AutoLEI: An XDS-based Graphical User Interface for Automated Real-time and Offline Batch 3D ED/MicroED Data Processing

Lei Wang.[a, †] Yinlin Chen.[a, †] Emma Scaletti Hutchinson.[b] Pål Stenmark.[b] Gerhard Hofer.[a, *] Hongyi Xu.[a, c, *] and Xiaodong Zou.[a, *]

[a] Department of Chemistry, Stockholm University, 106 91 Stockholm, Sweden

[b] Department of Biochemistry and Biophysics, Stockholm University, 106 91 Stockholm, Sweden

[c] Research School of Chemistry, the Australian National University, Acton, Australia

[†] These authors contributed equally to this work

* Correspondence e-mail: xiaodong.zou@su.se

Synopsis A graphical user interface for real-time and offline 3D ED/MicroED data processing by XDS was developed. The GUI aims to improve efficiency, minimize redundant data processing work, and provide users with real-time feedback during data collection.

Abstract Three-dimensional electron diffraction (3D ED), also known as microcrystal electron diffraction (MicroED), is an emerging method for determining structures of submicron-sized crystals. With the development of rapid and convenient data collection protocols, acquiring dozens of datasets in a single MicroED session has become routine. A fast and automated workflow for processing, scaling and merging a large number of MicroED datasets can significantly accelerate the structure determination process. Herein, we present an XDS-based graphical user interface for automated real-time and offline batch 3D ED/MicroED data processing (AutoLEI). We illustrate the functionality of the GUI through four examples, demonstrating both offline and real-time data processing capabilities. These examples include small organic molecules, metal-organic frameworks (MOFs), and proteins, showcasing the versatility and efficiency of the GUI in various applications.

Keywords: 3D ED, MicroED, Real-time data processing, Offline batch data processing, Data analysis, Beam sensitive material

1. Introduction

Three-dimensional electron diffraction (3D ED), also known as micro-crystal electron diffraction (MicroED), has been developed into a robust method for crystal structure determination over the last few decades (Kolb *et al.*, 2007; Zhang *et al.*, 2010; Gemmi *et al.*, 2019). The method enables rapid and accurate structure determination using commercially available transmission electron microscopes (TEM). A wide range of structures, including zeolite (Guo *et al.*, 2015; Cho *et al.*, 2023; Lu *et al.*, 2024), metal-organic framework (MOF) (Huang, Willhammar *et al.*, 2021; Huang, Grape *et al.*, 2021), pharmaceuticals (Van Genderen *et al.*, 2016; Lightowler *et al.*, 2022), and more recently, macromolecules (Shi *et al.*, 2013; Xu *et al.*, 2019; Martynowycz *et al.*, 2022), have been successfully solved with 3D ED/MicroED.

Benefiting from the strong interaction between electrons and matter, MicroED data can be collected from crystals that are too small for single-crystal X-ray diffraction. An ideal TEM grid typically contains hundreds of diffracting crystals or crystal fragments. With the latest MicroED protocol, it is possible to determine structures of high symmetry or stable samples through a single dataset within a few hours. However, due to the inherent limitations of the goniometer, data merging for crystals of low-symmetry is usually required to achieve high data completeness. For highly beam-sensitive material, a small-wedge data collection and merging strategy can also be employed to increase the signal-to-noise ratio and data completeness. Furthermore, selectively merging MicroED datasets can improve data quality by increasing redundancy (Xu *et al.*, 2018; Samperisi *et al.*, 2021; Bengtsson *et al.*, 2022). More recently, high-throughput data collection has been achieved using software such as *Instamatic* (Cichocka et al., 2018), *Automated EM Data Acquisition with SerialEM* (De La Cruz *et al.*, 2019), *Leginon* (Cheng et al., 2021), *EPUD* (Thermo Fisher Scientific), and *Latitude D* (GATAN) or dedicated electron diffractometer namely, *Synergy-ED* (Rigaku) and *ELDICO ED-1* (ELDICO Scientific). It is feasible to perform qualitative phase analysis by collecting and processing abundant datasets (Luo *et al.*, 2023; Unge *et al.*, 2024; Lightowler *et al.*, 2024). In all the above cases, an efficient, automated, and even real-time data processing solution is highly desirable to enhance the efficiency of structure determination and phase analysis using MicroED.

A large selection of well-established software such as X-ray detector software (XDS) (Kabsch, 2010), PETS2 (Palatinus *et al.*, 2019), CrysAlis^{Pro}-ED (Truong *et al.*, 2023), xia2.multiplex (Gildea *et al.*, 2022) and DIALS (Winter *et al.*, 2018, 2022), are available for processing MicroED data. Data processing can be performed via either command-line instructions or graphical user interfaces (GUI) (Brehm *et al.*, 2023), offering flexibility to users. However, format conversion, sorting, processing, analyzing and merging a large amount of MicroED datasets can be time-consuming and laborious. Automated batch processing and real-time processing capabilities are still underdeveloped for MicroED. In our previous work, Edtools (Wang *et al.*, 2019) was developed to process batch datasets in offline mode but only works in the command line. A processing pipeline Scipion-ED (Bengtsson *et al.*, 2022) embedded in

Scipion was also attempted. However, the complex operation and specialized platform requirement make them difficult for beginners to use.

In this work, we utilized XDS as the data processing engine and developed a graphical user interface for Automated Real-time and Offline Batch 3D ED/MicroED Data Processing (AutoLEI). Based on the fast data processing performance of XDS, AutoLEI enabled both offline and real-time batch MicroED data processing. This Python-based GUI is straightforward to install and provides users with the options and parameters that are essential for working with electron diffraction data. Meanwhile, data processing statistics and quality indicators are also optimized for MicroED. We demonstrate how to use AutoLEI to process data collected from different types of crystals, including an organic molecule, a metal-organic framework and two proteins, on different microscopes and detectors. An example of real-time data processing using AutoLEI is also presented. Furthermore, a detailed tutorial with training datasets is provided.

2. AutoLEI implementation

2.1. AutoLEI installation and data format supporting

AutoLEI is an open-access GUI written in Python. It can be downloaded from PyPI, Gitlab or Zenodo. In version 1.0.0 the native image format is SMV for XDS data processing. Automated image format conversion supports TIFF images collected on CheeTah (Amsterdam Scientific Instruments), OneView (GATAN), Timepix hybrid pixel detectors (Amsterdam Scientific Instruments) and MRC images collected by EPUD (Thermo Fisher Scientific). Real-time data processing is designed for raw MRC image data collected using EPUD (Thermo Fisher Scientific) as well as data collected using *Instamatic*. Future updates are expected for improved compatibility and increased support, while users can also add new functionality through the programming interface.

2.2. AutoLEI data processing workflow

The primary aim of AutoLEI is to process and analyze a batch of MicroED datasets using XDS offline, or in real-time. Figure 1 shows the workflow of AutoLEI. Data processing goes through the steps of ***XDSRunner***, ***CellCorr***, ***XDSRefine***, ***MergeData*** and ***Cluster&Output***. These steps are implemented as tabs in the AutoLEI interface as shown in Supporting information (SI) Figures S1–S8. By following this workflow, final HKL files are obtained for downstream structural determination. Videos 1–3 (SI) showcase the offline data processing of three examples: tyrosine, MOF SU-100 and the protein *H. sapiens* MutT homolog 1 (hMTH1) while Video 4 (SI) shows the real-time data processing of triclinic lysozyme.

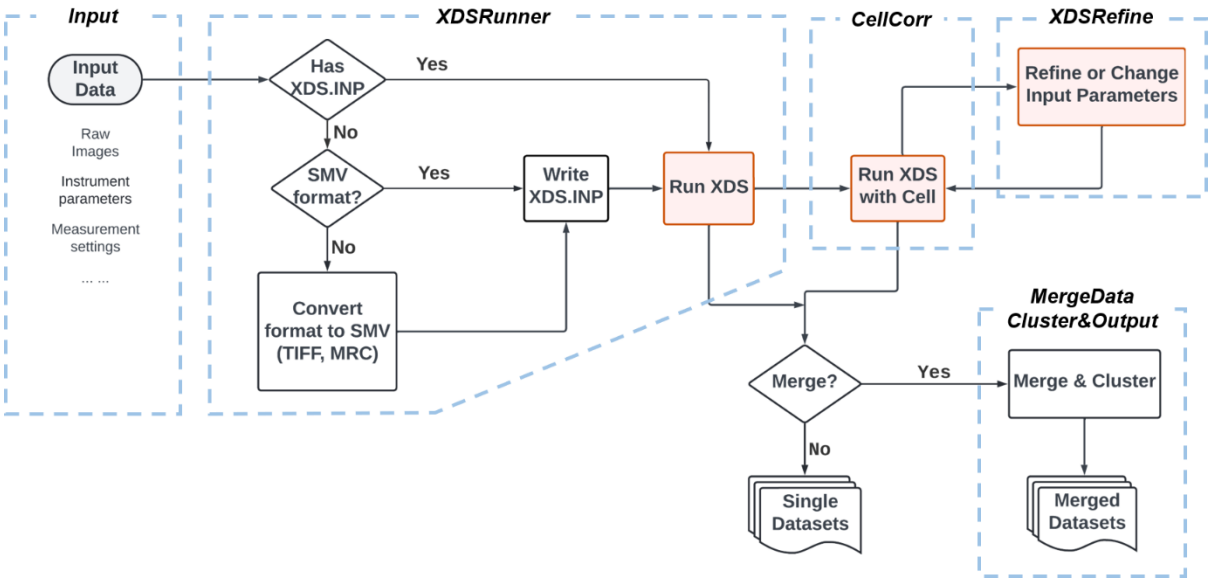


Figure 1 An overview of the AutoLEI workflow

2.2.1. Input

Input allows users to set up the work directory, instrument parameters, and measurement settings (Figure S1). The instrument parameters can be loaded from a predefined instrument model or existing input files. Alternatively, they can be given manually when datasets collected from different instruments and detectors are used. ***Input*** creates the "Input_parameters.txt" file, which will be used for generating input files "XDS.INP" for both real-time and offline MicroED data processing.

2.2.2. XDSrunner

XDSRunner generates the corresponding "XDS.INP" files for each dataset (Figure S2). It then launches XDS to process these datasets. No prior knowledge of the sample is required. AutoLEI processes all datasets sequentially under the work directory and parses statistics from XDS outputs. After data processing, AutoLEI will generate a summary table named "xdsrunner.xlsx". The table contains key parameters and quality indicators such as space group (SG), unit cell parameters (Unit cell), cell volume (Vol.), indexing rate (Index%), I/Sigma(I)^{asymptotic} (ISa), redundancy-independent merging R-factor (R_{meas}), correlation between reflection intensities of random half datasets ($CC_{1/2}$), completeness and estimated resolution (Reso.) of each dataset for inspection. SG, unit cell and ISa are read directly from XDS outputs directly while others are calculated by AutoLEI based on XDS HKL files. The high-resolution cut-off of each dataset is estimated by checking $CC_{1/2}$, R_{int} and signal-to-noise ratio of each resolution shell.

XDSRunner also contains supporting functions as shown below:

- *Format converter*: Convert raw image files to SMV and write the metadata into SMV files.

- *Find beam center*: Find the beam center automatically with or without a beam stop.
- *Estimate symmetry & cell-cluster*: Cluster data based on unit cell parameters and estimate the Laue group. This ensures data can be processed in the correct Laue group and unit cell.
- Details of these functions can be found in Supporting Information (S1. Supporting function implementation in AutoLEI).

2.2.3. CellCorr

CellCorr allows users to run XDS with given unit cell parameters and space group (Figure S3). It allows re-indexing the original HKL using space group and unit cell parameters suggested by either "Estimate symmetry & cell-cluster" in ***XDSRunner*** or determined by other software packages such as REDp (Wan *et al.*, 2013) and PETS2. This ensures that the datasets will be processed, scaled and merged in the correct Laue group and unit cell. ***CellCorr*** outputs a summary table named "xdsrunner2.xlsx" containing the same data quality indicators as "xdsrunner.xlsx".

2.2.4. XDSRefine

XDSRefine provides the opportunity to update XDS keywords in the input files for single, selected, or all datasets if users are not satisfied with the outputs from ***CellCorr*** (Figure S4). The step is useful for problematic data that cannot be processed using default settings. In AutoLEI v1.0.0, ***XDSRefine*** also provides automatic refinement of the rotation axis, and removal of scale outliers caused by crystal shifts during data collection. The reconstructed reciprocal lattice of each dataset can also be visually inspected. ***XDSRefine*** updates "xdsrunner2.xlsx" and the table will then be used for downstream data scaling and merging.

2.2.5. MergeData

MergeData analyzes statistics in "xdsrunner2.xlsx" and make initial selections of datasets for merging (Figure S5). Users can then inspect key indicators listed in the table and apply a filter for scaling/merging. The default value will be provided as a reference. The information on filtered data is written in "xdspicker.xlsx". ***MergeData*** will then initiate XSCALE and output scaled data as an HKL file.

2.2.6. Cluster&Output

Cluster&Output performs cluster analysis based on pair-wise correlation coefficients of reflection intensities in "XSCALE.LP" generated from the last step (Figure S6). A dendrogram will be produced for visual inspection. Users can set the distance threshold for further clustering. ***Cluster&Output*** will then perform scaling/merging of each cluster, output corresponding "HKL" files and data processing statistics (Figure S7). Webpage-based reports will also be generated for inspection of data reduction.

2.3. Real-time processing

Real-time MicroED data processing is a fully automated pipeline in AutoLEI (Figure S8). With default or customized settings, it provides crucial feedback to users for data quality analysis, evaluates data collection parameters, and generates result "HKL" files in real-time. The setup of the real-time processing function is found on the ***RealTime*** tab of AutoLEI, as shown in Figure 2. Real-time processing reads "Input_parameters.txt" from ***Input*** and continuously monitors the input folder where the data is being saved. Once a single dataset collection is completed, real-time processing will try to process data in either screening mode or merging mode:

a) *Screening mode*: When unit cell parameters and space group are not provided, the data is processed without performing data merging.

b) *Merging mode*: When unit cell parameters and space group are provided, the data will be processed in the specified cell and space group. If the data quality satisfies the filtering condition, it will be assigned as a valid dataset and merged with other valid datasets.

In the ***RealTime*** tab, estimated resolution, $CC_{1/2}$ and ISa for individual data or merged data are plotted and updated live, as shown in Figure 2a and 2b, respectively. Real-time processing currently supports the data collected using EPUD and Instamatic.

Figure 2 Real-time MicroED work page. (a) Screen Mode. (b) Merge Mode.

3. Examples

Table 1 and Figure 3 summarize four examples of data processing using AutoLEI. MicroED data were collected on a variety of microscopes and detectors using different software. For sample preparation details, please refer to the Supporting Information (S2). Among them, tyrosine (Video 1), SU-100 (Video 2) and hMTH1 (Video 3) use offline data processing protocol, while lysozyme (Video 4) showcases the online data processing capability.

Table 1 Summary of data used for demonstrating AutoLEI processing

Sample	Microscope	Detector	Data collection software	Total datasets	Data collection time	Data processing time[1]
Tyrosine ($P2_12_12_1$)	JEM-2100 LaB6 (JEOL)	Timepix hybrid pixel detector (ASI)	*Instamatic*	12	~ 1 h 20 min	~ 5 min
SU-100 ($C2/c$)	JEM-2100 LaB6 (JEOL)	Timepix hybrid pixel detector (ASI)	*Instamatic*	16	~ 2 h 15 min	~ 7 min
hMTH1 ($P2_12_12_1$)	Krios G3i (TFS)	Ceta-D (TFS)	EPUD (TFS)	38	~ 2 h 14 min	~ 12 min
Lysozyme ($P1$)	Krios G2 (TFS)	Ceta-D (TFS)	EPUD (TFS)	72	~ 6 h	Real-time

[1] Timing is from the start of data processing to obtaining usable "HKL" files (Videos 1–3)

JEOL: Japan Electron Optics Laboratory Co., LTD

ASI: Amsterdam Scientific Instruments

TFS: Thermo Fisher Scientific

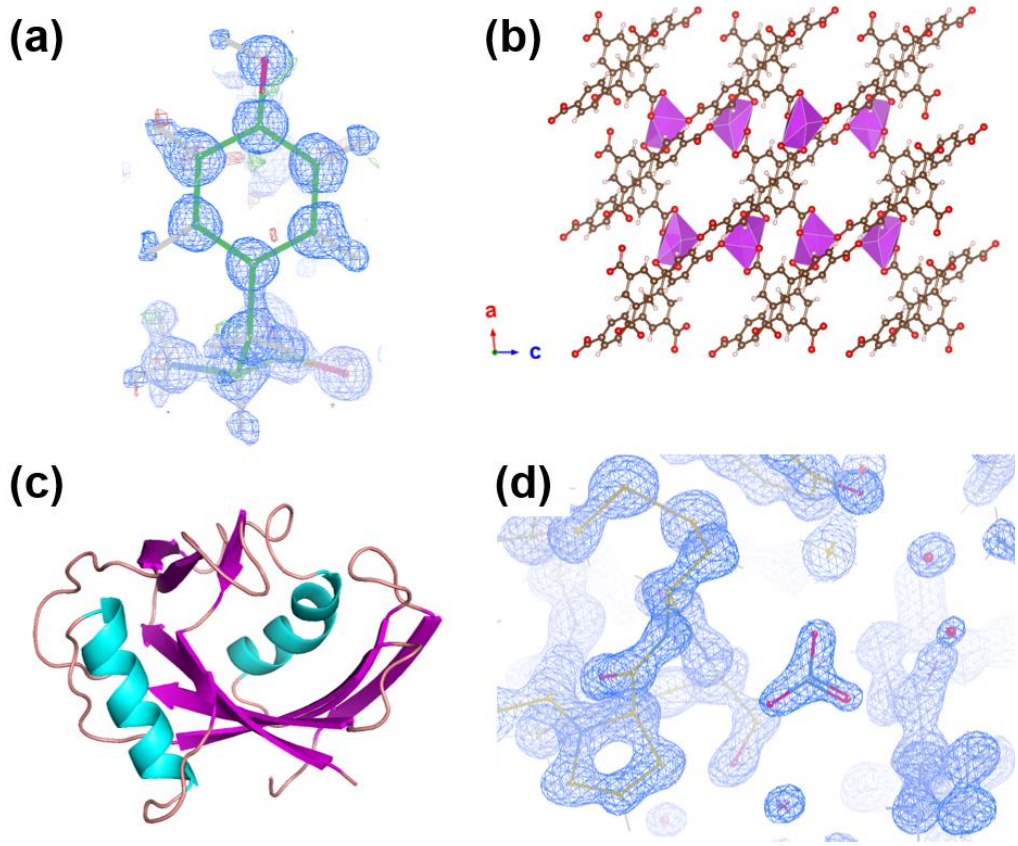


Figure 3 Summary of solved structures. (a) The structure of L-tyrosine using data from experiment_10. Electrostatic potential map (blue) and difference map (green and red) and tyrosine structure shown in COOT(Emsley *et al.*, 2010). The electrostatic potential map (blue) was contoured at 1σ-level while the difference map (red and green) was contoured at 3σ-level. (b) The structure framework of SU-100 viewed along the *b*-axis. (c) The structure of apo-hMTH1. (d) No-filled 2*F*o−*F*c electrostatic potential map (blue) with lysozyme model near a nitrate anion. The map was contoured at 1.5 r.m.s.d.

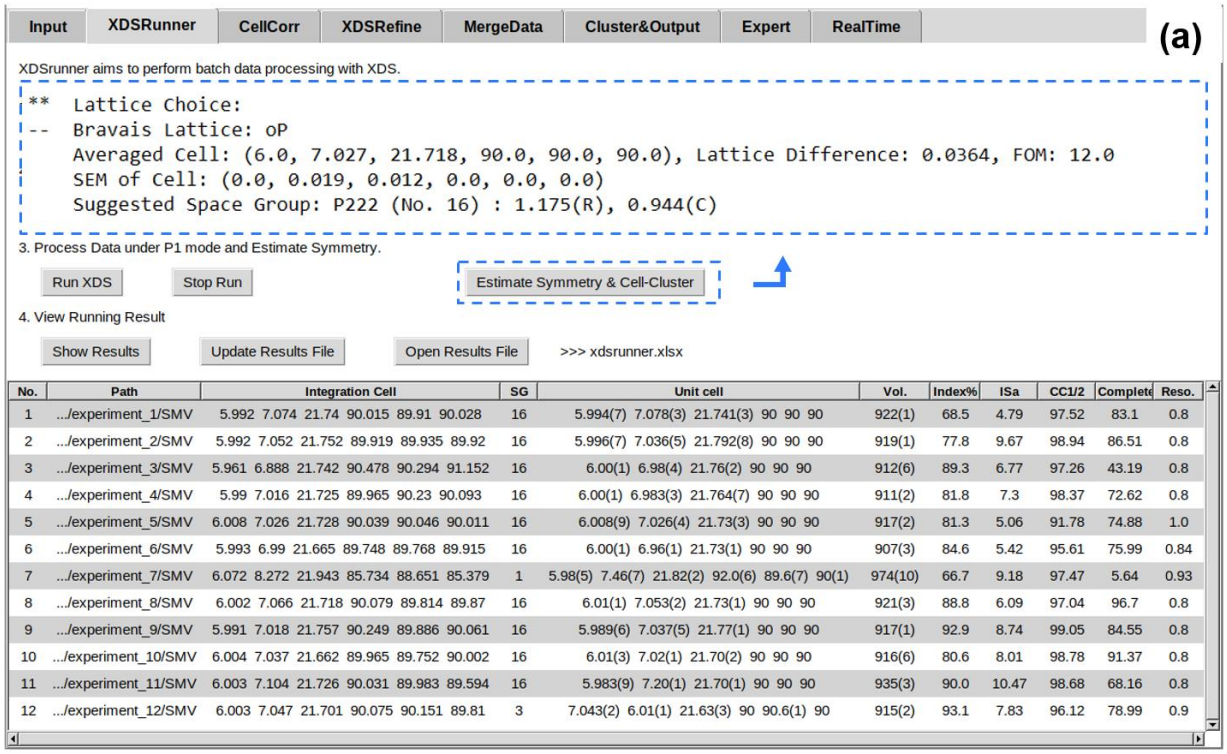
3.1. Single-data-based structure determination: rapid filtering of low-quality tyrosine data

Tyrosine is a naturally occurring amino acid and an essential protein building block. By collecting MicroED data over a large tilt range, structure determination from a single orthorhombic crystal is feasible. A total of 12 tyrosine datasets were collected and then processed with a default resolution range from 20 to 0.8 Å. Unit cell parameters, ISa, $CC_{1/2}$, completeness and estimated resolutions were determined and extracted from XDS outputs. As shown in Figure 4a, most of the datasets displayed favorable data quality and the estimated resolution reached the pre-defined high-resolution limit (0.8 Å) in ***XDSRunner*** tab. This suggested that the resolution range should be extended in the following procedure. However, both experiment_3 and experiment_7 suffered from unacceptable low completeness because of the low tilting range during data collection, indicating that these datasets might

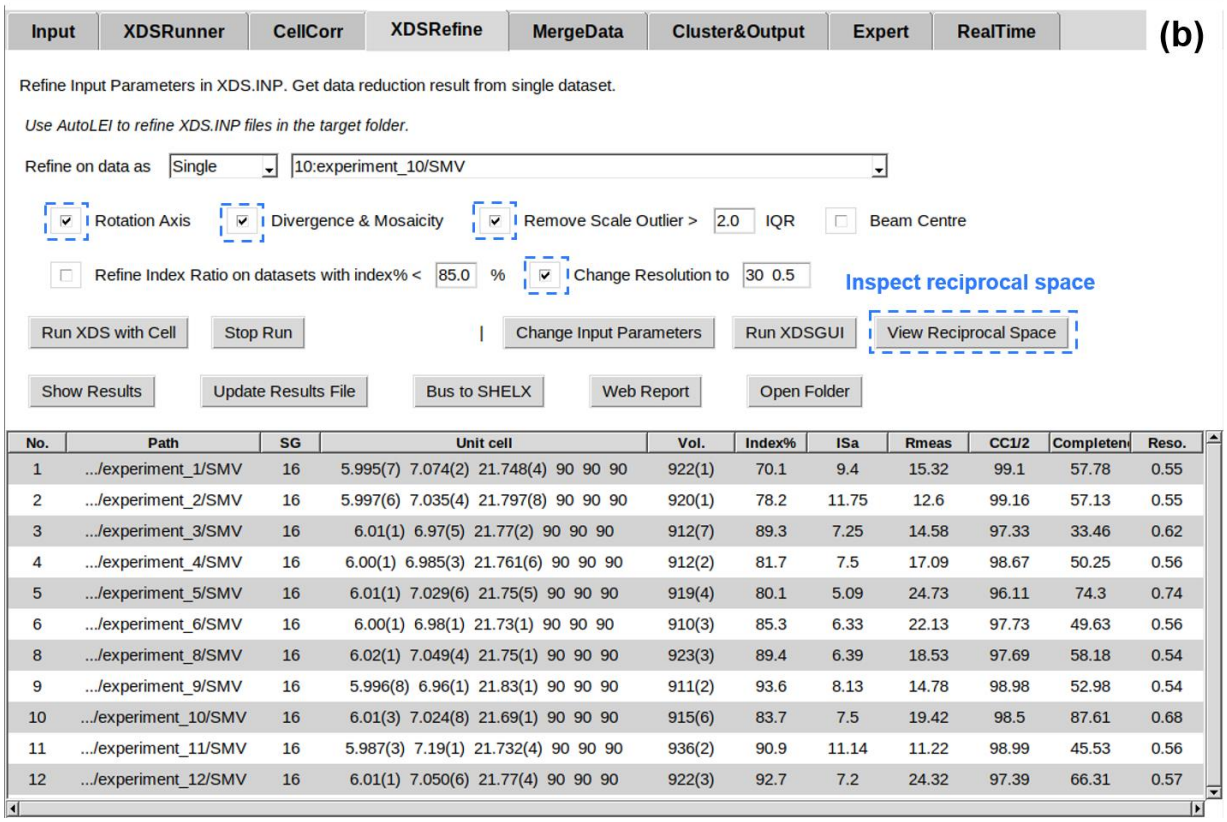
not be suitable. Subsequently, the unit cell parameters and Laue group were extracted from the output of "Estimate symmetry & cell clustering" and then used for XDS processing in ***CellCorr***. ***XDSrefine*** to refine the rotation axis, divergence and mosaicity, scale outliers and change resolution range to 30–0.5 Å, as shown in Figure 4b.

Different users may prioritize different key indicators when selecting the optimal datasets. Experientially, we recommend datasets with an ISa > 5, $CC_{1/2}$ > 95%, and completeness > 80% as being promising for single-data-based structure determination. We here selected experiment_10 due to its good data quality: Isa, $CC_{1/2}$, completeness and estimated resolution of this data were 7.51, 98.18%, 87.61% and 0.68 Å, respectively. AutoLEI also enables the systematic absence in reciprocal space to be visually inspected, as shown in Figure S9. It is noted that the 0*k*0 row was missing in most of datasets due to preferential crystal orientation. By using AutoLEI, it was possible to identify the missing 0*k*0 row in experiment_12 by checking the reciprocal space of all data rapidly. The space group could then be deduced as $P2_12_12_1$. The time used for data processing and analysis was approximately 5 minutes in total. The integrated data was then used for structure solution and kinematic refinement. The number of unique reflections and parameters used for refinement was 2592 and 162, respectively. The final R_1 converged to 11.90% (Table S1). All hydrogen atoms including the proton near the N atom could be found as shown in Figure 3a.

To further investigate the relationship between data quality and the structure information, we solved the structure using other datasets, as shown in Table S1. Except for experiment_3 and experiment_7, which had low completeness, all other datasets could be used for successful structure determination.



No.	Path	Integration Cell	SG	Unit cell	Vol.	Index%	ISa	CC1/2	Complete	Reso.
1	.../experiment_1/SMV	5.992 7.074 21.74 90.015 89.91 90.028	16	5.994(7) 7.078(3) 21.741(3) 90 90 90	922(1)	68.5	4.79	97.52	83.1	0.8
2	.../experiment_2/SMV	5.992 7.052 21.752 89.919 89.935 89.92	16	5.996(7) 7.036(5) 21.792(8) 90 90 90	919(1)	77.8	9.67	98.94	86.51	0.8
3	.../experiment_3/SMV	5.961 6.888 21.742 90.478 90.294 91.152	16	6.00(1) 6.98(4) 21.76(2) 90 90 90	912(6)	89.3	6.77	97.26	43.19	0.8
4	.../experiment_4/SMV	5.99 7.016 21.725 89.965 90.23 90.093	16	6.00(1) 6.983(3) 21.764(7) 90 90 90	911(2)	81.8	7.3	98.37	72.62	0.8
5	.../experiment_5/SMV	6.008 7.026 21.728 90.039 90.046 90.011	16	6.008(9) 7.026(4) 21.73(3) 90 90 90	917(2)	81.3	5.06	91.78	74.88	1.0
6	.../experiment_6/SMV	5.993 6.99 21.665 89.748 89.768 89.915	16	6.00(1) 6.96(1) 21.73(1) 90 90 90	907(3)	84.6	5.42	95.61	75.99	0.84
7	.../experiment_7/SMV	6.072 8.272 21.943 85.734 88.651 85.379	1	5.98(5) 7.46(7) 21.82(2) 92.0(6) 89.6(7) 90(1)	974(10)	66.7	9.18	97.47	5.64	0.93
8	.../experiment_8/SMV	6.002 7.066 21.718 90.079 89.814 89.87	16	6.01(1) 7.053(2) 21.73(1) 90 90 90	921(3)	88.8	6.09	97.04	96.7	0.8
9	.../experiment_9/SMV	5.991 7.018 21.757 90.249 89.886 90.061	16	5.989(6) 7.037(5) 21.77(1) 90 90 90	917(1)	92.9	8.74	99.05	84.55	0.8
10	.../experiment_10/SMV	6.004 7.037 21.662 89.965 89.752 90.002	16	6.01(3) 7.02(1) 21.70(2) 90 90 90	916(6)	80.6	8.01	98.78	91.37	0.8
11	.../experiment_11/SMV	6.003 7.104 21.726 90.031 89.983 89.594	16	5.983(9) 7.20(1) 21.70(1) 90 90 90	935(3)	90.0	10.47	98.68	68.16	0.8
12	.../experiment_12/SMV	6.003 7.047 21.701 90.075 90.151 89.81	3	7.043(2) 6.01(1) 21.63(3) 90 90.6(1) 90	915(2)	93.1	7.83	96.12	78.99	0.9

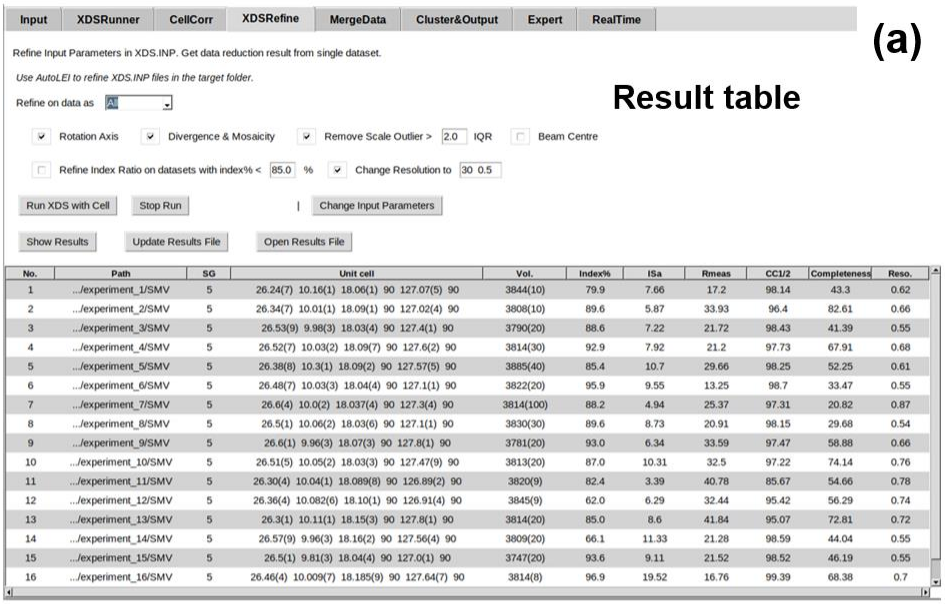


No.	Path	SG	Unit cell	Vol.	Index%	ISa	Rmeas	CC1/2	Completen	Reso.
1	.../experiment_1/SMV	16	5.995(7) 7.074(2) 21.748(4) 90 90 90	922(1)	70.1	9.4	15.32	99.1	57.78	0.55
2	.../experiment_2/SMV	16	5.997(6) 7.035(4) 21.797(8) 90 90 90	920(1)	78.2	11.75	12.6	99.16	57.13	0.55
3	.../experiment_3/SMV	16	6.01(1) 6.97(5) 21.77(2) 90 90 90	912(7)	89.3	7.25	14.58	97.33	33.46	0.62
4	.../experiment_4/SMV	16	6.00(1) 6.985(3) 21.761(6) 90 90 90	912(2)	81.7	7.5	17.09	98.67	50.25	0.56
5	.../experiment_5/SMV	16	6.01(1) 7.029(6) 21.75(5) 90 90 90	919(4)	80.1	5.09	24.73	96.11	74.3	0.74
6	.../experiment_6/SMV	16	6.00(1) 6.98(1) 21.73(1) 90 90 90	910(3)	85.3	6.33	22.13	97.73	49.63	0.56
8	.../experiment_8/SMV	16	6.02(1) 7.049(4) 21.75(1) 90 90 90	923(3)	89.4	6.39	18.53	97.69	58.18	0.54
9	.../experiment_9/SMV	16	5.996(8) 6.96(1) 21.83(1) 90 90 90	911(2)	93.6	8.13	14.78	98.98	52.98	0.54
10	.../experiment_10/SMV	16	6.01(3) 7.024(8) 21.69(1) 90 90 90	915(6)	83.7	7.5	19.42	98.5	87.61	0.68
11	.../experiment_11/SMV	16	5.987(3) 7.19(1) 21.732(4) 90 90 90	936(2)	90.9	11.14	11.22	98.99	45.53	0.56
12	.../experiment_12/SMV	16	6.01(1) 7.050(6) 21.77(4) 90 90 90	922(3)	92.7	7.2	24.32	97.39	66.31	0.57

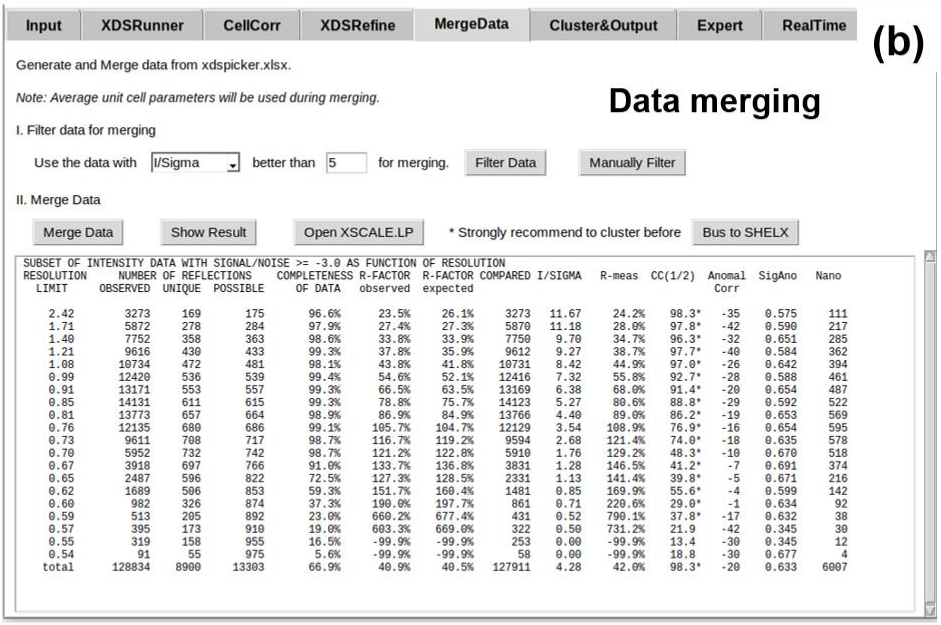
Figure 4 Datasets of 12 L-tyrosine crystals processed using AutoLEI. (a) Result table after running ***XDSrunner***. The suggested unit cell parameters and space group were extracted from the "Estimate Symmetry & Cell Cluster" function. (b) Result table after running ***XDSRefine***. "View Reciprocal Space" was designed to inspect the reciprocal space of a single dataset.

3.2. Interactive data merging of MOF SU-100

SU-100 is a monoclinic MOF material, and its flexible framework exhibits promising breathing behaviors (Grape *et al.*, 2020). In this study, a total of 16 datasets were collected using Instamatic, and all data were processable after ***XDSRefine*** as shown in Figure 5a. Data with an ISa higher than 5 were then merged during the ***XDSRefine*** step after specifying the space group and unit cell parameters (Figure 5b). We further conducted cluster analysis based on the pair-wise correlation coefficients of reflection intensities. As shown in Figure 5c, a threshold of 0.5 was selected due to the significant distance between four clusters. The final "HKL" file was obtained in 7 minutes. The resolution cut-off was 0.62 Å. As shown in Figure 3b, all non-H atoms could be directly determined from the solution. The number of unique reflections and parameters used for refinement was 7193 and 200, respectively. The final R_1 converged to 19.46% as shown in Table S2.



No.	Path	SG	Unit cell	Vol.	Index%	ISa	Rmeas	CC1/2	Completeness	Reso.
1	.../experiment_1/SMV	5	26.24(7) 10.16(1) 18.06(1) 90 127.07(5) 90	3844(10)	79.9	7.66	17.2	98.14	43.3	0.62
2	.../experiment_2/SMV	5	26.34(7) 10.01(1) 18.09(1) 90 127.02(4) 90	3808(10)	89.6	5.87	33.93	96.4	82.61	0.66
3	.../experiment_3/SMV	5	26.53(9) 9.98(3) 18.03(4) 90 127.4(1) 90	3790(20)	88.6	7.22	21.72	98.43	41.39	0.55
4	.../experiment_4/SMV	5	26.52(7) 10.03(2) 18.09(7) 90 127.6(2) 90	3814(30)	92.9	7.92	21.2	97.73	67.91	0.68
5	.../experiment_5/SMV	5	26.38(8) 10.3(1) 18.09(2) 90 127.57(5) 90	3885(40)	85.4	10.7	29.66	98.25	52.25	0.61
6	.../experiment_6/SMV	5	26.48(7) 10.03(3) 18.04(4) 90 127.1(1) 90	3822(20)	95.9	9.55	13.25	98.7	33.47	0.55
7	.../experiment_7/SMV	5	26.6(4) 10.0(2) 18.037(4) 90 127.3(4) 90	3814(100)	88.2	4.94	25.37	97.31	20.82	0.87
8	.../experiment_8/SMV	5	26.5(1) 10.06(2) 18.03(6) 90 127.1(1) 90	3830(30)	89.6	8.73	20.91	98.15	29.68	0.54
9	.../experiment_9/SMV	5	26.6(1) 9.96(3) 18.07(3) 90 127.8(1) 90	3781(20)	93.0	6.34	33.59	97.47	58.88	0.66
10	.../experiment_10/SMV	5	26.51(5) 10.05(2) 18.03(3) 90 127.47(9) 90	3813(20)	87.0	10.31	32.5	97.22	74.14	0.76
11	.../experiment_11/SMV	5	26.30(4) 10.04(1) 18.089(8) 90 126.89(2) 90	3820(9)	82.4	3.39	40.78	85.67	54.66	0.78
12	.../experiment_12/SMV	5	26.36(4) 10.082(6) 18.10(1) 90 126.91(4) 90	3845(9)	62.0	6.29	32.44	95.42	56.29	0.74
13	.../experiment_13/SMV	5	26.3(1) 10.11(1) 18.15(3) 90 127.8(1) 90	3814(20)	85.0	8.6	41.84	95.07	72.81	0.72
14	.../experiment_14/SMV	5	26.57(9) 9.96(3) 18.16(2) 90 127.56(4) 90	3809(20)	66.1	11.33	21.28	98.59	44.04	0.55
15	.../experiment_15/SMV	5	26.5(1) 9.81(3) 18.04(4) 90 127.0(1) 90	3747(20)	93.6	9.11	21.52	98.52	46.19	0.55
16	.../experiment_16/SMV	5	26.46(4) 10.009(7) 18.185(9) 90 127.64(7) 90	3814(8)	96.9	19.52	16.76	99.39	68.38	0.7



SUBSET OF INTENSITY DATA WITH SIGNAL/NOISE >= -3.0 AS FUNCTION OF RESOLUTION

RESOLUTION LIMIT	NUMBER OF REFLECTIONS OBSERVED	UNIQUE	POSSIBLE	COMPLETENESS OF DATA	R-FACTOR observed	R-FACTOR expected	COMPARED	I/SIGMA	R-meas	CC(1/2)	Anomal Corr	SigAno	Nano
2.42	3273	169	175	96.6%	23.5%	26.1%	3273	11.67	24.2%	98.3*	-35	0.575	111
1.71	5872	278	284	97.9%	27.4%	27.3%	5870	11.18	28.0%	97.8*	-42	0.590	217
1.40	7752	358	363	98.6%	33.8%	33.9%	7750	9.70	34.7%	96.3*	-32	0.651	285
1.21	9616	430	433	99.3%	37.8%	35.9%	9612	9.27	38.7%	97.7*	-40	0.584	362
1.08	10734	472	481	98.1%	43.8%	41.8%	10731	8.42	44.9%	97.0*	-26	0.642	394
0.99	12420	536	539	99.4%	54.6%	52.1%	12416	7.32	55.8%	92.7*	-28	0.588	461
0.91	13171	553	557	99.3%	66.5%	63.5%	13169	6.38	68.0%	91.4*	-20	0.654	487
0.85	14131	611	615	99.3%	78.8%	75.7%	14123	5.27	80.6%	88.8*	-29	0.592	522
0.81	13773	657	664	98.9%	86.9%	84.9%	13766	4.40	89.0%	86.2*	-19	0.653	569
0.76	12135	680	686	99.1%	105.7%	104.7%	12129	3.54	108.9%	76.9*	-16	0.654	595
0.73	9611	708	717	98.7%	116.7%	119.2%	9594	2.68	121.4%	74.0*	-18	0.635	578
0.70	5952	732	742	98.7%	121.2%	122.8%	5910	1.76	129.2%	48.3*	-10	0.670	518
0.67	3918	697	766	91.0%	133.7%	136.8%	3831	1.28	146.5%	41.2*	-7	0.691	374
0.65	2487	596	822	72.5%	127.3%	128.5%	2331	1.13	141.4%	39.8*	-5	0.671	216
0.62	1689	506	853	59.3%	151.7%	160.4%	1481	0.85	169.9%	55.6*	-4	0.599	142
0.60	982	326	874	37.3%	190.0%	197.7%	861	0.71	220.6%	29.0*	-1	0.634	92
0.59	513	205	892	23.0%	660.2%	677.4%	431	0.52	790.1%	37.8*	-17	0.632	38
0.57	395	173	910	19.0%	603.3%	669.0%	322	0.50	731.2%	21.9	-42	0.345	30
0.55	319	158	955	16.5%	-99.9%	-99.9%	253	0.00	-99.9%	13.4	-30	0.345	12
0.54	91	55	975	5.6%	-99.9%	-99.9%	58	0.00	-99.9%	18.8	-30	0.677	4
total	128834	8900	13303	66.9%	40.9%	40.5%	127911	4.28	42.0%	98.3*	-20	0.633	6007

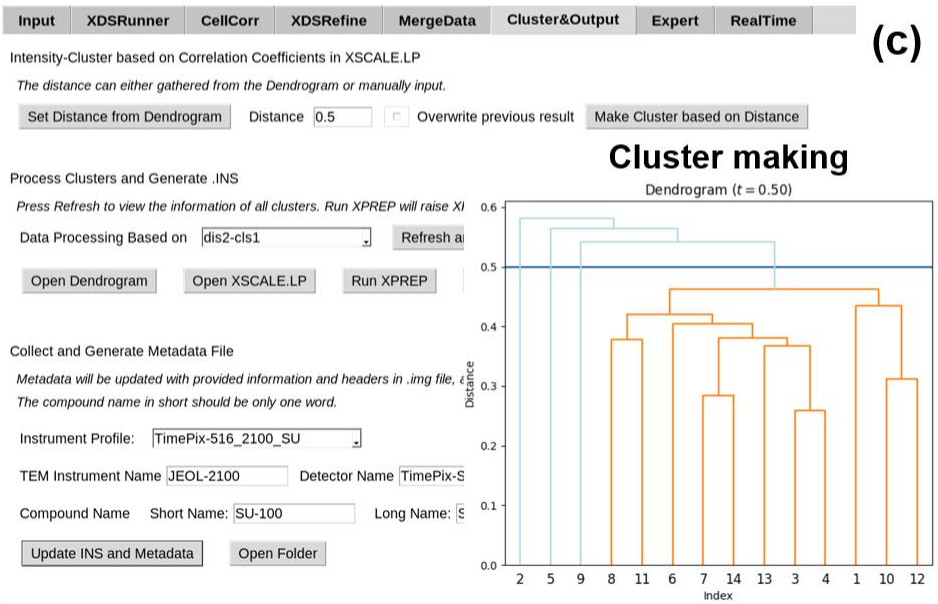
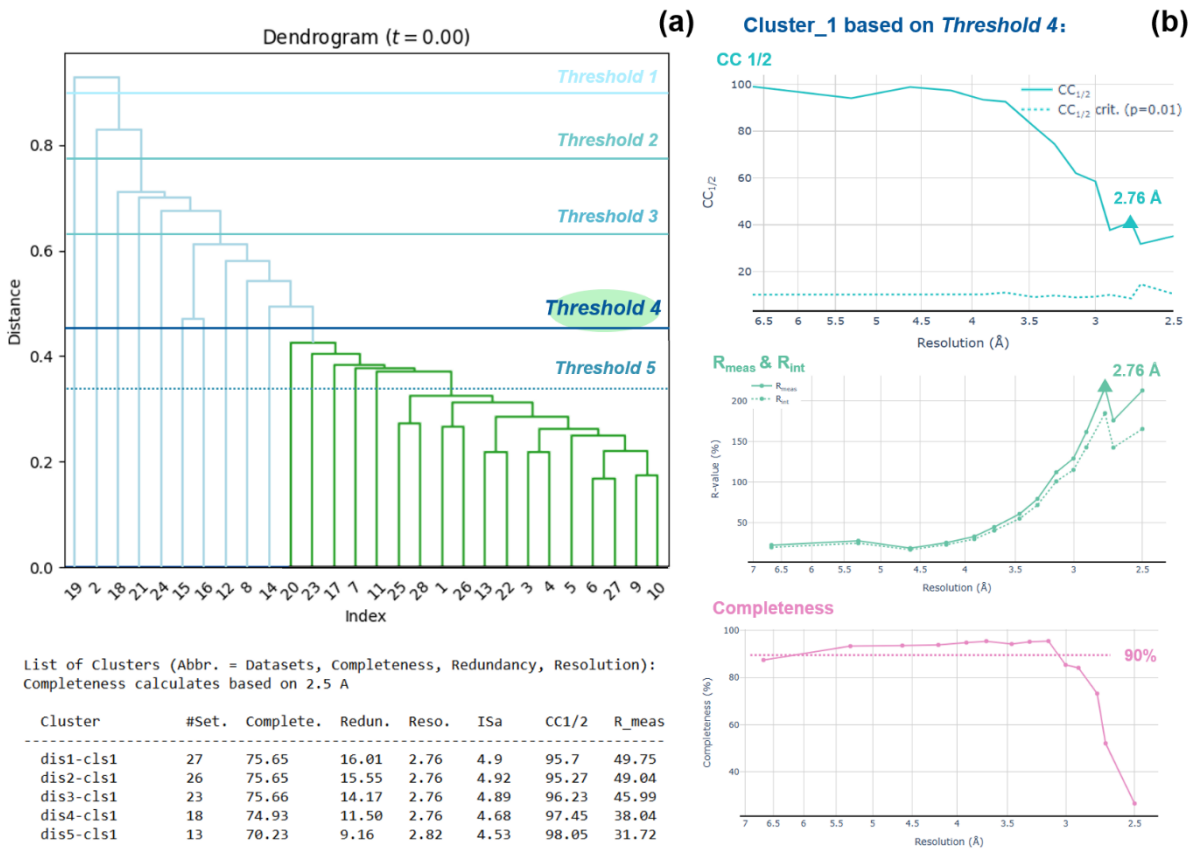


Figure 5 Data processing, analysis and merging of 16 SU-100 MOF datasets by AutoLEI. (a) ***XDSRefine*** work page showing the results table. (b) ***MergeData*** work page. The initial round of data merging was achieved after several clicks. (c) ***Cluster_Output*** work page. The clusters based on pair-wise reflection intensity correlations were made, and a final "HKL" file was prepared for structure determination.

3.3. Processing small wedge MicroED data of protein MutT homolog 1 (hMTH1)

Protein crystals are known for being extremely beam-sensitive during MicroED data collection (Shi *et al.*, 2013). To address this issue, a small-wedge data collection strategy was adopted for MicroED data collection. However, this strategy produces a large number of datasets that are of low completeness individually. Thus, repetitive data processing, data scaling and merging are required for structure determination, which is an ideal application for AutoLEI.

hMTH1 plays an important role in sanitizing oxidized dNTPs from the free nucleotide pool, preventing their incorporation into DNA, which reduces genotoxicity (Yoshimura *et al.*, 2003; Jemth *et al.*, 2018). Here, we used hMTH1 as an example to demonstrate how small wedge MicroED data can be effortlessly processed and merged with AutoLEI. A total of 38 datasets were collected using EPUD and the average rotation range of each crystal was limited to 15-20 degrees to boost the signal-to-noise ratio while reducing beam damage. Figure 6a shows the dendrogram of the clustering analysis after the ***XDSRefine*** step. AutoLEI also enables a stepwise clustering analysis, which is useful when processing protein data. Different distance thresholds were tested in order to find the optimal distance threshold. Based on this analysis, the first four groups containing 27, 26, 23 and 18 datasets showed comparable completeness while the completeness of the last group (contained 13 datasets) dropped by more than 4%. Thus, Cluster 1 based on threshold 4 (dis4-cls1) was selected for higher $CC_{1/2}$ and lower R_{meas}. A web report was then generated and used for data reduction of this cluster. Figure 6b indicates the relationships between $CC_{1/2}$, R_{meas} and completeness against resolution printed from the report of dis4-cls1. The estimated resolution was 2.76 Å and it took less than 12 minutes to obtain the final "HKL" file using AutoLEI. Molecular replacement and structure refinement were carried out using the Phenix package (Liebschner *et al.*, 2019). As shown in Table S3, the final R_{work} and R_{free} converged to 21.07% and 26.82%, respectively.




```
List of Clusters (Abbr. = Datasets, Completeness, Redundancy, Resolution):
Completeness calculates based on 2.5 A

  Cluster          #Set.  Complete.  Redun.  Reso.   ISa     CC1/2   R_meas
-------------------------------------------------------------------------------
  dis1-cls1        27     75.65      16.01   2.76    4.9     95.7    49.75
  dis2-cls1        26     75.65      15.55   2.76    4.92    95.27   49.04
  dis3-cls1        23     75.66      14.17   2.76    4.89    96.23   45.99
  dis4-cls1        18     74.93      11.50   2.76    4.68    97.45   38.04
  dis5-cls1        13     70.23      9.16    2.82    4.53    98.05   31.72
```

Figure 6 Data processing, analysis and merging of 38 hMTH1 datasets by AutoLEI. (a) Clustering analysis after data merging in the AutoLEI interface. A total of 5 cluster selections were made based on different distances. The summary of the key statistics of these 5 clusters was tabulated in the command window. (b) Web report plots of Cluster_1 based on threshold 4 (18 datasets in total). Plots were captured from the web report.

3.4. Real-time data process of triclinic lysozyme

Real-time data processing of single crystal X-ray diffraction is implemented at most major X-ray synchrotron facilities. This capability is also desirable for MicroED, as it could provide live information on current batch data collection, allowing the user to stop collection when enough data has been collected. Moreover, live processing statistics can also provide useful information for optimizing the data collection strategy. By combining AutoLEI with automated data collection protocols, rapid structure determination and phase analysis becomes possible. Here, we used triclinic lysozyme crystals to demonstrate the online MicroED data processing capability of AutoLEI.

Triclinic lysozyme is a challenging sample for MicroED, considering its low symmetry and preferred crystal orientation. Here, the small wedge data collection strategy was also employed to increase data quality. The average tilt range of individual crystals was limited to 15 degrees, making data collection

more challenging. As shown in Figure 7a, to achieve real-time data merging (Merge Mode), unit cell parameters and space group need to be given as prior knowledge. These two input parameters could be easily obtained by first collecting a high-tilt MicroED dataset. In the case of triclinic lysozyme, the expected resolution limit for inspecting live statistics was set to 1.3 Å. Data with $CC_{1/2}$ higher than 90% as well as ISa higher than 5 were considered to be good data and used for real-time merging iterations. These parameters could also be customized by editing the "strategy file settings". Furthermore, the resolution of each good dataset as well as the statistics of merged data completeness and $CC_{1/2}$ was updated live. The user can then use this information to check if the current data collection strategy or data filter strategy should be modified.

After approximately 6 hours of data collection, a total of 71 datasets were collected. The completeness of the data reached 95.8%, while the $CC_{1/2}$ was 98.68% at the resolution cut-off of 1.3 Å. Notably, during data collection, the real-time completeness plot revealed a period during which completeness was not significantly improved, indicating preferred crystal orientation. This suggests that the data collection strategy should be altered. We then looked for crystals with different morphologies on the TEM grids. Data collected from those crystals further improved the data completeness. In such an instance, MicroED data collected from high tilting angle might also be considered. The data collection was terminated after the completeness of the merged data reached higher than 95% and the $CC_{1/2}$ was above 98% at the pre-set resolution cut-off of 1.3 Å. The web report was then generated to estimate the real high-resolution cut-off. As shown in Figure 7b and 7c, the resolution cut-off was finally set at 1.1 Å, before the $CC_{1/2}$ dropped below 30%. The "HKL" file of the final merging iteration was used immediately for phasing and refinement using Phenix (McCoy *et al.*, 2007; Afonine *et al.*, 2012). The final refinement showed that the R_{work} and R_{free} converged to 20.42% and 23.45% respectively. The non-filled $2F_o-F_c$ electrostatic potential map with the lysozyme structure is presented in Figure 3d.

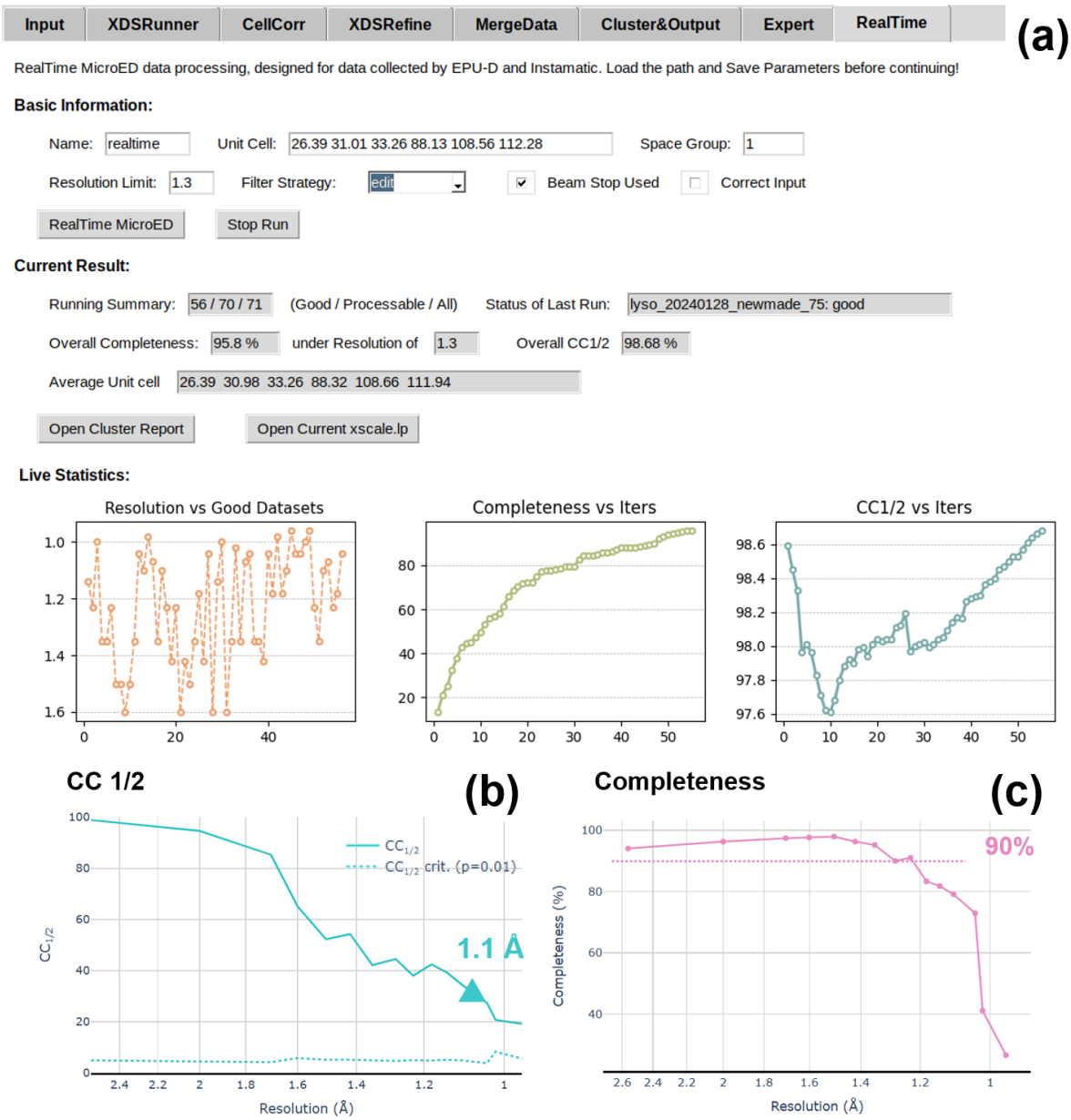


Figure 7 Real-time MicroED data processing of triclinic lysozyme datasets using AutoLEI. (a) The live processing results are displayed under "Live Statistics". (b) $CC_{1/2}$ and (c) Completeness against resolution plots captured from the web report.

4. Conclusion and outlook

In this paper, we developed an XDS-based GUI, AutoLEI for offline and real-time batch MicroED data processing. We presented four examples utilizing different types of samples to demonstrate the simplicity and usability of AutoLEI, illustrating its potential for fully automated high-throughput structure analyses, such as high-speed ligand screening for structure-based drug discovery and quantitative analysis of heterogeneous phases. The current version of AutoLEI still presents

opportunities for further development. We expect to incorporate additional data processing engines, enabling comparative analyses and data integration in subsequent versions. These developments will simplify MicroED data processing and lower the entry barriers for researchers new to the field of MicroED.

Acknowledgements This work was supported by the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement no. 956099 (NanED − Electron Nanocrystallography−H2020-MSCAITN). We also acknowledge financial support from the Swedish Research Council (VR, 2019-00815 (XZ), 2022-03681 (PS) and 2022-03596 (HY)), and the Knut and Alice Wallenberg Foundation (KAW, 2019.0124). We thank Dr. Stef Smith and Dr. Taimin Yang for sharing the code of Edtools. The MicroED data of protein samples were collected at the Cryo-EM Swedish National Facility funded by KAW, Family Erling Persson and Kempe Foundations, SciLifeLab, Stockholm University and Umeå University. We also thank Dr. Mathieu Coinçon, Dr. Dustin Morado and Dr. Marta Carroni for TEM support at Scilifelab.

Data available

The GUI is available from Zenodo (https://zenodo.org/records/15155988).

The data can be downloaded from Zenodo (https://zenodo.org/records/14536385).

References

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

Bengtsson, V. E. G., Pacoste, L., De La Rosa-Trevin, J. M., Hofer, G., Zou, X. & Xu, H. (2022). *J Appl Crystallogr* **55**, 638–646.

Brehm, W., Triviño, J., Krahn, J. M., Usón, I. & Diederichs, K. (2023). *J Appl Crystallogr* **56**, 1585–1594.

Cheng, A., Negro, C., Bruhn, J. F., Rice, W. J., Dallakyan, S., Eng, E. T., Waterman, D. G., Potter, C. S. & Carragher, B. (2021). *Protein Science* **30**, 136–150.

Cho, J., Willhammar, T. & Zou, X. (2023). *Microporous and Mesoporous Materials* **358**, 112400.

Cichocka, M. O., Ångström, J., Wang, B., Zou, X. & Smeets, S. (2018). *J Appl Crystallogr* **51**, 1652–1661.

De La Cruz, M. J., Martynowycz, M. W., Hattne, J. & Gonen, T. (2019). *Ultramicroscopy* **201**, 77–80.

Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. (2010). *Acta Crystallogr D Biol Crystallogr* **66**, 486–501.

Gemmi, M., Mugnaioli, E., Gorelik, T. E., Kolb, U., Palatinus, L., Boullay, P., Hovmöller, S. & Abrahams, J. P. (2019). *ACS Cent. Sci.* **5**, 1315–1329.

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 Crystallogr D Struct Biol* **78**, 752–769.

Grape, E. S., Xu, H., Cheung, O., Calmels, M., Zhao, J., Dejoie, C., Proserpio, D. M., Zou, X. & Inge, A. K. (2020). *Crystal Growth & Design* **20**, 320–329.

Guo, P., Shin, J., Greenaway, A. G., Min, J. G., Su, J., Choi, H. J., Liu, L., Cox, P. A., Hong, S. B., Wright, P. A. & Zou, X. (2015). *Nature* **524**, 74–78.

Huang, Z., Grape, E. S., Li, J., Inge, A. K. & Zou, X. (2021). *Coordination Chemistry Reviews* **427**, 213583.

Huang, Z., Willhammar, T. & Zou, X. (2021). *Chem. Sci.* **12**, 1206–1219.

Jemth, A.-S., Gustafsson, R., Bräutigam, L., Henriksson, L., Vallin, K. S. A., Sarno, A., Almlöf, I., Homan, E., Rasti, A., Warpman Berglund, U., Stenmark, P. & Helleday, T. (2018). *Nucleic Acids Research* https://doi.org/10.1093/nar/gky896.

Kabsch, W. (2010). *Acta Crystallogr D Biol Crystallogr* **66**, 125–132.

Kolb, U., Gorelik, T., Kübel, C., Otten, M. T. & Hubert, D. (2007). *Ultramicroscopy* **107**, 507–513.

Liebschner, D., Afonine, P. V., Baker, M. L., Bunkóczi, G., Chen, V. B., Croll, T. I., Hintze, B., Hung, L.-W., Jain, S., McCoy, A. J., Moriarty, N. W., Oeffner, R. D., Poon, B. K., Prisant, M. G., Read, R. J., Richardson, J. S., Richardson, D. C., Sammito, M. D., Sobolev, O. V., Stockwell, D. H., Terwilliger, T. C., Urzhumtsev, A. G., Videau, L. L., Williams, C. J. & Adams, P. D. (2019). *Acta Crystallogr D Struct Biol* **75**, 861–877.

Lightowler, M., Li, S., Ou, X., Cho, J., Liu, B., Li, A., Hofer, G., Xu, J., Yang, T., Zou, X., Lu, M. & Xu, H. (2024). *Angewandte Chemie International Edition* **63**, e202317695.

Lightowler, M., Li, S., Ou, X., Zou, X., Lu, M. & Xu, H. (2022). *Angew Chem Int Ed* **61**, e202114985.

Lu, P., Xu, J., Sun, Y., Guillet-Nicolas, R., Willhammar, T., Fahda, M., Dib, E., Wang, B., Qin, Z., Xu, H., Cho, J., Liu, Z., Yu, H., Yang, X., Lang, Q., Mintova, S., Zou, X. & Valtchev, V. (2024). *Nature* **636**, 368–373.

Luo, Y., Wang, B., Smeets, S., Sun, J., Yang, W. & Zou, X. (2023). *Nat. Chem.* https://doi.org/10.1038/s41557-022-01131-8.

Martynowycz, M. W., Clabbers, M. T. B., Hattne, J. & Gonen, T. (2022). *Nat Methods* **19**, 724–729.

McCoy, A. J., Grosse-Kunstleve, R. W., Adams, P. D., Winn, M. D., Storoni, L. C. & Read, R. J. (2007). *J Appl Crystallogr* **40**, 658–674.

Palatinus, L., Brázda, P., Jelínek, M., Hrdá, J., Steciuk, G. & Klementová, M. (2019). *Acta Crystallogr B Struct Sci Cryst Eng Mater* **75**, 512–522.

Samperisi, L., Jaworski, A., Kaur, G., Lillerud, K. P., Zou, X. & Huang, Z. (2021). *J. Am. Chem. Soc.* **143**, 17947–17952.

Shi, D., Nannenga, B. L., Iadanza, M. G. & Gonen, T. (2013). *eLife* **2**, e01345.

Truong, K.-N., Ito, S., Wojciechowski, J. M., Göb, C. R., Schürmann, C. J., Yamano, A., Del Campo, M., Okunishi, E., Aoyama, Y., Mihira, T., Hosogi, N., Benet-Buchholz, J., Escudero-Adán, E. C., White, F. J., Ferrara, J. D. & Bücker, R. (2023). *Symmetry* **15**, 1555.

Unge, J., Lin, J., Weaver, S. J., Sae Her, A. & Gonen, T. (2024). *Advanced Science* **11**, 2400081.

Van Genderen, E., Clabbers, M. T. B., Das, P. P., Stewart, A., Nederlof, I., Barentsen, K. C., Portillo, Q., Pannu, N. S., Nicolopoulos, S., Gruene, T. & Abrahams, J. P. (2016). *Acta Crystallogr A Found Adv* **72**, 236–242.

Wan, W., Sun, J., Su, J., Hovmöller, S. & Zou, X. (2013). *J Appl Crystallogr* **46**, 1863–1873.

Wang, B., Zou, X. & Smeets, S. (2019). *IUCrJ* **6**, 854–867.

Winter, G., Beilsten-Edmands, J., Devenish, N., Gerstel, M., Gildea, R. J., McDonagh, D., Pascal, E., Waterman, D. G., Williams, B. H. & Evans, G. (2022). *Protein Science* **31**, 232–250.

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 Crystallogr D Struct Biol* **74**, 85–97.

Xu, H., Lebrette, H., Clabbers, M. T. B., Zhao, J., Griese, J. J., Zou, X. & Högbom, M. (2019). *Sci. Adv.* **5**, eaax4621.

Xu, H., Lebrette, H., Yang, T., Srinivas, V., Hovmöller, S., Högbom, M. & Zou, X. (2018). *Structure* **26**, 667-675.e3.

Yoshimura, D., Sakumi, K., Ohno, M., Sakai, Y., Furuichi, M., Iwai, S. & Nakabeppu, Y. (2003). *Journal of Biological Chemistry* **278**, 37965–37973.

Zhang, D., Oleynikov, P., Hovmöller, S. & Zou, X. (2010). *Zeitschrift Für Kristallographie* **225**, 94–102.