

# *Multi-Crystal Methods*

Dr Amy Thompson

PDRA – VMXi Beamline – Diamond Light Source

xia<sup>2</sup>

diamond

DIALS  
Diffraction Integration for Advanced Light Sources

# *Why use Multi-Crystal Strategies for MX?*

In some cases, it is only possible to collect incomplete data from a single crystal due to radiation damage: obtain a complete dataset by combining data from multiple crystals.

This strategy allows for the use of:

- small crystals
- room-temperature data collection
- *In situ* data collection

Room temperature provides complementary information to your standard cryo-experiment:

- Cryo-cooling may hide biologically significant structural features
- Flexibility may show multiple states when large numbers of samples are collected

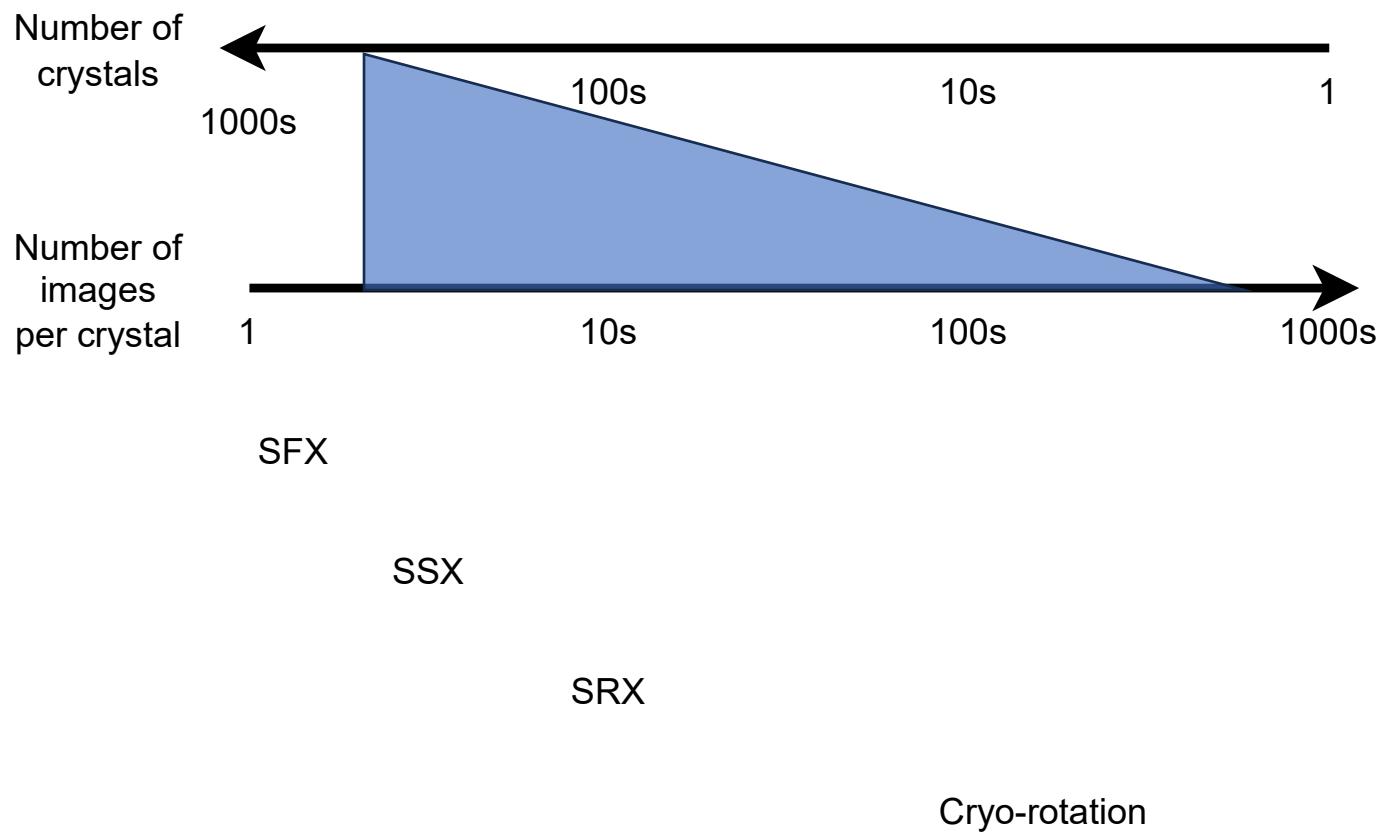


*The VMXi Beamline specializes in room temperature data collection *in-situ**

# *Multi-Crystal Approaches*

Three MX Beamlines at Diamond that make extensive use of multi-crystal data collections:

- I24
- VMXi
- VMXm



# *Advantages and Challenges of Multi-Crystal Strategies*

## **Advantages:**

- Improved signal-to-noise
- Better CC<sub>1/2</sub> at higher resolution
- Overcome challenges imposed by radiation sensitive samples

## **Challenges:**

- Consistent symmetry determination
- Non-isomorphism
- Preferential orientation

# Pitfalls with “Standard” Data Reduction Statistics

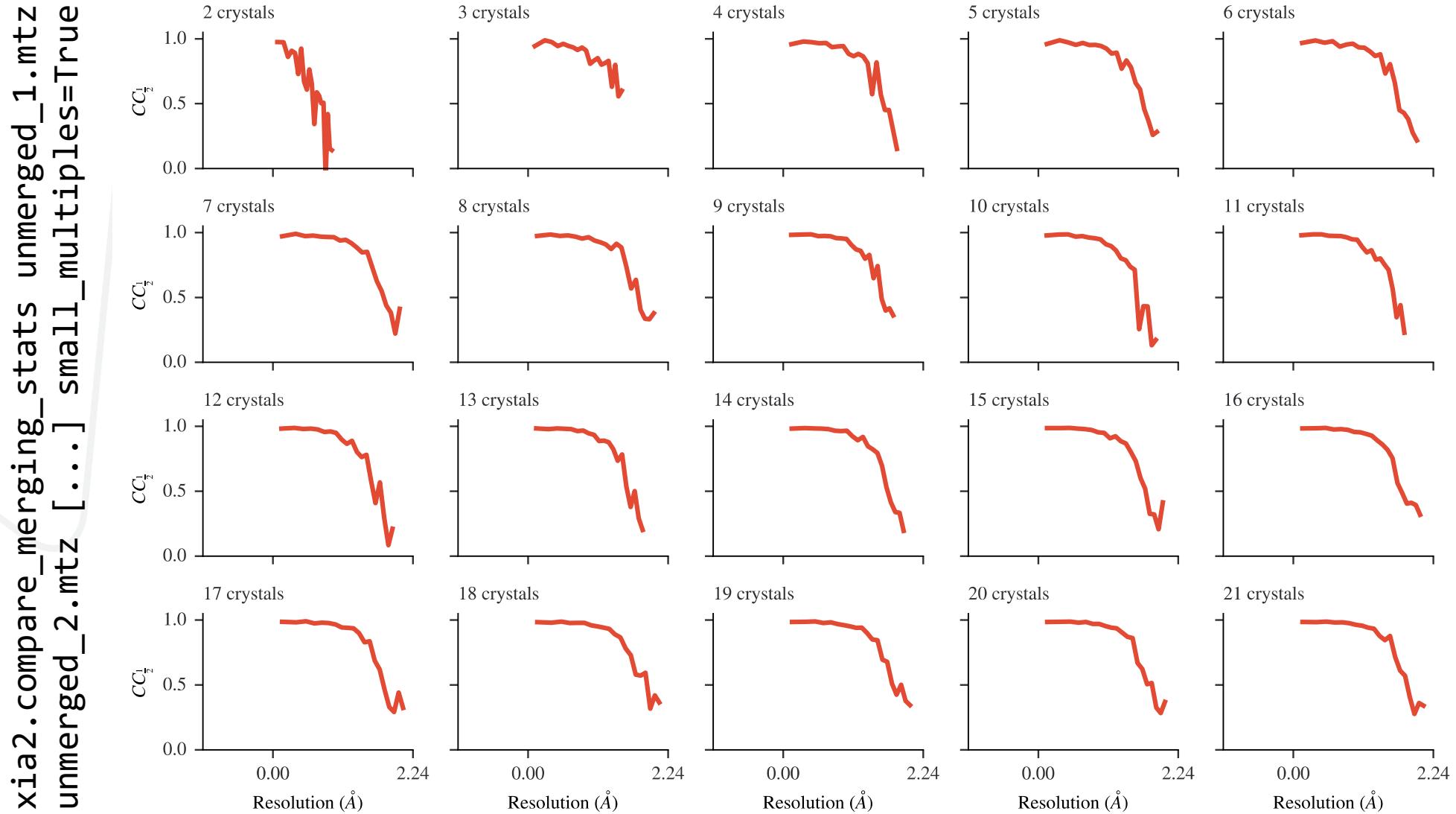
- Given only random errors, the standard error  $\sigma$  in a measurement is reduced by  $\sqrt{n}$  if the measurement is repeated  $n$  times
- Averaging improves accuracy and identifies / minimizes the effect of outliers
- As you increase the number of measurements, you become more certain of your final intensity value - **Means we get better signal-to-noise ratio!**

$$\sigma_x \approx \frac{S}{\sqrt{N}}$$

- Beware  $R_{\text{merge}}$  and  $R_{\text{meas}}$ !
- Use  $CC_{1/2}$  and  $\langle I/\sigma \rangle_{\text{mrgd}}$

Dataset	Big	Tiny	T100	Big+T100	Big2
Multiplicity	2	2	200	202	4
$\langle I/\sigma \rangle_{\text{ind}}$	2.0	0.2	0.2	0.22	2.0
$R_{\text{merge}}$	28%	280%	399%	395%	35%
$R_{\text{meas}}$	40%	400%	400%	396%	40%
$R_{\text{pim}}$	28%	280%	28%	28%	20%
$\langle I/\sigma \rangle_{\text{mrgd}}$	2.8	0.28	2.8	4.0	4.0
$CC_{1/2}$	0.66	0.04	0.66	0.80	0.80

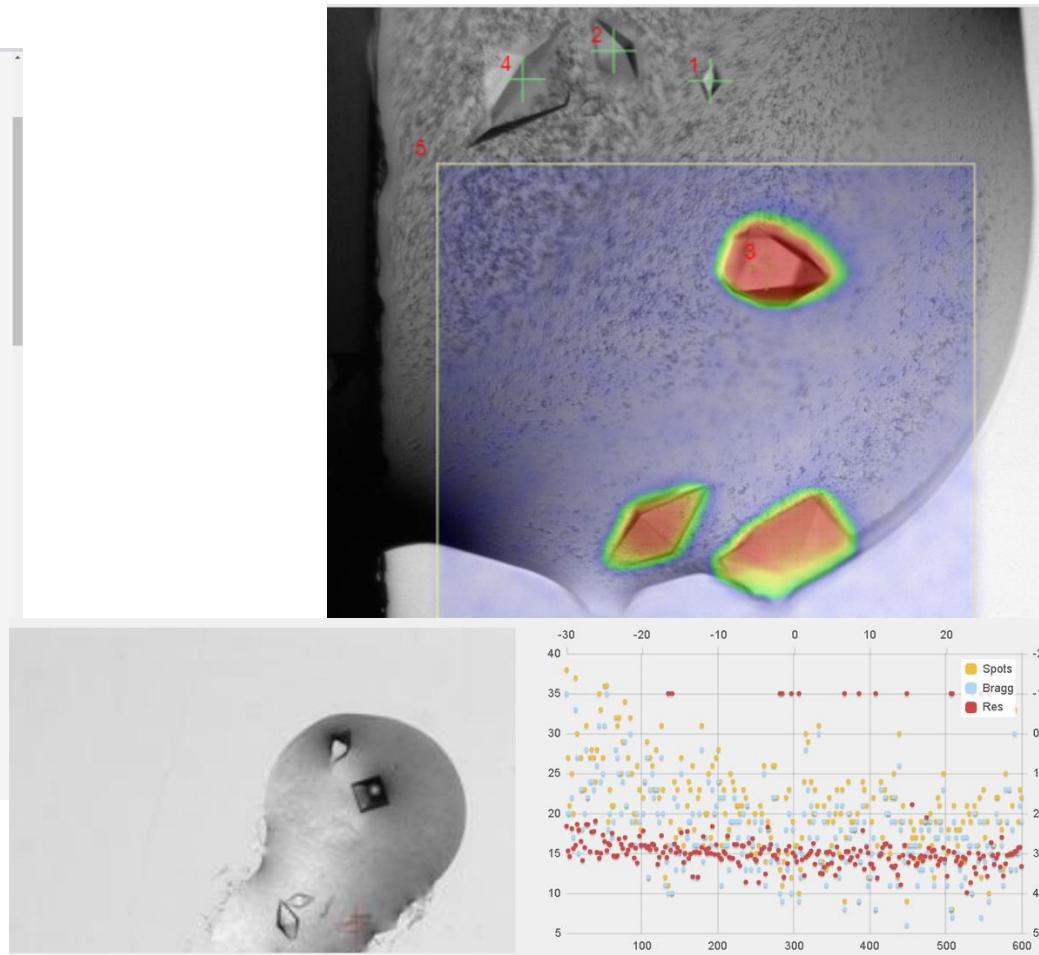
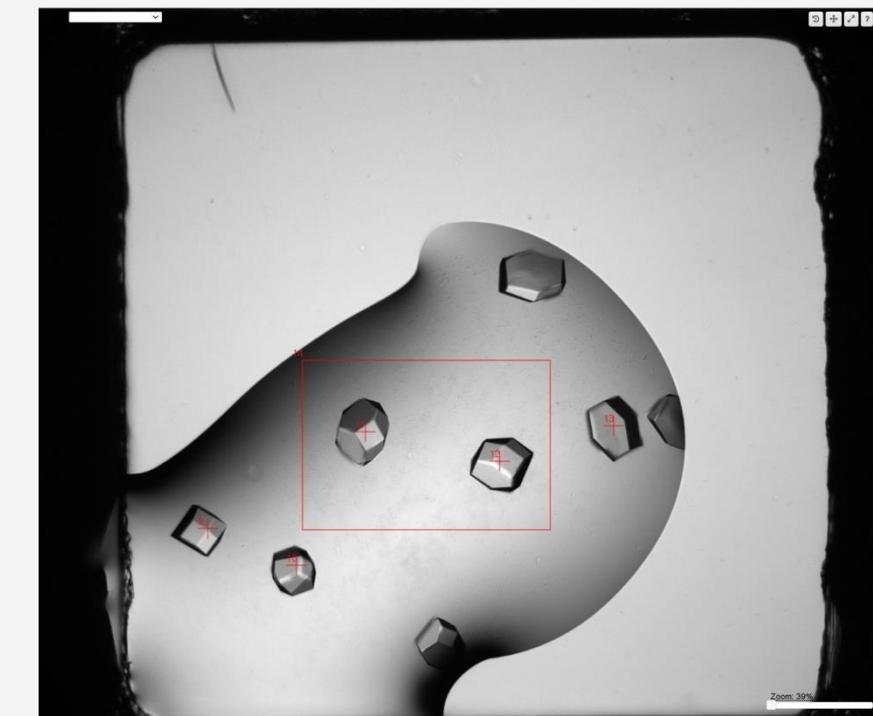
# Better CC1/2 at higher resolution



# VMXi – Collect at Room Temperature

Container: kangaroo\_AB2173\_lyso

	1	2	3	4	5	6	7	8	9	10	11	12
A	1 2	0 3	0 4	1 5	0 6	0 7	0 8	0 9	0 10	0 11	0 12	0 13
B	25 26	27 28	29 30	31 32	33 34	35 36	37 38	39 40	41 42	43 44	45 46	47 48
C	49 50	51 52	53 54	55 56	57 58	59 60	61 62	63 64	65 66	67 68	69 70	71 72
D	73 74	75 76	77 78	79 80	81 82	83 84	85 86	87 88	89 90	91 92	93 94	95 96
E	97 98	99 100	101 102	103 104	105 106	107 108	109 110	111 112	113 114	115 116	117 118	119 120
F	121 122	123 124	125 126	127 128	129 130	131 132	133 134	135 136	137 138	139 140	141 142	143 144
G	145 146	147 148	149 150	151 152	153 154	155 156	157 158	159 160	161 162	163 164	165 166	167 168
H	169 170	171 172	173 174	175 176	177 178	179 180	181 182	183 184	185 186	187 188	189 190	191 192



Discover the potential of your crystals without intervention  
Obtain many RT protein structures to explore dynamics, ligand binding etc.

Fisher, S.J., Levik, K. et al. (2015). SynchWeb: A Modern Interface For ISPyB. *J.Appl.Cryst.* 48, 927-932

Sandy, J., Mikolajek, H., et al. (2024). Crystallization and *In Situ* Room Temperature Data Collection Using the Crystallization Facility at Harwell and Beamline VMXi, Diamond Light Source. *J.Vis. Exp.* (205), e65964.



# *Advantages and Challenges of Multi-Crystal Strategies*

## **Advantages:**

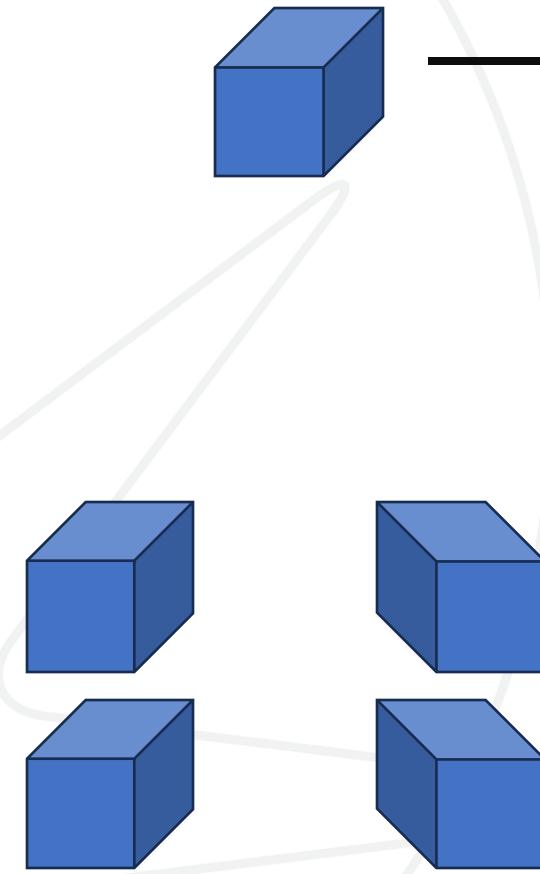
- Improved signal-to-noise
- Better CC<sub>1/2</sub> at higher resolution
- Overcome challenges imposed by radiation sensitive samples

## **Challenges:**

- Consistent symmetry determination
- Preferential orientation
- Non-isomorphism

# *Symmetry Determination (using dials.cosym)*

- Identification of consensus symmetry from narrow wedges or stills can be challenging
- Complicated by presence of potential indexing ambiguity

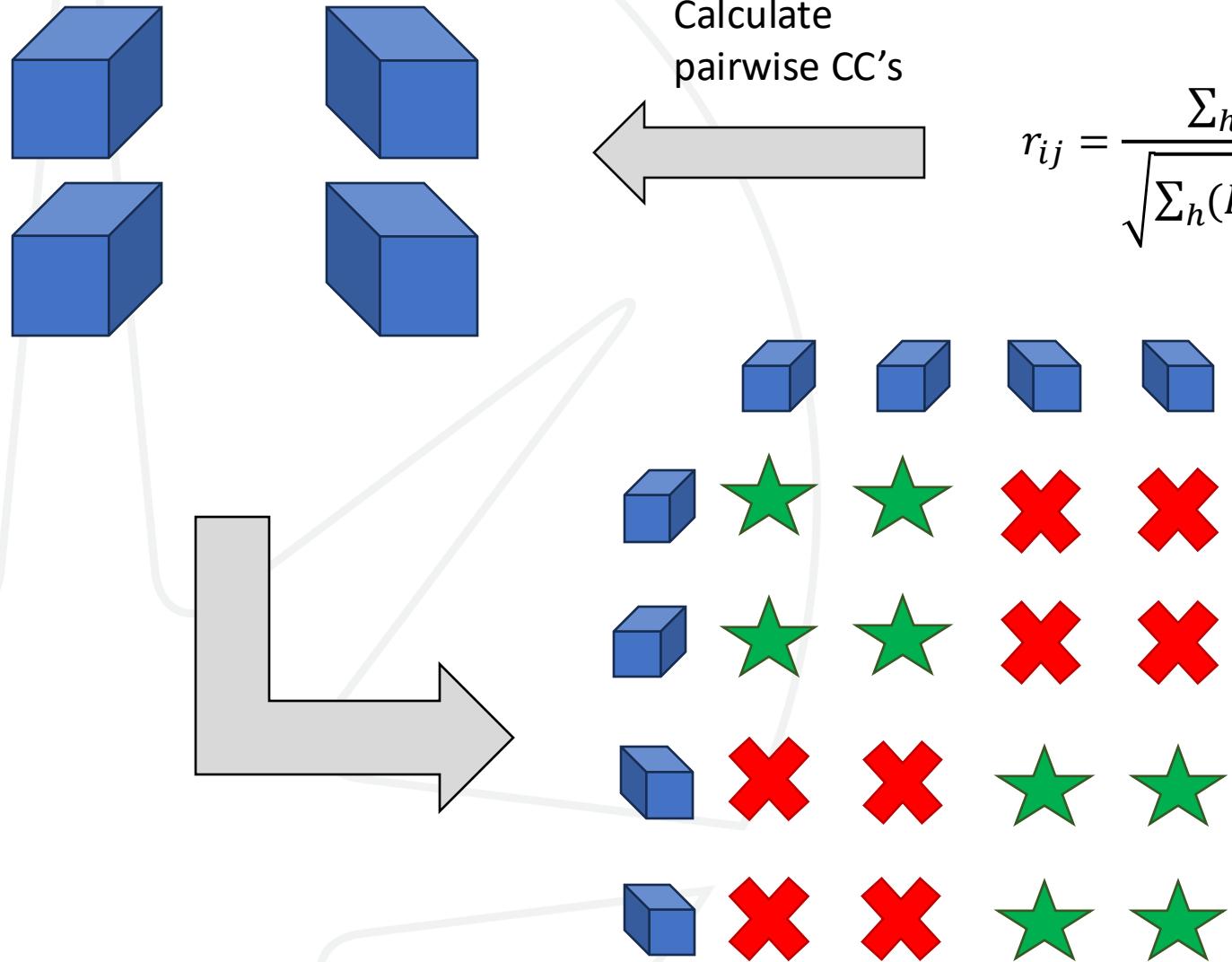


Has an indexing ambiguity

Point Group	Space Group	Point Group	Space Group
3	$P_3, P_{3_1}, P_{3_2}$	4	$P4_x (x = 0 \dots 3)$
6	$P6_x (x = 0 \dots 5)$	4	$I4, I4_1$
312	$P312, P_{3_1}12, P_{3_2}12$	3	$R3 (H3)$
321	$P321, P_{3_1}21, P_{3_2}21$	23	$P23, F23, I23, P_{2_1}3, I_{2_1}3$

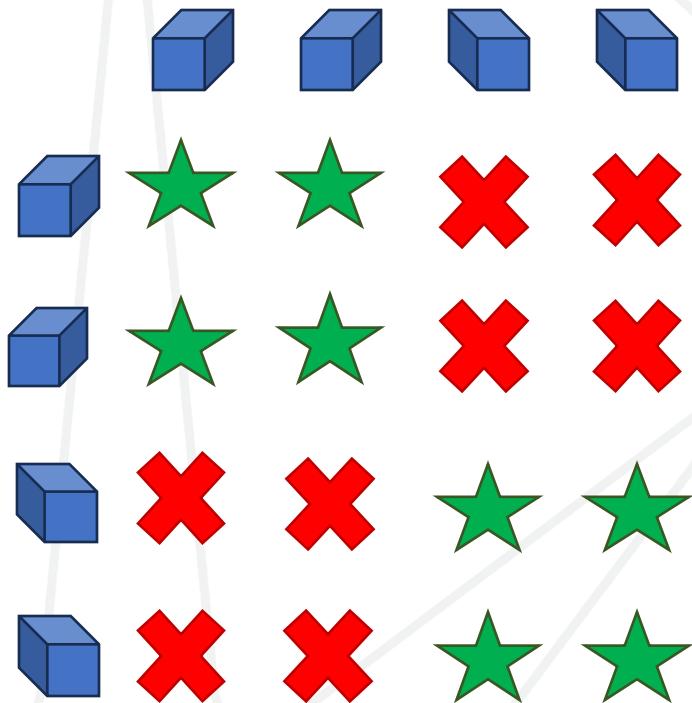
If we measure 4 incomplete datasets, how do we merge if they are indexed in different modes?

# Symmetry Determination



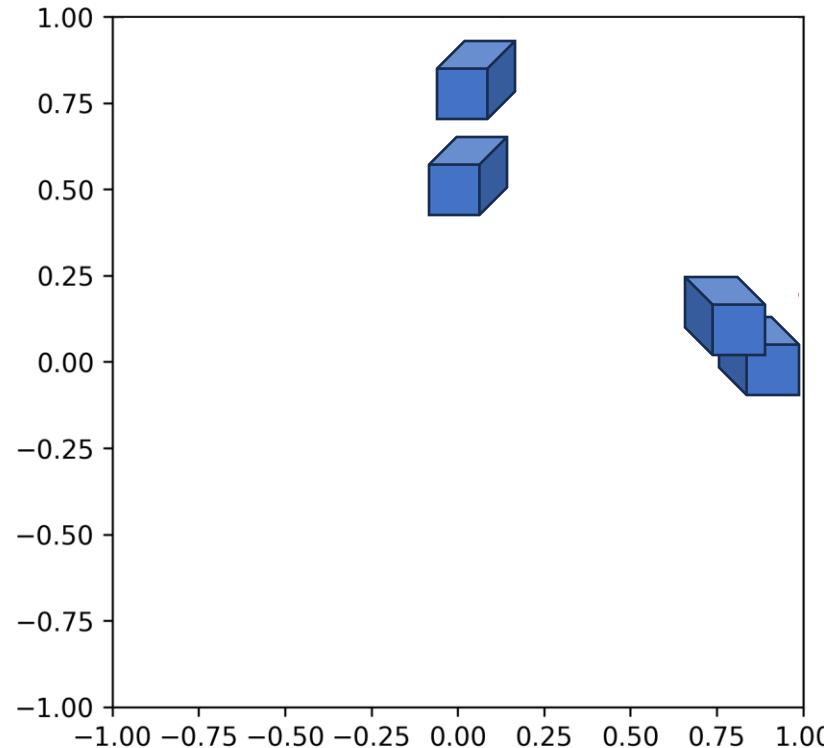
$$r_{ij} = \frac{\sum_h (I_i(h) - \bar{I}_i)(I_j(h) - \bar{I}_j)}{\sqrt{\sum_h (I_i(h) - \bar{I}_i)^2 \sum_h (I_j(h) - \bar{I}_j)^2}}$$

# Symmetry Determination



Represents a 4-dimensional problem – dimensionality reduction is needed (as datasets usually much larger than this!)

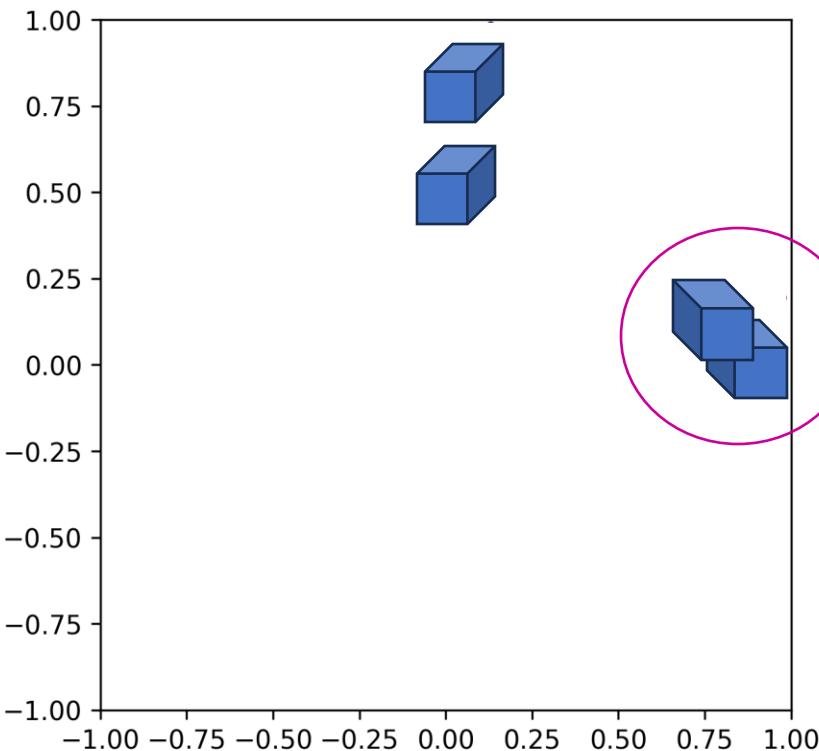
$$\Phi = \sum_{i=1}^{N-1} \sum_{j=i+1}^N (r_{ij} - \mathbf{x}_i \cdot \mathbf{x}_j)^2$$



Represent pairwise correlations as coordinates in reduced dimensional space (the lowest dimension that can represent the matrix with minimal residuals)

Plotting coords groups by like indexing

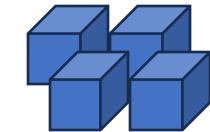
# Symmetry Determination



Apply twinning operator



Merge with  
other cluster

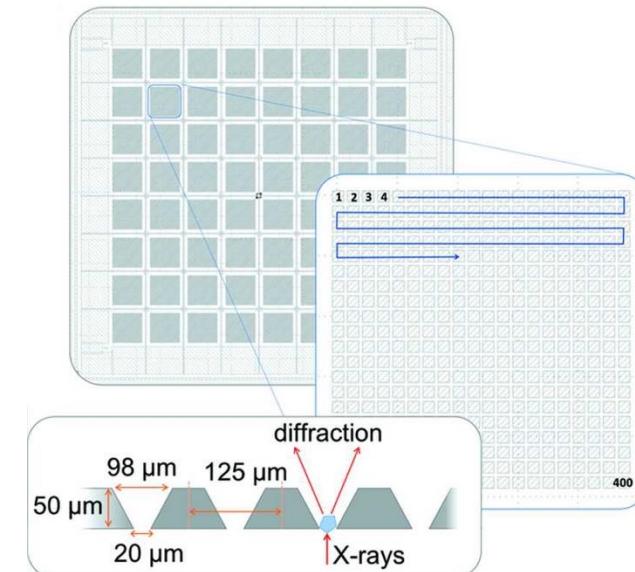
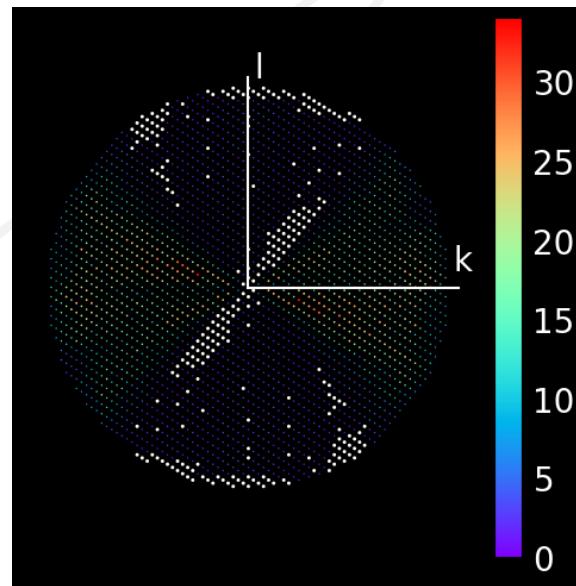


Non-twinned  
Dataset

dials.cosym extends these principles to also determine the symmetry of the dataset – it does not have to be previously known!

# *Preferential Orientation*

- Crystal symmetry and morphology combined with data collection conditions may lead to preferential crystal orientation
- May result in under-sampled regions of reciprocal space
- Check stereographic projection and multiplicity plots



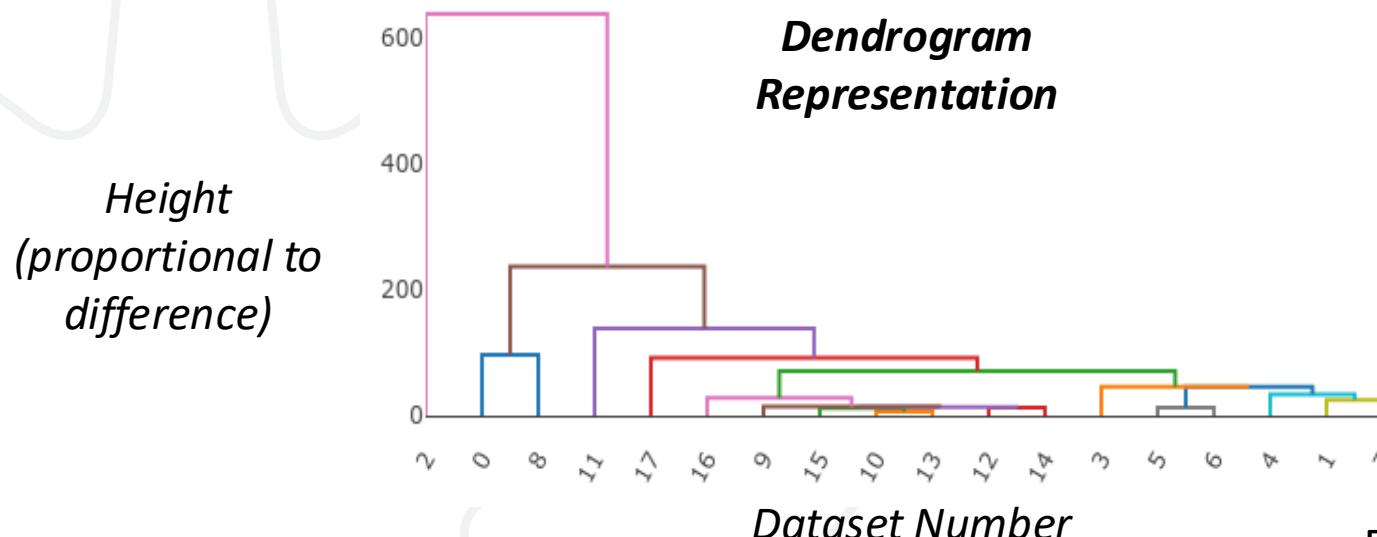
# *Clustering Methods to Achieve Isomorphous Data*

Achieving isomorphous data from multi-crystal collections is critical, although differences can still exist within structurally isomorphous crystals (ligand binding, alternative confirmations, etc).

First requires a known and consistent symmetry (methods from Brehm and Diederichs)

## **Unit Cell Clustering**

- Good to detect structural non-isomorphism
- Clusters on unit cell dimensions



## **Intensity-based Clustering**

- Good to detect differences in structurally isomorphous datasets
- Clusters on pair-wise correlation coefficients

Hierarchical Cluster Analysis (HCA) uses dendrograms to visualise similarity between datasets

# Hierarchical Clustering in *xia2.multiplex*

After consistent symmetry determination, there are three types of clustering available:

1. Unit Cell Clustering
2. Correlation Clustering (intensity-based)
3. Cosine Angle Clustering (intensity-based)

## Correlation Clustering

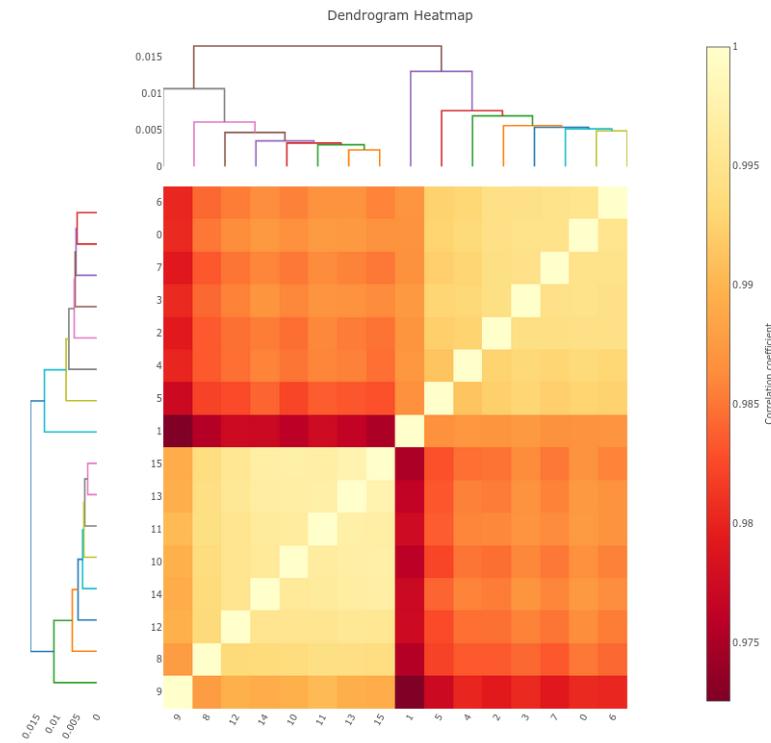
- Comparison of pairwise correlation coefficients
- Does not distinguish types of error resulting in differences in intensities

$$r_{ij} = \frac{\sum_h (I_i(h) - \bar{I}_i)(I_j(h) - \bar{I}_j)}{\sqrt{\sum_h (I_i(h) - \bar{I}_i)^2 \sum_h (I_j(h) - \bar{I}_j)^2}}$$

## Cosine Angle Clustering

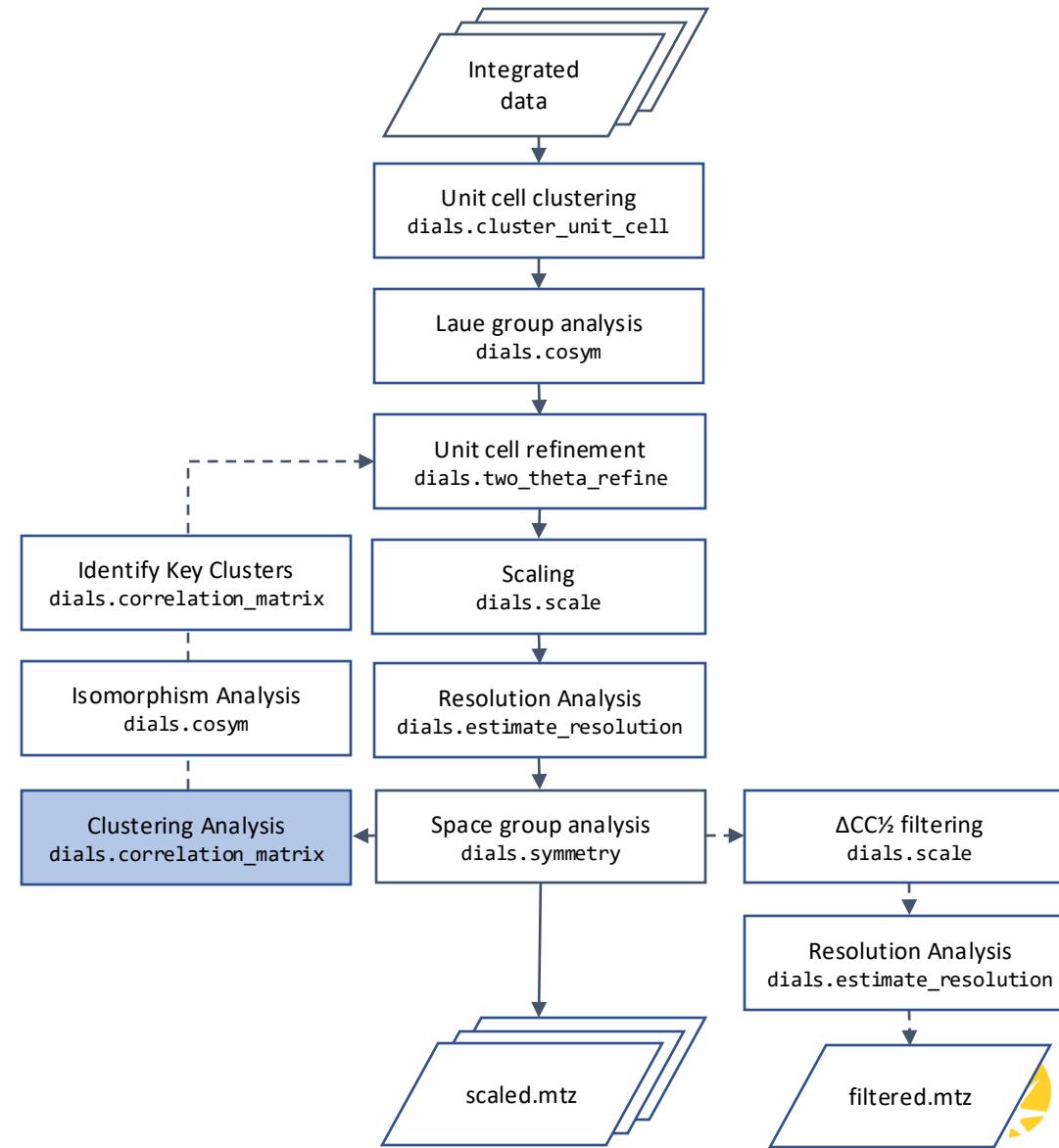
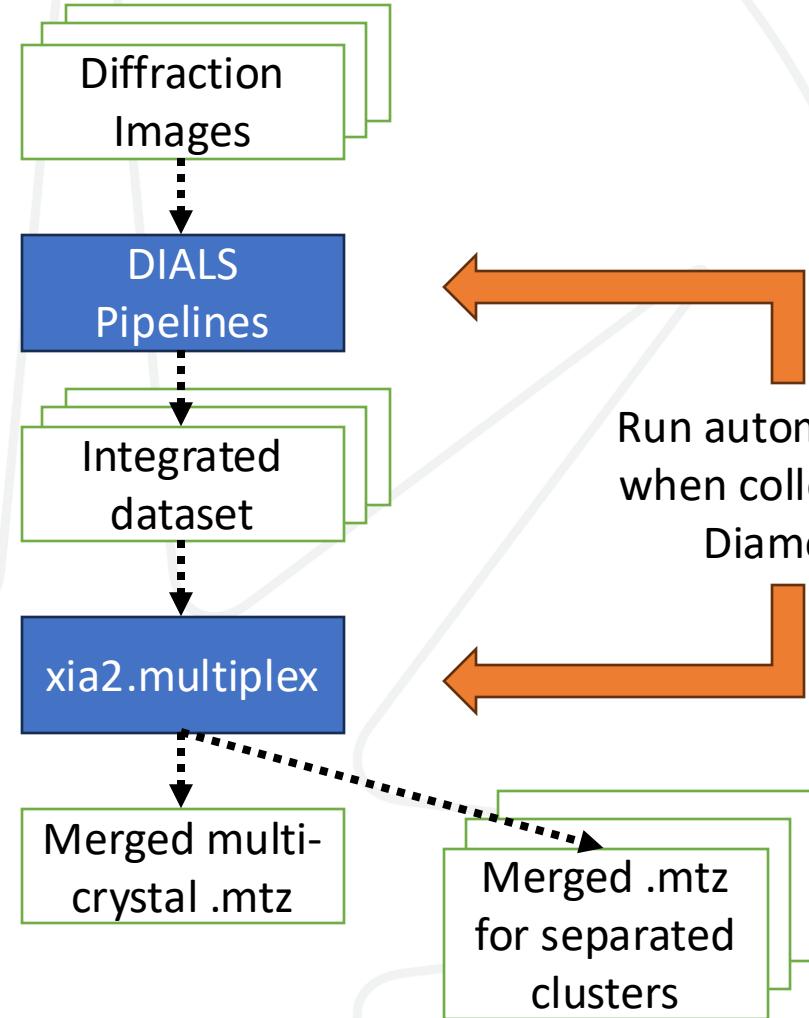
- Extension of correlation clustering
- Separates systematic and random error
- Same methods as indexing ambiguity resolution (dials.cosym)

$$\Phi = \sum_{i=1}^{N-1} \sum_{j=i+1}^N (r_{ij} - \mathbf{x}_i \cdot \mathbf{x}_j)^2$$



# Processing Multi-Crystal Data at Diamond

## Auto-processing Pipeline



# *Rotational Multi-Crystal Approaches*

Why rotation?

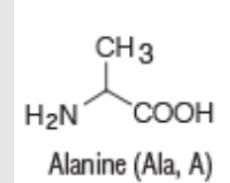
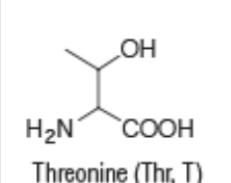
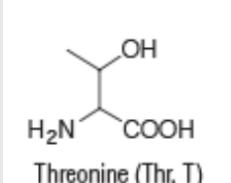
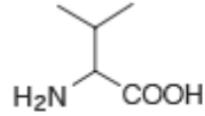
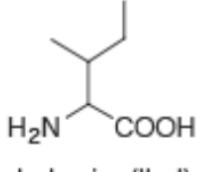
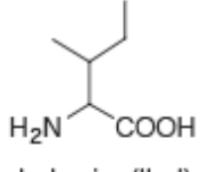
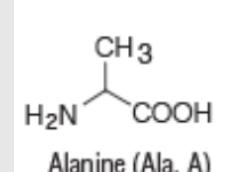
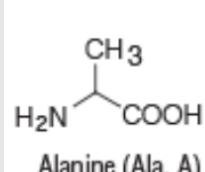
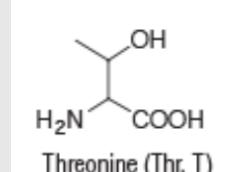
- More straight-forward data processing compared to still-shot data
- Radiation sensitive sample only allows small wedge of data prior to deteriorating
- Identify sub-populations within your sample easily
- Collect room temperature structures
  - No damaging cryo-protectants

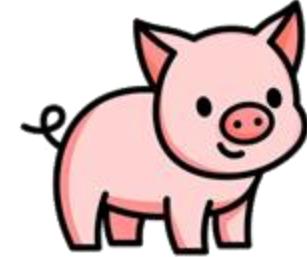
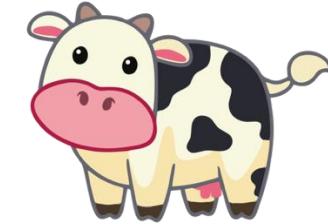
***What can we do with clustering?***



# *Introducing a Test Case... Cows, Pigs and People*

Human insulin, porcine insulin and bovine insulin differ subtly in their amino acid sequences and readily grow high-quality, cubic, isomorphous crystals

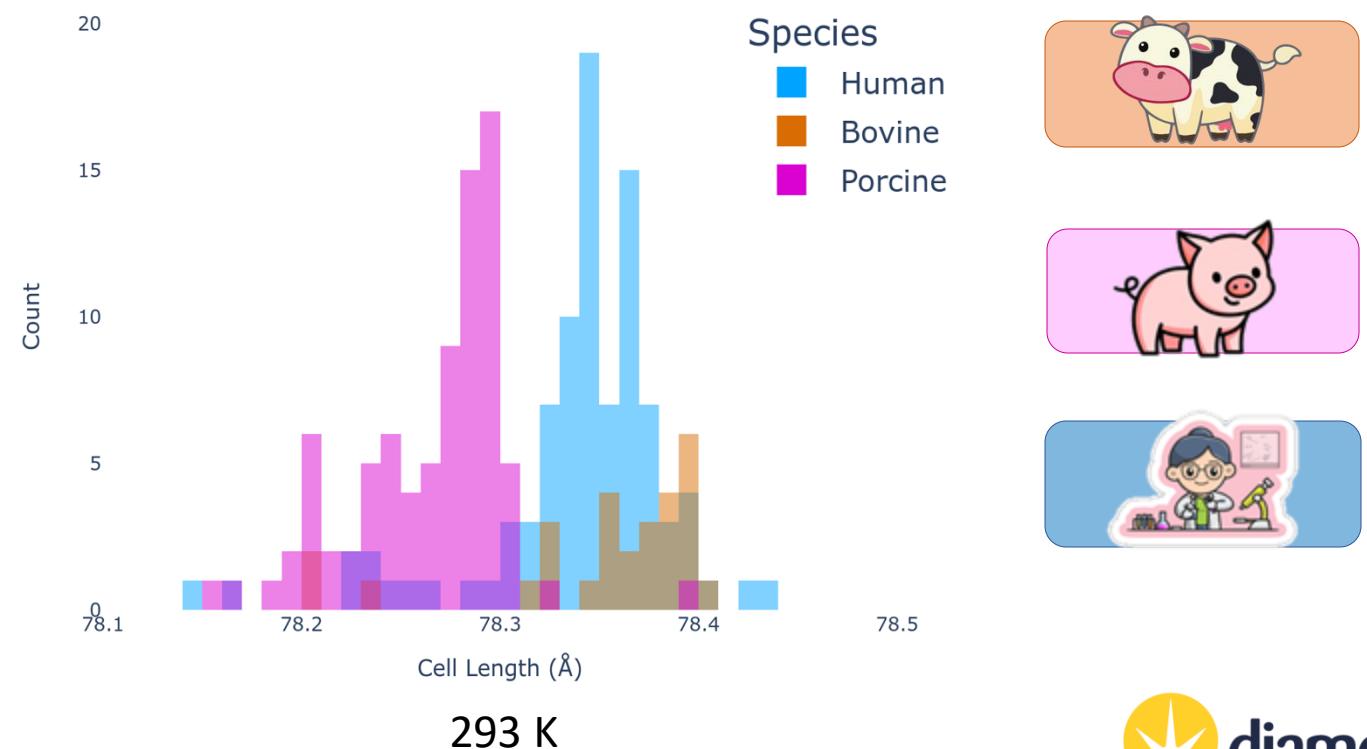
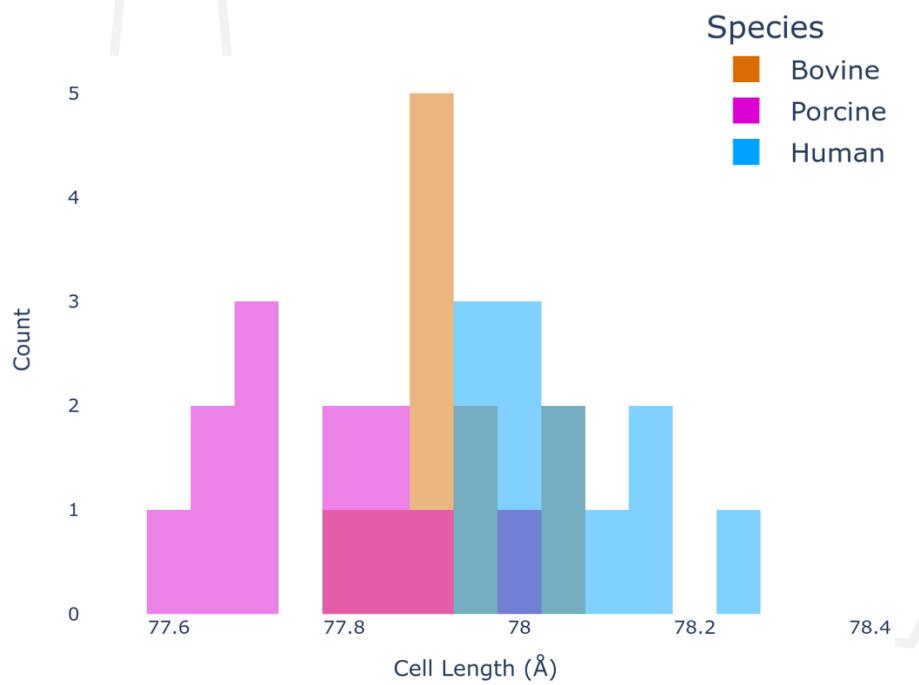
	Cows	Pigs	People
Chain A Residue 8	 Alanine (Ala, A)	 Threonine (Thr, T)	 Threonine (Thr, T)
Chain A Residue 10	 Valine (Val, V)	 Isoleucine (Ile, I)	 Isoleucine (Ile, I)
Chain B Residue 30 (terminus)	 Alanine (Ala, A)	 Alanine (Ala, A)	 Threonine (Thr, T)



# *Introducing a Test Case... Cows, Pigs and People*

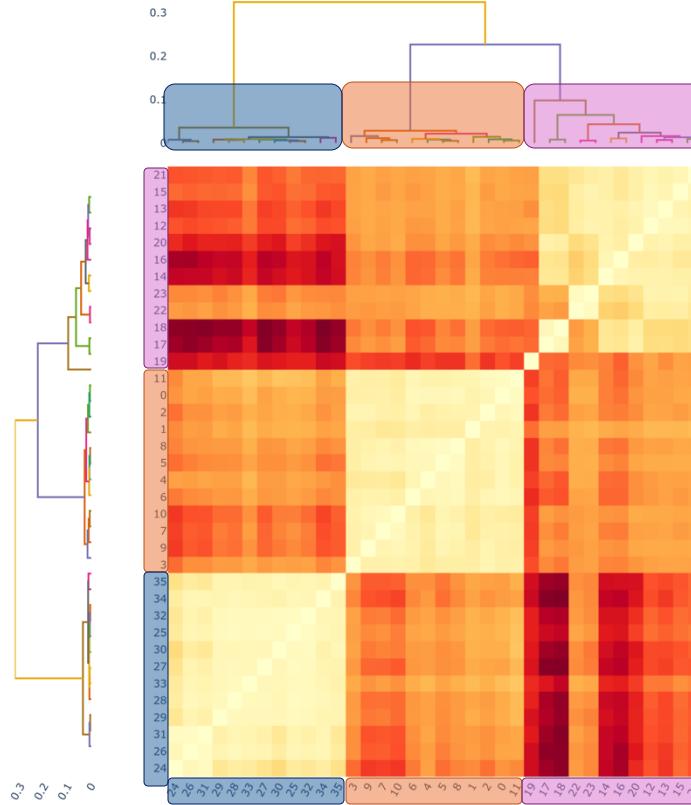
Isomorphous nature of these samples evident at both cryogenic and room temperature

- While subtle differences in unit cell are evident, natural spread across many crystals provides overlap, making them impossible to distinguish with unit cell-based clustering



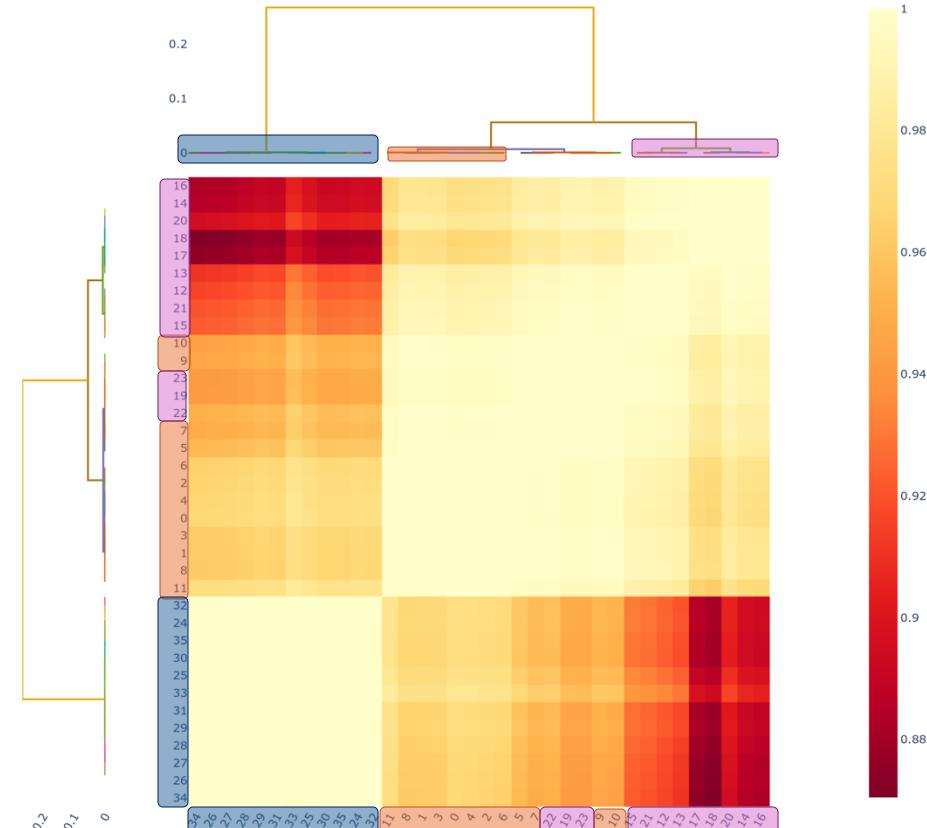
# Cows, Pigs and People – Cryo Temperature

- Correlation clustering can distinguish all three species
- Cosine angle clustering could not distinguish all three groups



Correlation Clustering

Thompson, A. J. et al. (2025) Acta Cryst. D81, 278-290.



Cosine Angle Clustering



# *Dimension Optimisation for Cosine Angle Clustering*

As xia2.multiplex uses dials.cosym methods for cosine angle clustering, it was optimized to determine a consensus symmetry, not cluster different groups. Therefore, the initial implementation was strictly performed in 2-dimensions.

However, the theory that these methods were based upon (Diederichs, K., 2017) mention that this can be tailored:

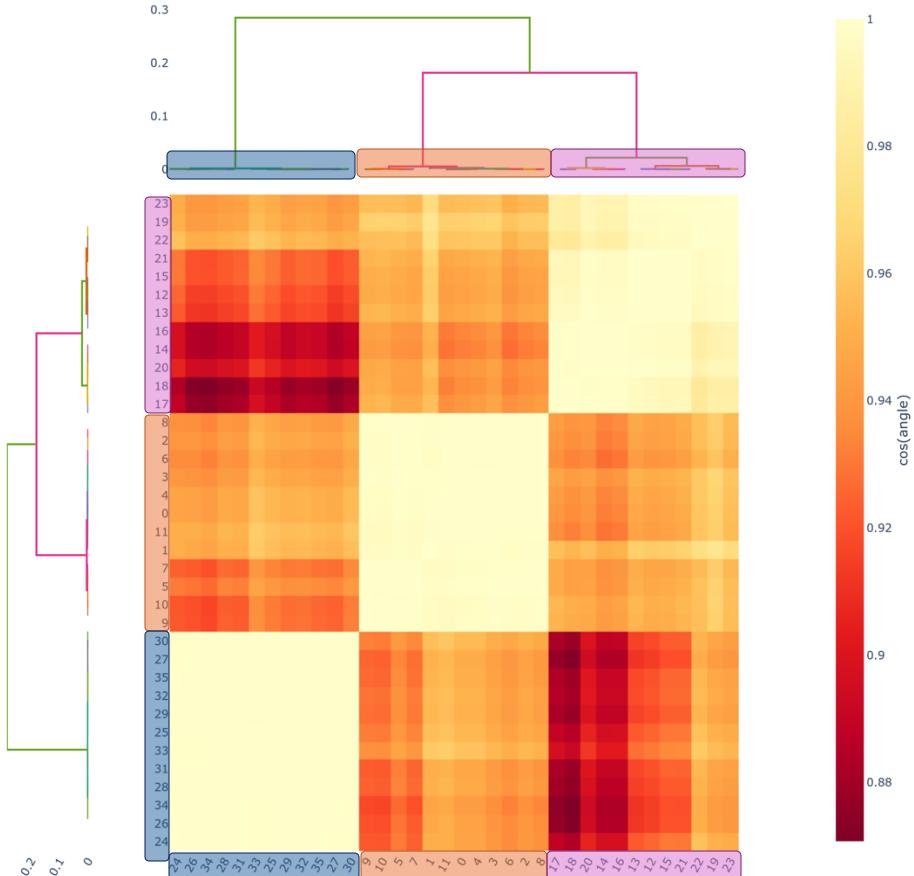
- **As more types of systematic differences are present, higher dimensions need to be used when minimizing**

$$\Phi = \sum_{i=1}^{N-1} \sum_{j=i+1}^N (r_{ij} - \mathbf{x}_i \cdot \mathbf{x}_j)^2$$

*For this dataset of cows, pigs and people, three were needed!!!*

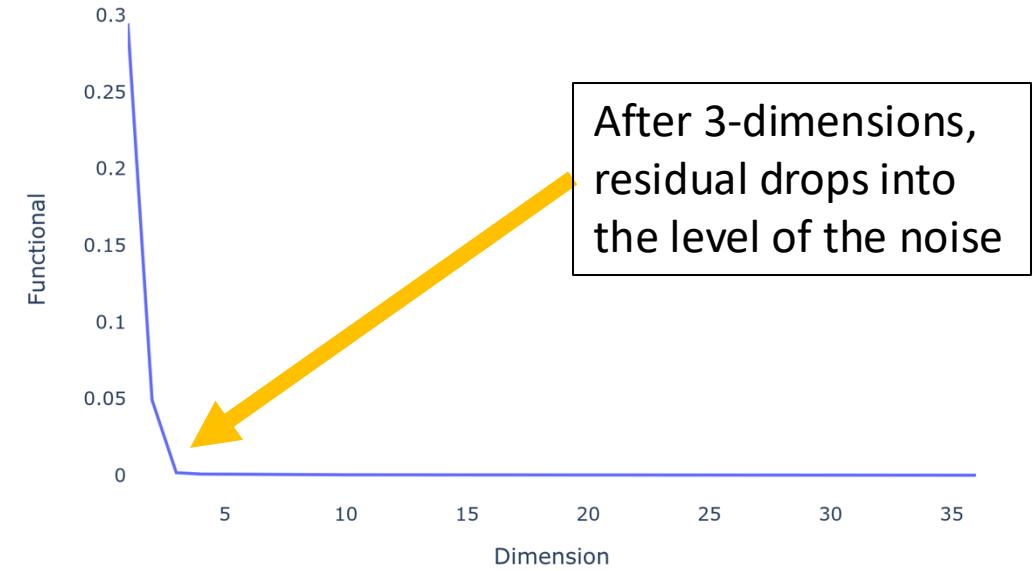
# Cows, Pigs and People – Cryo Temperature

- Performing the cosine angle clustering in three dimensions now resolves cows, pigs and people (including the outlier dataset identified in the correlation clustering)



**3D Cosine Angle Clustering**

Thompson, A. J. et al. (2025) Acta Cryst. D81, 278-290.



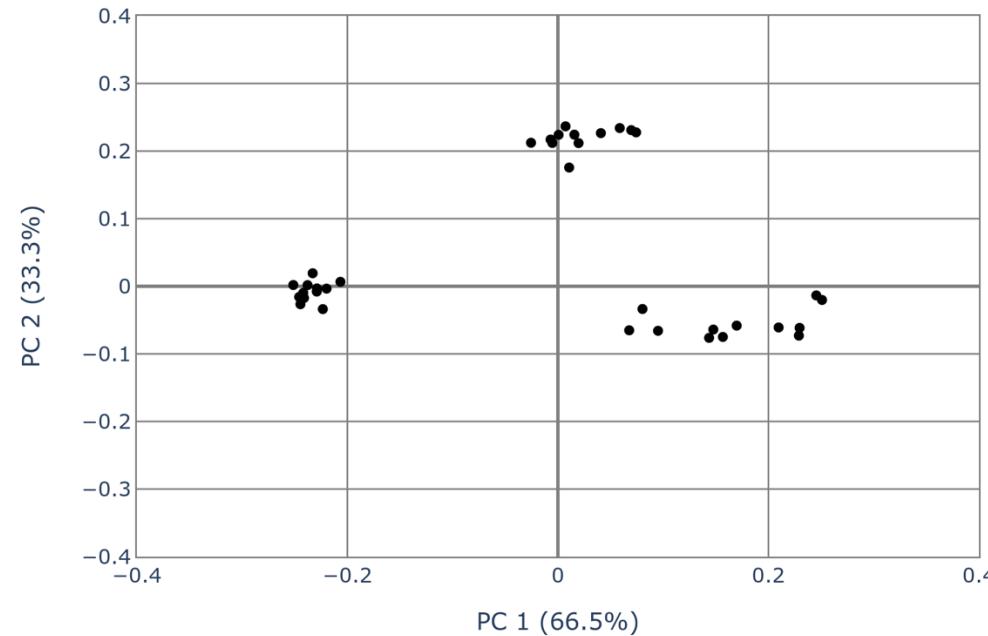
***But how do we get these clusters out automatically?***

# *Introducing the OPTICS algorithm*

- To perform the cosine-angle clustering, we represent each dataset as a point in some-dimensional space based on its correlation with other datasets

$$\Phi = \sum_{i=1}^{N-1} \sum_{j=i+1}^N (r_{ij} - \mathbf{x}_i \cdot \mathbf{x}_j)^2$$

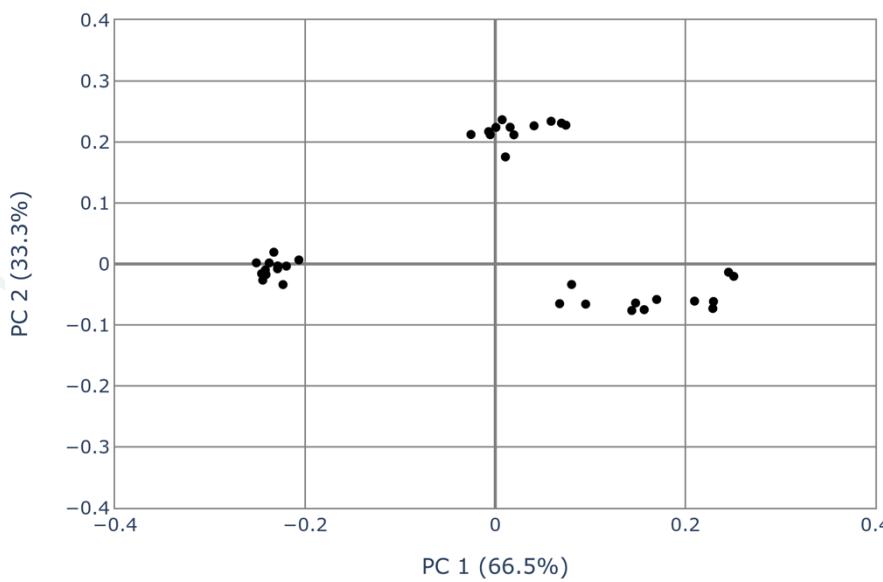
- This produces coordinate plots for visualisation
  - Points further away from the origin have less random error
  - Pairs of points with smaller angles have less systematic error between them
  - Rotated to align with principal components



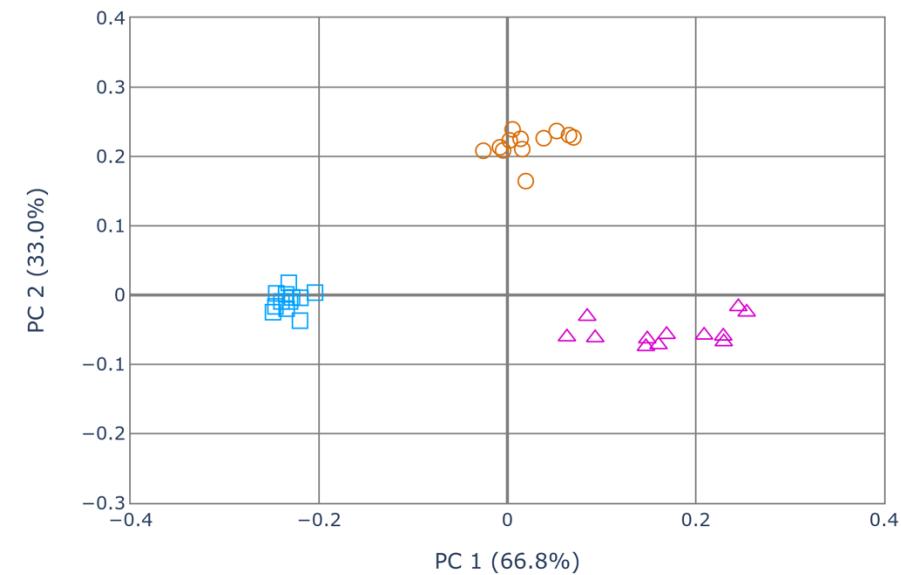
# *Introducing the OPTICS algorithm*

OPTICS is a density-based algorithm that automatically defines clusters with spatial relations (which we have calculated using cosine-angle clustering!)

- Requires no global density parameter (can detect clusters of varying density)
- Does not assume any specific number of clusters
- Can detect noise (does not require every dataset to be a part of a cluster)

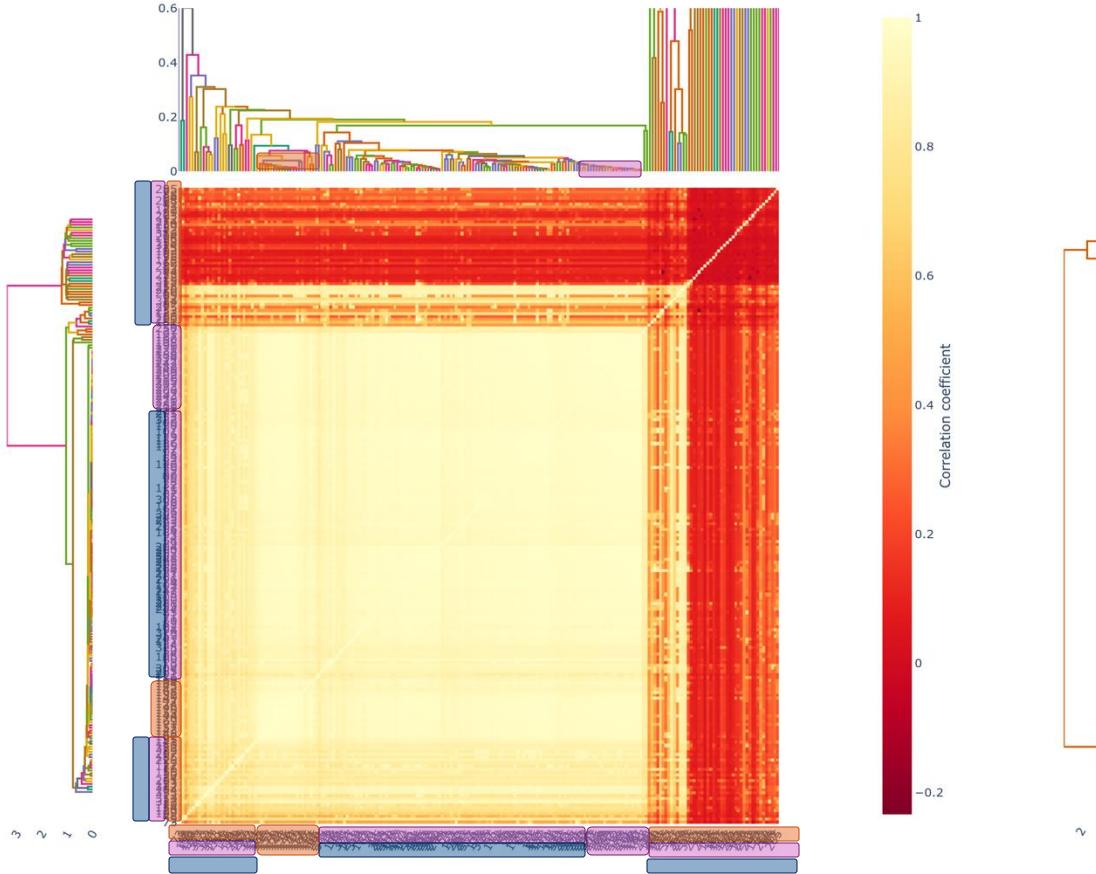


Apply  
OPTICS



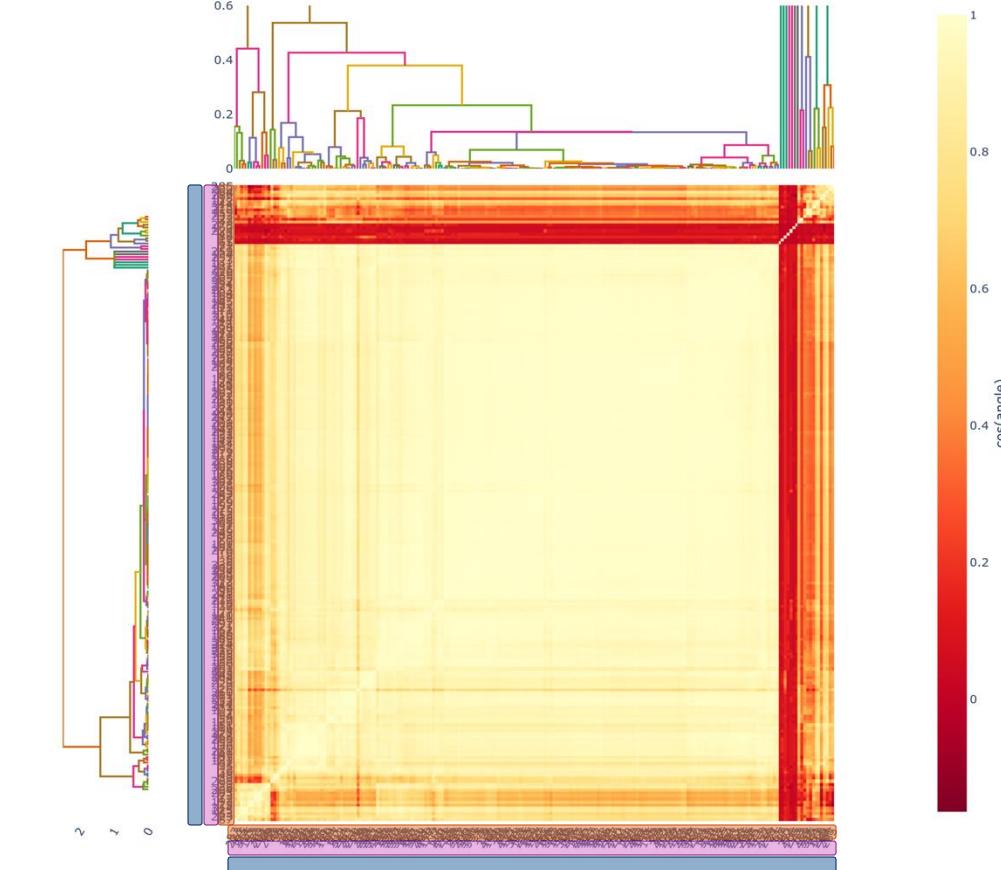
# *Cows, Pigs and People – Room Temperature*

- Initially, a large room temperature dataset did not result in any clusters...

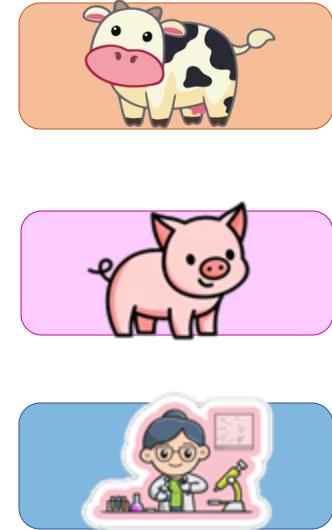


**Correlation Clustering**

Thompson, A. J. et al. (2025) Acta Cryst. D81, 278-290.



**Optimised-Dimension Cosine Angle Clustering**



# *Weighted Correlation Coefficient*

Issues with standard method of calculation correlation coefficient:

- Does not consider reliability of merged intensities
- May have vastly different numbers of reflections merged for each common index compared (particularly with small wedge data)
- Can be dominated by high-intensity data

$$CC = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}} = \frac{s_{xy}}{\sqrt{s_{xx} \times s_{yy}}}$$

**Add appropriate weights from intensity sigmas to help balance this:**

$$s_{xx} = \frac{\sum_i w_i (x_i - \bar{x})^2}{\sum_i w_i}, s_{yy} = \frac{\sum_i w_i (y_i - \bar{y})^2}{\sum_i w_i}, s_{xy} = \frac{\sum_i w_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i w_i}$$

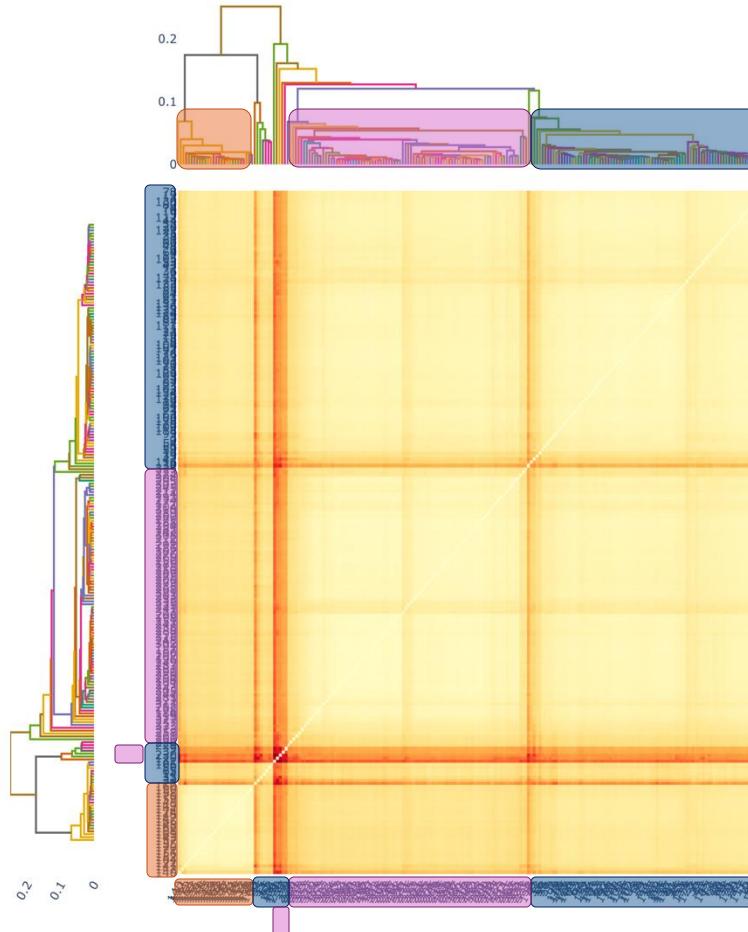
$$\bar{x} = \frac{\sum_i w_i x_i}{\sum_i w_i}, \bar{y} = \frac{\sum_i w_i y_i}{\sum_i w_i} \quad w_i = \frac{1}{\sigma_i^2}, \sigma_i^2 = \sigma_{x_i}^2 + \sigma_{y_i}^2$$

*Work by James Beilsten-Edmands*

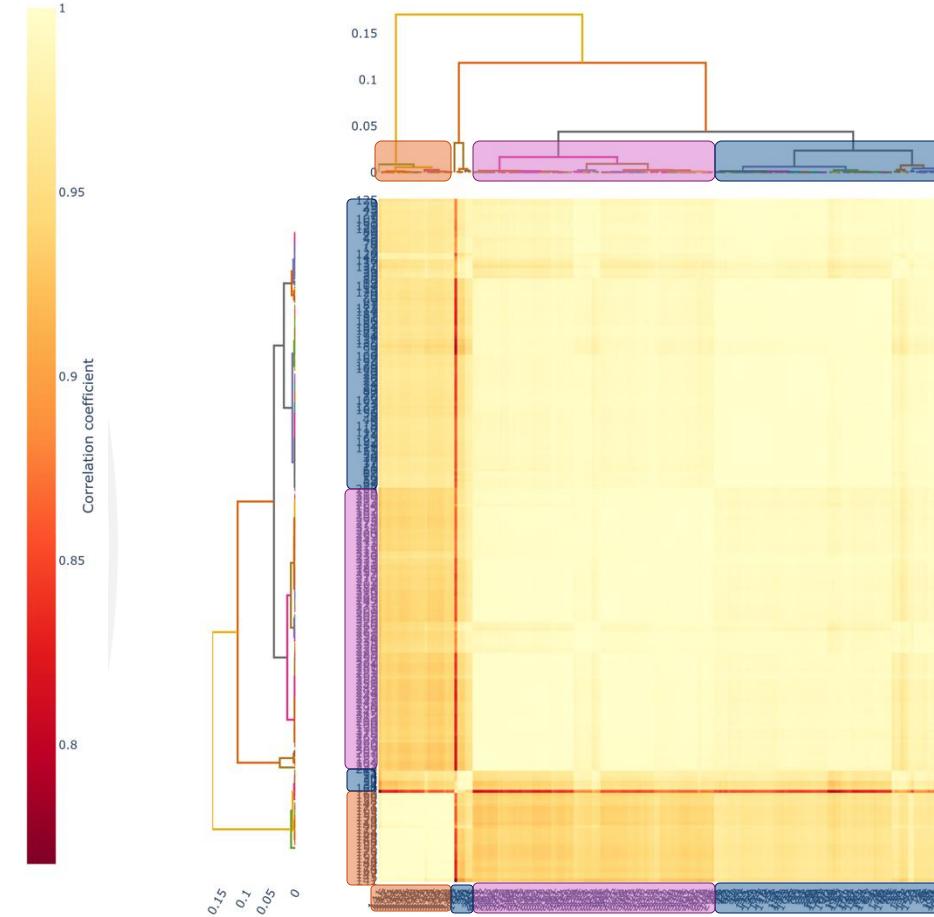


# *Cows, Pigs and People – Room Temperature*

- Weighted CC calculations mean the clusters can be separated into species



Correlation Clustering

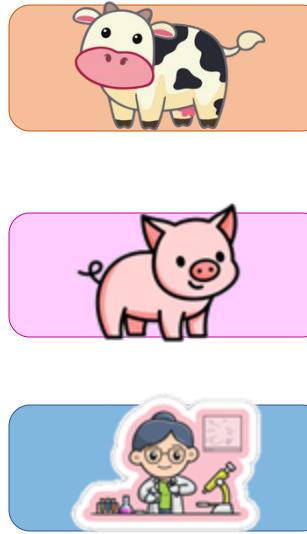
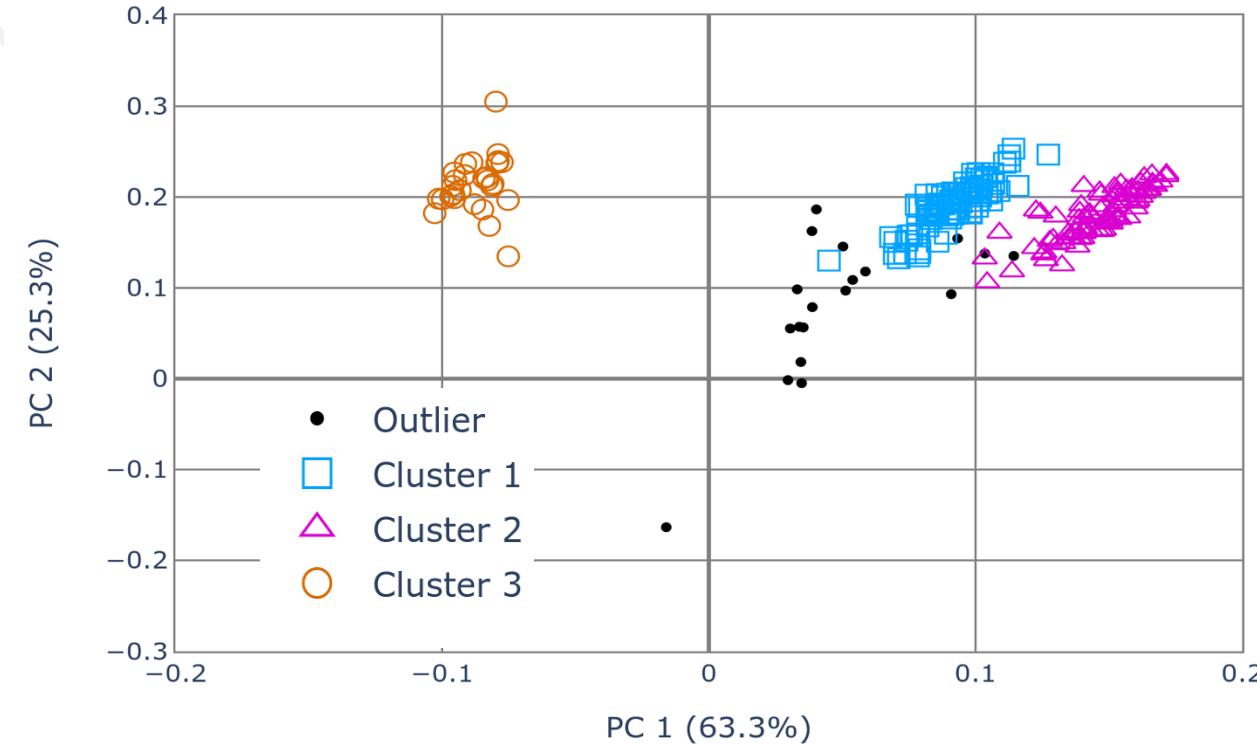
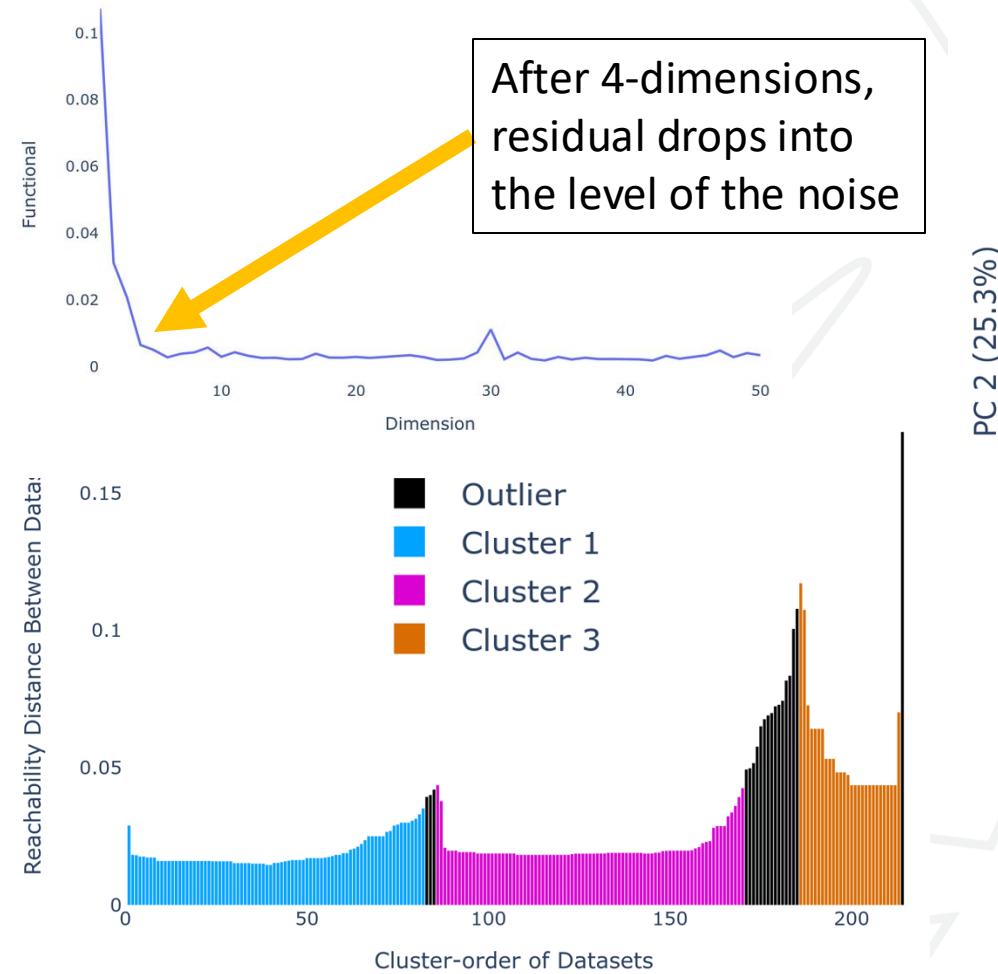


Optimised-Dimension Cosine Angle Clustering



# Cows, Pigs and People – Room Temperature

- Clustering optimised to 4-dimensions, has well separated pigs and people clusters, and an outlier group of datasets are identified



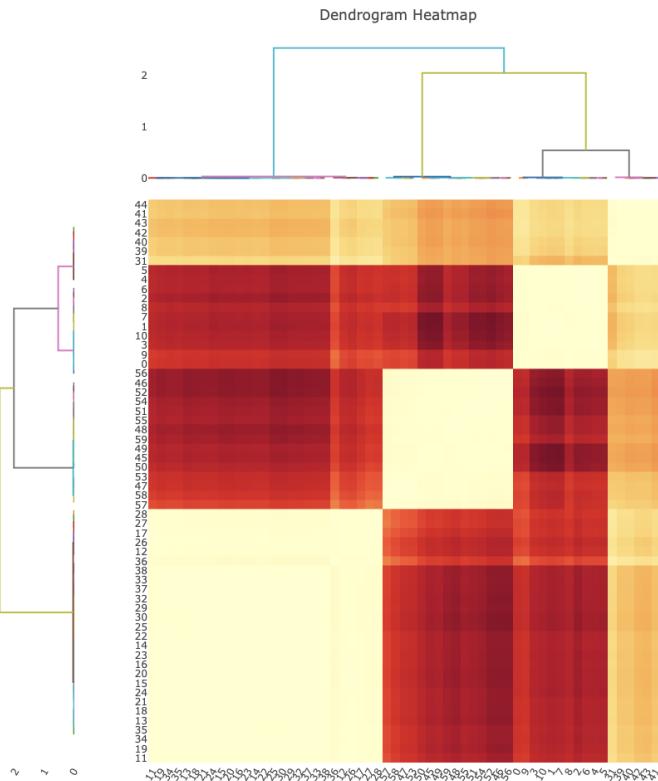
***OPTICS uses a "Reachability Plot" to classify data***  
controlled by  $\xi$ -parameter (steepness between  
reachability values for data to be considered a new group),  
and by setting the minimum number of datasets

**diamond**

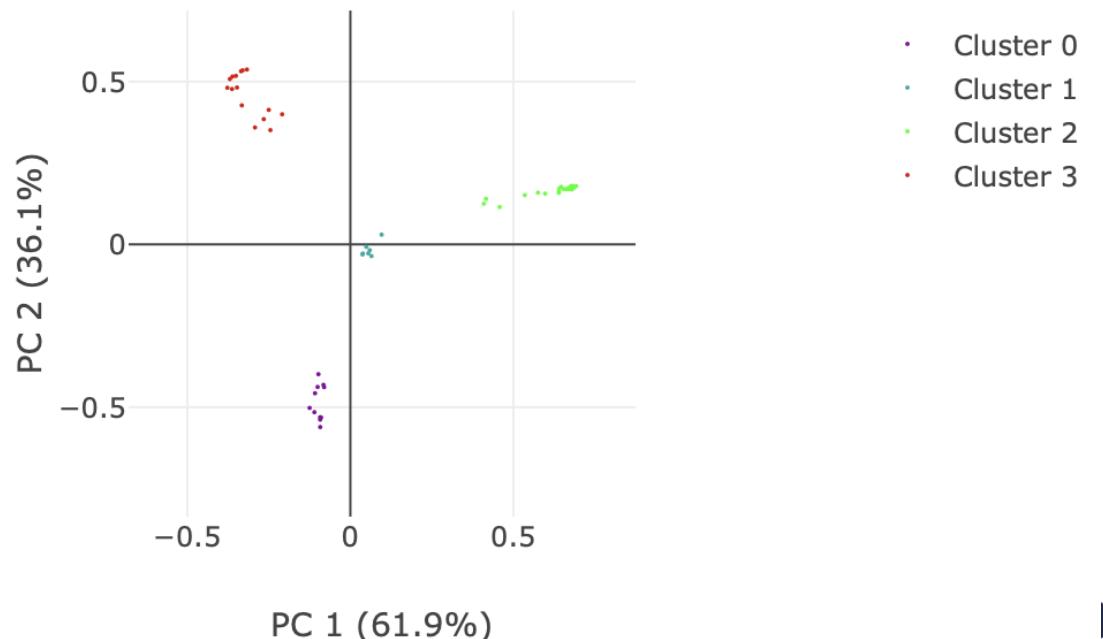
# *Example: Mislabelled Sample Vials!*

- Data measured on VMXm of 4 different isomorphous CPVs with subtly different sequences
- Multi-crystal data collection necessary due to beamline configuration – how do we know what to merge?
- BUT some ambiguous sample vial labelling.....

*Data from Jose Trincao (VMXm)*



Cosym Coordinates Rotated by Principal Components



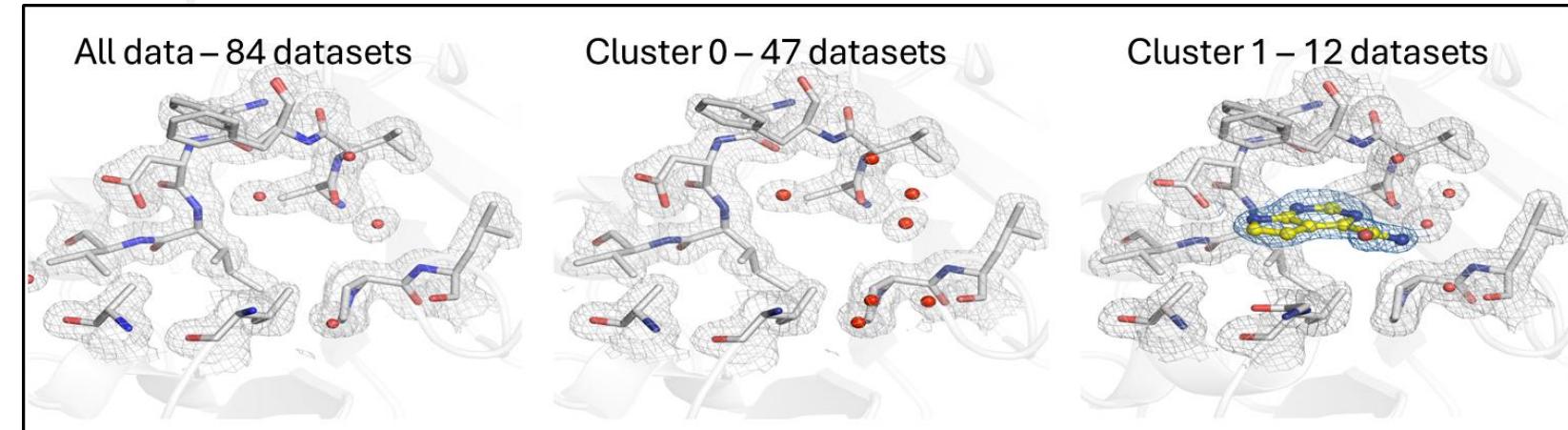
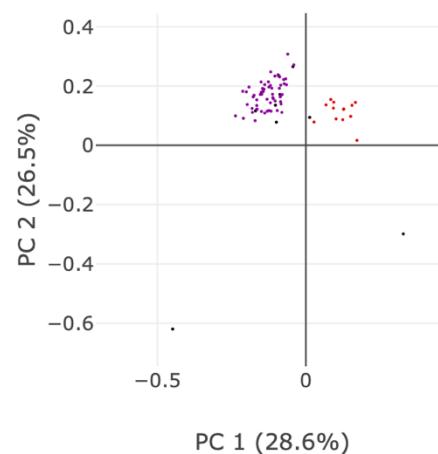
# *Example: Identification of a New Binding Event*

## **Identification of new fragment screening hit**

*Data from Megan Lambert (VMXi)*

- Identifying fragment hits is important for developing new drugs.
- VMXi are developing room temperature fragment screening – this needs a multi-crystal approach.
- Data first separated by which ligand was dispensed (sample-groups)
- Each sample-group run through xia2.multiplex with new clustering algorithms
- Able to identify new hits using improved clustering methods to separate apo and bound data.

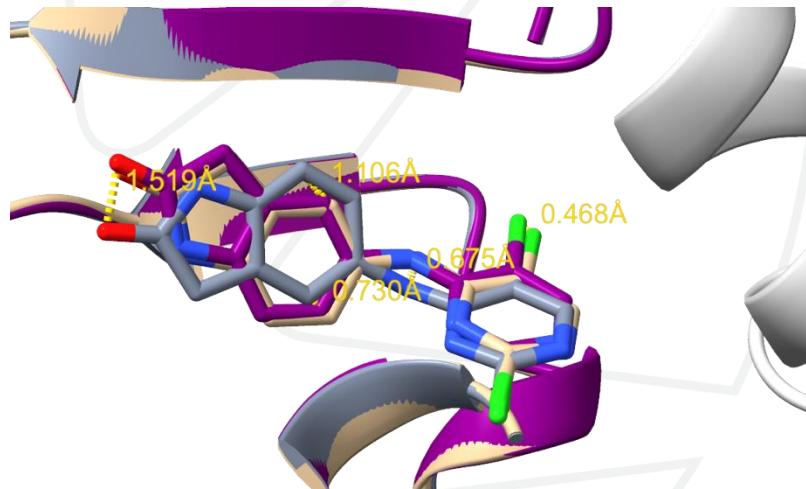
Cosym Coordinates Rotated by Principal Components



# *Example: Alternative Ligand Binding Conformations*

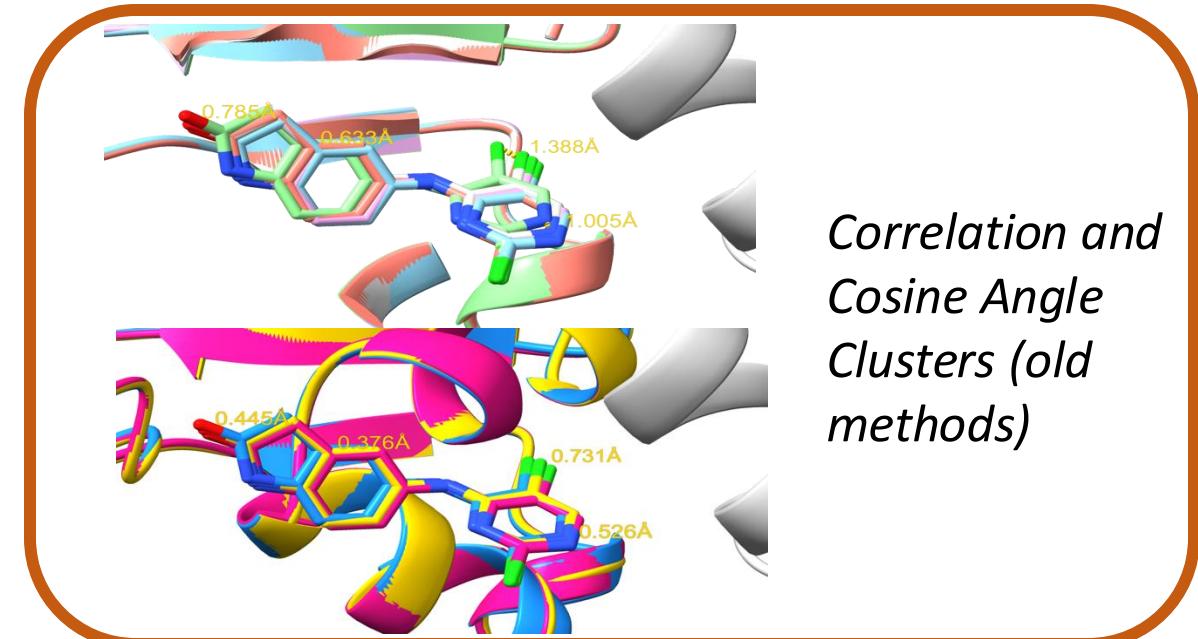
## **Identification of multiple binding states for drug discovery**

- XChem typically uses a one-crystal-one-fragment model, which means any variation in binding is not easily resolved.
- Any variation in binding state, however, gives key clues to the chemistry and thus critical to decision making for drug optimisation.
- Automatic clustering using new methods found distinct binding states not visible using original multiplex clustering methods.



*Clustering using new methods*

*Data from Jannik Wiebe  
(Institute for Cancer Research)*



*Correlation and  
Cosine Angle  
Clusters (old  
methods)*

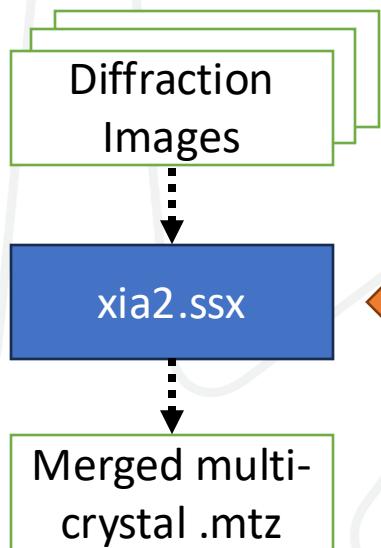
# *Still-Shot Multi-Crystal Approaches*

Why still shots?

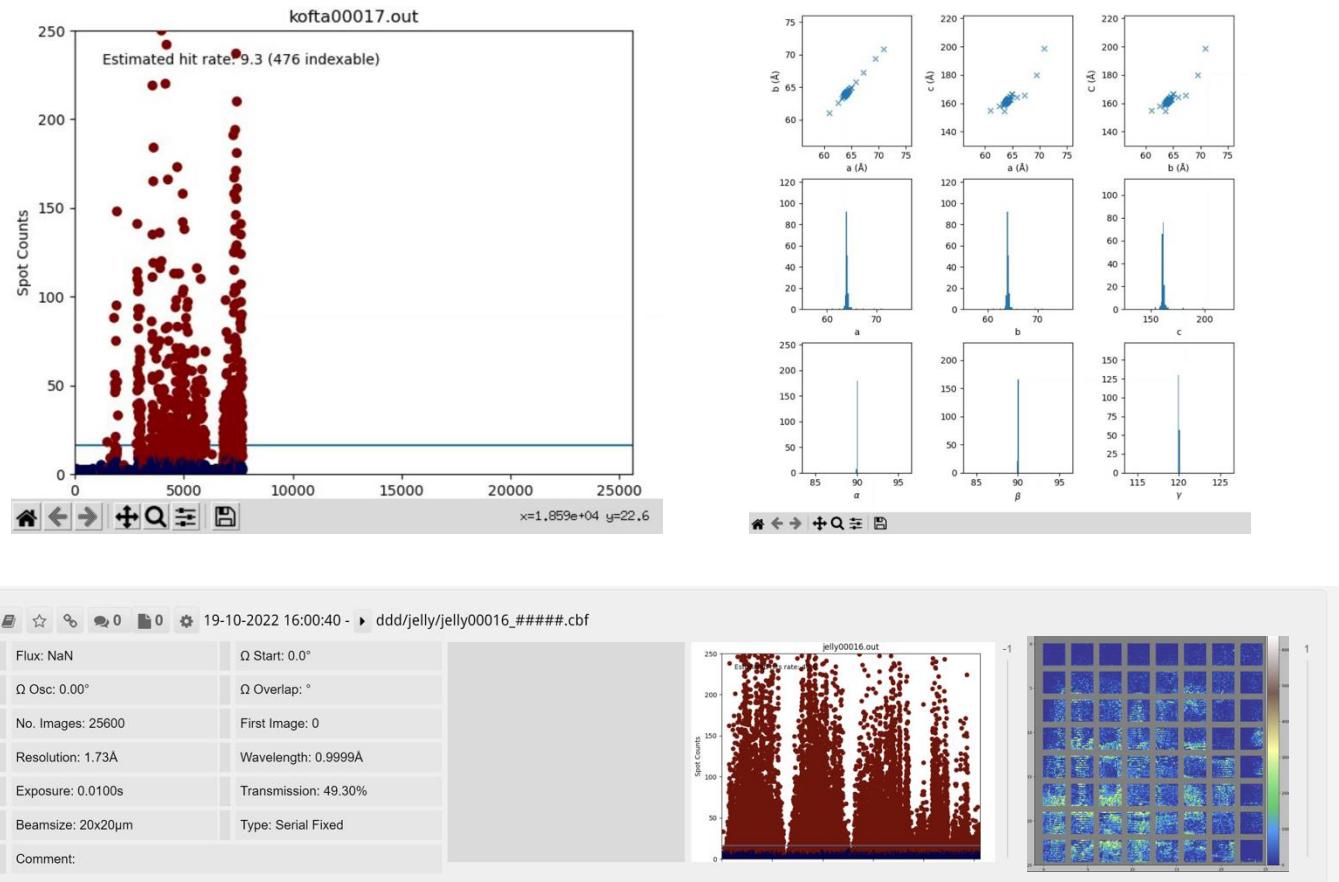
- Crystals extremely radiation sensitive
- Pump-probe experiments better with very small samples to ensure entire crystal affected by change
  - Ligand binding
  - Light activated reactions
- Time-resolved experiments
- Can use extremely high flux for weakly diffracting samples (diffraction before destruction)

# Processing Multi-Crystal Data at Diamond

Auto-processing Pipeline for still-shot  
Multi-Crystal Experiments



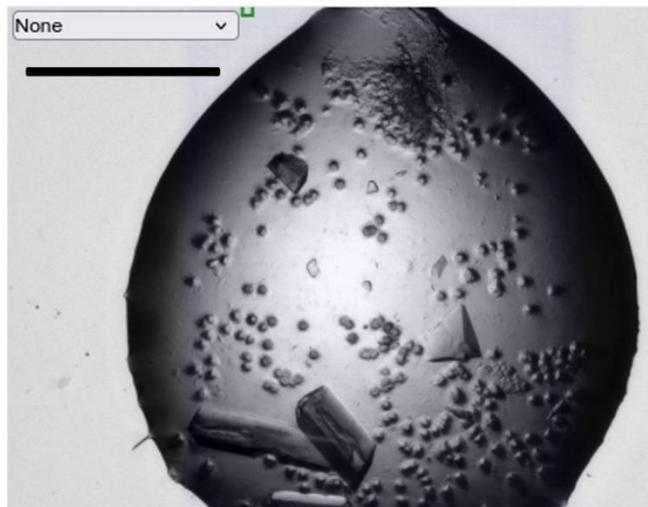
Runs  
automatically on  
I24 and grid scans  
on VMXi –  
otherwise need to  
run manually via  
command line



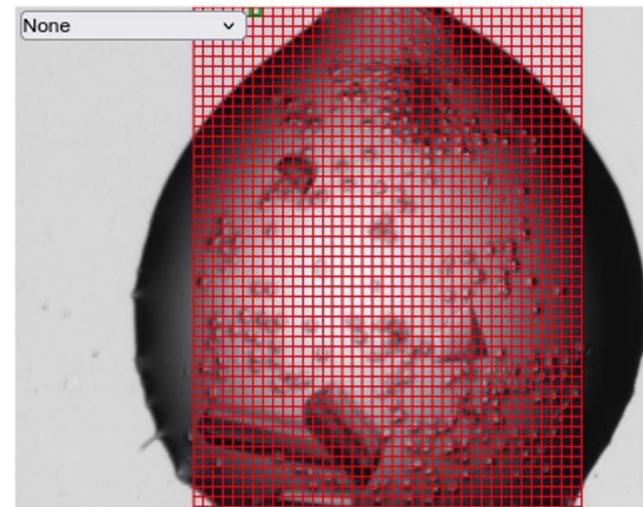
# *Using VMXi to screen for serial experiments*

- Most common use-case of raster scanning (aka grid-scanning) -> identify crystalline material
- This technique is fundamentally a still-shot serial crystallography experiment – we can process these as serial data!

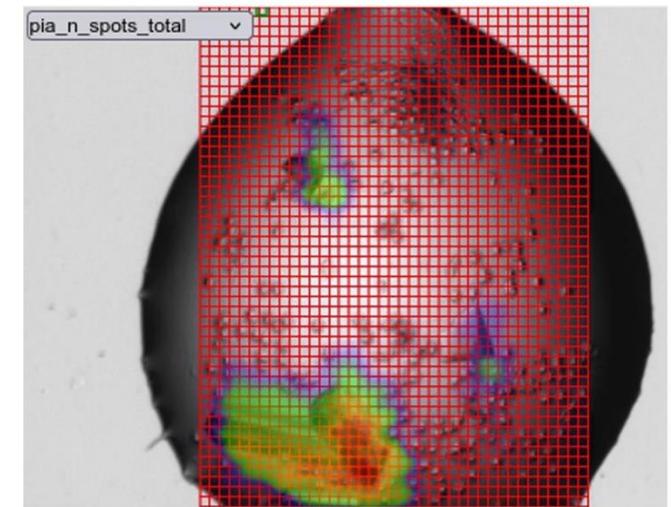
**(a)**



**(b)**

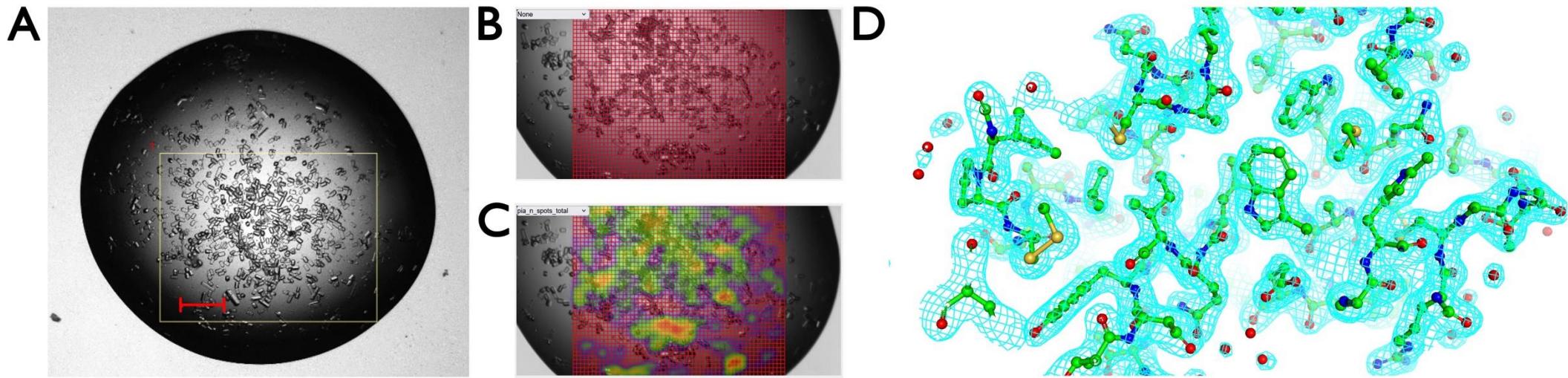


**(c)**



# Using VMXi to screen for serial experiments

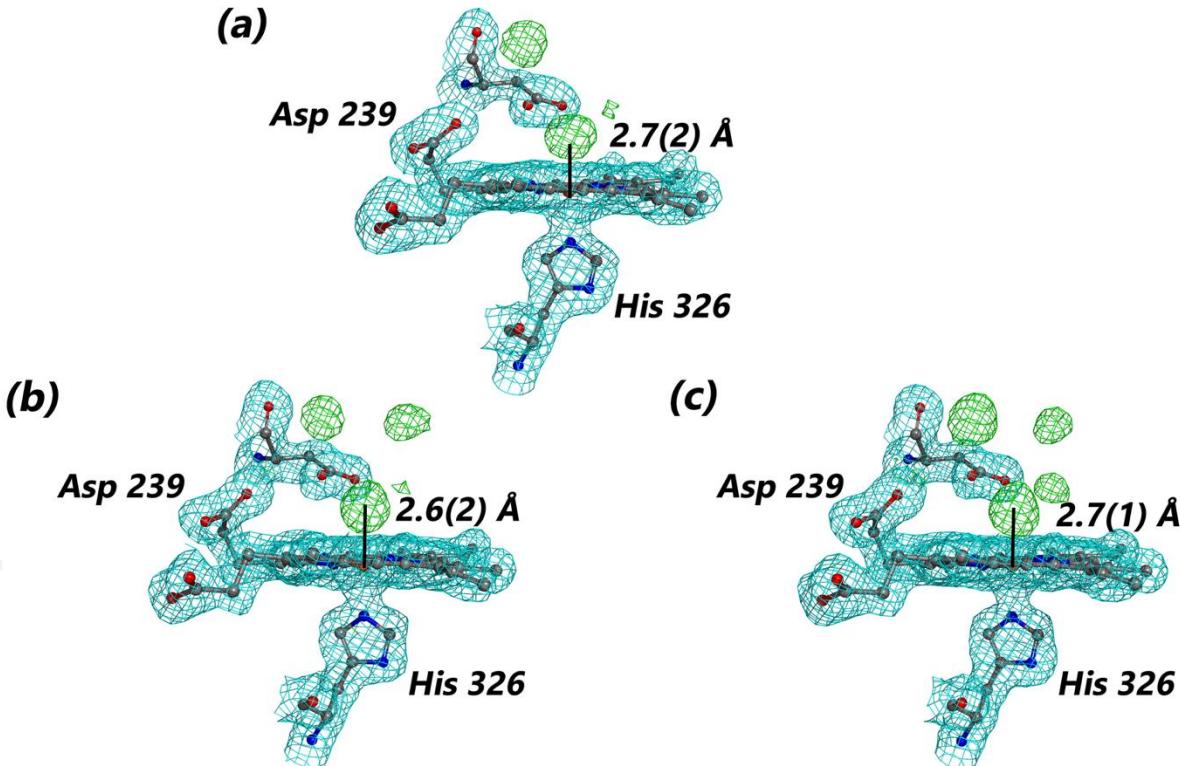
You can get decent quality serial data from grid scans on VMXi – example: lysozyme



Statistics for eight drops merged together	Completeness (%)	Multiplicity	$I/\sigma(I)$	$R_{\text{split}}$	$CC_{1/2}$	Unique Observations	Crystals merged
Overall (55.56 – 1.88)	99.27	110.9	19.1	0.083	0.997	10076	9891
High Resolution (1.95 – 1.88)	99.90	65.1	1.2	1.143	0.398	973	

# Using VMXi to screen for serial experiments

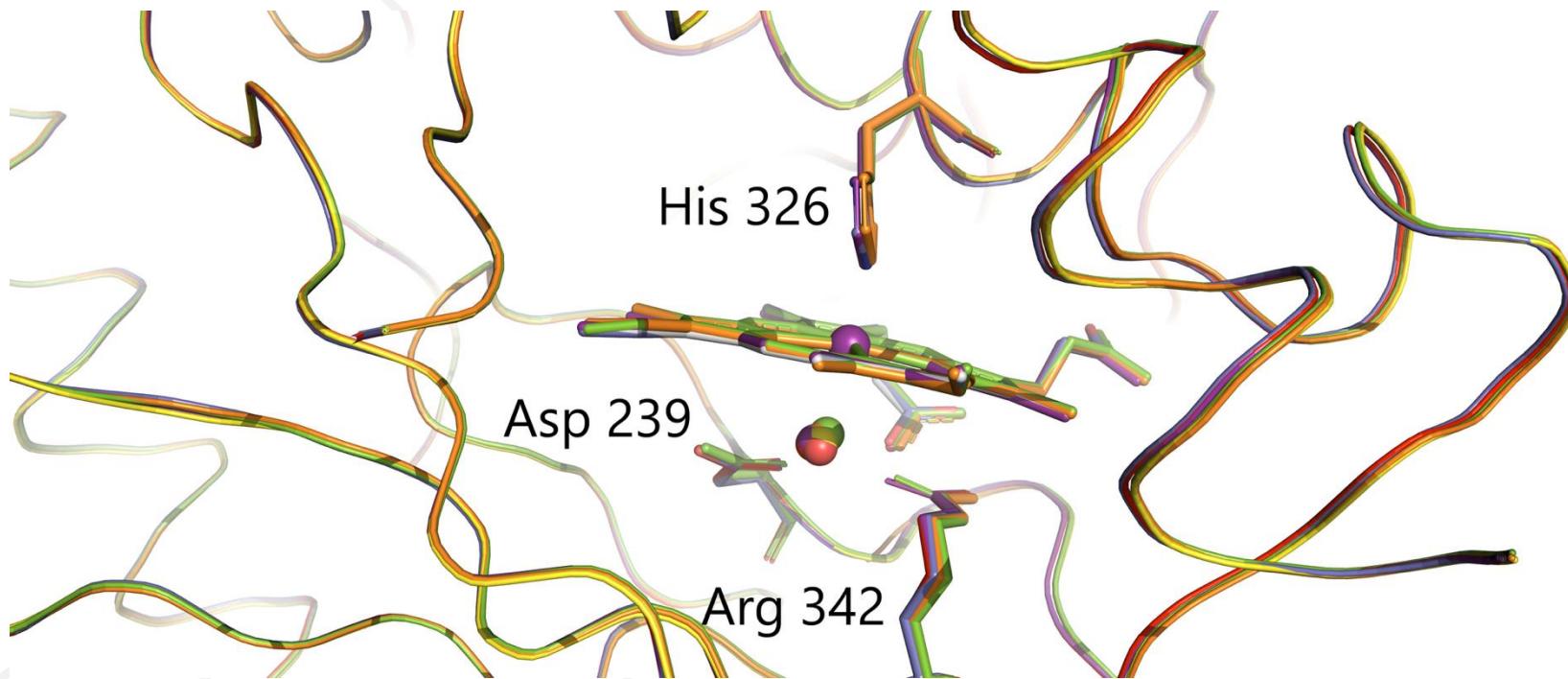
DtpAa: heme-containing dye-decolourising peroxidase  
(Space Group P<sub>2</sub><sub>1</sub>)



	(a) DtpAa (8 drops)	(b) DtpAa (12 drops)	(c) DtpAa (35 drops)
Resolution Range	69.76 – 2.07 (2.14 – 2.07)	69.73 – 1.88 (1.95 – 1.88)	69.77 – 1.79 (1.85 – 1.79)
Crystals Merged	5360	10054	22854
Volume Dispensed	0.8 µL	1.2 µL	3.5 µL
Unique Reflections	42538 (4203)	56729 (5631)	65821 (6543)
Multiplicity	28.7 (22.9)	48.0 (25.7)	98.9 (45.8)
Completeness	97.74 (84.49)	99.73 (99.36)	99.75 (99.69)
I/σ(I)	12.6 (2.0)	19.7 (2.2)	21.8 (2.0)
CC <sub>1/2</sub>	0.927 (0.329)	0.958 (0.302)	0.990 (0.315)
R <sub>work</sub>	0.2233 (0.3855)	0.2115 (0.2927)	0.1885 (0.3265)
R <sub>free</sub>	0.2748 (0.4277)	0.2420 (0.3491)	0.2281 (0.3893)

- Lower symmetry necessitates more drops for high quality maps
- High quality data with small volumes possible

# *Using VMXi to screen for serial experiments*

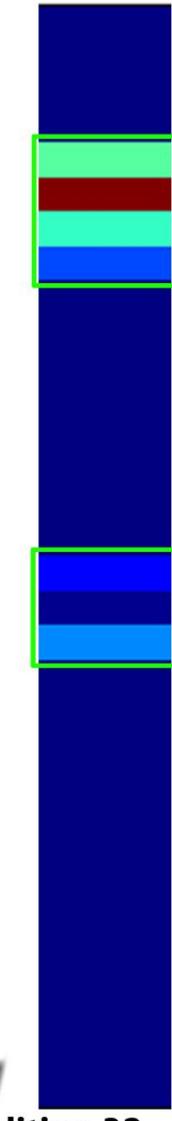


- Comparison to previously published XFEL and synchrotron serial structures revealed minimal differences
- Simple, in-plate serial collections are very viable for early stages of a serial project for initial structure determination

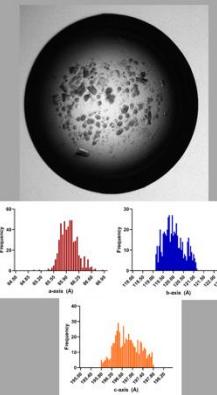
# Using VMXi to screen for serial experiments

Condition 1

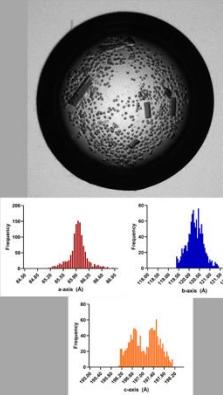
## Representative Images and Unit Cell Distributions of Auto-Processed Conditions



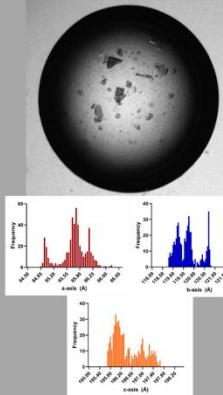
Condition 5  
540 crystals



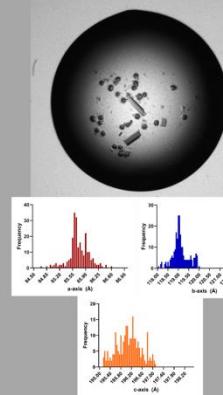
Condition 6  
1185 crystals



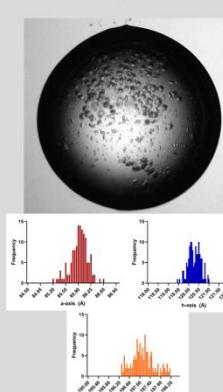
Condition 7  
490 crystals



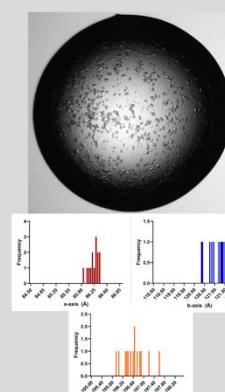
Condition 8  
235 crystals



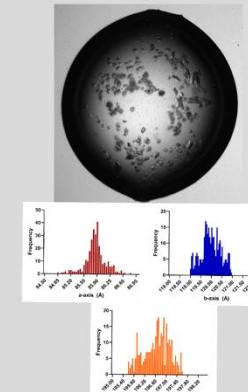
Condition 17  
134 crystals



Condition 18  
14 crystals

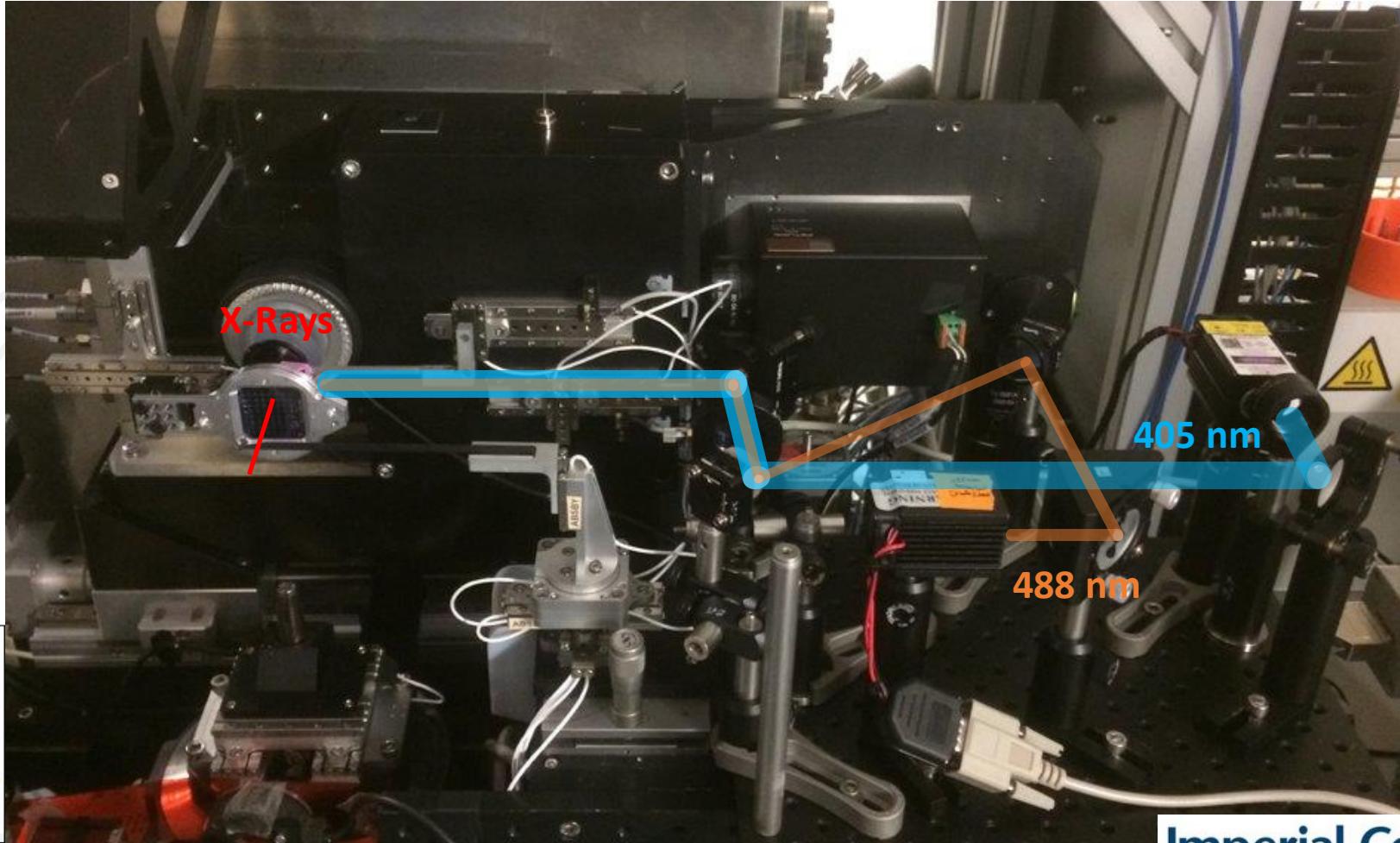
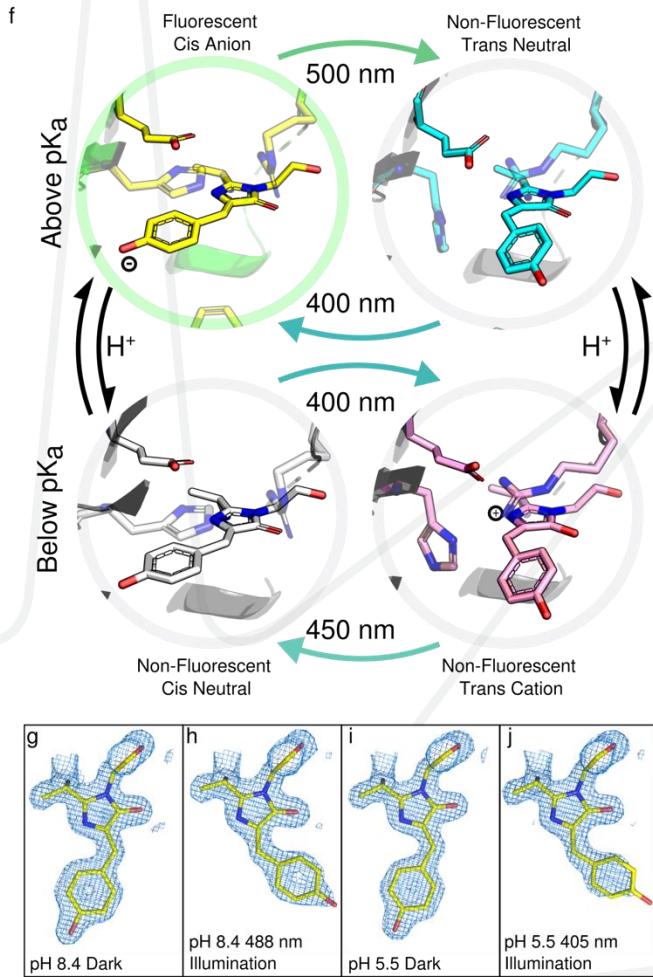


Condition 19  
306 crystals



- Very quickly determine diffraction quality of micro-crystallisation slurries with minimal loss of sample
- Get insight into homogeneity, isomorphism, crystal size, etc

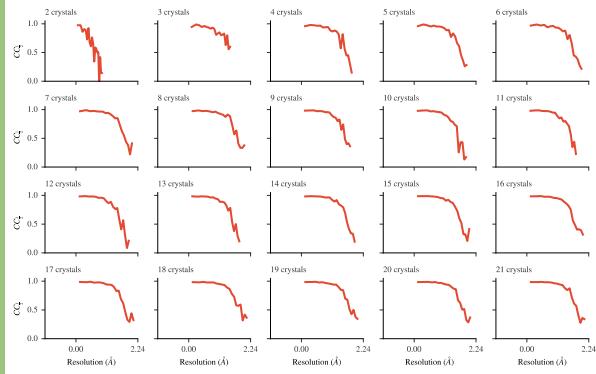
# Example: I24 SSX (Van Thor Group – Imperial College London)



# *Key Takeaways of Multi-Crystal Experiments*

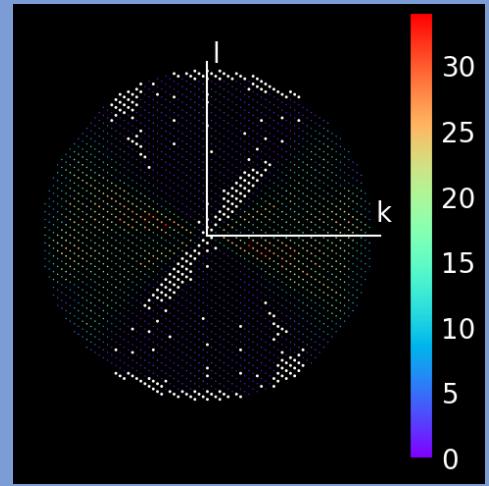
## **Advantages:**

- Improved signal-to-noise
- Push resolution of usable data
- Good data from radiation sensitive samples



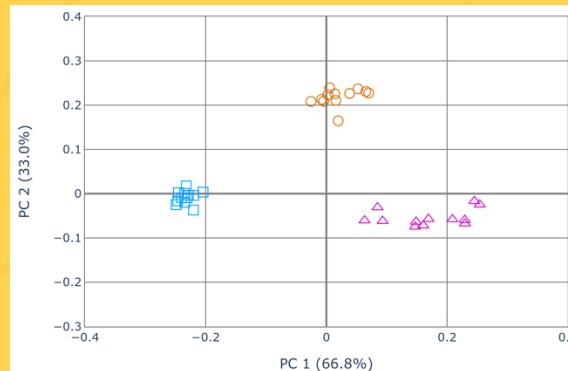
## **Challenges:**

- Obtain consistent symmetry
- Handle non-isomorphism
- Identify any preferential orientation



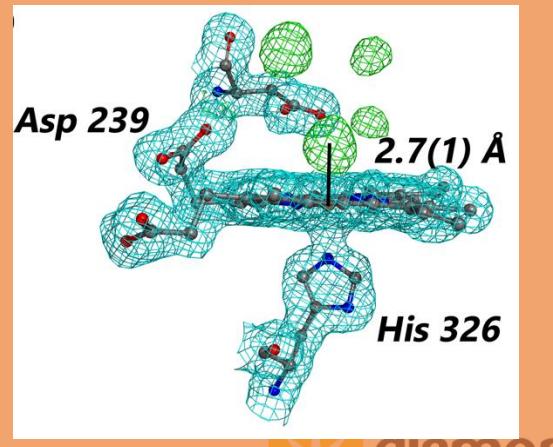
## **Clustering:**

- Unit cell good for large changes
- Intensity-based required for small changes
- Automatically identified in DIALS



## **Still-Shot Serial:**

- Diffraction before destruction
- Simple screening tools available
- Time-resolved experiments



# Acknowledgements

## VMXi:

- Michael Hough
- Juan Sanchez-Weatherby
- James Sandy
- Halina Mikolajek
- Megan Lambert
- Cicely Tam

## MX-Analysis Team:

- Graeme Winter (now at NE-CAT)
- James Beilsten-Edmands
- Phil Blowey
- Irakli Sikharulidze
- Nick Devenish
- Rowan Walker-Gibbons



## I24:

- Sofia Jaho
- Danny Axford
- Robin Owen

## ICR:

- Jannik Weibe
- Yann-Vai Le Bihan
- Rob Van Montfort

## VMXm:

- Jose Trincao

