Science and Technology Facilities Council

Scientific Computing

ada lovelace centre

Computational Biology

Martyn Winn 20th May 2024

SCD Computational Biology theme

Current facility activities

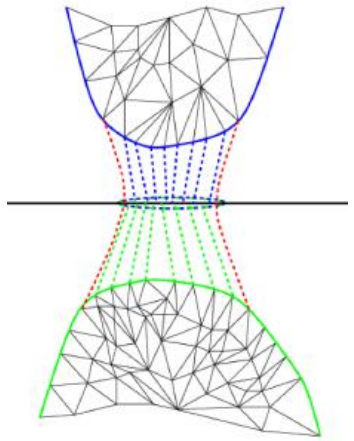
- CCP4 & CCP-EM
- Molecular simulation and modelling
- Data & machine learning
- Computed tomography

Future opportunities

Our history ….

- 1990s - CCP4 team at Daresbury in Synchrotron Radiation Department
- Ca. 2005 Molecular Dynamics simulation activity begun
- CCP4 move to Diamond House (2009) then Research Complex (2010)
- CCP-EM team created (2012) – help curate eBIC software
- 2020 Computed Tomography team added
- Jan 2024 Computational Biology Theme created

Computational Biology Theme



Macromolecular Crystallography

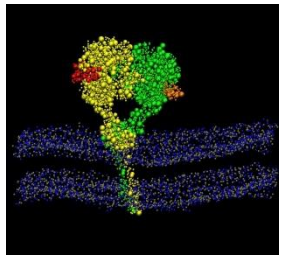
Methods development
Software suite

CCP4

Biomolecular simulation

Dynamics / *in silico* expts
Tools and HPC access

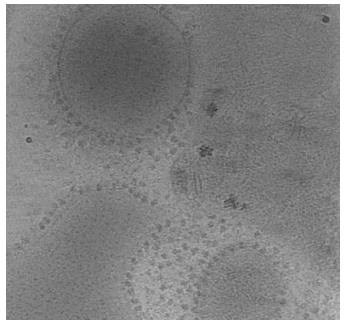
CCPBioSim



Molecular and Cellular Electron Microscopy

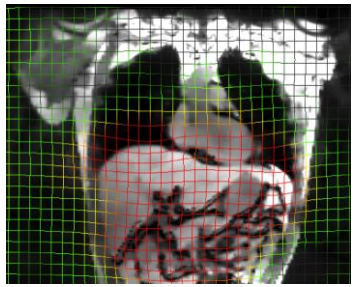
Methods development
Software suite

CCP-EM



Computed Tomography and Imaging

Software libraries
CCPi – materials
CCPSyneRBI – medical imaging



Elsewhere in SCD …

Materials – MD, mesoscale, QM/MM, electronic structure
Engineering – meshes, finite element
Maths - algorithms
SciML – machine learning
DAaaS - services
Cloud – hardware backend

Harwell facilities

UK and global communities

HECBioSim

SyneRBI

CCPBioSim

CCP-EM

CCPi
TOMOGRAPHIC IMAGING

CCP4

CoSeC
Computational Science Centre
for Research Communities

Our interests ….

- Largely molecular / structural biology
  - ➢ Also wider imaging and informatics
- Closely tied to experimental data
  - ➢ Modelling for interpretation and in silico experiments
- Grown organically over 30 years
- Variety of funding, mostly external

- Development of software, long term and stable
- Development of methods, though not deep theory
- Data / metadata handling important

In practice, software libraries and frameworks to support multiple workflows

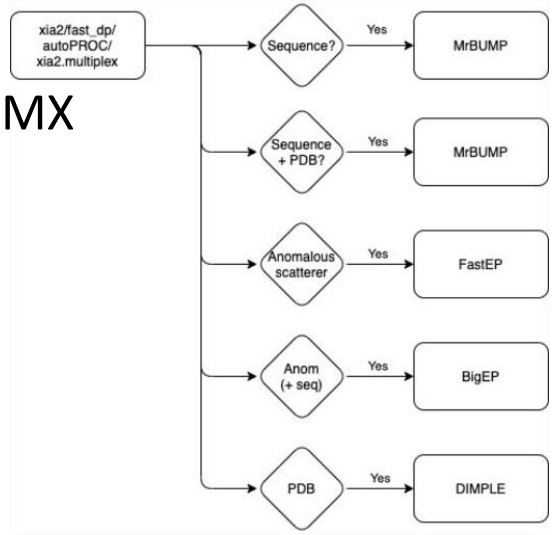
UKRI Medical Research Council
UKRI Biotechnology and Biological Sciences Research Council
UKRI Engineering and Physical Sciences Research Council
wellcome
Government Office for Technology Transfer
Chan Zuckerberg Initiative

	Diffraction	Microscopy	Tomography	Scattering
1. Expt.	CCP4, DIALS	CCP-EM	CCPi	MD for neutron reflectometry
2. Sim.	Electron diffraction	AlphaFold, Parakeet	Digital twins	Molecular Dynamics
3. Data	CCP4Cloud	CCP-EM Pipeliner	CIL	Provenance for MD
4. AI/ML	Solvent content prediction	2D/3D classification	De-noising	Learned potentials

CCP4 and CCP-EM

CCP4

- Macromolecular crystallography
- Originally X-ray, now neutron and electron
- Diffraction pattern to structure
- Large downloadable software suite
- Joint workshop with Diamond every Nov/Dec; Study Weekend
- Software used by Diamond MX
  - DIALS
  - MrBUMP
  - Crank2
  - etc



www.ccp4.ac.uk

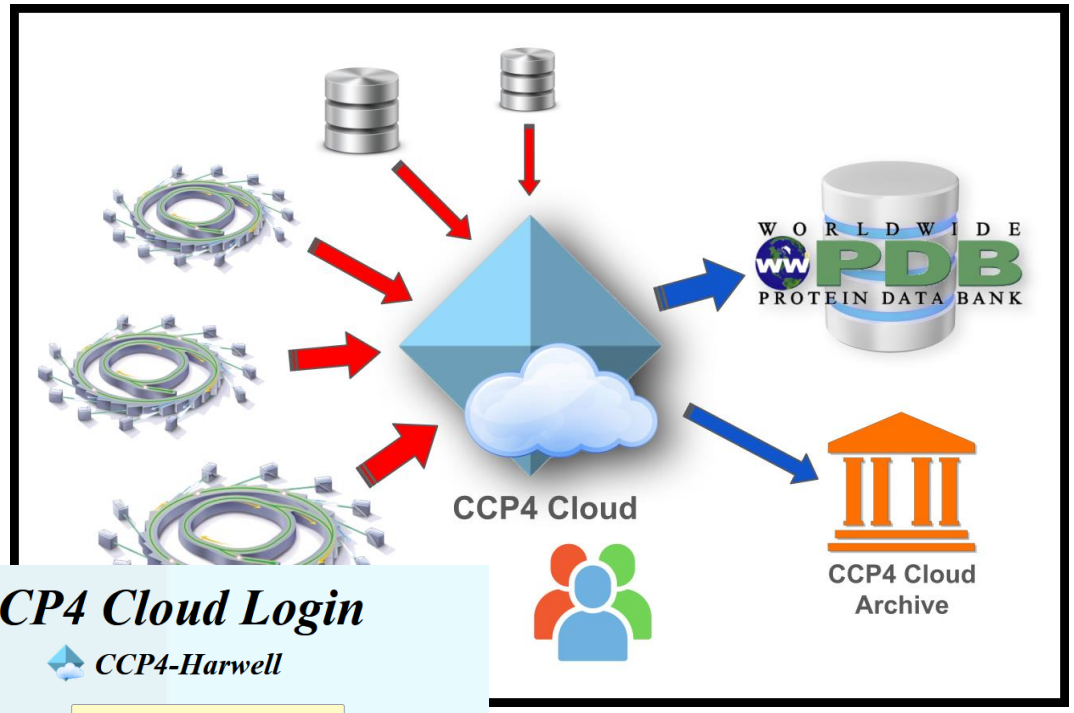
CCP-EM

- Cryogenic electron microscopy and tomography
- From frames/micrographs to map/structures
- Downloadable software suite
- Software used by eBIC
  - CCP-EM suite
  - Relion
  - Pipeliner based automated workflows

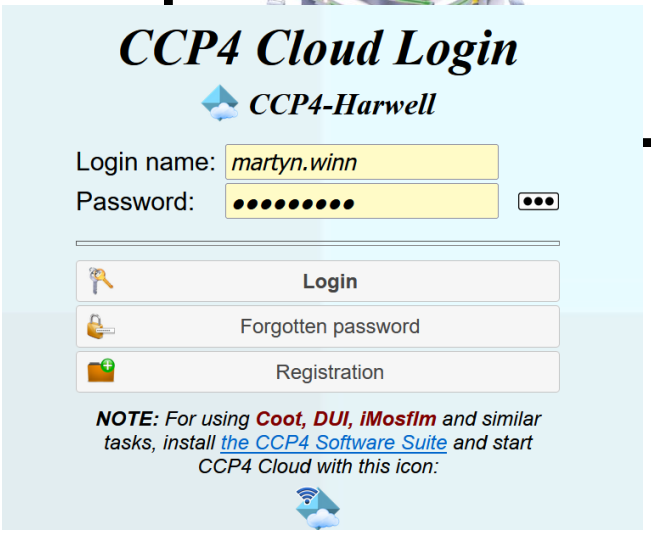
www.ccpem.ac.uk

CCP4Cloud

Jools Wills



- Online solution for crystallographic computation
  diffraction images → structure refinement → deposition in the PDB
- Main service at RAL, backed by SCD Cloud, but other instances have been created.
- Issue: upload of diffraction images difficult
  ⇒ usage typically starts with reduced data.



- With Diamond, working on DataLink between data collected at MX beamlines and CCP4 Cloud.
  - Pull-type link from CCP4Cloud.
  - Push-type link (using Node JS API).
  - Globus Endpoint.

Diffraction across the facilities

- Diffraction experiments on MX beamlines (X-ray), eBIC (electrons), ISIS (neutrons).
- Workflow similar in each case. Differences due to particle properties (e.g. flat Ewald sphere for electrons) or experimental setup (time-of-flight for neutrons).

- Processing of diffraction images to spot intensities carried out by DIALS software.
  - Originally for X-rays, replacing old CCP4 software
  - Adapted to neutron TOF diffraction experiments, starting Oct 2020
  - Adapted to ED from 2018 onwards

- DIALS core well-maintained.
- Collaboration Diamond, Lawrence Berkeley, CCP4.
- Customisations for particular experiment types.

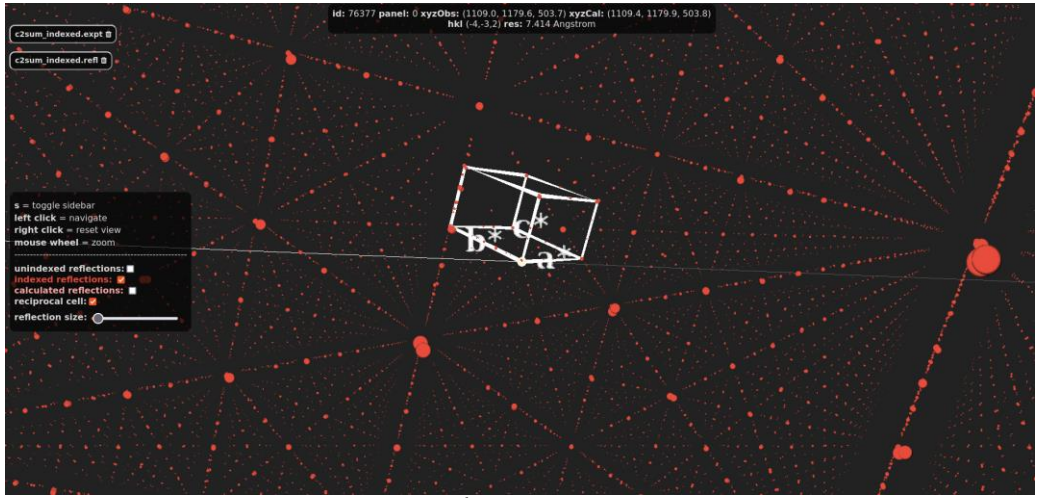
Gwyndaf Evans

David Waterman

DIALS
Diffraction Integration for Advanced Light Sources

https://dials.diamond.ac.uk/

Extending DIALS to Polychromatic Experiments



Reciprocal Lattice Viewer

DIALS
Diffraction Integration for Advanced Light Sources

three.js

New browser-based GUI

Science and Technology Facilities Council
ISIS Neutron and Muon Source

SXD
LMX

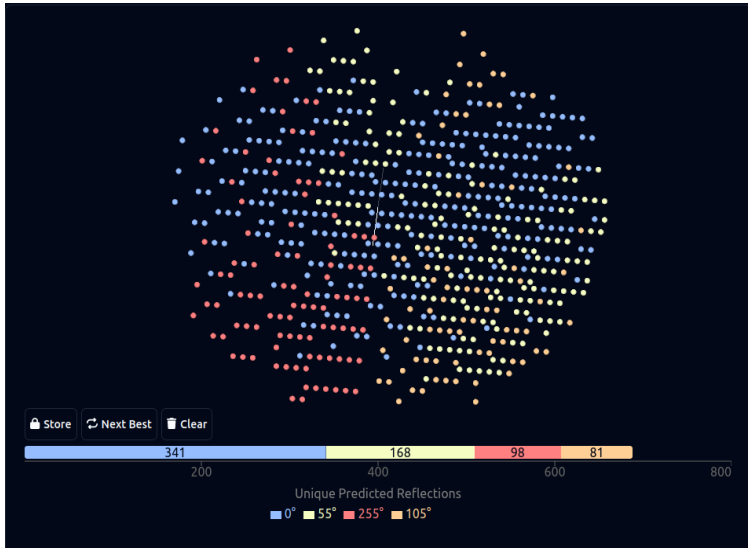
OAK RIDGE National Laboratory

MANDI
TOPAZ

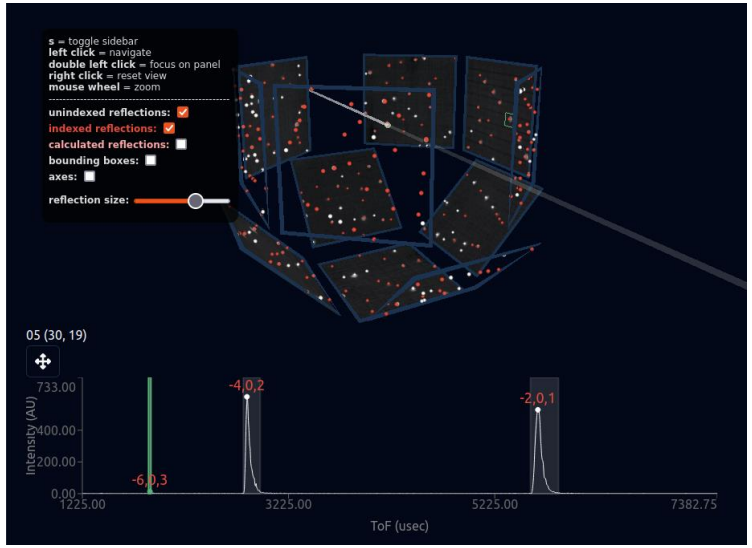
ess EUROPEAN SPALLATION SOURCE

NMX

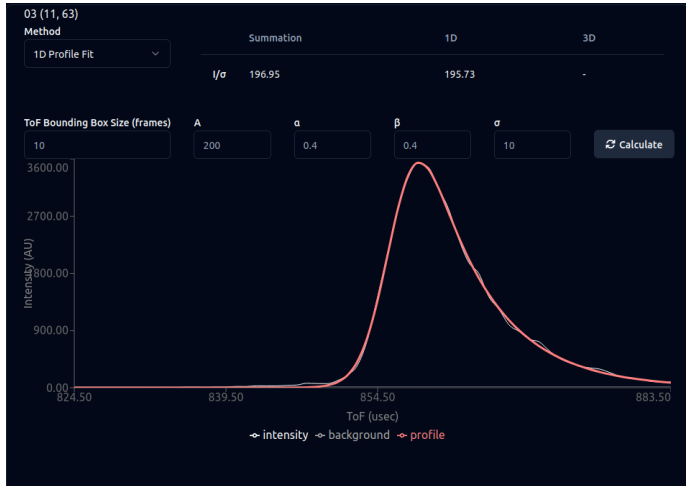
X-ray Laue



Experiment Planner



Experiment Viewer



Integration Profiler

David Mcdonagh

Silvia Capelli

Auto-processing of electron diffraction data


David Waterman


David Mcdonagh


Marko Petrovic

AutoED


detect


analyse

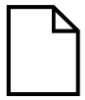

convert


process

eBIC

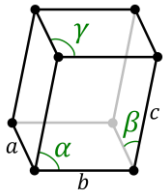
Microscope


ED data


metadata

Diamond filesystem

NeXus



unit cell


mtz file

Diamond cluster

xia2

DIALS


Daniel Hatton


Eilis Bragginton


David Owen

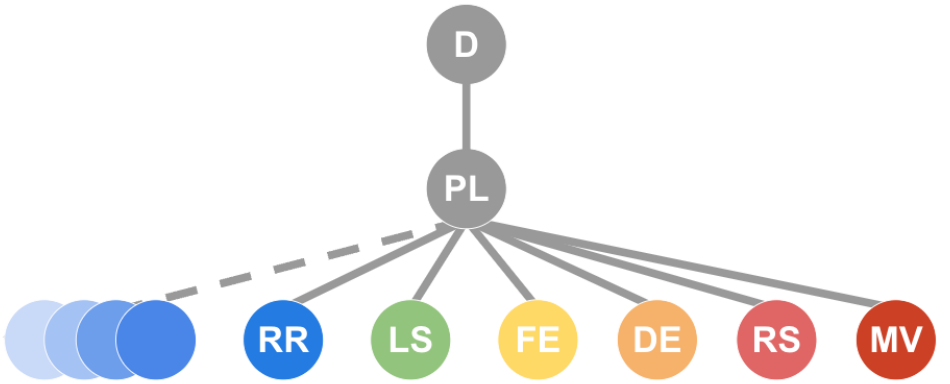
CCP-EM Pipeliner

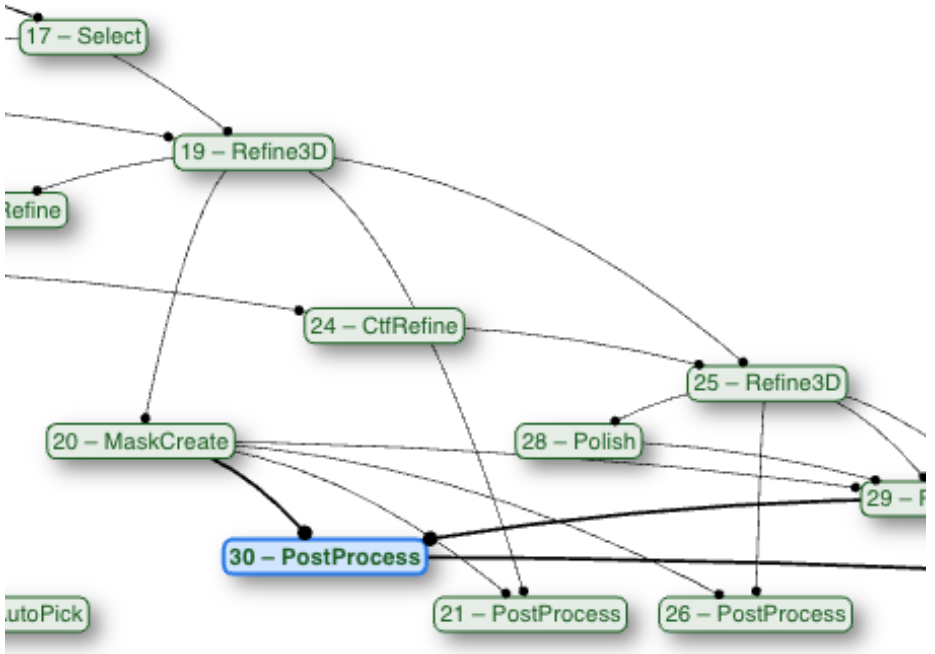
- ***Business logic layer decoupled from tasks and UI***
- Python 3
- MPL 2.0 licence
- https://gitlab.com/ccpem/ccpem-pipeliner
- Directed Acyclic Graph data flow
  - Metadata tracking
  - Jobs have input/output nodes
- Plugin architecture
  - 100-500 lines per app
  - Open - supports external apps
- CLI / UI / scripting APIs

Matt Iadanza

Sjors Scheres

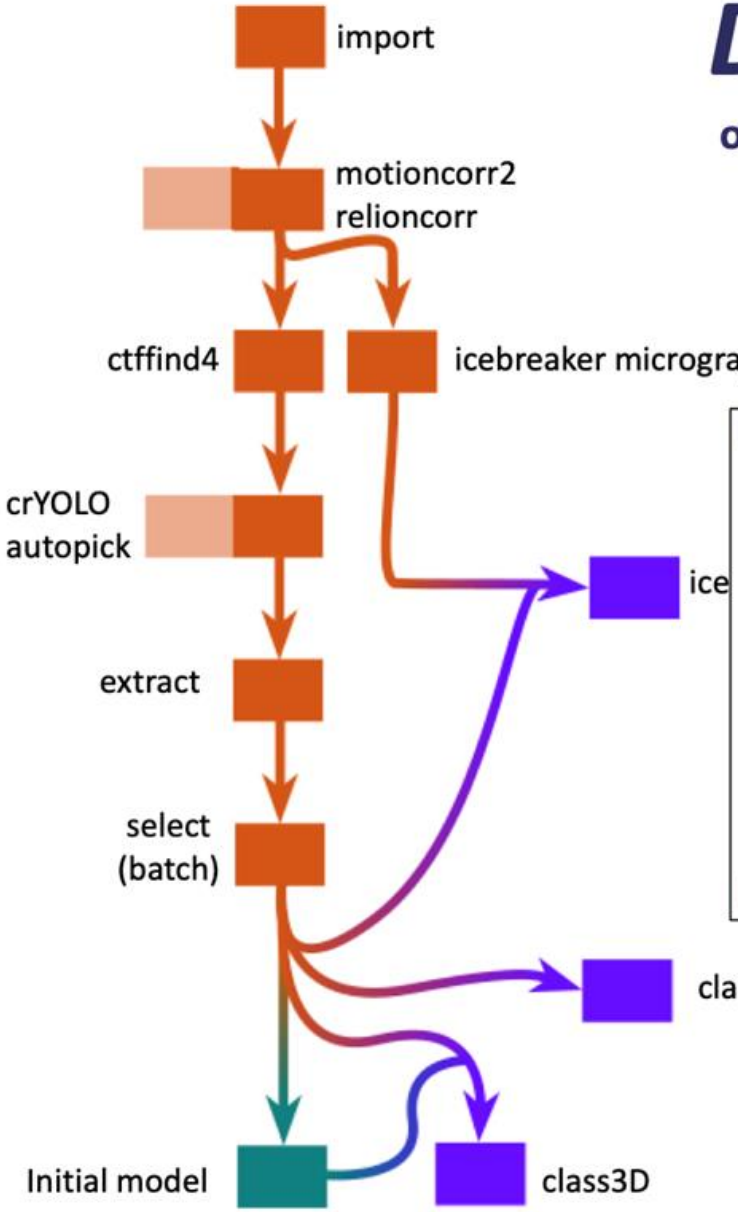
Gerard Bricogne



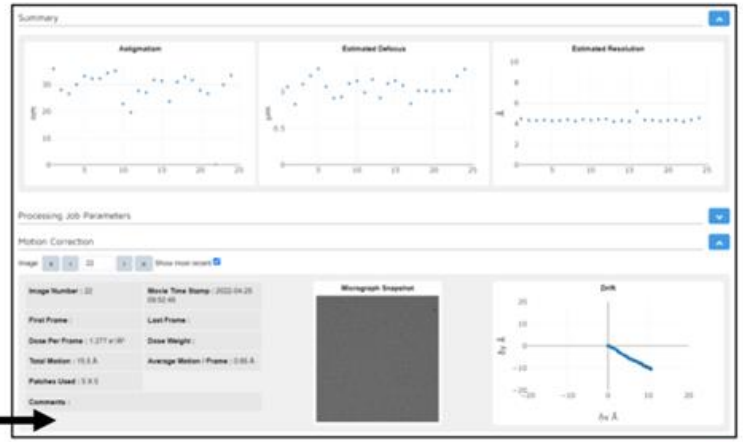


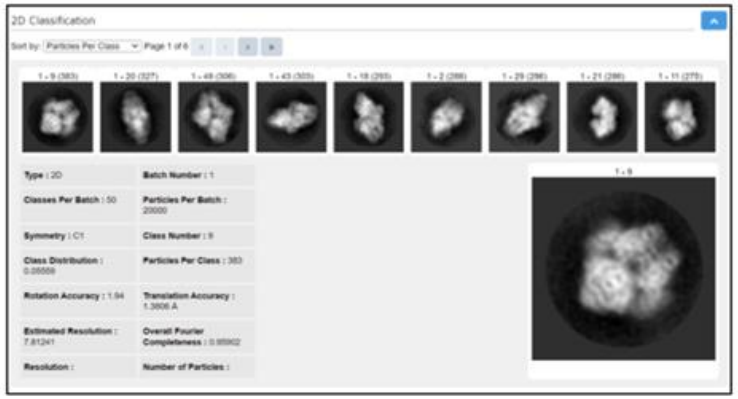
DLS/eBIC Processing

on-the-fly batched preprocessing and evaluation



ISPyB





Dan Hatton
Diamond Light Source

Deposition and archiving

Project archiving

Full Archive
Preserve the project and all associated files

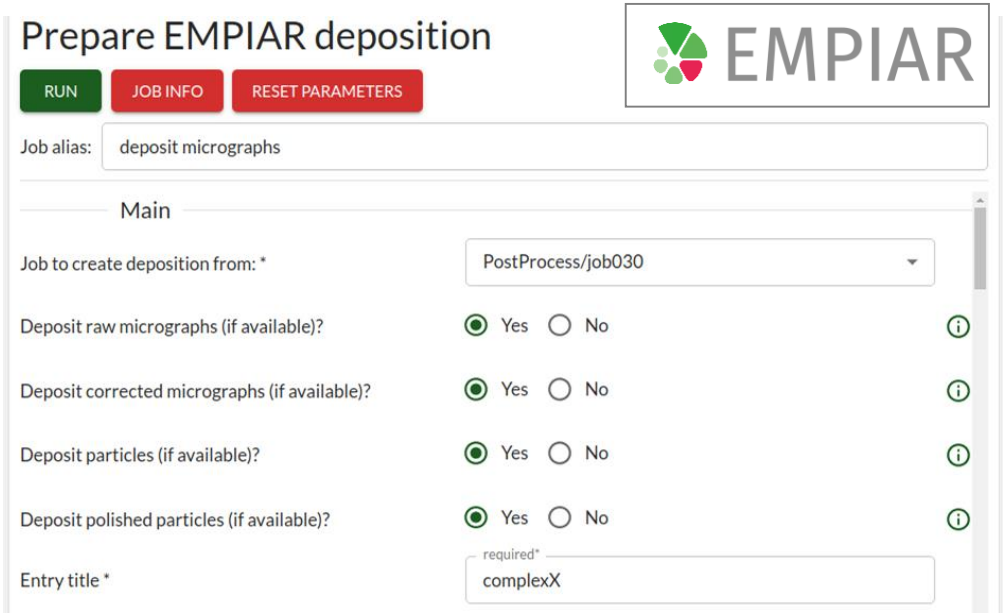
- Project directory structure
- All input files, except raw data
- All results files

Script archive
Preserve and repeat the workflow

- Project directory structure
- All input files, except raw data
- Script to re-run the project

Matt Iadanza

Kyle Morris



Use tracking to generate EMPIAR deposition files for raw data leading to the selected job.

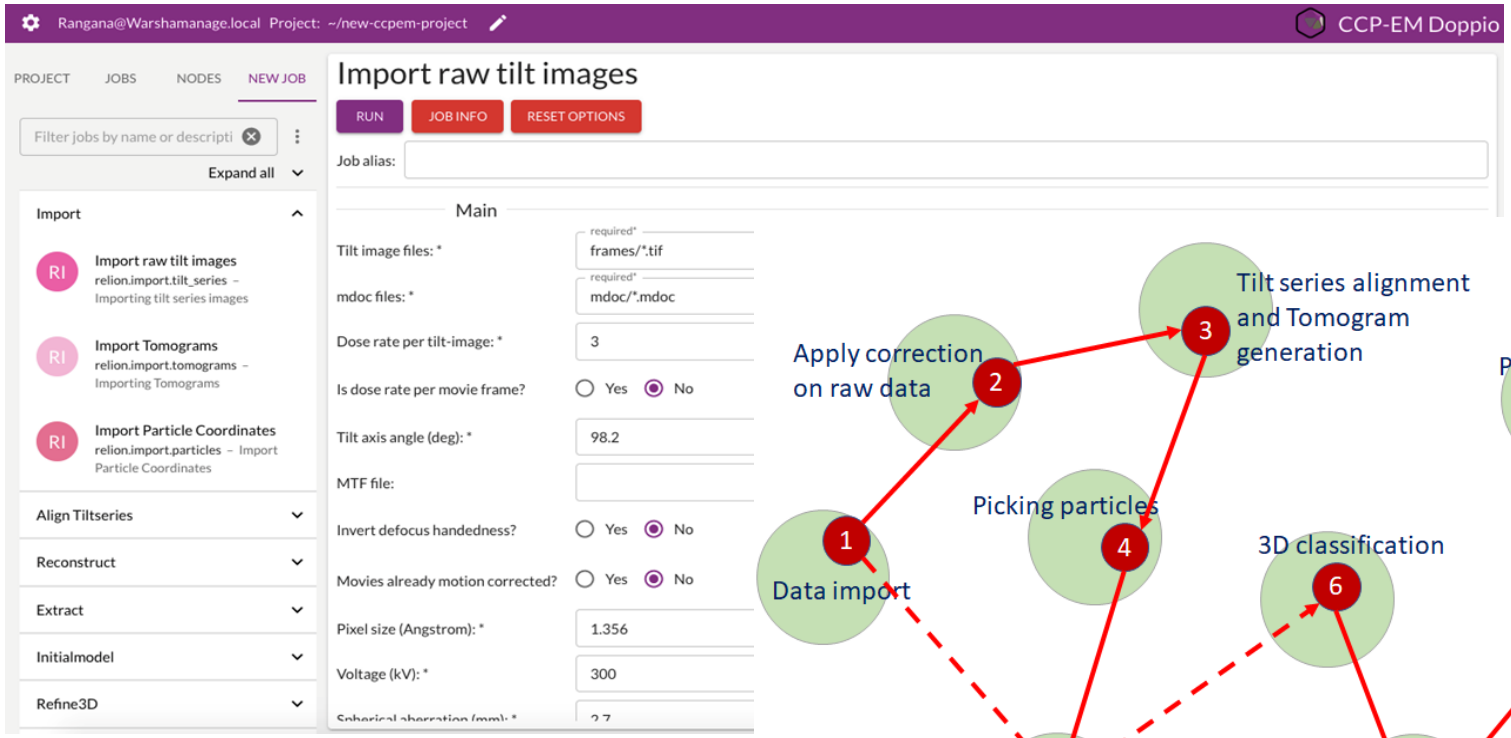
Automates deposition via empiar-depositor at EBI.

OneDep → wwPDB Worldwide PDB Protein Data Bank → EMDB Electron Microscopy Data Bank

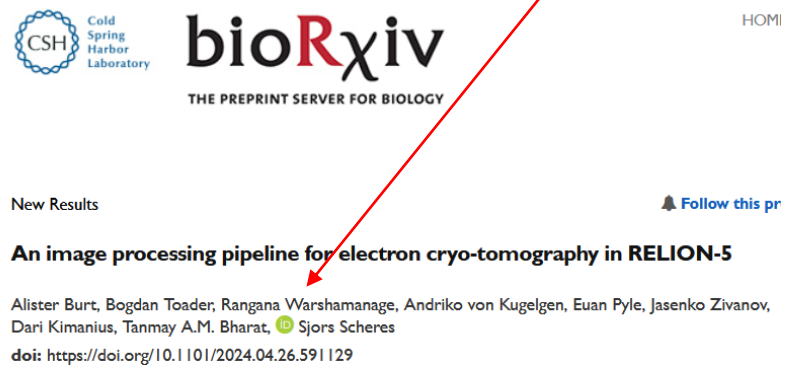
Under development: generating EMDB & PDB deposition files for final map/model submission to OneDep

cryoET pipelines

- First pipeline based on Relion5.
- Extend for other software and options.
- Collaboration with developer community, including eBIC and RFI.



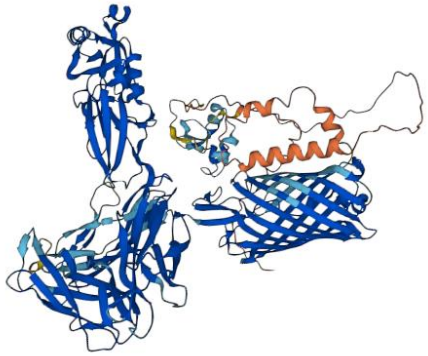
Rangana Warshamanage



Molecular simulation and modelling

Structural models:

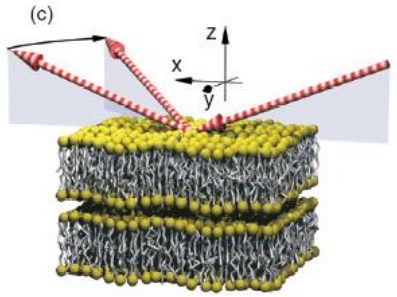
- Phasing in crystallography via Molecular Replacement
- Interpretation of cryoEM maps, SAS, etc
- Often trimming to high confidence regions
- From experimental structures, homology modelling, AlphaFold, etc
- Genome-wide modelling

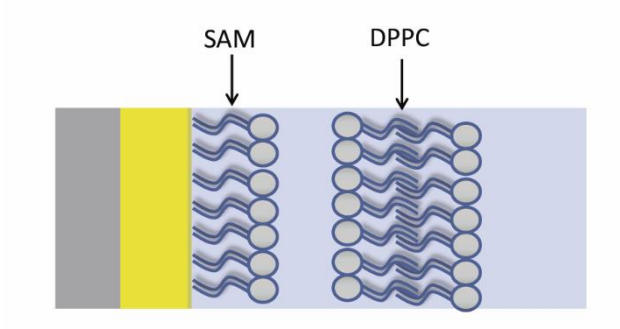


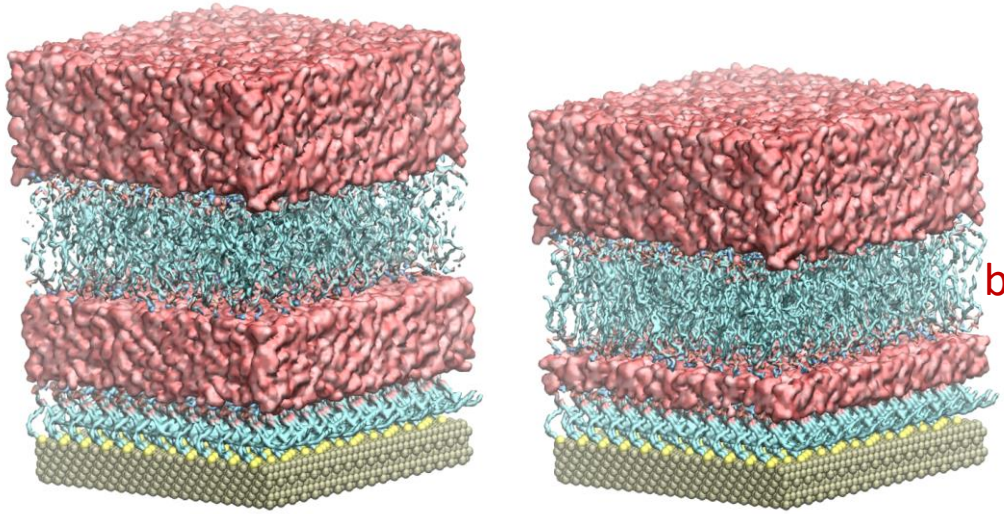
Molecular simulation:

- Previously a largely separate discipline
- Now used for generating conformations, understanding dynamics
- *In silico* experiments

Neutron reflectometry



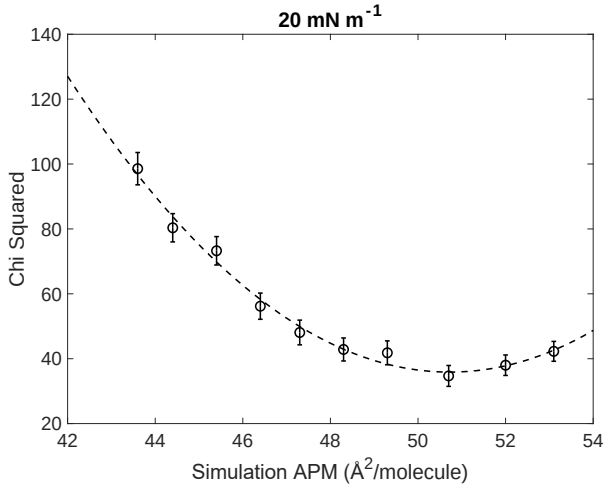




- NR considers z-profile of layered structures.
- Traditionally modelled with set of continuum layers, but more accurate models from MD (prior knowledge of chemistry).

(a) lipid bilayers, (b) monolayers at air-water interface, (c) self-assembled monolayers.

- Fit to experimental data (Rascal/RAT) very sensitive to choice of simulation cell.



The Analysis of Neutron Reflectivity from Langmuir Monolayers of Lipids Using Molecular Dynamics Simulations: The Role of Lipid Area.

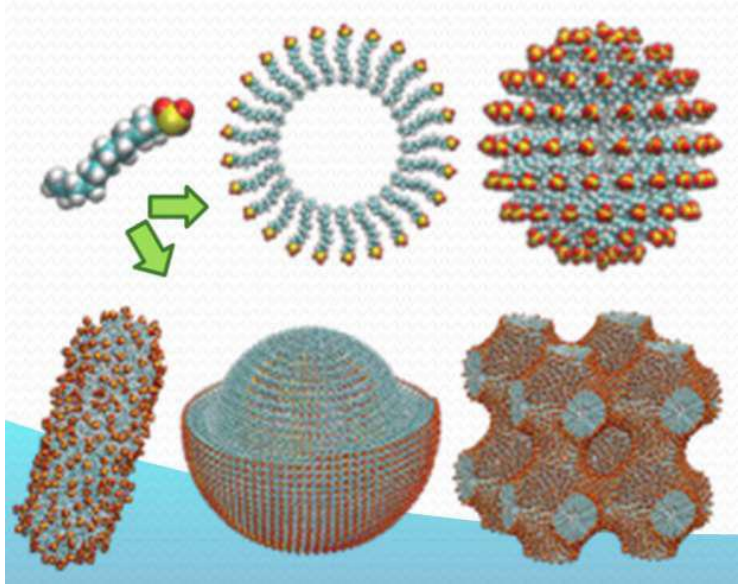
Arwel V. Hughes[a], Valeria Losasso[b] and Martyn Winn[c]. (submitted)

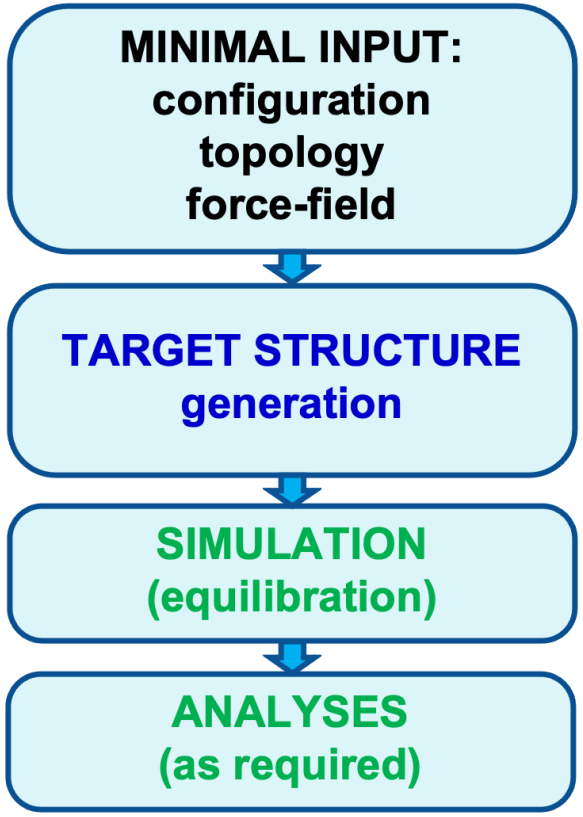
Valeria Losasso

Arwel Hughes

Shapespyer

Toolkit and workflows for molecular simulations of nanostructures.
Developed for soft matter SANS.





1. Generate molecular aggregates in various geometries (layered structures being added)
2. Setup and run MD in Gromacs (NAMD being added)
3. Cluster analyses, radii of gyration, hydration layer, cavity occupation

Andrey Brukhno

James Doutch

Tim Snow

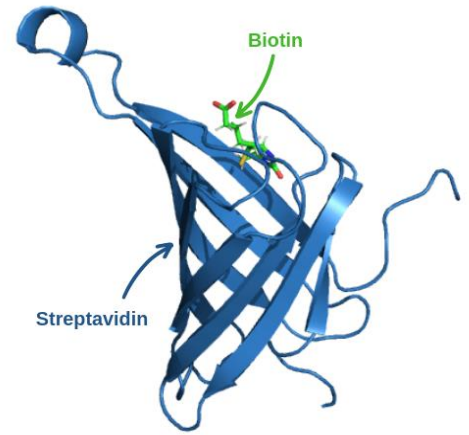
https://www.scd.stfc.ac.uk/Pages/Shapespyer.aspx

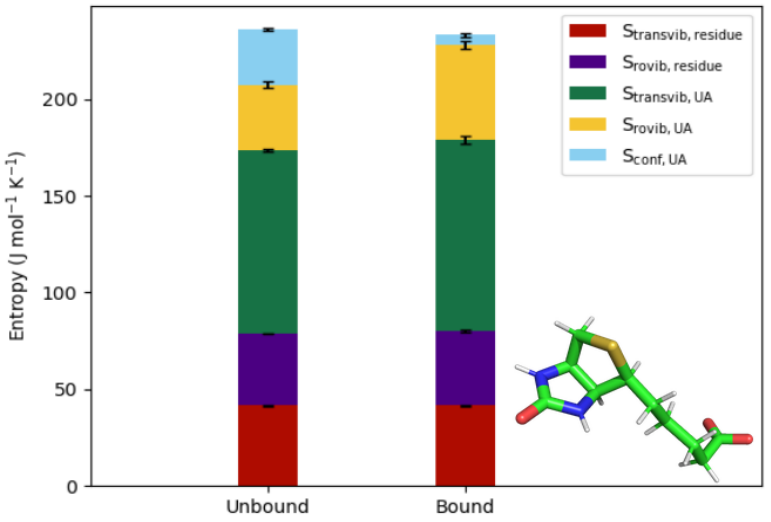
Binding Entropy of Protein-Ligand Systems

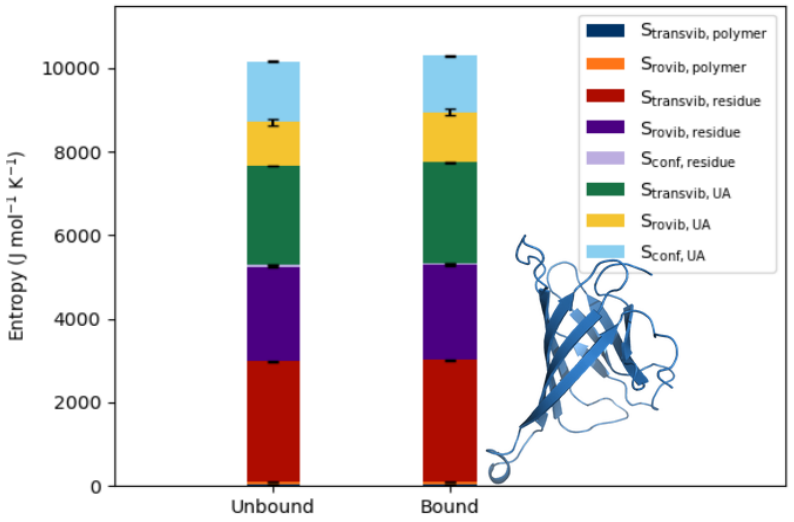
- Molecular dynamics (MD) simulations of the streptavidin-biotin system were computed on the JADE2 HPC using GROMACS
- **CodeEntropy** - a python package for computing the entropy of macromolecular systems from forces sampled from MD simulation trajectories using multiscale cell correlation (MCC) [1]
- **Quasielastic neutron scattering (QENS) data** obtained by Sarter et al. [2] was compared with the calculated binding entropies

{} CCPBioSim

Ioana, Sarah, Jas





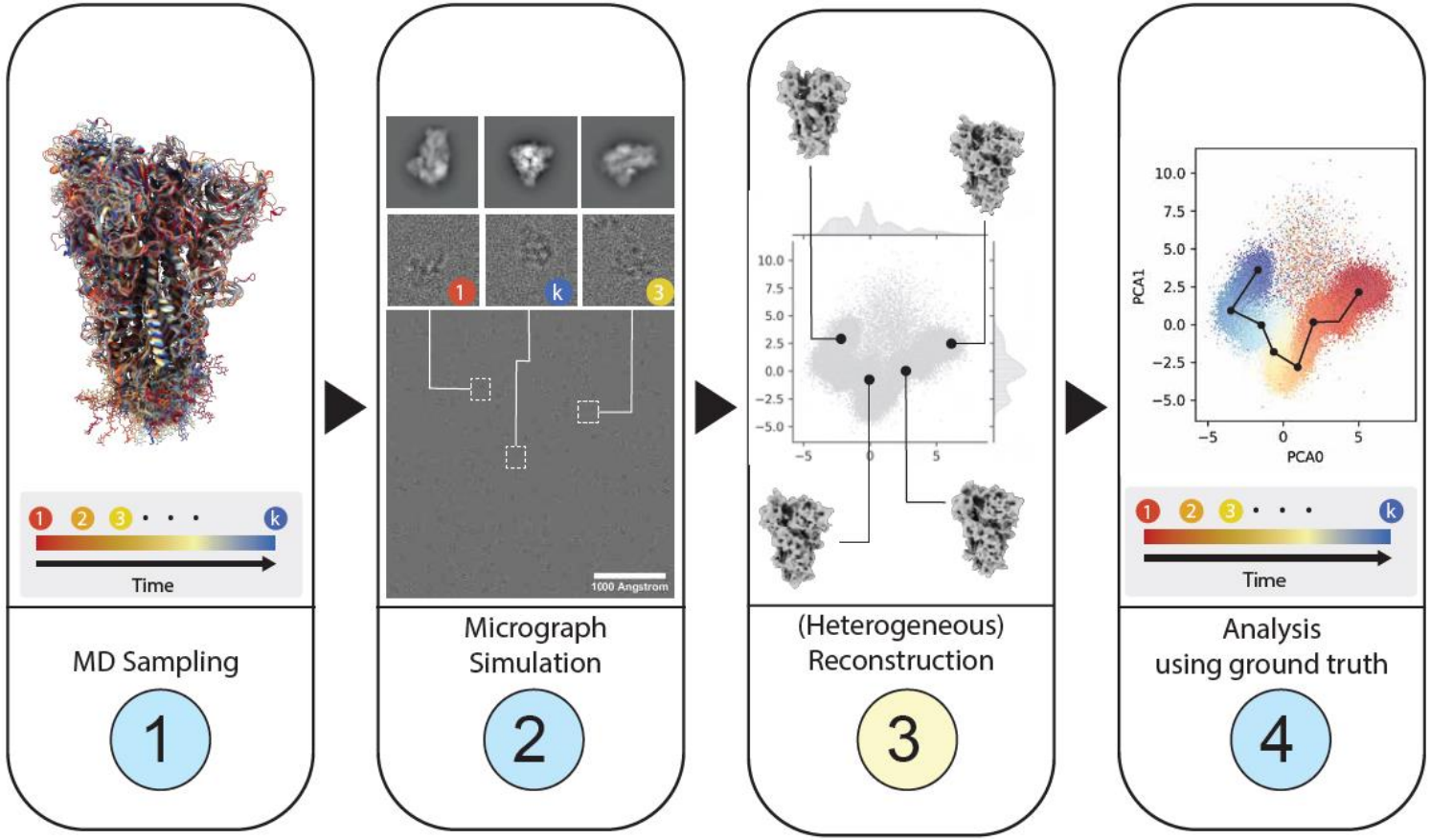


[1] Chakravorty, Arghya; Higham, Jonathan and Henchman, Richard H. (2020). *J. Chem. Inf. Model.*, **60**, 5540−5551.

[2] Sarter, Mona; Niether, Doreen; Koenig, Bernd. W; Lohstroh, Wiebke; Zamponi, Michaela; Jalarvo, Niina H.; Wiegand, Simone; Fitter, Jörg and Stadler, Andreas M. (2020) J. Phys. Chem. B, 2020, **124**, 324–335.

MD and cryoEM

- Molecular dynamics explores conformational states of a protein / complex.
  Ergodicity, simplified sample
- CryoEM can separate conformational states into discrete or continuous classes
  Signal to noise limit



Aim to benchmark heterogeneous reconstruction algorithms (e.g. CryoDRGN and 3DFlex) using synthetic particle sets from MD where ground truth known.

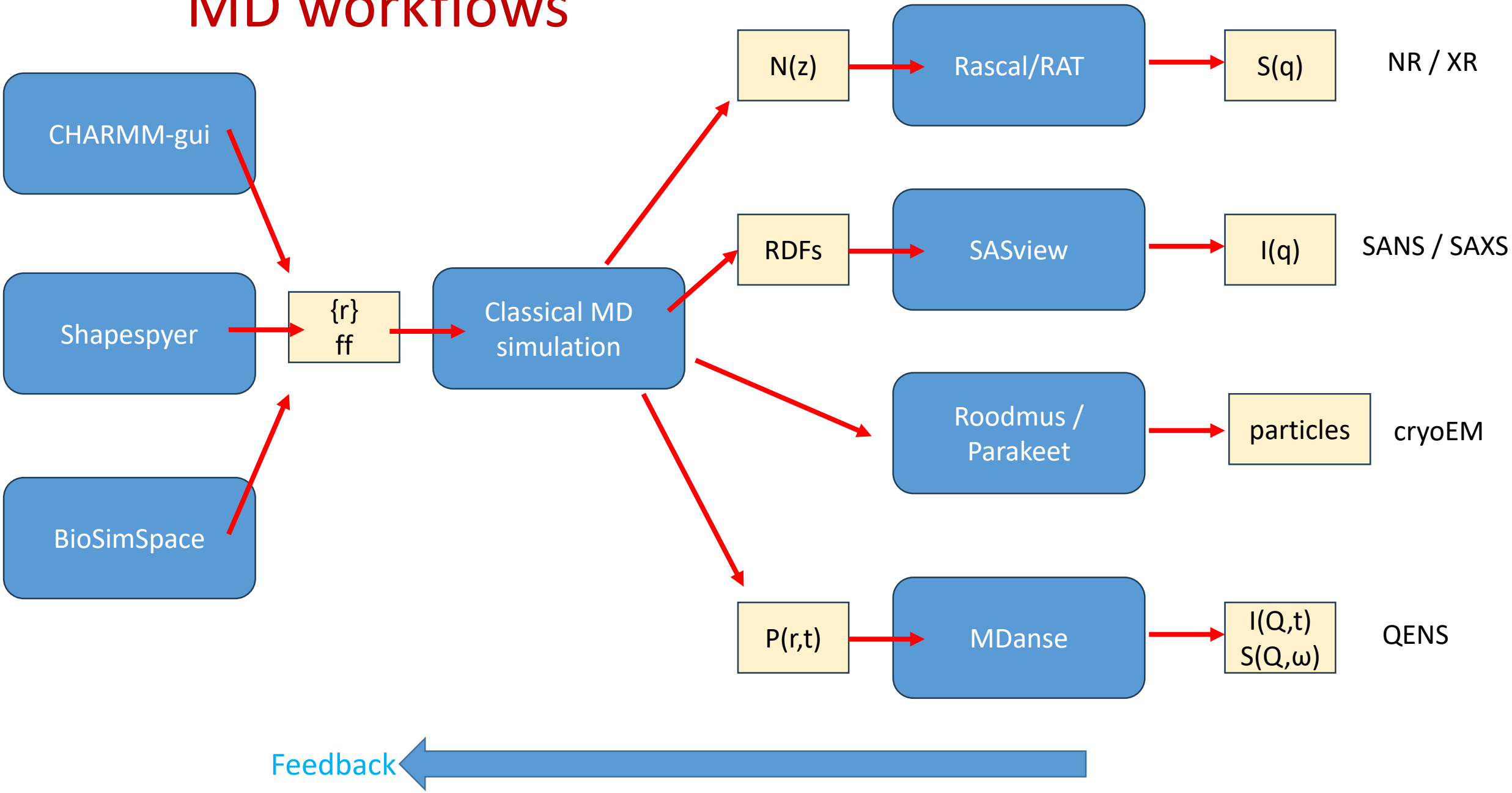
Roodmus: A toolkit for benchmarking heterogeneous electron cryo-microscopy reconstructions

Maarten Joosten,[a†] Joel Greer,[b†] James Parkhurst,[c,d] Tom Burnley[b*] and Arjen J. Jakobi [a*]

Delft, SCD, RFI

https://www.biorxiv.org/content/10.1101/2024.04.29.590932v1

MD workflows



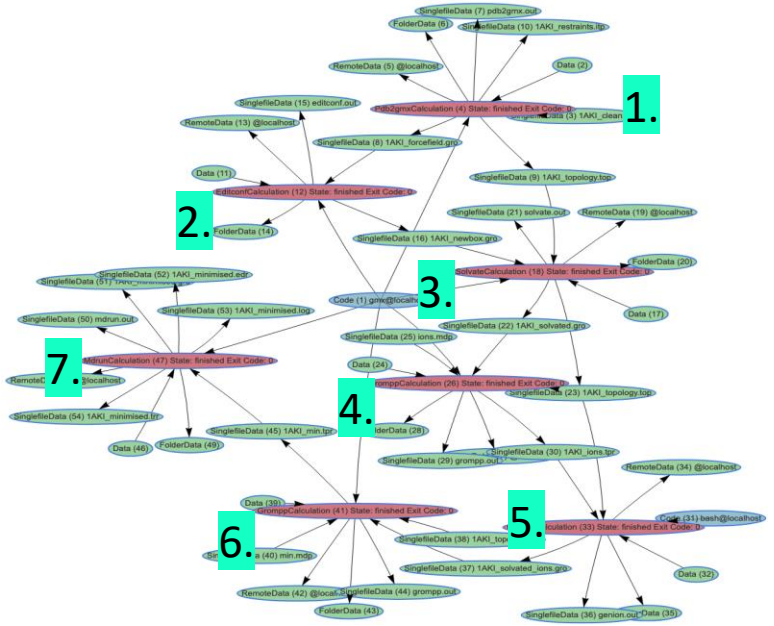
Data / metadata management

We do not handle data storage or data transfer.
Better experts in SCD!

PSDI
PHYSICAL SCIENCES
DATA INFRASTRUCTURE

Our workflows:

- Handle file format conversions
- Store and track metadata
- Connect to external data stores
- Deposit in international repositories



https://aiida-gromacs.readthedocs.io

Good data management a pre-requisite for ….

Machine learning

- Early adopters of ML for interpretation of scientific data
- Looking to adapt established models to our datasets / processing
- Have worked closely with SciML group and Alan Lowe's group at ATI

> 50% of any project is preparing / organising the training data …

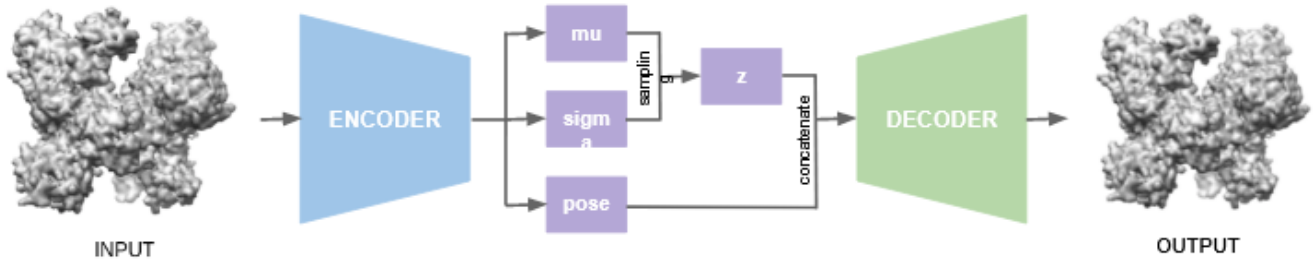
- Many methods coded in the Macromolecular Machine Learning Toolbox https://gitlab.com/ccpem/ml-protein-toolbox
- Besides our own tools, incorporate 3rd party tools e.g. AlphaFold, Blush (Relion 5)
- Working on integrating CryoDRGN

Affinity-VAE

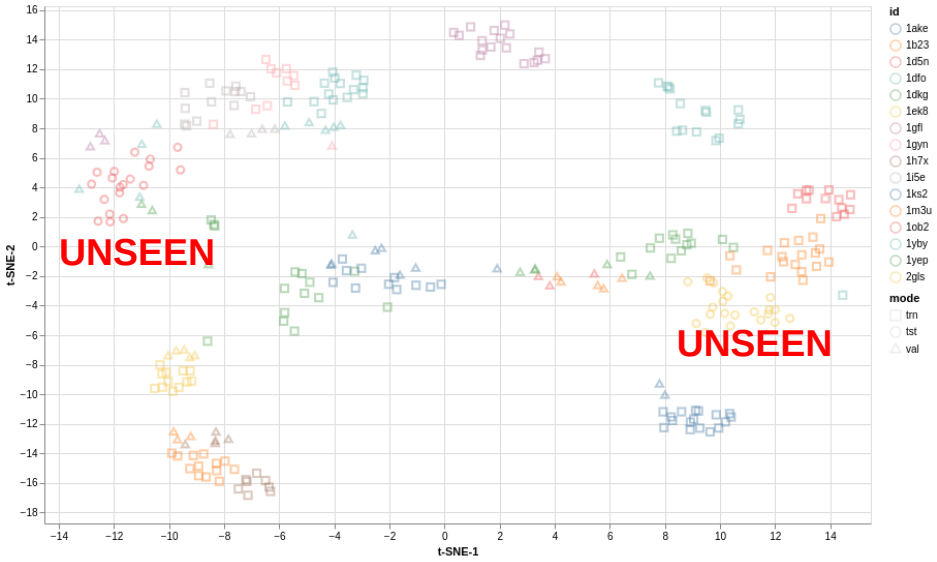
Jola

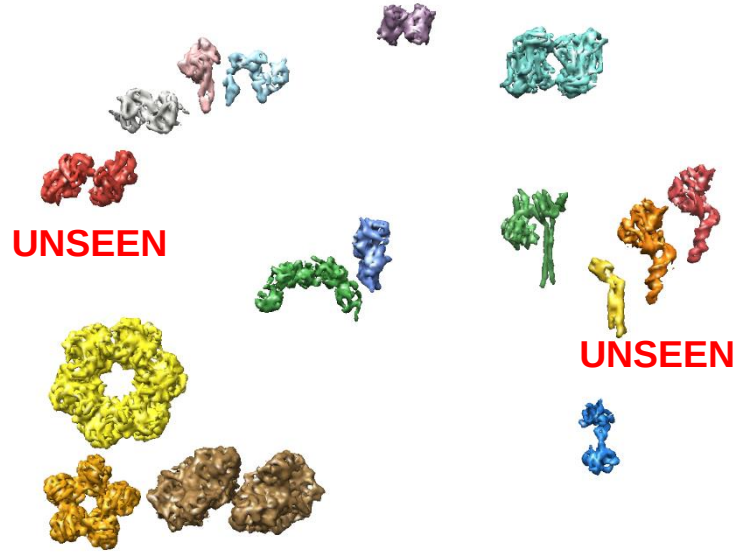
Identification of molecules in cryoET

- beta-VAE optimisation function plus affinity regulariser
- Based on similarity of two 3D objects
- Helps to disentangle of inter-class variation from intra-class (rot, trans)



CCP-EM, ATI, RFI, SciML



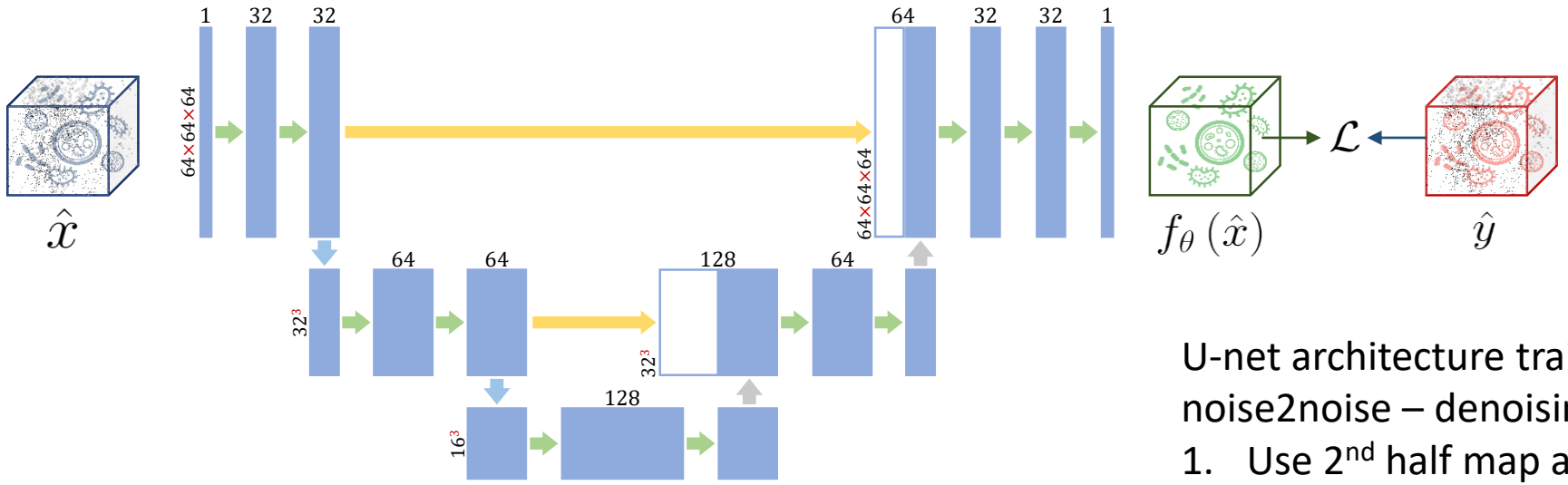


https://arxiv.org/abs/2209.04517

Denoising of electron tomograms


Jola

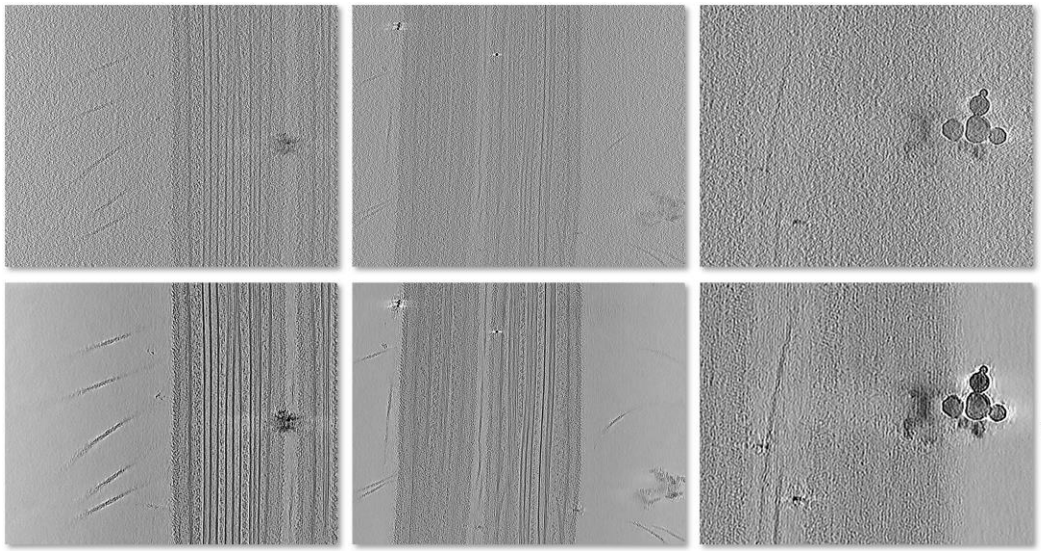

Ding



U-net architecture trained with 64 x 64 x 64 patches.
noise2noise – denoising without clean data:

1. Use 2nd half map as noisy target
2. Randomly remove high frequency components to give corrupted target

Pros and cons in efficacy, speed and data requirements.

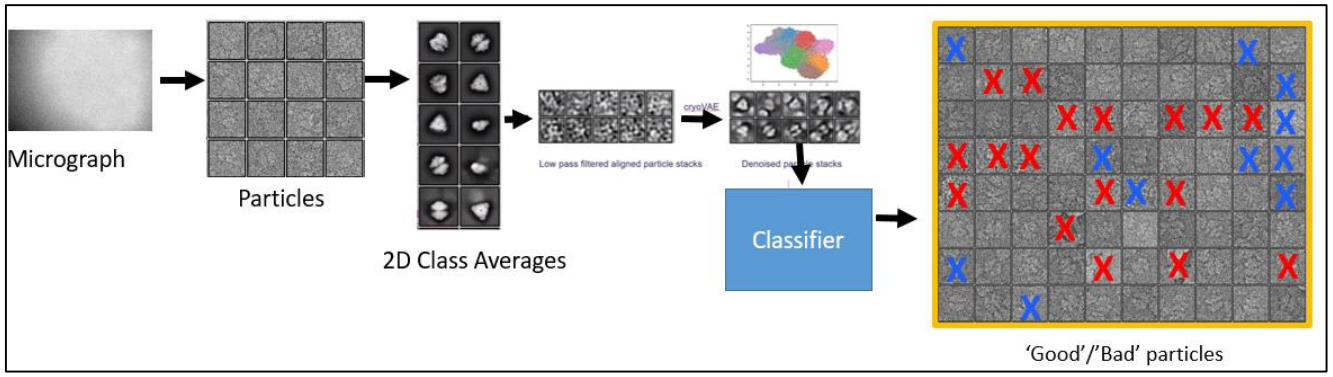


Feeds into tomography projects with eBIC and RFI.

cryoDANN: automated evaluation of particles

Sony

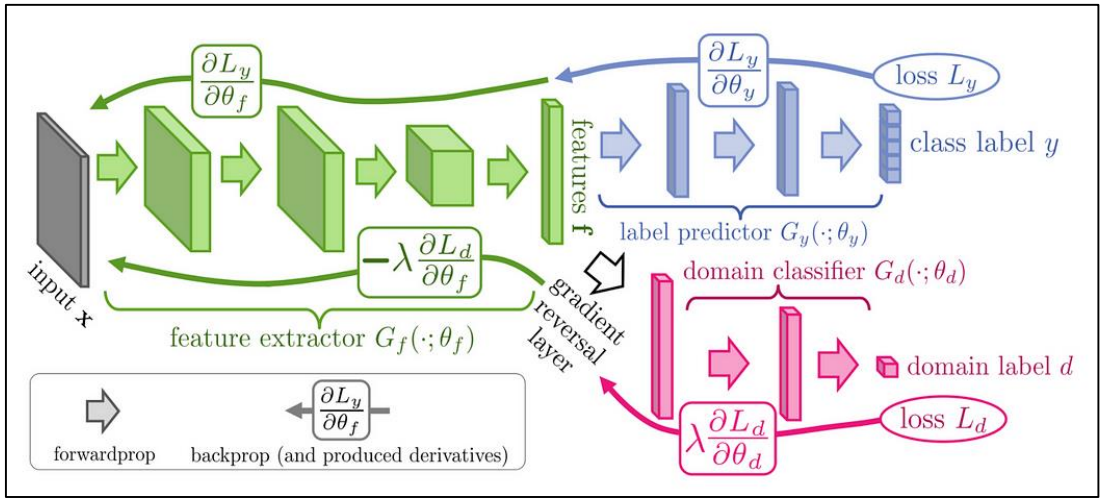
Objective: Annotating particle images



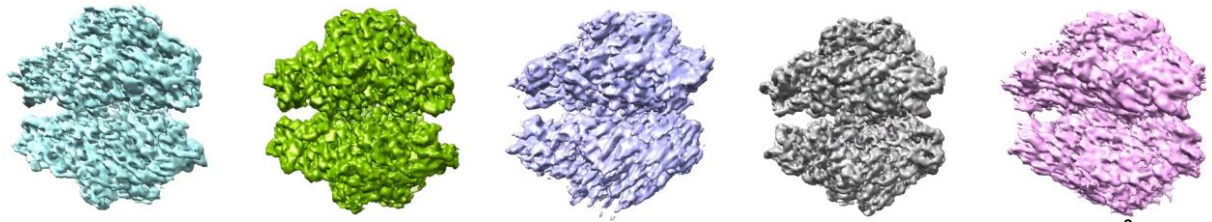
- Joint project with Diamond/eBIC, SciML
- ML based method for an automated evaluation of good/bad particle selection
- Tested the classifier on EMPIAR datasets
- Installed and being tested at eBIC (Yuriy Chaban, Dan Hatten)

Aim: Automate the particle selection step in data processing step

CryoDANN:



Manual ClassRanker Random CryoDann All



3.84Å,34k 3.84Å,30k 4.09Å,30k 3.79Å,30k 3.75Å,50k

Results: EMPIAR-10547

Computed tomography

- CCPi mainly materials, but formally in the Comp Bio theme and not covered elsewhere
- CCPSyneRBI is medical imaging
- On-going collaborations with ISIS, Diamond and EPAC

- Most software developments centred on the Core Imaging Library (CIL)

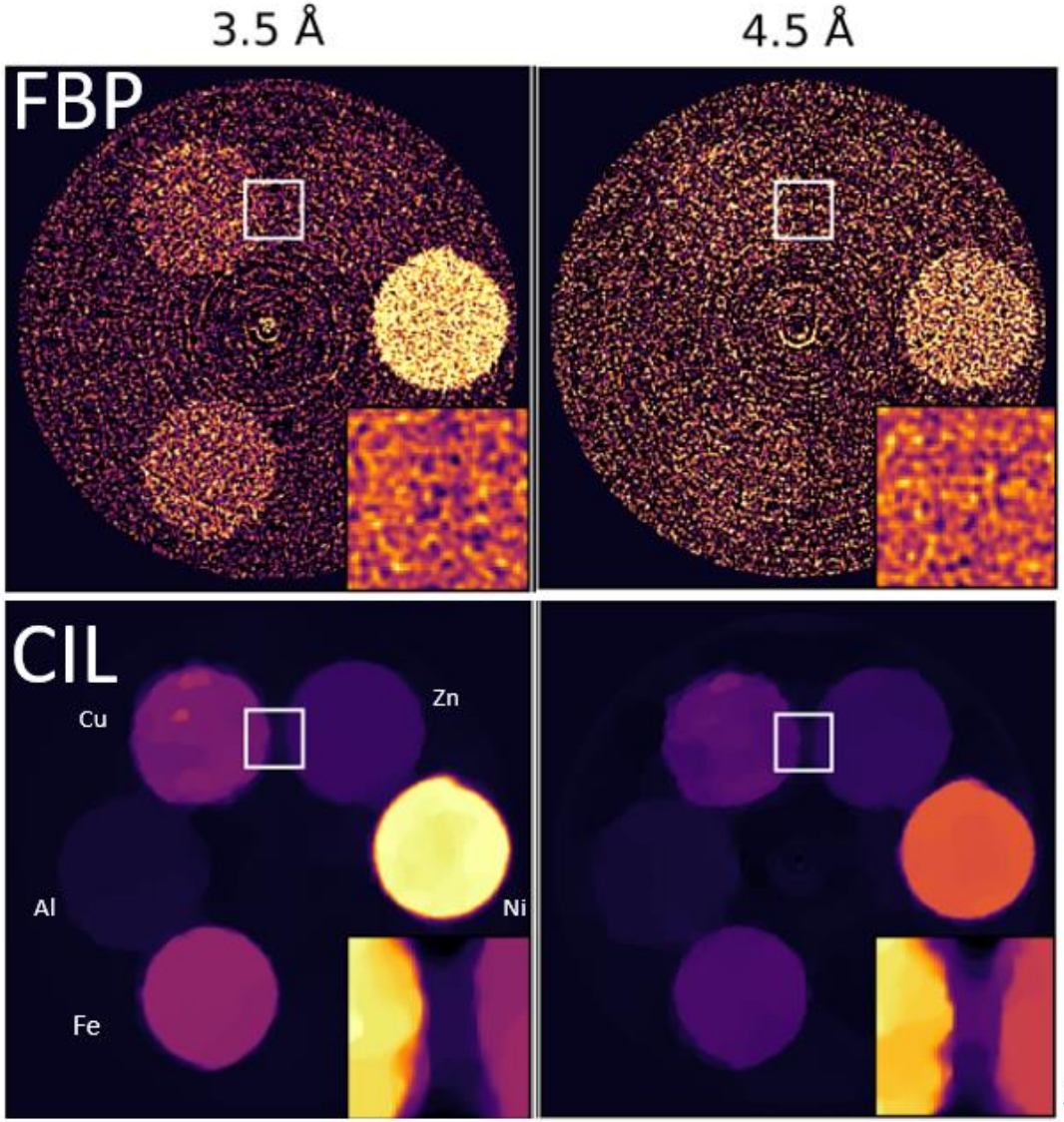
https://ccpi.ac.uk/cil/

- Underpinning for inverse problems

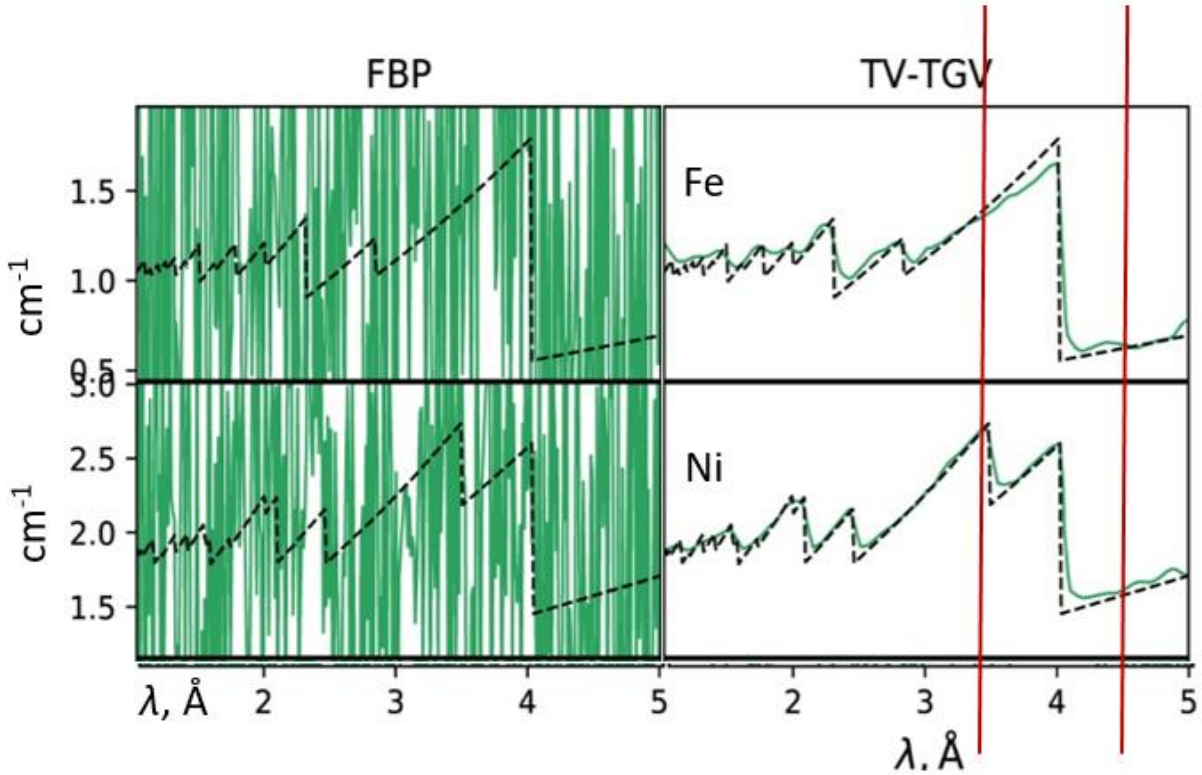
https://ccpi.ac.uk/

https://www.ccpsynerbi.ac.uk/

Energy-resolved neutron CT



- Data from ISIS/IMAT
- Proposed spatio-spectral TV-TGV regularization
- Enables clear identification of Bragg edges in 3D



Ametova et al. 2021: *Crystalline phase discriminating neutron tomography using advanced reconstruction methods*, J. Physics D, https://doi.org/10.1088/1361-6463/ac02f9

Digital Volume Correlation

Alignment / interpretation of 4D data.
Collaboration with Brian Bay (Oregon).
Porting, parallelisation, GUI.

iDVC

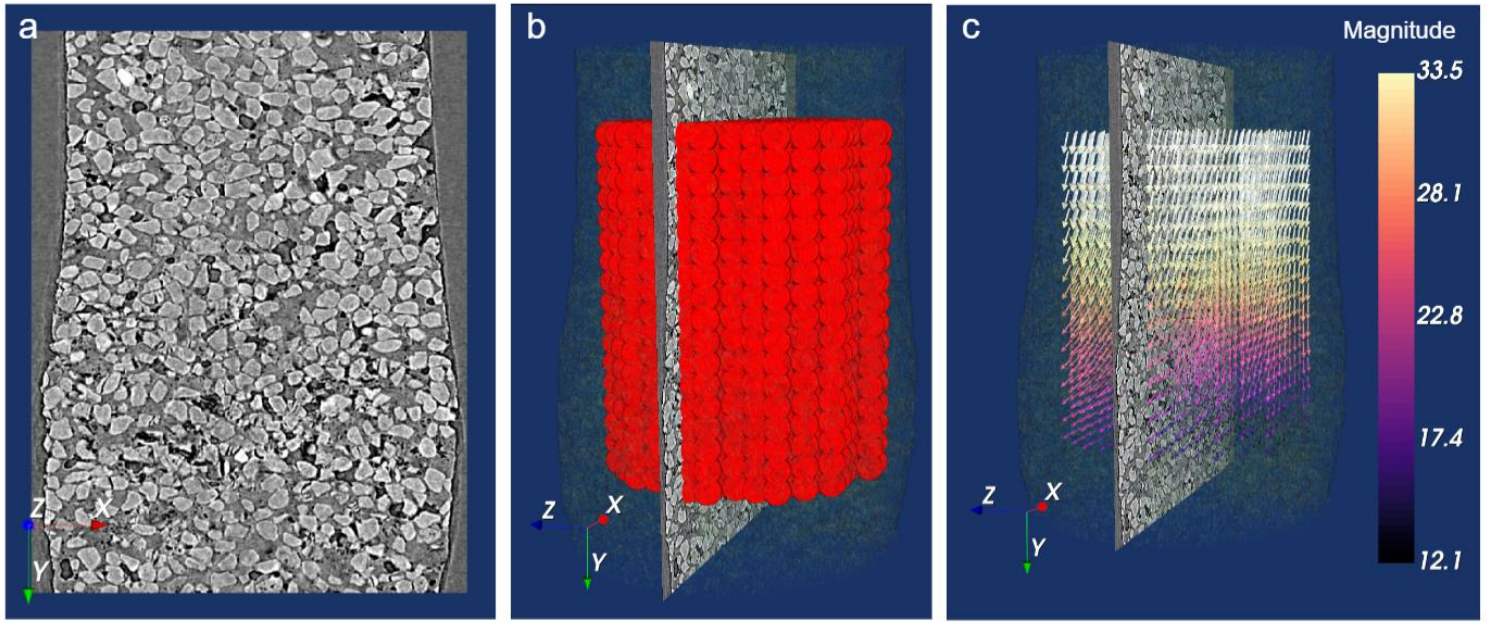


Figure 1: CT reconstruction of magma data. a) slice view, b) region of interest with point cloud and spherical subregions, c) displacement vector display in unit of pixels.

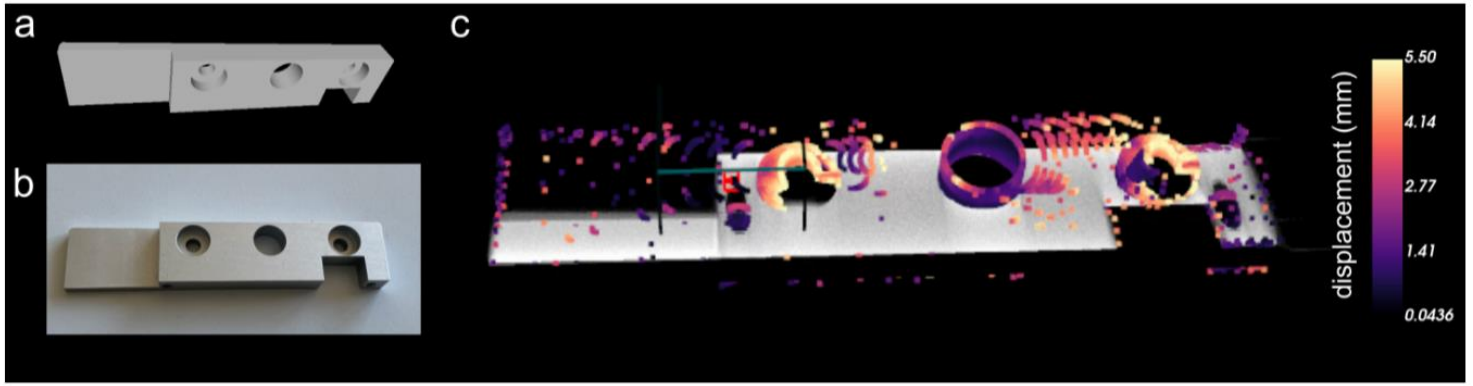
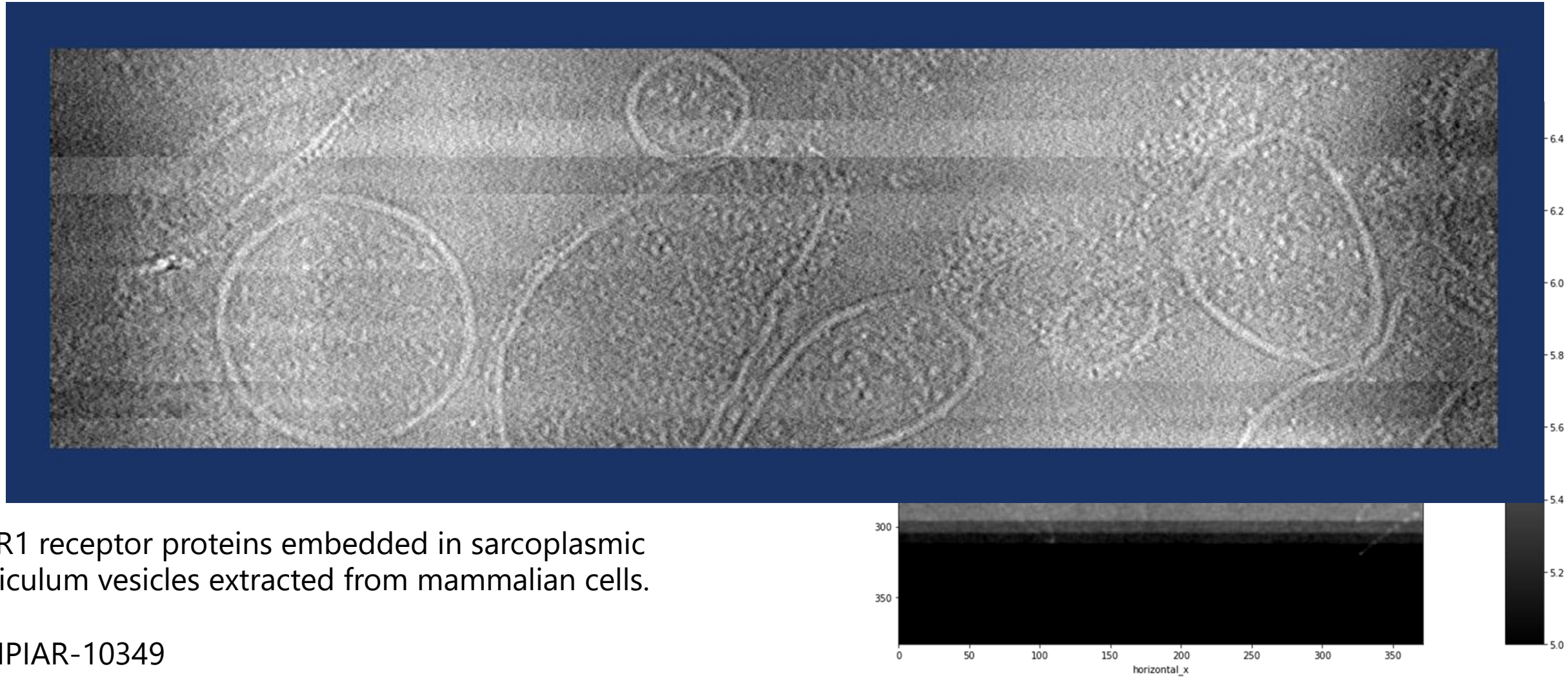


Figure 2: Mechanical component of an in-situ tensile testing machine. a) Original CAD model, b) component manufactured from the CAD model, c) displacement field between the 3D volumes of the experimental CT scan and of the virtual CT scan of the CAD model.

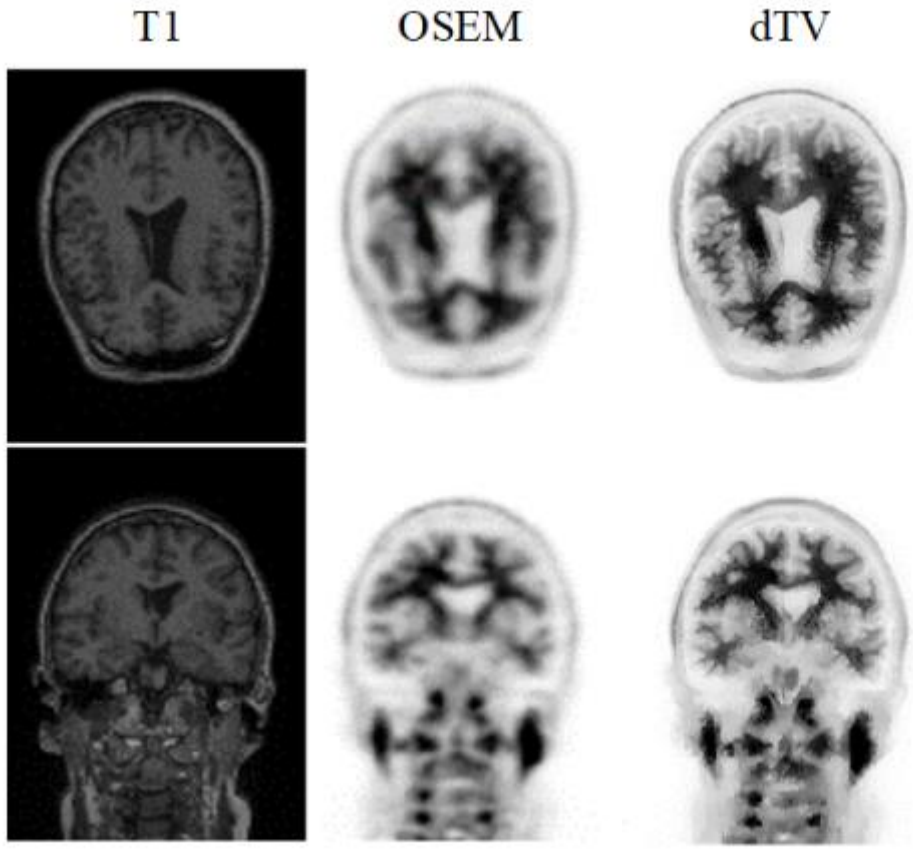
Reconstructions in CIL for cryoET



RyR1 receptor proteins embedded in sarcoplasmic reticulum vesicles extracted from mammalian cells.

EMPIAR-10349

Anatomically guided PET super resolution



https://www.medrxiv.org/content/10.1101/2023.04.23.23289004v1

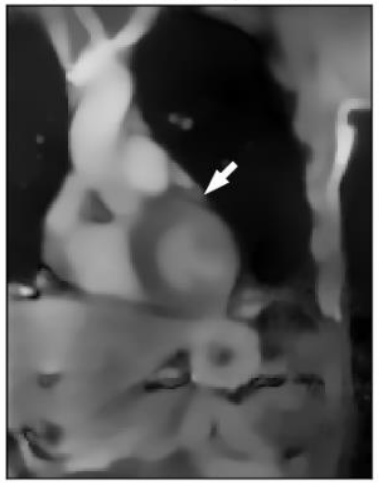
Super-resolution PET reconstructions, guided by MRI.
dTV-regularised, two stages.

Cardio-respiratory MRI motion compensated image reconstruction.

In preparation and https://royalsocietypublishing.org/doi/10.1098/rsta.2020.0208

Synergistic Image Reconstruction Framework (uses CIL)

SIRF



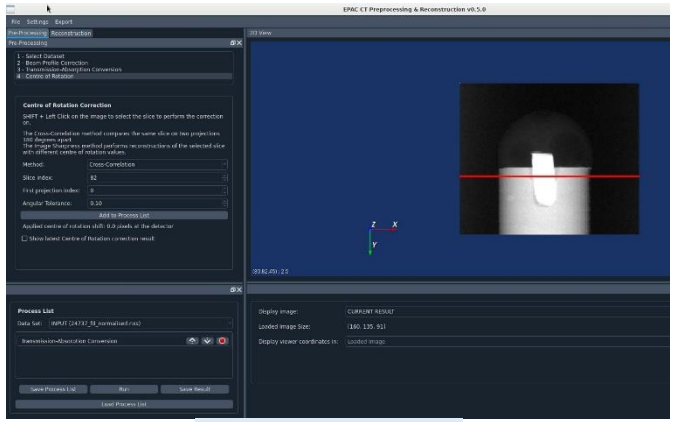
EPAC

GUI + Pipeline for CT recon with CIL

CORE IMAGING LIBRARY

- Handles Nexus & lab-based XCT data files
- Pre Processing
- Reconstruction

Danica

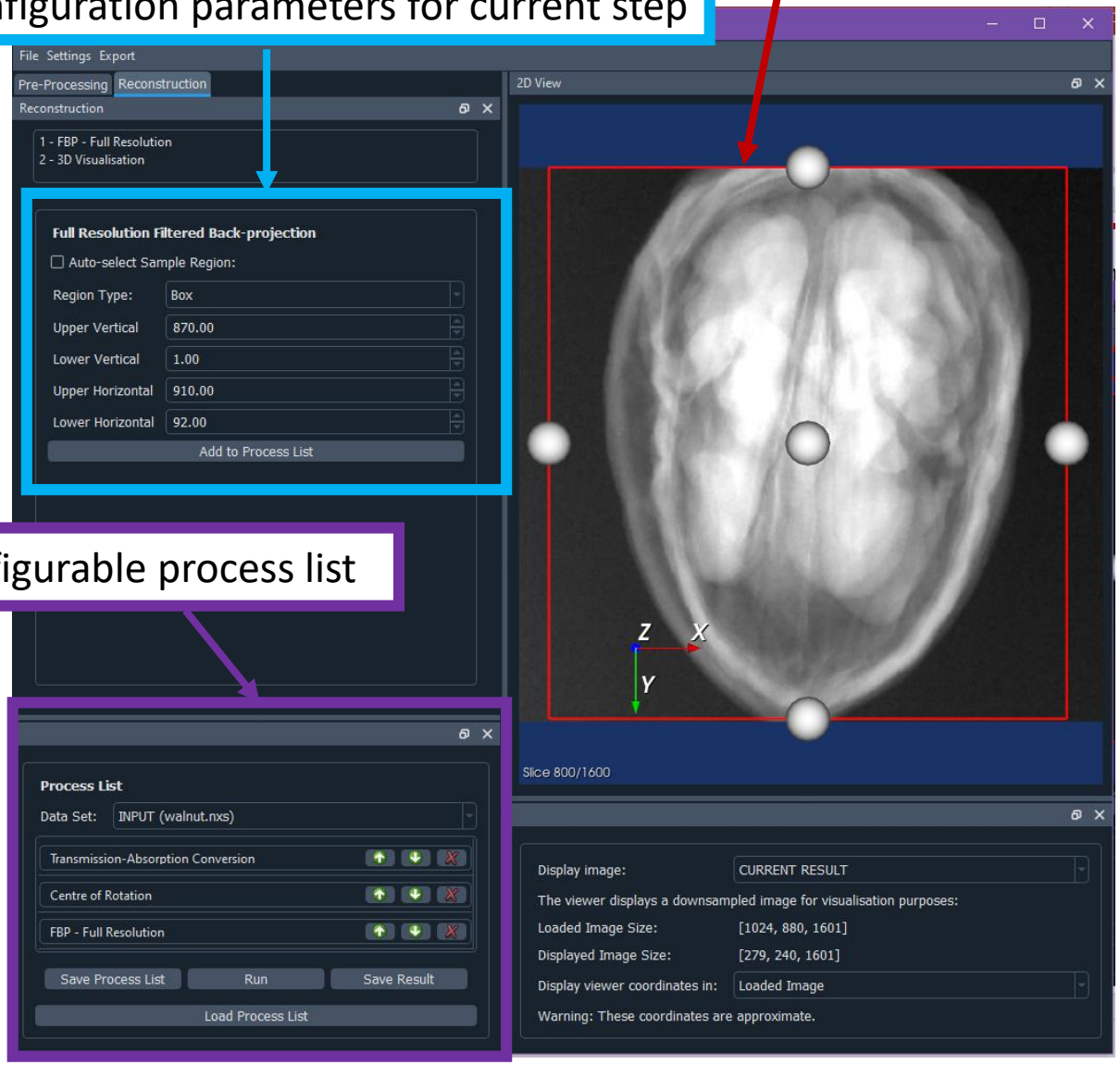


CIL-GUI

Visualisation + interaction with data at each step

Configuration parameters for current step

Configurable process list



https://www.clf.stfc.ac.uk/Pages/EPAC-X-Ray-radiography-and-X-ray-Computed-Tomography.aspx

Digital Twinning

A widely used and inconsistently used term …

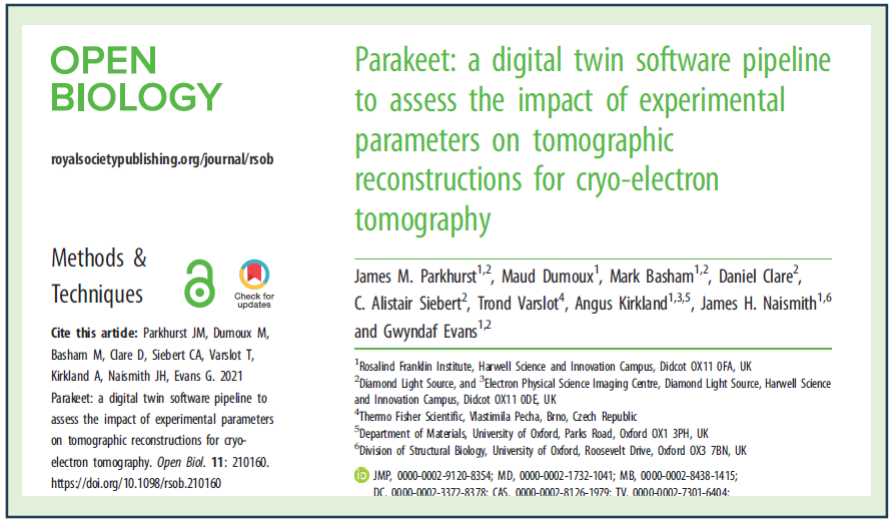
Here we consider simulated experiments:

- Generate large amounts of synthetic data
- Full control over sample and instrument parameters, allowing *in silico* trials

Doesn’t replace real data.
But useful for development and understanding.

Parakeet: Digital Twin for cryoEM/cryoET

James



OPEN BIOLOGY

royalsocietypublishing.org/journal/rsob

Methods & Techniques

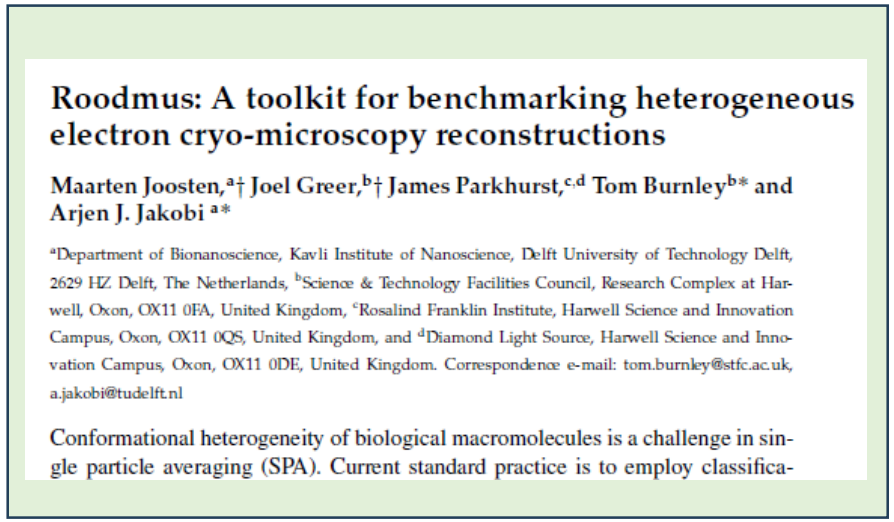
Cite this article: Parkhurst JM, Dumoux M, Basham M, Clare D, Siebert CA, Varslot T, Kirkland A, Naismith JH, Evans G. 2021 Parakeet: a digital twin software pipeline to assess the impact of experimental parameters on tomographic reconstructions for cryo-electron tomography. *Open Biol.* **11**: 210160. https://doi.org/10.1098/rsob.210160

Parakeet: a digital twin software pipeline to assess the impact of experimental parameters on tomographic reconstructions for cryo-electron tomography

James M. Parkhurst[1,2], Maud Dumoux[1], Mark Basham[1,2], Daniel Clare[2], C. Alistair Siebert[2], Trond Varslot[4], Angus Kirkland[1,3,5], James H. Naismith[1,6] and Gwyndaf Evans[1,2]

[1]Rosalind Franklin Institute, Harwell Science and Innovation Campus, Didcot OX11 0FA, UK
[2]Diamond Light Source, and [3]Electron Physical Science Imaging Centre, Diamond Light Source, Harwell Science and Innovation Campus, Didcot OX11 0DE, UK
[4]Thermo Fisher Scientific, Vlastimila Pecha, Brno, Czech Republic
[5]Department of Materials, University of Oxford, Parks Road, Oxford OX1 3PH, UK
[6]Division of Structural Biology, University of Oxford, Roosevelt Drive, Oxford OX3 7BN, UK

JMP, 0000-0002-9120-8354; MD, 0000-0002-1732-1041; MB, 0000-0002-8438-1415; DC, 0000-0002-3377-8378; CAS, 0000-0002-8176-1979; TV, 0000-0002-7301-6404;



Roodmus: A toolkit for benchmarking heterogeneous electron cryo-microscopy reconstructions

Maarten Joosten,[a]† Joel Greer,[b]† James Parkhurst,[c,d] Tom Burnley[b]* and Arjen J. Jakobi[a]*

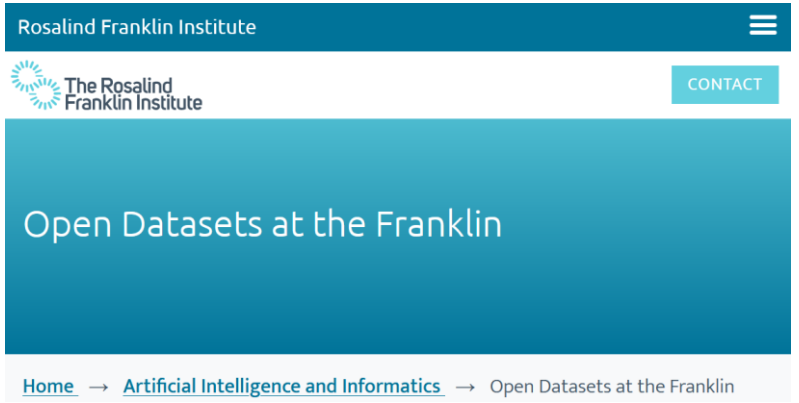
[a]Department of Bionanoscience, Kavli Institute of Nanoscience, Delft University of Technology Delft, 2629 HZ Delft, The Netherlands, [b]Science & Technology Facilities Council, Research Complex at Harwell, Oxon, OX11 0FA, United Kingdom, [c]Rosalind Franklin Institute, Harwell Science and Innovation Campus, Oxon, OX11 0QS, United Kingdom, and [d]Diamond Light Source, Harwell Science and Innovation Campus, Oxon, OX11 0DE, United Kingdom. Correspondence e-mail: tom.burnley@stfc.ac.uk, a.jakobi@tudelft.nl

Conformational heterogeneity of biological macromolecules is a challenge in single particle averaging (SPA). Current standard practice is to employ classifica-

Parakeet: initiative of the RFI.

Digital Twin of electron tomography experiment.
Includes geometry, microscope parameters, acquisition strategy, noise models, etc.

As part of **Roodmus**, we adapted Parakeet for single particle experiments (CCP-EM, TU Delft).
We generated simulated cryoEM datasets starting with atomic models generated in a Molecular Dynamics simulation.
Important for development (know ground truth).



RFI is hosting synthetic cryoEM datasets generated by RFI, CCP-EM and ATI.
Available as Globus collection.

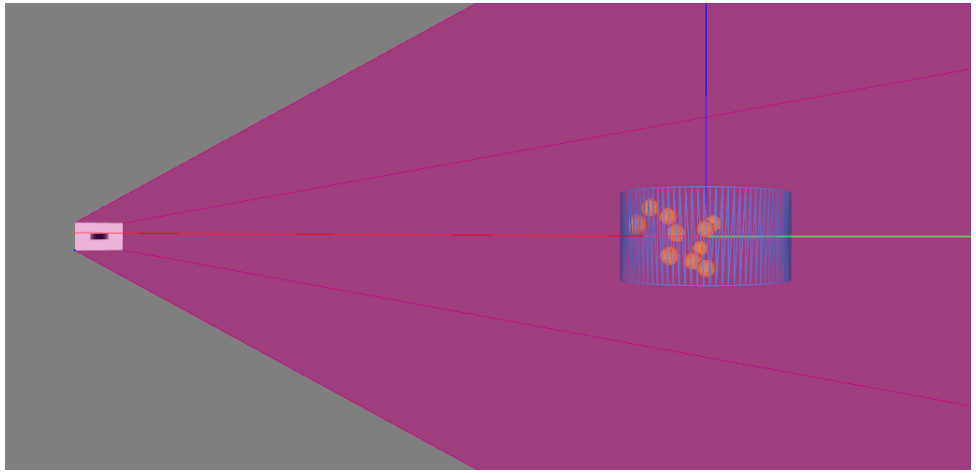
https://www.rfi.ac.uk/projects/open-datasets/

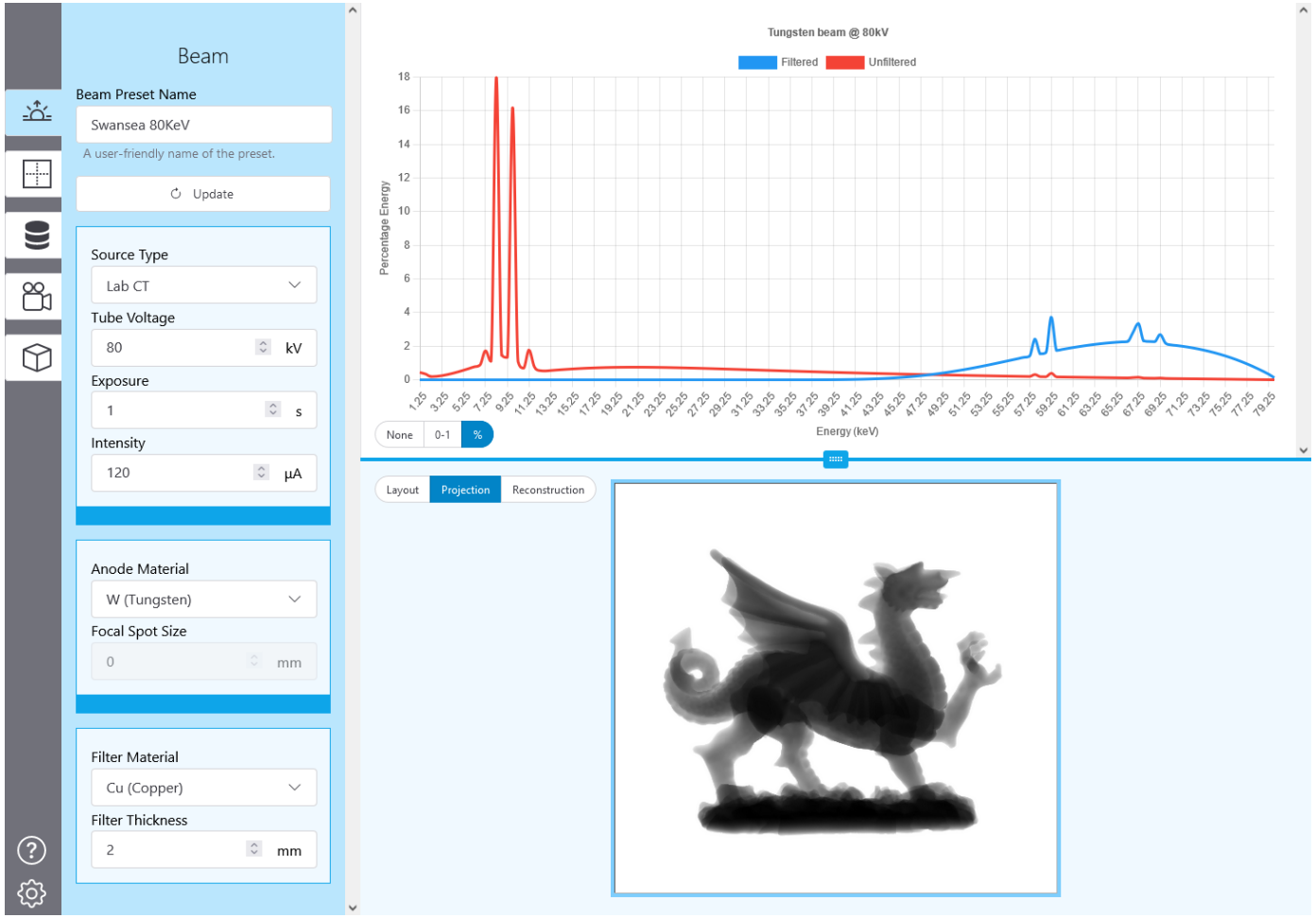
Computed Tomography Digital Twin

Franck

Developing a Digital Twin for CT beamlines.

Based on gVirtualXRay, a C++ library implemented on GPUs to simulate X-ray imaging.
Uses Beer-Lambert law to compute the absorption of light (i.e. photons) by 3D objects (here polygon meshes).

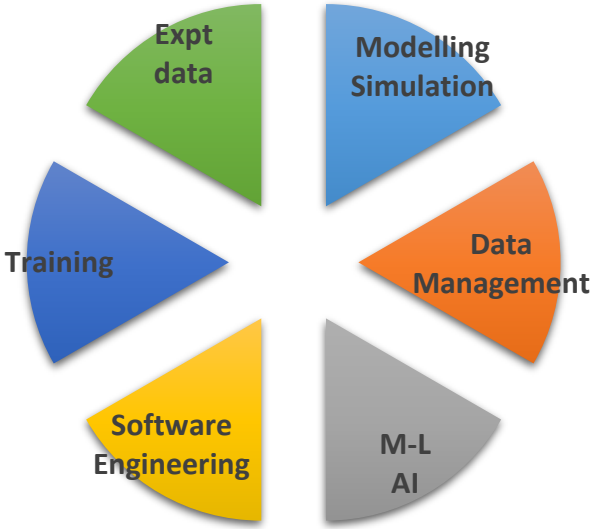




https://gvirtualxray.fpvidal.net/

Future engagement

- We want to build on our scientific strengths
  - ➢ but expand into new areas
- Software engineering skills
  - ➢ but we are not generic RSEs
- Increased use of machine learning
  - ➢ optimise data processing (efficiency)
- Automation via software pipelines
  - ➢ but bespoke application work for end users useful for engagement and experience
- Interested in studentships and placements



Software across facilities
MD simulation for interpretation
Imaging across lengthscales
Metadata capture and archiving

Thanks and contacts

Please come talk to us!

Edoardo Pasca

Eugene Krissinel

James Gebbie-Rayet

Martyn Winn

Tom Burnley

Science and
Technology
Facilities Council

Questions?