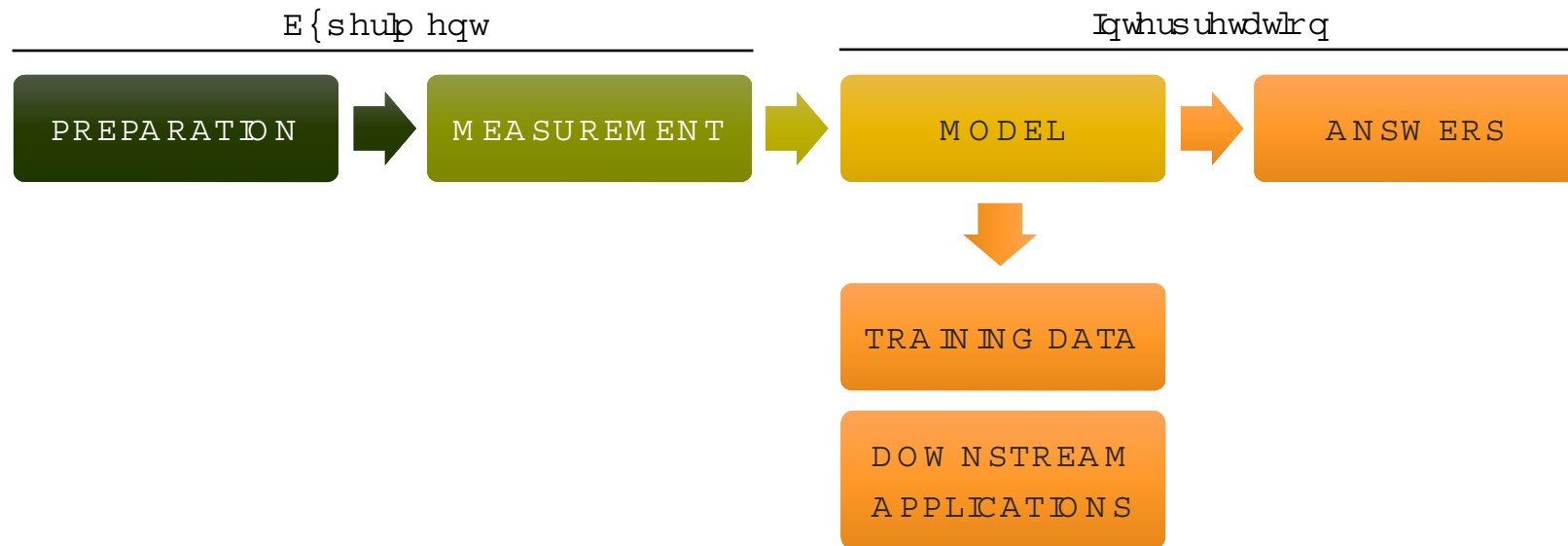


Visual Diagnostics for Macromolecular X-Ray Diffraction: AUSPEX

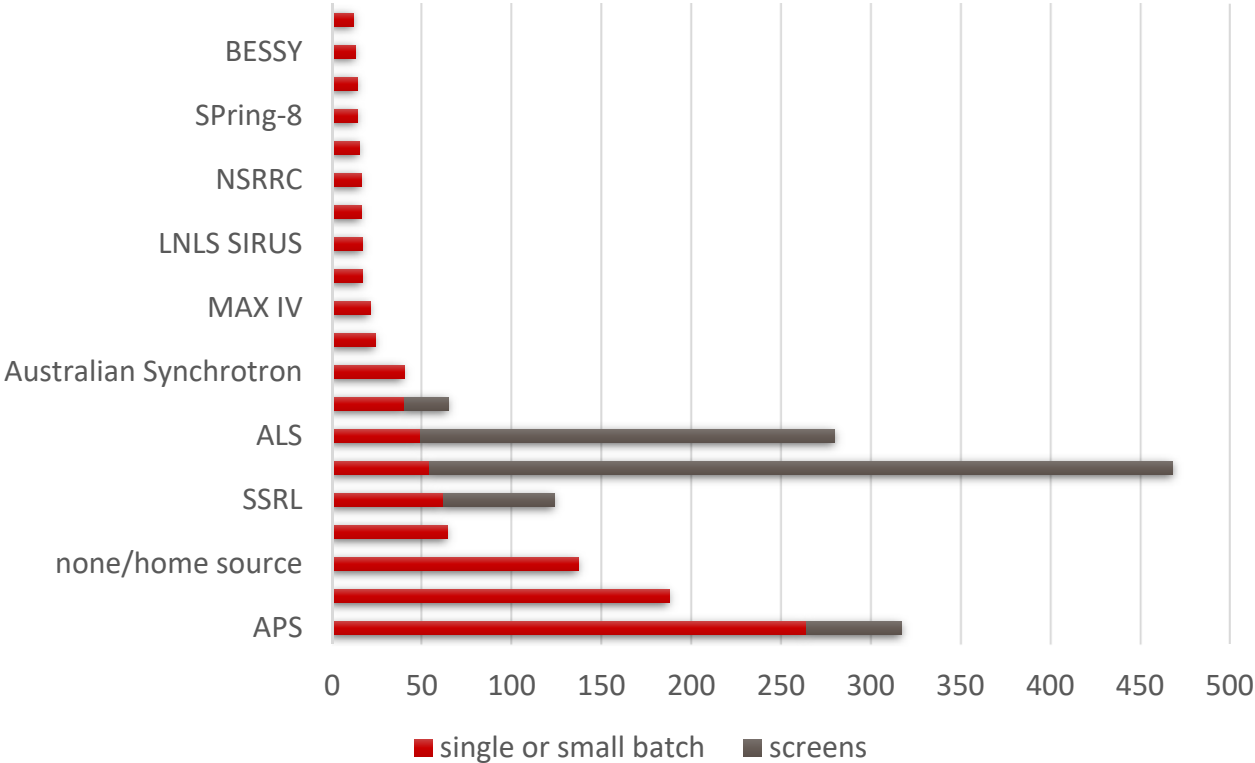
Andrea Thorn



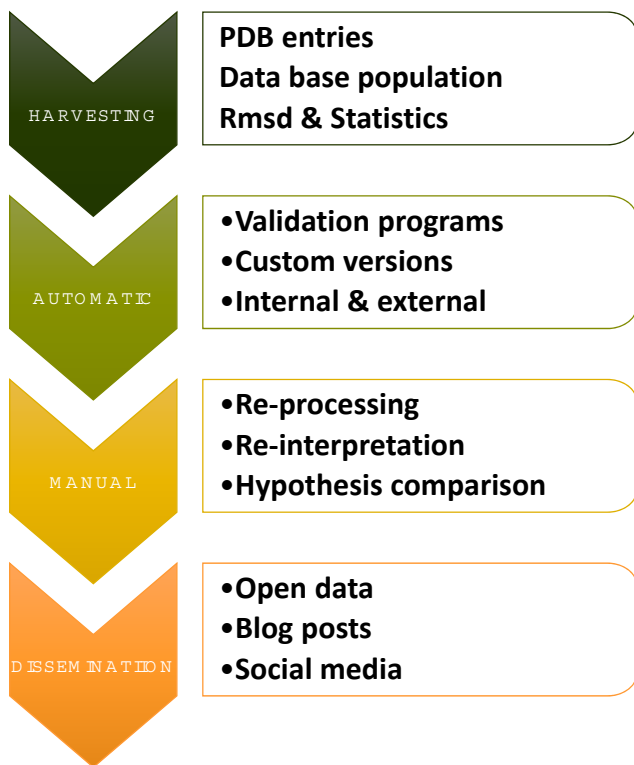
“Central Dogma” of Structural Biology



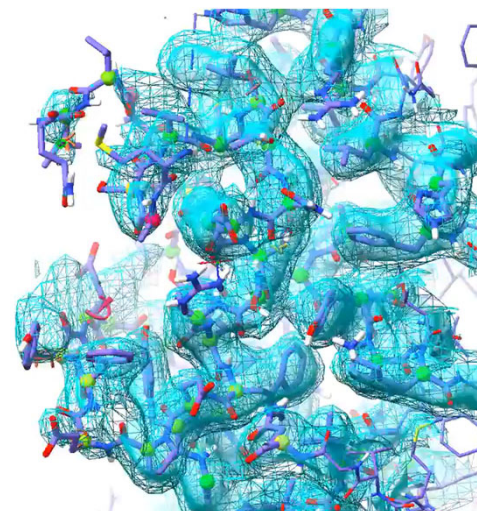
Quick aside: Large Facilities vs. COVID-19



Quick aside: Coronavirus Structural Task Force



Pds d.lj0dnh surwhdvh
+PDE 9z < f,



RNA
srd|p hudvh
erxqg wr
uhp ghvlylu
+PDE :ey5,



AUSPEX: Data Pathology Diagnostics

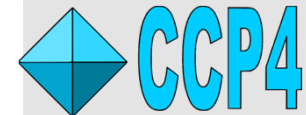
- Diffraction data problems more common than generally thought
- Implications for processing & structure solution
- **We need better indicators for diffraction data problems!**

Verbundforschungsprojekt:

AUSPEX - New AI-based, Visual and Automatic Diagnostics for Macromolecular Structure Determination at Large Facilities

Thorn, A.* et al. (2017) AUSPEX: a graphical tool for X-ray diffraction data analysis, *Acta Cryst D73*, 729-737

Collaborators:

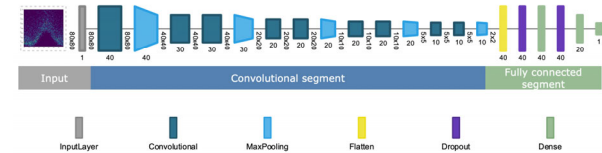
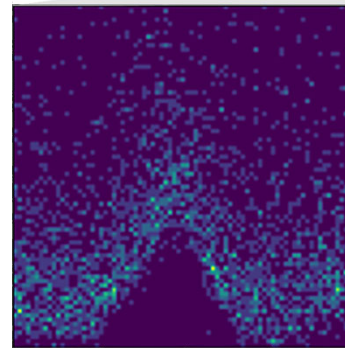
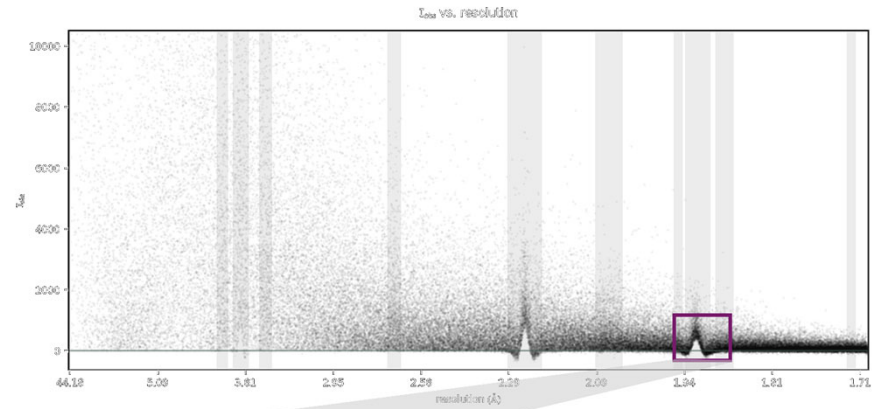
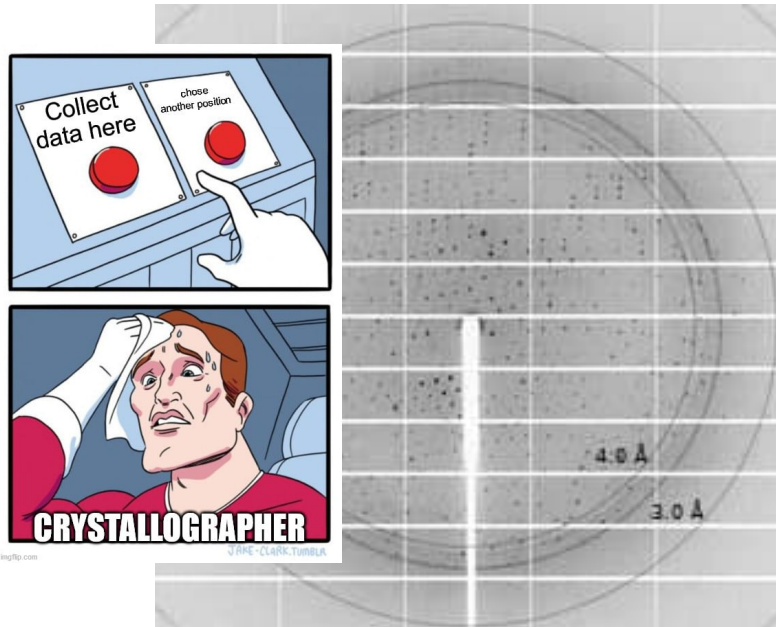


Funded by



Bundesministerium
für Bildung
und Forschung

Example: Ice rings



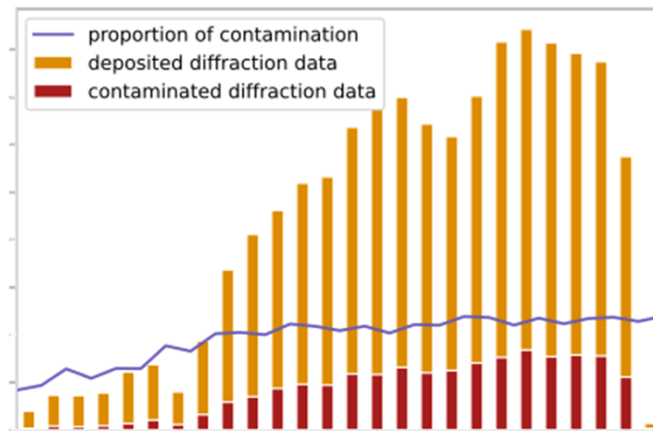
Nolte, K., Gao, Y., Stäb, S., Kollmannsberger, P., Thorn, A. (2022) Detecting ice artefacts in processed macromolecular diffraction data with machine learning, *Acta Cryst D78*, 187-195.

Diffraction image: Gianluca Santoni, ESRF



Neural network recognition

	True positives	True negatives	Accuracy	Sensitivity	Specificity
phenix.xtriage	13/44	142/156	78 %	30 %	91 %
CTRUNCATE	23/44	86/156	55 %	52 %	55 %
AUSPEX Icefinder	24/44	133/156	83 %	55 %	85 %
P _{ice}	29/44	147/156	88 %	66 %	94 %
AUSPEX Helcaraxe	38/44	153/156	96 %	86 %	98 %

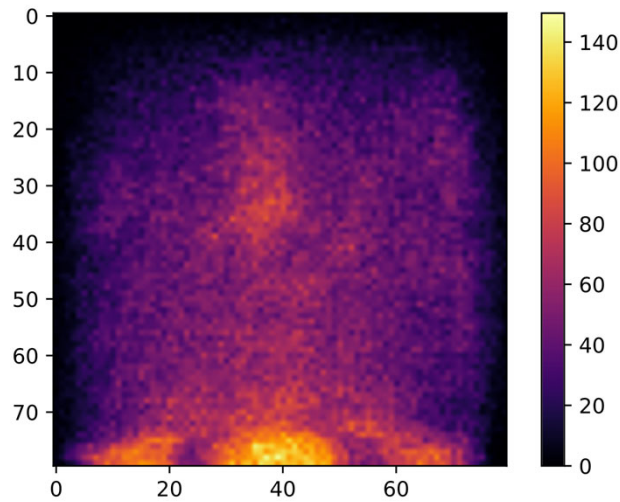
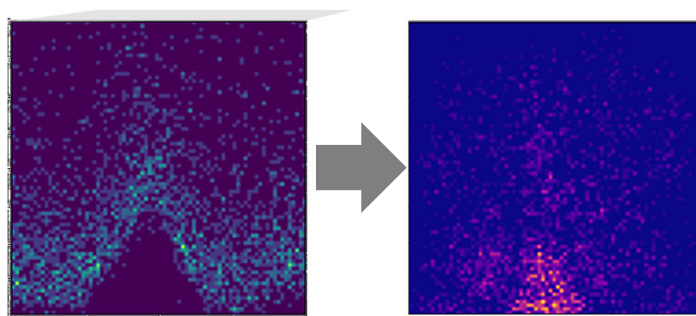


18.5% !

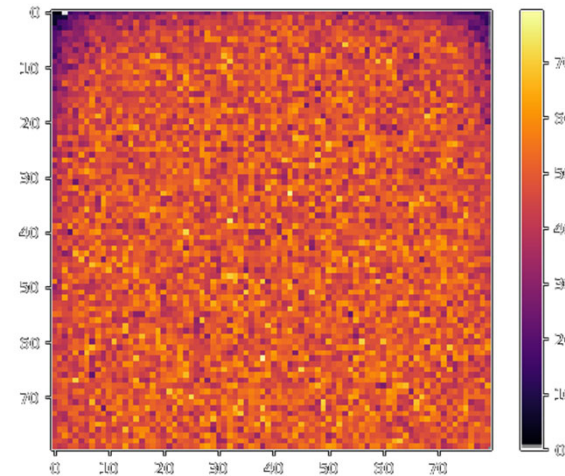


Nolte, K., Gao, Y., Stäb, S., Kollmannsberger, P., Thorn, A. (2022) Detecting ice artefacts in processed macromolecular diffraction data with machine learning, *Acta Cryst D78*, 187-195.

Explaining the black box: Sensitivity

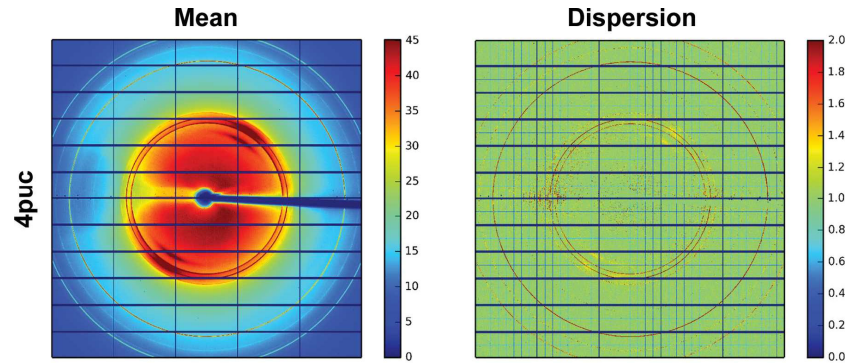


Edge weights =

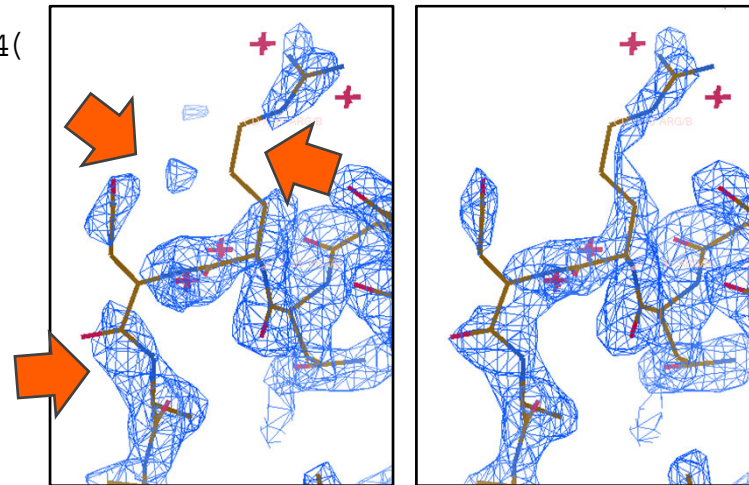


Smilkov, D., Thorat, N., Kim, B., Viégas, F. & Wattenberg, M. (2017). SmoothGrad: removing noise by adding noise *ArXiv*.

New background estimation in DIALS



$R_{i,h} = 5; \mathbb{A}(\dots)$



$R_{i,h} = 5.6 \mathbb{B}(\dots)$

Parkhurst, Thorn, Vollmar, Winter, Waterman, Fuentes-Montero, Gildea, Murshudov, Evans (2017) *IUCr* 4, 626-638.



Machine Learning in Cryo-EM and MX

Support Vector Machines

Algorithm for supervised classification or unsupervised clustering; training data can be sparse

Classification
Assignment of input to defined class

- Particle picking
- Sequence assignment in modelling
- MX model building & validation
- Cryo-EM Secondary Structure prediction
- Identification of metal ion type in MX models
- Prediction of structure solvability in MX

Convolutional Neural Networks

“Deep learning”; training data need to be large; can be supervised or unsupervised

Classification
Assignment of input to defined class

- Micrograph denoising; motion correction
- Contrast Transfer Function (CTF)
- Particle picking & elimination
- Reconstruction map resolution
- Reconstruction map improvement
- Secondary Structure recognition in Map
- Residue recognition in maps
- Crystal recognition during crystallization
- Crystallization condition optimization
- Finding pathology in diffraction image
- Neutron diffraction improvement
- Spot finding in serial crystallography
- Integration of MX intensities
- Model building in MX & Cryo-EM
- Model quality assessment

Object detection
Recognition and localization of defined object

- Particle picking
- Crystal centering

Regression
Adaptation of a mathematical function

- Reconstruction map resolution/quality
- Macromolecule trajectory in Cryo-EM

Bayesian Machine Learning

Training data can be sparse; priors can be used; supervised or unsupervised

Classification
Assignment of input to defined class

- Merging diffraction data
- Atomic displacement similarity
- Improvement in anomalous phasing

Clustering

Finding clusters in input data; training data can be sparse; unsupervised

- Particle picking
- Atomic displacement similarity

Other regression methods

K-nearest neighbours, random forest, gradient boosting machine etc.

Regression
Adaptation of a mathematical function

- Recognition of ligands in electron density
- Peptide flips in protein models
- Prediction of structure solvability in MX

Other types of neural networks

supervised or unsupervised; nodes may be added; no convolutional layers

- Diffraction quality estimation
- Integration of MX intensities
- Crystallization condition prediction

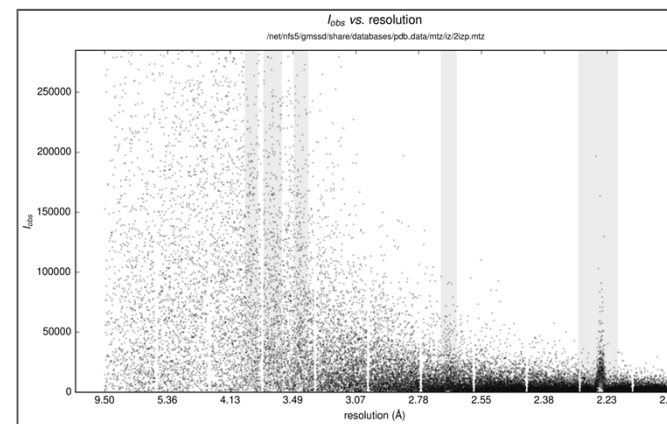
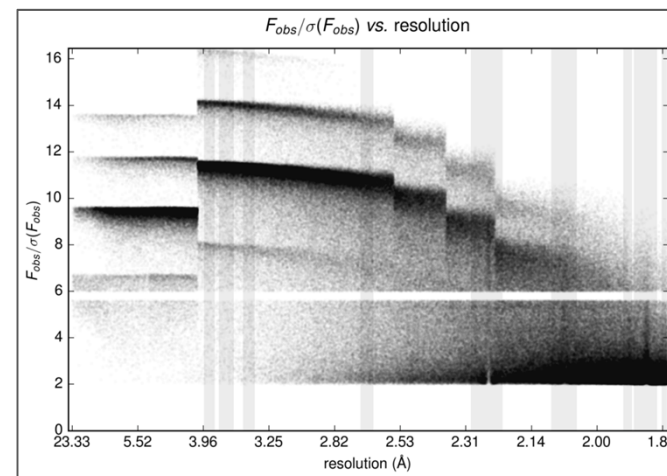
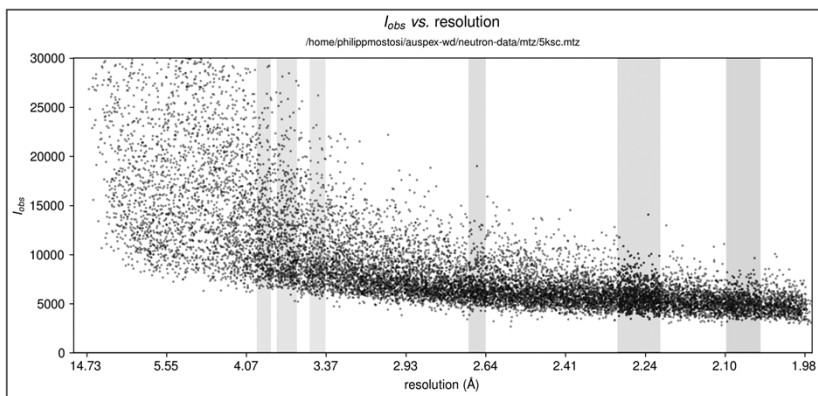
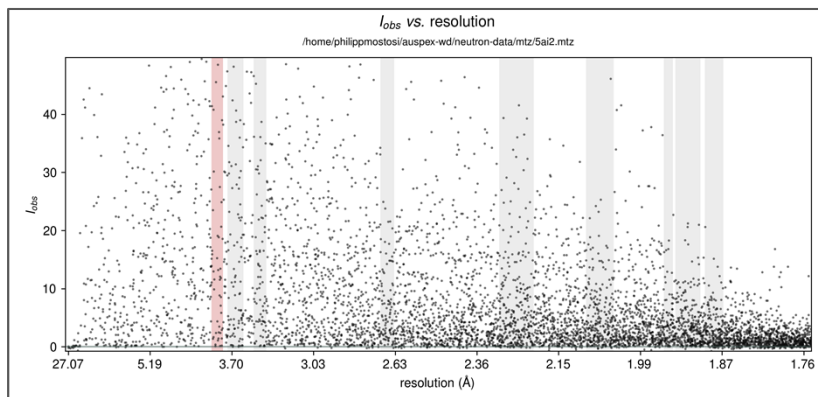
AI methods in experimental structural biology

Download poster &
add to list at
thorn-lab.com



Thorn, A.* Artificial intelligence in the experimental determination and prediction of macromolecular structures (2022) *Curr. Opin. Struct. Biol.* 74, 102368

More problems

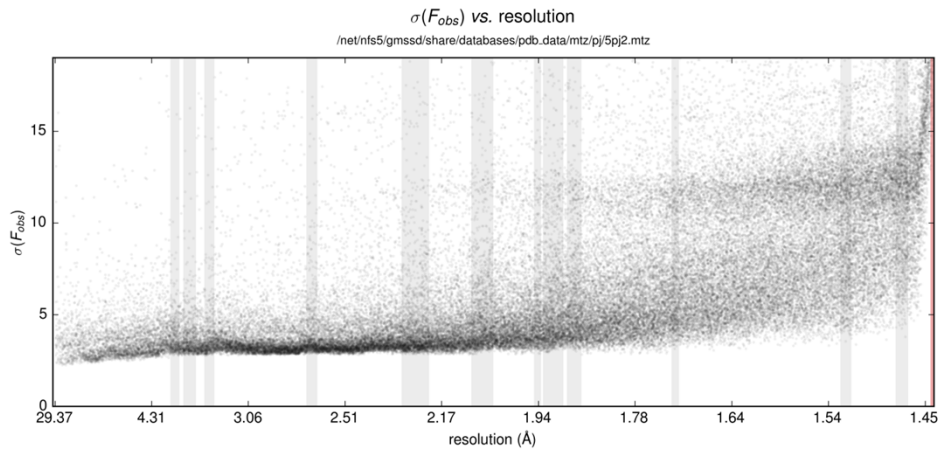


Thorn, A., Parkhurst, J.M., Emsley, P., Nicholls, r., Evans, G., Vollmar, M. & Murshudov, G.N. (2017) AUSPEX: a graphical tool for X-ray diffraction data analysis, *Acta Cryst D73*, 729-737.

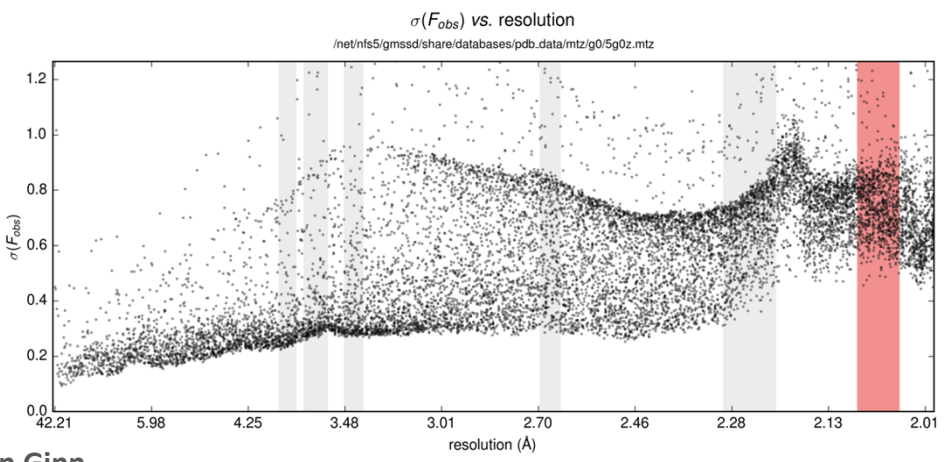


Polyhedrin: σ of XFEL data

P D E 8 P M 5
 C r q y h q w l r q d o [0 u d |
 [D S / A I M L E S S



P D E 8 G 3]
 [I E L
 C u | v w I E L V 1 3 1 8 1 5



Data kindly provided by Tom White and Helen Ginn





www.auspex.de

A service provided by



[HOME](#)

[ANALYZE PDB/MTZ](#)

[VIEW JOBS](#)

[EXAMPLES](#)

[DISCLAIMER](#)

AUSPEX is a diagnostic tool for graphical X-Ray data analysis, which enable users to visually and automatically detect ice-ring artefacts and other problems in integrated X-ray diffraction data.

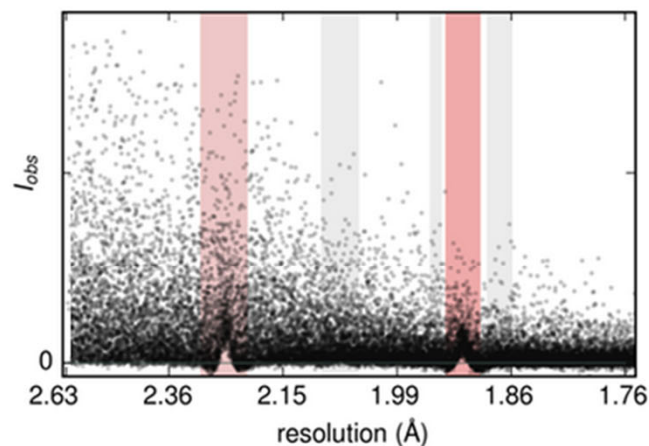
If you would like to plot your own data

[RUN AUSPEX](#)

Some examples of common pathologies are available

[EXAMPLES](#)

For a more detailed description of *AUSPEX* and our systematic study of the PDB, please look at the bottom of this page.



**PLEASE
CITE**

Thorn, A., Parkhurst, J.M., Emsley, P., Nicholls, R., Evans, G., Vollmar, M. & Murshudov, G.N. (2017)
AUSPEX: a graphical tool for X-ray diffraction data analysis, *Acta Cryst D73*, 729-737.

This helps us to develop AUSPEX as a free service.



Summary

- **Understanding experimental data (and errors) is crucial to avoid problems and pitfalls in macromolecular structure determination!**
- **We need better diagnostics to ensure best results from automation**
- **AUSPEX: tool for visual & easy diffraction data analysis**
Webserver: www.auspex.de
- **Well curated and open data are badly needed!**
- **Particularly for machine learning (limited by training data)**

There is much to be done!



Acknowledgements

Gianluca Santoni & Max Nanao, ESRF Grenoble, France

Manfred Weiss, HZB BESSY, Berlin

Philip Kollmannsberger, HHU Düsseldorf

Philipp Mostosi, Method Park, Erlangen

Henry Chapman & Adrian Mancuso, EuXFEL, Hamburg

Esko Oksanen, ESS Lund, Sweden

Kristopher Nolte, Hamburg University of Applied Sciences

Paul Emsley & Rob Nicholls, MRC-LMB Cambridge, UK

Melanie Vollmar, EBI Hinxton, UK

Gwyndaf Evans, Diamond, UK

James Parkhurst, Rosalind-Franklin Institute, UK



Andrea Thorn
Group Leader



Yunyun Gao
Postdoc



Pairoh Seeliger
Coordinator



Jan Schreiber
Student Assistant



Max Edich
PhD student



Oliver Kippes
Student assistant

