molecular crystals: their transforms and diffuse scattering* (Vol. 6). New York: Wiley.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I.N. & Bourne, P. E. (2000). *Nucleic acids research*, *28*(1), 235-242.
Bernstein, F. C., Koetzle, T. F., Williams, G. J., Meyer, E. F., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). *European Journal of Biochemistry* **80**, 319-324.

Bragg, W. H., & Bragg, W.L. (1913) *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character,* 428-438

Burnley, B. T., Afonine, P. V., Adams, P. D., & Gros, P. (2012) *Elife* **1**, e00311.

Chacko, S. & Phillips, G.N. Jr. (1992) *Biophysical Journal* **61**, 1256-1266

Clarage, J.B., Clarage, M.S., Phillips, W.C., Sweet, R.M., Caspar, D.L.D (1992) *Proteins* **12**, 145-157

Clarage, J. B., Phillips, G.N. Jr (1997) *Methods in Enzymology* **277**, 407-432

Faure, P., Micu, A., Perahia, D., Doucet, J., Smith, J.C., & Benoit, J. P. (1994) *Nature* **1**, 124-128

Fraser, J. S., van den Bedem, H., Samelson, A. J., Lang, P.T., Holton, J. M., Echols, N., Alber, T. (2011) *Proceedings of the National Academy of Sciences* **108**, 16247-16252

Fraser, J. S., Clarkson, M. W., Degnan, S. C., Erion, R., Kern, D. & Alber, T. (2009). *Nature* **462**, 669-673.

Frauenfelder, H., Petsko, G.A., and Tsernoglou, D. (1979) *Nature* **280**, 558-563.

Guinier, A. (1963). *X-ray diffraction in crystals, imperfect crystals, and amorphous bodies*. Courier Dover Publications.

Jackson, C. J., Carr, P. D., Liu, J. W., Watt, S. J., Beck, J. L. & Ollis, D. L. (2007). *Journal of molecular biology* **367**, 1047-1062.

Hery, S., Genest, D. & Smith, J.C (1998) *Journal of Molecular Biology* **279**, 303-319

Keedy, D.A., Kenner, L.R., Warkentin, M., Woldeyes, R.,Thompson, M.C., Brewster, A.S., Van Benschoten, A.H., Baxter, E.L., Hopkins, J.B., … Fraser, J.S (2015) *eLife (submitted)*

Liebschner, D., Dauter, M., Brzuszkiewicz, A., & Dauter, Z. (2013). *Acta Crystallographica Section D: Biological Crystallography* **69**, 1447-1462.

Mizuguchi, K., Kidera, A. & Go, N. (1994) *Proteins* **18**, 34-48

Painter, J., & Merritt, E. A. (2006a). *Acta Crystallographica Section D: Biological Crystallography* **62**, 439-450.

Painter, J. & Merritt, E.A. (2006b) *Journal of Applied Crystallography* **39**, 109-111

Perez, J., Faure, P. & Benoit, J.P. (1996) *Acta crystallographica. Section D, Biological Crystallography* **52**, 722-729

Phillips, G.N Jr, Fillers, J.P. & Cohen, C. (1980) *Biophysical Journal* **32**, 485-502

Riccardi, D., Cui, Q. & Phillips, G. N., Jr. (2010). *Biophysical journal* **99**, 2616-2625.

Ruschak, A. M. & Kay, L. E. (2012). *Proceedings of the National Academy of Sciences of the United States of America* **109**, E3454-3462.

Schomaker, V. & Trueblood, K. N. (1968). *Acta Crystallographica Section B* **24**, 63-76.

Tickle, I. J., & Moss, D. S. (1999). *IUCr99 Computing School, London, United Kingdom* [Online *http://people.cryst.bbk.ac.uk/~tickle/iucr99/iucrcs99.html*]

Thüne, T., & Badger, J. (1995) *Progress in biophysics and molecular biology*, *63*(3), 251-276.

Van Benschoten, A.H., Afonine P., Terwiliger T.T., Wall M.E., Jackson C.J., Sauter N.K., Adams, P.D., Urzhumtsev A., Fraser J.S. (2015) *Acta Crystallographica, Section D (accepted)*

van den Bedem, H., Bhabha, G., Yang, K., Wright, P. E. & Fraser, J. S. (2013). *Nature methods* **10**, 896-902.

Wall, M. E. (1996). PhD thesis*,* Princeton University.

Wall, M.E., Clarage J.B. & Phillips, G.N. Jr (1997) *Structure* **5**, 1599-1612

Wall, M. E., Ealick, S. E. & Gruner, S. M. (1997). *Proceedings of the National Academy of Sciences of the United States of America* **94**, 6180-6184.

Wall, M. E., Adams, P. D., Fraser, J. S. & Sauter, N. K. (2014). *Structure* **22**, 182-184.

Williams, B.B., Van Benschoten, A. H., Cimermancic, P., Donia, M. S., Zimmermann, M., Taketani, M., Ishihara, A., Kashyap, P. C., Fraser, J.S. & Fischbach, M. A. (2014) *Cell Host & Microbe,* **16**, 495-503

Winter, G. (2009) *Journal of applied crystallography* **43**, 186-190.

Woldeyes, R. A., Sivak, D. A., & Fraser, J. S. (2014). *Current opinion in structural biology* **28**, 56-62.

Publishing Agreement

It is the policy of the University to encourage the distribution of all theses, dissertations, and manuscripts. Copies of all UCSF theses, dissertations, and manuscripts will be routed to the library via the Graduate Division. The library will make all theses, dissertations, and manuscripts accessible to the public and will preserve these to the best of their abilities, in perpetuity.

Please sign the following statement:

I hereby grant permission to the Graduate Division of the University of California, San Francisco to release copies of my thesis, dissertation, or manuscript to the Campus Library to provide access and preservation, in whole or in part, in perpetuity.

Andrew Van Rooijen

Author Signature

5/26/15

Date