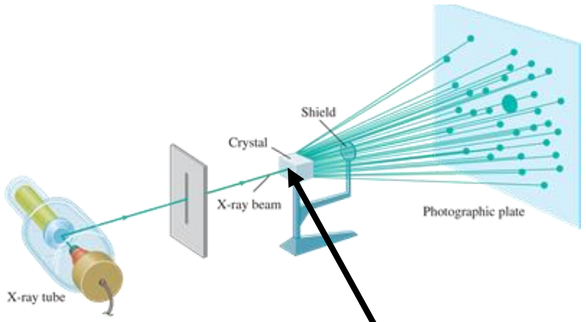
AI-Driven Voxel-wise Segmentation Pipeline for X-ray Tomography in Crystallography

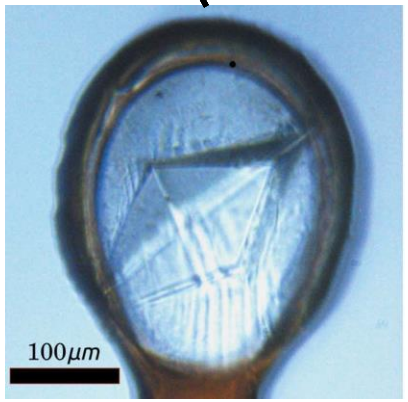
Yishun Lu, Wesley Armour,

Oxford e-Research Centre,

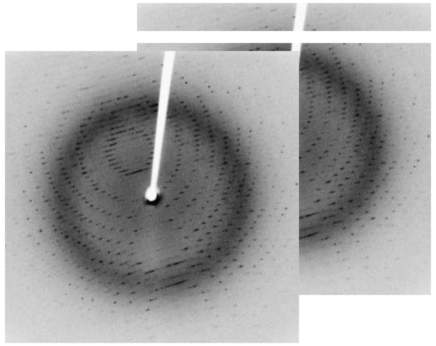
University of Oxford

X-ray Crystallography

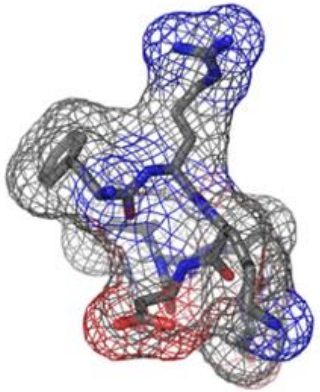




Detector measurement

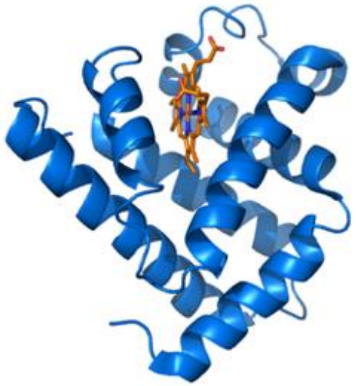


3D Inverse FFT



Electron density map

Fitting

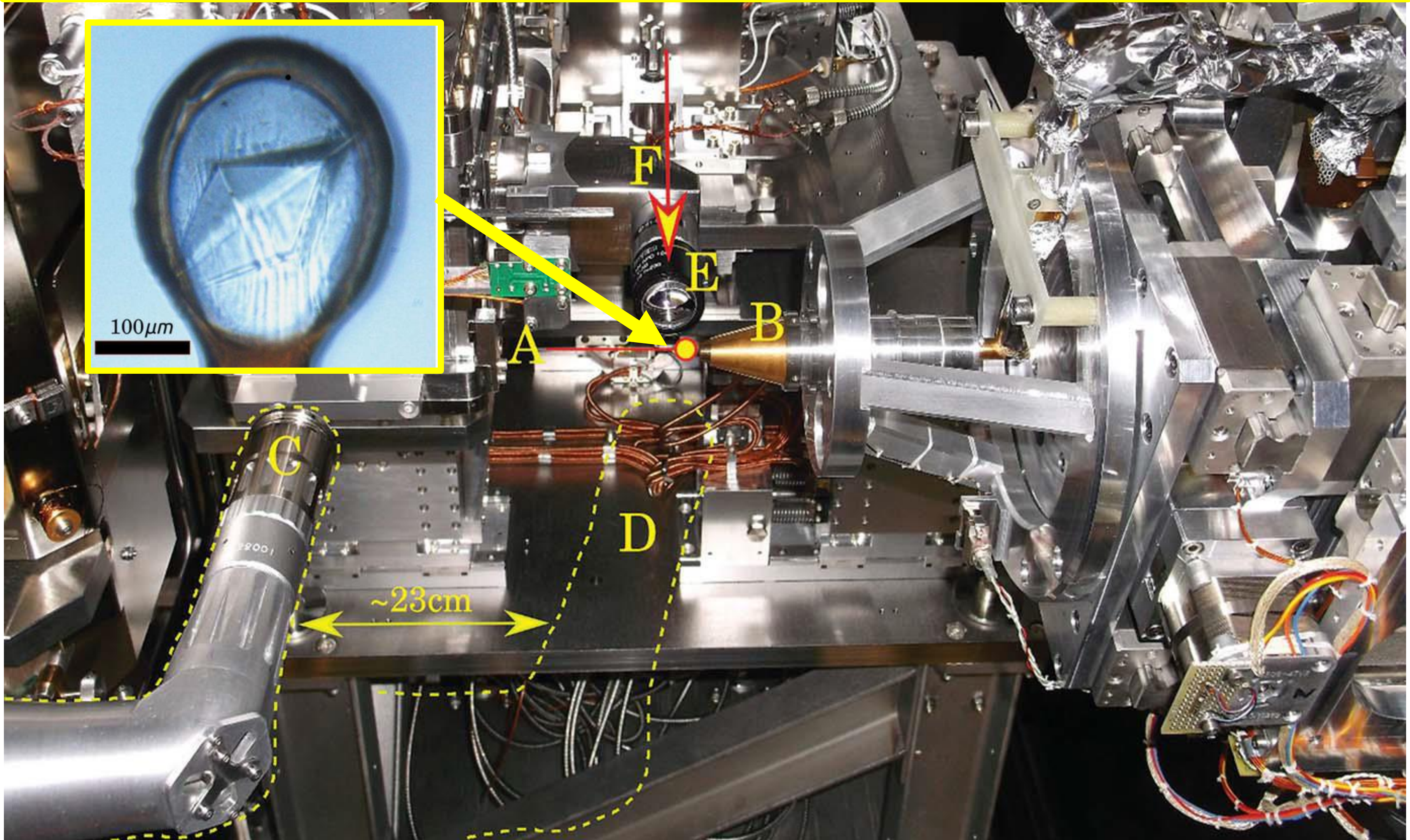


Atomic models

X-ray Crystallography is a mainstream protein structure-determination method

Image from https://macromoltek.medium.com/what-is-x-ray-crystallography-1e186bc3d180
https://www.creativebiomart.net/resource/principle-protocol-x-ray-crystallography-393.htm

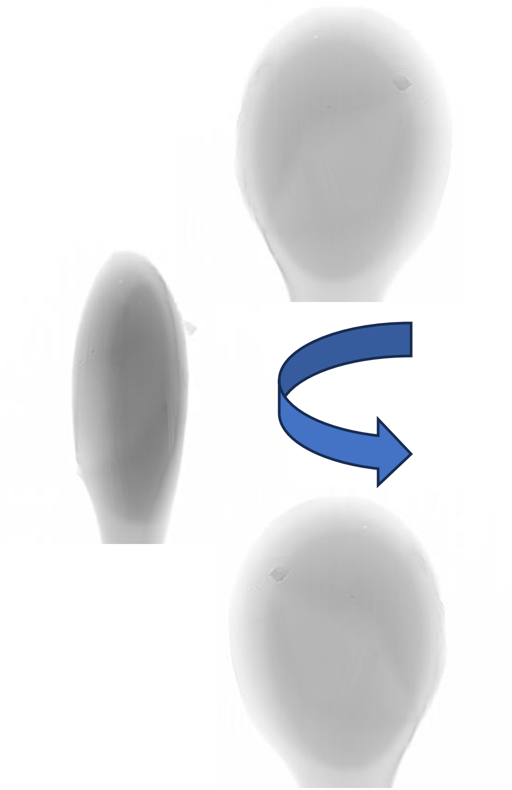
X-ray Tomography in Crystallography



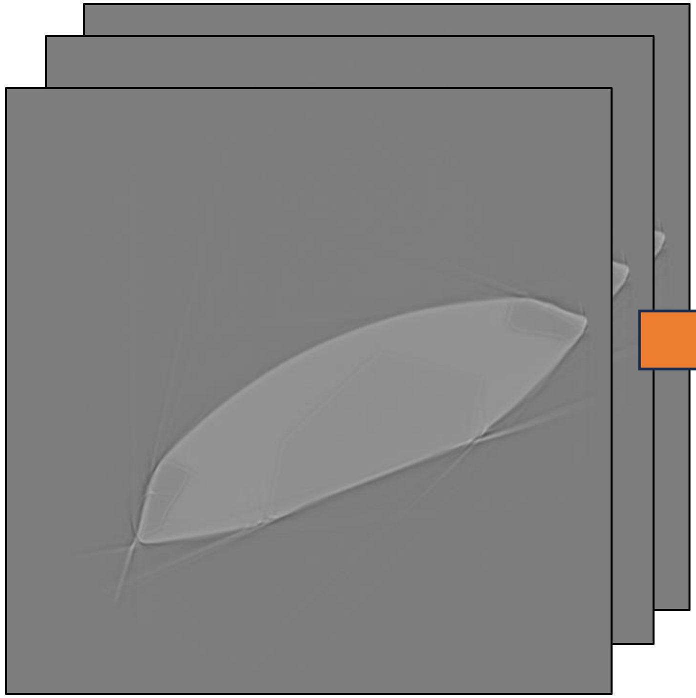
- Tunable-energy X-ray
  - 2.1keV – 18keV
- High resolution
  - 0.3µm/ pixel
- Horizontal goniometer rotational axis
  - Shared axis with diffraction experiment

Photo by Daniil Kazantsev et al. J. Synchrotron Rad. (2021). 28, 889–901

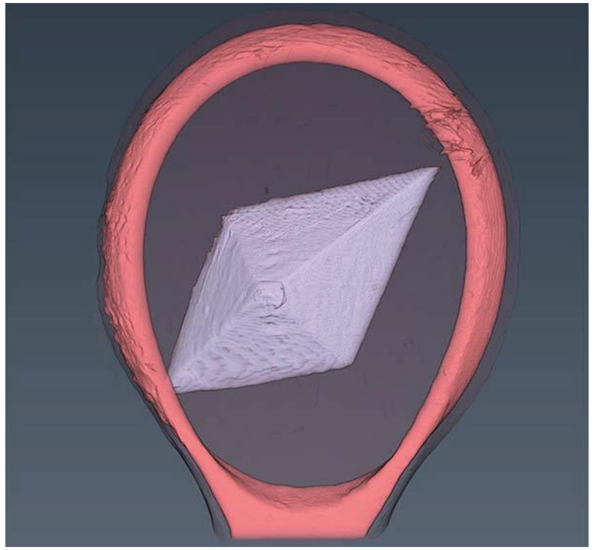
X-ray Tomography in Crystallography



0~180° tomography projection images



Tomography reconstruction slices



Manual segmented 3D model

Good, but it may take a scientist a day!!!

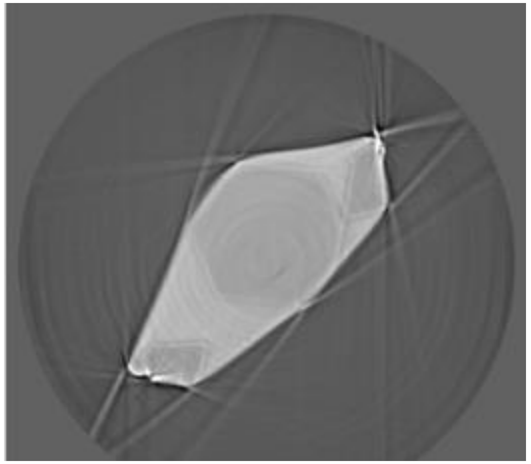
Challenges on segmentation

But… in limited and different experiments, we have **different:**

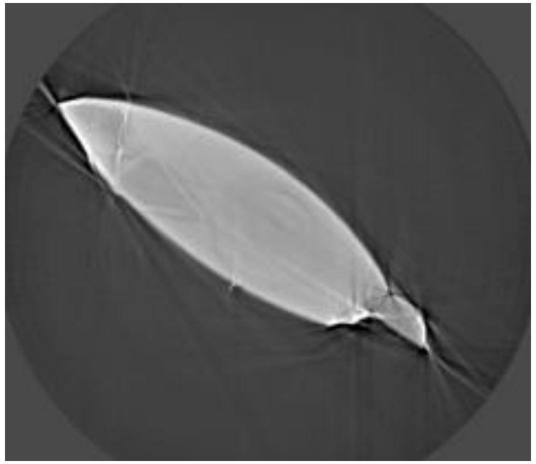
- ➢ Materials and shapes of the sample
- ➢ Absorption and refractive indices of the sample
- ➢ experimental parameters and setups

So… the reconstructions have **various:**

- ☒ Intensity distributions
- ☒ Reconstruction artefact
- ☒ Shapes and positions of materials



Sample A



Sample B

Towards accurate AI-Driven segmentation

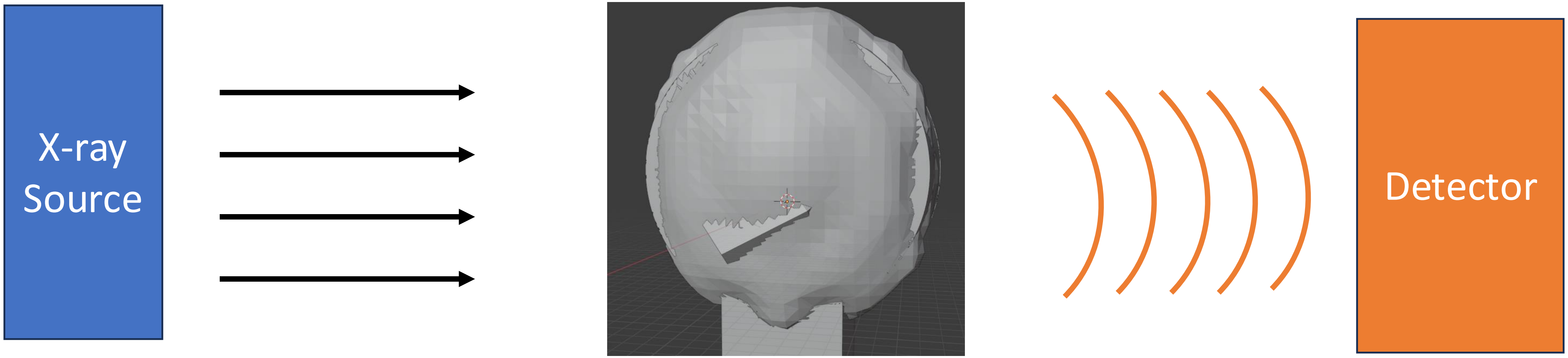
To achieve good automatic segmentation, the model needs to:

- Have enough **training data** to ensure sufficient data Data diversity
- Accurately distinguish the **boundaries** between materials
- Understand the **spatial relationships** between materials to classify the voxels into the correct categories

So… we need:

- ➢ A lot of training data
  - ✓ Experiments are expensive, but we can **simulate**!
- ➢ A powerful model
  - ✓ **ViT + Unet + Deformable convolution**

More data: Physics-based simulation



Simulated incident wavefield from tomography background

Assumed to be incident coherent beam

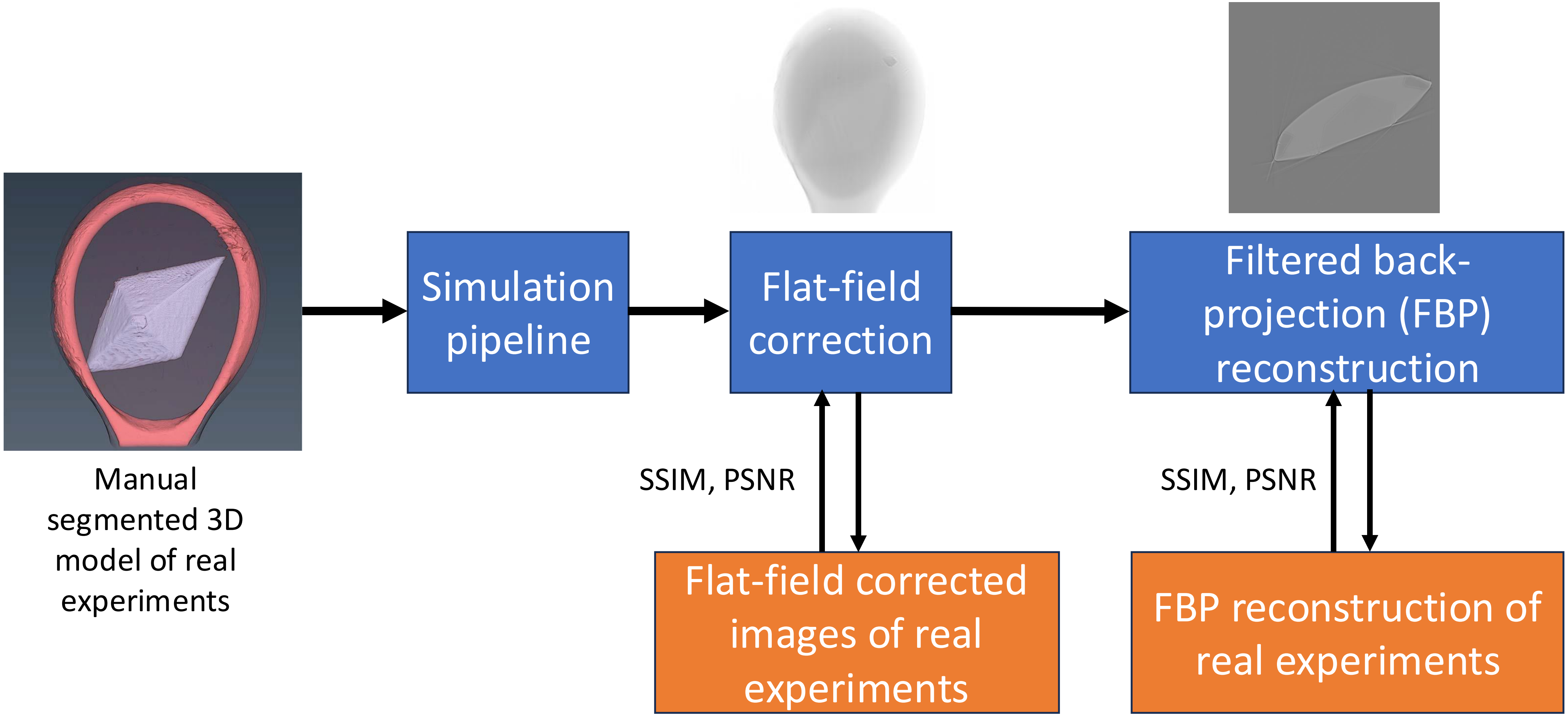
Propagation through simulated object by ***Blender***

Free-space propagation (Fresnel approximation)

Simulated detector assuming no loss due to the scintillator

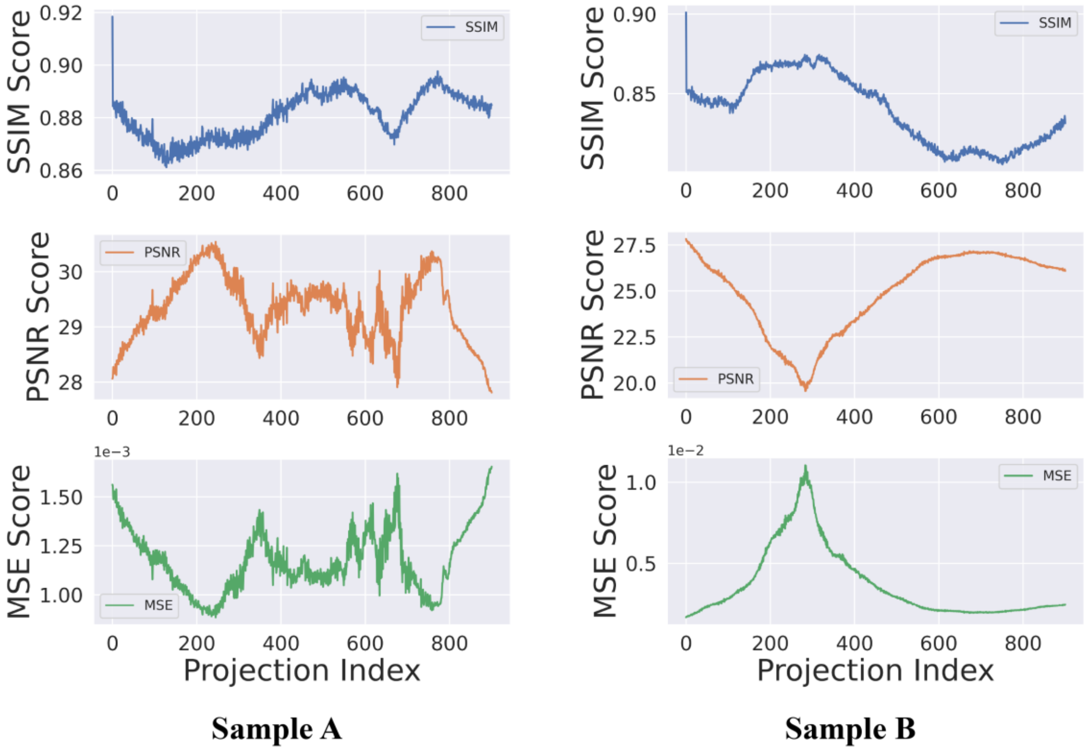
CUDA accelerated: ~0.7s for one tomography projection image

More data: Evaluation of simulation

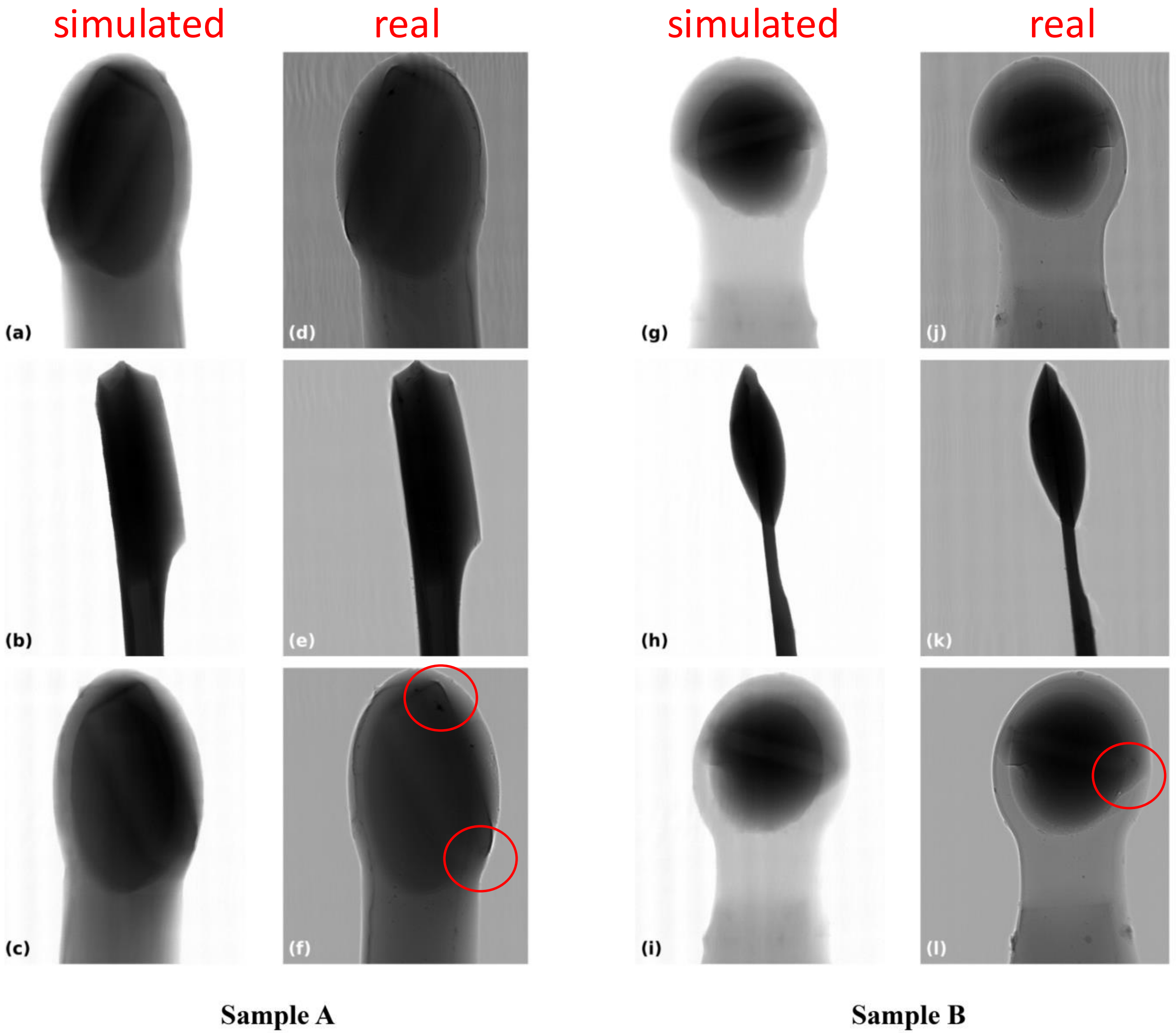


More data: Evaluation of simulation

Flat-fielded corrected projection images

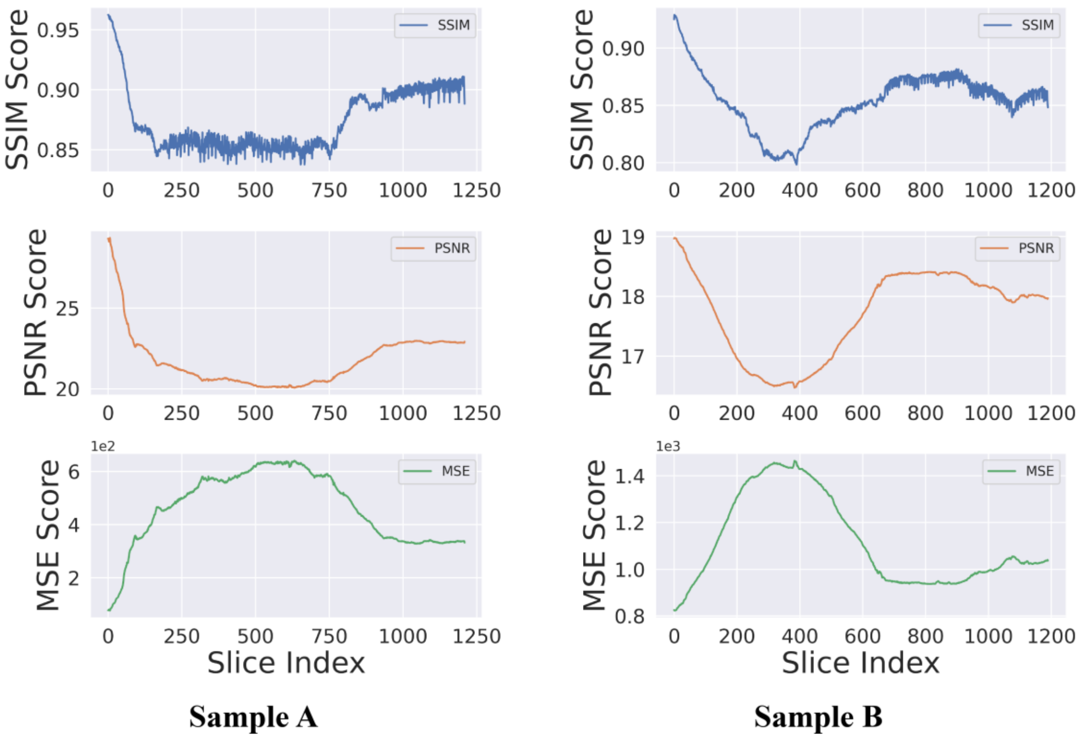


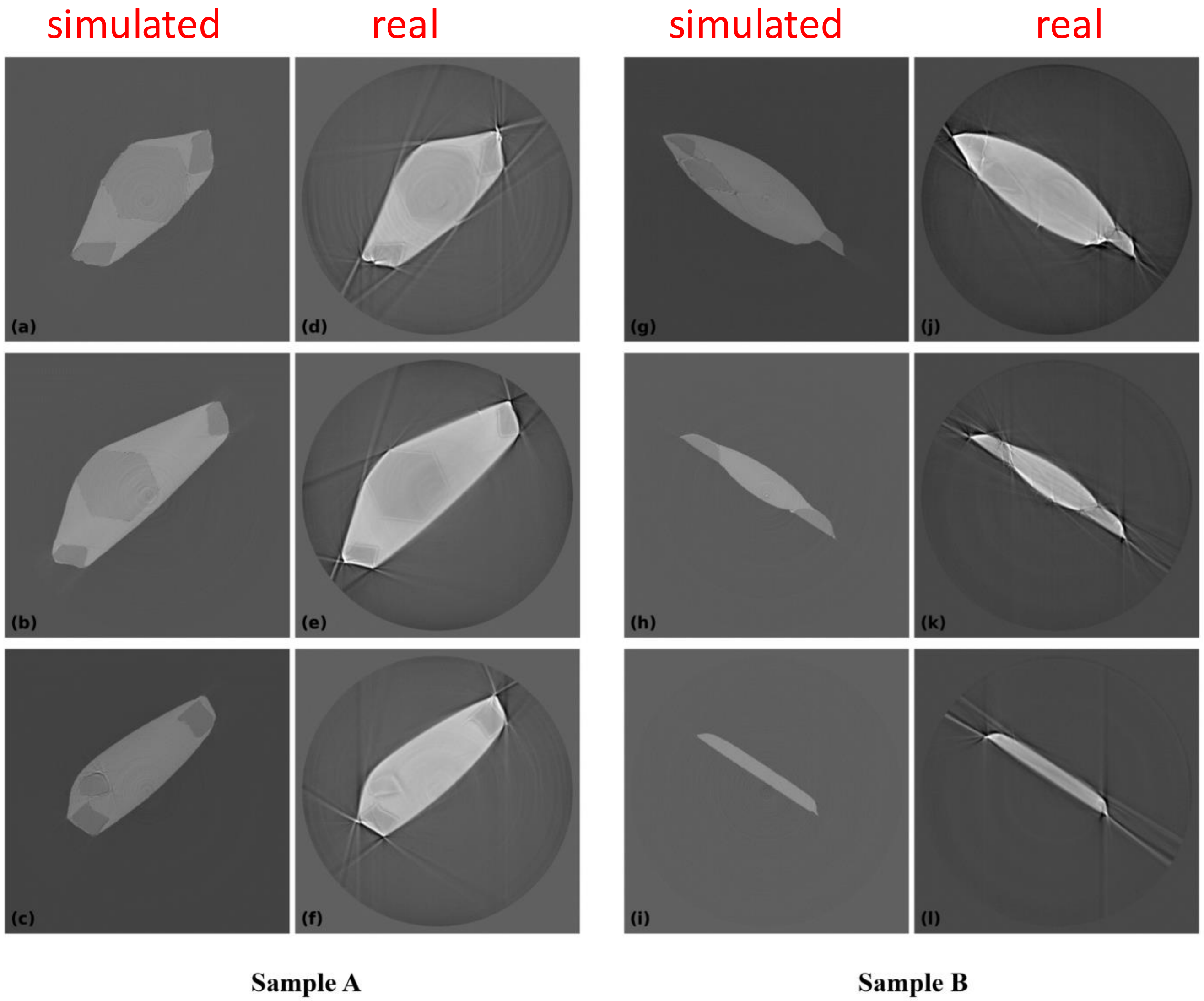
Rotation angle 0~180° (1 image for every 0.2°)



More data: Evaluation of simulation

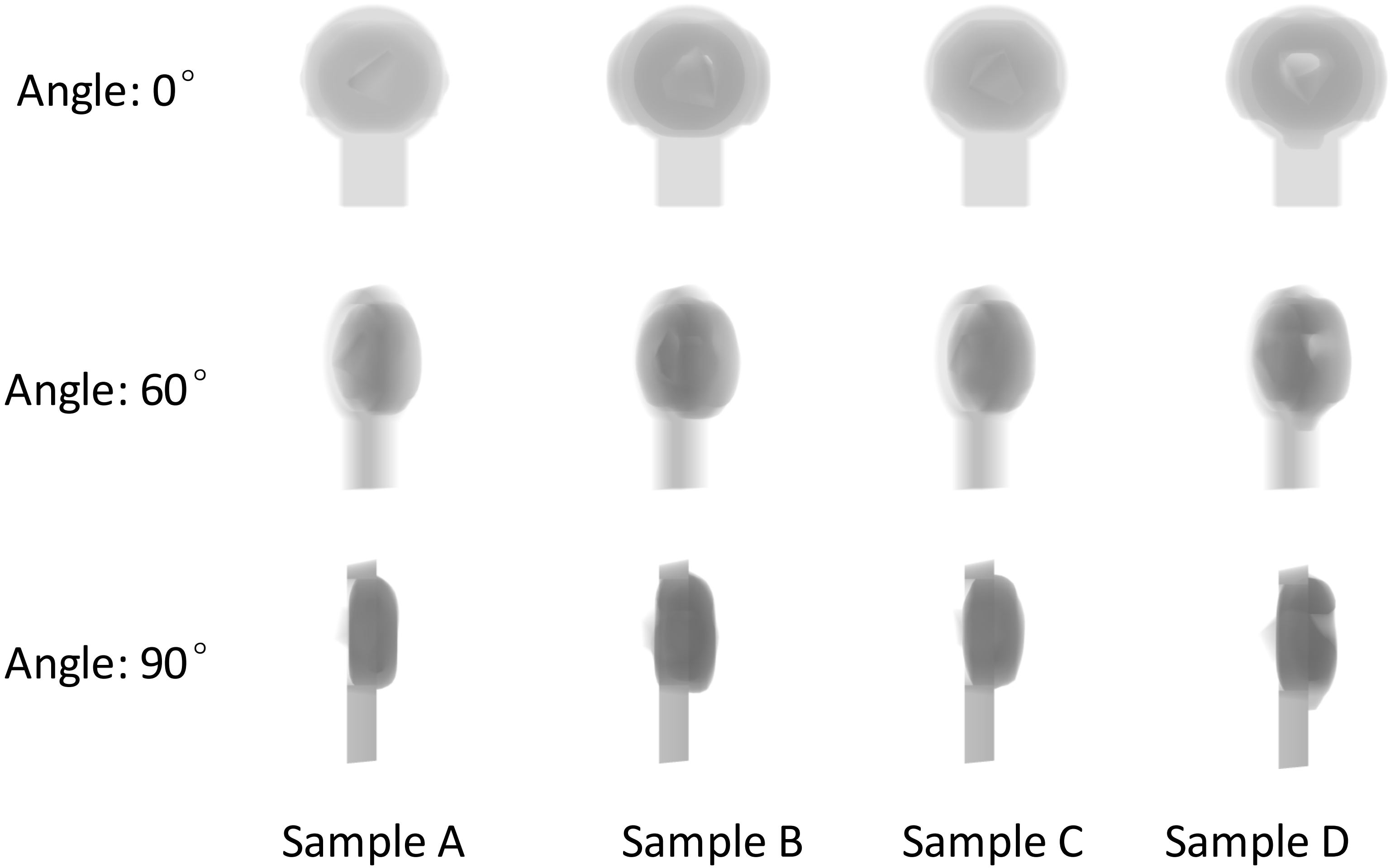
Filtered back-projection (FBP) reconstruction slices





More data: Evaluation of simulation

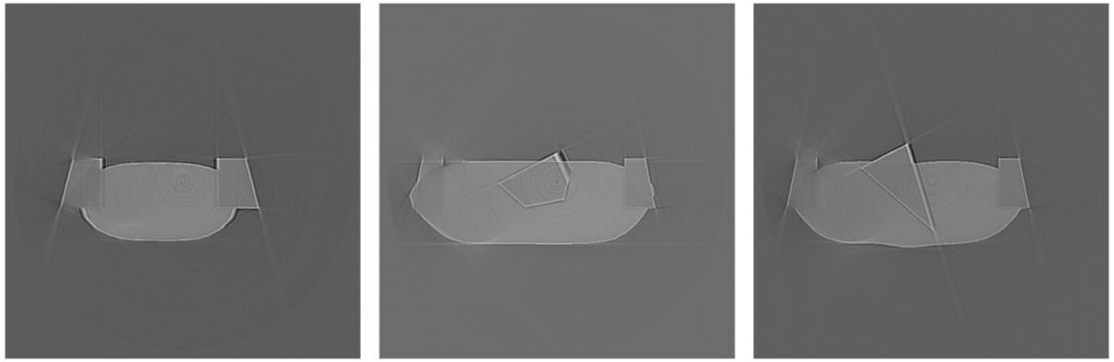
Synthetic flat-field corrected images



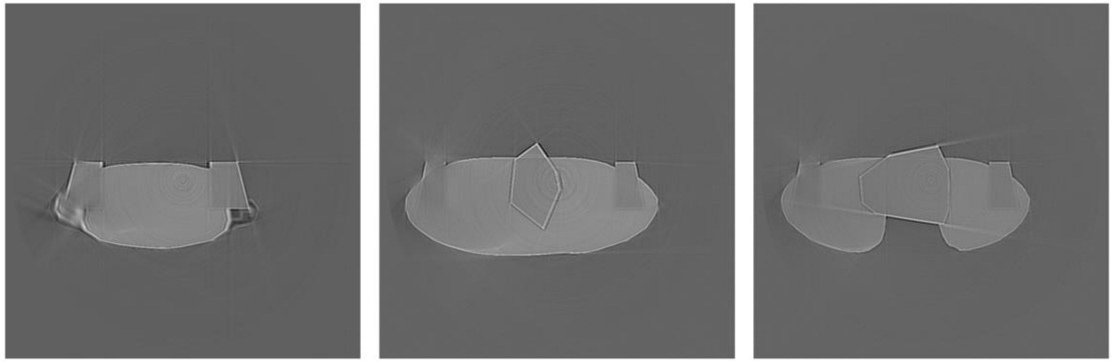
More data: Evaluation of simulation

Synthetic FBP reconstruction slice images

Sample A

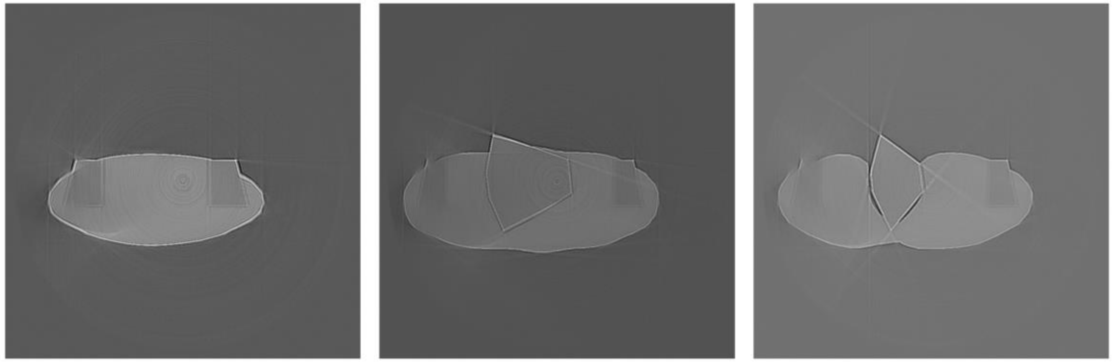


Sample C

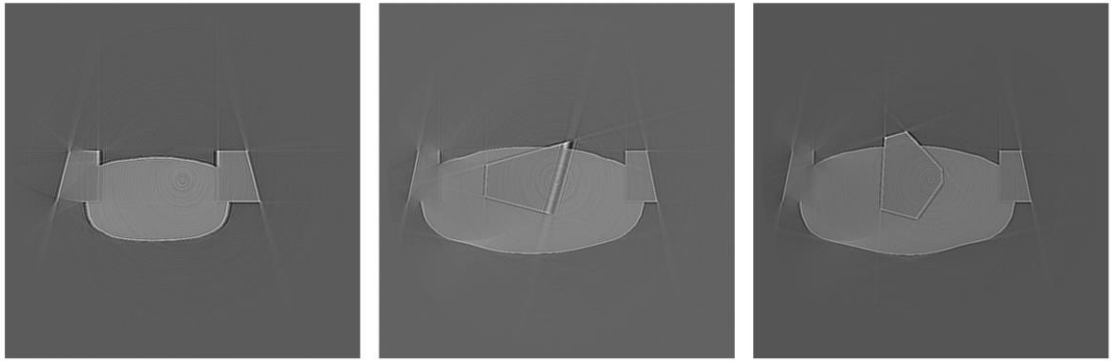


300 400 500 **Slice index** 300 400 500

Sample B

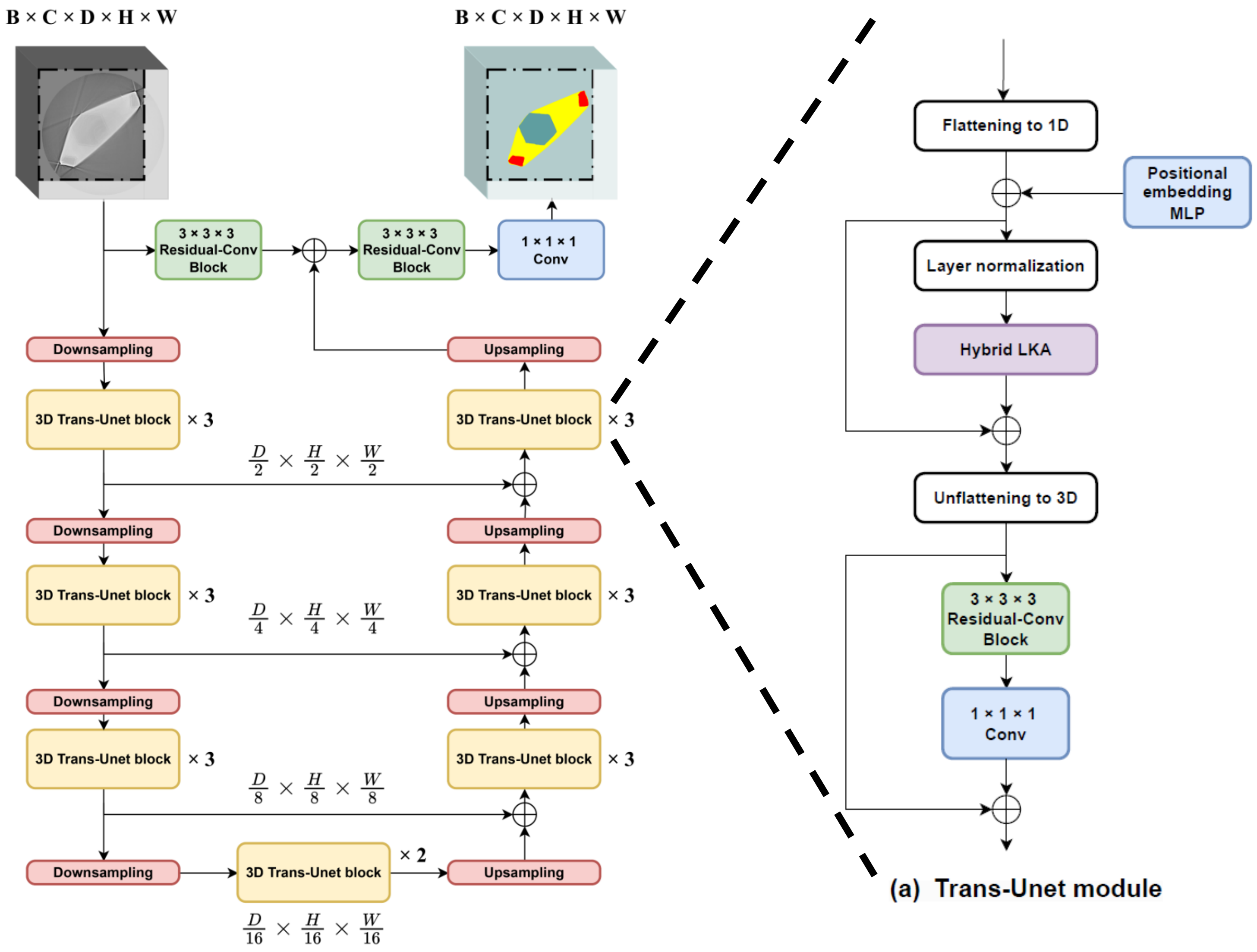


Sample D



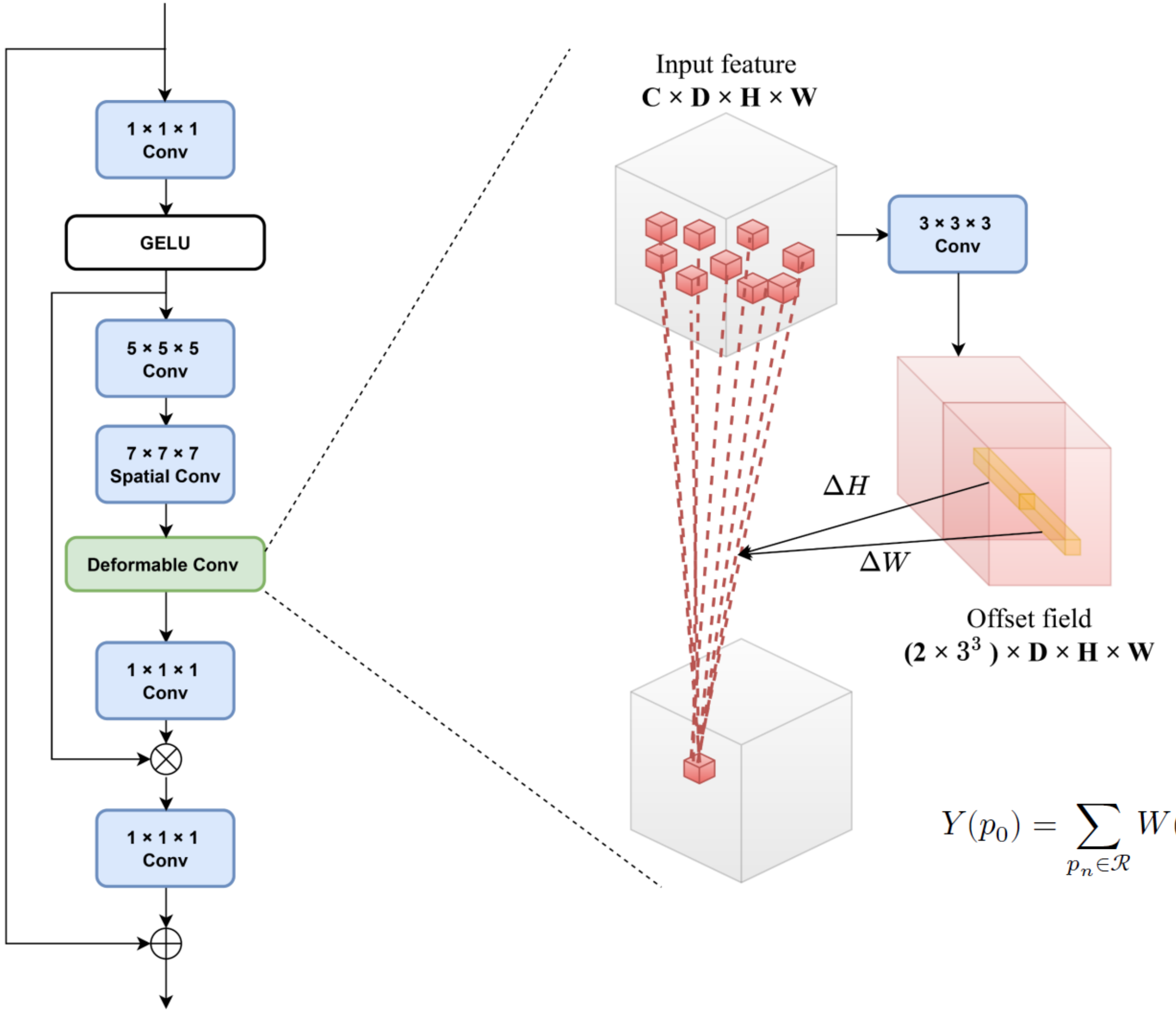
AI-driven 3D segmentation

AI model
architecture
AnACorNet



AI-driven 3D segmentation

Hybrid Large Kernel Attention (LKA)

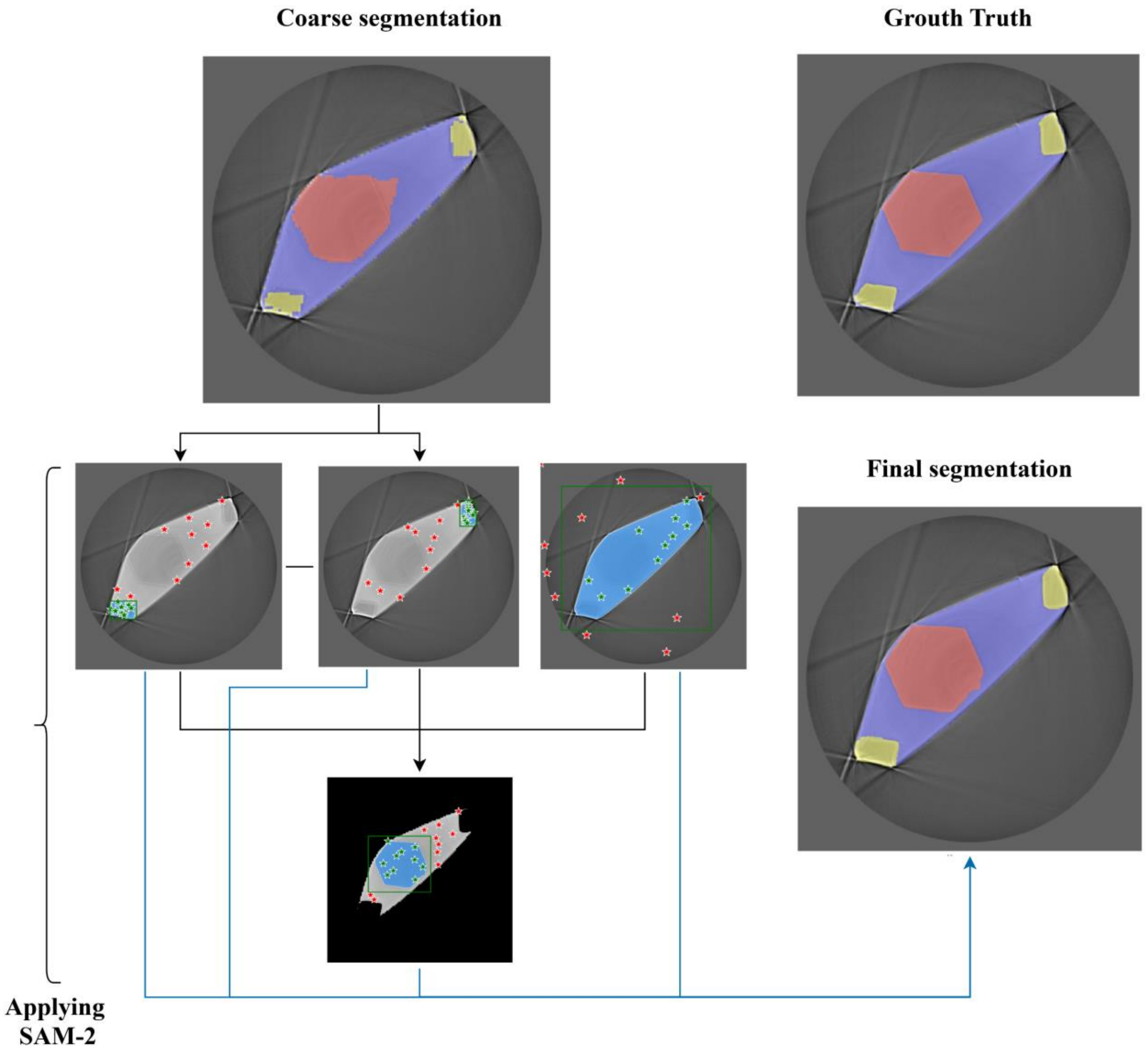


$$Y(p_0) = \sum_{p_n \in \mathcal{R}} W(p_n) \cdot X(p_0 + p_n + \Delta p_n)$$

(a) Deformable LKA module (b) Deformable Conv block

AI-driven 3D segmentation

Post refinement by SAM-2 slice by slice



Results

_R: training with only real dataset (12 datasets)
_RS: training with real and synthetic dataset (120 datasets)
_SAM: post refined by SAM-2 with the same setup

Table 2.1: Comparison of different segmentation models for Thermolysin and Thaumatin dataset in terms of accuracy (Dice Loss) for each class and Cross-entropy (CE) loss.

Method	**Accuracy (Dice Loss)**				**CE Loss**
	Background	**Mother Liquor**	**Loop**	**Crystal**	
Thermolysin					
AnACorNet_R	99.95% (0.0014)	82.68% (0.1770)	43.48% (0.4336)	80.43% (0.3134)	0.1839
AnACorNet_R_SAM	99.90% (0.0013)	88.15% (0.1258)	82.50% (0.1241)	88.26% (0.1394)	0.0706
AnACorNet_RS	99.88% (0.0013)	93.93% (0.1014)	78.10% (0.1426)	90.06% (0.0813)	0.0541
AnACorNet_RS_SAM	99.89% (0.0013)	93.47% (0.0712)	88.11% (0.0938)	96.54% (0.0382)	0.0327
Thaumatin					
AnACorNet_R	99.94% (0.0014)	86.51% (0.1989)	52.52% (0.3652)	69.42% (0.2892)	0.1341
AnACorNet_R_SAM	99.85% (0.0012)	93.80% (0.0978)	87.06% (0.1093)	88.38% (0.0776)	0.0360
AnACorNet_RS	99.82% (0.0012)	95.15% (0.0945)	81.40% (0.1286)	92.91% (0.0589)	0.0368
AnACorNet_RS_SAM	99.88% (0.0011)	93.38% (0.0716)	90.45% (0.0858)	96.42% (0.0415)	0.0222

Results

Method	**Inference running Time**	
	Thermolysin	**Thaumatin**
AnACorNet	25.59 seconds	20.85 seconds
AnACorNet + SAM	≈70 minutes	≈60 minutes
Manual Labor	$\approx\geq$ 4 hours	$\approx\geq$ 4 hours

Results

Application on analytical absorption correction factors

$$A_{\mathbf{h}} = \frac{1}{V}\int_V e^{-\mu(L_1(x,y,z)+L_2(x,y,z))}\, dV$$

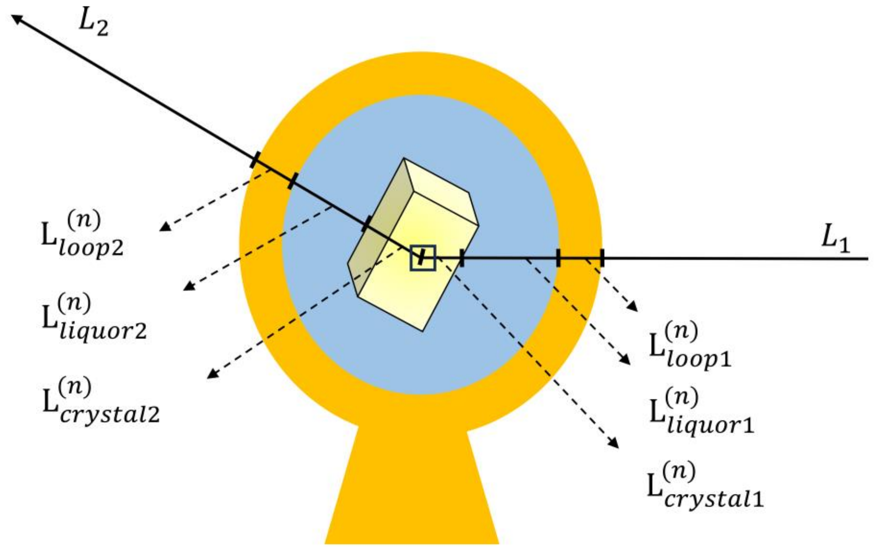
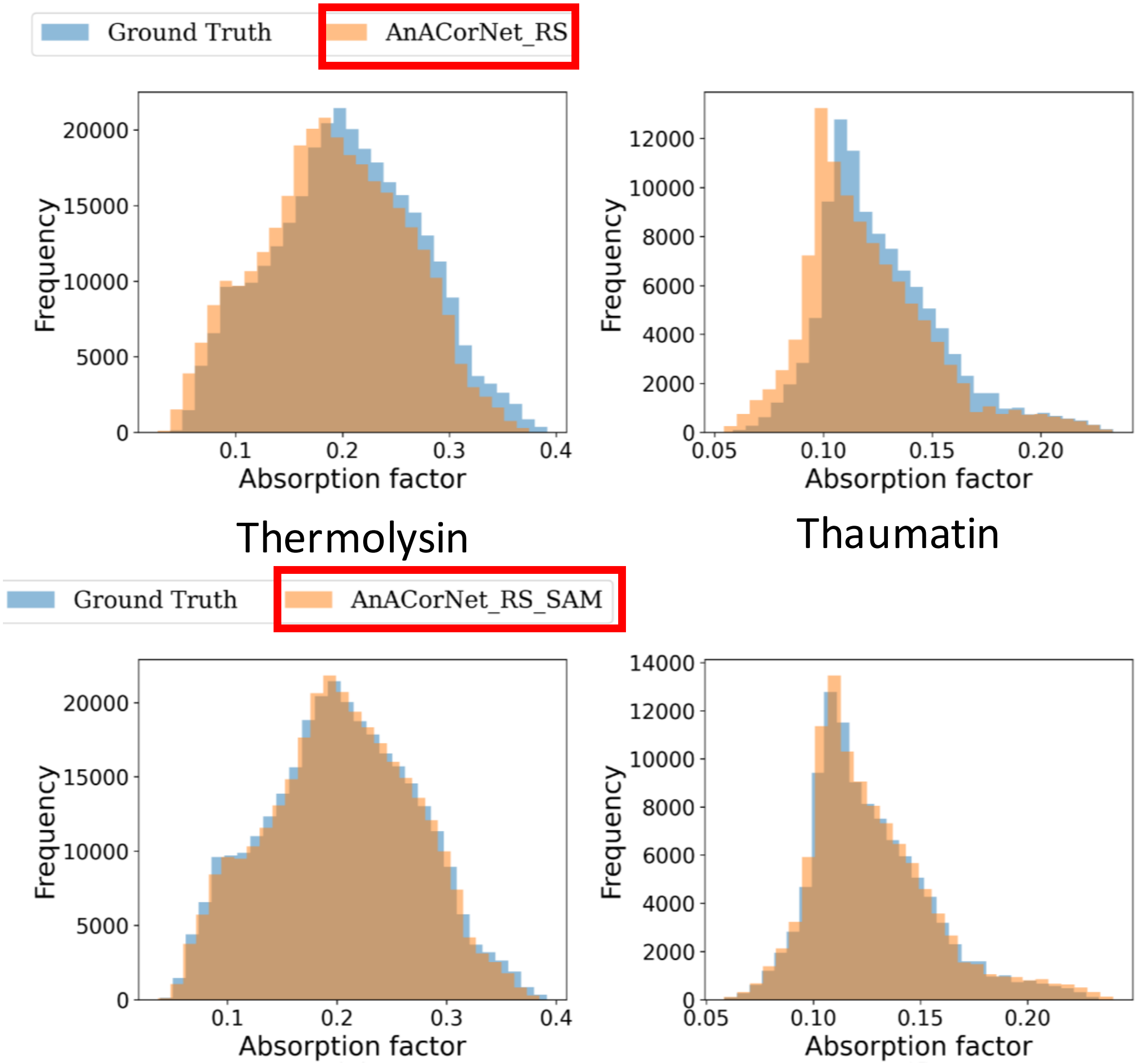


Photo by Yishun Lu et al., J. Appl. Cryst. (2024). 57, 649–658



Results

Application on analytical absorption correction in crystallography

Table 2.6: Comparison of merging statistics of Thermolysin results from Ground Truth, AnACorNet_RS, and AnACorNet_RS_SAM of AAC scaling method [59]. The values in brackets represent high-resolution statistics

Metric	Manual	AnACorNet_RS	AnACorNet_RS_SAM
Resolution limit	129.26 - 2.31 (2.35 - 2.31)	129.26 - 2.31 (2.35 - 2.31)	129.26 - 2.31 (2.35 - 2.31)
Completeness (%)	96.5 (90.2)	96.5 (90.2)	96.5 (90.2)
Multiplicity	21.3 (9.4)	21.3 (9.4)	21.3 (9.4)
I/sigma	26.6 (7.9)	26.0 (7.8)	26.5 (7.9)
Rmerge	0.134 (0.345)	0.141 (0.352)	0.133 (0.344)
Rmeas	0.137 (0.364)	0.144 (0.371)	0.135 (0.363)
Rpim	0.027 (0.113)	0.029 (0.115)	0.027 (0.113)
CC half	0.996 (0.943)	0.997 (0.939)	0.997 (0.944)
Anomalous correlation	-0.159 (-0.521)	-0.210 (-0.520)	-0.167 (-0.524)
Anomalous slope	1.096	1.081	1.092
Total reflection number	308760 (6189)	308944 (6189)	308778 (6189)
Unique reflection number	14513 (656)	14513 (656)	14513 (656)

Table 2.8: Comparison of merging statistics of Thaumatin results from Ground Truth, AnACorNet_RS, and AnACorNet_RS_SAM of AAC scaling method [59]. The values in brackets represent high-resolution statistics

Metric	Manual	AnACorNet_RS	AnACorNet_RS_SAM
Resolution limit	150.73 - 2.70 (2.75 - 2.70)	150.73 - 2.70 (2.75 - 2.70)	150.73 - 2.70 (2.75 - 2.70)
Completeness (%)	99.2 (90.8)	99.2 (90.8)	99.2 (90.8)
Multiplicity	13.9 (5.4)	13.9 (5.4)	13.9 (5.4)
I/sigma	37.7 (18.9)	35.5 (17.5)	37.6 (19.0)
Rmerge	0.082 (0.098)	0.086 (0.094)	0.084 (0.094)
Rmeas	0.085 (0.108)	0.089 (0.103)	0.087 (0.103)
Rpim	0.022 (0.043)	0.023 (0.041)	0.022 (0.041)
CC half	0.997 (0.990)	0.997 (0.992)	0.997 (0.992)
Anomalous correlation	0.572 (0.315)	0.551 (0.363)	0.565 (0.302)
Anomalous slope	2.781	2.615	2.835
Total reflection number	105385 (1873)	105392 (1873)	105381 (1873)
Unique reflection number	7580 (345)	7580 (345)	7580 (345)

Thank you for listening