Denoising Reveals Low-Occupancy Populations in Protein Crystals

Alisia Fadini[1*+], Virginia Apostolopoulou[2,3], Thomas J. Lane[2,3*], Jasper J. van Thor[1*]

[1] Department of Life Sciences, Faculty of Natural Sciences, Imperial College London, London SW7 2AZ, United Kingdom.

[2] Center for Free-Electron Laser Science CFEL, Deutsches Elektronen-Synchrotron DESY, Notkestr. 85, 22607 Hamburg, Germany.

[3] The Hamburg Centre for Ultrafast Imaging, Luruper Chaussee 149, 22761 Hamburg, Germany.

[+]Current address: Cambridge Institute for Medical Research, University of Cambridge, The Keith Peters Building, Hills Road, Cambridge CB2 0XY, United Kingdom.

[*]Correspondence

Abstract

Advances in structural biology increasingly focus on uncovering protein dynamics and transient or weak macromolecular complexes. Such studies require modeling of low-occupancy species, for instance time-evolving intermediates and bound ligands. In protein crystallography, difference maps that compare paired perturbed and reference datasets are

a powerful way to identify and aid modeling of low-occupancy species. Current methods to generate difference maps, however, rely on manually tuned parameters and, in cases of weak signal due to low occupancy, can fail to extract clear, chemically interpretable signals.

We address these issues, first by showing negentropy is an effective metric to assess difference map quality and can therefore be used to automatically determine parameters needed during difference map calculation. Leveraging this, we apply total variation denoising, an image restoration technique that requires a choice of regularization parameter, to crystallographic difference maps. We show that total variation denoising improves map signal-to-noise and enables us to estimate the latent phase contribution of low-occupancy states. This technology opens new possibilities for time-resolved and ligand-screening crystallography in particular, allowing the detection of states that previously could not be resolved due to their inherently low occupancy.

Introduction

A growing number of cutting-edge macromolecular crystallography methods aim to resolve weakly populated states rather than the primary protein conformation present in a crystal. Mechanistic time-resolved studies at XFELs[1] and synchrotrons[2], high-throughput fragment screens[3], or the study of functional protein conformational changes upon the external perturbation of a crystal (using pH, temperature[4], or an external field[5]) are important examples. Because full conversion to a new state through a transient stimulus, like optical excitation[6] or substrate and ligand diffusion[7–10], is difficult to achieve, such techniques rely on the observation of small differences between crystallographic datasets. This complication can limit their application: photoactivated systems with low quantum yields, weakly-bound ligands, and bound small molecules used for hit-to-lead drug discovery all produce signals with strengths comparable to the experimental noise and will benefit from new analysis methods that aid interpretation.

Provided that perturbation-driven structural changes are not too large (i.e. the data remain isomorphous[11]), the calculation of difference electron density (DED) maps can be used to reveal changes in the electron density between a new dataset, which we refer to here as the derivative dataset, and an unperturbed native dataset. DED maps are used to identify regions of structural change and to extrapolate Fourier coefficients to which new coordinates can be refined[12–14]. Any derivative structure factor **F'** is a superposition of the

initial native structure factor (**F**) and the one related to the perturbed state of interest ($\mathbf{F}^{pr}$), where by perturbed state we mean the pure structure of interest, i.e. ligand-bound, time-activated, *etc.*:

$$\mathbf{F'} = f \times \mathbf{F}^{pr} + (1 - f) \times \mathbf{F}$$

here f is the occupancy of the perturbed state and we indicate complex structure factors in bold. Combined with measurement noise, this partial occupancy (often below 30%) means that dataset-dependent changes can be small, making DED map features difficult to interpret. Specific communities have developed different approaches to tackle this issue: time-resolved crystallographers use weights (w) that reduce the amplitudes of difference structure factors (w x [F'_{obs} - F_{obs}]) if they have large experimental error and are deemed to be outliers[12,15,16]. The Xtrapol8 program[12] makes a variety of such weighting schemes available to users, together with occupancy estimation strategies. For crystallographic fragment screening, the PanDDA suite[17] has introduced an objective procedure to identify density associated with partially-occupied ligands. PanDDA looks for regions with statistically significant excess density as compared to reference, ligand-free datasets. While both strategies can identify weak signals in DED maps, experimental maps frequently remain dominated by noise (Figure 1). Moreover, while amplitude weighting schemes are powerful, they require the selection of appropriate weighting parameters, and this is left to the user's discretion. One common example is adjusting the outlier rejection term in k-weighting[12]. This manual intervention requires extensive trial-and-error that is based on visual inspection of maps and can introduce user bias.

Methods that further automate DED map estimation and increase the interpretability of low-occupancy density in an unbiased way are therefore desirable[18–20]; such tools will reduce user bias, provide faster feedback during time-resolved experiments or ligand screening campaigns, and allow accurate analysis of low-occupancy species where existing methods fail.

For this reason, we aim to develop methods that improve the signal-to-noise ratio in DED maps. To do so, we find it essential to establish an objective, quantitative, simple measure of map quality to compare different map generation strategies. Work in this direction is greatly aided by insights from other fields as well as open-source software that enables the development and distribution of new crystallographic analyses[21–23].

We therefore first focus on identifying a suitable and reliable statistic that reports the level of noise present in a difference density map. We evaluate statistics that measure deviations between the distribution of map voxel values and a Gaussian distribution as indicators of DED map quality, ultimately favoring negentropy. Negentropy is routinely applied as a measure of non-Gaussianity in independent component analysis (ICA)[24–26]. We show that maximizing negentropy is an effective approach for selecting parameters in models that aim to denoise DED maps.

We use this insight to propose a new way to denoise DED maps through total variation (TV) minimization. In crystallographic DED maps, we expect *a priori* that the signals of interest should consist of local regions of smoothly-varying signal where atoms have moved, appeared, or disappeared. The rest of the map should be empty. When such signal is corrupted by additive white Gaussian noise, the noise contribution dominates at high spatial frequencies. TV minimization is an established technique in image processing[27] that aims to clarify the true signal by reducing these fluctuations. The total variation is simply defined as the sum of the changes from each voxel to all neighboring voxels. TV denoising seeks a new map that minimizes these changes while remaining as faithful as possible to the original signal[27]. Therefore, TV denoising, applied as a density modification technique to a DED map, will search for a new map that is similar to the given input map but attempts to set noisy, non-signal regions to be a constant, most likely zero.

We apply TV denoising[28] to DED maps and use negentropy maximization to select a regularization parameter that effectively trades off between fidelity to the original map and a smoother result. For three distinct case studies, we show that a single pass of TV denoising boosts the signal-to-noise ratio of DED maps and improves interpretability. We also find that an iterative application of TV denoising can be used to estimate phases for the derivative **F'** dataset, with corresponding further improvement in map quality.

Finally, we demonstrate how TV-denoised maps can be used to obtain extrapolated maps of the perturbed state and also refine multi-state coordinates (combined reference and

perturbed state) to an **F'** dataset, revealing protein backbone and sidechain rearrangements in a fragment-bound COVID-19 main protease (M^{Pro}) structure[29] that were not previously discernible.

We compile this analysis as an open source python package, *meteor*: *m*ap *e*nhancement *t*ools for *e*phemeral *o*ccupancy *r*efinement. *meteor* includes code to generate difference maps, select weighting parameters based on map negentropy, and apply an appropriate TV denoising protocol. Importantly, our observations on using ~~of the use of~~ map negentropy and TV denoising to improve difference density signals are compatible and stackable with existing methods and suites that deal with low-occupancy species in crystal datasets.

Results

Negentropy Reports on the Interpretability of a Difference Map

We start by considering the voxel-value distribution for a difference map. In an ideal, error-free DED map where the atoms move large distances, we expect the distribution to be bimodal in nature, with a positive and a negative mode and otherwise zero-valued voxels (Figure 1(a)). In the more realistic case where positive and negative peaks from different motions overlay and partially cancel out, the density distribution may no longer be strictly bimodal but should remain non-Gaussian (Figure 1(b-c)). Noise, by the central limit

theorem, can be expected to be an additive Gaussian contribution to the difference map signal. Empirically, we indeed observe that the distribution of voxel values for experimental maps is notably more Gaussian than that of synthetic maps calculated with no errors, suggesting that noise contributions dominate the former (Figure 1(d)). We therefore postulate that looking for deviations from Gaussianity in the voxel-value distribution may enable maximization of DED features that are not noise.

Skewness, kurtosis, and negentropy are well-known measures for non-Gaussianity[24,25,30]. To investigate the suitability of these statistics as indicators of difference map quality, we extend the 1D model from Figure 1(a) by adding Gaussian noise to the bimodal signal (Supporting Note S1). We evaluate different signal-to-noise ratios and find that the negentropy decreases monotonically with the addition of noise, while kurtosis and skewness do not (Figure S1). On the basis of this test, we proceed with the proposal that negentropy could be a useful metric to evaluate the signal-to-noise ratio in difference density maps.

We observe, first of all, that by maximizing the negentropy we can ~~maximization can be used to~~ optimize parameters in the commonly used k-weighting[12,16] amplitude modification scheme (Figure S2). We note, however, that, even after k-weighting, the voxel value distribution for experimental difference maps often remains markedly Gaussian (see below) and hypothesize that further denoising could yield important improvements.

Total Variation Denoising Facilitates Interpretation of Weak Signals in Difference Maps

One such case where even the amplitude k-weighted DED map is extremely noisy is the 100 ps *trans*-to-*cis* photoisomerization of the Cl-rsEGFP2 protein chromophore reported in Fadini *et al.*[31] (PDB ID 8A6G) and shown in Figure 1(d). This dataset captures a spatially localized, light-induced change where the *cis* photoproduct coordinates are well-known from ground state synchrotron structures[32] and is therefore an excellent example for methods development. We use it here to evaluate how TV denoising can assist in the interpretation of noisy maps.

Our hypothesis is that total variation denoising should remove unwanted high-frequency noise, while preserving the signal features of difference density that vary more smoothly. A necessary step in TV denoising is to choose the degree of smoothing, dictated by a regularization parameter, λ (Supporting Note S2). To gain intuition into the effect of different choices of regularization parameter, we compare the original k-weighted Cl-rsEGFP2 DED map to two denoised maps, where we set λ manually based on visual interpretability of the resulting DED map (Figure 2). TV denoising with a moderate level of regularization (λ = 0.008) produces a map with stronger signals on the protein chromophore and removes much of the noise from the original map. An overly-aggressive denoising (λ = 0.02), leads to a less interpretable, overly smooth map (Figure 2(b)). To support this subjective assessment, we plot the power spectra of the DED maps, which show the amplitude of each spatial frequency component in the map. The power spectra show that the moderate value of λ

(0.008) restores signal and recovers the expected resolution-dependent behavior of a noise-free synthetic map, while the higher λ (0.02) leads to a power spectrum that deviates from the "ground truth" spectrum computed from the synthetic example.

We conclude that TV denoising is a promising method for suppressing noise in DED maps, but would benefit from a robust, objective, and automated choice of the regularization parameter. To test whether maximizing the negentropy of DED maps as a function of λ could effectively serve this purpose, we simulate a *trans*-to-*cis* difference map that includes additive noise and track map negentropy while screening a range of λ values (Figure S3). We observe that the value of λ that maximizes negentropy closely matches the value that maximizes the real-space correlation coefficient between the denoised map and the noise-free synthetic map (Figure S3(c)). We therefore propose that the regularization parameter λ can be chosen by finding the value that generates the highest negentropy difference map (Figure 2(c)).

While the TV denoising step modifies the Fourier amplitudes of the map, it also alters the Fourier phases (Figure S4). This is notable, as the phases traditionally used in DED maps are approximate: the **F'** phases, which have a one-to-one correspondence to the perturbed state phases ~~and~~ are unknown. These phases are usually approximated by the native dataset phases coming from a well-characterized reference model (ϕ_c)[11]. This assumption, however, approximately halves the signal-to-noise ratio in the corresponding map as compared to what would be obtained if the perturbed state phases were known[33], an

undesirable effect, particularly when the perturbed signal is weakened by partial occupancy. Our observations suggest that TV denoising, which modifies the phases, may partially correct for this approximation.

This insight points to the potential use of TV denoising for iterative phase improvement, employing a cycle similar to that used for solvent flattening in crystallography[34] or phase retrieval in coherent diffractive imaging[35,36]. Inspired by this prior work, we use the set of TV-denoised difference structure factors to better estimate the phases for the derivative dataset, **F'**, through an iterative algorithm that we name iterative-TV (it-TV) (see Methods and Figure 3). We again first validate the it-TV approach on a synthetic noisy map, where the noise-free, ground truth map is known (Supporting Note S3 and Figure S3). In Figure 3, we show the experimental Cl-rsEGFP2 map before and after it-TV, together with the cumulative phase change and negentropy values for the difference maps generated at each iteration. Map negentropy reaches a stable positive value, having started below 10^{-4} for the original experimental map, and the final it-TV map contains much stronger density on the protein chromophore. This first result indicates that application of TV denoising as a density modification approach could find improved phase estimates for low occupancy states in DED maps.

Test Cases Demonstrate the Power of Negentropy-Guided TV Denoising to Recover Time-Resolved and Ligand-Binding Signals

To benchmark the single-pass and iterative TV denoising techniques, we select three distinct science cases:

- the 100 ps time-resolved crystallography Cl-rsEGFP2 dataset from Fadini *et al*.[31] (PDB ID 8A6G)
- the example of M^{pro} bound to tegafur (fragment SW7-401, PDB ID: 7AWR), identified as a potential binder to an allosteric site with a modeled occupancy of 0.50[29]
- the example of M^{pro} bound to a small electrophilic fragment (U1G, PDB ID: 5RGO) with a modeled occupancy of 0.42[37].

For Cl-rsEGFP2 and the M^{pro}-tegafur complex, we compute difference maps using the sets of observed amplitudes for the available derivative and native datasets ($F'_{obs} - F_{obs}$). For the M^{pro}-U1G complex, we illustrate the case where the difference map is computed between the observed amplitudes and the ones calculated from the model ($F_{obs} - F_{calc}$), when a reference native dataset is not available. Both M^{pro} structures score very poorly for ligand goodness of fit to experimental data in their respective PDB depositions[38], suggesting that the ligand density could benefit from further improvement.

The first two columns in Figure 4 show the effect of TV denoising on DED maps for our three test cases: (a) Cl-rsEGFP2 (b) M^{pro}-tegafur complex (c) M^{pro}-U1G complex. Features like the outline of the chlorophenolate ring from the Cl-rsEGFP2 *cis* chromophore are more easily identifiable within the protein structure and more chemically interpretable: at ±3 rms, a large

negative feature is visible on the imidazolinone ring after TV denoising, and the positive density on the *cis* chlorophenolate ring extends to four carbons and the hydroxyl group. Relevant signals on the chromophore appear significantly stronger following TV denoising. For instance, the positive and negative peaks on the chlorine atom reach ±11 rms in the TV-denoised map, compared to ±5 rms in the original map. For the M^{pro}-U1G complex, the denoising removes a large portion of uninterpretable density next to the ligand. In each case, visual improvements in the denoised maps are supported by increased negentropy for the voxel value distributions. In the Cl-rsEGFP2 case, the map negentropy increases from less than 10^{-4} to 0.059 (for comparison, the Cl-rsEGFP2 noise-free synthetic map shown in Figure 1(c) is associated with a negentropy value of 0.35). An analysis of the changes introduced by the denoising step reveals that the largest modifications involve high resolution structure factor amplitudes and phases (Figure S4). Finally, supporting the statement that the denoising step removes noise from the maps, the histograms and probability plots show that voxel value distributions are less Gaussian after TV denoising (Figure S5).

The it-TV maps for our three test cases can be found in the third column in Figure 4. These show clear differences that can be explicitly assigned to molecular structure. They are all characterized by an increased negentropy and less Gaussian voxel-value distributions (Figure S7) compared to their respective originals and single-pass TV maps. Note, for example, that the Cl-rsEGFP2 it-TV map in Figure 4 shows a clear outline of the chlorophenolate ring from the *cis* photoproduct, even though *cis* phases calculated from an atomic model are never introduced. Similarly, in the M^{pro}-tegafur complex map, an almost-

complete outline of the fragment's fluorouracil ring can be seen at 3.5 rms without ever using the fragment structure in the analysis. For the M^{pro}-U1G F_{obs} - F_{calc} case, it-TV increases the strength of positive signals on the side chain atoms and ligand. Once more, the strongest modifications by the algorithm occur in higher resolution shells (Figure S4), effectively recovering the underlying high resolution signal from the starting map.

Denoised Difference Maps Can Guide the Refinement of New Coordinates

Because DED maps do not require a model of the perturbed state, they provide unbiased and immediate information about the structural changes that occur after a perturbation, such as ligand binding or light-induced protein motion. This is invaluable to understand if the experiment has been successful and propose models that explain the results. The ultimate aim of most crystallographic experiments, however, is to interpret these changes chemically i.e. in terms of the positions of atoms and bonds. This requires an atomic model of the perturbed state, which can be refined using an extrapolated map. The extrapolated map should ideally reveal the electron density of the perturbed state without contributions from the reference state, making it easier to identify perturbation-specific structural change[12,17,39].

To produce an extrapolated map, an accurate estimate of the perturbed state occupancy is essential: once an estimate for the occupancy is known, the extrapolated map can be computed by performing an appropriate addition in real space (between a reference map

and a difference map) or in reciprocal space (between reference structure factors and difference structure factors)[40].

During this procedure, TV-denoised maps are useful for ~~in~~ identifying the region of the difference signal to focus on for initial occupancy estimation (Figure S8). They can also improve the extrapolated map resolution: Figure 5(a) displays the extrapolated map obtained by real-space addition between the reference M^{pro} state (PDB ID 7AR6) map and the M^{pro}-tegafur it-TV map (see Methods). The map shows a clear shift of the backbone and sidechains in the ligand-binding pocket and can be used to refine new atomic coordinates for the ligand-bound state. This low-occupancy conformation was not observable in the original PanDDA event map (Figure S10(a)). We can refine a multi-state model of ligand bound and unbound states directly to the ligand-bound dataset. Figure 5(b) compares the model from this refinement with the deposited structure for the ligand-bound state (PDB ID 7AWR)[29]. We note that refining a single model to the ligand-bound dataset results in atomic positions that are effectively an average between the two states in the multi-state model and are less representative of the underlying system.

An additional use for it-TV maps in the process of modeling perturbed-state coordinates is to suggest an initial ligand pose for refinement. We use the it-TV denoised map for the M^{pro}-U1G complex to manually remodel the ligand binding pose (Figure 4) and improve model fit to the data compared to both the deposited structure and a simple re-refinement of the deposited coordinates (Figure S11).

Discussion

Exciting applications of crystallography, such as time-resolved mechanistic studies or high-throughput fragment screens[3], struggle with the challenge of analyzing signal from low-occupancy species, which is at the same level as the noise in the data. Therefore, methods to measure and suppress noise in these datasets can unlock new scientific opportunities.

To this end, we find negentropy to be an effective reporter of difference map quality when seeking optimal amplitude weighting and denoising parameters. Negentropy maximization alone, however, cannot be used for map denoising without other restrictions to the procedure. Consider a putative DED map with half of the voxels randomly set to +x and half to -x. Such a map will have a very high negentropy, but no information content. Instead, negentropy, which measures how noise-like a signal is, should be thought of as a measure of the signal-to-noise ratio in a map. It can tell if the map contains signal, but not if that signal is faithful to any crystallographic or biological reality. Therefore, negentropy should only be used to evaluate competing DED maps generated by methods that can maintain fidelity to the original crystallographic data. As examples here, we show the need to choose the k parameter in k-weighting or the λ parameter in TV denoising, where negentropy maximization replaces manual selection, improving automation and reducing user bias.

We hypothesized that extending a single pass of TV denoising (single-TV) to an iterative density modification algorithm (it-TV) could improve estimates of perturbed-state phases, potentially doubling the signal-to-noise ratio in the difference structure factors set compared to the native-phase approximation used in F_{obs}-F_{obs} DED maps. Across all three test cases, the it-TV output maps exhibit greater chemical interpretability and higher negentropy than those produced by a single pass of TV denoising, suggesting that iterative refinement adds significant value. Our current it-TV implementation, based on the Gerchberg-Saxton iterative projection algorithm[41], achieves this improvement effectively; however, it may be susceptible to local minima in the presence of experimental noise. To address this, a feedback-based algorithm, such as Feinup's hybrid-input output or its descendants[35], could potentially enhance it-TV robustness against noise-induced errors.

The largest gains from TV denoising occur for F'_{obs}-F_{obs} maps, which are expected to benefit the most from the replacement of the single-phase approximation in it-TV and also contain weaker signals compared to F_{obs}-F_{calc} maps. We therefore anticipate that our denoising methods will be the most helpful for interpreting datasets from low-yield time-resolved studies or ligands that are weakly bound. Future versions of DED map denoising could incorporate more sophisticated models for perturbed-state phases or for the errors introduced by the experimental measurement and reference-state coordinates[42,43]; an important additional point is that TV minimization-based techniques will filter high-frequency noise but will not treat scaling errors or other systematic differences between datasets, which difference maps are sensitive to. Deep learning models, such as

convolutional neural networks (CNNs), have the capacity to learn more complex patterns and features from data and have been shown to perform better than traditional methods in image restoration and denoising tasks[44–46], making them potential candidates for future map denoising algorithms. Another advantage of such models is that, once trained, they do not rely on iteration and can therefore quickly process inputs. Nonetheless, it is important to note that the TV denoising tools we present here are fast, computationally cheap, and can be run from most workstations without the need to train an additional set of parameters.

The field of macromolecular crystallography is moving towards novel and technically ambitious data collection strategies: cutting-edge experiments are now aimed at capturing transient reaction intermediates, leveraging high-throughput facilities for small molecule screens, and studying the conformational variability that underlies protein dynamics. Analysis methods need to match these advances with increased sensitivity, robustness to noise, and improved automation. The negentropy metric and total variation denoising techniques reduce human bias and noise in the generation of difference density maps, as well as enhance signal by proposing new phase estimates for low-occupancy species. We show that our analysis particularly improves data interpretation in cases that struggle with weaker signals. For these reasons we believe such treatment of difference maps could unlock the study of minorly-populated states that are currently discarded in fragment screens or not attempted in time-resolved studies. The use of negentropy as a way of systematically scoring difference maps should also be useful for quick decision-making

during online experiments and for processing large crystallographic datasets from high-throughput beamlines.

Methods

Total Variation Denoising. To conduct TV denoising on a difference map, we employ Chambolle's algorithm[28] implemented in scikit-image[47]. TV denoising is performed in real-space on difference map arrays. To fit the λ parameter, map negentropy is minimized using golden section search[48,49]. For the skimage.restoration.denoise_tv_chambolle function, the tolerance for the stop criterions is set to $5x10^{-8}$ and the maximum number of iterations to 50.

Iterative Total Variation Denoising Algorithm. Native (F_{obs}) and derivative (F'_{obs}) state amplitudes are provided as input, together with a reference model. A difference map is computed using the phases calculated from the reference model and the difference structure factor amplitudes (F'_{obs} - F_{obs}). This starting difference map is denoised using Chambolle's TV algorithm as described above after a small λ value optimization. The TV denoised map is then inverse Fourier-transformed to obtain complex difference structure factors. These Fourier differences are projected onto the set of input amplitudes and phases to obtain an updated estimate of the latent perturbed state (Figure S6). These in turn are Fourier transformed back to real space for TV denoising, and the procedure is iterated until the average phase change no longer increases by more than 10^{-3}. The default in *meteor* is then to k-weight and TV denoise the it-TV map to produce a final output.

Perturbed State Extrapolation. To demonstrate how TV-denoised maps can be used in the process of extrapolating perturbed-state density, we generate two different types of extrapolated maps (Figure 5 and Figures S8-S10), described below. First, to estimate a perturbed state occupancy, we re-implement the background subtraction strategy used by PanDDA[17]. This procedure involves identifying a local region of change caused by the perturbation and finding an occupancy value that can produce a map that maximally differs from its reference in such localized region but is similar to the reference elsewhere.

Extrapolation for perturbed state structure factors ($\mathbf{F}^{pr}$) or density (ρ^{pr}) can be carried out in real or reciprocal space.

$$\text{Reciprocal:} \quad \mathbf{F}^{pr} = F_{obs} + \alpha^{-1} w\Delta F \times e^{i\phi c}$$

$$\text{Real:} \quad \rho^{pr} = \rho^{ref} + \alpha^{-1}\Delta\rho$$

where $\Delta F = F'_{obs} - F_{obs}$, w represents error-based amplitude weights, ϕ_c are the calculated phases from the reference model, and α is related to the true perturbed state occupancy (f) by $f = 2\alpha$ when the constant phase approximation is used. To estimate α (with final choice written $\hat{\alpha}$), the strategy from PanDDA[17] finds the value that maximizes the difference in Pearson correlation coefficient (calculated from the reference state and the extrapolated map) between the entire protein and a specific local region of change. We showcase this for our photoisomerization example in Figure S8(a), where the local region is set as a sphere of

5 Å centered around the chromophore double bond. The entire protein is defined as a global region after a solvent mask is applied. For a range of values of $0 < \alpha < 1$, the Pearson correlation coefficient between the respective $\mathbf{F}^{pr}/\rho^{pr}$ map and the map obtained from the reference model is computed. The value that maximizes the difference between these two correlation coefficients is chosen as $\hat{\alpha}$ and the corresponding map is saved (Figure S8(b)).

For the M^{pro}-tegafur complex, we use the higher-negentropy it-TV map as $\Delta\rho$ for real space extrapolation and show the extrapolated result in Figure 5. We also carry out a reciprocal space extrapolation, which does not use new phases from either single-TV or it-TV but is guided by the denoised maps in the choice of local region. The corresponding map is shown in Figure S10(b). Figure S9 contains the analysis to estimate $\hat{\alpha}$ for these two extrapolation types.

Multi-State Model Refinement. As outlined by Pearce *et al.*[50], we combine the models through the use of alternate conformers, we constrain the occupancy of the atoms within each state to be fixed, and we set the sum of the occupancies between states to be equal to 1. As a starting occupancy for the ligand-bound state we choose the value that minimizes R-free (Figure 5(b)), and proceed to refine B-factors and coordinates with phenix.refine[51]. For the R-factors reported in the table in Figure 5(b), the anisotropic B-factors are removed from the originally deposited models (PDB ID 7AR6/7AWR) and the B-factors are re-refined with phenix.refine. This ensures a consistent procedure to compare the metrics obtained with the multi-state model refinement.

Code Availability. Negentropy-driven k-weighting, single pass TV denoising, and the it-TV algorithm are implemented as executables in the *meteor* package, requiring initial MTZ files and a reference model as input. *meteor* is written in python and depends on the GEMMI[22], reciprocalspaceship[23], numpy[52], scikit-learn[53], and panda[54] packages. The code is available at: https://github.com/rs-station/meteor.

Acknowledgements

AF and JJvT acknowledge funding from the Imperial College President's PhD Scholarship and the Biotechnology and Biological Sciences Research Council (BBSRC) (BB/P00752X/1). TJL acknowledges the support of the Helmholtz Association through a YIG award. VA and TJL were supported by the Cluster of Excellence "Advanced Imaging of Matter," Deutsche Forschungsgemeinschaft (DFG), EXC 2056, project ID 390715994. We acknowledge Nick Pearce for discussion on the background subtraction estimation. We would also like to deeply thank Kevin Dalton for scientific exchange and for his help integrating *meteor* into the *Reciprocal Space Station* platform.

Author Contributions

Conceptualization: AF, TJL; Methodology: AF, TJL; Investigation: AF, TJL, VA; Visualization: AF, TJL, VA, JJvT; Funding acquisition: JJvT; Writing: AF, TJL, VA, JJvT.

Competing Interests

TJL is a shareholder of CHARM Therapeutics.

References

1. Orville, A. M. Recent results in time resolved serial femtosecond crystallography at XFELs. *Curr. Opin. Struct. Biol.* **65**, 193–208 (2020).
2. Pearson, A. R. & Mehrabi, P. Serial synchrotron crystallography for time-resolved structural biology. *Curr. Opin. Struct. Biol.* **65**, 168–174 (2020).
3. Budziszewski, G. R., Snell, M. E., Wright, T. R., Lynch, M. L. & Bowman, S. E. J. High-Throughput Screening to Obtain Crystal Hits for Protein Crystallography. *J. Vis. Exp.* e65211 (2023). doi:10.3791/65211
4. Keedy, D. A. *et al.* Mapping the conformational landscape of a dynamic enzyme by multitemperature and XFEL crystallography. *Elife* **4**, (2015).
5. Hekstra, D. R. *et al.* Electric-field-stimulated protein mechanics. *Nature* **540**, 400–405 (2016).
6. Moffat, K. Time-resolved crystallography and protein design: signalling photoreceptors and optogenetics. *Philos. Trans. R. Soc. B Biol. Sci.* **369**, 20130568–20130568 (2014).
7. Käck, H., Gibson, K. J., Lindqvist, Y. & Schneider, G. Snapshot of a phosphorylated substrate intermediate by kinetic crystallography. *Proc. Natl. Acad. Sci.* **95**, 5495–5500 (1998).
8. Calvey, G. D., Katz, A. M., Schaffer, C. B. & Pollack, L. Mixing injector enables time-resolved crystallography with high hit rate at X-ray free electron lasers. *Struct. Dyn.* **3**, 054301 (2016).
9. Schmidt, M. Reaction Initiation in Enzyme Crystals by Diffusion of Substrate. *Crystals* **10**, 116 (2020).
10. Olmos, J. L. *et al.* Enzyme intermediates captured 'on the fly 'by mix-and-inject serial crystallography. *BMC Biol.* **16**, 59 (2018).
11. Rould, M. A. & Carter, C. W. Isomorphous Difference Methods. *Methods Enzymol.* **374**, 145–163 (2003).
12. De Zitter, E., Coquelle, N., Oeser, P., Barends, T. R. M. & Colletier, J.-P. Xtrapol8 enables automatic elucidation of low-occupancy intermediate-states in crystallographic studies. *Commun. Biol.* **5**, 640 (2022).
13. Pandey, S. *et al.* Time-resolved serial femtosecond crystallography at the European XFEL. *Nat.*

Methods **17**, 73–78 (2020).

14. Pande, K. *et al.* Femtosecond structural dynamics drives the trans/cis isomerization in photoactive yellow protein. *Science (80-.).* **352**, 725–729 (2016).
15. Ursby, T. & Bourgeois, D. Improved Estimation of Structure-Factor Difference Amplitudesfrom Poorly Accurate Data. *Acta Crystallogr. Sect. A Found. Crystallogr.* **53**, 564–575 (1997).
16. Ren, Z. *et al.* A Molecular Movie at 1.8 Å Resolution Displays the Photocycle of Photoactive Yellow Protein, a Eubacterial Blue-Light Receptor, from Nanoseconds to Seconds †. *Biochemistry* **40**, 13788–13801 (2001).
17. Pearce, N. M. *et al.* A multi-crystal method for extracting obscured crystallographic states from conventionally uninterpretable electron density. *Nat. Commun.* **8**, 15123 (2017).
18. Carolan, C. G. & Lamzin, V. S. Automated identification of crystallographic ligands using sparse-density representations. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **70**, 1844–1853 (2014).
19. Smart, O. S. *et al.* Validation of ligands in macromolecular structures determined by X-ray crystallography. *Acta Crystallogr. Sect. D, Struct. Biol.* **74**, 228–236 (2018).
20. Brändén, G. & Neutze, R. Advances and challenges in time-resolved macromolecular crystallography. *Science (80-.).* **373**, (2021).
21. Adams, P. D. *et al.* PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **66**, 213–221 (2010).
22. Wojdyr, M. GEMMI: A library for structural biology. *J. Open Source Softw.* **7**, 4200 (2022).
23. Greisman, J. B., Dalton, K. M. & Hekstra, D. R. reciprocalspaceship : a Python library for crystallographic data analysis. *J. Appl. Crystallogr.* **54**, 1521–1529 (2021).
24. Jutten, C. & Herault, J. Blind separation of sources, part I: An adaptive algorithm based on neuromimetic architecture. *Signal Processing* **24**, 1–10 (1991).
25. Comon, P. Independent component analysis, A new concept? *Signal Processing* **36**, 287–314 (1994).
26. Hyvärinen, A. & Oja, E. Independent Component Analysis: Algorithms and Applications. *Neural Networks* **13**, 411–430 (2000).
27. Rudin, L. I., Osher, S. & Fatemi, E. Nonlinear total variation based noise removal algorithms. *Phys. D*

Nonlinear Phenom. **60**, 259–268 (1992).

28. Chambolle, A. An Algorithm for Total Variation Minimization and Applications. *J. Math. Imaging Vis.* **20**, 89–97 (2004).

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

30. Ben-David, A., Hausegger, S. Von & Jackson, A. D. Skewness and kurtosis as indicators of non-Gaussianity in galactic foreground maps. *J. Cosmol. Astropart. Phys.* **2015**, 019–019 (2015).

31. Fadini, A. *et al.* Serial Femtosecond Crystallography Reveals that Photoactivation in a Fluorescent Protein Proceeds via the Hula Twist Mechanism. *J. Am. Chem. Soc.* **145**, 15796–15808 (2023).

32. Chang, J., Romei, M. G. & Boxer, S. G. Structural Evidence of Photoisomerization Pathways in Fluorescent Proteins. *J. Am. Chem. Soc.* **141**, 15504–15508 (2019).

33. Henderson, R. & Moffat, J. K. The difference Fourier technique in protein crystallography: errors and their treatment. *Acta Crystallogr. Sect. B Struct. Crystallogr. Cryst. Chem.* **27**, 1414–1420 (1971).

34. Terwilliger, T. C. Reciprocal-space solvent flattening. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **55**, 1863–1871 (1999).

35. Marchesini, S. Invited Article: A unified evaluation of iterative projection algorithms for phase retrieval. *Rev. Sci. Instrum.* **78**, 011301 (2007).

36. Butola, M., Rajora, S. & Khare, K. Complexity-guided Fourier phase retrieval from noisy data. *J. Opt. Soc. Am. A* **38**, 488 (2021).

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

38. Shao, C. *et al.* Simplified quality assessment for small-molecule ligands in the Protein Data Bank. *Structure* **30**, 252-262.e4 (2022).

39. Genick, U. K., Soltis, S. M., Kuhn, P., Canestrelli, I. L. & Getzoff, E. D. Structure at 0.85 Åresolution of an early protein photocycle intermediate. *Nature* **392**, 206–209 (1998).

40. Schmidt, M. Practical considerations for the analysis of time-resolved x-ray data. *Struct. Dyn.* **10**, (2023).

41. Gerchberg, R. A practical algorithm for the determination of phase from image and diffraction plane pictures. *Optik (Stuttg)*. (1972).

42. Read, R. J. Improved Fourier coefficients for maps using phases from partial structures with errors. *Acta Crystallogr. Sect. A Found. Crystallogr.* **42**, 140–149 (1986).

43. Read, R. J. & McCoy, A. J. A log-likelihood-gain intensity target for crystallographic phasing that accounts for experimental error. *Acta Crystallogr. Sect. D Struct. Biol.* **72**, 375–387 (2016).

44. Lim, B., Son, S., Kim, H., Nah, S. & Lee, K. M. Enhanced Deep Residual Networks for Single Image Super-Resolution. in *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* **2017-July**, 1132–1140 (IEEE, 2017).

45. Nah, S., Kim, T. H. & Lee, K. M. Deep Multi-scale Convolutional Neural Network for Dynamic Scene Deblurring. in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* **2017-Janua**, 257–265 (IEEE, 2017).

46. Zhang, K., Zuo, W., Chen, Y., Meng, D. & Zhang, L. Beyond a Gaussian Denoiser: Residual Learning of Deep CNN for Image Denoising. *IEEE Trans. Image Process.* **26**, 3142–3155 (2017).

47. van der Walt, S. *et al.* scikit-image: image processing in Python. *PeerJ* **2**, e453 (2014).

48. Kiefer, J. Sequential minimax search for a maximum. *Proc. Am. Math. Soc.* **4**, 502–506 (1953).

49. Virtanen, P. *et al.* SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* **17**, 261–272 (2020).

50. Pearce, N. M., Krojer, T. & von Delft, F. Proper modelling of ligand binding requires an ensemble of bound and unbound states. *Acta Crystallogr. Sect. D Struct. Biol.* **73**, 256–266 (2017).

51. Afonine, P. V. *et al.* Towards automated crystallographic structure refinement with phenix.refine. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **68**, 352–367 (2012).

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

53. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).

54. The pandas development team. pandas-dev/pandas: Pandas. (2023). doi:10.5281/ZENODO.7658911

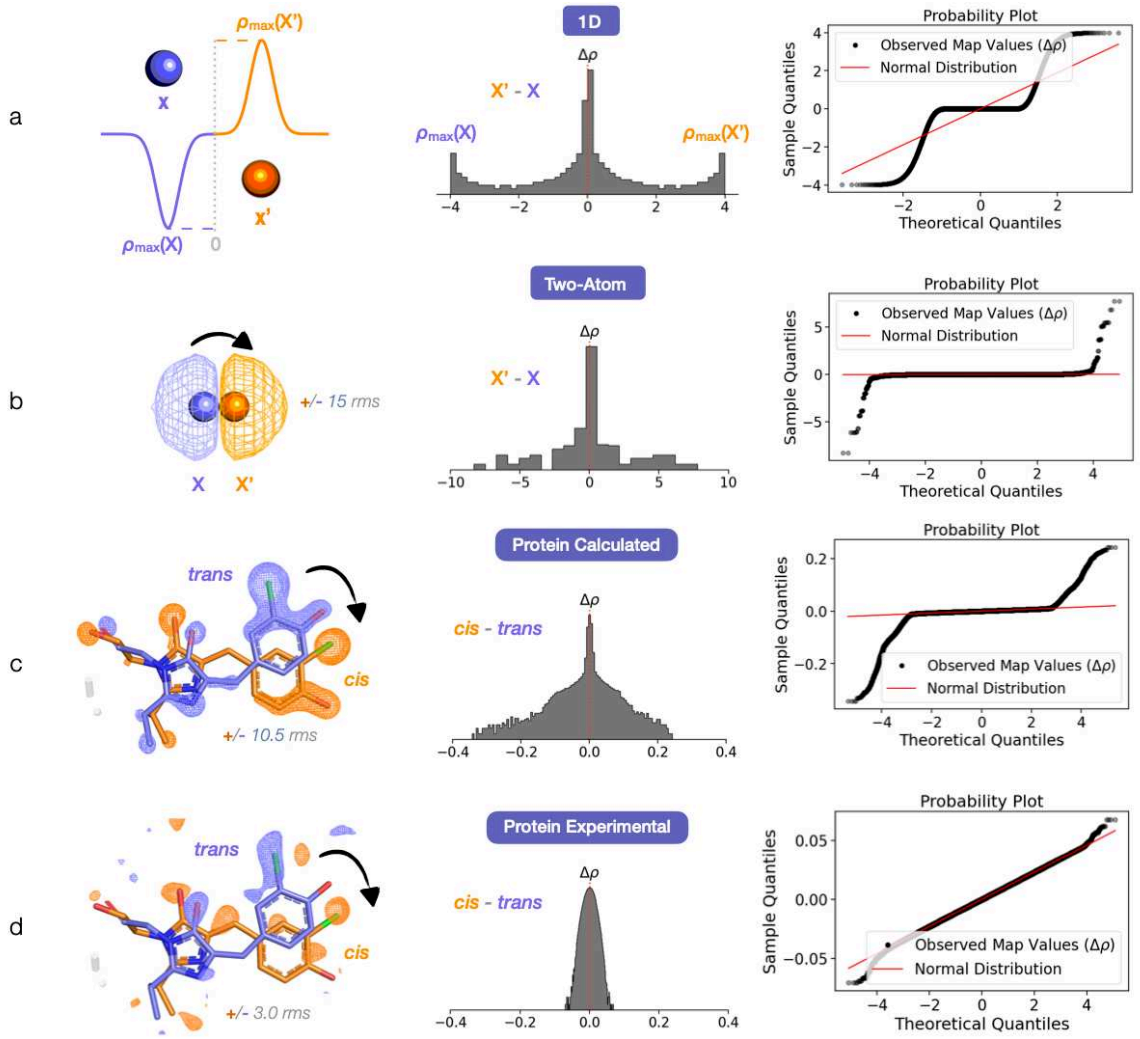


Figure 1: The voxel distribution of a difference map reflects its signal-to-noise ratio. (a) A 1D signal with positive and negative peaks at positions at X' and X respectively can serve as a simple model for a difference map where an atom moves from X to X'. The corresponding

distribution for the difference map voxel values ($\Delta\rho$) is non-Gaussian, with a mode at 0 and two modes at ρ_{max}(X) and ρ_{max}(X') (the histogram is plotted with a log-scale on the y-axis). On the right column, the normal probability plot compares the map voxel value distribution to a perfect Gaussian: the observed data (sample quantiles) is ordered and plotted against the expected values of the ordered statistics for a sample from a standard normal distribution of the same size as the data (theoretical quantiles). The red line denotes the behavior for a sample that is normally distributed. Deviations from this straight line indicate deviations from a Gaussian distribution. **(b)** A more realistic case is that of a carbon atom translating from X to X' in three-dimensional space. Here we show the case where the positive and negative electron density signals from the displacement of the atom overlay and partially cancel. The resulting difference map voxel value distribution ($\Delta\rho$) is no longer strictly bimodal but remains non-Gaussian. **(c)** Deviation from a Gaussian distribution for the histogram of voxel values is also noticeable for the calculated difference map of a known light-induced protein structural change: we display a noise-free synthetic difference electron density (DED) map for the 100 ps *trans*-to-*cis* photoisomerization of the Cl-rsEGFP2 protein chromophore (PDB ID 8A6G). To match experimental data[31], the map is calculated by converting 13% of the reference dark *trans* population to *cis*. As the dark state already contains 14% of the *cis* species, the final state (in orange here) is at 27% *cis* occupancy. We show the calculated difference map at a contour level that is comparable to the experimental map shown below in **(d)**. In contrast to the calculated map, the experimental map[31] is dominated by Gaussian noise, even after error-based amplitude k-weighting. This is

apparent from the histogram and probability plot of the map voxel value distribution shown on the right.

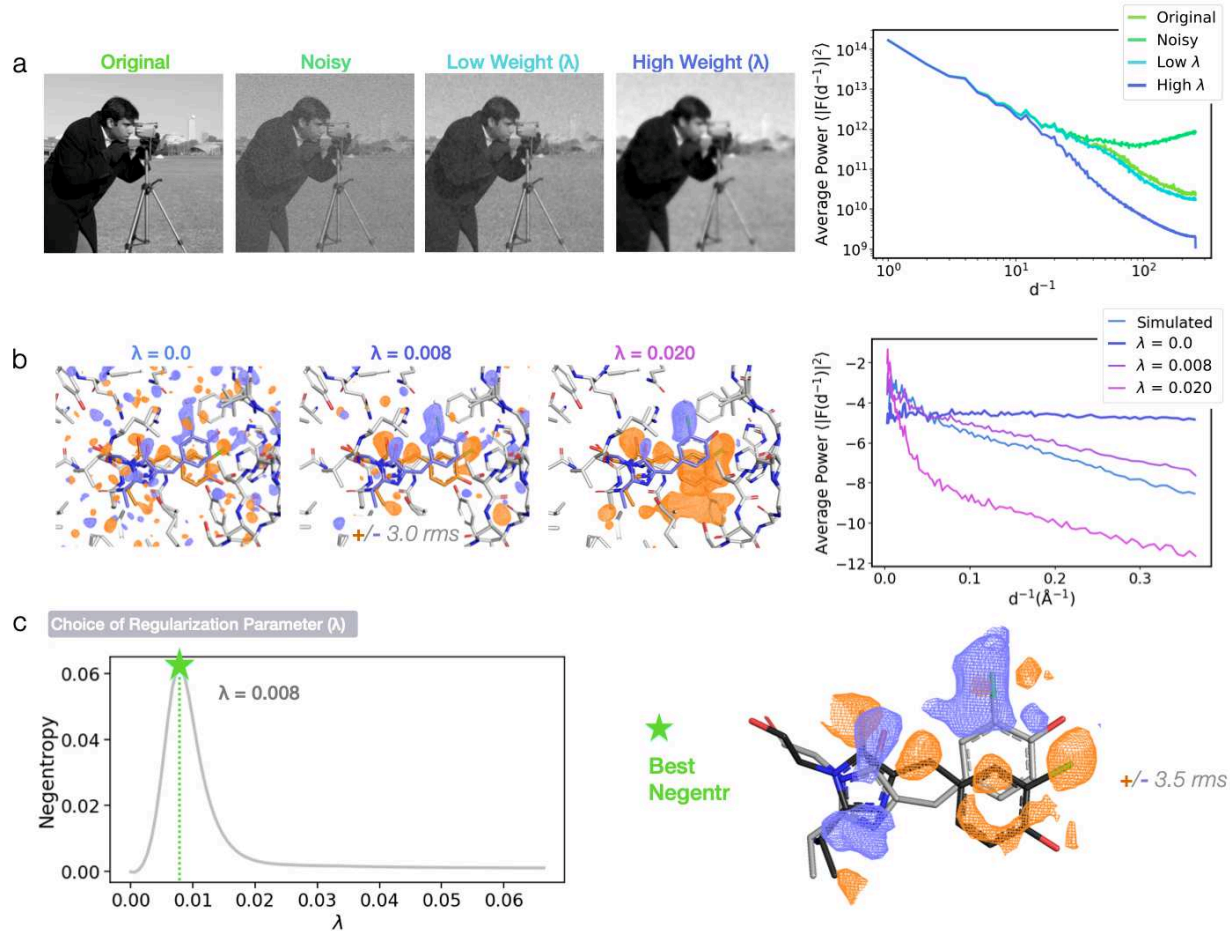


Figure 2: Total variation (TV) minimization can effectively denoise experimental DED maps. (a) TV denoising is used in signal processing to remove spurious noise from data. This is exemplified here by artificially adding white noise (sampled from a normal distribution with a standard deviation ≈10% of the maximum pixel value in the image) to the cameraman image and applying Chambolle's total variation minimization algorithm to retrieve the underlying signal[28]. A regularization parameter λ determines the degree of denoising: a value of λ that is too high will produce an image that is overly "smoothed" compared to the original. On the right, the power spectrum for each image illustrates the effect of the added noise and of subsequent TV denoising to recover signal. **(b)** TV denoising using two different manually

chosen values of λ is shown for the weighted Cl-rsEGFP2 *trans*-to-*cis* photoisomerization DED map, along with the original experimental map (λ = 0.0). As for the 2D images in (a), power spectra show that, while the starting experimental map contains little signal, an appropriately regularized denoising yields map coefficients that approach those for the simulated map displayed in Figure 1(c). **(c)** To refine an appropriate value of λ, we denoise the experimental Cl-rsEGFP2 map for a range of regularization levels (0 ≤ λ ≤ 0.08) and identify the λ value that maximizes map negentropy.

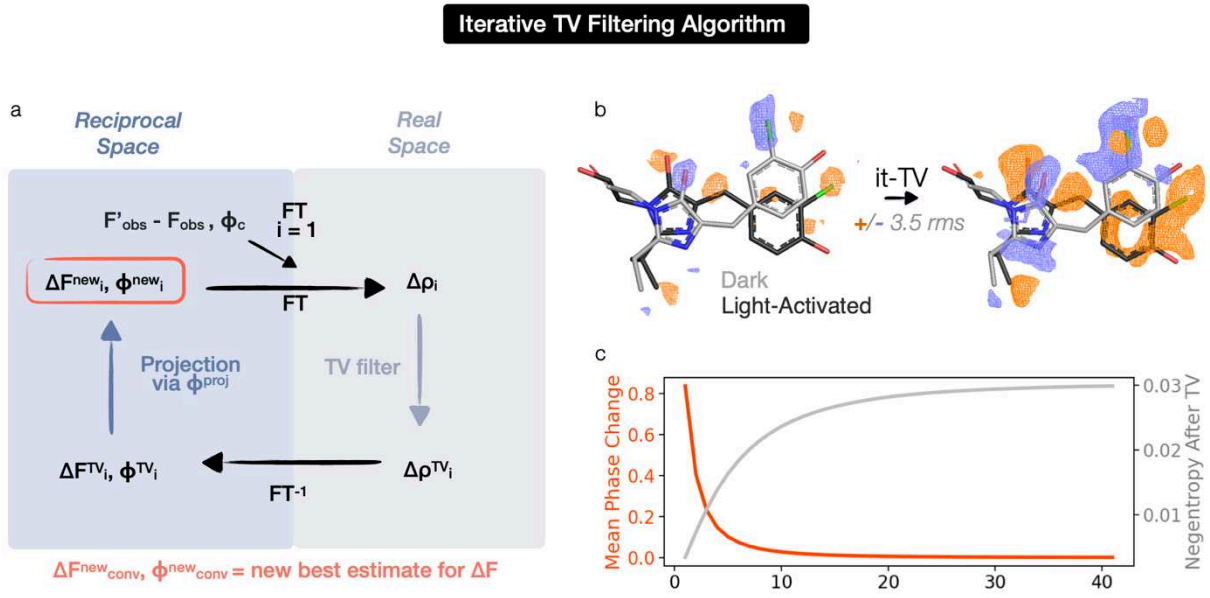


Figure 3: An iterative TV minimization algorithm estimates the phases of the structure factor contribution from low occupancy states. (a) The unknown perturbed state phases are estimated through an iterative procedure (it-TV, Figure S6). The initial difference map is

computed from the observed derivative and native amplitudes (F'_{obs} and F_{obs}) and the phases from the reference state model (ϕ_c). A step of difference density (Δρ) TV denoising is followed by an inverse Fourier transform (FT) to reciprocal space. We then project the $\mathbf{\Delta F^{TV}}$ set onto the phase circle with magnitude F'_{obs} (see also Figure S6) as a way of finding a new phase estimate for **F'**. The new phases are used in the next iteration. Iterations are run until convergence (see Methods). **(b)** The experimental Cl-rsEGFP2 map is shown before and after it-TV (in purple and orange for negative and positive density respectively, ±3 rms), together with the cumulative phase change and negentropy values for the difference maps generated at each iteration in **(c).** The mean phase change at each iteration is plotted in red. The negentropy as a function of iteration is plotted in gray.

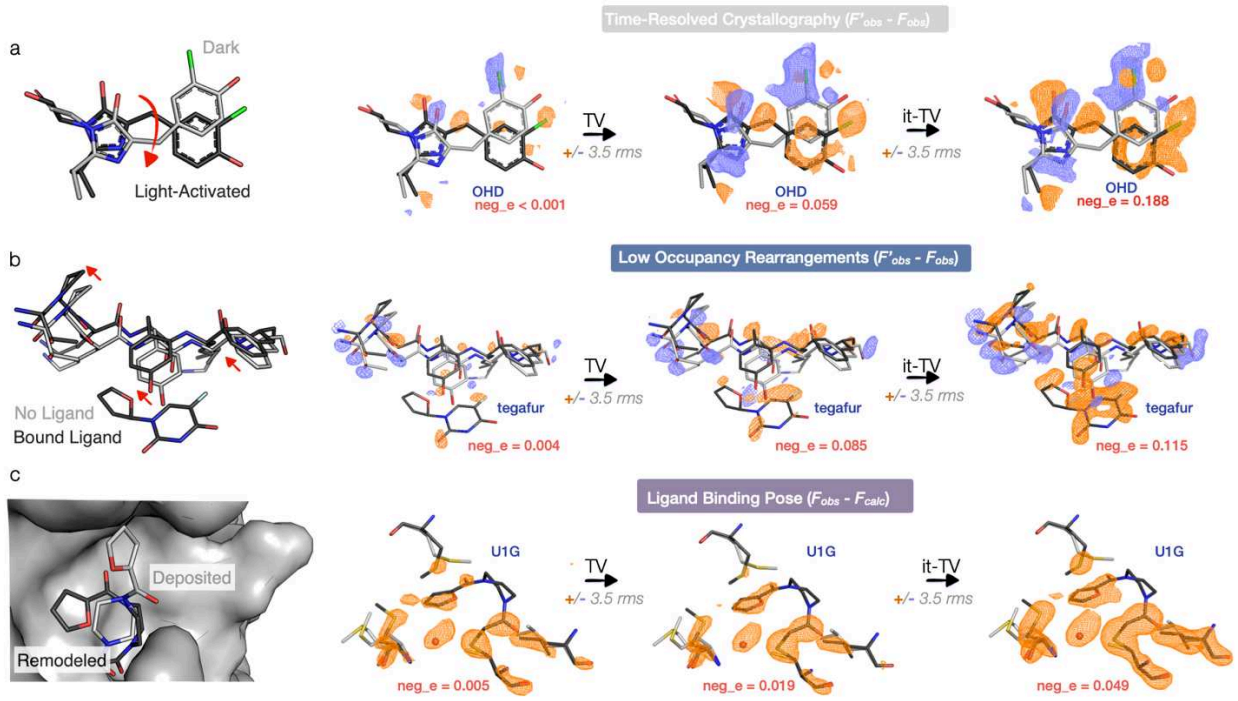
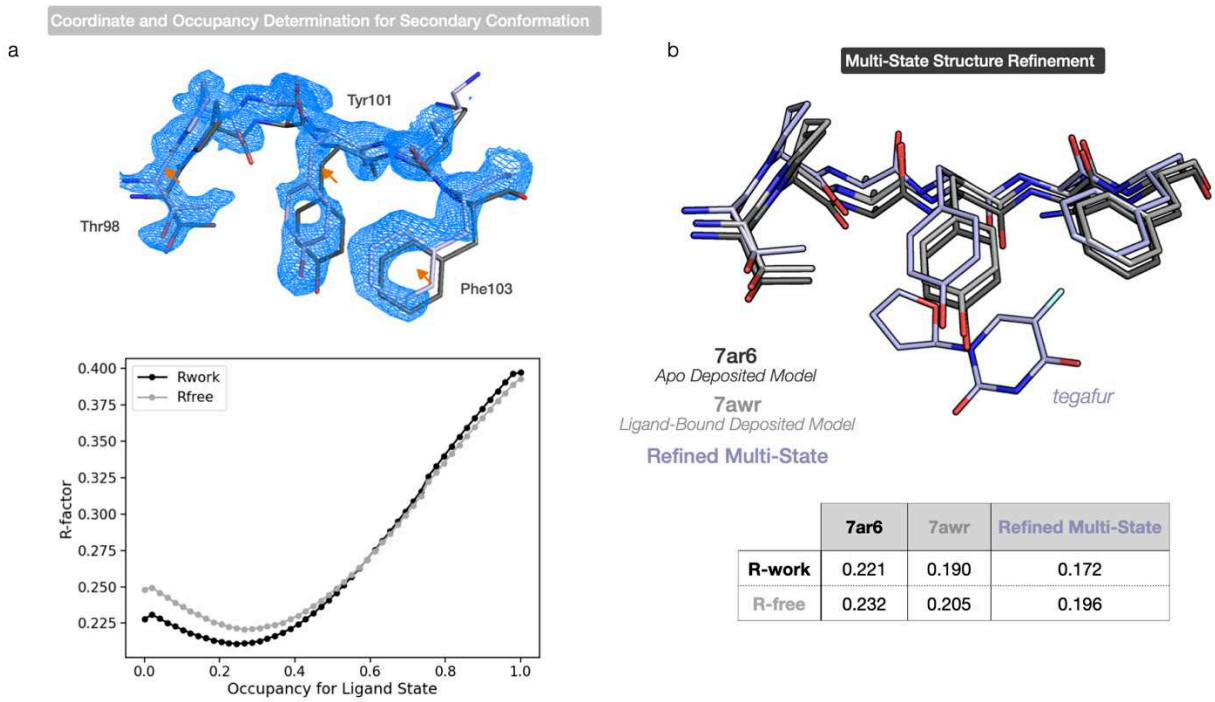


Figure 4: Test cases demonstrate the power of negentropy-guided TV denoising to recover time-resolved and ligand-binding signals. We show single-pass TV denoising and iterative-TV maps with their associated negentropy values for our three test datasets: the 100 ps photoisomerization of the OHD chromophore in the Cl-rsEGFP2 protein (PDB ID 8A6G) (a), the example of M^{pro} bound to tegafur (PDB ID 7AWR), which was identified as a potential binder to an allosteric site with a modeled occupancy of 0.50 (b), and the example of the M^{pro}-U1G complex (PDB ID 5RGO), with a modeled fragment occupancy of 0.42 (c). Reference state structures are shown in gray, while structures that we refined to the perturbed dataset are shown in black. Ligand outlines become stronger and more chemically interpretable for all three test cases. For the M^{pro}-tegafur complex, there are

strong signals for rearrangements of side chains and backbone atoms near the ligand binding pocket. For the M^{pro}-U1G complex, the it-TV map suggests an alternative fragment pose. The largest gains from the denoising step occur when the initial map is close to being normally distributed (as for the signal on the OHD chromophore in Cl-rsEGFP2 or in the M^{pro}-tegafur complex). We show maps at ±3.5 rms, which we find is an appropriate rms cutoff to highlight signal for these examples. However, because TV-denoised maps are intentionally non-Gaussian, the second moment (rms) alone does not fully capture the distribution of voxel values. It's therefore important to keep in mind that common visualization thresholds, like ±3 rms, may not directly relate when trying to compare denoised and standard maps.



	7ar6	7awr	Refined Multi-State
R-work	0.221	0.190	0.172
R-free	0.232	0.205	0.196

Figure 5: Denoised difference maps can guide the refinement of new coordinates and uncover low-occupancy conformations. (a) Extrapolated map for the M^{pro}-tegafur complex. This is obtained through addition of the it-TV map shown in Figure 4 to the $2mF_o$-DF_c map from the reference state (PDB ID 7AR6, black). We highlight the rearrangement of the backbone and Thr98/Tyr101/Phr103 sidechain atoms close to the ligand binding pocket. We refine coordinates for the bound state to this extrapolated map (light blue) and screen R-factor values to initiate the occupancy of the ligand-bound state in a multi-state model, choosing an occupancy of 0.29 for initial refinement. **(b)** Binding pocket coordinates from the deposited ligand bound (PDB ID 7AWR) and unbound (PDB ID 7AR6) models and the multi-state model generated here are shown. The table reports final refinement R-factors for the deposited ligand bound and unbound models and the multi-state model when compared to the ligand-bound dataset. For the refined multi-state model, the reference chain is overlaid with the 7AR6 model.