Sensitive Detection of Structural Differences using a Statistical Framework for Comparative Crystallography

Doeke R. Hekstra[1,2]*, Harrison K. Wang[1,4], Margaret A. Klureza[1,3], Jack B. Greisman[1,5], Kevin M. Dalton[1,6,7]*

[1]Department of Molecular and Cellular Biology, [2]School of Engineering and Applied Sciences, [3]Department of Chemistry and Chemical Biology, Harvard University, Cambridge, MA 02138, USA; [4]Graduate Program in Biophysics, Harvard University, Boston, MA 02115, USA; [5]Current address: D. E. Shaw Research New York, NY 10036, USA; [6]New York University, New York, NY 10003, USA; [7]SLAC National Accelerator Laboratory, Menlo Park, CA 94025

Abstract

Chemical and conformational changes underlie the functional cycles of proteins. Comparative crystallography can reveal these changes over time, over ligands, and over chemical and physical perturbations in atomic detail. A key difficulty, however, is that the resulting observations must be placed on the same scale by correcting for experimental factors. We recently introduced a Bayesian framework for correcting (scaling) X-ray diffraction data by combining deep learning with statistical priors informed by crystallographic theory. To scale comparative crystallography data, we here combine this framework with a multivariate statistical theory of comparative crystallography. By doing so, we find strong improvements in the detection of protein dynamics, element-specific anomalous signal, and the binding of drug fragments.

* Corresponding Authors: doeke_hekstra@harvard.edu, kmdalton@slac.stanford.edu

Introduction

Proteins and their assemblies mediate chemical catalysis, molecular transport, signal transduction, and the allosteric control of these processes. X-ray crystallography is rapidly expanding experimental access to these processes, driven by the advent of new X-ray sources including fourth-generation synchrotrons[1] and X-ray free-electron lasers (XFELs)[2,3], by new time-resolved methods such as mix-and-inject studies of enzyme catalysis[4-8], and by high-throughput methods permitting screening of the interactions of thousands of molecules with proteins of interest[9-14]. These methods enable tracking of enzymatic intermediates in atomic detail, mapping of the concerted motions of proteins[10,15-17], and the identification of protein-binding drug fragments (small-molecule templates for drug design). Each of these experiments is comparative and seeks to detect subtle changes in average structure. The success of such experiments hinges not only on billion-dollar instruments, but also on the sensitivity of the algorithms used to detect changes in structure from the X-ray diffraction data.

X-ray diffraction images obtained from macromolecular crystals consist of patterns of many small spots, or reflections. The true intensities of these reflections are proportional to the square amplitudes of the "structure factors"—the Fourier components of the electron density in the crystal. The observed intensities, however, also depend on a range of multiplicative factors ("scales") due to the intensity and polarization of the incident beam, the volume of crystal exposed to the beam, intrinsic and beam-induced crystal defects, and absorption of the beam by surrounding material, air, and the detector itself. Neither the true reflection intensities nor these scales are directly observed, yet it is critical that before inferring changes in structure from changes in observed intensities, these intensities must be corrected. The correction procedure is known as scaling.

Approaches to scaling rely on comparison of redundant observation of equivalent reflections which should yield the same structure factor amplitude up to variation in the scale factor. When sample or beamtime are limited, the need for scaling to compare different conditions generates a tradeoff: one can either accurately estimate scales by maximizing redundancy for a few conditions, or observe many conditions but be left with residual systematic

differences in the scales of each dataset. These remaining scaling inaccuracies are often addressed after the fact, for example by applying local scaling in SOLVE[18], SCALEIT[19], or the "Isomorphous Difference Map" utility of PHENIX (e.g., in [20,21]).

Here we introduce a framework that enables statistically efficient comparison of related datasets. We first observe that the structures of proteins at different timepoints or under different chemical or physical conditions are often very similar, and therefore could provide nearly redundant measurements. To exploit this similarity, we adapted an algorithm we recently introduced for scaling and merging of X-ray diffraction data[22]. This algorithm, implemented in the software package *Careless*, applies to diffraction data from all predominant experimental approaches. *Careless* is based on approximate Bayesian inference combined with a neural network that learns scale factors from experimental metadata such as the position of each diffraction spot on the detector, Miller indices, and frame or crystal number. To adapt this approach to comparative applications, we developed a multivariate prior distribution that accurately captures correlations between related crystallographic datasets, and extend the *Careless* formalism to use such prior information. We show that this approach leads to strong improvements in the detection of signals in several examples—detection of protein dynamics from a polychromatic, time-resolved experiment, detection of specific elements in enzymes from serial XFEL and polychromatic synchrotron data, and the detection of bound drug fragments from a crystallographic screen. In forthcoming work[23,24], we further show that this approach makes it possible to follow enzyme catalysis in a serial, rapid-mixing experiment.

A statistical framework for comparative crystallography

Our statistical model is based on three concepts and explained in mathematical detail in the **Supplementary Information**: **First**, each structure factor, or Fourier component of the electron density, can be thought of as the sum of contributions from all the constituent atoms[25]. The sum of these contributions is well-approximated by the central limit theorem. The structure factors, which have real and imaginary components, therefore follow a bivariate normal distribution in the complex plane centered at the origin (**Figure 1a**, black arrows). The structure factor amplitudes, which correspond to the distance from the origin in **Figure 1a** (dashed circle), follow the well-known Wilson distribution[26]. **Second**, when comparing two related datasets, the

contributions of corresponding atoms in the two crystals will be similar and therefore the sums of their contributions to their respective structure factors will be similar. When considered together, they will closely follow a multivariate normal distribution characterized by a correlation parameter, henceforth the *double-Wilson r*[27,28], that quantifies the correlation between the real components of pairs of structure factors. Importantly, the joint distribution of complex structure factors F_1, F_2 can be factorized such that $P(F_1, F_2) = P(F_1|F_2) \cdot P(F_2)$ with $P(F_1|F_2)$ only dependent on the difference in phases of F_1 and F_2. Integrating this unobserved phase difference yields a conditional probability density $P(F_1|F_2)$ that follows the Rice distribution[28-30] (where $F = |F|$). We will refer to this statistical model as the double (or bivariate) Wilson model, and to the resulting joint distribution of structure factor amplitudes as the bivariate Wilson distribution.

Figure 1b compares structure factor amplitudes for crystals of human Protein Tyrosine Phosphatase 1B (PTP-1B) in the presence and absence (apo) of the inhibitor TCS-401[31]. A fair degree of correlation is evident. A fit to the double-Wilson model (**Supplementary Notebook 3**) shows that $r \approx 0.85$, with the implied Pearson correlation coefficient between structure factor amplitudes $\approx r^2$ (**Supplementary Notebook 2a**)[32]. Indeed, histograms of structure factor amplitudes for the inhibitor-bound data (**Figure 1c-e and Figure S1**) given the binned amplitudes of the apo data (blue shading in **Figure 1b**) are better fit by this model (red curves in **Figure 1c-e**) than by assuming statistical independence (blue curves). We likewise find that the bivariate Wilson distribution describes other comparative crystallography datasets well, including time-resolved X-ray crystallography experiments (**Figure S2**), replicate data sets across labs (**Figure S3**), and across temperatures (**Figure S4**), providing confidence that the formalism generally applies to related datasets.

Third, this approach does not generalize to more than two datasets, as it is not known how to integrate over the unknown phase differences. However, for many crystallographic experiments, it is reasonable to assume some conditional independence. For instance, when structure factor amplitudes A and C are both correlated with structure factor amplitudes B, their joint probability is often well described as the product $P(A, B, C) \approx P(A|B) \cdot P(C|B) \cdot P(B)$ rather than the exact expression $P(A, B, C) = P(A|B, C) \cdot P(C|B) \cdot P(B)$, as long as C

does not provide additional information about A not contained in B. Assumptions about conditional (in)dependence can be summarized in a Bayesian network or graphical model[33] (**Figure 2a**). In this representation, each node represents a dataset, and each edge (arrow) represents conditional dependence. Whenever the resulting graph is acyclic, the resulting joint probability of structure factor amplitudes can be calculated analytically (**Supplementary Information**). We will refer to this distribution as the multivariate Wilson distribution. We note that conditional dependence is reciprocal, and the orientation of the displayed arrows is irrelevant for our purposes. Indeed, the expected correlations in many comparative crystallography experiments can be approximated by such networks (**Figure 2b**). For example, for a crystallographic time series, when estimating structure factors at time 3, it is a good assumption that we would not gain anything from knowing the structure factor amplitudes at time 1 if we already have access to their values at time 2. We further note that not all nodes need to have been observed—conditional independence mediated by a 'dummy' has found application in the analysis of multiple isomorphous replacement[34].

Scaling of related datasets

We recently introduced a Bayesian statistical framework for simultaneous inference of scales and structure factor amplitudes from reflection metadata (such as Miller indices, image number, position on the detector, and X-ray wavelength) and observed intensities[22]. Until now, estimation of each structure factor amplitude has been independently constrained by a Wilson prior distribution on the amplitudes. For comparative applications, this approach is inefficient: during scaling, more accurate scales enable more accurate estimates of structure factor amplitudes, and vice versa. It would be natural to incorporate a joint prior distribution on structure factor amplitudes during scaling that reflects the expected correlations between conditions. Different scenarios for a pair of related structure factor amplitudes F_1, F_2 illustrate the anticipated benefits of a multivariate approach: (i) if only noisy observations are available for F_1 and F_2, the prior should force estimates of F_1 and F_2 to be highly similar and, therefore, their difference small; (ii) when F_1 can be estimated accurately, the anticipated similarity of the structure factor amplitudes will facilitate more accurate estimation of the scales for the other

dataset by narrowing down the range of probable values of F_2; (iii) when accurate observed intensities are available for both datasets, the anticipated similarity helps calibrate the relative scales in both datasets. With these advantages in mind, we implemented the multivariate Wilson distribution as a prior in Careless and show in four examples that structure factor differences can, indeed, be much better estimated upon scaling amplitudes with a multivariate prior. Operationally, our implementation expects users to supply the topology of conditional dependencies between datasets (**Figure 2b**) and an estimate of the double-Wilson parameter r.

Laue anomalous diffraction

Although many X-ray sources are intrinsically polychromatic, most of their photons are typically discarded by use of monochromators. For studies of dynamics, however, it can be essential to use much brighter polychromatic pulses, giving rise to Laue diffraction. To assess whether the use of multivariate priors can improve the processing of Laue data, we first addressed the extraction of anomalous signal, which can be recovered from small differences between Bijvoet pairs—pairs of structure factors that are identical in the absence of element-specific electronic resonances (that is, anomalous effects) but that differ when such effects are present.

In this example, all data were collected on a single crystal of hen egg white lysozyme soaked with sodium iodide using polychromatic X-ray pulses (about 5% spectral bandwidth, from 1.02 to 1.20 Å wavelength; see **Methods**). As these data were obtained from a single crystal, one might expect little inconsistency in artifacts across the data and minimal benefit from imposing a bivariate prior on Bijvoet pairs. Nevertheless, we compared scaling and merging in Careless with a univariate Bayesian prior (i.e., the standard Wilson distribution) against a bivariate prior that assumes varying levels of correlation, r, between the Bijvoet mates. For consistency, we phased all sets of structure factor amplitudes using a reference model refined against monochromatic data collected on the same day (see **Methods**). In the resulting anomalous difference map, we find strong gains in peak heights of both iodide ions and sulfur atoms when using a bivariate prior (**Figure 3a, Figure S5a**), even though the anomalous scattering strength of sulfur is only about 1/4th the scattering from a single electron (f'' = 0.26 at the dominant wavelength of the incident X-ray spectrum, 1.04 Å). Importantly, the signal strength is maximal at an intermediate r of 0.999, with an increase in anomalous peak height of about 50% over the

use of a univariate prior (**Figure 3b, Figure S5b**). Cross-validation measures of data processing quality ($CC_{1/2}$[35], CC_{pred} [36], CC_{anom}[37]) are also improved by the introduction of the bivariate prior (**Figure 3c-d**).

Time-resolved Laue diffraction

Laue diffraction at synchrotron sources provides access to the dynamics of proteins on timescales of 100 picoseconds and longer. To determine whether a bivariate prior can improve scaling of time-resolved signal, we processed a time-resolved Laue dataset collected on a single crystal of photoactive yellow protein (PYP) at BioCARS 14-ID at the Advanced Photon Source. The light-induced *trans*-to-*cis* isomerization of the active site chromophore (**Figure 4a**) has been studied extensively[38-40]. Our dataset contains 20 images of PYP without laser exposure and 20 images 2 milliseconds after exposure to a blue laser pulse, at which point PYP is expected to adopt the pB1 state[40]. Using Careless, we jointly scaled the OFF and 2ms datasets while imposing a joint prior with varying levels of the correlation parameter, *r*. Merging statistics improve when scaling with a bivariate prior (**Figure 4b**). One such statistic is the correlation coefficient, CC_{pred}, between observed intensities and those implied by the scaling model and the estimated structure factor amplitudes[22]. The gap between the CC_{pred} for training data and a cross-validation test set is reduced with increasing *r*. This means that the bivariate prior protects the scaling model against overfitting. The extensive literature on PYP enables us to quantify the quality of the resulting electron density difference maps: based on models of the ground-state and excited-state conformations from previous studies, we can calculate real-space correlation coefficients with a predicted difference map (that is, the correlation with an $F_c^{2ms} - F_c^{dark}$ map based on conformations from Protein Data Bank (PDB) entry 1TS0, see ref. [40]). At the best value of *r* (about 0.999), the real-space CC (Correlation Coefficient) of the observed difference map is more than 0.50 within 10 Å of the chromophore, about twice as strong as for an observed difference map obtained by scaling with a univariate prior (**Figure 4c**). Consistent with this quantitative assessment, the difference map near the chromophore is more readily visually interpretable after scaling with a bivariate prior (**Figure 4d**; see also **Figure S6**). We also attempted to place ON and OFF datasets on the same scale using local scaling or SCALEIT[41] (**Methods**). In our hands, neither approach was effective (**Figure 4c**, on the left).

It is instructive to compare our scaling approach with commonly used[16,22,42,43] weights that are heuristic modifications of a weighting scheme proposed in ref[44], inspired by Bayesian statistics (**Methods**). These weights reduce the contributions of estimated structure factor amplitude differences (ΔF) that are large or have large estimated errors ($\sigma(\Delta F)$). A coefficient α, used in calculating the weights, tunes the contribution of large ΔF to the difference map (**Methods**). In the case of the PYP data (**Figure 4c**) we find that suppression of large ΔF ($\alpha > 0$) is necessary to achieve a similar effect as the bivariate prior. That is, the bivariate prior suppresses spuriously large contributions to the difference map without the need for heuristic weighting schemes.

Serial Femtosecond Crystallography

Serial femtosecond X-ray crystallography (SFX) of microcrystals probed by intense x-ray free-electron laser (XFEL) pulses has dramatically expanded the range of systems and timescales accessible to time-resolved studies. To determine whether multivariate priors could also improve the accuracy of SFX data processing, we sought to extract anomalous signal from an SFX dataset collected for thermolysin, an enzyme containing a catalytic zinc (Zn) ion in its active site. Specifically, we processed 3,160 images from a much larger dataset (CXIDB 81[45,46]). To account for the serial monochromatic data collection strategy, we included an Ewald offset and per-image layers in our scaling model, as before[22]. We split Bijvoet pairs into two half-datasets before relating these half-datasets by a joint prior during scaling in Careless. The anomalous difference map peak height of the catalytic zinc ion (ZN317) increases from 15σ without a bivariate prior to 28σ for the optimal *r* value (**Figure 5a**). Additionally, as *r* is varied, more associated calcium ions become detectable. Using a univariate prior, only two of four calcium ions appear above 3σ (an approximate noise threshold). After scaling with a multivariate prior, all four previously modeled calcium ions are detectable (**Figure 5a-b**). In prior analyses of a thermolysin dataset with four times as many images, anomalous difference signal has previously been detected at a site adjacent to the bound ZN317[45,47]. When scaling our smaller dataset with a uniform prior, this site was not detectable[22]. Upon scaling with a multivariate prior, however, this site can be clearly identified (**Figure 5b, black arrow**). We note that, again, these anomalous difference maps can also be improved by weighting (**Figure S7**), but that this is

substantially less effective than the use of a multivariate prior during scaling. In summary, a bivariate prior strongly improves the detection of anomalous signal from a small SFX dataset.

Fragment screening

Finally, crystallographic drug fragment screens now enable the rapid discovery of pharmacological lead compounds against proteins of interest, e.g., SARS CoV-2 proteins[9,11,13,48]. In these studies, thousands of crystals of a protein of interest are each soaked with a "drug fragment"—a small-molecule building block for larger drug molecules. To determine if joint scaling can improve the detection of bound fragments in difference maps, we analyzed data from a fragment screening experiment totaling one *apo* dataset and sixteen *holo* datasets identified by PanDDA[49]. This dataset contains structures of SARS-CoV2 nonstructural protein 3 macromolecular domain 1 (Nsp3 Mac1) with ligands from a fragment screening library bound to the Mac1 adenosine-binding site. We ran *Careless* on unmerged intensities, adding several metadata keys including one-hot encodings of each dataset (**Methods**). By one-hot encoding, we allow the model to express differences in scale between fragment-bound datasets. Additionally, we set each *holo* dataset to depend on the *apo* dataset through a joint prior, while assuming conditional independence between *holo* datasets (**Figure 2b**). We find that difference density maps visually improve upon scaling with a multivariate prior (**Figure 6a, Figure S8**). We tracked the largest ligand peaks in the difference map for a given dataset while we varied r. We find that in all datasets, this signal is strongest when scaling with a multivariate prior (**Figure 6b**). The r value for the tallest peak differs for each *holo* dataset. Without special modifications to scaling and merging, we obtain peak heights on par with those produced by PanDDA, which relies on real-space post-processing (**Table S2**). To determine if the *apo-holo* dataset structure factor correlation, measured by r, relates to ligand occupancy, we examined the relationship between optimal r and the PanDDA background density correction factor (BDC, a parameter which scales inversely with ligand occupancy and accounts for crystal-to-crystal variation in soaking and data quality)[12]. We find a moderate correlation with r (Pearson $r = 0.505$, $p < 0.05$) (**Figure S9a**). Merging statistics also improve when scaling with a bivariate prior (**Figure S9b**): in particular, the gap between the training and test CC_{pred} is reduced for higher r, indicating that the bivariate prior reduces overfitting of the scales.

9

This example shows that the difference map quality improves when scaling high-resolution monochromatic data with a bivariate prior. It would be natural to expect that the benefits of a multivariate prior during scaling would keep increasing with the number of datasets included. At present, this is not the case, likely because we do not account for crystal-to-crystal variation in the average unit cell.

Discussion

Bayesian statistics have been applied extensively in macromolecular crystallography, playing an important role in solutions to the phase problem[28] and maximum-likelihood-based refinement[50]. Here, we establish that a Bayesian formalism enables the sensitive detection of subtle chemical and time-dependent signals and ligand binding from comparative crystallographic measurements by introducing a multivariate prior describing the anticipated correlations among related datasets. Although this formalism proves effective even in the simple form presented here, there are several promising extensions to further improve the detection of structural change. For example, high-quality reference datasets are commonly measured for serial time-resolved crystallography experiments, but these data are not commonly used to improve scaling. Such datasets could be readily added to the Bayesian networks shown in **Figure 2** to better condition scaling. Second, we currently treat the correlation parameter r as a fixed, global hyper-parameter, yet correlations may vary between pairs of datasets (e.g., as a function of ligand occupancy) and vary with resolution. These correlations could be estimated from the data. Third, as not every node in **Figure 2** needs to be observed, our approach allows for interpolation of missing observations.

The effects of perturbations in comparative crystallography, whether due to laser excitation[51], temperature[15,52], electric field[16], or small-molecule binding[11], often only affect a fraction of the molecules in the crystal. As an important consequence, the resulting mixed structure factor amplitudes tend to be slightly smaller than the ground-state (OFF) amplitudes[53], and the resulting ON-OFF difference maps tend to resemble a negative image of the OFF electron density. It is common practice to rescale scaled data to place OFF and ON structure factor amplitudes on a common scale (as we do, implicitly, in the present work). In

the developed statistical framework, correlation of an isomorphous difference map with the OFF electron density, the true ON-OFF difference map, and the map error can all be calculated exactly (**Supplementary Notebook 6**), supporting the common observation that correlation with the OFF map can be removed with minimal effect on difference map quality (**Figure S12**).

Finally, we observe that the developed framework also implies a joint distribution of the unobserved crystallographic phases across related datasets (**SI** and **Figure S13**). As the traditional crystallographic data processing paradigm begins to give way to machine learning methods, we anticipate that these implied correlations can play a key role in learning crystallographic movies directly and efficiently from time-resolved diffraction data.

Figures
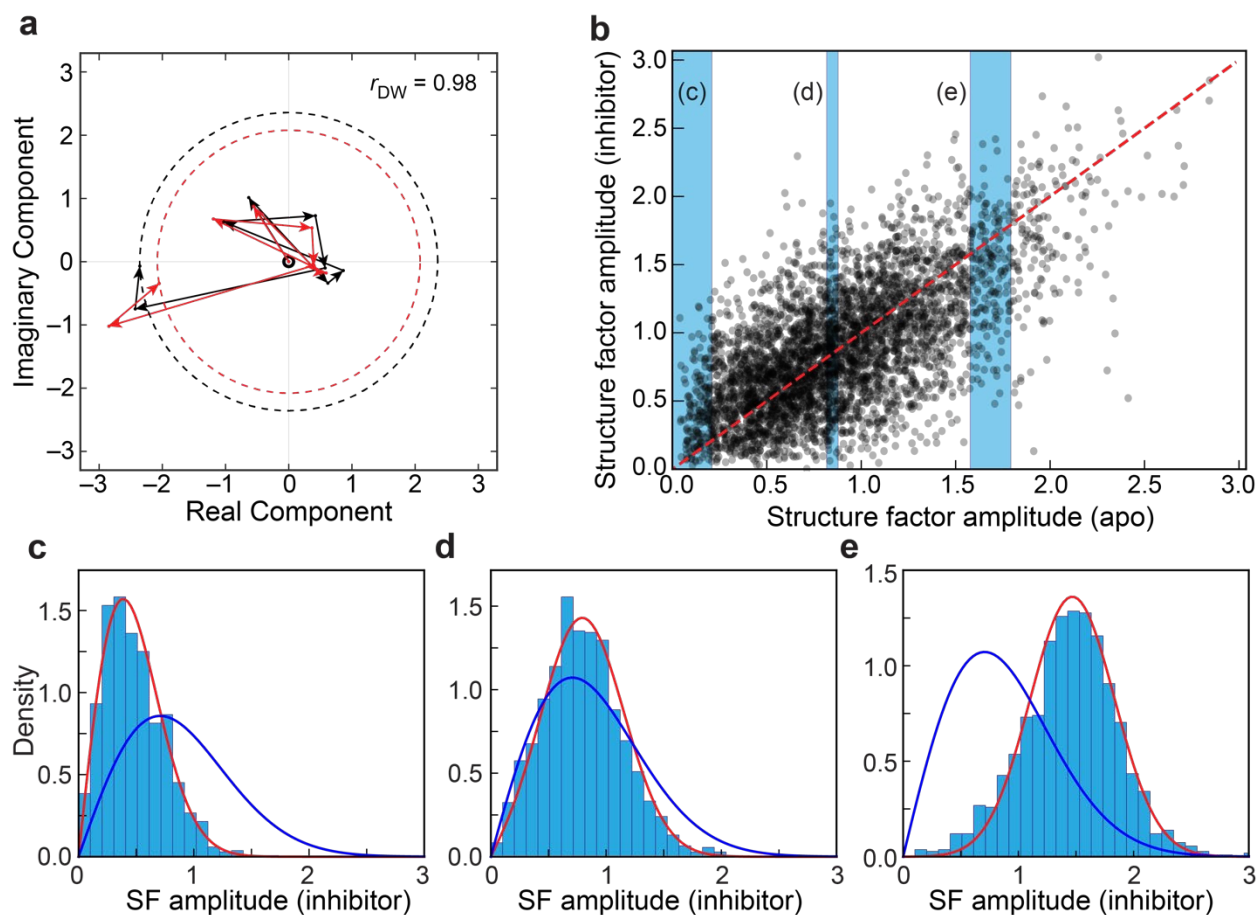
Figure 1: The double-Wilson model. a) In the double-Wilson model, the structure factors of related structures (black and red) are built up from pairs of random contributions (arrows) with correlation r_{DW} between their real components and likewise between their imaginary components. The experimentally observed amplitudes of the structure factors are described by the radii of the dashed circles. **b)** Scatter plot for structure factor amplitudes for isomorphous PTP-1B datasets for *apo* and inhibitor-bound forms (for a random subset of Miller indices). Blue slices indicate data points for which histograms are shown in panels c-e. **c-e)** Histograms for slices through panel b are better approximated by the Rice distributions (red) parametrized by a global r_{DW} (here 0.85) than by the unconditional, univariate Wilson distribution (blue).
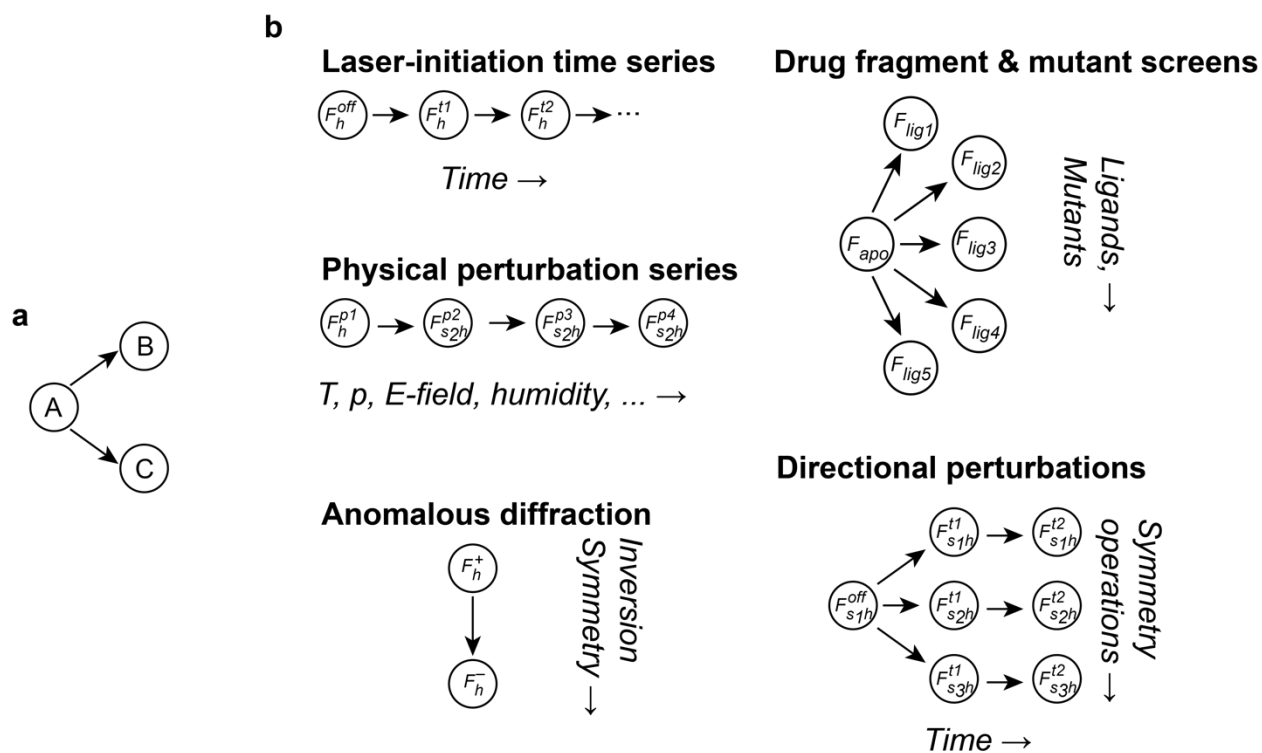
Figure 2. Modeling correlations of structure factors in comparative crystallography experiments. a. A graphical model (or Bayesian network) illustrating the relationship between three datasets (or, more generally, sets of structure factor amplitudes). Arrows indicate conditional dependence; the lack of an arrow between datasets B and C indicates conditional independence. **b.** Correlations in many experiments are well approximated by acyclic graphical models for which the joint prior distribution of structure factor amplitudes F_h can be calculated analytically. $t_1, t_2, \ldots$ represent timepoints; $s_1, s_2, \ldots$ represent symmetry operations; $p_1, p_2, \ldots$ represent perturbations due to, e.g., temperature T, pressure p, or electric field; F^+ and F^- are reflections related by Friedel's law.
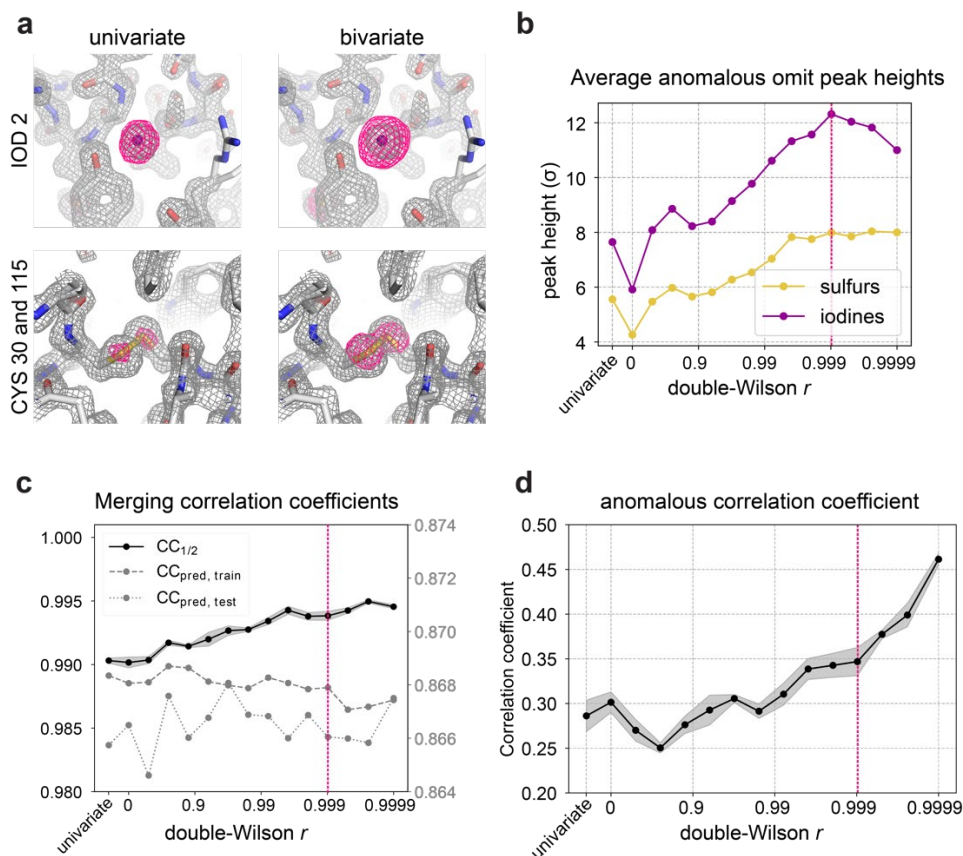
Figure 3. Use of a bivariate prior improves anomalous signal in a Laue diffraction experiment. a) Comparisons between anomalous omit peaks after scaling and merging with univariate or bivariate priors ($r = 0.999$ for the bivariate prior, marked in panels **b-d** with a dashed magenta line). **b)** Dependence of average iodine and sulfur anomalous peak height on r. **c)** Merging correlation coefficients of the lysozyme dataset across double-Wilson r values. The y-axis labels for the $CC_{1/2}$ are on the left, and the y-axis labels for the CC_{pred} are on the right. A test set of 10% of observations were held out during scaling and merging to evaluate performance of the scaling model, yielding $CC_{pred, test}$ for the test set, and $CC_{pred, train}$ for the 90% of data used during scaling. The shaded confidence interval of the $CC_{1/2}$ curve represents the standard deviation over three half-dataset repeats. **d)** Anomalous correlation coefficient of the lysozyme dataset across r. The shaded confidence interval represents the standard deviation over three half-dataset repeats.
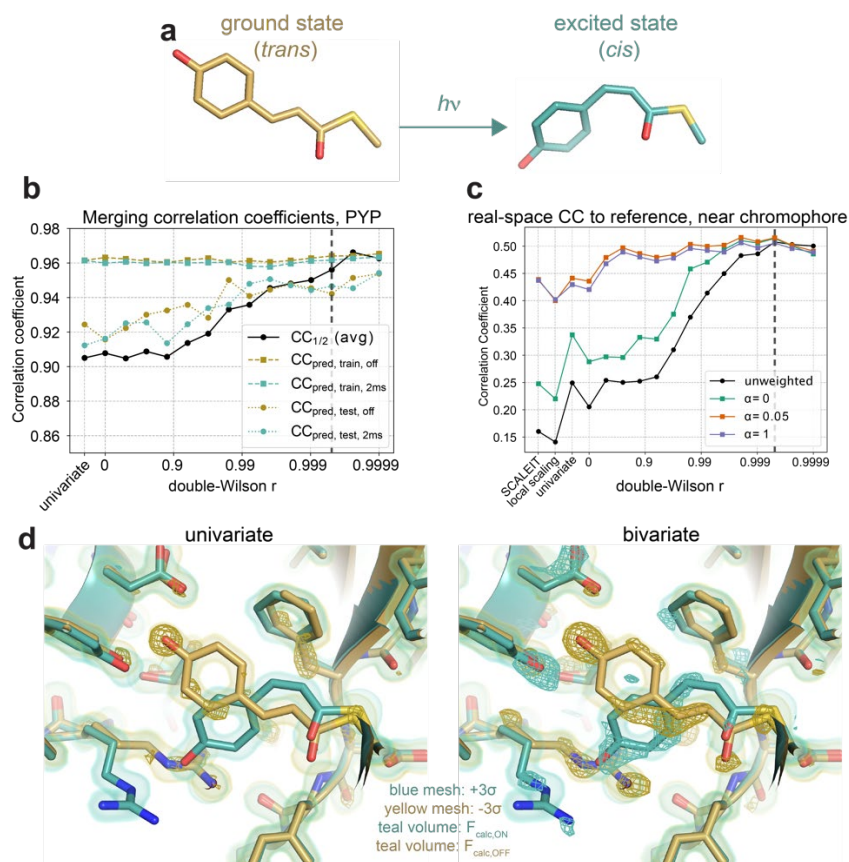
Figure 4: **Use of a bivariate prior improves time-resolved difference signal. a)** The *para*-coumaric acid chromophore of PYP isomerizes when exposed to blue light. **b)** Merging statistics of the 2ms and OFF structure factor amplitudes as a function of *r*. **c)** Correlation coefficient of the weighted (α=0, green; α=0.05, red; α=1, blue) and unweighted (black) observed difference maps with the 2ms−OFF (dark) calculated difference map ($F_{C,2ms} - F_{C,OFF}$), in a region 10 Å around the chromophore. α weights the contributions of large structure factor differences (see **Methods**). The correlation coefficient is plotted for the univariate prior, with and without applying local scaling or SCALEIT[19], and as a function of the double-Wilson parameter *r*. Vertical dashed black line in panels **b-d** indicate the *r* value, 0.9995, that parametrizes the bivariate prior for merged data displayed in the following panel. **d) left:** Observed difference map (teal and yellow mesh, contoured to 3σ, unweighted), from merging with a univariate prior. The map is overlaid with models (2ms and dark models colored with teal and yellow sticks, respectively) and the electron density at 2ms after excitation and in the dark state (teal and yellow surfaces, respectively; contoured to 1.5σ), based on calculated structure factors F_{2ms} and F_{off} from PDB entry 1TSO. **Right:** Observed 2ms−OFF difference map (yellow and teal mesh, contoured to 3σ, unweighted), from merging with a bivariate prior, *r* = 0.9995. F_{2ms} map and F_{OFF} map as in left panel.
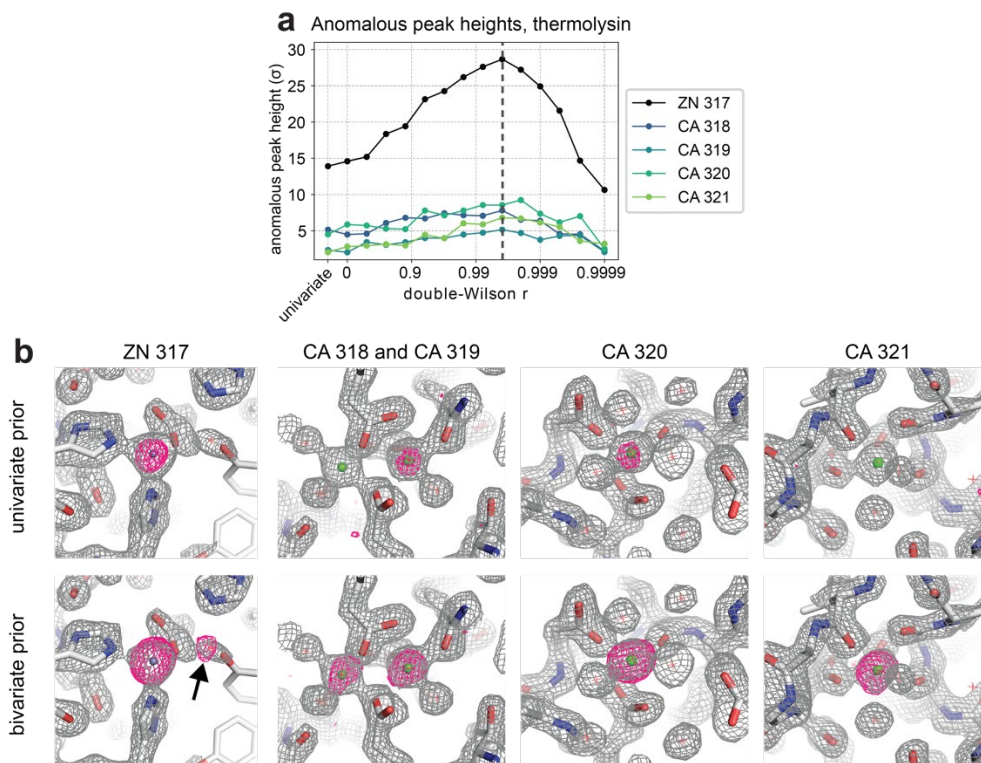
Figure 5: **Scaling and merging with a bivariate prior improves the anomalous signal in a serial XFEL experiment. a)** Anomalous difference map peak heights in thermolysin, shown for zinc and four calcium sites, as a function of the double-Wilson parameter r. Vertical dashed black line in panel **a** indicates the r value, 0.9961, that parametrizes the bivariate prior for merged data displayed in the following panel. **b) first column:** Region near the thermolysin-bound zinc atom, with observed electron density map ($2mF_o$-DF_c) in gray, contoured at 2σ, and anomalous difference omit map in magenta, contoured at 5σ. Arrow indicates the location of an alternative zinc binding site revealed by merging with the bivariate prior. **Next three columns:** Anomalous difference omit maps of thermolysin calcium sites, carved to 1.5 Å of the model and contoured to 4σ (magenta).
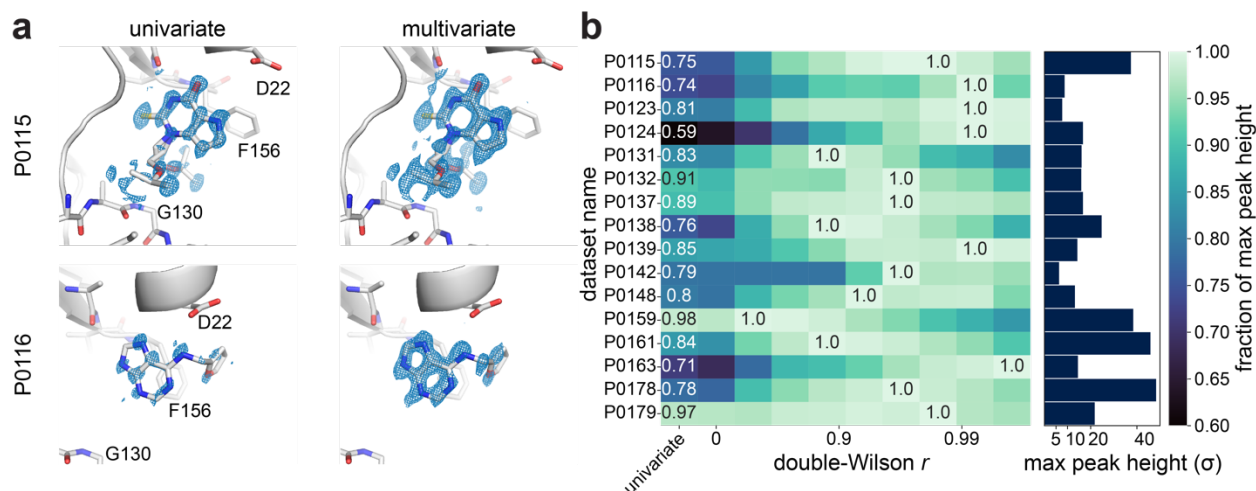
16

Figure 6. Use of a multivariate prior improves bound fragment signal in a drug fragment screen. a) *holo − apo* isomorphous difference maps from fragment screening of Mac1, scaled with a univariate or multivariate prior. All maps carved to 1.5 Å of the ligand and contoured to +3σ. **b)** Tallest positive peak in each *holo − apo* difference map as a function of the double-Wilson *r* parameter. For a given *holo* dataset, the tallest positive peak is measured. The maximum peak height of that peak, across all double-Wilson *r* values, is normalized to 1.0. The maximum peak height is labeled and reported in a bar chart (**right**). Also labeled is the fraction of the maximum peak height for the tallest peak in the univariate datasets (**left**). The r value, for each dataset's maximum peak height, parametrizes the bivariate prior for merging the electron density displayed in panel **a**.

17

Methods

Implementation details. The multivariate Wilson distribution we present here is implemented as a prior distribution in the Careless software package. Careless uses gradient-based optimization to maximize a Bayesian objective function[54,55] which is the sum of the expected log likelihood of the observed data plus the Kullback-Leibler divergence between the structure factors and the prior distribution. The likelihood of the data describes the probability of the observed diffraction intensities, $I_{h,i}$, conditional on scale $\Sigma_{h,i}$, and structure factor amplitudes F_h. The likelihood also takes into account the estimated errors $\sigma_{I_{h,i}}$ in the observed intensities. The objective function is computed as described in equation 21 and 22 of Dalton et al.[22] using reparameterized samples[56] from a truncated normal surrogate posterior (variational distribution). For clarity, we reproduce this objective function,

$$\text{ELBO}\big(\mu_{q_F}, \sigma_{q_F}, \theta\big) \approx \sum_{s=1}^{S} \left[\sum_i \sum_h \log p\big(I_{h,i} | F_{h,s}^2 \Sigma_{h,i,s}, \sigma_{I_{h,i}}\big) - \sum_h \big(\log q_{F_h}(F_{h,s}) - \log p(F_{h,s}) \big) \right].$$

The Evidence Lower BOund (ELBO) depends on the parameters of the surrogate posterior distributions, q_F and q_Σ. q_Σ is parameterized by a neural network with parameters θ. q_F is parameterized by independent truncated normal distributions for each reflection with location and scale parameters $\{\mu_{q_{F_h}}, \sigma_{q_{F_h}}\}$ supported on $[0, \infty)$ for centric reflections and $(0, \infty)$ for acentrics. $I_{h,i}$ refers to the intensity of a particular observation i for reflection h. s indexes reparameterized samples from the surrogate posteriors,

$$\begin{aligned} F_{h,s} &\sim q_{F_h} \\ \Sigma_{h,i,s} &\sim q_{\Sigma_{h,i}} \end{aligned}$$

Reparameterization is implemented in the TensorFlow Probability library[57]. The prior distribution enters the objective function through the ultimate term, $\log p(F_{h,s})$. Previously,[22] we exclusively used a univariate Wilson distribution as a prior for this term,

$$p(F_{h,s}) = \begin{cases} \sqrt{\dfrac{2}{\pi \epsilon_h}} \exp\left(-\dfrac{F_{h,s}^2}{2\epsilon_h}\right) & h \text{ is centric} \\[4mm] \dfrac{2}{\epsilon} F_{h,s} \exp\left(-\dfrac{F_{h,s}^2}{\epsilon_h}\right) & h \text{ is acentric,} \end{cases}$$

where ϵ_h is the multiplicity of the reflection h, an integer value determined by the crystal's symmetry. Note that because the scale function, Σ_h, is parametrized separately within Careless, it does not appear in expressions for $p(F_{h,s})$.

One of the features of variational inference is its flexibility. In this work we leverage the inherent flexibility of reparameterization-based variational inference to extend the Careless package with a multivariate prior distribution. To do so, the Careless command line interface requires users to supply separate unmerged data sets for each node they wish to model (as in **Figure 2**). Careless asks the user to specify the parent of each node and the expected double-Wilson parameter r. Internally, this changes how the prior is calculated as follows,

$$P(F_{h,s}) = \begin{cases} \text{FoldedNormal}\left(F_{h,s} | rF_{Pa(h),s}, \sqrt{\epsilon_h(1-r^2)}\right) & h \text{ is centric} \\ \text{Rice}\left(F_{h,s} | rF_{Pa(h),s}, \sqrt{\dfrac{\epsilon_h}{2}(1-r^2)}\right) & h \text{ is acentric,} \end{cases}$$

where $Pa(h)$ denotes the parent reflection of h. Where a particular reflection has no parent because it is the root node, the resolution range of parent and child differ, or the parent is systematically absent, the prior defaults to the univariate Wilson model. The folded normal distribution's probability density function is

$$\begin{aligned} \text{FoldedNormal}(x|\mu,\sigma) &= \text{Normal}(x|\mu,\sigma) + \text{Normal}(x|-\mu,\sigma) \\ &= \frac{1}{\sigma\sqrt{2\pi}}\left\{\exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] + \exp\left[-\frac{1}{2}\left(\frac{x+\mu}{\sigma}\right)^2\right]\right\}, \end{aligned}$$

while that of the Rice distribution is

$$\text{Rice}(x|\nu,\sigma) = \frac{x}{\sigma^2}\exp\left(-\frac{(x^2+\nu^2)}{2\sigma^2}\right)I_0\left(\frac{x\nu}{\sigma^2}\right),$$

where I_0 denotes the modified Bessel function of the first kind with order zero. For numerical stability, both probability densities are computed directly in log space.

PYP. Collection of the Laue diffraction data for photoactive yellow protein was described previously[58]. Indexing, geometry refinement, wavelength assignment, and integration were performed in Precognition v5.2.2. Precognition integrated intensities were written to mtz files and then scaled and merged with Careless v0.4.1, with use of a bivariate prior on the OFF and 2ms timepoint mtz files. Fifteen different double-Wilson *r* values were used, with *r* varied for each Careless run as $1 -$

$0.5^{\text{run number}}$. Correlation coefficients were computed using *careless.cchalf*, *careless.ccpred*, and custom scripts based on *gemmi*[59] and *reciprocalspaceship*[60]. Isomorphous difference maps were first phased with an OFF model refined in Phenix 1.20.1[61] from PDB ID 2PHY[62], then weighted using *reciprocalspaceship* according to

$$w_{hkl} = \left(1 + \frac{\sigma^2(\Delta F)}{\langle \sigma^2(\Delta F) \rangle} + \alpha \frac{|\Delta F|^2}{\langle |\Delta F|^2 \rangle}\right)^{-1}$$

for structure factor amplitude differences ΔF (short for ΔF_{hkl}) and their errors $\sigma(\Delta F)$, with angled brackets indicating averages over Miller indices. Local scaling was performed using a custom script available at (https://github.com/Hekstra-Lab/dw) and SCALEIT (CCP4 software suite version 7.1) was run from *rs-booster* version 0.1.2 (https://github.com/rs-station/rs-booster) as *rs.scaleit* (**Supplementary Information, "Partial Excitation"**).

Difference maps were visualized in *PyMOL*[63]. Data processing scripts can be found in the Zenodo deposition.

Lysozyme data collection: NaI-soaked lysozyme crystals were prepared as described previously[64]. Monochromatic data were collected on the Northeastern Collaborative Access Team (NE-CAT) beamline 24-ID-C (Advanced Photon Source, Argonne National Laboratory) on 12 August 2020. Diffraction data were collected at ambient temperature (about 295 K). Three 1440-image passes (720°) were collected from a single crystal of lysozyme with an exposure time of 0.1 s and an oscillation angle of 0.5°. The incident X-ray intensity, at an energy of 11.95 keV, was attenuated to 0.5% transmission, which corresponds to an estimated flux of 2.4×10^{10} photons s^{-1}. Data were collected using helical acquisition, so that dose was evenly distributed along the crystal. The PILATUS 6M-F detector (Dectris) was positioned at the minimal distance of 150 mm.

Laue data were collected on the same day at BioCARS (Advanced Photon Source, Argonne National Laboratory), similar to data collection described in previous work[16], except that Laue stills were collected in 1° steps, for a total of 3,000 images. Data were collected at ambient temperature (about 295 K).

Thermolysin and lysozyme: Careless 0.4.1 was run with a bivariate prior, sweeping over fifteen double-Wilson *r* values on two datasets, one of thermolysin containing 3160 images collected from a serial XFEL experiment, CXIDB81[45], and one of lysozyme containing the first 999 images from the above data collection. Lysozyme data were indexed and integrated using *laue-dials* (https://github.com/rs-station/laue-dials; manuscript in preparation). Thermolysin unmerged intensities (in DIALS *.pickle*

format) were obtained from CXIDB 81. Both datasets were split according to Friedel symmetry and scaled and merged in Careless, applying the bivariate prior to relate anomalous half-datasets. Next, the anomalous half-datasets were combined into single datasets of merged structure factor amplitudes. Correlation coefficients on these datasets were computed using *careless.cchalf*, *careless.ccpred*, and *careless.ccanom*[36]. Phenix version 1.20.1[61] was used to phase the resultant structure factor amplitudes by isomorphous replacement with PDBID 2TLI in the case of thermolysin and with a high-resolution monochromatic model in the case of lysozyme (**Table S1**). Additionally, an anomalous omit map was phased. The phases and omit map were then used for visualization of anomalous omit peaks with *PyMOL* and measurement of peak heights with *reciprocalspaceship*. Data processing scripts can be found in the Zenodo deposition.

Fragment screening of Mac1. Data collection and analysis of a crystallographic drug fragment screen of SARS-CoV-2 non-structural protein 3 (Nsp3) macrodomain 1 (Mac1) have been previously described[11]. For the work described here, one unmerged *apo* dataset and sixteen unmerged *holo* datasets HKL files were converted to MTZ files and then processed with Careless 0.4.1, with use of a univariate or multivariate prior to relate the OFF and sixteen holo files. The double-Wilson *r* was set to a uniform value for all sixteen *apo-holo* dependencies, and was varied between Careless runs as $1 - 0.5^{\text{run number}}$. Additionally, we introduced a control *apo* dataset, a duplicate of the reference that also depended on the reference *apo* dataset, and difference maps were computed by subtracting each *holo* dataset from the control *apo* dataset. To compensate for the duplication of the reference *apo* dataset, we multiply the errors of the intensities by $\sqrt{2}$. Metadata keys included observed Miller indices, detector position and positional encoding, image number, the azimuthal angle PSI[65], diffracted beam direction parameters ALF1 and BET1[66], and dataset one-hot encoding. Correlation coefficients were computed using *careless.ccpred*. *Apo* phases were obtained by refinement of a model against the *apo* mtz scaled using a univariate prior. Refinement was performed using PHENIX v1.20.1. The *apo* phases were then used for creating isomorphous difference maps of each *holo* minus *apo* dataset. We created difference maps and calculated peak heights using *rs-booster* (https://github.com/rs-station/rs-booster). PanDDA 1-background density correction (1-BDC) values were taken from an analysis that included all previously-reported datasets[11,49]. Data processing scripts can be found in the Zenodo deposition.

Data Availability

Supplementary notebooks, including analyses used for generating **Figures 1-2** and **S1-4,10-13** are available in a GitHub repository at https://github.com/Hekstra-Lab/dw. The monochromatic sodium-iodide-soaked lysozyme structure presented here has been deposited in the Protein Data Bank[67] with PDB ID 9B7C. All merged structure factors, crossvalidation data, and analysis used to generate **Figures 3-6** and **S5-9** have been deposited in Zenodo under accession code 10.5281/zenodo.11099739 at https://doi.org/10.5281/zenodo.11099739. The analyses, without merged structure factors, are included in a GitHub repository at https://github.com/Hekstra-Lab/dw-examples.

Code Availability

The source code used for generating all figures and tables are available in the GitHub repositories https://github.com/Hekstra-Lab/dw and https://github.com/Hekstra-Lab/dw-examples as well as deposited in Zenodo at https://doi.org/10.5281/zenodo.11099739. The multivariate prior scaling model is implemented in the Careless python package, available in the GitHub repository https://github.com/rs-station/careless.

Acknowledgements

Author contributions

D.R.H. and K.M.D. conceived the project. D.R.H. developed the theory and K.M.D. implemented the algorithm. J.B.G. and K.M.D. collected lysozyme anomalous diffraction data. D.R.H., H.K.W., and M.A.K. analyzed the data. D.R.H., H.K.W., and K.M.D. drafted the manuscript with input and edits from all authors.

References

1   Shin, S. New era of synchrotron radiation: fourth-generation storage ring. *AAPPS Bulletin* **31**, 21 (2021).

2   Emma, P. *et al.* First lasing and operation of an ångstrom-wavelength free-electron laser. *nature photonics* **4**, 641-647 (2010).

3   Reschke, D. & Hamburg, D. Recent Progress with Eu-XFEL. *SFR, Canada, USA* **2015** (2015).

4   Olmos, J. L. *et al.* Enzyme intermediates captured "on the fly" by mix-and-inject serial crystallography. *BMC Biology* **16**, 59, doi:10.1186/s12915-018-0524-5 (2018).

5   Schmidt, M. Mix and inject: Reaction initiation by diffusion for time-resolved macromolecular crystallography. *Advances in Condensed Matter Physics* **2013** (2013).

6   Calvey, G. D., Katz, A. M. & Pollack, L. Microfluidic mixing injector holder enables routine structural enzymology measurements with mix-and-inject serial crystallography using X-ray free electron lasers. *Analytical chemistry* **91**, 7139-7144 (2019).

7   Pandey, S. *et al.* Observation of substrate diffusion and ligand binding in enzyme crystals using high-repetition-rate mix-and-inject serial crystallography. *IUCrJ* **8**, 878-895 (2021).

8   Stagno, J. *et al.* Structures of riboswitch RNA reaction states by mix-and-inject XFEL serial crystallography. *Nature* **541**, 242-246 (2017).

9   Douangamath, A. *et al.* Crystallographic and electrophilic fragment screening of the SARS-CoV-2 main protease. *Nature communications* **11**, 5047 (2020).

10   Keedy, D. A. *et al.* An expanded allosteric network in PTP1B by multitemperature crystallography, fragment screening, and covalent tethering. *eLife* **7**, e36307, doi:10.7554/eLife.36307 (2018).

11   Schuller, M. *et al.* Fragment binding to the Nsp3 macrodomain of SARS-CoV-2 identified through crystallographic screening and computational docking. *Science advances* **7**, eabf8711 (2021).

12   Mehlman, T. S. *et al.* Room-temperature crystallography reveals altered binding of small-molecule fragments to PTP1B. *Elife* **12**, e84632 (2023).

13   Newman, J. A. *et al.* Structure, mechanism and crystallographic fragment screening of the SARS-CoV-2 NSP13 helicase. *Nature communications* **12**, 4848 (2021).

14   Günther, S. *et al.* X-ray screening identifies active site and allosteric inhibitors of SARS-CoV-2 main protease. *Science* **372**, 642-646, doi:doi:10.1126/science.abf7945 (2021).

15   Wolff, A. M. *et al.* Mapping protein dynamics at high spatial resolution with temperature-jump X-ray crystallography. *Nature Chemistry* **15**, 1549-1558, doi:10.1038/s41557-023-01329-4 (2023).

16   Hekstra, D. R. *et al.* Electric-field-stimulated protein mechanics. *Nature* **540**, 400-405, doi:10.1038/nature20571 (2016).

17   Keedy, D. A. *et al.* Mapping the conformational landscape of a dynamic enzyme by multitemperature and XFEL crystallography. *Elife* **4**, e07574 (2015).

18   Terwilliger, T. C. & Berendzen, J. Automated MAD and MIR structure solution. *Acta Crystallographica Section D: Biological Crystallography* **55**, 849-861 (1999).

19   Howell, P. & Smith, G. Identification of heavy-atom derivatives by normal probability methods. *Journal of applied crystallography* **25**, 81-86 (1992).

20   Cao, H. & Skolnick, J. Time-resolved x-ray crystallography capture of a slow reaction tetrahydrofolate intermediate. *Structural Dynamics* **6** (2019).

21   Nogly, P. *et al.* Retinal isomerization in bacteriorhodopsin captured by a femtosecond x-ray laser. *Science* **361**, eaat0094 (2018).

22   Dalton, K. M., Greisman, J. B. & Hekstra, D. R. A unifying Bayesian framework for merging X-ray diffraction data. *Nature Communications* **13**, 7764 (2022).

23   Zielinski, K. A. *et al.* Resolving DJ-1 Glyoxalase Catalysis Using Mix-and-Inject Serial Crystallography at a Synchrotron. *Submitted.* (2024).

24   Zielinski, K. A. *et al.* Scaling and Merging Time-Resolved Laue Data with Variational Inference. *Submitted.* (2024).

25   Yü, S. Determination of absolute from relative X-ray intensity data. *Nature* **150**, 151-152 (1942).

26   Wilson, A. J. C. The Probability Distribution of X-Ray Intensities. *Acta Crystallogr* **2**, 318-321, doi:Doi 10.1107/S0365110x49000813 (1949).

27   Read, R. J. Structure-factor probabilities for related structures. *Acta Crystallogr A* **46**, 900-912 (1990).

28   Bricogne, G. Bayesian statistical viewpoint on structure determination: Basic concepts and examples. *Method Enzymol* **276**, 361-423, doi:Doi 10.1016/S0076-6879(97)76069-5 (1997).

29   Rice, S. O. Mathematical analysis of random noise. *The Bell System Technical Journal* **23**, 282-332 (1944).

30   Rice, S. O. Mathematical analysis of random noise. *The Bell System Technical Journal* **24**, 46-156 (1945).

31   Greisman, J. B. *et al.* Native SAD phasing at room temperature. *Acta Crystallographica Section D: Structural Biology* **78**, 986-996 (2022).

32   Hauptman, H. A. On integrating the techniques of direct methods and isomorphous replacement. I. The theoretical basis. *Acta Crystallographica Section A: Crystal Physics, Diffraction, Theoretical and General Crystallography* **38**, 289-294 (1982).

33   Murphy, K. P. Machine Learning: A Probabilistic Perspective. *Machine Learning: A Probabilistic Perspective*, 1-1067, doi:Book_Doi 10.1002/3527606696 (2012).

34   Read, R. J. New ways of looking at experimental phasing. *Acta Crystallographica Section D: Biological Crystallography* **59**, 1891-1902 (2003).

35   Karplus, P. A. & Diederichs, K. Linking Crystallographic Model and Data Quality. *Science* **336**, 1030-1033, doi:doi:10.1126/science.1218231 (2012).

36   Aldama, L. A., Dalton, K. M. & Hekstra, D. R. Correcting systematic errors in diffraction data with modern scaling algorithms. *Acta Crystallographica Section D: Structural Biology* **79** (2023).

37   Evans, P. Scaling and assessment of data quality. *Acta Crystallographica Section D: Biological Crystallography* **62**, 72-82 (2006).

38   Tenboer, J. *et al.* Time-resolved serial crystallography captures high-resolution intermediates of photoactive yellow protein. *Science* **346**, 1242-1246, doi:doi:10.1126/science.1259357 (2014).

39   Pande, K. *et al.* Femtosecond structural dynamics drives the trans/cis isomerization in photoactive yellow protein. *Science* **352**, 725-729, doi:doi:10.1126/science.aad5081 (2016).

40   Ihee, H. *et al.* Visualizing reaction pathways in photoactive yellow protein from nanoseconds to seconds. *Proceedings of the National Academy of Sciences* **102**, 7145-7150, doi:doi:10.1073/pnas.0409035102 (2005).

41   Matthews, B. & Czerwinski, E. W. Local scaling: a method to reduce systematic errors in isomorphous replacement and anomalous scattering measurements. *Acta Crystallographica Section A: Crystal Physics, Diffraction, Theoretical and General Crystallography* **31**, 480-487 (1975).

42   Šrajer, V. *et al.* Protein conformational relaxation and ligand migration in myoglobin: a nanosecond to millisecond molecular movie from time-resolved Laue X-ray diffraction. *Biochemistry* **40**, 13802-13815 (2001).

43   De Zitter, E., Coquelle, N., Oeser, P., Barends, T. R. M. & Colletier, J.-P. Xtrapol8 enables automatic elucidation of low-occupancy intermediate-states in crystallographic studies. *Communications Biology* **5**, 640, doi:10.1038/s42003-022-03575-7 (2022).

44   Ursby, T. & Bourgeois, D. Improved Estimation of Structure-Factor Difference Amplitudesfrom Poorly Accurate Data. *Acta Crystallographica Section A Foundations of Crystallography* **53**, 564-575, doi:10.1107/S0108767397004522 (1997).

45   Kern, J. *et al.* Taking snapshots of photosynthetic water oxidation using femtosecond X-ray diffraction and spectroscopy. *Nature Communications* **5**, 4371, doi:10.1038/ncomms5371 (2014).

46   Brewster, A. S. *et al.* Improving signal strength in serial crystallography with DIALS geometry refinement. *Acta Crystallographica Section D: Structural Biology* **74**, 877-894 (2018).

47   Uervirojnangkoorn, M. *et al.* Enabling X-ray free electron laser crystallography for challenging biological systems from a limited number of crystals. *Elife* **4**, e05421 (2015).

48   Günther, S. *et al.* X-ray screening identifies active site and allosteric inhibitors of SARS-CoV-2 main protease. *Science* **372**, 642-646 (2021).

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

50   Murshudov, G. N., Vagin, A. A. & Dodson, E. J. Refinement of macromolecular structures by the maximum-likelihood method. *Acta Crystallographica Section D: Biological Crystallography* **53**, 240-255 (1997).

51   Barends, T. R. M. *et al.* Influence of pump laser fluence on ultrafast myoglobin structural dynamics. *Nature* **626**, 905-911, doi:10.1038/s41586-024-07032-9 (2024).

52   Greisman, J. B. *et al.* Perturbative diffraction methods resolve a conformational switch that facilitates a two-step enzymatic mechanism. *Proceedings of the National Academy of Sciences* **121**, e2313192121 (2024).

53   Vorontsov, I. I. & Coppens, P. On the refinement of time-resolved diffraction data: comparison of the random-distribution and cluster-formation models and analysis of the light-induced increase in the atomic displacement parameters. *Journal of synchrotron radiation* **12**, 488-493 (2005).

54   Blei, D. M., Kucukelbir, A. & McAuliffe, J. D. Variational inference: A review for statisticians. *Journal of the American statistical Association* **112**, 859-877 (2017).

55   Jordan, M. I., Ghahramani, Z., Jaakkola, T. S. & Saul, L. K. An introduction to variational methods for graphical models. *Machine learning* **37**, 183-233 (1999).

56   Kingma, D. P. & Welling, M. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* (2013).

57   Dillon, J. V. *et al.* Tensorflow distributions. *arXiv preprint arXiv:1711.10604* (2017).

58   Schmidt, M. *et al.* Protein energy landscapes determined by five-dimensional crystallography. *Acta Crystallographica Section D: Biological Crystallography* **69**, 2534-2542 (2013).

59   Wojdyr, M. GEMMI: A library for structural biology. *Journal of Open Source Software* **7**, 4200 (2022).

60   Greisman, J. B., Dalton, K. M. & Hekstra, D. R. Reciprocalspaceship: a Python library for crystallographic data analysis. *Journal of Applied Crystallography* **54**, 1521-1529 (2021).

61   Adams, P. D. *et al.* PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallographica Section D: Biological Crystallography* **66**, 213-221 (2010).

62   Borgstahl, G. E. O., Williams, D. R. & Getzoff, E. D. 1.4 .ANG. Structure of Photoactive Yellow Protein, a Cytosolic Photoreceptor: Unusual Fold, Active Site, and Chromophore. *Biochemistry* **34**, 6278-6287, doi:10.1021/bi00019a004 (1995).

63   DeLano, W. L. Pymol: An open-source molecular graphics tool. *CCP4 Newsl. Protein Crystallogr* **40**, 82-92 (2002).

64   Takeda, K. *et al.* Multi-wavelength anomalous diffraction method for I and Xe atoms using ultra-high-energy X-rays from SPring-8. *Journal of applied crystallography* **37**, 925-933 (2004).

65   Schwarzenbach, D. & Flack, H. D. On the definition and practical use of crystal-based azimuthal angles. *Journal of applied crystallography* **22**, 601-605 (1989).

66   Kabsch, W. xds. *Acta Crystallographica Section D: Biological Crystallography* **66**, 125-132 (2010).

67   Berman, H. M. *et al.* The protein data bank. *Acta Crystallographica Section D: Biological Crystallography* **58**, 899-907 (2002).

68   Winter, G. *et al.* DIALS: implementation and evaluation of a new integration package. *Acta Crystallographica Section D: Structural Biology* **74**, 85-97 (2018).

69   Hunter, J. D. Matplotlib: A 2D graphics environment. *Computing in science & engineering* **9**, 90-95 (2007).

70   Harris, C. R. *et al.* Array programming with NumPy. *Nature* **585**, 357-362 (2020).

71   McKinney, W. pandas: a foundational Python library for data analysis and statistics. *Python for high performance and scientific computing* **14**, 1-9 (2011).

72   Virtanen, P. *et al.* SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature methods* **17**, 261-272 (2020).

73   Waskom, M. L. Seaborn: statistical data visualization. *Journal of Open Source Software* **6**, 3021 (2021).

74   Abadi, M. *et al.* in *12th USENIX symposium on operating systems design and implementation (OSDI 16).* 265-283.