Assessing the applicability of Bayesian inference for merging small molecule microED data

Huanghao Mai,[a] Ariana Peck,[b] Kevin M. Dalton,[c] Lygia Silva de Moraes,[a] Jessica E. Burch,[a] Frédéric Poitevin[c] and Hosea M. Nelson [a]*

[a]*Division of Chemistry and Chemical Engineering, California Institute of Technology, CA 91125, United States,* [b]*Chan Zuckerberg Imaging Institute, Redwood City, CA 94065, United States, and* [c]*SLAC National Accelerator Laboratory, Menlo Park, CA 94025, United States. E-mail: hosea@caltech.edu*

Abstract

Microcrystal electron diffraction (MicroED) is an emerging technique for characterizing small molecule structures from nanoscale crystals. Merging data from multiple crystals is a particularly challenging step in the microED workflow. A common practice is to manually curate datasets and apply scaling programs conventionally utilized in rotational X-ray diffraction (XRD), but this could be time-consuming and risks introducing human bias in data analysis. Recently, a Bayesian inference program named Careless (Dalton *et al.*, 2022) has demonstrated excellent performance in merging macromolecular XRD data. Here, the applicability of Careless to small molecule microED data is evaluated and an investigation of the impact of dataset curation is performed. Benchmarking against XDS/XSCALE shows that Careless is an effective complementary approach that merges data to a higher $CC_{1/2}$ value at high resolution.

Furthermore, merging outcomes are not significantly improved by curating datasets either manually or with an automated extension to Careless, cautioning against the common practice of manual dataset curation.

1. Introduction

Structural characterization is critical for understanding small molecule properties and advancing research in chemistry fields, including organic chemistry, natural product chemistry, and drug discovery. For decades, single-crystal X-ray diffraction (SCXRD) has been the gold standard for determining the bond connectivity of molecules with high precision. However, the need to obtain large (10–100 μm) single crystals has severely limited the applicability of SCXRD. To overcome this limitation, new methods have been developed to solve structures from smaller crystals (Smith *et al.*, 2012; Gemmi *et al.*, 2019). Synchrotrons now offer micro-focus beamlines that can reduce the beam width to match micron-sized crystals (Grimes *et al.*, 2018). X-ray free electron laser (XFEL) facilities enable studying even smaller crystals using serial crystallography by delivering extremely bright, femtosecond-long pulses that have the additional benefit of outrunning radiation damage (Chapman *et al.*, 2011). However, few structures of small molecules have been elucidated using these resources, and limited XFEL sources and micro-focus beamlines render these techniques unsuitable for routine analysis compared to in-house instruments, which offer rapid turnaround times.

Microcrystal electron diffraction (microED) (Shi *et al.*, 2013), also known as continuous rotational electron diffraction (cRED) (Wang *et al.*, 2017) and a sub-method of 3D electron diffraction (3D ED), provides a powerful alternative for small molecule structure determination (Jones *et al.*, 2018; Gruene *et al.*, 2018; Wang *et al.*, 2017; Kim *et al.*, 2021; Bruhn *et al.*, 2021; Park *et al.*, 2022; E. Gorelik *et al.*, 2022). Compared to photons, electrons interact more strongly with matter, enabling this technique to mea-

sure diffraction signals from nanoscale crystals at sub-Ångstrom resolution. For small molecules, such nanocrystals are generally far easier to obtain than micron-sized or larger crystals required by other techniques, and they can be found in seemingly amorphous powders as well as crude natural products extracts (Jones *et al.*, 2018; Delgadillo *et al.*, 2024). MicroED is also advantageous due to the broad availability of transmission electron microscopes (TEM) and the potential to use this technique with increasing throughput (Cichocka *et al.*, 2018; Wang *et al.*, 2019; Lightowler *et al.*, 2023; Delgadillo *et al.*, 2024; Unge *et al.*, 2024). However, the widespread adoption of microED calls for improvements in data processing (Powell *et al.*, 2021). Current microED workflows typically leverage software such as XDS which was originally developed for rotational XRD experiments (Kabsch, 2010*b*), but several steps of microED data processing still require time-consuming manual intervention.

A particularly challenging step of microED data processing is data merging. In SCXRD, merging refers to the process where measured symmetry-equivalent reflection intensities are reduced to a set of unique values after being scaled to correct systematic errors, a process also known as data reduction. The scaled and merged reflections can then be phased to solve the structure. In small molecule crystallography, *ab initio* phasing is a standard practice, which requires accurate estimates of the structure factor amplitudes at 1.2 Å or higher resolution (Sheldrick, 1990) from the merging output. Compared to SCXRD, merging in microED is inherently more challenging. The background noise is higher due to non-negligible diffuse and inelastic scattering (Nannenga & Gonen, 2014). Dynamical scattering events increase the variation among intensities that should be theoretically equivalent, effectively contributing to the errors in conventional merging (Palatinus *et al.*, 2015; Khouchen *et al.*, 2023). Moreover, most TEMs restrict the accessible tilt range to less than $\pm 70^{\circ}$, resulting in a missing wedge of information where data cannot be measured. Although crystallographic symmetry

should in principle overcome this low completeness, in practice, data from multiple crystals often need to be merged due to radiation damage that compromises intensities in a way that is reflection-dependent and non-monotonic in dose (Saha *et al.*, 2024).

Multi-crystal merging is often a trial-and-error process conducted under the assumption that few crystals are sufficiently isomorphous and of high enough quality to yield correct *ab initio* phasing solutions and acceptable refinement statistics. Thus, even though high multiplicity from redundant measurements is considered helpful in macromolecular XRD (Karplus & Diederichs, 2015), empirically, many microED small molecule structures have been solved by merging only several crystals out of the tens to hundreds collected. In other emerging fields of diffraction experiments such as serial femtosecond crystallography (SFX) from multiple crystals, specialized software has been developed (Uervirojnangkoorn *et al.*, 2015; White *et al.*, 2016) to merge data that conventional methods find challenging. Nevertheless, this is a computationally expensive approach. Small-wedge serial crystallography has prompted iterative approaches (Beilsten-Edmands *et al.*, 2020; Gildea *et al.*, 2022) that combine preliminary clustering and then outlier rejection (Assmann *et al.*, 2016). In the microED field, clustering-based heuristics (Giordano *et al.*, 2012; Foadi *et al.*, 2013; Yamashita *et al.*, 2018; Wang *et al.*, 2019) and brute-force enumeration of dataset combinations (Unge *et al.*, 2024) have been deployed, but the manual curation of datasets remains a dominant practice.

A recent machine learning (ML)-based merging program, Careless, promises a unifying framework through Bayesian inference for any type of diffraction experiment (Dalton *et al.*, 2022). It has been successfully applied to a variety of macromolecular XRD experiments but has not been validated in small molecule or microED studies. Compared to conventional programs, Careless has the potential to be adapted for multi-crystal merging with minimal human bias in selecting datasets or setting

cutoff values for clustering and filtering datasets. Here, we reprocess 17 molecules from previous studies on natural products and pharmaceutical compounds to test the applicability of Careless in comparison to a conventional scaling and merging program XDS/XSCALE (Kabsch, 2010*b*). In addition, the popular practice of manual dataset curation motivates us to explore an extension to the ML algorithm that automates this process. We compare the merging outcomes of dataset curation, whether manual or automated, with a naive merging of all datasets. Finally, we show how well results from different merging protocols are translated to *ab initio* structures using standard phasing programs, and present recommendations to microED practitioners.

2. Results

2.1. Careless as an alternative merging tool for microED data

To assess whether Bayesian inference generalizes well to microED data processing, we compile existing microED datasets comprised of a diverse set of 17 small molecules (Fig. S1), from simple cases such as calcium oxalate (Fig. S1-**15**) (Delgadillo *et al.*, 2024) to challenging cases such as fischerin (Fig. S1-**5**) (Kim *et al.*, 2021). They span a range of crystallographic complexity, including 7 space groups and unit cell volumes from 10^3 to 10^4 Å^3 (Table S1). 6 of the molecules were solved from single-crystal datasets processed by XDS, and the rest were solved after individually processing all datasets in XDS and then merging a manually curated subset in XSCALE (Jones *et al.*, 2018; Kim *et al.*, 2021; Burch *et al.*, 2021; Chhetri *et al.*, 2023; Delgadillo *et al.*, 2024; Abad *et al.*, 2024).

Careless is first evaluated using the single-crystal datasets and manually curated datasets in multi-crystal merging. Previously published structures (Fig. S1 and Table S1) are used as reference structures to assess merging performance. For consistency, in multi-crystal merging, we adhere to the same dataset curation manually done by

the authors of the reference structures, and the effect of manual dataset curation is examined in the next section (2.2). Merging outcomes are evaluated by examining the internal consistency of the intensities measured by $CC_{1/2}$ as well as $CC_{F_oF_c}$, which indicates the accuracy relative to calculated structure factors from the reference structure (Methods section 4.5). To compare the average performance of different merging protocols, we use a stringent criterion of whether the 95% bootstrap confidence intervals (CI) for the mean overlap or not to assess statistically significant differences.

Compared to XDS/XSCALE (Kabsch, 2010*a*; Kabsch, 2010*b*), a conventional merging software, Careless merging yields lower overall $CC_{1/2}$ but comparable overall $CC_{F_oF_c}$ for single-crystal merging (Table 1) and multi-crystal merging (Table 3). This suggests that even though the Careless merging output is less precise among symmetry-equivalent measurements, it is not necessarily less accurate. In the highest resolution bin of each dataset, Careless performance is on average similar to XDS/XSCALE as indicated by the 95% CIs of the $CC_{1/2}$ and $CC_{F_oF_C}$ of all 17 molecules (Table 2).

2.2. Effects of dataset curation on multi-crystal merging

After demonstrating that Careless achieves comparable accuracy given curated datasets, we next sought to automate dataset curation within Careless. This was motivated by current practices in the microED field, where multi-crystal merging is commonly performed on datasets selected by manual inspection of the completeness, $CC_{1/2}$, R_{merge}, and other summary statistics. Manual curation of datasets could be time-consuming with a large number of datasets collected and risk introducing human bias to data analysis. Nevertheless, a naive merge of all datasets may compromise data quality. For effective comparisons, we first evaluate the impact of manual curation relative to the baseline that omits dataset curation.

Naively merging all datasets that could be indexed to the expected space group and unit cell maximizes the completeness (Fig. 1a) and multiplicity of the data. Using the naive merging as a baseline for comparison, manual dataset curation has the opposite effect on $CC_{1/2}$ for the two merging regimes studied: it is beneficial for XDS/XSCALE but harmful for Careless (Fig. 1b). Within each merging program, the effect of dataset curation is statistically significant for the overall $CC_{1/2}$ but less pronounced in the highest resolution bin. With uncurated merging, Careless achieves similar overall $CC_{1/2}$ with XDS/XSCALE but has the additional benefit of significantly better $CC_{1/2}$ in the highest resolution bin (Fig. 1b and Table 4).

Among the four merging protocols — XDS/XSCALE vs. Careless using manually curated vs. all datasets, the common practice of merging manually curated datasets by XDS/XSCALE still gives the best overall $CC_{1/2}$, while merging all datasets by Careless gives the best $CC_{1/2}$ in the highest resolution bin. Despite the different performances according to $CC_{1/2}$, on average, $CC_{F_oF_c}$ is minimally affected both overall and at high resolution regardless of the merging protocol used (Fig. 1c). This suggests that the accuracy of merging is not significantly improved by manual curation.

Although we find that manual dataset curation does not benefit Careless, it is still of interest to investigate whether a fully automated curation would improve the outcome. This approach, referred to as MC-Careless for Multi-Crystal Careless, uses ML principles to learn an optimal weighting among datasets to account for the variability of data quality across datasets collected from different crystals (section 4.2). This weight modulates the effective uncertainty of the intensities during model training for multi-crystal merging and is optimized jointly with the structure factor amplitudes (Fig. 4). It provides an alternative to the manual curation of datasets and reduces human bias in evaluating summary statistics and filtering datasets. Nevertheless, MC-Careless achieves similar performance on both $CC_{1/2}$ and $CC_{F_oF_c}$ to the original Careless that

naively merges all datasets (Fig. 1b and c). This result is consistent with the observation above that manual curation of datasets does not improve Careless merging outcomes.

2.3. Phasing and initial maps

Finally, we perform *ab initio* phasing on the outputs from all five merging protocols using SHELXT or SHELXD with the same phasing parameters that previously led to the preliminary solutions of the reference structures. In the microED field, *ab initio* phasing could be challenging, especially for large organic molecules lacking heavy atoms (Bruhn *et al.*, 2021). The preliminary structure from *ab initio* phasing is often corrupted by missing or extra atoms as well as mis-assignment of elements due to the difference between X-ray and electron scattering (Mott & Bragg, 1997; Dorset, 1996). Consequently, naive structural alignment with the reference structures for quantitative comparisons is difficult. Here, we manually classify phasing as successful or not by inspecting the overall connectivity or recognizable fragments for cases with disorder, and present several visual examples of the raw phasing structures.

As expected from the analysis of $CC_{F_oF_c}$ in the section above, XDS/XSCALE merging outcomes could be successfully phased for all 17 molecules regardless of dataset curation (Fig. 2). Even though Careless merges data with similar $CC_{F_oF_c}$ to XDS/XSCALE, the outputs could be more challenging for conventional phasing programs. In Careless, scaling and merging are jointly performed, which requires estimating structure factors independent of scaling by physical factors (Dalton *et al.*, 2022). A consequence of this modeling approach is that structure factors are outputted on an arbitrary scale that is flat across resolution bins (Dalton *et al.*, 2022), whereas conventional programs output intensities that decay over increasing resolution. Nevertheless, correct structural information could still be retrieved from Careless outputs

in most cases (Fig. 2). At identical contour levels, the maps from phasing are often sharper than those from XDS/XSCALE merging (Fig. 3), which is likely because the parallel inference of scaling and structure factors in Careless has an analogous effect to B-factor sharpening (DeLaBarre & Brunger, 2006).

The only phasing solution from Careless outputs that contains almost no recognizable fragments is the naive merging of fischerin (Fig. S1-**5**) datasets (Kim *et al.*, 2021). Despite achieving comparable overall $CC_{F_oF_c}$ and further improvement in the highest resolution shell compared to the curated merging by Careless or either protocol of XDS/XSCALE merging (Fig. S2), the phasing result is visually worse (Fig. S3). This was a particularly difficult case where flexible molecular conformations and preferred orientation of the crystals demanded recrystallization and more than 6 months of manual processing in previous work (Kim *et al.*, 2021).

3. Discussion and conclusion

The successful generalization from macromolecular XRD studies (Dalton *et al.*, 2022) to small molecule microED data in this work highlights the flexibility and impact of Bayesian inference in emerging structural studies. Through benchmarking against XDS/XSCALE, we also show that this approach has some benefits in merging small molecule microED data. Careless merges reflections to higher $CC_{1/2}$ at high resolution, and comparable accuracy with respect to the reference structure is achieved both overall and at high resolution. Moreover, for the examples presented here, dataset curation, whether manual or automated, is not necessary for Careless, as the naive merging of all datasets achieves the best $CC_{1/2}$, comparable $CC_{F_oF_c}$, and the highest completeness. Thus, merging by Careless eliminates an opportunity for human bias in data processing and maximally leverages information from all datasets. Even though automated curation by MC-Careless does not further improve merging outcomes, it

shows that Careless could be easily extended for future methods development.

For microED practitioners, we caution against the common practice of manually curating datasets. We find that $CC_{1/2}$ is elevated in XDS/XSCALE merging using dataset curation inherited from previous work, yet we see no significant differences in $CC_{F_oF_c}$ or the *ab initio* phasing structures. Our findings suggest that data quality is not always compromised when naively merging all datasets, indicating that useful signal can be missed in manual curation of datasets. For example, the three lowest overall $CC_{F_oF_c}$ from molecules Py-469 (Fig. S1-**4**) (Kim *et al.*, 2021), demethoxyviridin (Fig. S1-**1**) (Delgadillo *et al.*, 2024), and AMG7 (Fig. S1-**10**) (Burch *et al.*, 2021) are improved by more than 8% when including all datasets in XDS/XSCALE merging (Tables 3 and 4). From a practical perspective, the preliminary structures from *ab initio* phasing seem robust against dataset curation, although the impact on refinement statistics is beyond the scope of this work.

In conclusion, using experimental datasets from previous studies, we demonstrate that Careless could robustly merge microED and small molecule crystallography data. Careless could improve multi-crystal merging outcomes with reduced human bias and is a flexible framework for methods development in diffraction data processing. In most cases examined here, Careless outputs lead to similar preliminary structural solutions with sharpened initial maps compared to XDS/XSCALE outputs. For challenging cases, additional optimization of merging and phasing parameters might be necessary to obtain the correct phasing solutions. As Bayesian inference and other ML approaches have only recently been introduced to crystallography, continuing method developments are warranted. Additional case studies and future investigation in refinement outcomes may help improve the integration of Careless into existing data processing pipelines.

4. Methods

4.1. Merging algorithms

To contextualize the ML approach in Dalton *et al.* (2022) and our extension to automate dataset curation in Careless, we briefly describe the formalism of scaling and merging in crystallography. Readers are referred to Aldama *et al.* (2023) for a more detailed review.

4.1.1. Merging by weighted average: Conventionally, the true intensity $I_{\mathbf{h}}$ at Miller index $\mathbf{h}$ is estimated by computing the weighted average of redundant measurements across all images after correcting for systematic errors. This corresponds to the maximum likelihood estimate of the mean intensity. The functional form of the weights w determines the error model with normally-distributed being the most common choice. Each measurement $\hat{I}_{\mathbf{h},i}$ on image i is corrected by estimating a scaling factor $K_{\mathbf{h},i}$:

$$\hat{I}_{\mathbf{h},i} = K_{\mathbf{h},i} I_{\mathbf{h},i}. \tag{1}$$

Established programs in XRD data processing such as XDS (Kabsch, 2010*a*; Kabsch, 2010*b*) and AIMLESS (Evans & Murshudov, 2013) use sophisticated models to parameterize $K_{h,i}$ and minimize the least-squares loss for scaling and merging:

$$\Phi = \sum_{\mathbf{h}} \sum_{i} w_{\mathbf{h},i} (I_{\mathbf{h}} - \hat{I}_{\mathbf{h}} / K_{\mathbf{h},i})^2. \tag{2}$$

4.1.2. Merging by Bayesian inference: Careless works on the same premise in eq. (1) that systematic errors can be corrected by scaling $I_{\mathbf{h},i}$, but uses an alternative inference approach. Under the kinematical approximation, the true intensity is $I_{\mathbf{h},i} = F_{\mathbf{h}}^2$, where F denotes the structure factor amplitude. Careless uses variational Bayesian inference (Blei *et al.*, 2017; Jordan *et al.*, 1999) to reformulate merging as estimating $p(F, K|I)$, the posterior distribution of the scaling and structure factors conditioned

on the observed intensities. As $p(F, K|I)$ is generally intractable, it is approximated by a parametric surrogate function q. The standard modeling objective in variational Bayesian inference is to minimize the difference between q and $p(F, K|I)$ by maximizing the Evidence Lower BOund (ELBO) (Kingma & Welling, 2014). The ELBO typically consists of an expected log-likelihood term that encourages fitting q to the data and a Kullback–Leibler (KL) term as a regularization to penalize deviations from a prior distribution. The exact form used in Dalton *et al.* (2022) is:

$$\text{ELBO} = \mathbb{E}_q \left[\log p(I|F, K)\right] - D_{\text{KL}} \left[q_F \| p(F)\right] \tag{3}$$

where q is assumed to be factorizable,

$$p(F, K|I) \simeq \prod_{\mathbf{h}} \left[q_{F_\mathbf{h}} \prod_i q_{K_{\mathbf{h},i}} \right], \tag{4}$$

and q_K is further parameterized by a multi-layer perceptron (MLP) that takes the metadata of observed reflections as the input. Parameters of q_K are optimized without regularization from a prior distribution. A modified version of the Wilson distribution — the intensity distribution if atoms are uniformly distributed within the unit cell (Wilson, 1949) — is used as the prior $p(F)$ for estimating the structure factor amplitudes independent of the scale (Dalton *et al.*, 2022).

4.2. Extending Careless for multi-crystal merging

The uncertainty of the observed intensity, σ_I, is important for estimating data quality (Evans, 2011; Evans & Murshudov, 2013) and directly affects merging in the conventional approach as described in eq. (2). In the variational Bayesian inference approach, σ_I also modulates the contribution of each measurement to the training loss. Specifically, the log-likelihood term in eq. (3) is a parametric distribution where the mean is $\hat{K}\hat{F}^2$ obtained from drawing Monte Carlo samples $\hat{F} \sim q_F$ and $\hat{K} \sim q_K$, and the standard deviation or scale in the case of Student's t-distribution, is the σ_I

estimated by integration programs.

In MC-Careless, to account for the different quality of each dataset in multi-crystal merging, we adjust the uncertainty of observed intensities σ_I inversely by a weight w that is sparsely parameterized as a categorical distribution q_w over the N crystals to merge (Fig. 4). The distribution is normalized such that w averages to 1 across all unmerged reflections. The contribution of a crystal to merging is decreased as w becomes smaller than 1, consequently increasing the uncertainty of observed intensities from that crystal. The modified training objective is:

$$\text{wELBO} = \mathbb{E}_q \left[\log p \left(I | F, K; \frac{\sigma_I}{w} \right) \right] - D_{\text{KL}} \left[q_F \| p(F) \right] - D_{\text{KL}} \left[q_w \| p(w) \right] \tag{5}$$

where w is learned jointly with F and K, and is regularized by a prior distribution $p(w)$. The prior distribution is the discrete uniform distribution that represents no adjustment to σ_I and equal weights among crystals as treated in the naive merging of all input datasets. Code that implements MC-Careless is available at `https://github.com/DorisMai/careless/tree/multi_xtal_sig`.

4.3. Data processing workflow with XDS/XSCALE

Each rotational diffraction movie was collected in SER format and converted to SMV as previously described (Hattne *et al.*, 2015). Spot finding, indexing, integration, and correcting/scaling are performed using XDS (Kabsch, 2010*b*). XDS is a standard crystallography program that has proven effective for small molecule microED data (Jones *et al.*, 2018), although other programs such as DIALS (Winter *et al.*, 2018; Clabbers *et al.*, 2018), Jana2020 (Petříček *et al.*, 2023), and CrysAlis$^{\text{Pro}}$ (Ito *et al.*, 2021) are also applicable. The instruction file for initial processing by XDS is generated using an in-house Python script (`https://github.com/jess-burch/microed`) for greater automation as previously described (Burch *et al.*, 2021; Delgadillo *et al.*, 2024). To benchmark merging performance with minimal confounding errors from other pro-

cessing steps, here all datasets are reprocessed by XDS using previously reported space group and unit cell parameters. XSCALE (Kabsch, 2010*a*; Kabsch, 2010*b*) is used to merge data from multiple crystals. The resolution cutoff from previous work is used whenever possible in the reprocessing but relaxed by 0.05 Å for 6β-hydroxyeremophilenolide (Fig. S1-**17**), calcium oxalate (Fig. S1-**15**), and peyssobaricanoside B (Fig. S1-**6**), by 0.1 Å for AMG3 (Fig. S1-**8**) and 4 (Fig. S1-**9**), and by 0.15 Å for AMG7 (Fig. S1-**10**) to reproduce phasing outcomes.

4.4. Data processing workflow with Careless

4.4.1. Data preprocessing: XDS_ASCII.HKL files from reprocessing as described above are converted to .mtz format using `careless.xds2mtz` before merging. Each dataset is then standardized such that the intensity I has unit variance. This is achieved by scaling the observed intensities $I'_{\mathbf{h},i}$ and uncertainty $\sigma'_{I_{\mathbf{h},i}}$ by k, where

$$k = \frac{1}{\sqrt{\frac{\sum_{\mathbf{h},i}\left(I'_{\mathbf{h},i} - \bar{I}'\right)^2}{N_{\text{unmerged}}}}}. \tag{6}$$

This standardization supports stable training. Unit cell parameters are averaged across all crystals to be merged.

4.4.2. Model training: Metadata features used for model training include the image number, resolution, and X/Y positions of each observed intensity on the image. For multi-dataset merging, intensities from all datasets are concatenated, and the index of the source dataset is supplied as an additional feature. A non-negative scaling factor is enforced during model training. Training steps are 30,000 and 50,000 for single-crystal and multi-crystal merging, respectively. Training for all cases can run on an NVIDIA Tesla P100 GPU in under 1.5 hours, with the exception of merging all 89 fischerin (Fig. S1-**5**) datasets which could take 7.5 hours.

4.4.3. Hyperparameter selection: The original training objective of Careless described by eq. (3) is approximated using Monte Carlo sampling with 1 sample per training step as the default:

$$\text{ELBO} \approx \sum_{s=1}^{S} \left[\sum_{i} \sum_{\mathbf{h}} \log p\left(I_{i,\mathbf{h}} | F_{\mathbf{h}}, K_{i,\mathbf{h}}, \sigma_I; \nu\right) - \sum_{\mathbf{h}} \left(\log q_F - \log p\left(F\right)\right) \right]. \quad (7)$$

We increase to $S = 20$ Monte-Carlo samples to improve convergence with a minimal increase in total training time. The log-likelihood term in eq. (7) is modeled as Student's t-distribution with the degree of freedom ν as a hyperparameter to adjust the sensitivity to outliers. This error model becomes a normal distribution as ν approaches infinity and becomes a Cauchy distribution when $\nu = 1$. We keep $\nu = 16$, which was found to be optimal by cross-validation in Dalton *et al.* (2022) and empirically robust by other users of Careless.

The relative weight between the log-likelihood term and the KL term in eq. (7) defaults to the average multiplicity of the datasets $m = N_{\text{unmerged}}/N_{\text{merged}}$. As of Careless version 0.3.4, this weight is adjustable through the hyperparameter λ_F. Empirically, we find that a small value of λ_F generally works well, possibly because observed intensities in small molecule 3D ED do not always obey ideal statistics described by the Wilson distribution (Fig. S4) which is used as a prior distribution in Careless as described in eq. (3). In this work, $\lambda_F = 0.01$ is used for all cases, except for fischerin (Fig. S1-**5**) datasets where the optimal value between 0.01 and 0.001 is chosen by cross-validation. In MC-Careless, we also introduce λ_w to adjust the relative weight of the second KL term in eq. (5) such that:

$$\begin{aligned} \text{wELBO} \approx \sum_{s} \Bigg[& \sum_{i} \sum_{\mathbf{h}} \log p\left(I_{i,\mathbf{h}} | F_{\mathbf{h}}, K_{i,\mathbf{h}}, \frac{\sigma_I}{w_{i,\mathbf{h}}}; \nu\right) \\ & - m\lambda_F \sum_{\mathbf{h}} \left(\log q_F - \log p\left(F\right)\right) \\ & - m\lambda_w \sum_{i} \sum_{i,\mathbf{h}} \left(\log q_w - \log p\left(w\right)\right) \Bigg]. \end{aligned} \quad (8)$$

The optimal value of λ_w is found over 0.001, 0.01, 0.1, 1, and 10 by cross-validation.

4.5. Evaluation of merging outcomes

Merging quality is assessed based on the following metrics: completeness, $CC_{1/2}$, and $CC_{F_oF_c}$. The $CC_{F_oF_c}$ metric is calculated as the uncertainty-weighted Pearson correlation coefficient between the estimated structure factor amplitude F_o from merging and the F_c calculated from the reference structures. The CCDC numbers of reference structures are available in Table S1 of the supplementary information. F_c is calculated using `gemmi`, accounting for electron form factors and anisotropic atomic displacement parameters. An additional non-negative global B-factor is fit when calculating $CC_{F_oF_c}$ because Careless outputs F_o on the same scale across resolution bins, unlike XDS/XSCALE and other conventional data reduction programs. The completeness and $CC_{1/2}$ are extracted from CORRECT.LP (single-crystal) or XSCALE.LP (multi-crystal) for XDS/XSCALE and from `careless.completeness` and `careless.cchalf` for Careless. Both the overall statistic and the statistic in the highest resolution bin are reported. Resolution bins are determined by default in XDS/XSCALE and in Careless to distribute reflections evenly across 10 bins.

4.6. Structure determination

Intensities merged by XDS/XSCALE are converted to SHELX format using the XDSCONV program. Intensities merged by Careless are scaled and reformatted by a separate Python script that uses `reciprocalspaceship` (Greisman *et al.*, 2021) to parse the .mtz output file from Careless. *Ab initio* phasing is then performed using SHELXT (Sheldrick, 2015) or SHELXD (Usón & Sheldrick, 1999). Phasing parameters are kept unchanged from previous processing that led to the reference structures. We run SHELXD on 32 CPUs for 15 minutes for all cases except for fischerin (Fig. S1-**5**) where the run time is extended to 1 hour. $2F_o - F_c$ maps from *ab initio* phasing are generated using the `shelx2map` program, and visualized in Pymol (Schrödinger,

LLC, 2022) with contouring at 1.5σ and carving at 1.2Å.

5. Data availability

MicroED data used in this work are available at `10.5281/zenodo.12775590`, `10.5281/zenodo.12797270`, `10.5281/zenodo.8206533`, `10.5281/zenodo.10059796`, `10.5281/zenodo.10059842`, and `10.5281/zenodo.10059864`, except for AMG3 (Fig. S1-**8**), AMG4 (Fig. S1-**9**), AMG7 (Fig. S1-**10**), AMG10 (Fig. S1-**13**), and AMG11 (Fig. S1-**11**), which are available upon request. Specific datasets used in single-crystal merging and manually curated multi-crystal merging are described in `curated_movie_id.csv` in `10.5281/zenodo.12775590`.

Acknowledgements

H.M. thanks Douglas C. Rees, Michael R. Sawaya, Jose A. Rodriguez, William M. Clemons, Stephen L. Mayo, and Vignesh C. Bhethanabotla for helpful discussions, Dmitry B. Eremin, Kunal K. Jha, and David A. Delgadillo for feedback on the manuscript, and David A. Delgadillo, Lee Joon Kim, and Christopher G. Jones for sharing raw microED data. Compounds AMG3 (Fig. S1-**8**), AMG4 (Fig. S1-**9**), AMG7 (Fig. S1-**10**), AMG10 (Fig. S1-**13**), and AMG11 (Fig. S1-**11**) were provided through the Amgen Process Development Academic Interface Team collaboration. This work is sponsored by the NSF Center for Computer-Assisted Synthesis, an NSF Center for Chemical Innovation (CHE-2202693). This work also used computational resources from the Resnick High Performance Computing Center, a facility supported by Resnick Sustainability Institute at the California Institute of Technology.

References

Abad, A. N. D., Seshadri, K., Ohashi, M., Delgadillo, D. A., de Moraes, L. S., Nagasawa, K. K., Liu, M., Johnson, S., Nelson, H. M. & Tang, Y. (2024). *Journal of the American Chemical Society*, **146**(21), 14672–14684. Publisher: American Chemical Society. https://doi.org/10.1021/jacs.4c02142

Aldama, L. A., Dalton, K. M. & Hekstra, D. R. (2023). *Acta Crystallographica Section D: Structural Biology*, **79**(9), 796–805. Publisher: International Union of Crystallography. https://journals.iucr.org/d/issues/2023/09/00/qi5002/

Assmann, G., Brehm, W. & Diederichs, K. (2016). *Journal of Applied Crystallography*, **49**(3), 1021–1028. Number: 3 Publisher: International Union of Crystallography. //scripts.iucr.org/cgi-bin/paper?zw5005

Beilsten-Edmands, J., Winter, G., Gildea, R., Parkhurst, J., Waterman, D. & Evans, G. (2020). *Acta Crystallographica Section D: Structural Biology*, **76**(4), 385–399. Number: 4 Publisher: International Union of Crystallography. https://journals.iucr.org/d/issues/2020/04/00/di5035/

Blei, D. M., Kucukelbir, A. & McAuliffe, J. D. (2017). *Journal of the American Statistical Association*, **112**(518), 859–877. ArXiv:1601.00670 [cs, stat]. http://arxiv.org/abs/1601.00670

Bruhn, J. F., Scapin, G., Cheng, A., Mercado, B. Q., Waterman, D. G., Ganesh, T., Dallakyan, S., Read, B. N., Nieusma, T., Lucier, K. W., Mayer, M. L., Chiang, N. J., Poweleit, N., McGilvray, P. T., Wilson, T. S., Mashore, M., Hennessy, C., Thomson, S., Wang, B., Potter, C. S. & Carragher, B. (2021). *Frontiers in Molecular Biosciences*, **8**, 648603. Publisher: Frontiers. https://www.frontiersin.org/articles/10.3389/fmolb.2021.648603

Burch, J. E., Smith, A. G., Caille, S., Walker, S. D., Wurz, R., Cee, V. J., Rodriguez, J., Gostovic, D., Quasdorf, K. & Nelson, H. M. (2021). Putting MicroED to the Test: An Account of the Evaluation of 30 Diverse Pharmaceutical Compounds. https://chemrxiv.org/engage/chemrxiv/article-details/61670e747d3da50c42f692b9

Chapman, H. N., Fromme, P., Barty, A., White, T. A., Kirian, R. A., Aquila, A., Hunter, M. S., Schulz, J., DePonte, D. P., Weierstall, U., Doak, R. B., Maia, F. R. N. C., Martin, A. V., Schlichting, I., Lomb, L., Coppola, N., Shoeman, R. L., Epp, S. W., Hartmann, R., Rolles, D., Rudenko, A., Foucar, L., Kimmel, N., Weidenspointner, G., Holl, P., Liang, M., Barthelmess, M., Caleman, C., Boutet, S., Bogan, M. J., Krzywinski, J., Bostedt, C., Bajt, S., Gumprecht, L., Rudek, B., Erk, B., Schmidt, C., Hömke, A., Reich, C., Pietschner, D., Strüder, L., Hauser, G., Gorke, H., Ullrich, J., Herrmann, S., Schaller, G., Schopper, F., Soltau, H., Kühnel, K.-U., Messerschmidt, M., Bozek, J. D., Hau-Riege, S. P., Frank, M., Hampton, C. Y., Sierra, R. G., Starodub, D., Williams, G. J., Hajdu, J., Timneanu, N., Seibert, M. M., Andreasson, J., Rocker, A., Jönsson, O., Svenda, M., Stern, S., Nass, K., Andritschke, R., Schröter, C.-D., Krasniqi, F., Bott, M., Schmidt, K. E., Wang, X., Grotjohann, I., Holton, J. M., Barends, T. R. M., Neutze, R., Marchesini, S., Fromme, R., Schorb, S., Rupp, D., Adolph, M., Gorkhover, T., Andersson, I., Hirsemann, H., Potdevin, G., Graafsma, H., Nilsson, B. & Spence, J. C. H. (2011). *Nature*, **470**(7332), 73–77. Number: 7332 Publisher: Nature Publishing Group. https://www.nature.com/articles/nature09750

Chhetri, B. K., Mojib, N., Moore, S. G., Delgadillo, D. A., Burch, J. E., Barrett, N. H., Gaul, D. A., Marquez, L., Soapi, K., Nelson, H. M., Quave, C. L. & Kubanek, J. (2023). *ACS Omega*, **8**(15), 13899–13910. Publisher: American Chemical Society. https://doi.org/10.1021/acsomega.3c00301

Cichocka, M. O., Ångström, J., Wang, B., Zou, X. & Smeets, S. (2018). *Journal of Applied Crystallography*, **51**(6), 1652–1661. Publisher: International Union of Crystallography. https://journals.iucr.org/j/issues/2018/06/00/yr5038/

Clabbers, M. T. B., Gruene, T., Parkhurst, J. M., Abrahams, J. P. & Waterman, D. G. (2018). *Acta Crystallographica Section D: Structural Biology*, **74**(6), 506–518. Publisher: International Union of Crystallography. https://scripts.iucr.org/cgi-bin/paper?rr5161

Dalton, K. M., Greisman, J. B. & Hekstra, D. R. (2022). *Nature Communications*, **13**(1), 1–13. Number: 1 Publisher: Nature Publishing Group. https://www.nature.com/articles/s41467-022-35280-8

DeLaBarre, B. & Brunger, A. T. (2006). *Acta Crystallographica Section D: Biological Crystallography*, **62**(8), 923–932. Publisher: International Union of Crystallography. //journals.iucr.org/paper?dz5074

Delgadillo, D. A., Burch, J. E., Kim, L. J., de Moraes, L. S., Niwa, K., Williams, J., Tang, M. J., Lavallo, V. G., Khatri Chhetri, B., Jones, C. G., Rodriguez, I. H., Signore, J. A., Marquez, L., Bhanushali, R., Woo, S., Kubanek, J., Quave, C., Tang, Y. & Nelson, H. M. (2024). *ACS Central Science*, **10**(1), 176–183. Publisher: American Chemical Society. https://doi.org/10.1021/acscentsci.3c01365

Dorset, D. L. (1996). *Acta Crystallographica Section B: Structural Science*, **52**(5), 753–769. Publisher: International Union of Crystallography. https://journals.iucr.org/b/issues/1996/05/00/an0524/

Evans, P. R. (2011). *Acta Crystallographica Section D: Biological Crystallography*, **67**(4), 282–292. Publisher: International Union of Crystallography. //scripts.iucr.org/cgi-bin/paper?ba5158

Evans, P. R. & Murshudov, G. N. (2013). *Acta Crystallographica Section D: Biological Crystallography*, **69**(7), 1204–1214. Publisher: International Union of Crystallography.
//scripts.iucr.org/cgi-bin/paper?ba5190

E. Gorelik, T., E. Tehrani, K. H. M., Gruene, T., Monecke, T., Niessing, D., Kaiser, U., Blankenfeldt, W. & Müller, R. (2022). *CrystEngComm*, **24**(33), 5885–5889. Publisher: Royal Society of Chemistry.
https://pubs.rsc.org/en/content/articlelanding/2022/ce/d2ce00707j

Foadi, J., Aller, P., Alguel, Y., Cameron, A., Axford, D., Owen, R. L., Armour, W., Waterman, D. G., Iwata, S. & Evans, G. (2013). *Acta Crystallographica Section D: Biological Crystallography*, **69**(8), 1617–1632. Publisher: International Union of Crystallography.
//scripts.iucr.org/cgi-bin/paper?dz5278

Gemmi, M., Mugnaioli, E., Gorelik, T. E., Kolb, U., Palatinus, L., Boullay, P., Hovmöller, S. & Abrahams, J. P. (2019). *ACS Central Science*, **5**(8), 1315–1329. Publisher: American Chemical Society.
https://doi.org/10.1021/acscentsci.9b00394

Gildea, R. J., Beilsten-Edmands, J., Axford, D., Horrell, S., Aller, P., Sandy, J., Sanchez-Weatherby, J., Owen, C. D., Lukacik, P., Strain-Damerell, C., Owen, R. L., Walsh, M. A. & Winter, G. (2022). *Acta Crystallographica Section D: Structural Biology*, **78**(6), 752–769. Publisher: International Union of Crystallography.
https://journals.iucr.org/d/issues/2022/06/00/gm5092/

Giordano, R., Leal, R. M. F., Bourenkov, G. P., McSweeney, S. & Popov, A. N. (2012). *Acta Crystallographica Section D: Biological Crystallography*, **68**(6), 649–658. Number: 6 Publisher: International Union of Crystallography.
//scripts.iucr.org/cgi-bin/paper?lv5017

Greisman, J. B., Dalton, K. M. & Hekstra, D. R. (2021). *Journal of Applied Crystallography*, **54**(5), 1521–1529. Publisher: International Union of Crystallography.
https://journals.iucr.org/j/issues/2021/05/00/te5079/

Grimes, J. M., Hall, D. R., Ashton, A. W., Evans, G., Owen, R. L., Wagner, A., McAuley, K. E., von Delft, F., Orville, A. M., Sorensen, T., Walsh, M. A., Ginn, H. M. & Stuart, D. I. (2018). *Acta Crystallographica Section D: Structural Biology*, **74**(2), 152–166. Publisher: International Union of Crystallography.
https://scripts.iucr.org/cgi-bin/paper?ba5286

Gruene, T., Wennmacher, J. T. C., Zaubitzer, C., Holstein, J. J., Heidler, J., Fecteau-Lefebvre, A., De Carlo, S., Müller, E., Goldie, K. N., Regeni, I., Li, T., Santiso-Quinones, G., Steinfeld, G., Handschin, S., van Genderen, E., van Bokhoven, J. A., Clever, G. H. & Pantelic, R. (2018). *Angewandte Chemie International Edition*, **57**(50), 16313–16317. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/anie.201811318.
https://onlinelibrary.wiley.com/doi/abs/10.1002/anie.201811318

Hattne, J., Reyes, F. E., Nannenga, B. L., Shi, D., de la Cruz, M. J., Leslie, A. G. W. & Gonen, T. (2015). *Acta Crystallographica Section A: Foundations and Advances*, **71**(4), 353–360. Number: 4 Publisher: International Union of Crystallography.
//scripts.iucr.org/cgi-bin/paper?mq5031

Ito, S., White, F. J., Okunishi, E., Aoyama, Y., Yamano, A., Sato, H., Ferrara, J. D., Jasnowski, M. & Meyer, M. (2021). *CrystEngComm*, **23**(48), 8622–8630. Publisher: The Royal Society of Chemistry.
https://pubs.rsc.org/en/content/articlelanding/2021/ce/d1ce01172c

Jones, C. G., Martynowycz, M. W., Hattne, J., Fulton, T. J., Stoltz, B. M., Rodriguez, J. A., Nelson, H. M. & Gonen, T. (2018). *ACS Central Science*, **4**(11), 1587–1592. Publisher: American Chemical Society.
https://doi.org/10.1021/acscentsci.8b00760

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S. & Saul, L. K. (1999). *Machine Learning*, **37**(2), 183–233.
https://doi.org/10.1023/A:1007665907178

Kabsch, W. (2010*a*). *Acta Crystallographica Section D: Biological Crystallography*, **66**(2), 133–144. Number: 2 Publisher: International Union of Crystallography.
//scripts.iucr.org/cgi-bin/paper?dz5178

Kabsch, W. (2010*b*). *Acta Crystallographica Section D: Biological Crystallography*, **66**(2), 125–132. Publisher: International Union of Crystallography.
//scripts.iucr.org/cgi-bin/paper?dz5179

Karplus, P. A. & Diederichs, K. (2015). *Current Opinion in Structural Biology*, **34**, 60–68.
https://www.sciencedirect.com/science/article/pii/S0959440X15000937

Khouchen, M., Klar, P. B., Chintakindi, H., Suresh, A. & Palatinus, L. (2023). *Acta Crystallographica Section A: Foundations and Advances*, **79**(5), 427–439. Publisher: International Union of Crystallography.
https://journals.iucr.org/a/issues/2023/05/00/pl5027/

Kim, L. J., Ohashi, M., Zhang, Z., Tan, D., Asay, M., Cascio, D., Rodriguez, J. A., Tang, Y. & Nelson, H. M. (2021). *Nature Chemical Biology*, **17**(8), 872–877. Number: 8 Publisher: Nature Publishing Group.
https://www.nature.com/articles/s41589-021-00834-2

Kingma, D. P. & Welling, M. (2014). *arXiv*, pp. 1312.6114, ver. 10. ArXiv:1312.6114 [cs, stat] version: 1.
http://arxiv.org/abs/1312.6114

Lightowler, M., Li, S., Ou, X., Cho, J., Li, A., Hofer, G., Xu, J., Yang, T., Zou, X., Lu, M. & Xu, H. (2023). *ChemRxiv*, p. ver. 1.
https://chemrxiv.org/engage/chemrxiv/article-details/6466d172a32ceeff2ddee872

Mott, N. F. & Bragg, W. L. (1997). *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, **127**(806), 658–665. Publisher: Royal Society.
https://royalsocietypublishing.org/doi/10.1098/rspa.1930.0082

Nannenga, B. L. & Gonen, T. (2014). *Current Opinion in Structural Biology*, **27**, 24–31.
https://www.sciencedirect.com/science/article/pii/S0959440X14000268

Palatinus, L., Petříček, V. & Corrêa, C. A. (2015). *Acta Crystallographica Section A: Foundations and Advances*, **71**(2), 235–244. Publisher: International Union of Crystallography.
//scripts.iucr.org/cgi-bin/paper?td5023

Park, J.-D., Li, Y., Moon, K., Han, E. J., Lee, S. R. & Seyedsayamdost, M. R. (2022). *Angewandte Chemie International Edition*, **61**(4), e202114022. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/anie.202114022.
https://onlinelibrary.wiley.com/doi/abs/10.1002/anie.202114022

Petříček, V., Palatinus, L., Plášil, J. & Dušek, M. (2023). *Zeitschrift für Kristallographie - Crystalline Materials*, **238**(7-8), 271–282. Publisher: De Gruyter (O).
https://www.degruyter.com/document/doi/10.1515/zkri-2023-0005/html

Powell, S. M., Novikova, I. V., Kim, D. N. & Evans, J. E. (2021). *bioRxiv*, p. ver. 1. Pages: 2021.12.13.472146 Section: New Results.
https://www.biorxiv.org/content/10.1101/2021.12.13.472146v2

Saha, A., Mecklenburg, M., Pattison, A. J., Brewster, A. S., Rodriguez, J. A. & Ercius, P. (2024). *arXiv*, pp. 2404.18011, ver. 1. ArXiv:2404.18011 [cond-mat].
http://arxiv.org/abs/2404.18011

Schrödinger, LLC (2022). The PyMOL Molecular Graphics System, Version 2.5.

Sheldrick, G. M. (1990). *Acta Crystallographica Section A: Foundations of Crystallography*, **46**(6), 467–473. Number: 6 Publisher: International Union of Crystallography.
//scripts.iucr.org/cgi-bin/paper?an0278

Sheldrick, G. M. (2015). *Acta Crystallographica Section A: Foundations and Advances*, **71**(1), 3–8. Publisher: International Union of Crystallography.
//scripts.iucr.org/cgi-bin/paper?sc5086

Shi, D., Nannenga, B. L., Iadanza, M. G. & Gonen, T. (2013). *eLife*, **2**, e01345. Publisher: eLife Sciences Publications, Ltd.
https://doi.org/10.7554/eLife.01345

Smith, J. L., Fischetti, R. F. & Yamamoto, M. (2012). *Current Opinion in Structural Biology*, **22**(5), 602–612.
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3478446/

Uervirojnangkoorn, M., Zeldin, O. B., Lyubimov, A. Y., Hattne, J., Brewster, A. S., Sauter, N. K., Brunger, A. T. & Weis, W. I. (2015). *eLife*, **4**, e05421. Publisher: eLife Sciences Publications, Ltd.
https://doi.org/10.7554/eLife.05421

Unge, J., Lin, J., Weaver, S. J., Sae Her, A. & Gonen, T. (2024). *Advanced Science*, **11**, 2400081. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/advs.202400081.
https://onlinelibrary.wiley.com/doi/abs/10.1002/advs.202400081

Usón, I. & Sheldrick, G. M. (1999). *Current Opinion in Structural Biology*, **9**(5), 643–648.
https://www.sciencedirect.com/science/article/pii/S0959440X99000202

Wang, B., Zou, X. & Smeets, S. (2019). *IUCrJ*, **6**(5), 854–867. Publisher: International Union of Crystallography.
https://journals.iucr.org/m/issues/2019/05/00/fc5033/

Wang, Y., Takki, S., Cheung, O., Xu, H., Wan, W., Öhrström, L. & Ken Inge, A. (2017). *Chemical Communications*, **53**(52), 7018–7021. Publisher: Royal Society of Chemistry.
https://pubs.rsc.org/en/content/articlelanding/2017/cc/c7cc03180g

White, T. A., Mariani, V., Brehm, W., Yefanov, O., Barty, A., Beyerlein, K. R., Chervinskii, F., Galli, L., Gati, C., Nakane, T., Tolstikova, A., Yamashita, K., Yoon, C. H., Diederichs, K. & Chapman, H. N. (2016). *Journal of Applied Crystallography*, **49**(2), 680–689. Publisher: International Union of Crystallography.
//scripts.iucr.org/cgi-bin/paper?zd5001

Wilson, A. J. C. (1949). *Acta Crystallographica*, **2**(5), 318–321. Publisher: International Union of Crystallography.
https://journals.iucr.org/q/issues/1949/05/00/a00174/

Winter, G., Waterman, D. G., Parkhurst, J. M., Brewster, A. S., Gildea, R. J., Gerstel, M., Fuentes-Montero, L., Vollmar, M., Michels-Clark, T., Young, I. D., Sauter, N. K. & Evans, G. (2018). *Acta Crystallographica Section D: Structural Biology*, **74**(2), 85–97. Publisher: International Union of Crystallography.
https://scripts.iucr.org/cgi-bin/paper?di5011

Yamashita, K., Hirata, K. & Yamamoto, M. (2018). *Acta Crystallographica Section D: Structural Biology*, **74**(5), 441–449. Publisher: International Union of Crystallography.
https://scripts.iucr.org/cgi-bin/paper?wa5117

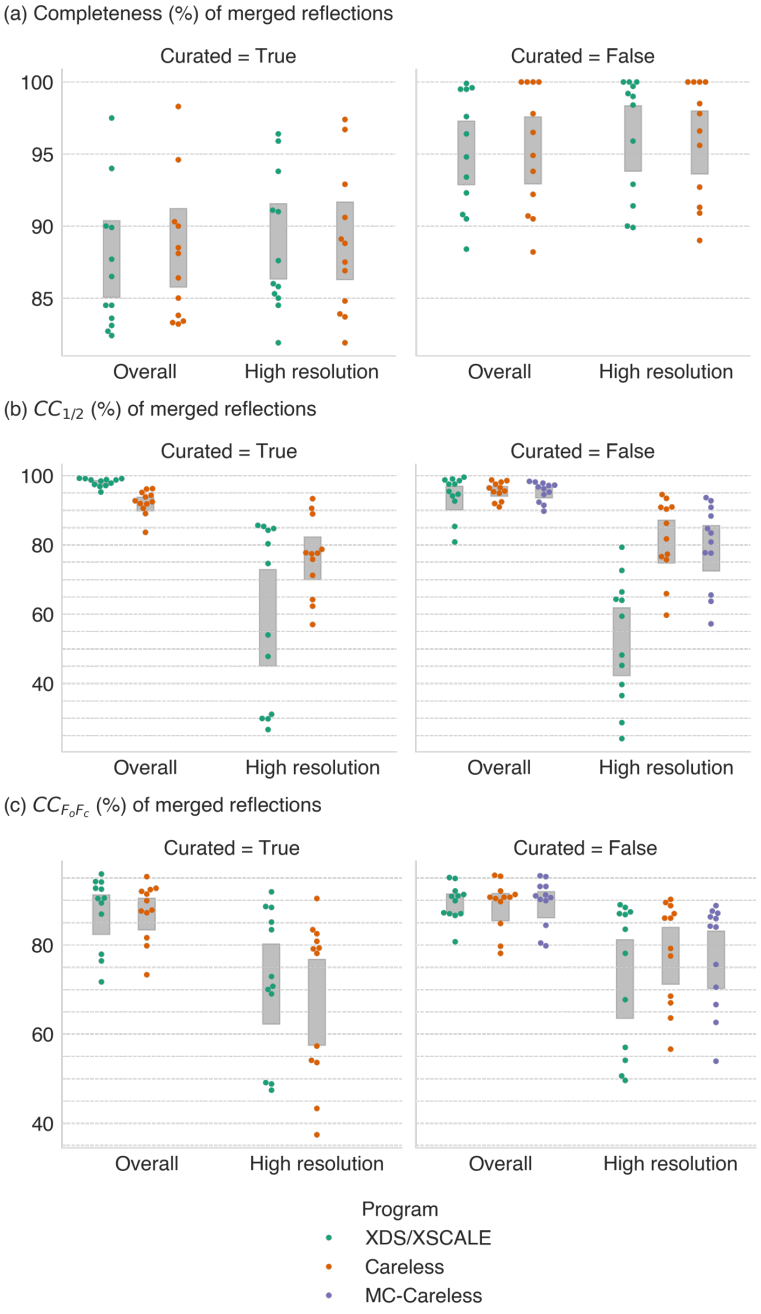

Fig. 1. Statistics of multi-crystal merging with (Curated=True) and without (Curated=False) manual dataset curation. Grey-shaded regions represent 95% confidence intervals for the mean from bootstrapping. (a) Data completeness is maximized when using all datasets. (b) Manual dataset curation has the opposite effects on $CC_{1/2}$ in XDS/XSCALE and Careless. Automated curation by MC-Careless does not improve results. Careless consistently achieves higher $CC_{1/2}$ than XDS/XSCALE in the highest resolution bin. (c) All merging protocols achieve similar accuracy as indicated by $CC_{F_oF_c}$.

IUCr macros version 2.1.17: 2023/10/19

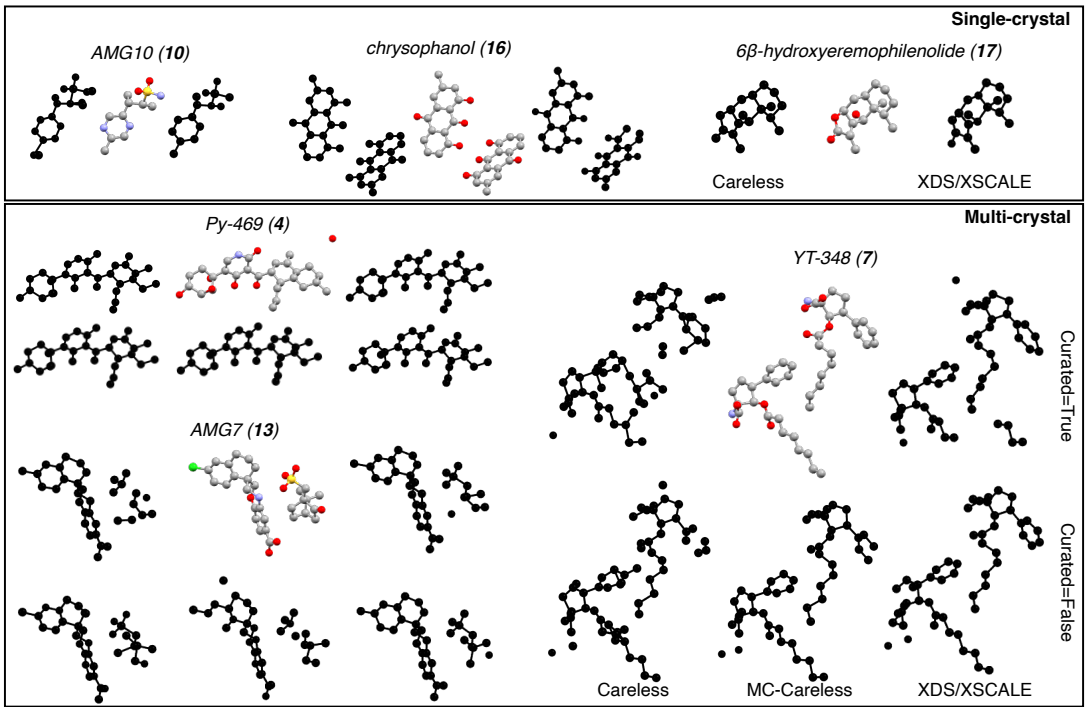

Fig. 2. Representative examples of *ab initio* phasing outcomes (black) show that correct structural information is extracted from XDS/XSCALE and Careless merging outcomes by standard phasing programs regardless of dataset curation. Reference structures (colored by elements) are presented for comparison.

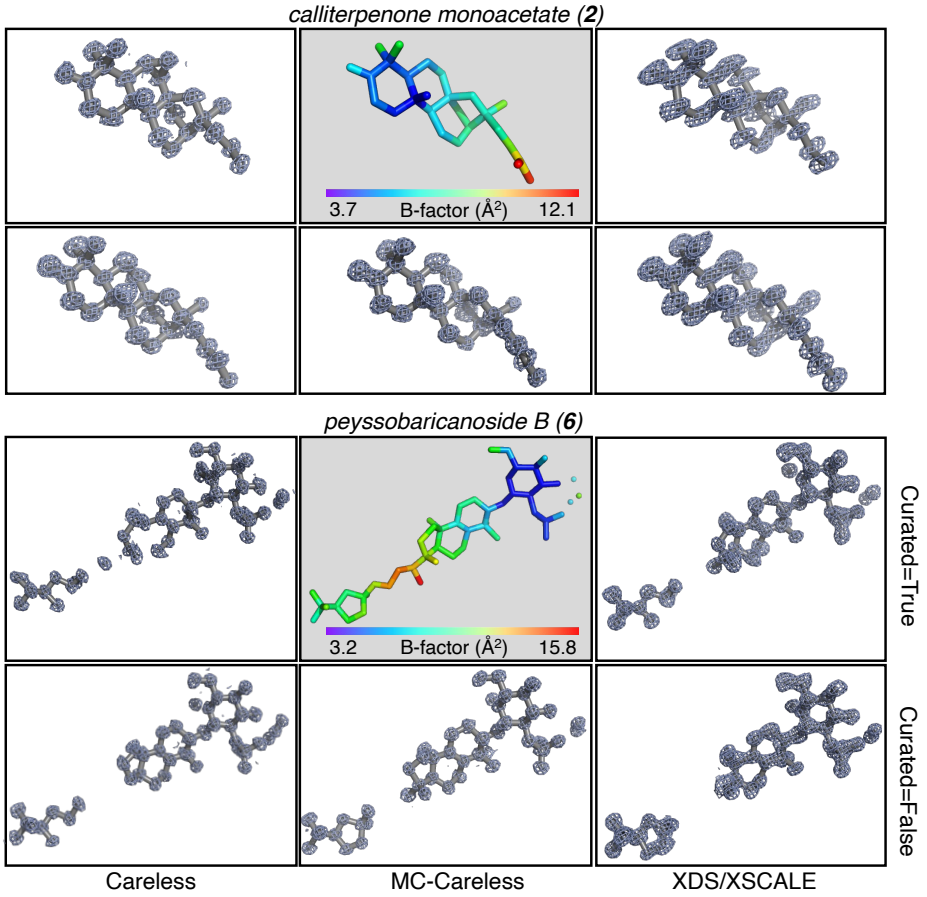

Fig. 3. Data merged by Careless and MC-Careless yield sharper $2F_o - F_c$ maps after *ab initio* phasing. Example *ab initio* structures and maps from all 5 merging protocols are shown for calliterpenone acetate (left) and peyssobaricanoside B (right). Reference structures are colored by atomic B-factors to show flexible parts where phasing might be challenging.

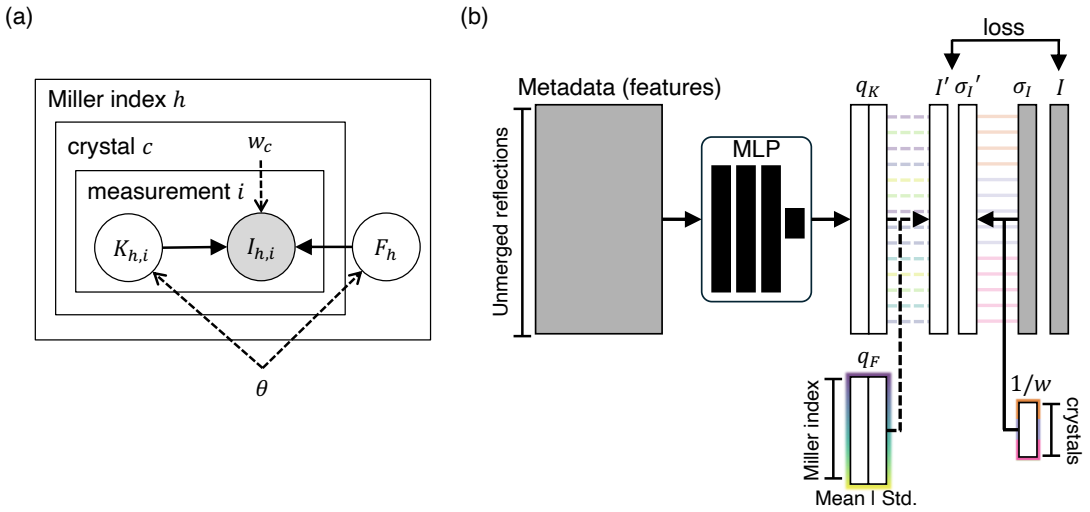

Fig. 4. Schematic of multi-crystal extension to Careless. (a) Probabilistic graphical model of merging diffraction data using variational Bayesian inference algorithm. Solid lines denote the generative process of the observed intensity I from the scaling factor K and structure factor amplitude F. Dashed lines represent variational Bayesian inference of K and F parameterized by model parameters θ, where the uncertainty of I is adjusted by a per-crystal weight w. (b) MC-Careless model architecture. Posterior estimates of K and F are variationally approximated as q_K and q_F respectively, and q_K is further parameterized by an MLP transformation from the metadata of unmerged reflections. Dashed lines denote the reparameterization process, where samples $\hat{K}$ and $\hat{F}$ are drawn from q_K and q_F to compute the loss between observed intensity I and predicted intensity $\hat{I} = \hat{K}\hat{F}^2$ with adjusted uncertainty σ_I/w.

Table 1. *Single-crystal* merging results.*

	molecule	AMG10 (**13**)	mannitol (**14**)	calcium oxalate (**15**)	chryso-phanol (**16**)	6β-hydroxy-eremoph-ilenolide (**17**)
XDS	highres† bin (Å)	0.9-0.85	1.01-0.95	1.01-0.95	0.9-0.85	0.9-0.85
	Completeness (%)	98.5 (98.6)	82.2 (87.3)	87.7 (88.9)	91.8 (92.4)	89.1 (95.1)
	$CC_{1/2}$ (%)	98.4 (57)	99.8 (21.9)	98.7 (74.8)	98.1 (88)	98.9 (38.9)
	$CC_{F_oF_c}$ (%)	76.6 (71.9)	95.4 (61.2)	86.8 (60.9)	88.1 (85)	89.4 (30.5)
	phasing	✓‡	✓‡	✓‡	✓‡	✓‡
Careless	highres bin (Å)	0.89-0.85	0.98-0.95	1-0.95	0.88-0.85	0.88-0.85
	Completeness (%)	99.7 (98)	84.6 (86.2)	89.8 (94.7)	92.2 (92.6)	89.7 (94.4)
	$CC_{1/2}$ (%)	87 (61)	97.9 (27.5)	92.9 (85.4)	92.2 (78.1)	94.7 (49.5)
	$CC_{F_oF_c}$ (%)	75.1 (80.9)	92.3 (50.2)	92.1 (55.7)	88.2 (76.5)	85 (38.6)
	phasing	✓‡	✓	✓‡	✓	✓

* For biotin single-crystal results, see Table 3.
† Statistics in the highest resolution (highres) bin are shown in parentheses.
‡ Phased by SHELXT. Otherwise phasing is performed using SHELXD.

Table 2. *95% CI of bootstrapped mean of single-crystal and manually curated merging results.*

metric (%)	XDS	Careless
$CC_{1/2}$ (overall)	97.58 - 98.66	90.56 - 93.87
$CC_{1/2}$ (highres)	46.52 - 69.46	62.10 - 77.94
$CC_{F_oF_c}$ (overall)	83.57 - 90.48	83.86 - 89.80
$CC_{F_oF_c}$ (highres)	60.35 - 76.30	57.40 - 73.72

Table 3. *Multi-crystal merging results using manually selected datasets.*

	molecule	demeth-oxyviridin (**1**)	calliter-penone acetate (**2**)	pachybasin (**3**)	Py-469 (**4**)	fischerin (**5**)	peysso-barican-oside B (**6**)
	# crystals	2	3	3	2	4	3
XDS/XSCALE	highres bin (Å)	1.04-1	1.04-1	0.93-0.9	0.94-0.9	1.09-1.05	1.14-1.1
	Completeness (%)	90 (95.9)	82.4 (85.8)	83.6 (81.9)	84.5 (84.5)	89.9 (91.1)	94.0 (93.8)
	$CC_{1/2}$ (%)	97.1 (54)	97.8 (84.2)	98.7 (47.8)	98.8 (84.7)	99.2 (29.9)	99.1 (31.1)
	$CC_{F_oF_c}$ (%)	71.7 (48.8)	94.2 (88.4)	92.7 (70.0)	77.9 (85.1)	90.4 (49.1)	92.5 (72.9)
	phasing	✓	✓	✓‡	✓	✓	✓
Careless	highres bin (Å)	1.04-1	1.04-1	0.93-0.9	0.94-0.9	1.09-1.05	1.14-1.1
	Completeness (%)	90.3 (96.7)	83.3 (83.9)	83.8 (81.9)	85 (84.8)	90 (90.6)	94.6 (92.9)
	$CC_{1/2}$ (%)	92.7 (77.6)	96.2 (90.5)	96.1 (64.2)	91.9 (88.9)	83.6 (75.8)	93.8 (77.4)
	$CC_{F_oF_c}$ (%)	81.6 (54.1)	92.4 (79.1)	91.4 (37.4)	79.8 (79.3)	73.3 (82.5)	92.7 (83.4)
	phasing	✓	✓	✓	✓	✓	✓

Table 3. *Continued*

	molecule	YT-348 (**7**)	AMG3 (**8**)	AMG4 (**9**)	AMG7 (**10**)	AMG11 (**11**)	biotin (**12**)
	# crystals	9	2	2	3	2	1
XDS/XSCALE	highres bin (Å)	1.04-1	1.04-1	1.04-1	1.03-1	0.93-0.9	0.9-0.85
	Completeness (%)	87.7 (87.6)	82.7 (85.0)	83.1 (85.3)	84.5 (86.0)	86.5 (91)	97.5 (96.4)
	$CC_{1/2}$ (%)	98.7 (26.7)	98.4 (74.6)	99.1 (85.3)	95.2 (85.6)	97.4 (29.8)	96.8 (80.3)
	CCF_oF_c (%)	89.4 (47.4)	94.1 (83.4)	95.9 (91.9)	76.4 (88.6)	90.5 (69)	86.9 (70.7)
	phasing	✓	✓	✓	✓	✓‡	✓‡
Careless	highres bin (Å)	1.04-1	1.04-1	1.04-1	1.03-1	0.93-0.9	0.89-0.86
	Completeness (%)	88.1 (88.8)	83.4 (86.9)	83.2 (83.7)	86.4 (89.1)	88.5 (87.5)	98.3 (97.4)
	$CC_{1/2}$ (%)	94.3 (57.0)	91.8 (78.7)	95.1 (93.3)	89 (77.7)	90.5 (71.2)	92.4 (62.3)
	$CC_{F_oF_c}$ (%)	87.8 (57.3)	92 (78.1)	95.3 (90.4)	87.2 (80.8)	89.9 (53.6)	87.6 (43.3)
	phasing	✓	✓	✓	✓	✓	✓‡

Table 4. *Multi-crystal merging results using all datasets.*

	molecule	demeth-oxyviridin (**1**)	calliter-penone acetate (**2**)	pachybasin (**3**)	Py-469 (**4**)	fischerin (**5**)	peysso-barican-oside B (**6**)
	# crystals	6	9	6	21	89	16
XDS/XSCALE	highres bin (Å)	1.04-1	1.04-1	0.93-0.9	0.94-0.9	1.09-1.05	1.14-1.1
	Completeness (%)	92.3 (99)	96.4 (99.2)	99.5 (100)	94.8 (95.9)	90.8 (91.4)	97.6 (98.4)
	$CC_{1/2}$ (%)	92.6 (64)	80.8 (66.4)	97.5 (48.2)	85.3 (64.3)	99.5 (28.7)	95.4 (45.2)
	$CC_{F_oF_c}$ (%)	80.7 (78.1)	87.2 (86.8)	92.1 (67.7)	86.6 (89)	90.9 (50.6)	89.9 (57)
	phasing	✓	✓	✓‡	✓	✓	✓
Careless	highres bin (Å)	1.04-1	1.04-1	0.93-0.9	0.94-0.9	1.09-1.05	1.14-1.1
	Completeness (%)	92.9 (97.8)	96.5 (96.6)	100 (100)	94.9 (95.6)	90.7 (91.3)	97.8 (98.5)
	$CC_{1/2}$ (%)	91.9 (77.3)	94.9 (90.8)	96.5 (59.7)	98.1 (93.4)	97.5 (90.9)	95.5 (75.7)
	$CC_{F_oF_c}$ (%)	78.1 (67)	90.7 (87)	89.7 (63.6)	92.1 (89.5)	90.7 (86)	91.3 (77.5)
	phasing	✓	✓	✓	✓	×	✓
MC-Careless	$CC_{1/2}$ (%)	91.4 (80.8)	96.7 (88.3)	96.2 (57.2)	98.1 (93.6)	97.8 (90.8)	94.5 (63.7)
	$CC_{F_oF_c}$ (%)	79.8 (66.6)	91 (86.3)	90.2 (62.6)	93.1 (88.8)	89.9 (84)	93.1 (75.6)
	phasing	✓	✓‡	✓	✓	×	✓

Table 4. *Continued*

	molecule	YT-348 (**7**)	AMG3 (**8**)	AMG4 (**9**)	AMG7 (**10**)	AMG11 (**11**)	biotin (**12**)
	# crystals	17	18	21	10	8	3
XDS/XSCALE	highres bin (Å)	1.04-1	1.04-1	0.93-0.9	0.94-0.9	1.09-1.05	1.14-1.1
	Completeness (%)	90.5 (90)	93.4 (92.9)	88.4 (89.9)	99.5 (99.7)	99.9 (100)	99.6 (100)
	$CC_{1/2}$ (%)	98.5 (39.7)	99 (79.3)	98.7 (59.4)	94.1 (72.6)	97.5 (24.1)	94.6 (36.5)
	$CC_{F_oF_c}$ (%)	91 (49.6)	94.9 (87)	95.1 (88.4)	87 (87.4)	91.3 (83.5)	87 (54.1)
	phasing	✓	✓	✓	✓	✓‡	✓‡
Careless	highres bin (Å)	1.04-1	1.04-1	0.93-0.9	0.94-0.9	1.09-1.05	1.14-1.1
	Completeness (%)	90.5 (90.9)	93.8 (92.7)	88.2 (89)	100 (100)	100 (100)	100 (100)
	$CC_{1/2}$ (%)	92.4 (65.9)	98.7 (94.5)	98.5 (90.3)	90.9 (81.7)	95.3 (76.6)	96.4 (86.2)
	$CC_{F_oF_c}$ (%)	84.8 (56.6)	95.4 (86)	95.6 (90.2)	90.4 (88.8)	79.7 (79.2)	90.7 (68.5)
	phasing	✓	✓	✓	✓	✓	✓‡
MC-Careless	$CC_{1/2}$ (%)	89.7 (65.5)	98.3 (92.7)	97.2 (84.7)	92.3 (83.4)	95.2 (77.7)	97.1 (77.6)
	$CC_{F_oF_c}$ (%)	80.4 (53.9)	95.5 (87.1)	95.3 (87.6)	90.6 (85.9)	91.2 (84.2)	84.4 (70.5)
	phasing	✓	✓	✓	✓	✓	✓‡

Synopsis

Demonstration of variational Bayesian inference for merging multi-crystal small molecule microED data.