Supervised and Unsupervised Machine Learning for Large, Noisy STEM Data

Paul M. Voyles
Materials Science and Engineering

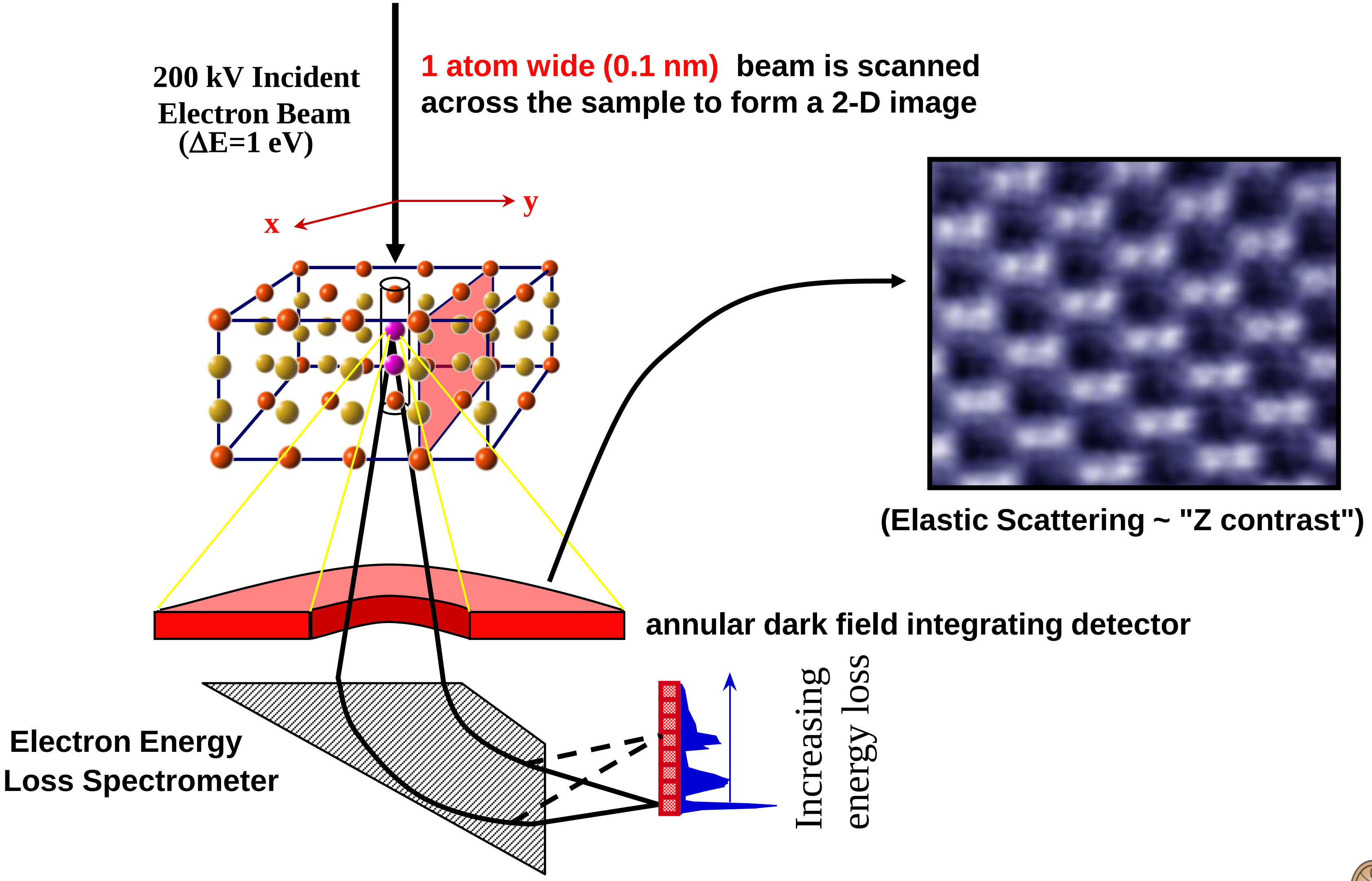
WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON

NSF DMR-1720415

DE-FG02-08ER46547

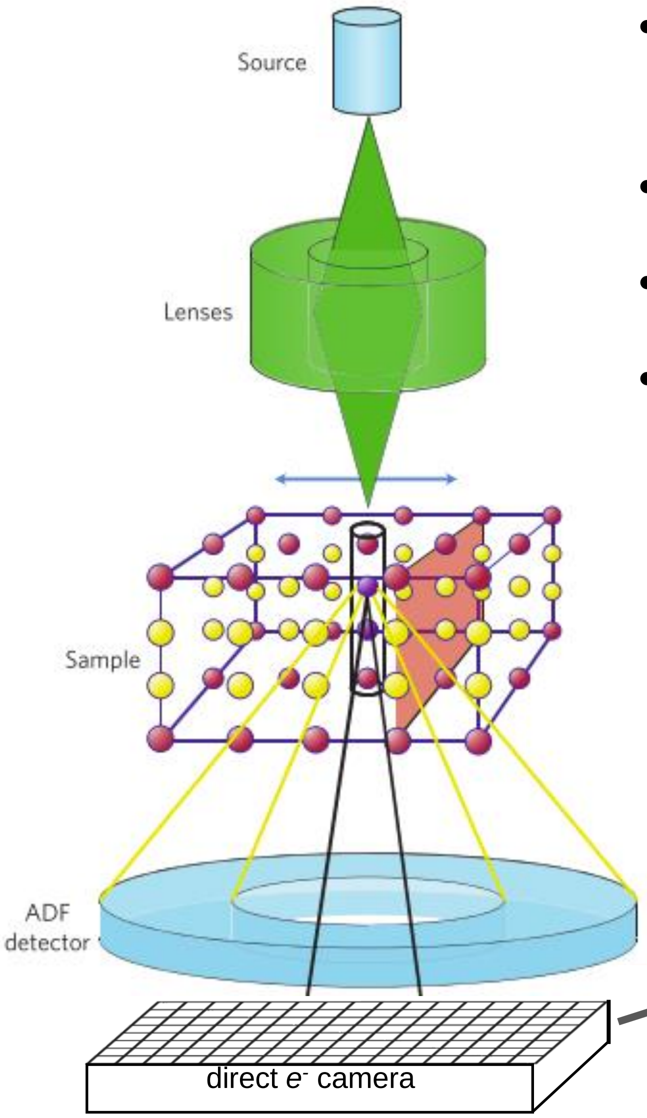
Scanning Transmission Electron Microscopy

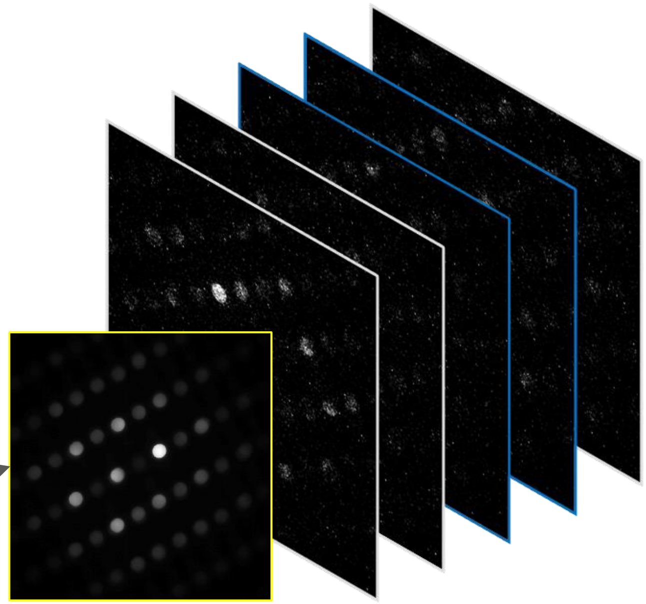
- Modern commercial instrument offers:
  - 1 nA into a 1 Å probe
  - 9000 EEL spectra/sec
  - Atomic-resolution STEM images at 100 nsec / pixel
  - Atomic-resolution TEM images at 3000 fps
- Routine acquisition of O(10 GB) datasets
- Specialize instruments go (much) faster



4D STEM Data: $I(r_x, r_y, k_x, k_y)$

- Read out a fast, pixelated camera at every position of a STEM probe scan
- Revolutionizing materials TEM & STEM
- Ultrafast cameras at UW-Madison and LBL
- TB+ datasets are routine, requiring large scale computing and AI / ML



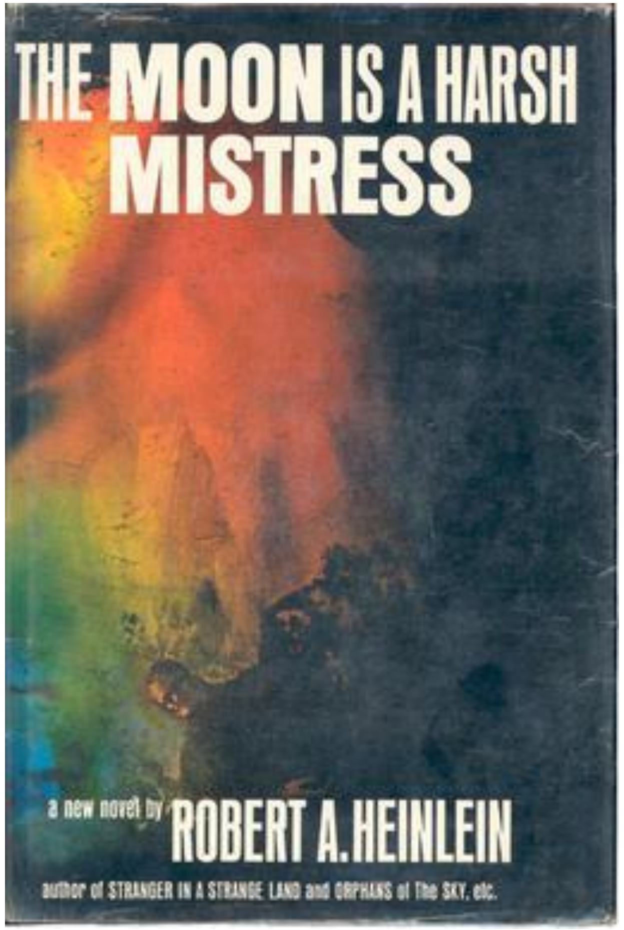


Maximum Readout Speed	
Readout size	frames per second
1024×1024	1,900
512×512	7,380
256×256	26,000
256×128	49,000
256×64	86,000

TANSTAAFL

There ain't no such thing as a free lunch

- ML / AI methods always rely on some form of prior information about
- Unsupervised learning:
  - prior information about the mathematical structure of the data
  - applications in distortion correction, denoising, spectral unmixing, and signal clustering
- Supervised learning:
  - prior information from example, already analyzed data
  - applications in finding features in images, connecting images and spectra to physical quantities of interest
- Examples of both as they apply to STEM

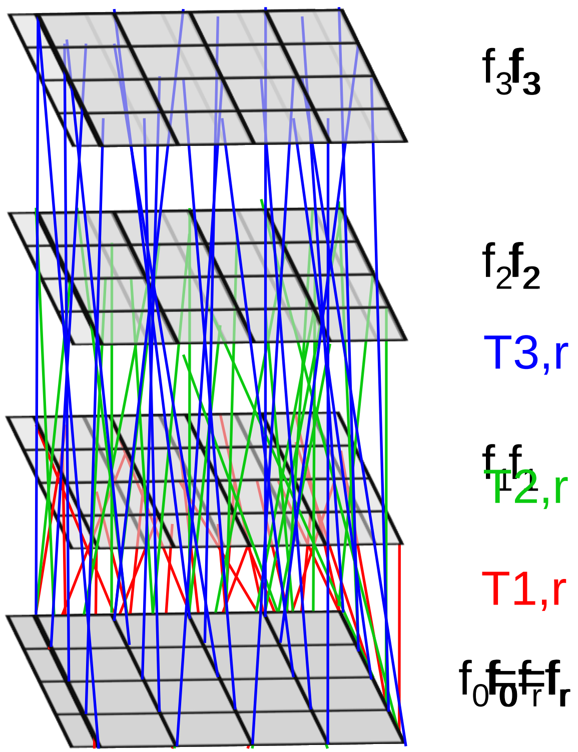


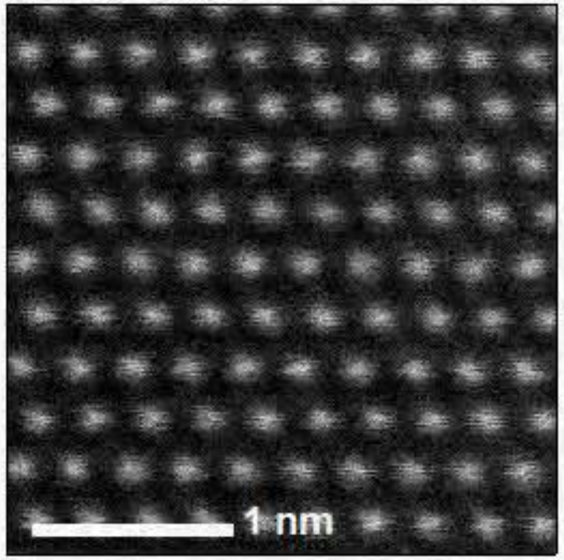
Distortion Correction in Scanning Images

- Imperfect instruments

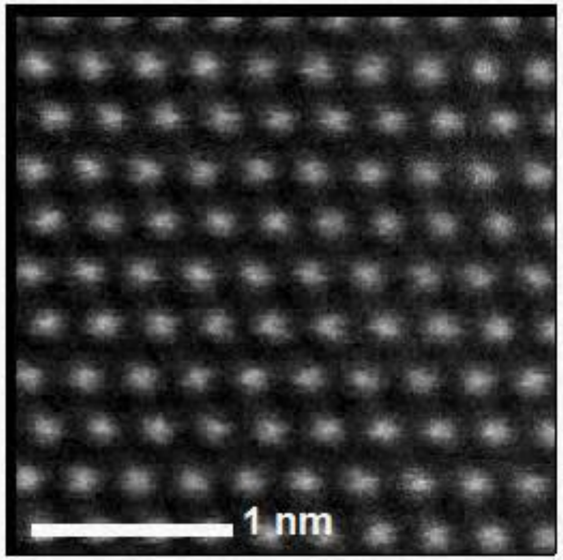
- Scanning images are subject to distortion arising from instrumental instabilities
- Distortions can be corrected from a series of images if the object is unchanged and the distortions are random
- More prior information:
  - Higher frequency distortions are smaller in magnitude (*e.g.* electronic jitter vs floor vibrations)
  - Lowest frequency distortions are rigid motion of the sample + shear of the image
  - Higher frequency distortions have zero mean over many images

Distortion Correction by Non-Rigid Registration

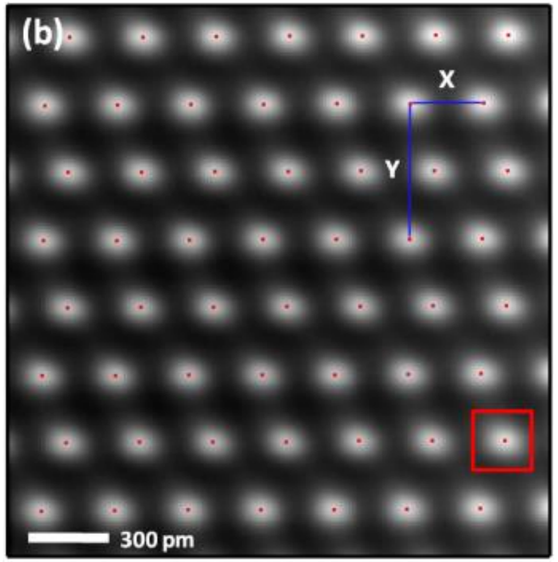




Raw series



Non-rigid aligned series



average image

- Average image has high SNR and low distortion
- Enables measurement of atom positions with <1 pm precision

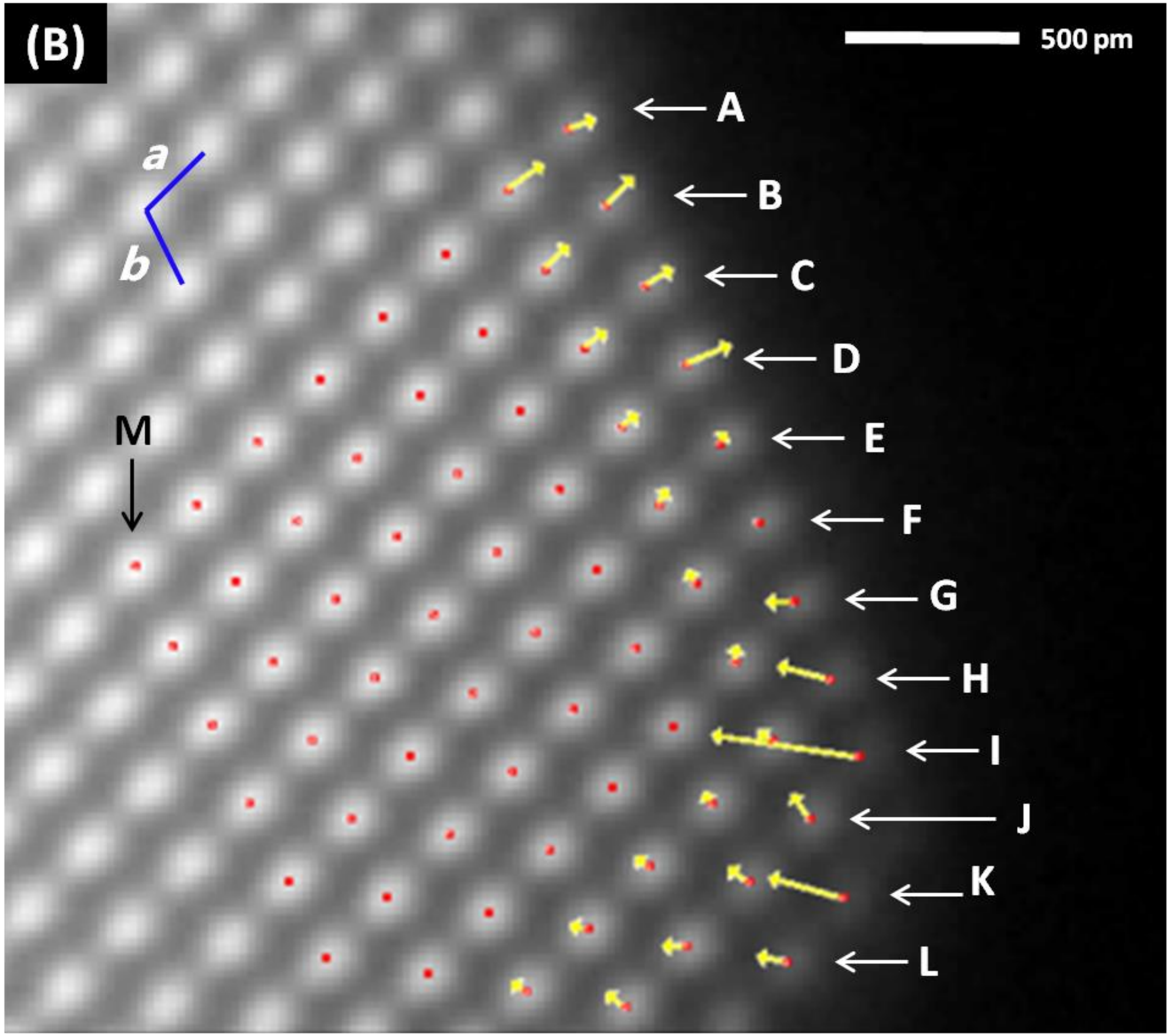
Andrew Yankovich

Benjamin Berkels

Nat. Comm. **5**, 5144 (2014)

Pt on SiO_2 Catalyst

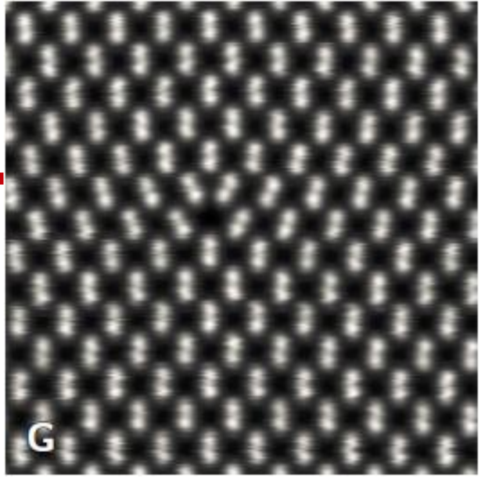
- Catalysis happens preferentially at corners and edges of nanoparticles
- Atoms at corners and edges lack some neighboring atoms
- They have shorter bonds that atoms inside the particle
- We can measure those bond lengths more accurately that it is possible to calculate them.



(C) Atom	Δd (pm)
A	9 ± 2
B	12 ± 4
C	10 ± 2
D	15 ± 3
E	4 ± 2
F	2 ± 2
G	8 ± 3
H	16 ± 5
I	44 ± 10
J	9 ± 3
K	23 ± 7
L	8 ± 4

Nat. Comm. **5**, 5144 (2014)

Non-local Means Denoising

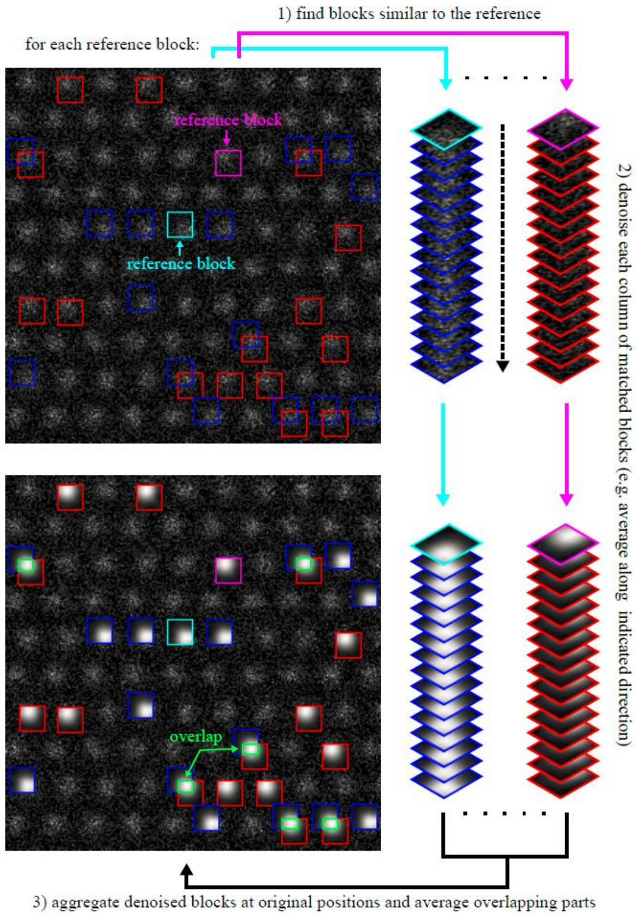


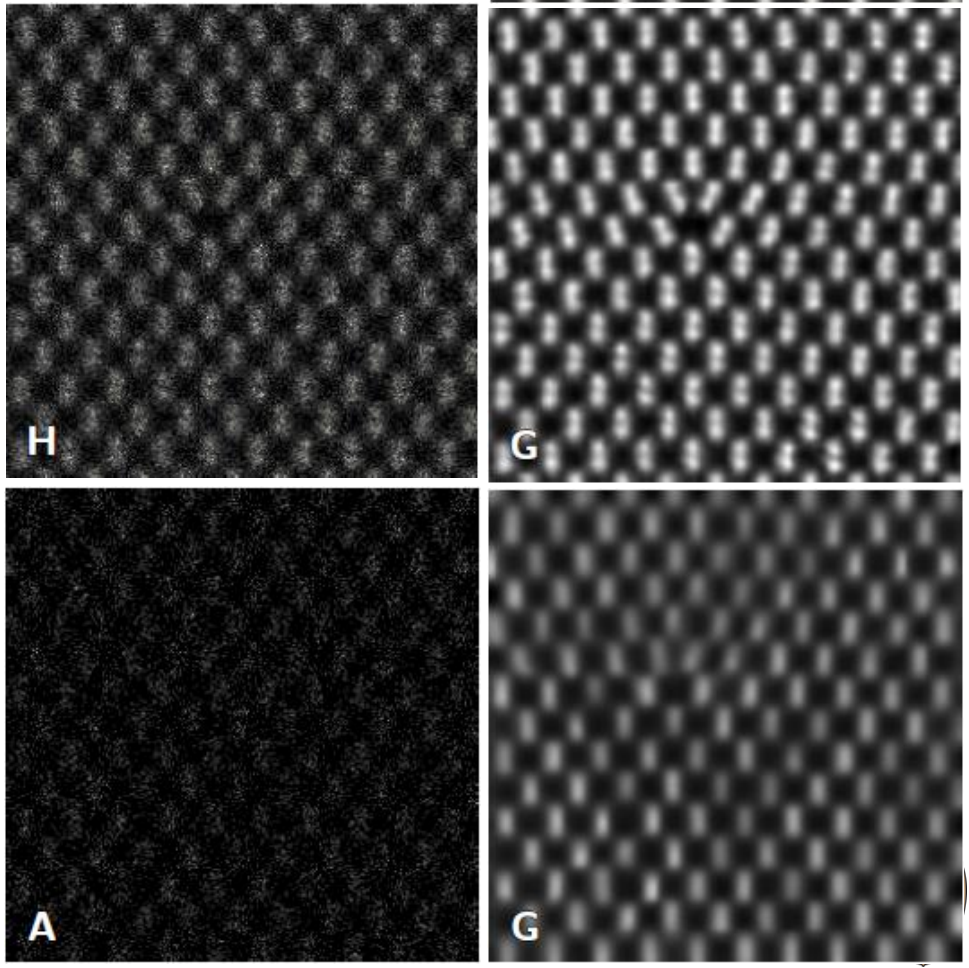
- Prior information:
  - High-resolution images of crystals contain many repeating features
  - Electron detection experiments are corrupted by Poisson noise
- Result:
  - Non-local means with periodic block matching
  - Similarity measure for Poisson noise
  - Better performance than state-of-the-art BM3D.

N. Mevenkamp, *Adv. Struct. Chem. Imaging* **1,** 3 (2015).

Niklas Mevenkamp

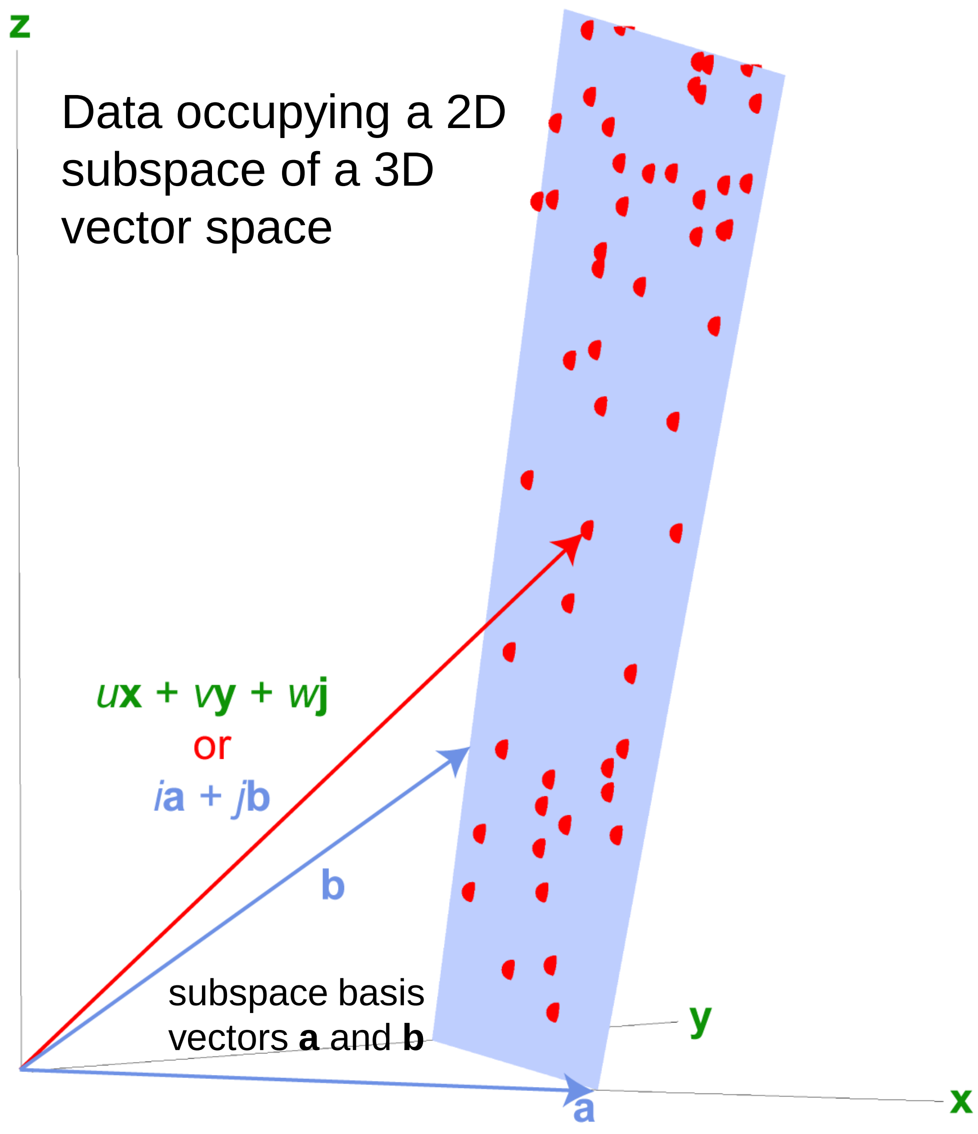
Benjamin Berkels





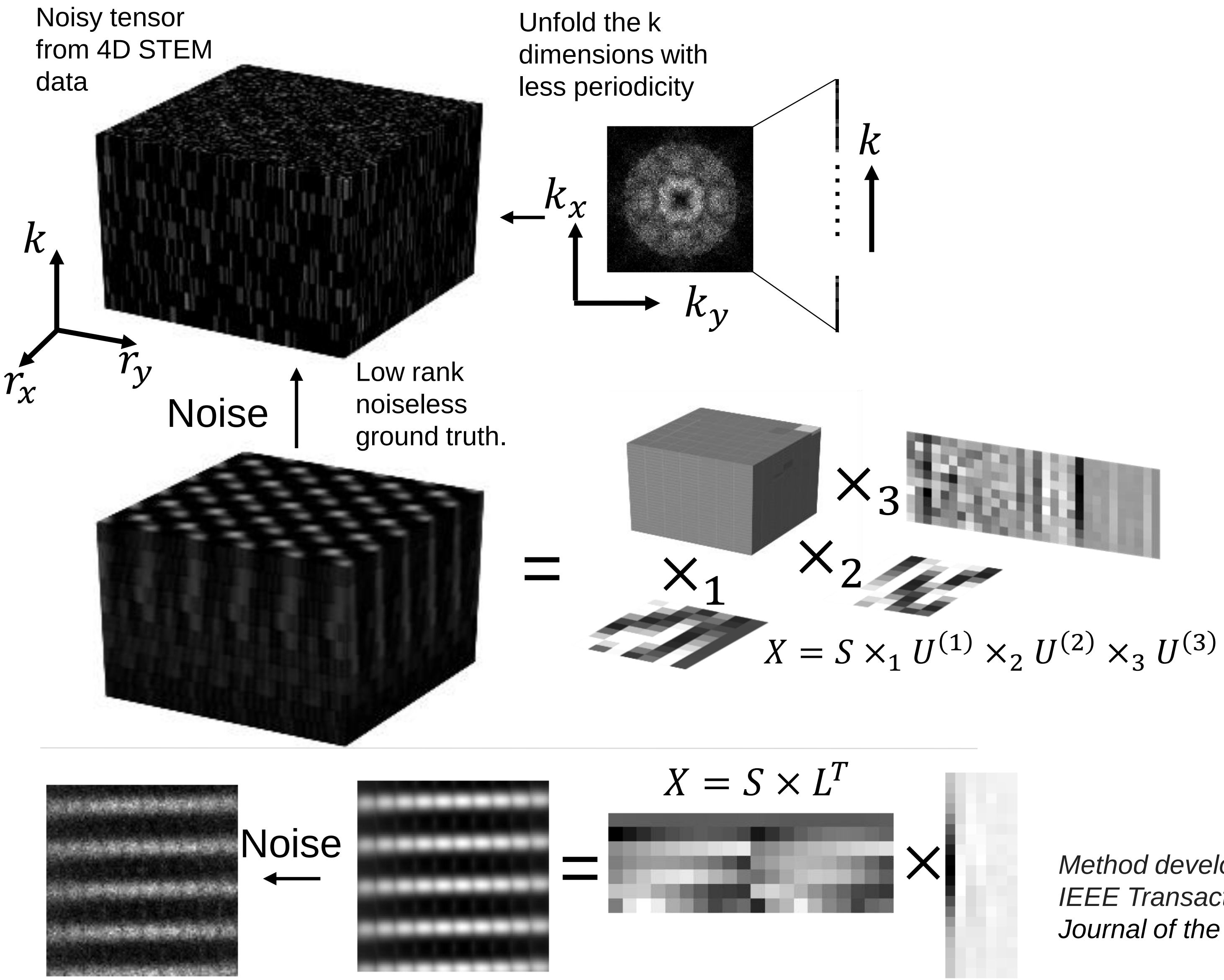
Dimensionality Reduction

- Prior information: data occupy a subspace in the high dimensional vector space of the set of possible measurements
- Applications:
  - denoising by throwing away the signal outside the subspace
  - spectral unmixing to discover prototype signals or for mapping
- Methods for 2D matrices like PCA are widely used, but do not exploit the full structure of higher-dimensional data



M. Bosman, *Ultramicroscopy* **106**, 1024–32 (2006)

Tensor Singular Value Decomposition

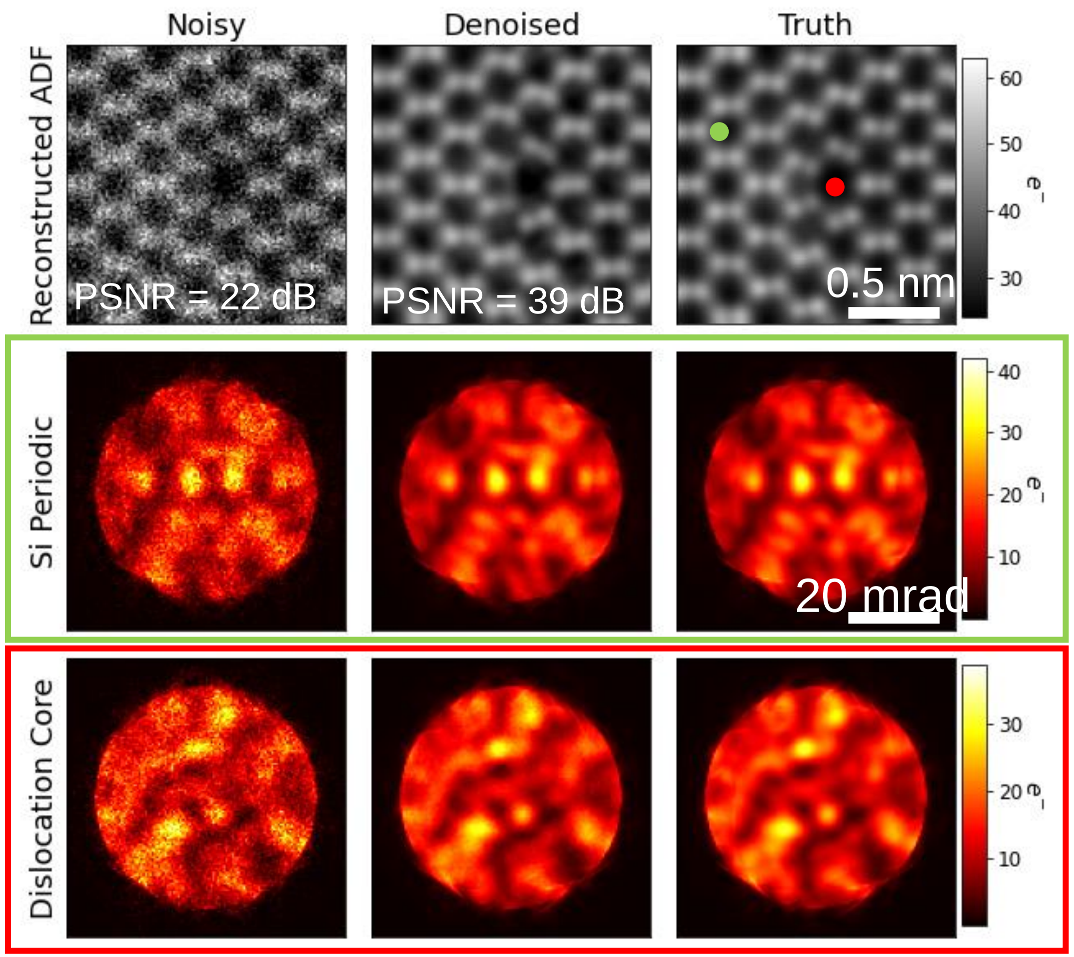


- Generalization of SVD to work on tensors of any shape
- Can be applied to arbitrary high dimensional data while maintaining structure along all dimensions.
- Preserves inherent structure in the data, aiding learning when the data are highly redundant, like atomic-resolution 4D STEM data
- Iterative estimate, not closed-form solution like 2D SVD.
- Find low rank ground truth from noisy input data.

Method developed by Rungang Han and Arun Zhang:
IEEE Transactions on Information Theory 64.11 (2018): 7311-7338.
Journal of the American Statistical Association 114.528 (2019): 1708-1725.

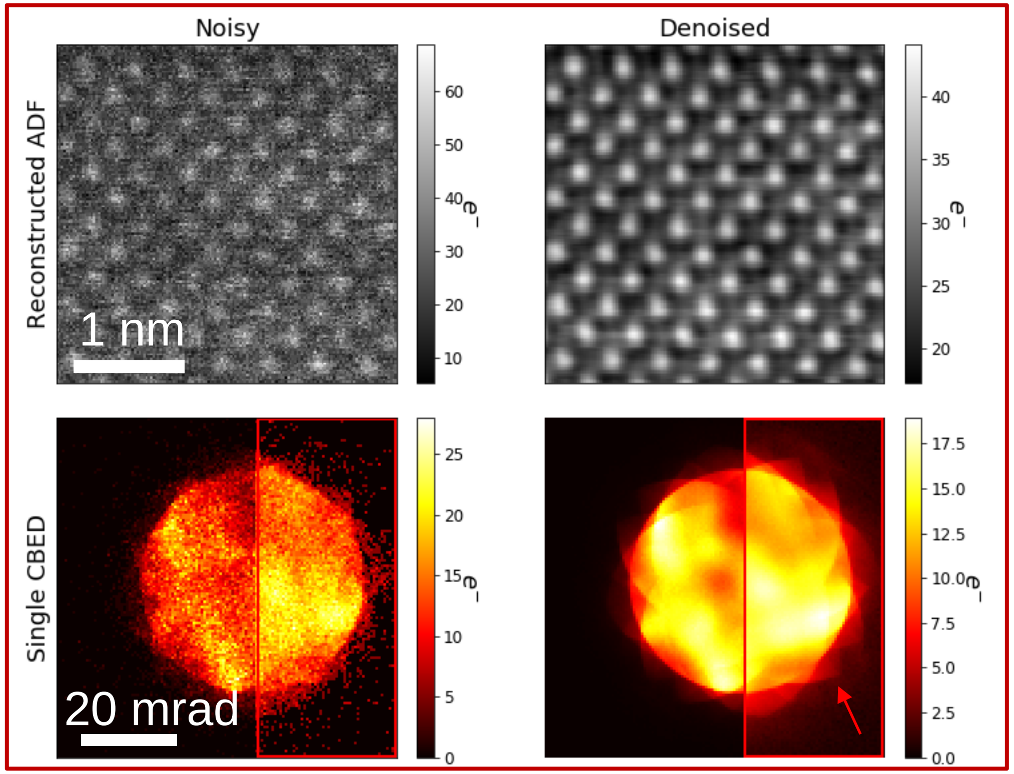
Tensor SVD Performance: Simulated 4D STEM

Simulations: Si [110] dislocation core



Input size: 1.6 GB Processing time: 525.6 sec

Experiments: $SrTiO_3$ [100]

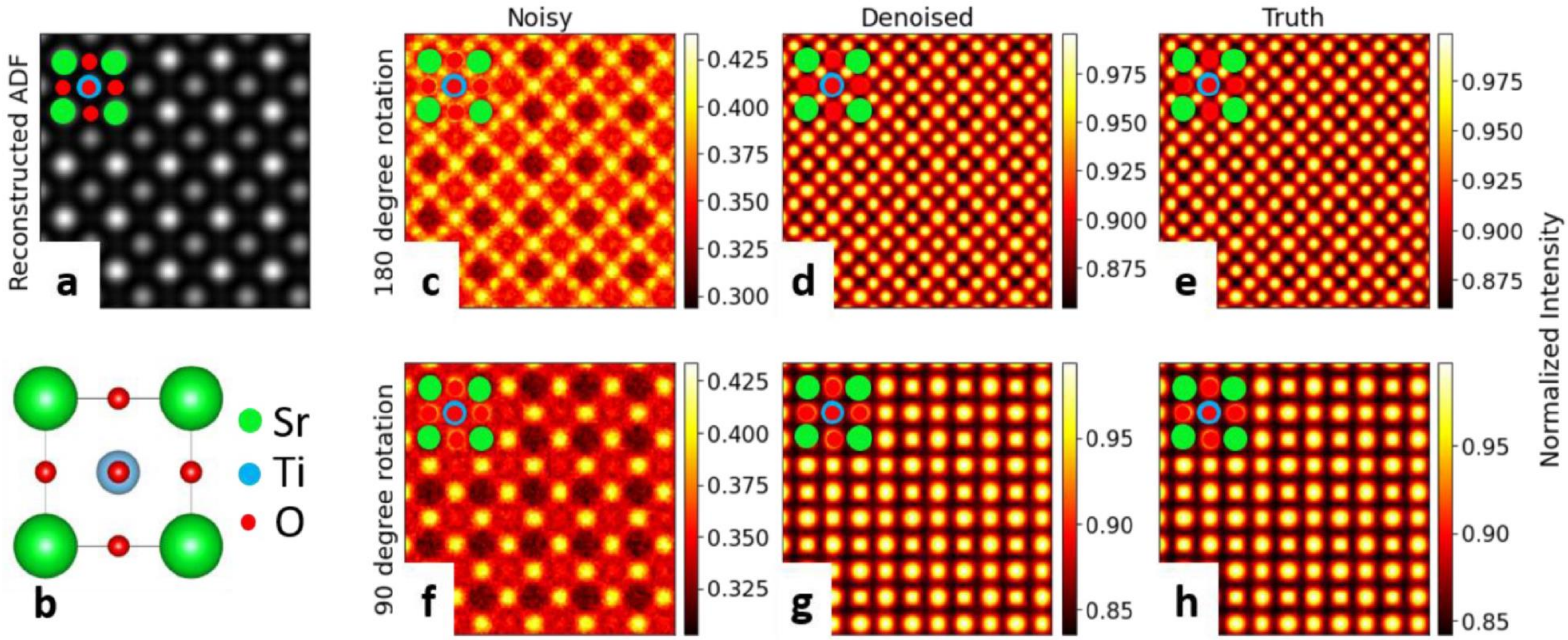


Input size: 2.8 GB Processing time:538.9 sec

- Processing time on a desktop with moderate computing power (single Xeon E5-2603 CPU).

C. Zhang, *Ultramicroscopy* **219**, 113123 (2020); DOI: 10.1016/j.ultramic.2020.113123

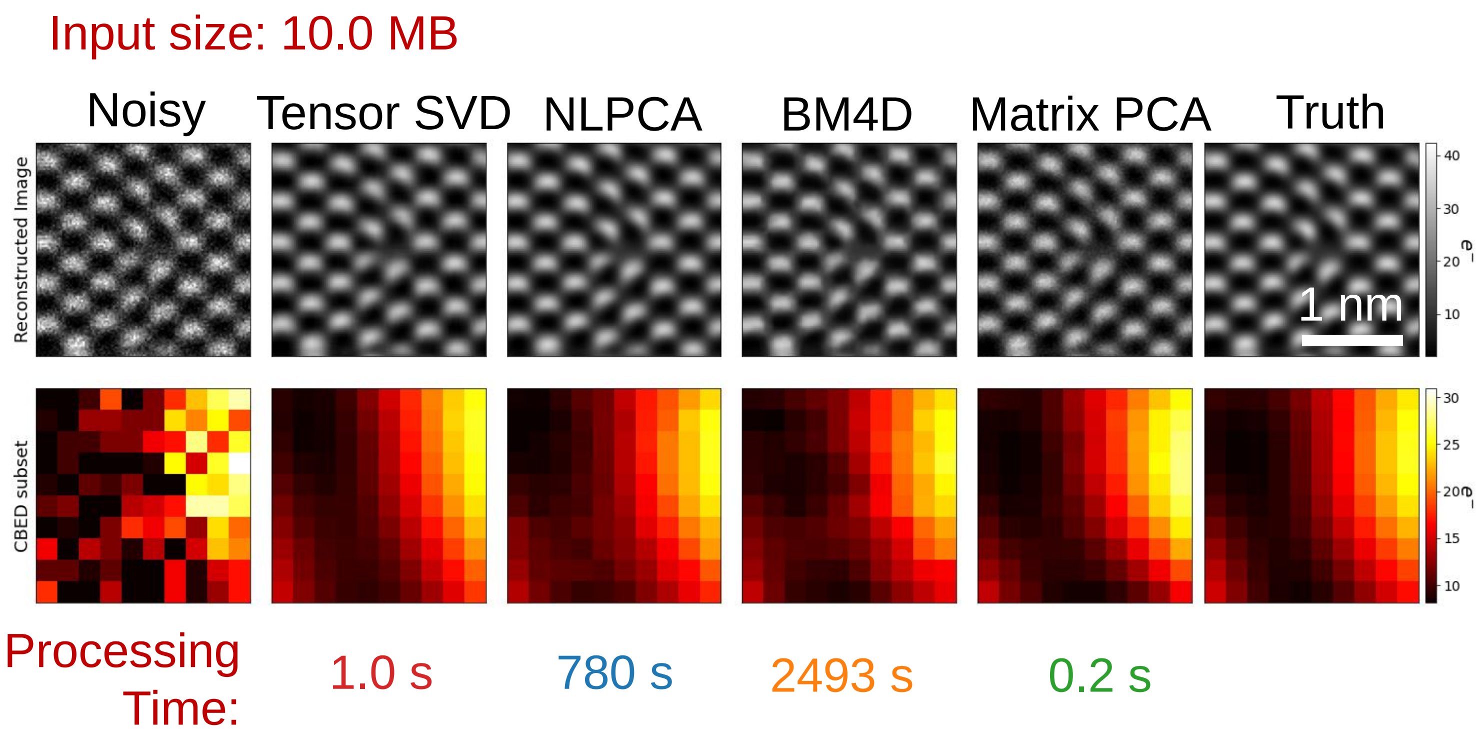
Tensor SVD Improves Symmetry Information

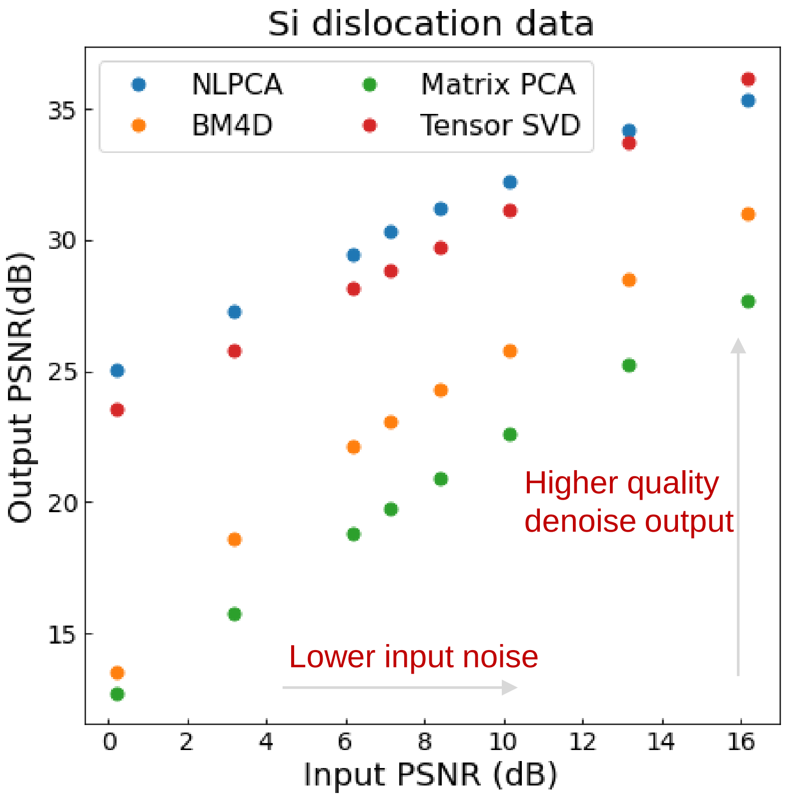


- Symmetry STEM is a new method to extract crystallographic point symmetries from 4D STEM data.
- Noisy 4D STEM data do not report the correct 4-fold symmetry for Sr sites, but denoised data do.

M. Krajnak, J. Etheridge, *Proc. Natl. Acad. Sci.* **117**, 27805–27810 (2020).

Comparison to Other Denoising Methods





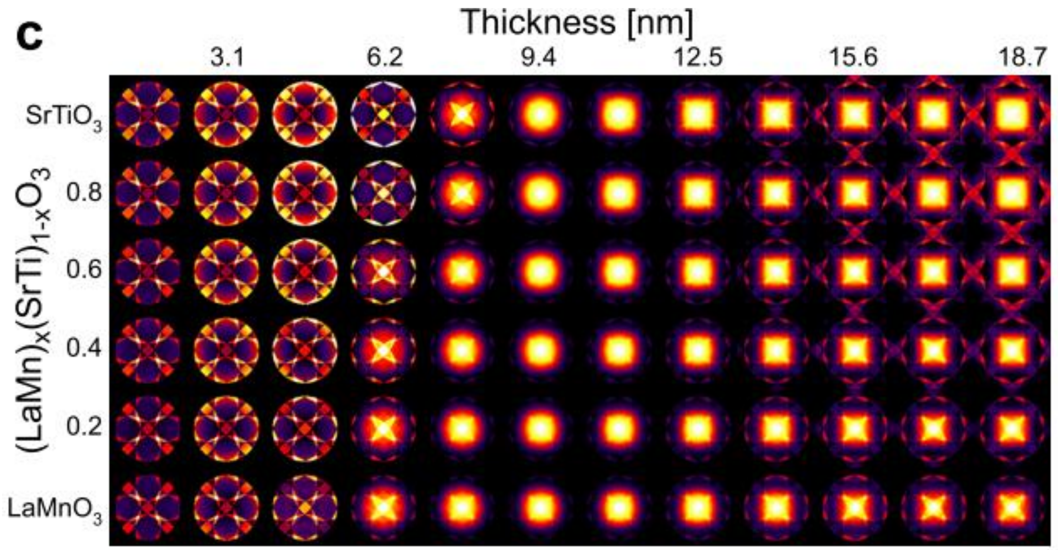
- Tensor SVD is tested against non-local principal component analysis (NLPCA), block matching and 4D filtering (BM4D), and matrix PCA.
- Tensor SVD has the best or close to the best denoising performance.
- Tensor SVD is fast and suitable for multi-GB hyperspectral data.

Nanotechnology 27.36 (2016): 364001.
IEEE transactions on image processing 22.1 (2012): 119-133.

Supervised Learning with Neural Networks

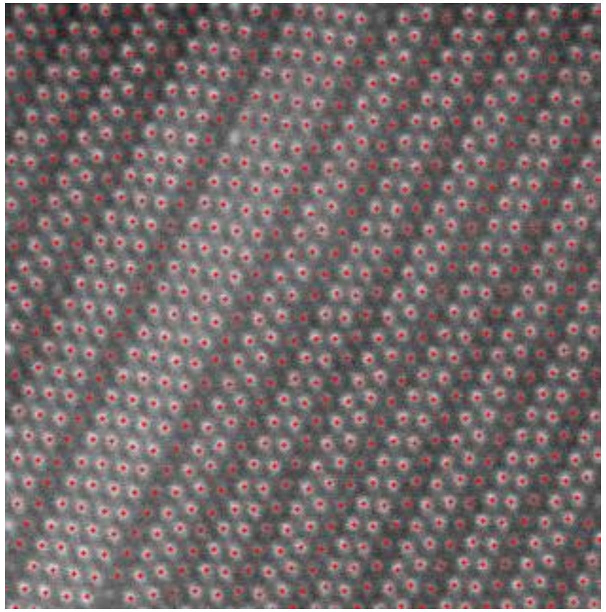
- Prior knowledge is example data, labeled with the result of the analysis
- For STEM, training data can come from simulations
- Limits of the resulting network are not very well determined

Determining sample thickness from 4D STEM data

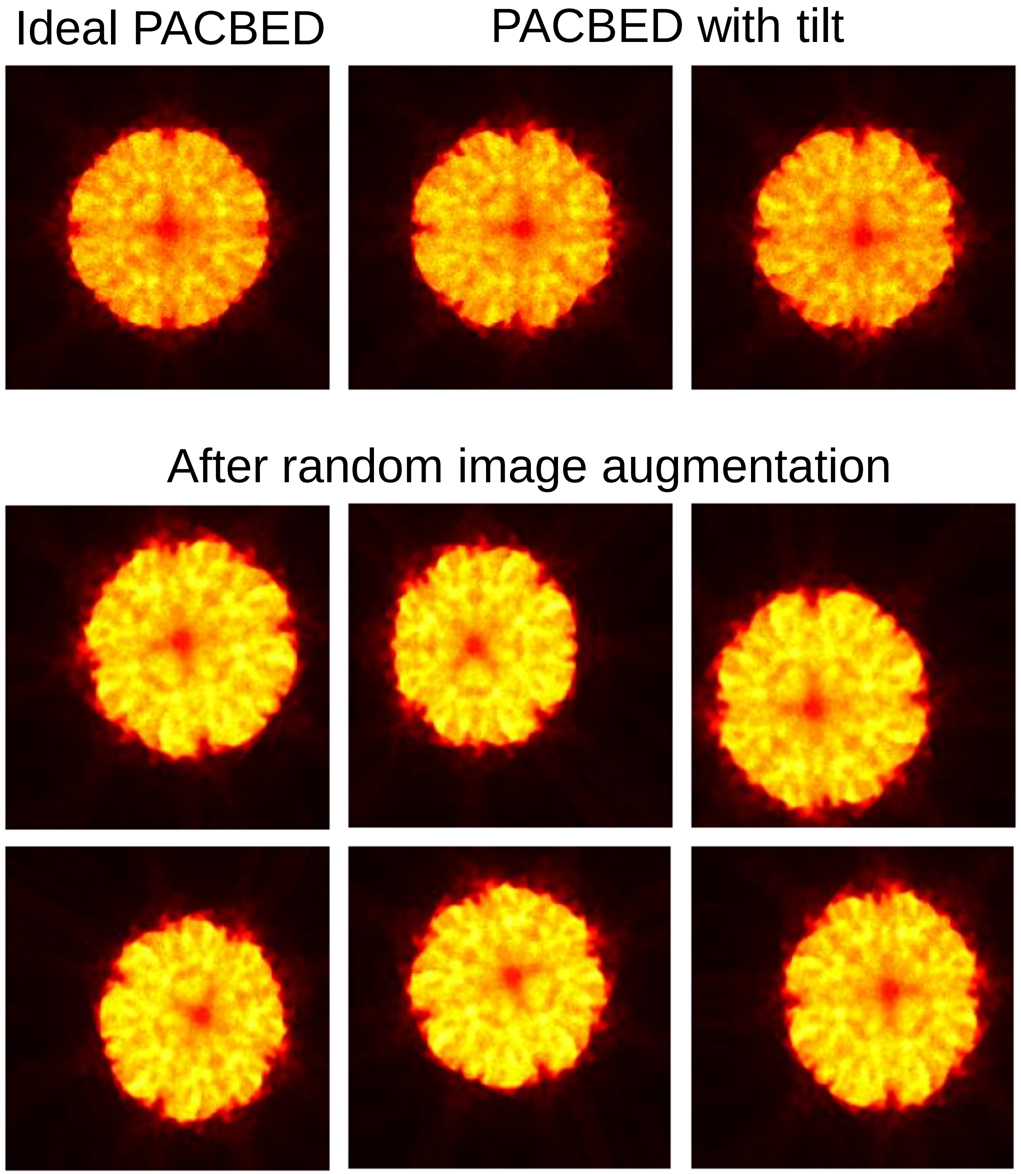


CBED patterns changing with sample composition and thickness. C. Ophus, *Appl. Phys. Lett.* **110**, 063102 (2017).

Finding atomic column locations in HRSTEM images

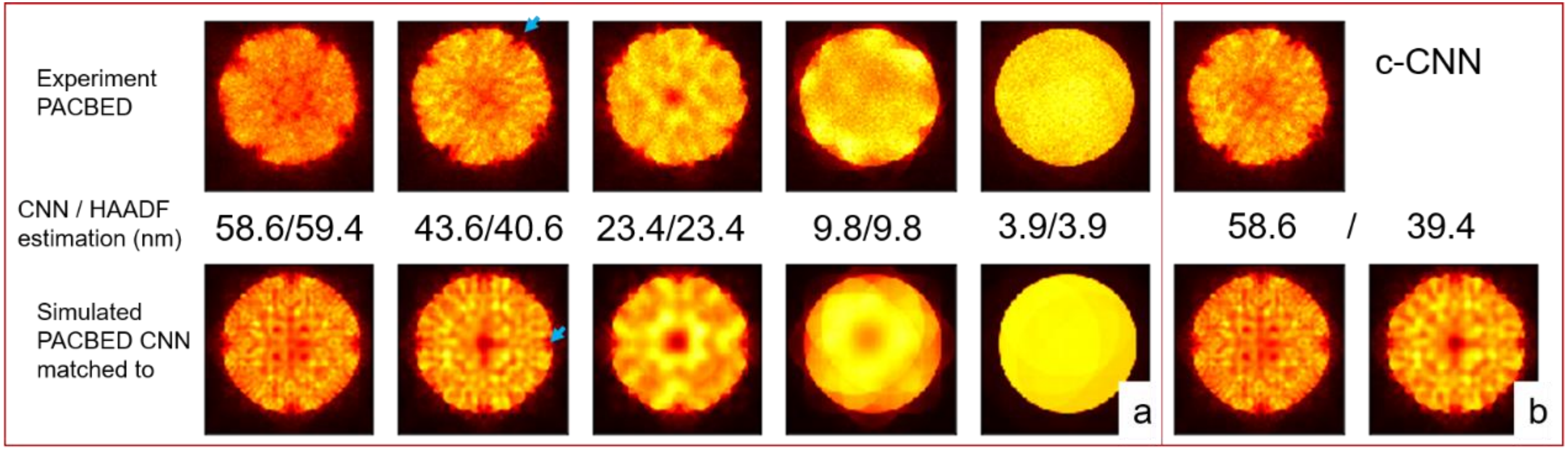


Simulated Training Data for CNN

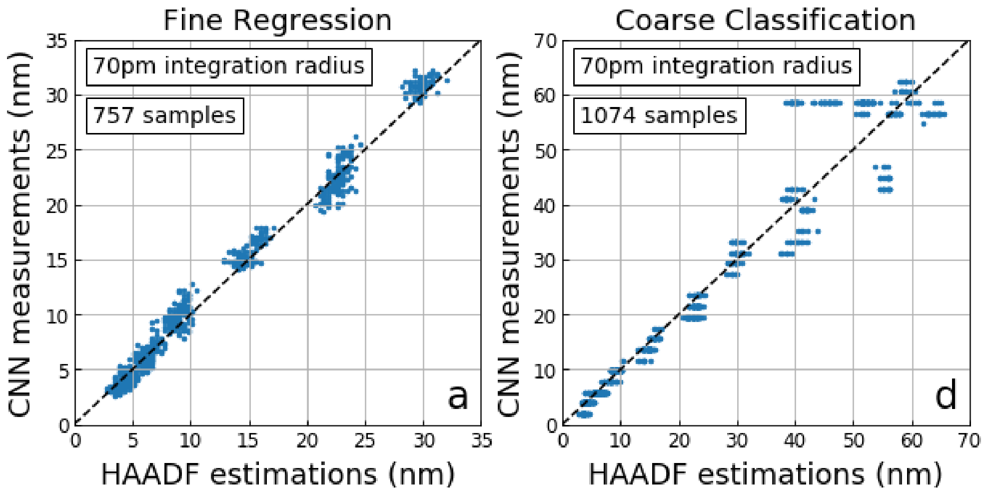


- **Use multislice simulations to generate automatically labeled training data:**
  - Cover a wide range of possible experiment conditions in simulations, including thickness but also crystal tilt
  - Augment the images after simulations by adding noise, and distortions including shift, zoom, rotation, shear, *etc.*
- Transfer learning:
  - use a vgg-16 network pretrained to recognize features in natural images
  - retrain just the fully connected final layers at first, then tweak the convolutional layers only at the end
- Full training data set is about 750 GB

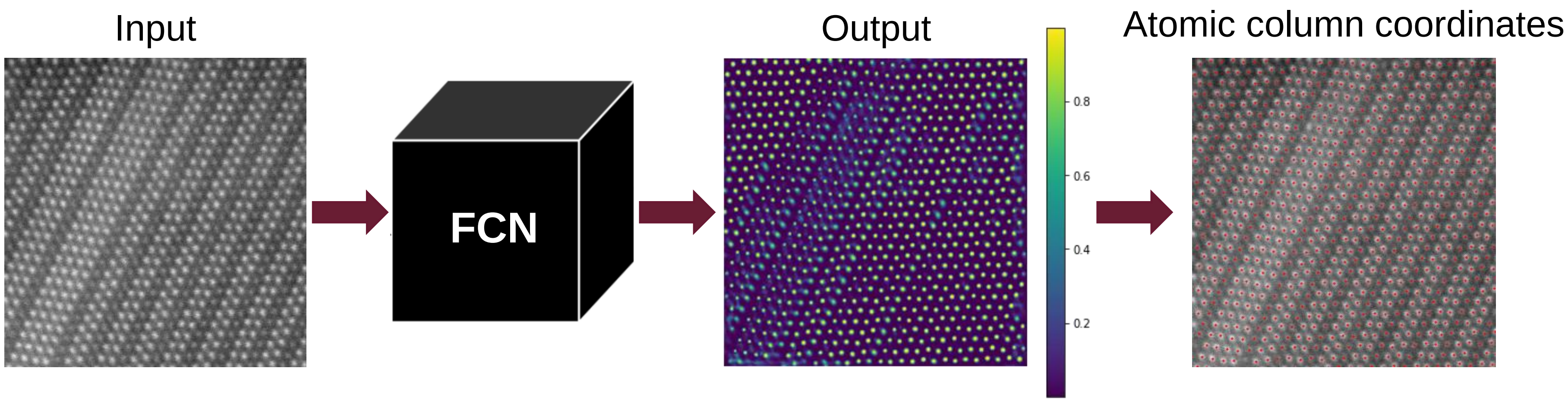
Feature Identification Successes and Failures



- RMS deviation for experimental data on the same crystal and orientation is ±1 nm
- Larger thicknesses and more complicated image features work less well.
- Network fails for thicknesses outside training data set.
- Network fails for other crystals or even other orientations of the same crystal.



Atom Finding: A Common Problem



Lin's AtomSegNet

Scientific Reports 11.5386 (2021): 1-15

- Functionalities: **atom segmentation**, noise reduction, background removal, and super-resolution processing.
- Trained on 15 crystal lattices (e.g. $SrTiO_3$, graphene)

Ziatdinov's AtomNet

ACS Nano 11.12 (2017): 12742–12752

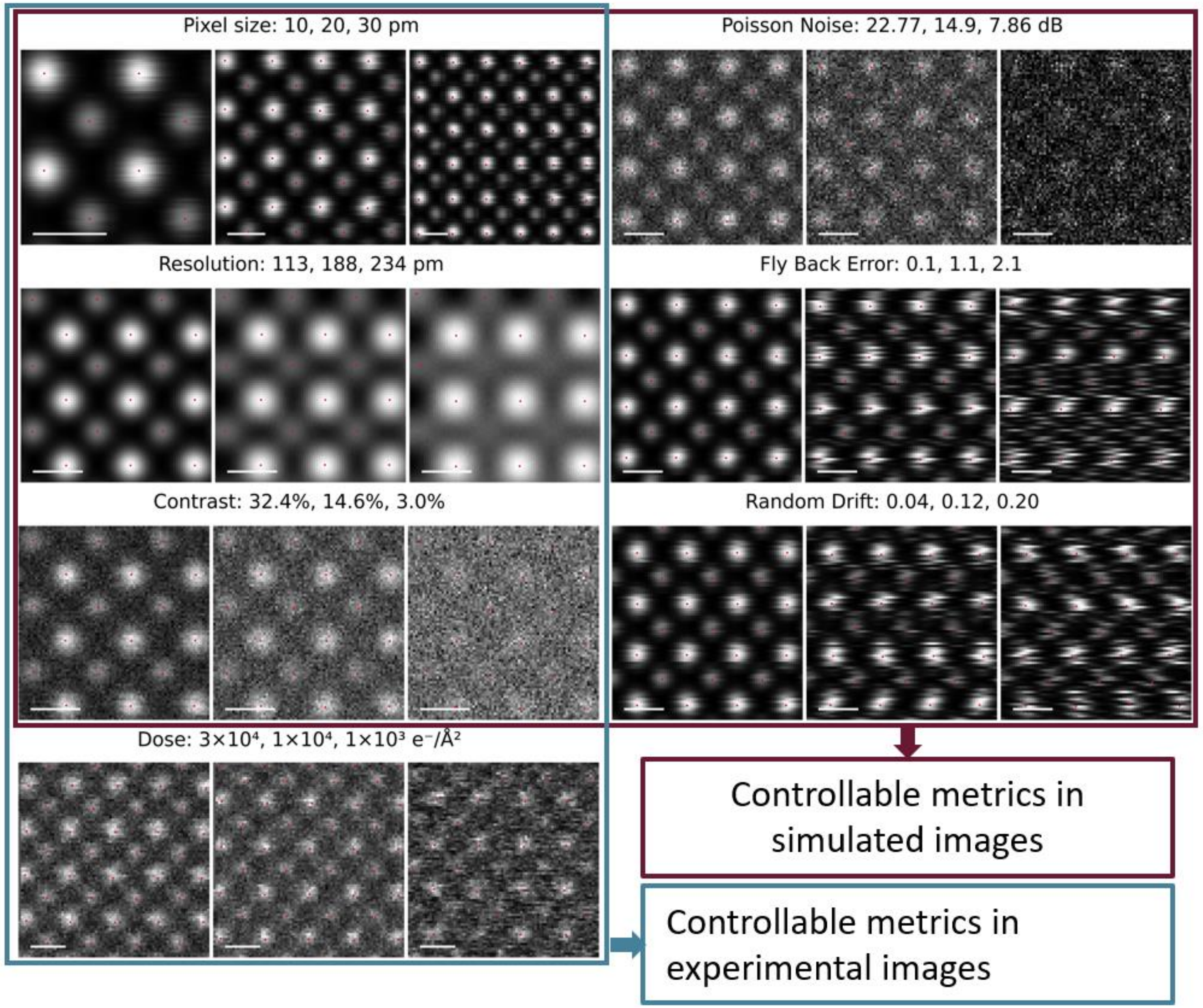
- Functionalities: **atom segmentation**, detecting atom species and defects.
- One trained on crystal lattices (e.g. $SrTiO_3$), another trained on hexagonal lattices.

Ziatdinov's AtomAI

https://github.com/ziatdinovmax/atomai

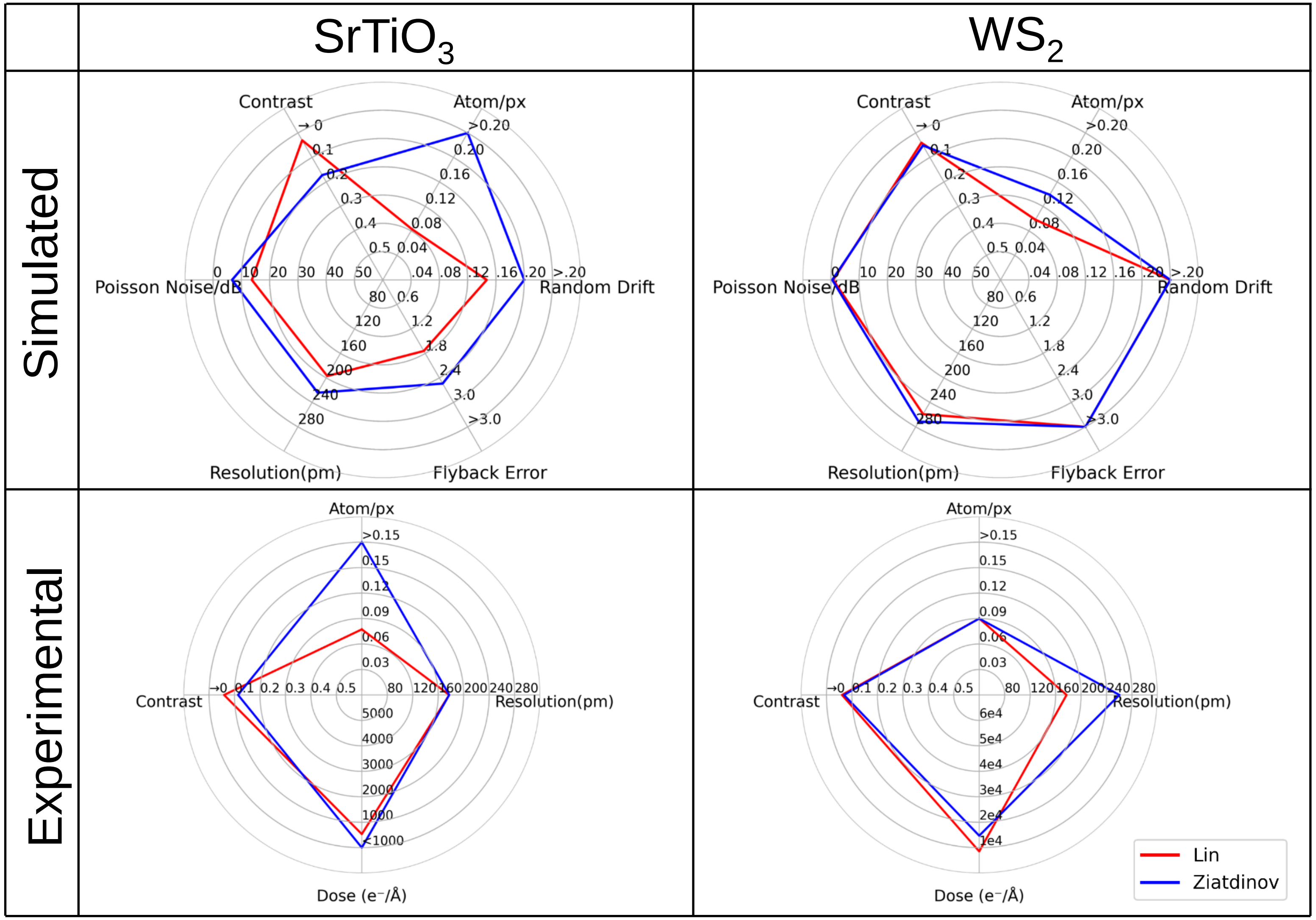
- Pytorch-based package for training new models for new problems
- We trained a new U-net model on 5 crystal lattices using AtomAI

Which Model is "Best"?



- We wanted to use the best network with the least investment of time, but we found no way to evaluate network performance outside their own training and test data.
- Created a benchmark data set with varying image quality:
  - In simulations, vary pixel size, contrast, Poisson noise level, scan distortion
  - In experiments, vary pixel size, spatial resolution, electron dose
  - WS_2 and $SrTiO_3$
- ~40 experimental images of various crystal lattices, defects, interfaces
- DOI: 10.18126/e73h-3w6n

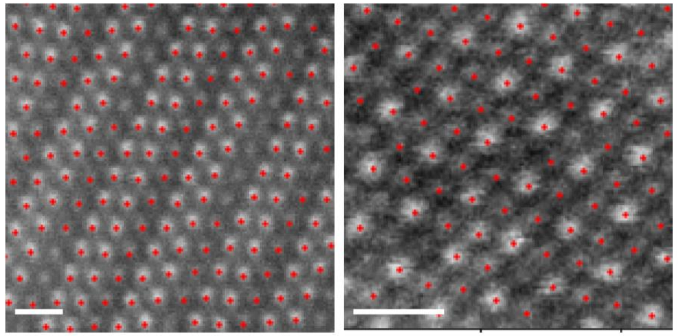
Model Performance vs Image Quality



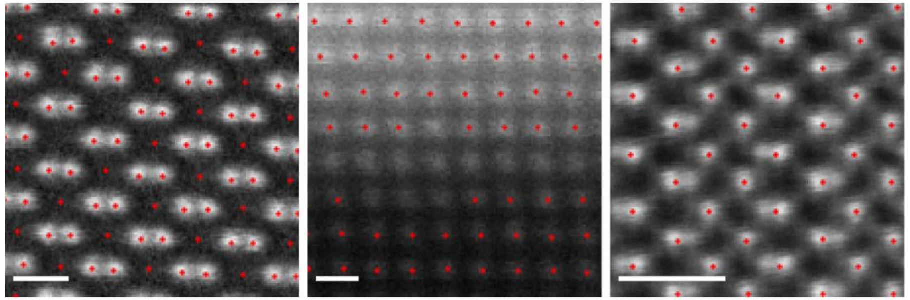
- Define acceptable performance as ID-recall > 0.90, ID-precision > 0.95 and Δd < 0.3 Å
- Larger blue polygons Ziatdinov's model is more forgiving of poor image quality
- Potential trade-off between model overall performance and general applicability

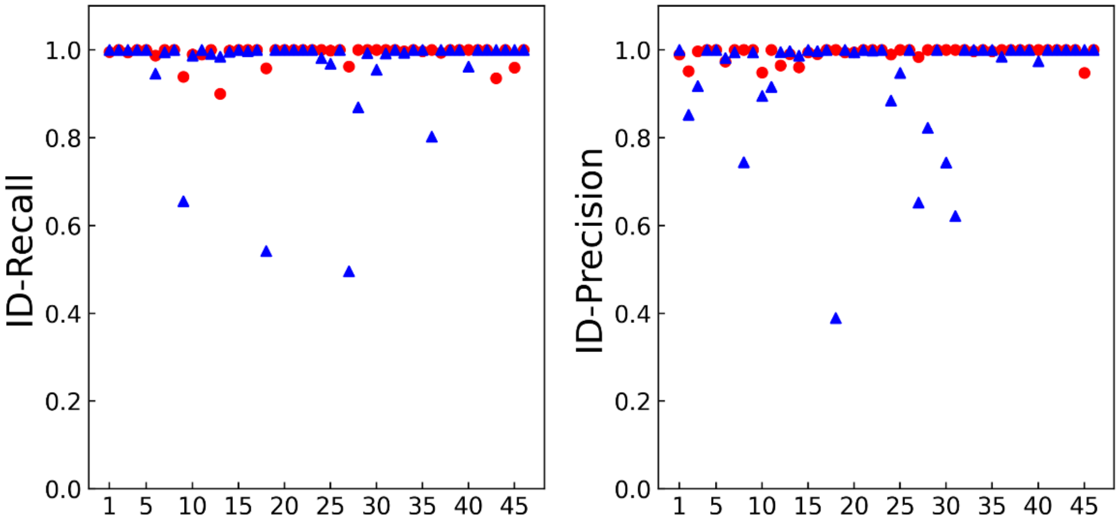
Applicability Outside Training Crystals

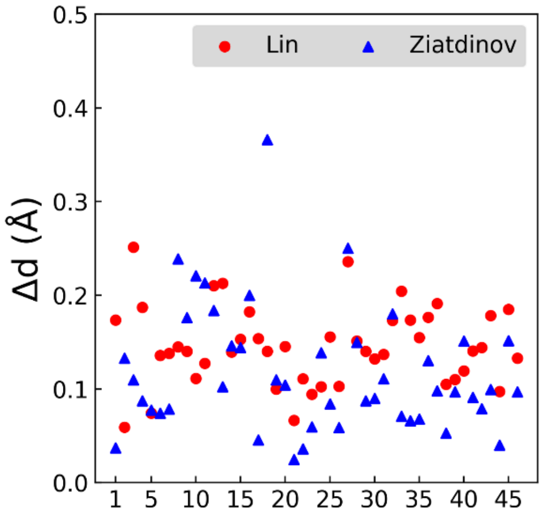
- Models applicable for most crystal lattice, defects, interface, etc.
- Poor cases for Lin's model due to low SNR



- Poor cases for Ziatdinov's model including FPs in background, TNs in areas of varying contrast and overlapping atoms.

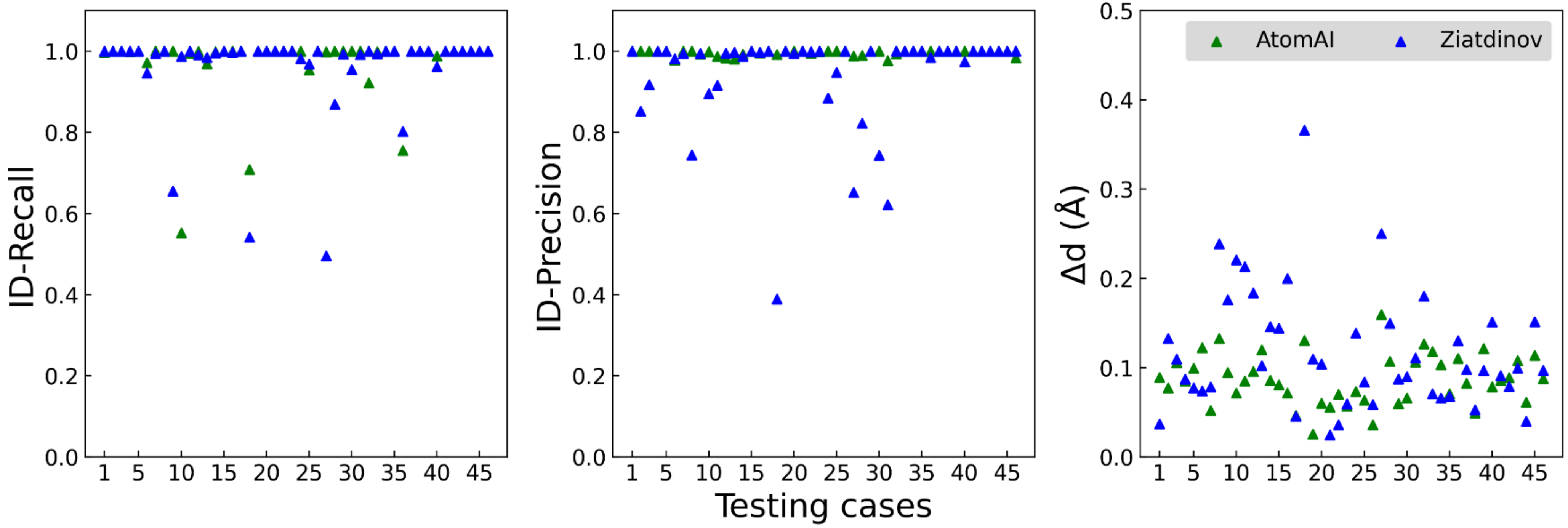






Toward a More General Network

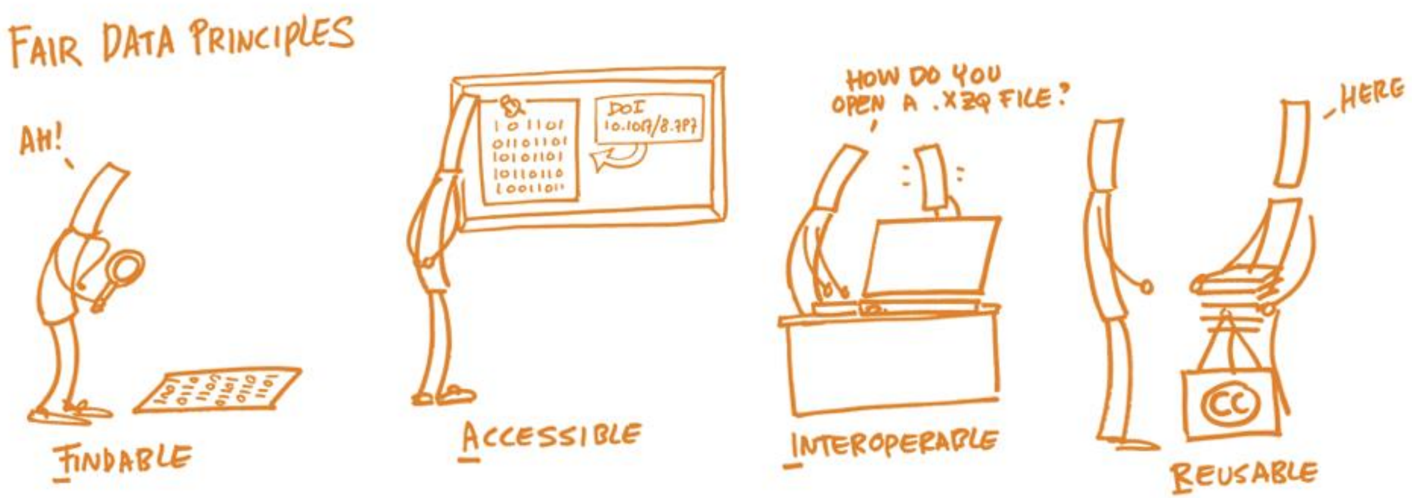
- Used the AtomAI framework to train a network on simulated images from 5 crystal lattices, plus augmentation
- More general than Ziatdinov model while maintaining robustness against image quality.



Making Materials ML FAIR

- Findable
- Accessible
- Interoperable
- Reusable

- At least in my corner of materials science, ML models are not FAIR
  - Easy find (Github) but hard to run
  - Prior knowledge / training data is often unspecified or unavailable



M. D. Wilkinson, The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* **3**, 160018 (2016). DOI: 10.1038/sdata.2016.18

https://www.force11.org/group/fairgroup/fairprinciples

FAIR Data and Models

- Tools for distributing data like figshare and Materials Data Facility are well developed
  - Non-rigid registration: 10.6084/m9.figshare.12592466.v1
  - Non-local denoising: 10.6084/m9.figshare.12592457.v1
  - tensor SVD: 10.18126/vh9q-i1l6
  - 4D STEM CNN: 10.18126/4nm2-0g70
  - Atom finding test data: 10.18126/e73h-3w6n
- Need to be more widely used

- Tools for software exist and are widely used
  - NRR and tensor SVD have python modules compatible with HyperSpy
  - Non-local denoising and the 4D STEM CNN are available on Github
- How often does research-grade software off Github actually work to solve a problem?
- How often can you test the software on the data used to develop it?

Foundry Infrastructure for Materials ML

FOUNDRY
DATA, MODELS, SCIENCE

- Containerized ML models permanently associated with data sets
- Radically reduced barriers to reuse, meta-studies, benchmarking, and more
- Atom finder dataset available now
  - DOI: 10.18126/e73h-3w6n
  - Standard dataset description interface
  - Queriable format (hdf5)
  - Highly accessible metadata

Consumers

Science!

```
From foundry import Foundry
f = Foundry()

X,y = f.load(“dataset1”, v=“1.0”)
y_pred = f.run(“model1”, v=“1.0”, X)
```

- Models run locally or on distributed endpoints
- Capabilities to pull datasets to desired location or move compute to desired location

Dataset

Function

API layer

API layer

Data Publishers

Data Provider

Models / Functions

Model Publishers

```
f.data.publish(“./”
“dataset1”, v=“1.1”)
```

```
f.model.publish(“./”
“model1”, v=“1.1”)
```

MATERIALS DATA FACILITY

NIST CHiMaD

DLHub
Data and Learning Hub for Science
https://www.dlhub.org

U.S. DEPARTMENT OF ENERGY

Argonne NATIONAL LABORATORY

Dane Morgan, Paul Voyles, Michael Ferris, Marcus Schwarting, **Ben Blaiszik**

Summary

- STEM data are growing in rapidly in size and complexity
- ML / AI methods are essential and developing quickly
- Example applications:
  - distortion correction
  - non-local denoising
  - low-dimensional representations for tensor data
  - determining sample characteristics directly from 4D STEM data
  - atom finding in high-resolution images
- Data and models are all available from the bibliography at tem.msae.wisc.edu
- ***TANSTAAFL and make it FAIR***