bioRxiv preprint doi: https://doi.org/10.1101/2021.02.11.429988; this version posted March 4, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Understanding conformational dynamics from macromolecular crystal diffuse scattering

Parichita Mazumder and Kartik Ayyer*
Max Planck Institute for the Structure and Dynamics of Matter, 22761 Hamburg, Germany
Center for Free Electron Laser Science, 22761 Hamburg, Germany and
The Hamburg Center for Ultrafast Imaging, 22761 Hamburg, Germany

All macromolecular crystals contain some extent of disorder. The diffraction from such crystals contains diffuse scattering in addition to Bragg peaks and this scattering contains information about correlated displacements in the constituent molecules. While much work has been performed recently in decoding the dynamics of the crystalline ordering, the goal of understanding the internal dynamics of the molecules within a unit cell has been out-of-reach. In this article, we propose a general framework to extract the internal conformational modes of a macromolecule from diffuse scattering data. We combine insights on the distribution of diffuse scattering from short- and long-range disorder with a Bayesian global optimization algorithm to obtain the best fitting internal motion modes to the data. To illustrate the efficacy of the method, we apply it to a publicly available dataset from triclinic lysozyme. Our mostly parameter-free approach can enable the recovery of a much richer, dynamic structure from macromolecular crystallography.

X-ray crystallography remains the preeminent technique to determine the structures of macromolecules at atomic resolution. During a crystallography experiment, in addition to the bright and sharp Bragg peaks used for crystallographic analysis, one also observes weaker scattering at all angles, collectively termed diffuse scattering. The Bragg peak intensities encode information about the average electron density in the unit cell of the crystal, which in turn is used to generate a model of the average atomic positions and their variance (B-factors or ADPs) [1]. Part of the diffuse scattering is composed of trivial contributions from the Compton scattering, bulk solvent scattering, air scatter and other parasitic scattering from beam line components. But crucially, when the crystal is not perfectly ordered, it also contains photons scattered from the proteins themselves. For brevity, in the rest of the article, we will refer to this disorder-induced scatter as diffuse scattering, while keeping in mind that the other contributions must also be accounted for while preprocessing the data [2, 3].

If one assumes the disorder is not too large, the displacements of individual atoms are expected to follow a normal distribution (an assumption borne out by the success of the B-factor formalism in crystallographic refinement). Within this limit, the measured diffuse scattering is a function of the two-point correlation function of all atomic displacements. This means that one has direct access to equilibrium dynamics of proteins via correlations in their atomic displacements. In recent years, there has been an upsurge in interest in such data, driven by the desire to understand proteins as more than static objects since their equilibrium dynamics play a crucial role in their function. With modern detectors and X-ray sources, we can now measure these weak intensities with sufficient quality to enable interpretation [4].

Over the past three decades many models of motion have been proposed to describe the diffuse scattering including liquid-like motion [5–8], rigid-body motion [9–12], lattice dynamics [3, 13] and normal mode analysis [14–17]. Another approach has been to examine the diffuse scattering predicted by molecular dynamics (MD) simulations of ever-larger systems [3, 18–21]. In the latter case, if MD predictions agree with the observed diffuse scattering, one can analyze the trajectories to gain insights into eqilibrium protein dynamics.

It has also been observed that agreement between experimental and calculated diffuse maps improve when motion due to coupled displacements of neighboring molecules in the lattice is included [22]. Recently, Meisburger, Case and Ando (MCA) reported a carefully processed dataset on triclinic lysozyme and analyzed it with a lattice dynamics normal-mode model [3], leading to the highest correlation coefficients with the experimental data yet. Unfortunately, these strong results still do not provide insights into functionally relevant dynamics of the proteins, since intermolecular interactions in a crystal are not likely to be relevant in the native environment. The long term goal of extracting a dynamical structure from macromolecular crystallography data requires a different approach, which we describe below.

In order to understand the conformational landscape sampled by crystallized proteins from their diffuse scattering, we employ three concepts which have not been utilized previously in the field. The first is to capitalize on the separation of diffuse scattering from long- and short-range correlations in reciprocal space. Secondly, we develop a Bayesian optimization pipeline to refine the dynamical parameters of our model which minimizes the number of computationally expensive calculations of the scattering from given parameters. Finally, we use restrained molecular dynamics to provide a basis for fitting against the measured diffuse scattering which better identifies the dominant modes in crystallized proteins.

We will expand upon each of these pillars as well as de-

* kartik.ayyer@mpsd.mpg.de

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.11.429988; this version posted March 4, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

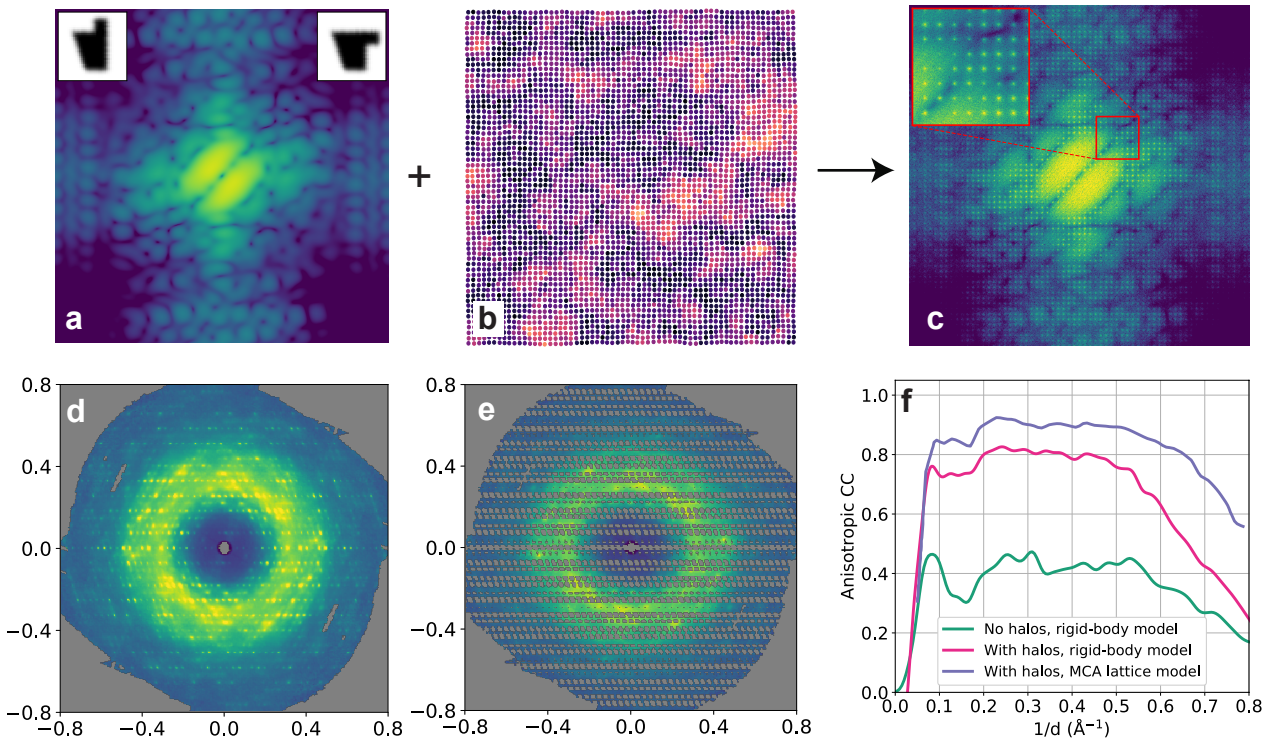


FIG. 1. Toy model and sample system. (a) Diffuse scattering expected from an equally populated two-state model combined with rigid body translational disorder. The two states are shown in the insets. (b) Schematic of lattice dynamics of disorder (exponentially decaying correlations). Lighter colors represent lattice points further from their ideal positions. (c) Diffuse scattering when the unit cells are displaced according to the lattice in (b) and each unit cell is randomly in one of the two states. One can clearly see that other than the 'halos' around Bragg peaks (shown in the red box), the smooth diffuse intensity is the same as the model shown in (a). Both (a) and (c) intensities are shown on a logarithmic scale for clarity. (d) Slice through the experimental 3D diffuse intensity distribution, showing both sharp and smooth features. (e) Same slice as (d), but with the sharp 'halos' surrounding each Bragg peak masked out. (e) Pearson cross-correlation between a simple disorder model with rigid molecules and exponentially decaying intermolecular correlations and the experimental data with and without the halo features. Higher CC values obtained when the halo features are included contain the relatively trivial information that the 'halos' are bright and that they surround reciprocal lattice points. We also show the CC values obtained by MCA using their more sophisticated lattice model for intermolecular displacements.

scribing the dataset and system on which we will apply our approach in the results section. We show a significant improvement in the fit to the measured data upon inclusion of an optimized internal dynamics model. This analysis can also be extended to other systems and experimental datasets.

RESULTS

Short- and long-range correlations. The first property of diffuse scattering we exploit is that long-range correlations in displacements produce sharp intensity features and vice versa. As an extreme example, if the displacements are completely uncorrelated between atoms (correlation length is zero), the diffuse scattering is unstructured and contains no information apart from the average amount of disorder [23]. The other extreme is if the displacements of all atoms are fully correlated, in which case there is no disorder, and the "diffuse" scattering consists of just the Bragg peaks.

A special crossover happens when the correlation length increases past the unit cell size. Above this limit, if we assume that the crystal is homogeneous, then even though the atoms in each unit cell are displaced randomly, the probability distribution of the displacements described by the covariance matrix is the same for all unit cells. This means that the covariance matrix of atomic displacements has the symmetry of the crystal, and so too should the diffuse scattering produced by such long-range disorder, resulting in scattering in and around the reciprocal lattice points. This form of diffuse scattering has been observed and attributed to long range liquid-like correlations [22] and thermal acoustic phonon dynamics [3, 13]. In fact, such 'halos' around Bragg peaks are expected whenever intermolecular correlations monotonically decay with distance smoothly, as we expect if no special phonon modes are excited (see Methods section

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.11.429988; this version posted March 4, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

'Diffuse scattering from decaying intermolecular correlations' for a mathematical justification).

If the dynamics of each unit cell are independent, then the diffuse scattering does not 'see' the reciprocal lattice at all. One example of this is uncorrelated rigid body translational (RBT) disorder [10] where the diffuse scattering is just the molecular transform weighted by the complement of the Debye-Waller factor. But this could also be coupled with conformational variations as described by simple models like liquid-like motion (LLM) which extend the RBT model to allow relaxation of strict rigidity when pairs of atoms are far apart. Such intramolecular correlations are the ones relevant to the structural flexibility of the molecule itself.

In general, when both intra- and inter-molecular correlations are present, their relative contributions to the diffuse scattering at a given $\mathbf{q}$ are strongly dependent upon its distance to the nearest reciprocal lattice point. In Fig. 1(a-c) we illustrate the reciprocal space separation of diffuse scattering from long- and short-range correlations with a toy model. Figure 1(a) shows the diffuse scattering from a crystal where the unit cell can be in one of two states (shown in inset) with random but equal occupancy and with some random translational disorder. This translational disorder is correlated across unit cells and is shown in Fig. 1(b) where only lattice points are shown and colored by their displacement magnitude. If one replaces each lattice point randomly by one of the two states, the crystal diffraction is shown in Fig. 1(c), where one can see the 'halos' from the distorted lattice on top of the smooth scattering from the two-state model. This kind of separation is derived from some basic assumptions in the Methods section 'Diffuse scattering from separable correlations'.

Thus, in many cases, one can separate the scattering from inter- and intra-unit cell correlations in reciprocal space. The smooth diffuse scattering far from the Bragg peak 'halos' are sensitive to intra-unit cell or internal displacement correlations. This has been implicitly recognized earlier in attempts to "filter-out" these 'halos' to access the internal dynamics [2, 7]. We adapt an approach of masking out the halo features to avoid artifacts due to filtering and since our optimization algorithm can tolerate some missing data.

Another important point which one notes is that the long-range intermolecular correlations do have an effect on the scattering far from Bragg peaks. This is in the form of a rigid body translation component to the disorder within a unit cell. This is not to say that the molecules are rigid, rather that a significant part of the total disorder may be incorporated in a concerted motion due to lattice effects, but each molecule may be conformationally different on top of that.

Model system. We apply our approach to the dataset collected by MCA on triclinic lysozyme crystals (PDB: 6o2h, CXIDB: 128) [3]. A central slice through their measured intensity distribution is shown in Fig. 1(d). In addition to being carefully analyzed in terms of background characterization and merging, the triclinic $P1$ lattice also has the advantage of having a single asymmetric unit per unit cell, avoiding possible complications due to molecules being in different orientations. The authors observe that the dominant features were 'halos' surrounding Bragg peaks showing a $1/|q-q_{\mathrm{Bragg}}|^2$ intensity falloff, resembling the well-known behavior of thermal acoustic phonons. These are modes where the displacement of one molecule induces displacements in its neighbors, resulting in longitudinal waves with long correlation lengths compared to the lattice constants. MCA analyze these 'halos' to reconstruct the connectivity network of the crystal while assuming the molecules themselves to be rigid. A small improvement is observed by allowing the molecular structure to vary according to a previously predicted normal mode.

Since our primary interest lies in understanding the internal dynamics, we mask the regions of reciprocal space close to reciprocal lattice points and use the rest to optimize our dynamics model for a single protein. This has the additional computational benefit of needing to only simulate the MD trajectory of a single molecule, rather than a super-cell [3, 21], making it more computationally accessible as a technique. Figure 1(e) shows a slice through the masked intensities for the lysozyme model dataset. One effect of this masking is that the correlation coefficient between the measured and predicted data is lower than if the 'halos' were retained. This is due to two reasons, first that the measured intensities in this region are weaker, and hence noisier. The other, more important reason is that any model with 'halos' will have a relatively high Pearson correlation coefficient (CC) simply because it predicts high values near the Bragg peaks.

This is highlighted in Fig. 1(f) where the CC is calculated with and without masking. A rigid-body translational disorder model is used for the intra-cell dynamics, whose predicted scattering is just the q-weighted molecular transform. The anisotropic CC is around 0.4 for most of the resolution range when comparing against masked data. However, the introduction of even a simple common multiplicative halo function increases the CC to 0.75 (see Methods section for estimation of 'halos'). The "lattice" model with more sophisticated 'halos' from MCA is shown for comparison.

Refinement framework. We now turn our attention to determining what internal variability modes best describe the diffuse scattering far from the Bragg 'halos'. In the rest of this study, we will focus on this masked diffuse scattering since this limits us to correlations at length scales much less than the lattice constants.

As discussed above, within the harmonic approximation, this diffuse scattering is a function of the pair correlation tensor $C_{aa'}$ between atoms a and a' within the molecule. Unfortunately, this mapping is not invertible since the number of independent measurements, which is of the order $\mathcal{O}\left[(d_{\mathrm{protein}}/d_{\mathrm{resolution}})^3\right]$ is much smaller

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.11.429988; this version posted March 4, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

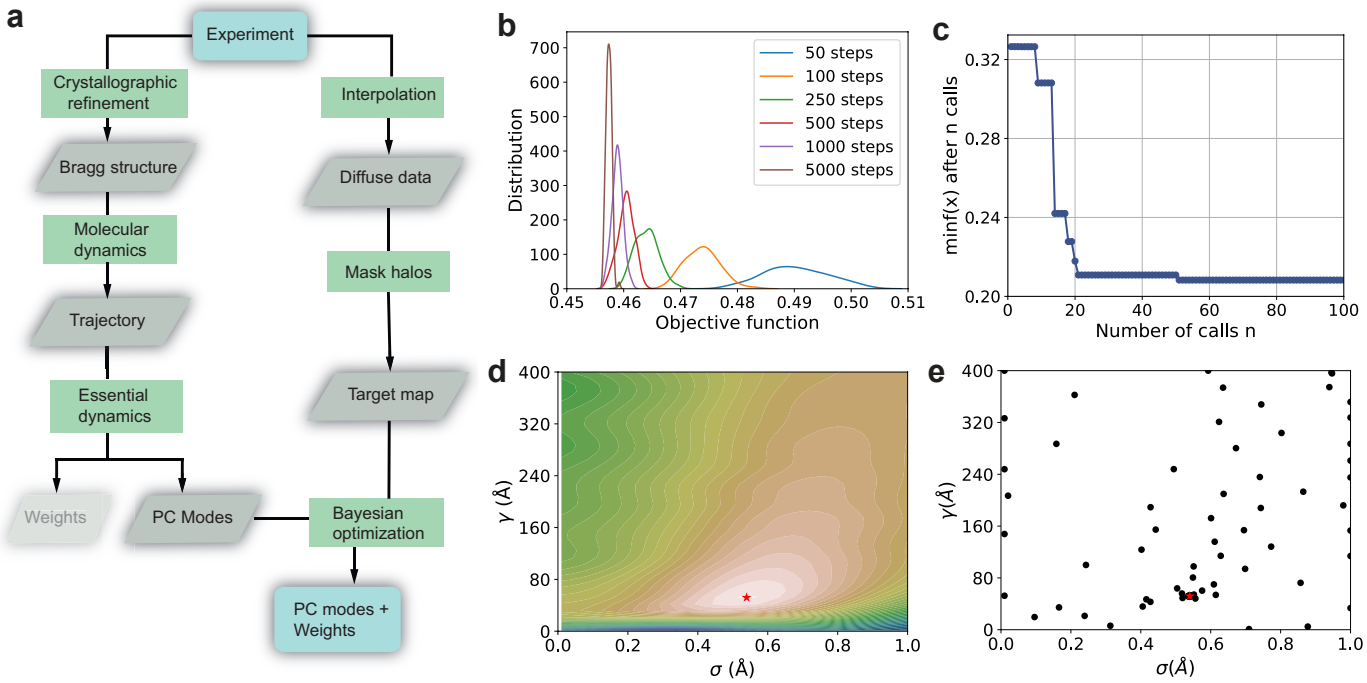


FIG. 2. Reconstruction pipeline. (a) Flow-chart of pipeline to estimate principal component (PC) internal dynamics modes and their weights. Molecular dynamics is used to estimate energetically favorable internal modes and their weights are determined using Bayesian optimization by fitting against the measured diffuse scattering data. (b) Distribution of objective function values from different Monte Carlo runs with the same parameters. The width of this distribution is used to parameterize the Gaussian process regressor used in the Bayesian optimization. (c) Convergence plot showing the evolution of the best objective function with iteration number. This is the output for a 2-parameter optimization of liquid-like motion parameters, σ (rms displacement) and γ (correlation length) for demonstration. (d) Reconstructed surrogate objective function. (e) Points sampled in $\sigma - \gamma$ space, with denser sampling close to the optimum.

than the number of unknowns, $\mathcal{O}[n^2_{\text{atoms}}]$ for protein crystals. This means that there should be many $C_{aa'}$ solutions which generate the same diffuse scattering. But almost all those solutions will be highly unphysical, corresponding to motions which have a very high free energy cost. Thus, one would like to optimize the $C_{aa'}$ matrix within a subspace of physically reasonable distortions of the molecule.

We use molecular dynamics (MD) to generate this subspace. Any MD trajectory can be converted to a $(3N \times 3N)$ $C_{aa'}$ matrix by calculating the covariance of the displacements of all 3 components of all atoms over the trajectory. This matrix can then be diagonalized to calculate the so-called "essential dynamics" of the trajectory [24]. It has been observed that most of the $3N$ modes generated as the eigenvectors of this diagonalization are occupied with low amplitude, collectively giving rise to effects similar to uncorrelated motion of the atoms. There are only a few essential modes which describe the low free-energy parts of the motion in the trajectory.

Now the trajectory itself can be used to predict the diffuse scattering, but this has had limited success predicting scattering features [20]. We adopt the hybrid approach of relying on MD to generate the basis of likely modes, but refine the weights of the modes against the measured diffuse scattering.

The first step in performing this refinement is the forward calculation of the diffuse scattering given a set of modes and their relative weights. By *weight*, we mean the amplitude of random oscillations along that mode. If the average position and structure of the molecule is represented by the $3N$-dimensional vector $\langle \mathbf{r} \rangle$ and the i-th displacement mode is Δ_i, the distorted molecule has the coordinates

$$\mathbf{R} = \langle \mathbf{R} \rangle + \sum_i \alpha_i \Delta_i \tag{1}$$

where α_i is a scalar which is the magnitude of the distortion along mode i. The weights to be optimized are the standard deviations of the random numbers α_i. In order to calculate the diffuse scattering, a large number of randomly distorted molecules are generated to make an ensemble of structures corresponding to a given set of mode weights. The Guinier equation is then used to calculate the diffuse scattering as follows

$$I_D(\mathbf{q}) = \langle |F_n(\mathbf{q})|^2 \rangle_n - |\langle F_n(\mathbf{q}) \rangle_n|^2 \tag{2}$$

where $F_n(\mathbf{q})$ is the molecular transform of the n-th distorted molecule. One immediate feature of this kind of

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.11.429988; this version posted March 4, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

calculation is that it is noisy, on top of the noise in the measured data. This Monte Carlo sampling of the molecular distortions means that in order to reduce the noise in the calculation, one requires a large number of samples, making it expensive to evaluate.

An optimization method purpose-built for such problems is Bayesian global optimization (BGO) [25–28]. In this method, a noise-aware surrogate function is fit against the measurements. This surrogate function is fast to evaluate and so can be optimized using standard methods. The method then also predicts the optimal choice for the next sample. Thus, it is ideal for situations where one has a black-box, expensive to evaluate, noisy forward function without gradients. The goal is not just to optimize the objective function, but also to minimize the number of evaluations of our expensive function. The objective function we employ is the anisotropic cross-correlation coefficient in a relevant resolution range (Eq. 13).

A flow chart of the whole reconstruction algorithm is shown in Fig. 2(a). Figure 2(b) shows the noise in the objective function for a particular set of weights as a function of the number of Monte Carlo samples. To illustrate the method, we show the results of the algorithm for a 2-parameter optimization of the liquid-like motion parameters σ (rms of displacement) and γ (correlation length). Figure 2(c) shows the convergence of the best objective function with iteration number. Figure 2(d) and (e) show the reconstructed surrogate function as a function of σ and γ and the sampled points respectively. Details of the optimization pipeline are given in the Methods section.

Position-restrained molecular dynamics. In order to apply the framework designed above, we need a basis set of dynamical modes which can describe the covariance matrix of atomic displacements. We generate this basis from a molecular dynamics simulation using the ideas of 'essential dynamics' [24]. Briefly, one generates a molecular dynamics (MD) trajectory, from which the displacement covariance matrix is calculated. This matrix is diagonalized to generate eigen-modes and one observes that, usually, only a few modes account for the bulk of the correlated motion.

Proteins in crystalline environments have more constraints to their motion than those in solution due to the presence of the neighboring molecules. One way to account for this is to simulate a super-lattice with multiple unit cells of the crystal [3, 21], which can be computationally expensive. We choose an alternative strategy where we simulate a single molecule but apply additional positional restraints to all non-hydrogen atoms. The strength of the restraining force is a function of the crystallographically refined B-factor of that atom and the effective B-factor of the atom from an unrestrained MD simulation. If the root mean squared fluctuation (RMSF) of an atom from an MD trajectory after rigid body alignment is σ, the effective B-factor is $4\pi^2\sigma^2$.

For an isolated atom simulated at temperature T, a force constant of F results in a B-factor of k_BT/F [29]. One can use this to estimate the effective force constant on each atom i from the intrinsic MD force fields, F_i^{MD}. The total restraining force to produce the measured B-factors should be $F_i^{\mathrm{exp}} = k_BT/B_i^{\mathrm{exp}}$. For our model system, unrestrained MD almost always produces much higher atomic fluctuations than the experimental data. Thus, one can calculate an additional restraining force, $F_i^{\mathrm{res}} = F_i^{\mathrm{exp}} - F_i^{\mathrm{MD}}$. But this is true only in the case where atomic displacements are independent. The presence of structured diffuse scattering means that displacements are correlated and so restraining just one atom reduces the fluctuations of many others by suppressing otherwise favored modes of motion.

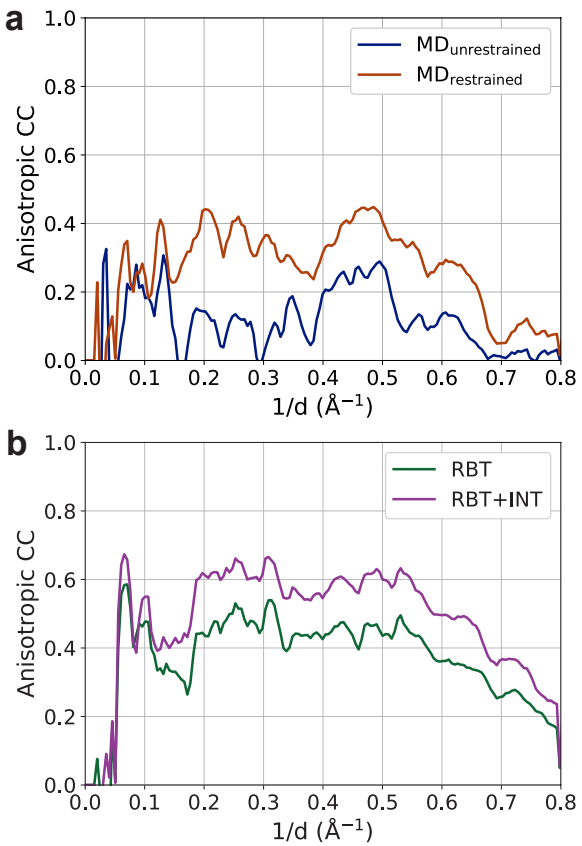
In our system, we found a reasonable agreement between restrained-MD and experimental B-factors by using force constants 200 times lower than those predicted by the above calculation (see Fig. S1 for details) The agreement was not perfect, but is sufficient to significantly improve the correlation with the experimental data.

Application to triclinic lysozyme. The experimental data was masked as described above in the section on short- and long-range correlations by masking out a spherical region around each reciprocal lattice point whose radius was 3/7-th of the inter-peak spacing. This region was spherical in *hkl* space, but ellipsoidal in isotropic reciprocal space since the lattice was triclinic. A slice through the masked intensities is shown in Fig. 1(b).

In order to generate an ensemble of structures from which to calculate the principal component modes, we performed a 1 µs simulation of a single solvated lysozyme molecule with average structure given by the crystallographic structure (PDB: 6o2h). Position restraints were applied as described above in the restrained MD section and the correlation coefficient (CC) with the masked diffuse scattering of the trajectory itself is shown in Fig. 3(a). The covariance matrix of fluctuations of the positions was diagonalized to generate the principal component dynamical modes and their relative weights.

This data was compared against the masked experimental data. Various Bayesian optimization runs were performed, and the results are summarized in Fig. 3(b). The base model was isotropic rigid body translation (RBT), which generates an average CC of 0.44 in the (3.3Å-1.6Å) resolution range. The weights of the principal component modes were optimized along with rigid body translation as described above. That resulted in a significant increase in the average CC in the same resolution range to 0.60 with the inclusion of this internal dynamics (RBT+INT). With the inclusion of liquid-like motion (LLM) and no internal dynamics, the average CC in that range was 0.60 (RBT+LLM). The optimization of internal dynamics in the presence of liquid-like motion increased the CC to 0.62 (RBT+LLM+INT) (data not shown for clarity). The more modest improvement is partially explained by the effect that LLM-like effects can

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.11.429988; this version posted March 4, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



c

	RBT	RBT+ INT$_{rstrn}$	RBT+ INT$_{rstrn,free}$	RBT+ INT$_{unrstrn}$
<CC>$_d$	0.44	0.60	0.60	0.47

FIG. 3. Cross-Correlation between masked experimental data and simulated diffuse intensity. (a) Diffuse scattering calculated from unrestrained and restrained MD trajectories. (b) Different short-range models: rigid-body translation (RBT) and RBT + internal modes (RBT+INT) with RBT $\sigma = 0.46$ Å. (c) Summary of average cross-correlation (CC) in the resolution range d, 3.3Å-1.6Å for different models.

result from many equally weighted modes, each of whom involve the concerted motion of a part of the molecule.

For validation of our INT model we performed Bayesian optimization with top 10 unrestrained MD modes which gave a CC of 0.44 in the same resolution range, which is the same as what we get with RBT only, showing that the wrong choice of basis does not improve the fit to the data even though additional parameters are refined. We also performed an optimization by excluding the Brillouin zones around 10% of the Bragg peaks and calculated the CC_{work} and CC_{free} in the standard way (see Fig. S2). The average CC_{work} was 0.61 while CC_{free} calculated just from the excluded voxels in the dataset was 0.60.

Figure 4 illustrates the most prominent mode found in the (RBT+INT) optimization. This mode describes motions between conformations which are present in crystallized lysozyme Fig. 4(a). To describe the internal motion of this mode, the displacement of the C_α backbone is shown in Fig. 4(b). The β-lobe (THR40 to CYS94) shows higher motion compare to the α-lobe (ARG5 to SER36 and ILE98 to LEU129). The displacement is low for active site GLU35 and ASP52. Some internal deformation is observed in the β-lobe Fig. 4(c-d). Around the hinge region, ASN39 shows significant movement. This and other prominent modes are illustrated in Supplementary Movie S1.

DISCUSSION

Proteins are inherently dynamic objects, with their conformational variation often playing a crucial role in their function. While conventional crystallography can routinely provide high resolution static structures, they provide no direct information about correlated motion within the molecule [4]. Diffuse scattering, which is always present but usually discarded, is affected by the pair correlation of the displacements, and hence is strongly influenced by whether different atoms are displaced synchronously or asynchronously and to what extent.

In this article, we demonstrate a method to fit the internal dynamics of a protein to its diffuse scattering. The first step in this process is the partitioning of the effect of conformational motions from lattice dynamics in reciprocal space for homogeneous crystals. This separation can be detected in the form of e.g. 'halos' around reciprocal lattice points, and the signal far from such lattice-induced features is dominated by internal structural variation of the protein within a unit cell. Once the relevant part of the experimental data is identified, we use molecular dynamics to obtain a basis space for dynamical modes whose weights are optimized using a Bayesian optimization procedure.

As a proof of principle, we apply this method to the diffuse scattering from triclinic lysozyme crystals [3]. Our first observation is that the basis space generated from the MD simulation of isolated proteins in solution does not fit the data well, suggesting that the dominant modes in solution are poorly represented in the crystalline environment. Better agreement with the data was found by applying additional restraining force fields which more closely align the fluctuations in the atomic positions to the experimental B-factors. The optimized dominant modes improve the correlation coefficient with the experimental data to 0.60 from the 0.45 value obtained from a rigid body model of the protein. Note that we have to include rigid body motion along with the internal modes to obtain good agreement, which may explain the poor agreement observed in previous studies incorporating only internal dynamics which hold the center-of-mass of the molecule fixed [22]. The strongest conformational mode in the crystalline environment predicted from the optimization is shown in Fig. 4.

Apart from the initial analysis of the 'halos', the whole

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.11.429988; this version posted March 4, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

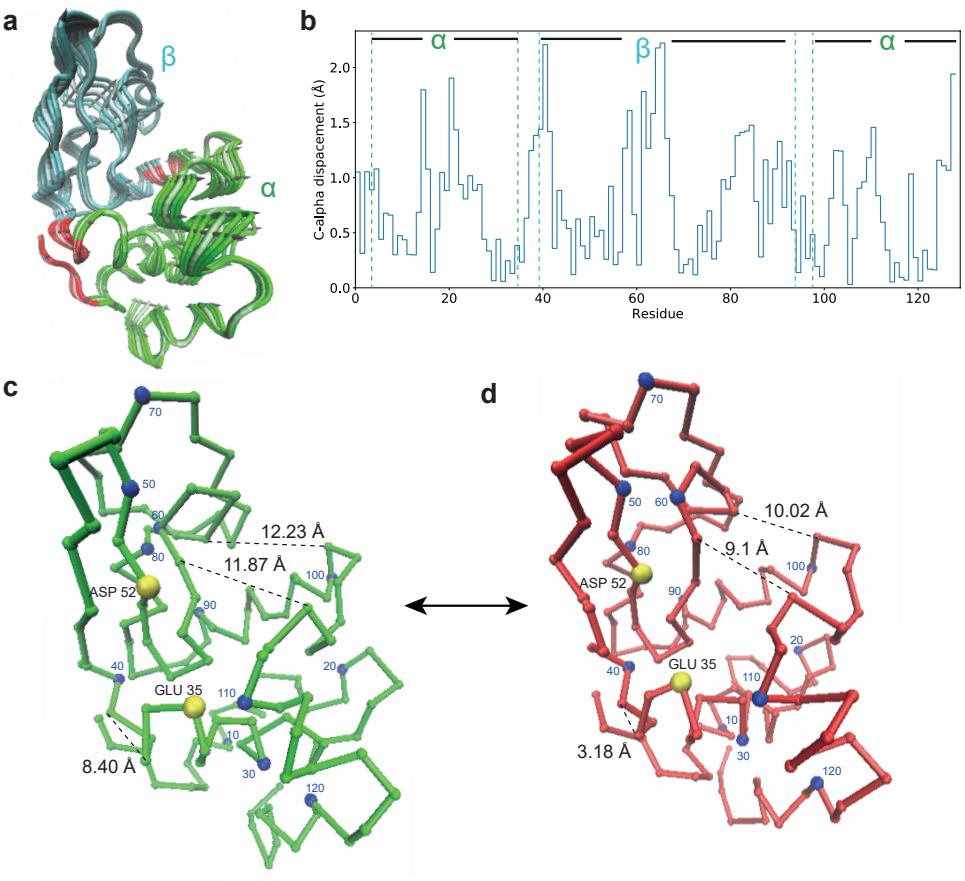


FIG. 4. Visualization of dominant mode of motion. (a) Conformations along the dominant principal mode. The mode is shown in both directions black arrows. (b) C_α-atom displacement for this mode. The α and β lobe atoms are shown within cyan and green dotted lines. (c-d) C_α atoms displacement in two different conformations shown in green(far) and red(close).The distance between ASN37-ASN39, TRP63-ASP101, ASN59-ALA107 are shown to understand the motion. Every 10th C_α atom is colored in blue. The active site residue GLU35 and ASP52 shown in yellow. See Movie S1 for a visualization of this and other modes with non-negligible weights.

procedure is mostly parameter-free. With further refinements to the pipeline, we envision the analysis of such internal dynamics to be routinely performed on crystallographic datasets to create a dynamic model of the molecule, and not just the coordinates plus B-factor static picture we currently extract. One future direction of improvement is a better scheme for restrained MD simulations such that the first few modes are the most relevant to the crystalline dynamics. At the other end, the optimization scheme could also be improved to allow us to increase the number of modes which can be simultaneously optimized. Of course, this also requires improvements in the collection and processing of the experimental data, especially with regards to reducing and characterizing the background scattering.

METHODS

Diffuse scattering from decaying intermolecular correlations. Following the convention of [23], the position of atom a in cell c is $\mathbf{r}_{ac} = \mathbf{r}_c + \mathbf{r}_a$ and the total intensity from a disordered crystal is given by

$$I(\mathbf{q}) = \sum_{cc'} L_{cc'}(\mathbf{q}) \sum_{aa'} [I_{aa'}(\mathbf{q}) \times e^{-2\pi^2 \mathbf{q}^\intercal (U_a + U_{a'})\mathbf{q}} e^{4\pi^2 \mathbf{q}^\intercal C_{aca'c'} \mathbf{q}}] \quad (3)$$

where $L_{cc'}(\mathbf{q}) = e^{2\pi i(\mathbf{r}_c - \mathbf{r}_{c'})}$

$$I_{aa'}(\mathbf{q}) = f_a(\mathbf{q}) f^*_{a'}(\mathbf{q}) e^{2\pi i(\mathbf{r}_a - \mathbf{r}_{a'})}$$

where U_a is the 3x3 variance matrix of atomic displacements, also called the anisotropic displacement parameter (ADP) and $C_{aca'c'}$ is the covariance matrix of the displacements of atoms ac and $a'c'$.

Here we analyze the diffuse scattering when the intermolecular correlations are non-negative and smoothly decaying as a function of the separation between the unit cells. For simplicity the molecules themselves are assumed to translate as rigid units. The next subsection discusses the case when this latter assumption is not true. This generalizes the derivation of scattering for liquid-like

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.11.429988; this version posted March 4, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

motion in [23]. From the above mentioned assumptions, let $C_{aca'c'} = C_{cc'}$ be a function of only $(\mathbf{r}_c - \mathbf{r}_{c'})$ for all a, a'. Then, Eq. 3 becomes

$$I(\mathbf{q}) = \sum_{cc'} L_{cc'}(\mathbf{q}) e^{4\pi^2 \mathbf{q}^\intercal C_{cc'} \mathbf{q}} \sum_{aa'} I_{aa'}(\mathbf{q}) e^{-2\pi^2 \mathbf{q}^\intercal (U_a + U_{a'}) \mathbf{q}} \tag{4}$$

$$= I_B(\mathbf{q}) \sum_{cc'} L_{cc'}(\mathbf{q}) e^{4\pi^2 \mathbf{q}^\intercal C_{cc'} \mathbf{q}} \tag{5}$$

where $I_B(\mathbf{q})$ is just the Bragg intensity distribution obtained from the Fourier transform of the average unit cell. The exponential in the weighted lattice sum can be expanded to give

$$I(\mathbf{q}) = I_B(\mathbf{q}) \sum_{n=0}^{\infty} \frac{(2\pi)^{2n}}{n!} \sum_{cc'} L_{cc'}(\mathbf{q}) \left(\mathbf{q}^\intercal C_{cc'} \mathbf{q}\right)^n \tag{6}$$

Expanding the $L_{cc'}$ term and including the assumption that $C_{cc'}$ is only a function of $\mathbf{r}_{cc'} = \mathbf{r}_c - \mathbf{r}_{c'}$, we can see that the sum only depends upon the separation $\mathbf{r}_{cc'}$. This allows us to insert the integral term $\int d^3\mathbf{u}\, \delta(\mathbf{u} - \mathbf{r}_{cc'})$ and get

$$I(\mathbf{q}) = I_B(\mathbf{q}) \sum_{n=0}^{\infty} \Bigg[\frac{(2\pi)^{2n}}{n!} \times \sum_{cc'} \int d\mathbf{u}\, \delta(\mathbf{u} - \mathbf{r}_{cc'}) e^{2\pi i \mathbf{q}.\mathbf{u}} \left(\mathbf{q}^\intercal C(\mathbf{u}) \mathbf{q}\right)^n \Bigg] \tag{7}$$

The integral is now just a Fourier transform integral,

$$I(\mathbf{q}) = I_B(\mathbf{q}) \sum_{n=0}^{\infty} \Bigg[\frac{(2\pi)^{2n}}{n!} \times \mathcal{F}_{\mathbf{u}} \left(\left(\mathbf{q}^\intercal C(\mathbf{u}) \mathbf{q}\right)^n \sum_{cc'} \delta(\mathbf{u} - \mathbf{r}_{cc'}) \right)(\mathbf{q}) \Bigg] \tag{8}$$

where $\mathcal{F}$ represents the Fourier transform operation. From the convolution theorem,

$$I(\mathbf{q}) = I_B(\mathbf{q}) \sum_{n=0}^{\infty} \frac{(2\pi)^{2n}}{n!} \mathcal{F}_{\mathbf{u}} \left[\left(\mathbf{q}^\intercal C(\mathbf{u}) \mathbf{q}\right)^n \right](\mathbf{q}) \otimes L(\mathbf{q}) \tag{9}$$

where $L(\mathbf{q})$ is the lattice sum generated by the Fourier transform of an ideal lattice. The $n = 0$ term corresponds to the Bragg intensities while the other terms contribute to the diffuse scattering. This series represents a sum of broadened reciprocal lattices where the higher order terms are weighted towards high resolution due to the q^{2n} dependence. $C^n(\mathbf{u})$ gets sharper as n increases which has the reverse effect on its Fourier transform, broadening the convolution kernel for higher n.

Thus, we get the result that when lattice interactions result in correlations which decay with inter-molecular separation, the diffuse scattering is given by a series of broadened Bragg peaks ('halos') where the peaks get broader at higher resolution.

Diffuse scattering from separable correlations. In this section, we analyze the diffuse scattering caused by displacement correlations that can be separated into intra- and inter-molecular terms. Thus, the covariance matrix $C_{aca'c'}$ between atoms ac and $a'c'$ can be described as the sum of two terms

$$C_{aca'c'} = \delta_{cc'} C_{aa'} + C_{cc'} \tag{10}$$

In this formulation, atoms within a molecule have correlated displacements governed by $C_{aa'}$ but the interaction between molecules is governed purely by $C_{cc'}$ for all atoms in the molecule. Physically, this corresponds to different energy scales or mechanisms for internal interactions versus those for lattice interactions.

Then Eq. 3 becomes

$$I(\mathbf{q}) = \sum_{cc'} L_{cc'}(\mathbf{q}) e^{4\pi^2 \mathbf{q}^\intercal C_{cc'} \mathbf{q}} \times \sum_{aa'} I_{aa'}(\mathbf{q}) e^{-2\pi^2 \mathbf{q}^\intercal (U_a + U_{a'}) \mathbf{q}} e^{4\pi^2 \delta_{cc'} \mathbf{q}^\intercal C_{aa'} \mathbf{q}} \tag{11}$$

Separating the two cases when $c = c'$ and $c \neq c'$ and then completing the sum, we get

$$I(\mathbf{q}) = I_B(\mathbf{q}) \sum_{cc'} L_{cc'}(\mathbf{q}) e^{4\pi^2 \mathbf{q}^\intercal C_{cc'} \mathbf{q}} + N_c \sum_{aa'} I_{aa'}(\mathbf{q}) e^{-2\pi^2 \mathbf{q}^\intercal (U_a + U_{a'}) \mathbf{q}} \left(e^{4\pi^2 \mathbf{q}^\intercal C_{aa'} \mathbf{q}} - 1 \right) \tag{12}$$

The first term results in the 'halos' as discussed in the previous subsection while the second term is independent of the lattice periodicity and only depends on the internal dynamics of the molecule, $C_{aa'}$.

The target smooth diffuse scattering map which corresponds to the short-range correlations, was generated from experimental data by masking out a spherical region around each reciprocal lattice point('halos').

Simple model for estimation of 'halos'. To illustrate the effects of including 'halos' on the CC metric, a simple uniform 'halo' model was estimated directly from the experimental data. The 'halos' were simulated by broadening the Bragg intensities by a 'halo' kernel, whose shape was estimated as described below.

The signal in each Brillouin zone was averaged, leading to a $7 \times 7 \times 7$ voxel volume. The minimum value was subtracted and the peak was exponentiated by a variable exponent α and tiled out to generate a lattice-sum function. This function was multiplied to the rigid-body diffuse scattering to generate the predicted intensities with 'halos'. A one-dimensional scan was performed to optimize the α parameter, leading to a value of 3.8 used in estimating the CC values shown in Fig. 1(c).

Molecular dynamics simulations. MD simulations were performed using GROMACS 2019.4 [30] using

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.11.429988; this version posted March 4, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

OPLS-AA/L all-atom [31] parameters for protein and OPLS-2009L parameters for nitrate ions [32]. The starting structure was generated using coordinates from PDB ID 6o2h (conformation A for lysozyme with seven nitrate ions, one chloride ion and crystal waters). Next, this model was solvated with a TIP3P water box of dimension $60 \times 67 \times 74$ Å and neutralized with one chloride ion. After minimization (steepest decent), the model was equilibrated for 1 ns with position restraints on protein atoms. Then, two sets of simulation runs were performed in NPT conditions to keep the temperature and pressure stable around 295 K and 1 bar with 2 fs time steps. For unrestrained-MD, we gradually reduced the position restraints on protein atoms to zero over 0.2 µs and then recorded coordinates every 10 ps for 1.7 µs.

For restrained-MD, we applied positional restraints on protein atoms to the equilibrated model (see Fig. S1 for details) during the 1 µs production run to maintain the experimental B-factors.

Diffuse scattering calculation. For diffuse scattering calculation from internal dynamics, the trajectory was fitted with respect to the backbone of the starting configuration, comprising the last 150000 frames for unrestrained-MD and last 90000 frames for restrained-MD. To generate the essential modes of motion from MD trajectory we calculated a $(3N \times 3N)$ covariance matrix of the displacements of all 3 components of all N atoms over the trajectory. This matrix was then diagonalized to calculate the essential modes of the trajectory. We used VMD, NMWiz to visualize the modes [33, 34].

For a given set of coordinates or mode variance vector (or covariance matrix), the diffuse scattering was calculated using the Guinier equation (Eq. 2, in results 'Refinement framework' section).

Bayesian global optimization (BGO). We implemented Bayesian optimization with Gaussian process regressors using the `scikit-optimize` Python package (v0.8)[35]. Since the goal was to maximize the correlation between experimental and simulated diffuse intensity, the objective function was defined as,

$$\mathrm{obj}(\mathbf{s}) = 1. - \mathrm{corrcoef}(I_{\mathrm{calc}}(\mathbf{s}), I_{\mathrm{target}}) \qquad (13)$$

where the correlation was calculated after subtracting the radial averages of both terms and was averaged over the resolution range of 3.3-1.6Å.

The larger the number of Monte Carlo simulation steps used to generate I_{calc}, the lower and less noisy the value of the objective function (see Fig. 2(b)). To compromise between speed and precision, we used 1000 samples during the optimization while the final CC was calculated using 5000 samples. We used BGO to find the minimum of the objective function over the range of one or multiple input parameters, $\mathbf{s}$, in as few iterations as possible. The Gaussian process surrogate model gives an estimate of the objective function which can be used to direct future sampling. For acquiring more samples we used `gp_hedge` acquisition function to minimize over the Gaussian posterior distribution with the `lbfgs` acquisition optimizer. A noise level of 1×10^{-7} was used during optimization with `n_restarts_optimizer` = 2. We performed various optimization runs with respect to masked diffuse data with different short-range models, e.g., 1-parameter optimization of rigid-body translation (RBT) σ, 2-parameters optimization of the liquid-like motion (LLM) parameters σ and γ, 10-parameters optimization of internal modes weights along with RBT $\sigma = 0.46$ Å (see Fig. S3) etc. For internal modes optimization we ran BGO 128 times from random initial starting estimates. We noted that excluding some failed runs, most of the optimized $\mathbf{s}$ values could be clustered around a single point. For convenience, we estimated the best fit internal mode weights by calculating the median value of the results from the 128 runs.

Data availability

The experimental data used in this study were previously published on the CXIDB (ID: 128). The molecular dynamics trajectories and optimization outputs are available from the corresponding author upon reasonable request.

Code availability

The code used to perform the calculations here can be found at `https://github.com/AyyerLab/diffuser`.

[1] Dunitz, J. D., Schomaker, V. & Trueblood, K. N. Interpretation of atomic displacement parameters from diffraction studies of crystals. *The Journal of Physical Chemistry* **92**, 856–867 (1988). URL https://doi.org/10.1021/j100315a002.

[2] Chapman, H. N. *et al.* Continuous diffraction of molecules and disordered molecular crystals. *Journal of applied crystallography* **50**, 1084–1103 (2017).

[3] Meisburger, S. P., Case, D. A. & Ando, N. Diffuse x-ray scattering from correlated motions in a protein crystal. *Nature communications* **11**, 1–13 (2020).

[4] Van Benschoten, A. H. *et al.* Measuring and modeling diffuse scattering in protein x-ray crystallography. *Proceedings of the National Academy of Sciences* **113**, 4069–4074 (2016).

[5] Caspar, D. L. D., Clarage, J., Salunke, D. M. & Clarage, M. Liquid-like movements in crystalline insulin. *Nature*

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.11.429988; this version posted March 4, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

332, 659–662 (1988). URL https://doi.org/10.1038/332659a0.

[6] Clarage, J. B., Clarage, M. S., Phillips, W. C., Sweet, R. M. & Caspar, D. L. D. Correlations of atomic movements in lysozyme crystals. *Proteins: Structure, Function, and Bioinformatics* **12**, 145–157 (1992). URL https://doi.org/10.1002/prot.340120208.

[7] Wall, M. E., Ealick, S. E. & Gruner, S. M. Three-dimensional diffuse x-ray scattering from crystals of staphylococcal nuclease. *Proceedings of the National Academy of Sciences* **94**, 6180–6184 (1997).

[8] Wych, D. C., Fraser, J. S., Mobley, D. L. & Wall, M. E. Liquid-like and rigid-body motions in molecular-dynamics simulations of a crystalline protein. *Structural Dynamics* **6**, 064704 (2019).

[9] Pérez, J., Faure, P. & Benoit, J.-P. Molecular Rigid-Body Displacements in a Tetragonal Lysozyme Crystal Confirmed by X-ray Diffuse Scattering. *Acta Crystallographica Section D* **52**, 722–729 (1996). URL https://doi.org/10.1107/S0907444996002594.

[10] Ayyer, K. *et al.* Macromolecular diffractive imaging using imperfect crystals. *Nature* **530**, 202–206 (2016).

[11] Morgan, A. J. *et al.* Ab initio phasing of the diffraction of crystals with translational disorder. *Acta Crystallographica Section A: Foundations and Advances* **75**, 25–40 (2019).

[12] De Klijn, T., Schreurs, A. & Kroon-Batenburg, L. Rigid-body motion is the main source of diffuse scattering in protein crystallography. *IUCrJ* **6**, 277–289 (2019).

[13] Polikanov, Y. S. & Moore, P. B. Acoustic vibrations contribute to the diffuse scatter produced by ribosome crystals. *Acta Crystallographica Section D: Biological Crystallography* **71**, 2021–2031 (2015).

[14] Faure, P. *et al.* Correlated intramolecular motions and diffuse x–ray scattering in lysozyme. *Nature Structural Biology* **1**, 124–128 (1994). URL https://doi.org/10.1038/nsb0294-124.

[15] Mizuguchi, K., Kidera, A. & Gō, N. Collective motions in proteins investigated by x-ray diffuse scattering. *Proteins: Structure, Function, and Bioinformatics* **18**, 34–48 (1994). URL https://doi.org/10.1002/prot.340180106.

[16] Wall, M. E., Clarage, J. B. & Phillips Jr, G. N. Motions of calmodulin characterized using both bragg and diffuse x-ray scattering. *Structure* **5**, 1599–1612 (1997).

[17] Riccardi, D., Cui, Q. & Phillips Jr, G. N. Evaluating elastic network models of crystalline biological molecules with temperature factors, correlated motions, and diffuse x-ray scattering. *Biophysical journal* **99**, 2616–2625 (2010).

[18] Héry, S., Genest, D. & Smith, J. C. X-ray diffuse scattering and rigid-body motion in crystalline lysozyme probed by molecular dynamics simulation11edited by r. huber. *Journal of Molecular Biology* **279**, 303–319 (1998). URL http://www.sciencedirect.com/science/article/pii/S0022283698917545.

[19] Meinhold, L. & Smith, J. C. Correlated dynamics determining x-ray diffuse scattering from a crystalline protein revealed by molecular dynamics simulation. *Physical Review Letters* **95**, 218103 (2005). URL https://doi.org/10.1103/PhysRevLett.95.218103.

[20] Wall, M. E. *et al.* Conformational dynamics of a crystalline protein from microsecond-scale molecular dynamics simulations and diffuse x-ray scattering. *Proceedings of the National Academy of Sciences* **111**, 17887–17892 (2014).

[21] Wall, M. E. Internal protein motions in molecular-dynamics simulations of Bragg and diffuse X-ray scattering. *IUCrJ* **5**, 172–181 (2018).

[22] Peck, A., Poitevin, F. & Lane, T. J. Intermolecular correlations are necessary to explain diffuse scattering from protein crystals. *IUCrJ* **5**, 211–222 (2018).

[23] Ayyer, K., Yefanov, O. M. & Chapman, H. N. Structure determination by continuous diffraction from imperfect crystals. In *X-ray Free Electron Lasers*, 253–300 (Springer, 2018).

[24] Amadei, A., Linssen, A. B. & Berendsen, H. J. Essential dynamics of proteins. *Proteins: Structure, Function, and Bioinformatics* **17**, 412–425 (1993).

[25] Pelikan, M., Goldberg, D. E. & Cantú-Paz, E. Boa: The bayesian optimization algorithm. In *Proceedings of the 1st Annual Conference on Genetic and Evolutionary Computation - Volume 1*, GECCO'99, 525–532 (Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1999).

[26] Snoek, J., Larochelle, H. & Adams, R. P. Practical bayesian optimization of machine learning algorithms (2012). 1206.2944.

[27] Shahriari, B., Swersky, K., Wang, Z., Adams, R. P. & de Freitas, N. Taking the human out of the loop: A review of bayesian optimization. *Proceedings of the IEEE* **104**, 148–175 (2016). URL https://doi.org/10.1109/JPROC.2015.2494218.

[28] Frazier, P. I. A tutorial on bayesian optimization (2018). 1807.02811.

[29] Li, D.-W. & Brüschweiler, R. All-atom contact model for understanding protein dynamics from crystallographic b-factors. *Biophysical Journal* **96**, 3074–3081 (2009). URL http://www.sciencedirect.com/science/article/pii/S0006349509004731.

[30] Abraham, M. J. *et al.* Gromacs: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **1-2**, 19 – 25 (2015).

[31] Kaminski, G. A., Friesner, R. A., Tirado-Rives, J. & Jorgensen, W. L. Evaluation and reparametrization of the opls-aa force field for proteins via comparison with accurate quantum chemical calculations on peptides. *The Journal of Physical Chemistry B* **105**, 6474–6487 (2001).

[32] Doherty, B., Zhong, X., Gathiaka, S., Li, B. & Acevedo, O. Revisiting opls force field parameters for ionic liquid simulations. *Journal of Chemical Theory and Computation* **13**, 6131–6145 (2017).

[33] Humphrey, W., Dalke, A. & Schulten, K. VMD – Visual Molecular Dynamics. *Journal of Molecular Graphics* **14**, 33–38 (1996).

[34] Bakan, A., Meireles, L. M. & Bahar, I. Prody: Protein dynamics inferred from theory and experiments. *Bioinformatics* **27**, 1575–1577 (2011). URL https://doi.org/10.1093/bioinformatics/btr168.

[35] Head, T., Kumar, M., Nahrstaedt, H., Louppe, G. & Shcherbatyi, I. scikit-optimize/scikit-optimize (2020). URL https://doi.org/10.5281/zenodo.4014775.

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.11.429988; this version posted March 4, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

ACKNOWLEDGMENTS

This work is supported by the Cluster of Excellence 'CUI: Advanced Imaging of Matter' of the Deutsche Forschungsgemeinschaft (DFG) - EXC 2056 - project ID 390715994. The computations in this work were partially performed at the Max Planck Computing & Data Facility. We thank Henry N. Chapman and Thomas J. Lane for valuable comments on a draft of this article.

Author contributions

K.A. conceived the project. K.A. and P.M. wrote the analysis program. P.M performed the molecular dynamics simulations and the analysed the output. K.A and P.M. wrote the manuscript.

Competing interests

The authors declare no competing interests.