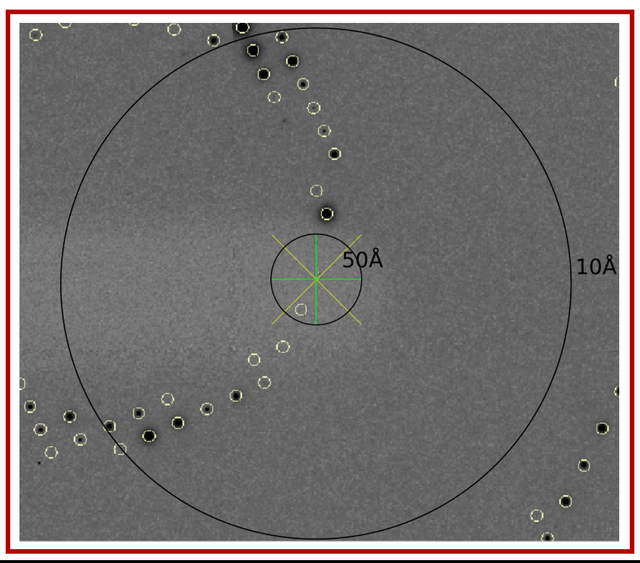
Yunyun Gao[a], Helen Ginn[a,b], Andrea Thorn[a]
a. Institut für Nanostruktur und Festkörperphysik, Universität Hamburg, Germany
b. Center for Free-Electron Laser Science, CFEL, Deutsches Elektronen-Synchrotron DESY

Finding NEMOs: A Story about Lost in Autoprocessing (and the Redemption)

Try me: How will the pipeline/you mask the beamstop

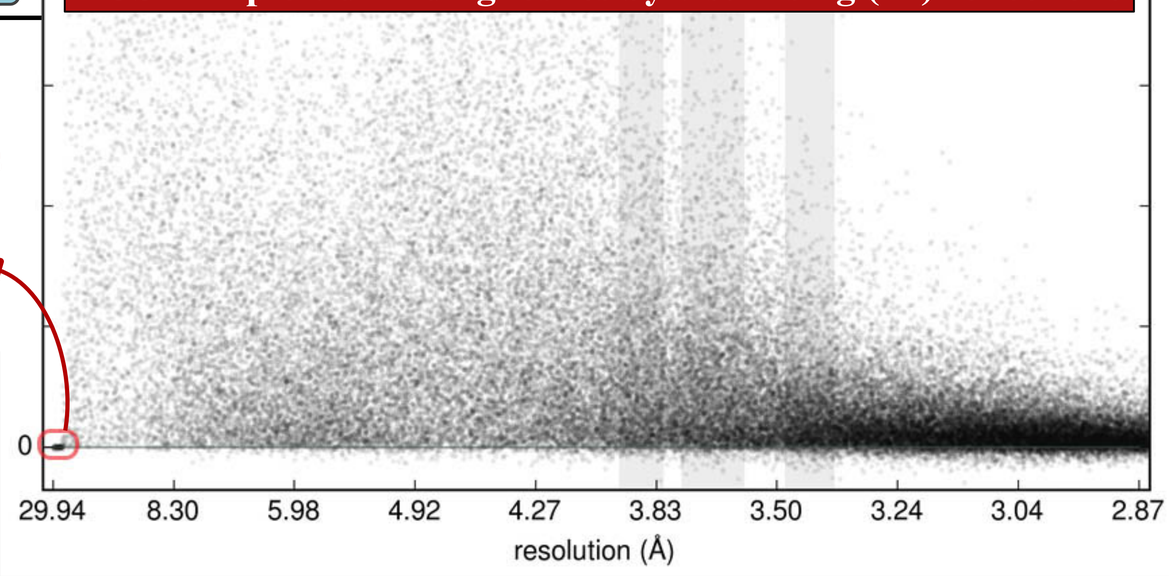


NEMOs

Not-Excluded-unMasked-Outliers

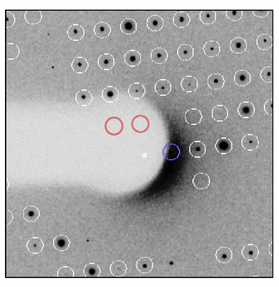
cluster(s) of weak observations at the low-resolution end in processed datasets

AUSPEX plot: data diagnostics by visualising (all) reflections



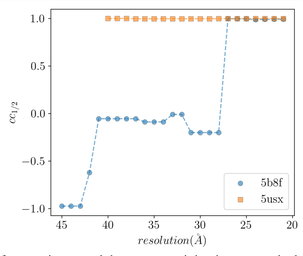
How NEMOs Escape

the Data Reduction Workflow Indexing->Prediction->Integration



Two types of low-angle outliers related to beamstop. Predicted locations of reflections are circled. The blue circle and red circle highlight the Type 1 and Type 2 outliers, respectively. Type 1 outliers are typically strong and detectable using Wilson statistics. Detection and exclusion of the Type 2 outliers remains more difficult.

"Binning-based" Data Statistics



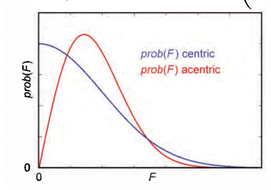
$CC_{1/2}$ of two re-integrated datasets containing beamstop shadow outliers. Each point in the figure is calculated by incorporating all observations below its corresponding resolution. In the case of PDB entry 5usx, no low-angle data should be omitted according to $CC_{1/2}$. For PDB entry 5b8f, setting a low-resolution cutoff at 27.5 Å results in the exclusion of 9 beamstop shadow outliers and 12 well-behaved unique observations from the merged dataset, causing unwanted loss of information. Similar challenges can be encountered with other statistical tools where binning is imperative.

Perfect Beamstop Mask for the Entire Dataset Does That Even Exist?

Crystallographical Statistics

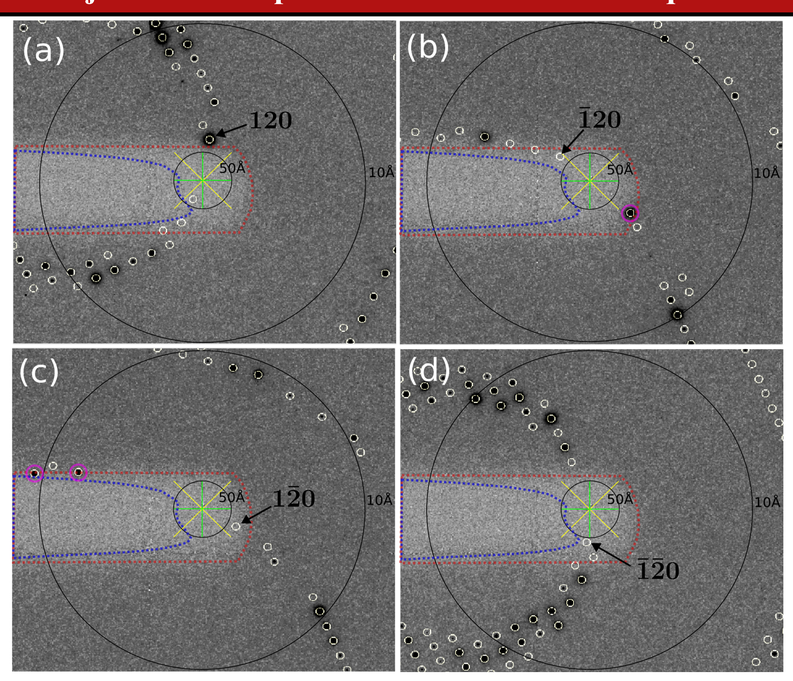
$$p_{\mathrm{a}}\left(E_{\mathrm{O}} < E_{\mathrm{O,meas}}\right) = 1 - \exp\left(-E^2_{\mathrm{O,meas}}\right)$$

$$p_{\mathrm{c}}(E_{\mathrm{O}} < E_{\mathrm{O,meas}}) = 1 - \mathrm{erfc}\left(\frac{E_{\mathrm{O,meas}}}{2^{1/2}}\right)$$



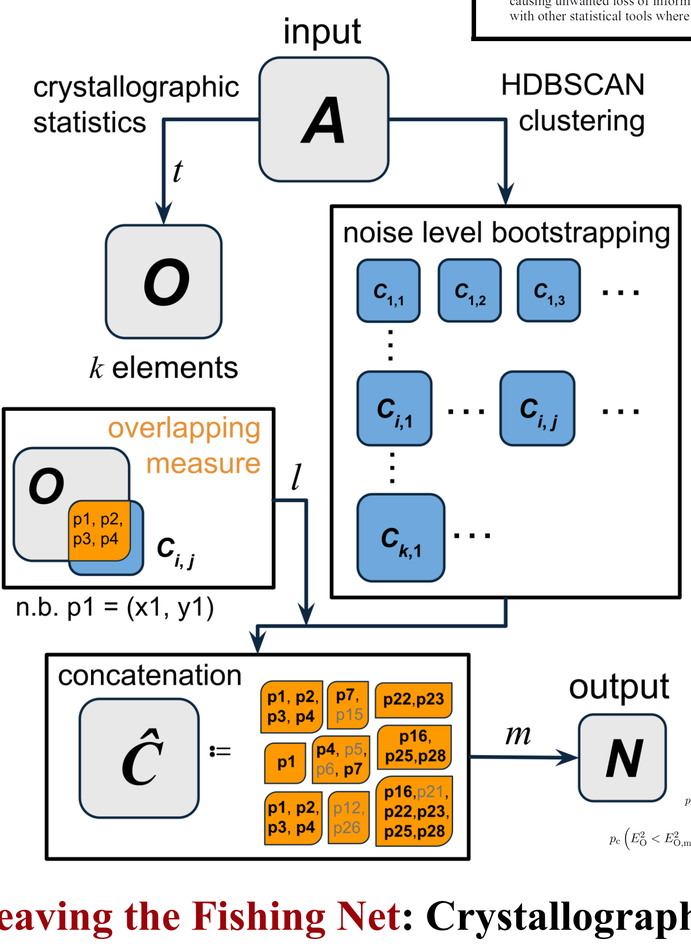
The persistence of NEMOs as clusters of weak signals is exacerbated as a result of applying the French-Wilson method when estimating the structure factors from the intensities. Given the high likelihood of weak intensities in the Wilson distribution, the posterior distribution for weak reflections with significant errors, such as NEMOs, tends to be dominated by the prior Wilson distribution. For amplitudes that are inferred from the Wilson distribution, the minimum ratios of the French-Wilson posterior amplitude and standard deviation for centric and acentric data are 1.324 and 1.913, respectively. Upon inspecting Protein Data Bank (PDB) entries, all NEMOs clusters slightly surpass these lower limits.

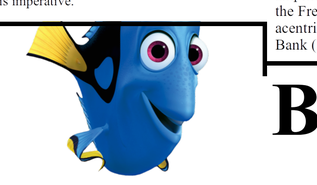
Outlier Rejction : Multiple Measurement of an Unique Reflection



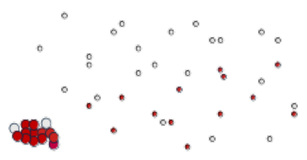
Example of the experimental origin of NEMOs. The images are the corresponding raw detector frames of PDB entry 5b8f. The blue dashed polygon highlights the edge of excluded shaded regions recognised by the best possible *XDS-DEFPIX* trial without masking unobstructed areas on the detectors. The red dashed polygon highlights the edge of how one intuitively would define a mask, manually created with the aid of *dials.image_viewer*. Using the blue mask, the similarity of observations in (b), in (c) and in (d) result in the exclusion of observation in (a). The resulting merged unique reflection then would have a value close to 0, but an "algorithm acceptable" uncertainty. Using the red mask, the unique reflection can be properly recorded, as the other symmetry equivalent observations are completely masked. However, in (b) and (c) such a mask results in the masking of unbiased observations (highlighted with magenta circles).

Find NEMOs





But Keep Dori



1. Constructing a pool for potential beamstop shadow outlier

$P(\mathrm{E<Eo}) < t$, d > 10 Å

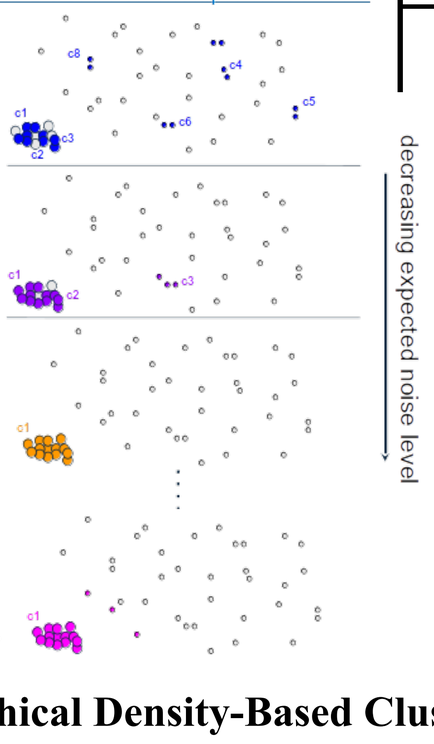
$t > 10^{-2}$

set O =
{p2,p3,p4,p5,p6,p7,p8,p9,p10,
p11,p13,p14,p15,
p21, p22,...,
p26, c28,...,
p38,...,
p49....}

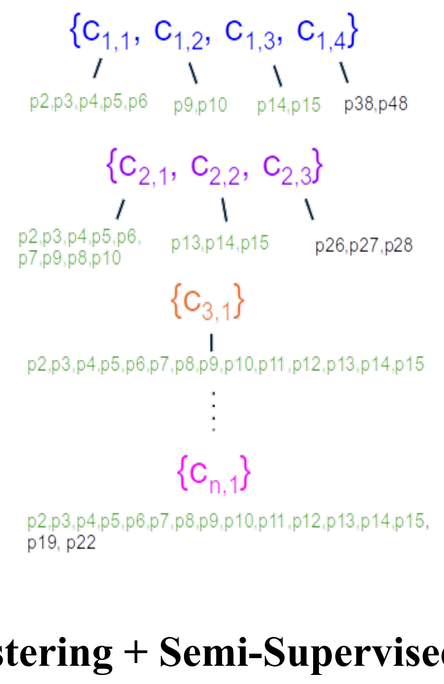
$$p_{\mathrm{a}}\left(E^2_{\mathrm{O}} < E^2_{\mathrm{O,meas}}\right) = \tfrac{1}{2}\left[\mathrm{erfc}\left(\frac{-E^2_{\mathrm{O,meas}}}{2^{1/2}\sigma(E^2_{\mathrm{O}})}\right) - \exp\left(\frac{\sigma(E^2_{\mathrm{O}}) - 2E^2_{\mathrm{O,meas}}}{2}\right)\mathrm{erfc}\left(\frac{\sigma(E^2_{\mathrm{O}}) - E^2_{\mathrm{O,meas}}}{2^{1/2}\sigma(E^2_{\mathrm{O}})}\right)\right]$$

$$p_{\mathrm{c}}\left(E^2_{\mathrm{O}} < E^2_{\mathrm{O,meas}}\right) = \int_{-\infty}^{E^2_{\mathrm{O,meas}}}\int_{0}^{\infty}\frac{1}{\left(2\pi\sigma^2_{E^2_{\mathrm{O}}}\right)^{1/2}}\exp\left[-\frac{\left(E^2_{\mathrm{O}} - E^2\right)^2}{2\sigma^2_{E^2_{\mathrm{O}}}}\right]\frac{1}{(2\pi E^2)^{1/2}}\exp\left(-\frac{E^2}{2}\right)\mathrm{d}E^2\mathrm{d}E^2_{\mathrm{O}}$$

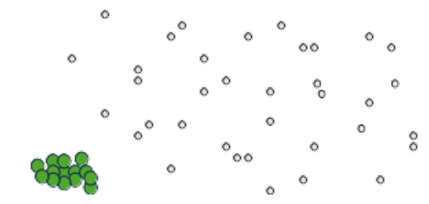
2. HDBSCAN on multiple noise level

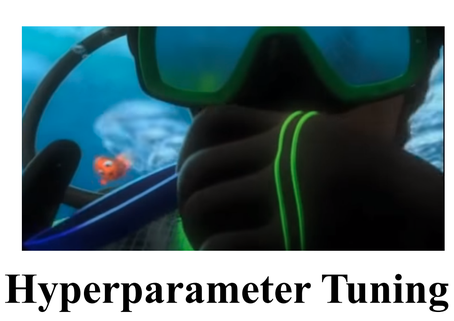


3.robust assignment



set N =
{p2,p3,p4,p5,p6,p7,p8,p9,p10
,p11,p12,p13,p14,p15}



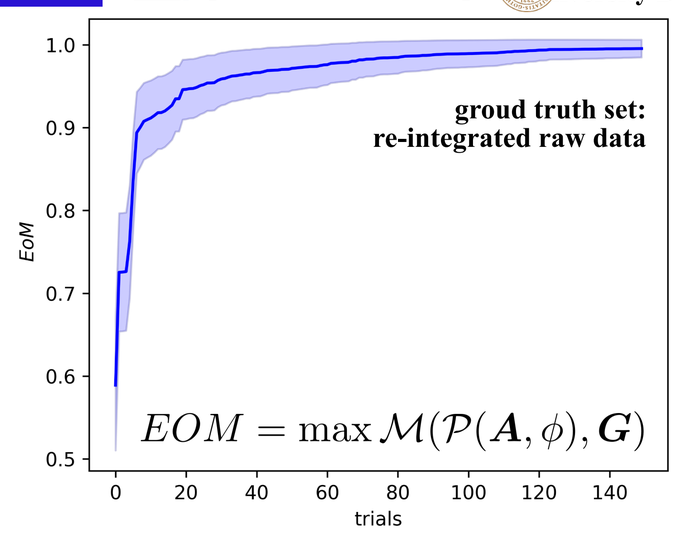


Weaving the Fishing Net: Crystallographical Statistics + Hierachical Density-Based Clustering + Semi-Supervised Hyperparameter Tuning

	input	precision	accuracy	sensitivity	specificity
beamstop_shadow_outlier in *cctbx*	F_{obs}	0.809 ± 0.047	0.729 ± 0.053	0.794 ± 0.048	0.584 ± 0.058
AUSPEX clustering-derived detection	$F_{\mathrm{obs}}/\sigma(F_{\mathrm{obs}})$	$\mathbf{0.974 \pm 0.019}$	$\mathbf{0.960 \pm 0.023}$	$\mathbf{0.959 \pm 0.023}$	$\mathbf{0.960 \pm 0.023}$
AUSPEX clustering-derived detection	$I_{\mathrm{obs}}/\sigma(I_{\mathrm{obs}})$	0.958 ± 0.040	0.935 ± 0.050	0.955 ± 0.042	0.875 ± 0.067

MORE RAW IMAGES/UNTRUNCATED PROCESSED DATA PLS!

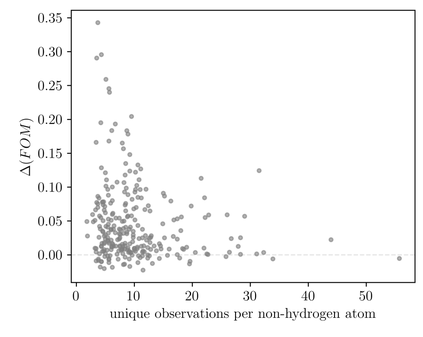
zenodo NIH proteindiffraction.org University Depos



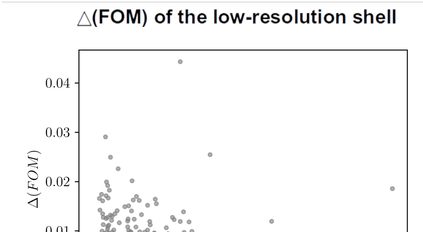
Why You Should Care About NEMOs

	range
NEMOs %	0.75 - 0.0015
resolution	3.38 - 0.97
R-free	0.341 - 0.1191
# unique observations per non-hydrogen atoms	1.8 - 55.7
solvent %	32.25 - 81.83
synchrotron	NSLS-II,ESRF,BESSY, PETRA III,APS,Diamond, australian synchrotron, ALS,SSRL,LNLS
data reduction	XDS,HKL-3000, HKL-2000,DIALS, MOSFLM,autoPROC,DENZO

- 270 PDB entries
- data contain NEMOs
- confirmed beamstop outliers through re-integration
- no twinning or tNCS
- re-refine & rebuild by PDB-REDO 8.03
- model: deposited coordinates
  + 0.25 shake with phenix.pdbtools
- data: original vs. NEMOs removed

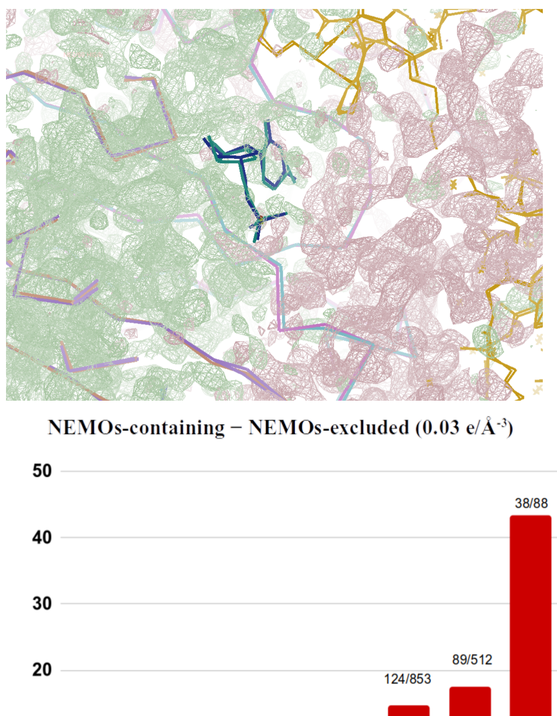


△(FOM) of the low-resolution shell



△(FOM) of the whole dataset

figure-of-merit (mean phase error) improves after removing NEMOs



NEMOs Found, What Now

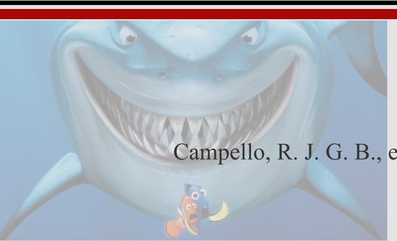
Guide for Low-resolution Cutoff

Minimize Information Loss

Refine Processed Data *post-mortem*

Objective Metrics for Automatic Beamstop Mask Generation

(Central) Beamstop Mask Free When Reaching 1.0 Accuracy?



Thank Arwen Pearson, Dale Tronrud, Kay Diederichs and Randy Read for Helping us Find NEMOs

Ref. Read, R. J. (1999). Acta Crystallogr. D Biol. Crystallogr. 55, 1759–1764.
Read, R. J. & McCoy, A. J. (2016). Acta Crystallogr. D Struct. Biol. 72, 375–387.
Campello, R. J. G. B., et. al. (2013). Advances in Knowledge Discovery and Data Mining, Vol. 7819. Springer Berlin Heidelberg.
Mishra, S., Monath, N., Boratko, M., Kobren, A. & McCallum, A. (2022). AAAI. 36, 7788–7796.

Bundesministerium für Bildung und Forschung

Help Us Improve the Code
https://github.com/thorn-lab/AUSPEX
Hate to Build? Try Our Web Service!