THE UNIVERSITY of TENNESSEE KNOXVILLE

Lecture 14: Clustering for imaging and spectroscopic data

Instructor: Sergei V. Kalinin

Spectroscopic Imaging

Advancements in imaging led to a broad spectrum of the spectroscopic imaging techniques, in which response spectra are measured in each spatial location giving rise to 3- and higher dimensional data.

Scanning probe microscopy:

- Force-distance curve measurements
- Current-voltage measurements

Electron microscopy:

- Electron Energy Loss Spectroscopy

Optical microscopy:

- Hyperspectral imaging
- Time resolved measurements

Mass-spectrometry:

- Secondary ion MS imaging

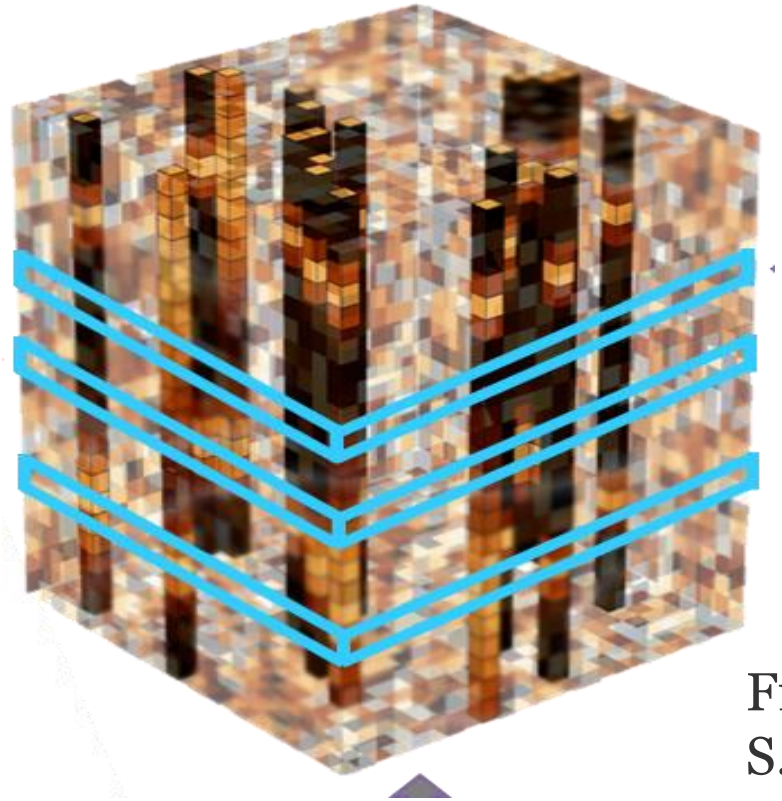
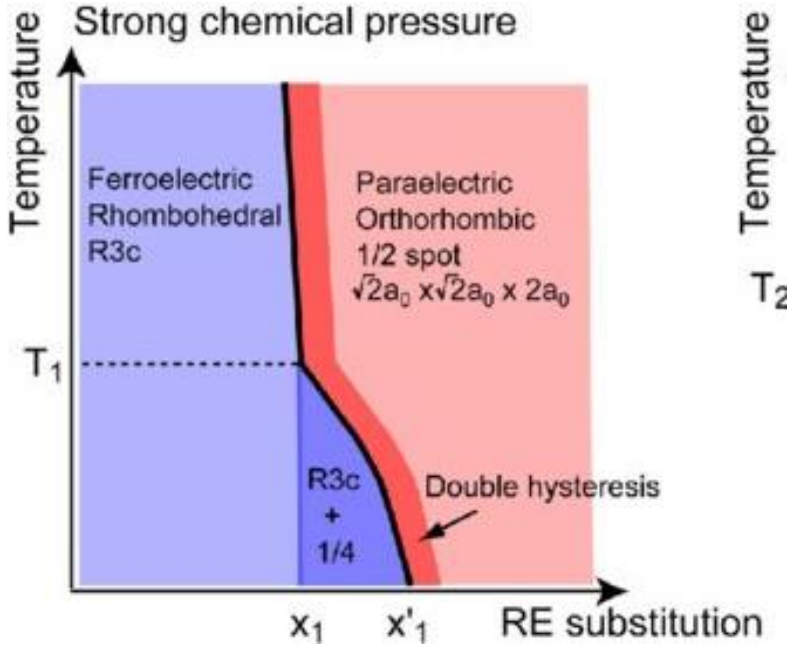
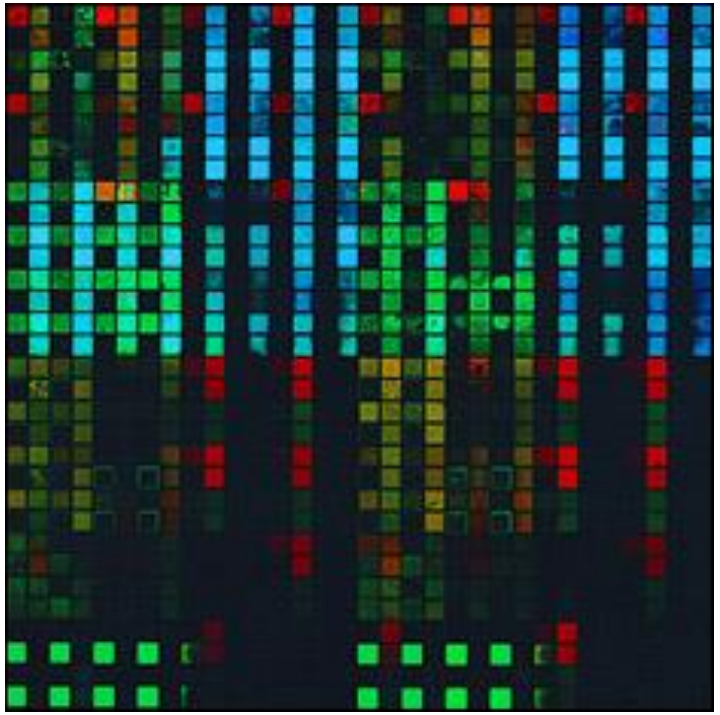


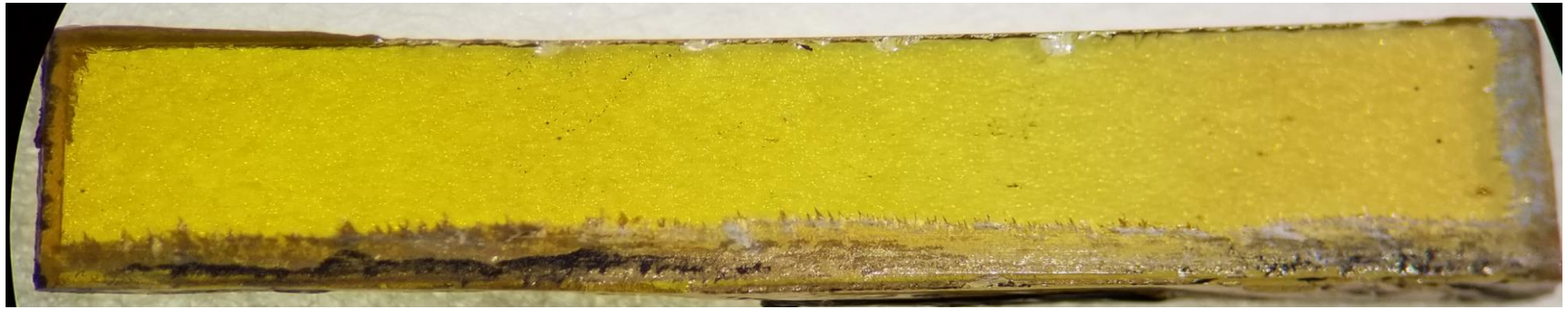
Figure by S. Jesse

Combinatorial Libraries



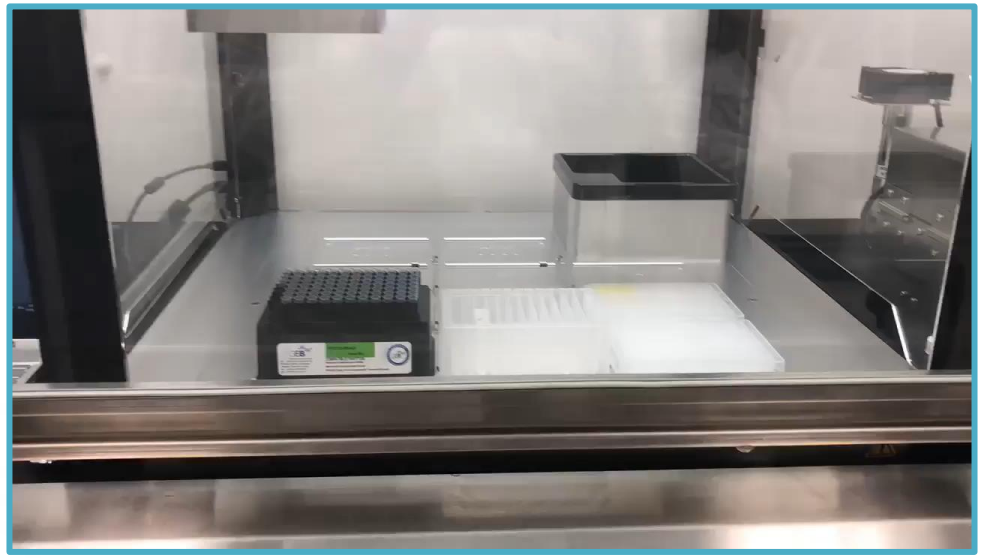


https://mse.umd.edu/research/spotlight/combinatorial

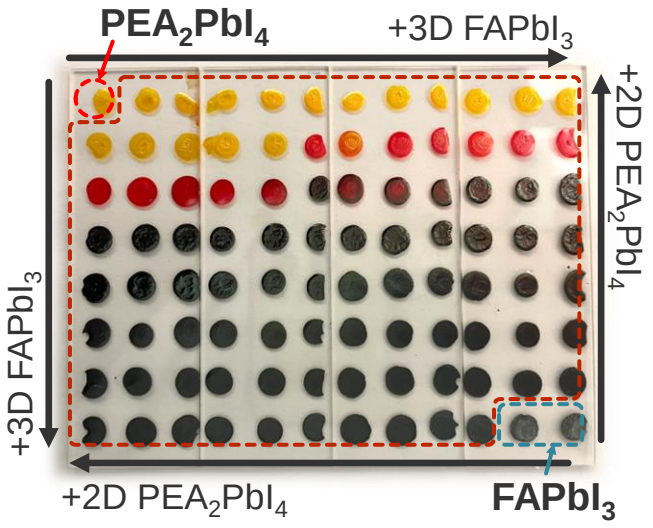


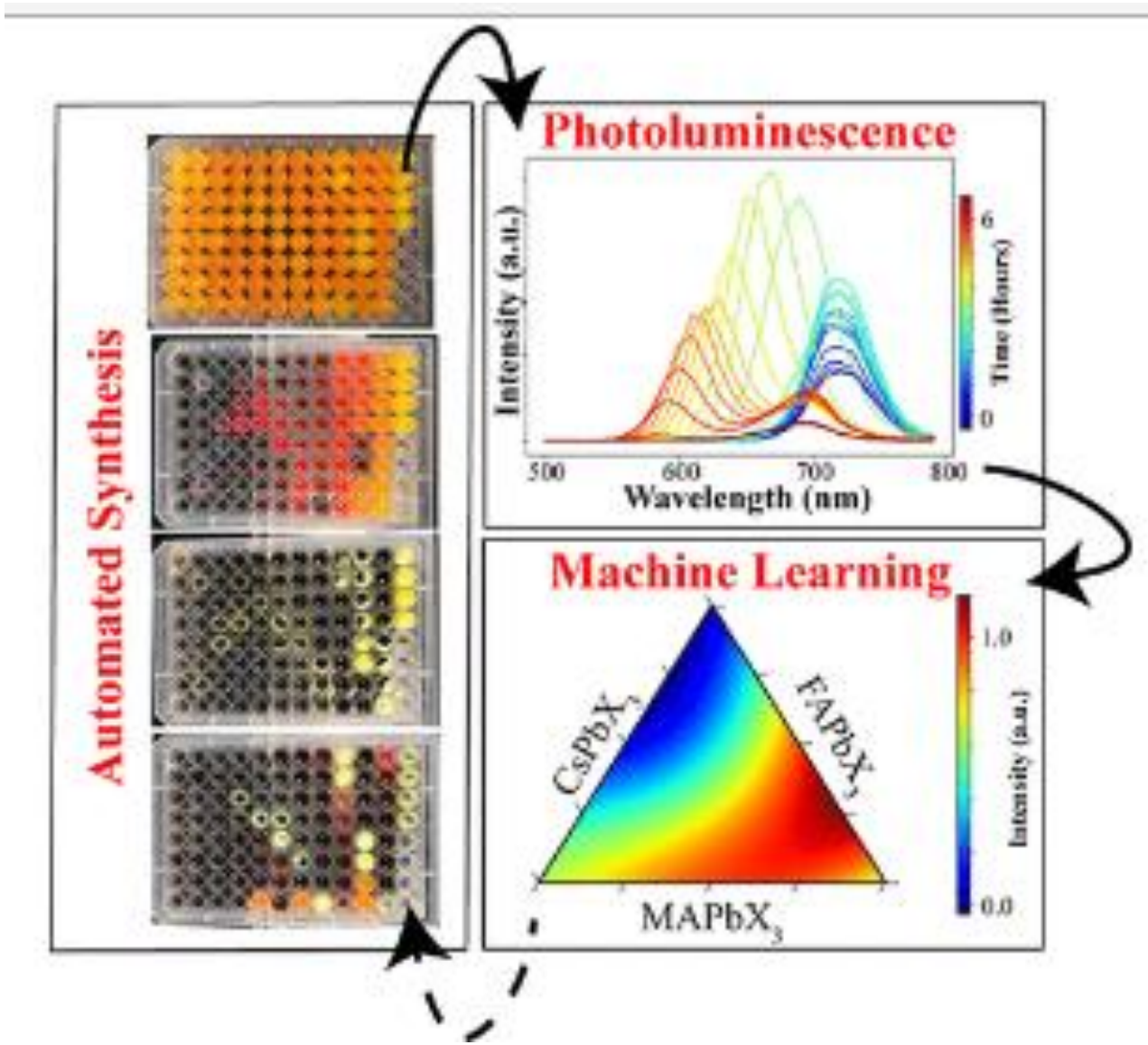
$BiFeO_3$ Linear est. 7%Sm $BiFeO_3$ 20%Sm $BiFeO_3$

Combinatorial Libraries

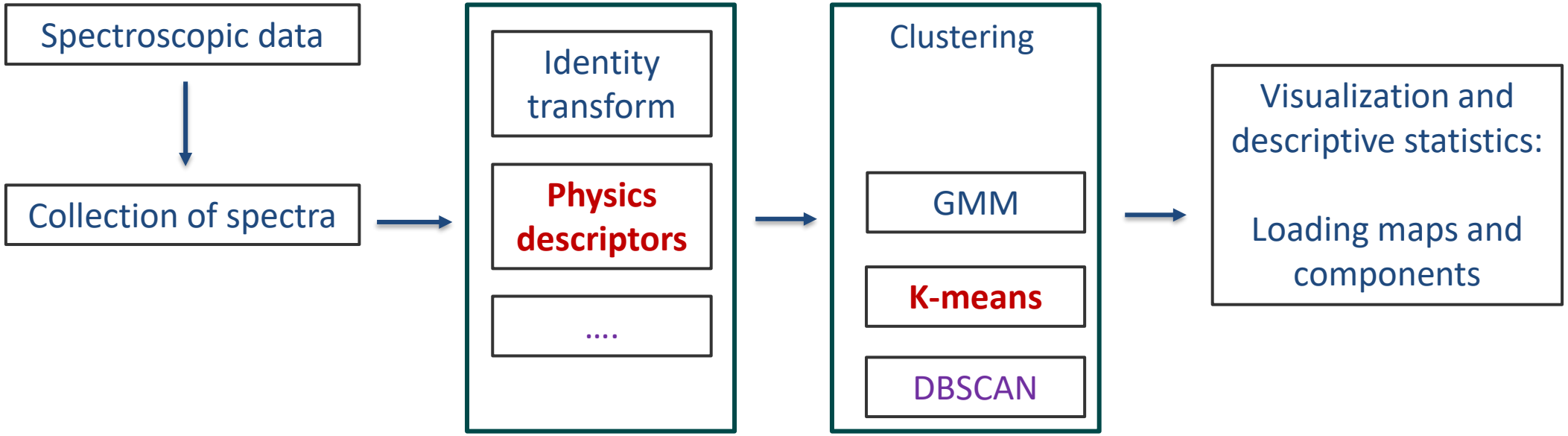


Quasi-2D MHP Film Array





Analysis pipeline

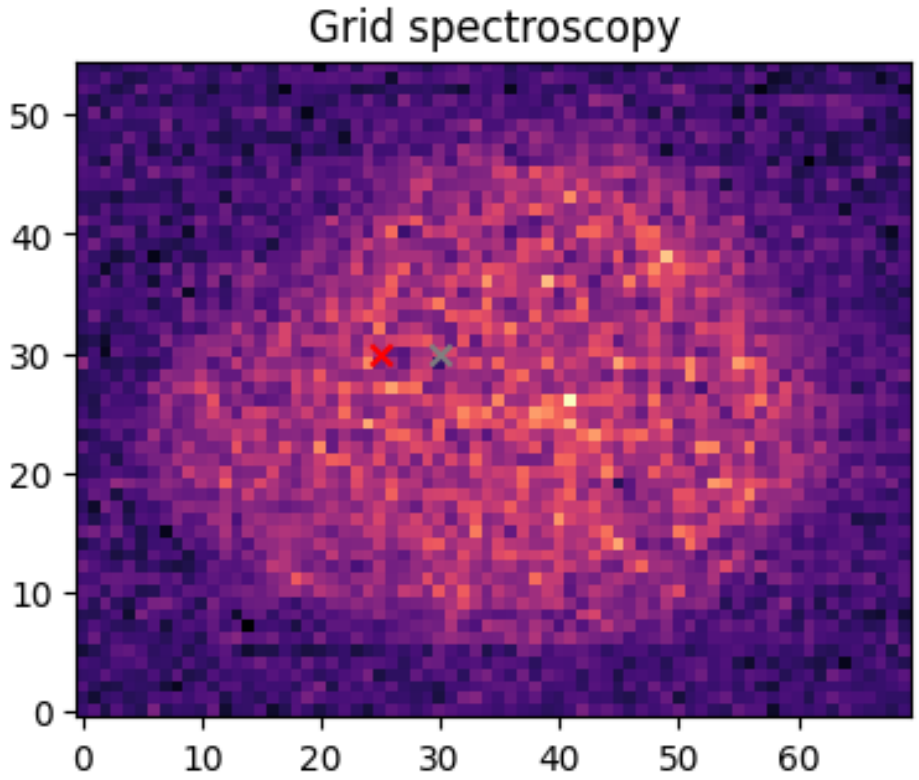


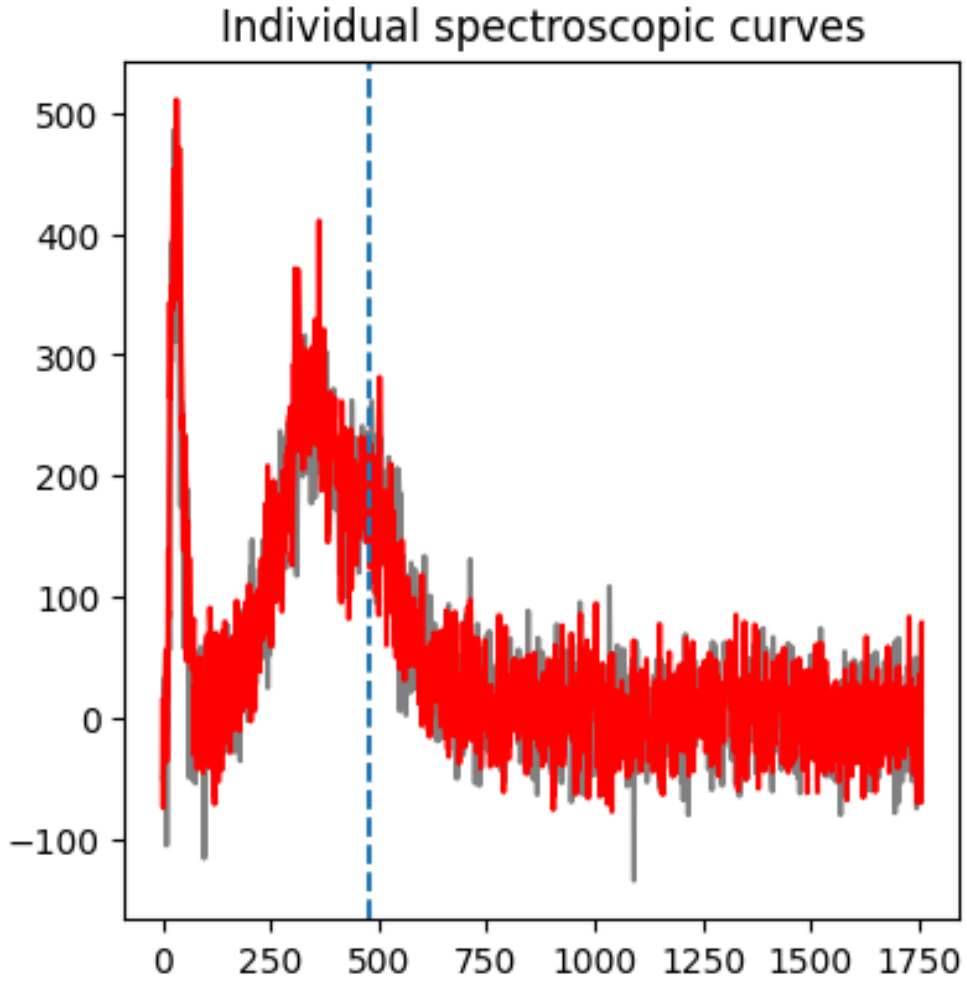
Pipelines are defined to

- Make analysis traceable, repeatable, explainable, and transferable
- Allow for hyperparameter tuning and optimization
- Efficiently use the memory

EELS Colab

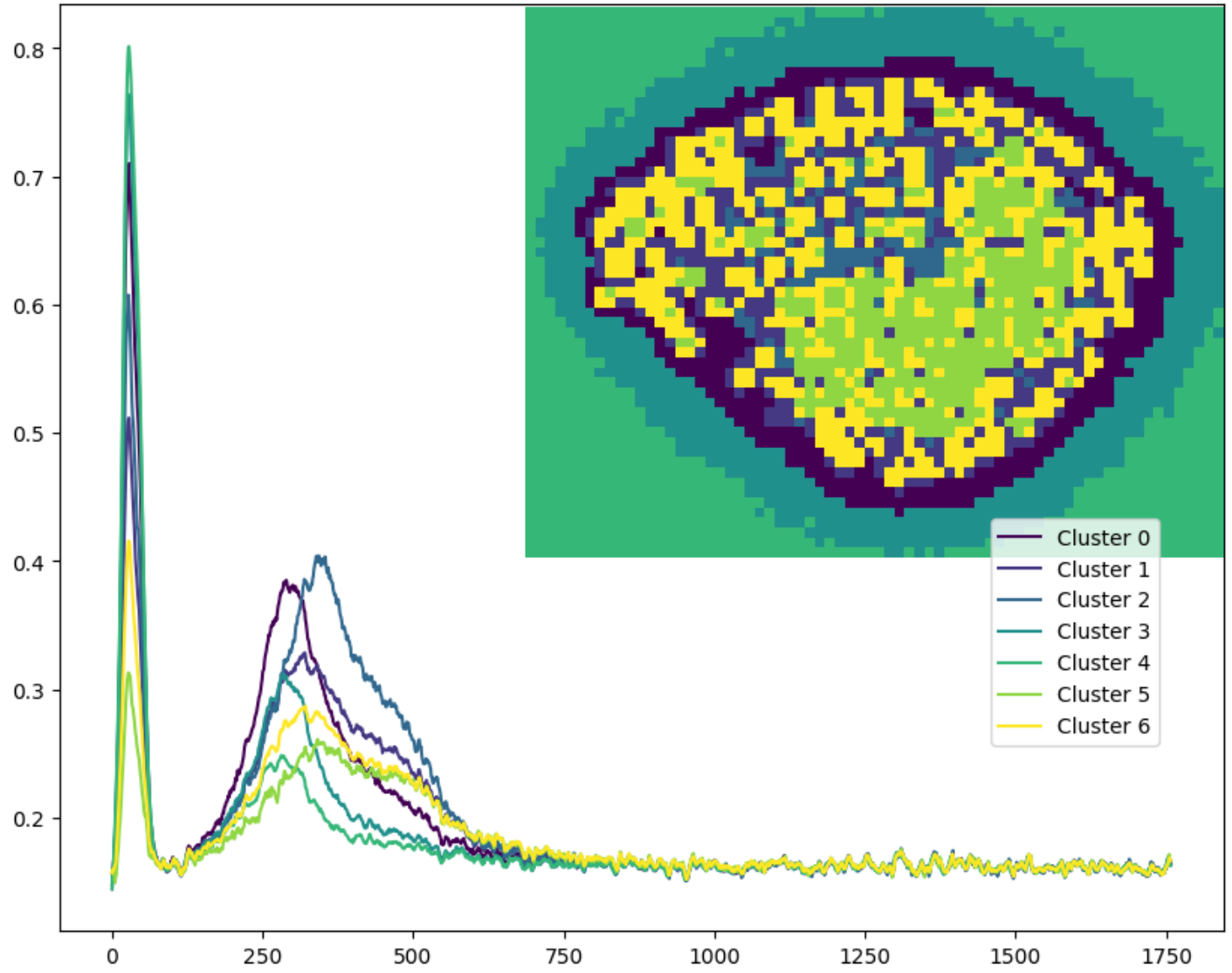
Clustering of spectral data





- The hyperspectral data set contains spectrum at each spatial position on the dense rectangular grid
- We use clustering to establish internal structure of this dataset

Clustering of spectral data

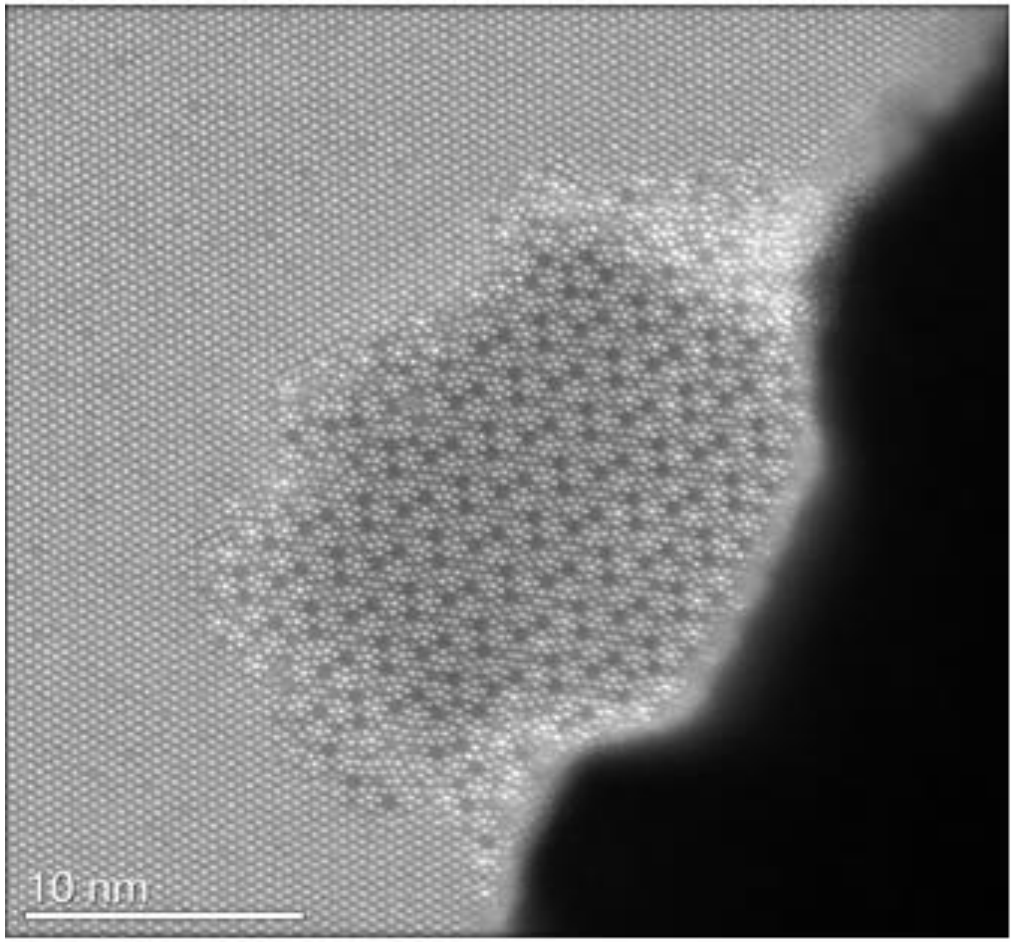


- Experiment with number of clusters
- Based on domain experience, explore the behavior of the components and images of class labels
- This is already “real” research

But what about images?

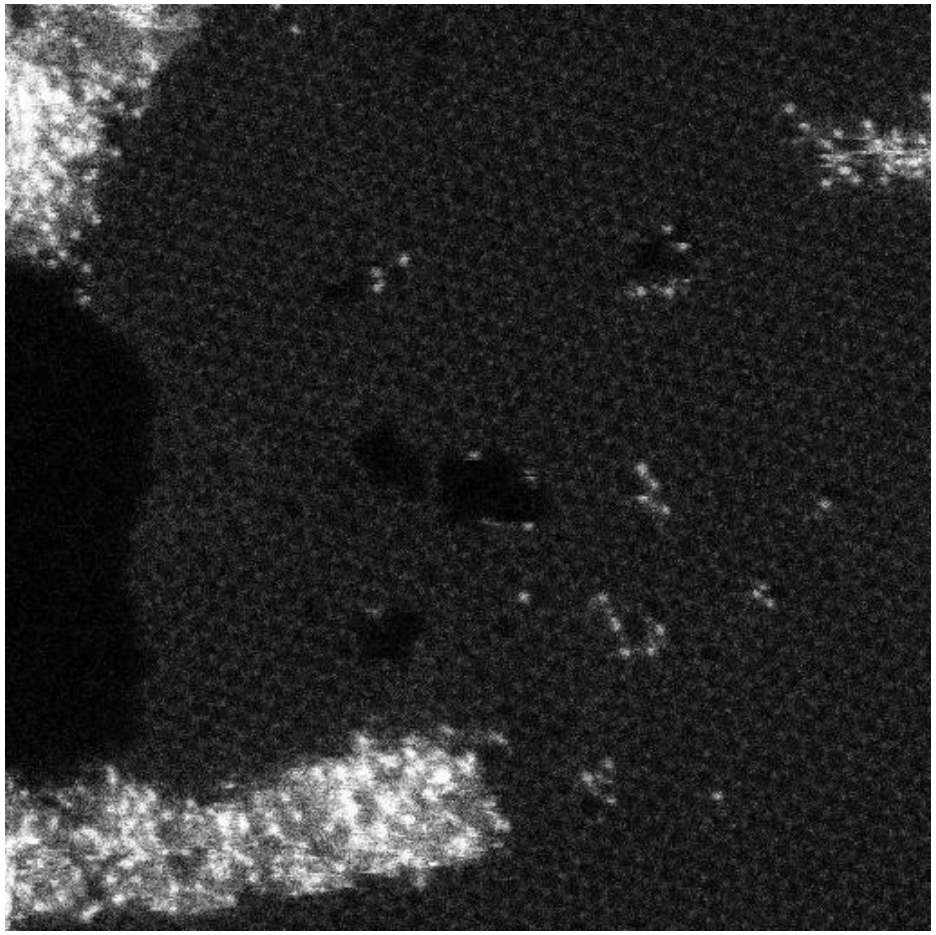
Chemically disordered systems

Mo-V-Ta complex oxide



Q. He et al, ACS Nano 9, 3470-3478

Si in graphene

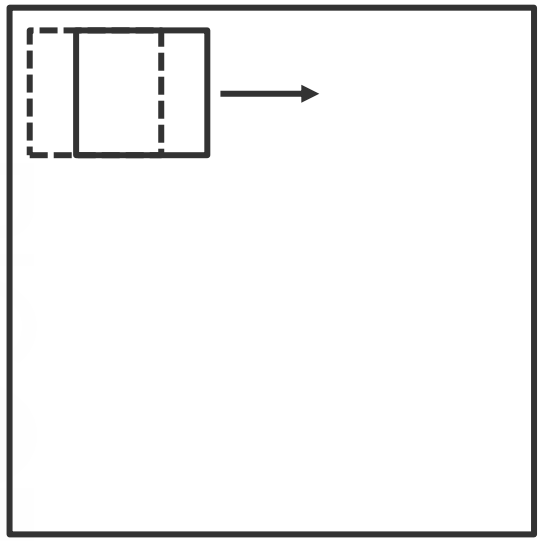


Data collected by O. Dyck (ORNL)

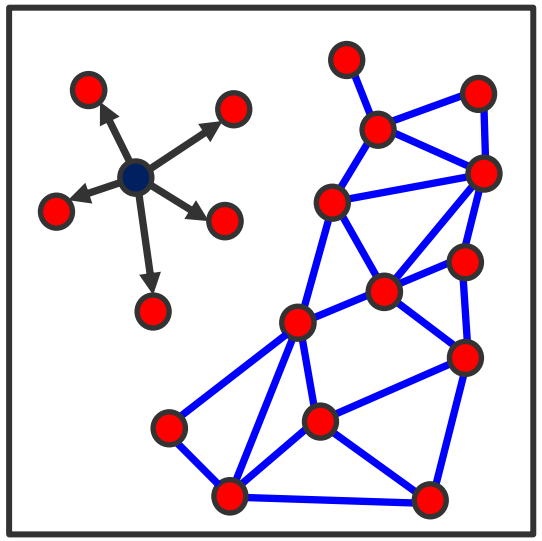
- What is the nature of the building blocks and relevant atomic configurations?
- Can we define single-phase regions and phase boundaries?

Constructing the descriptors

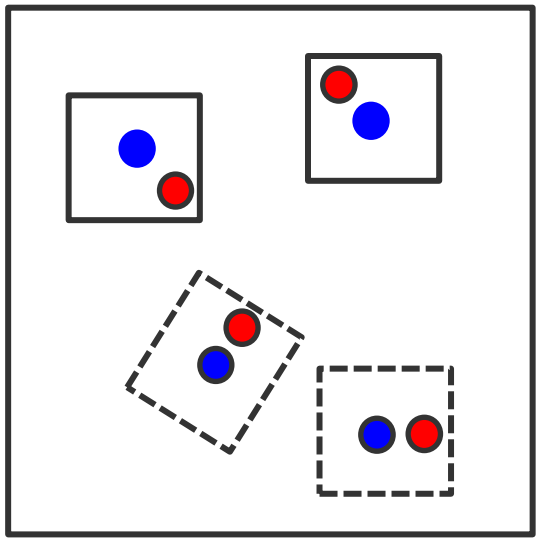
Continuous translational symmetry



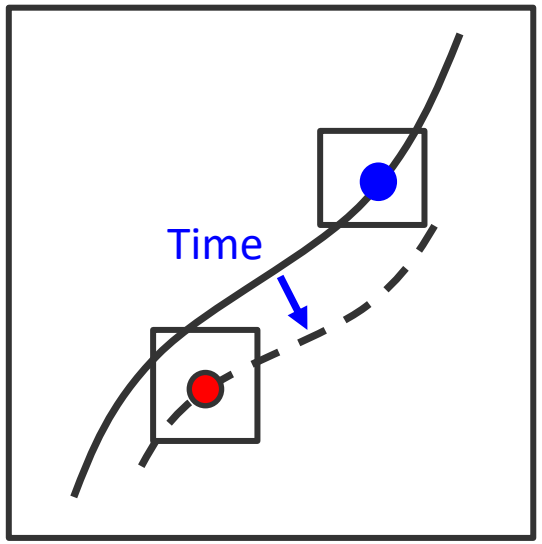
Atom based descriptions



Localized sub-images



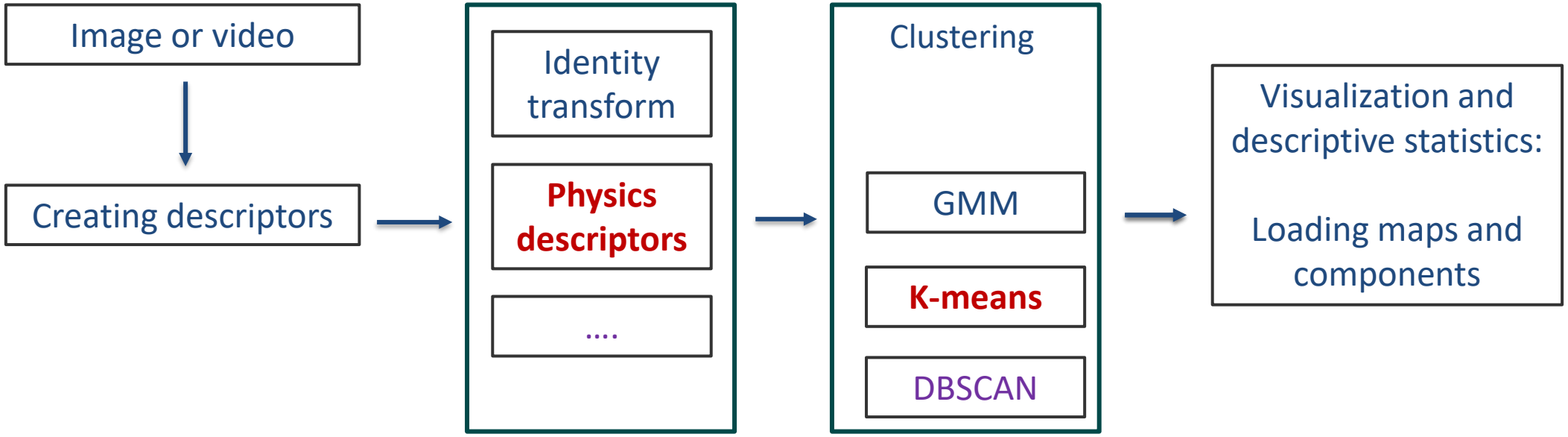
Time-delayed descriptors



The choice of the descriptor:

- Defines physical inferential biases and allows to introduce prior knowledge
- Determines the physical meaning of the analysis
- Establishes the analysis pipeline

Example of analysis pipeline

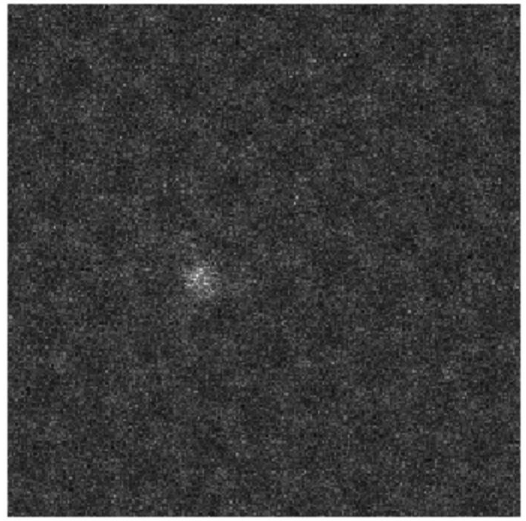


Pipelines are defined to

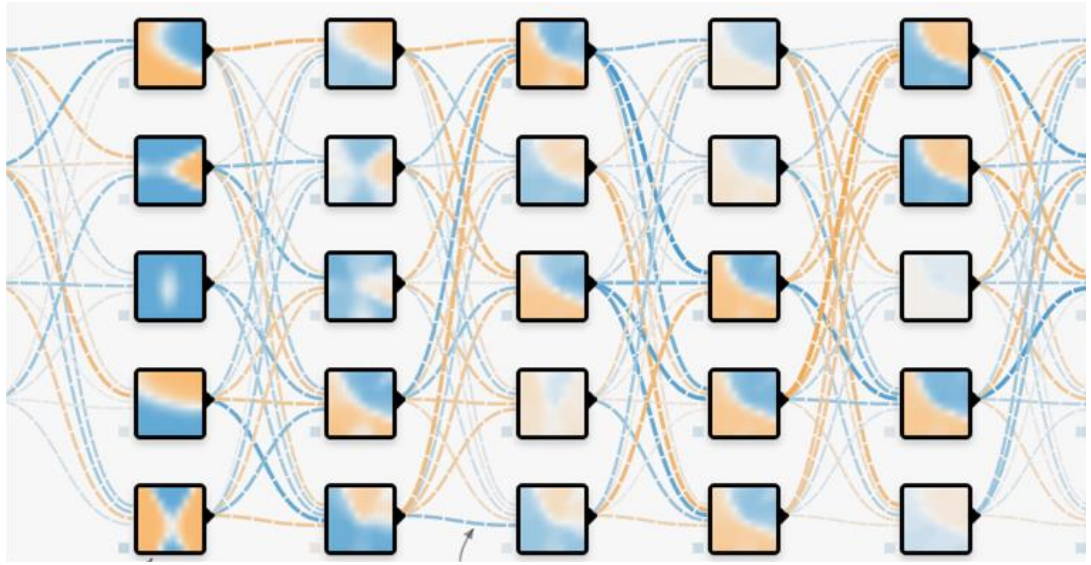
- Make analysis traceable, repeatable, explainable, and transferable
- Allow for hyperparameter tuning and optimization
- Efficiently use the memory

Machine learning can help!

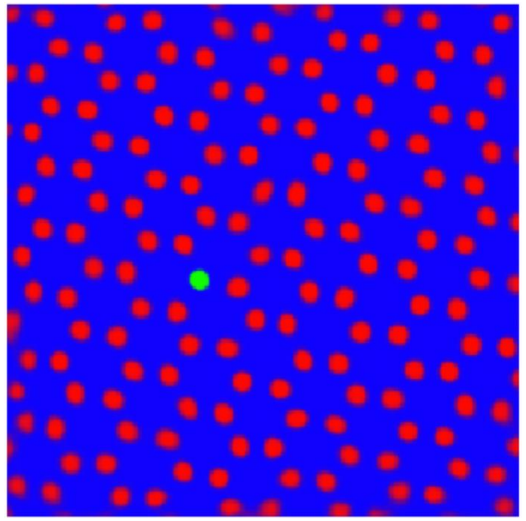
Input



Deep Neural Network (aka deep learning)



Output

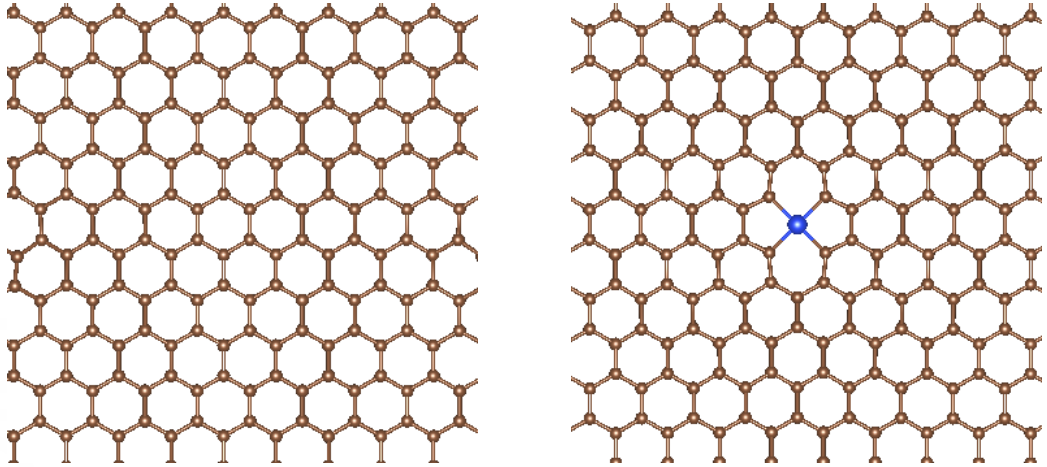


The unique aspect of data in high resolution electron microscopy is the presence of well-defined ground truth (atomic positions) and reasonably well understood physics of the imaging process (allows a creation of a training set with problem-specific physical constraints.

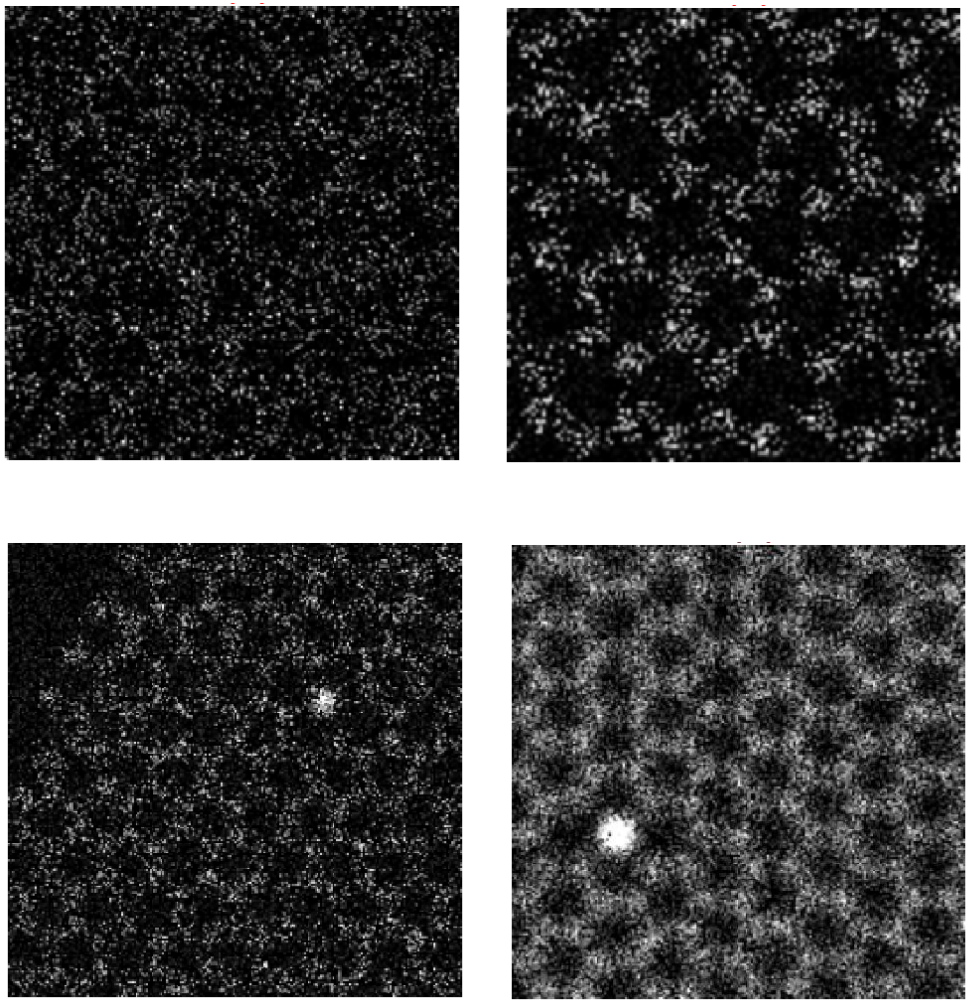
Slide by M. Ziatdinov

Generating training data

Theoretical structures



Simulated images (Training data)

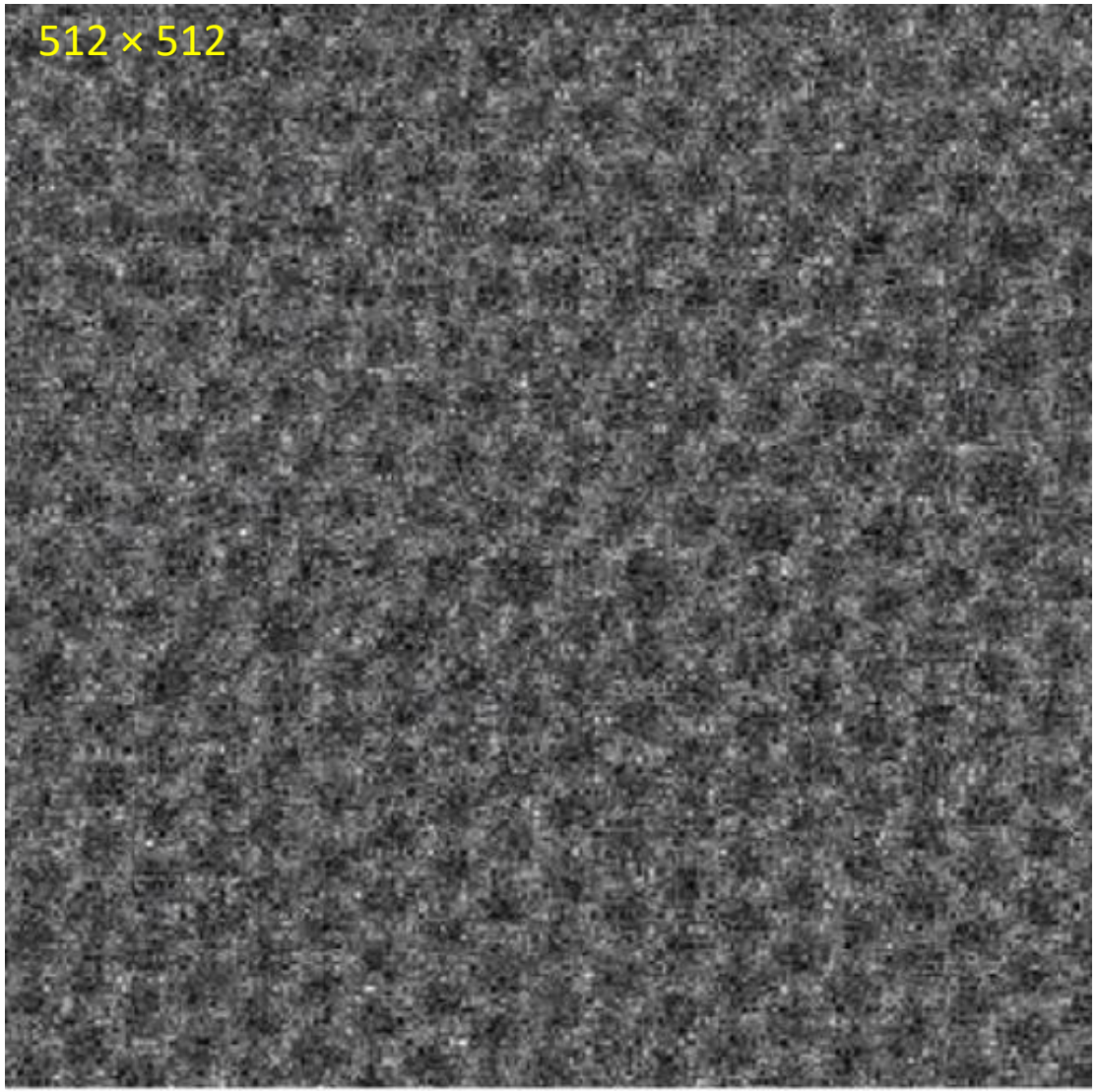


- Atomic coordinates from DFT or MD
- Each structure “augmented” by applying different strain
- Image simulation:
  - Multislice algorithm
  - Atoms as 2D Gaussians
- Image augmentation to account for instrumental factors

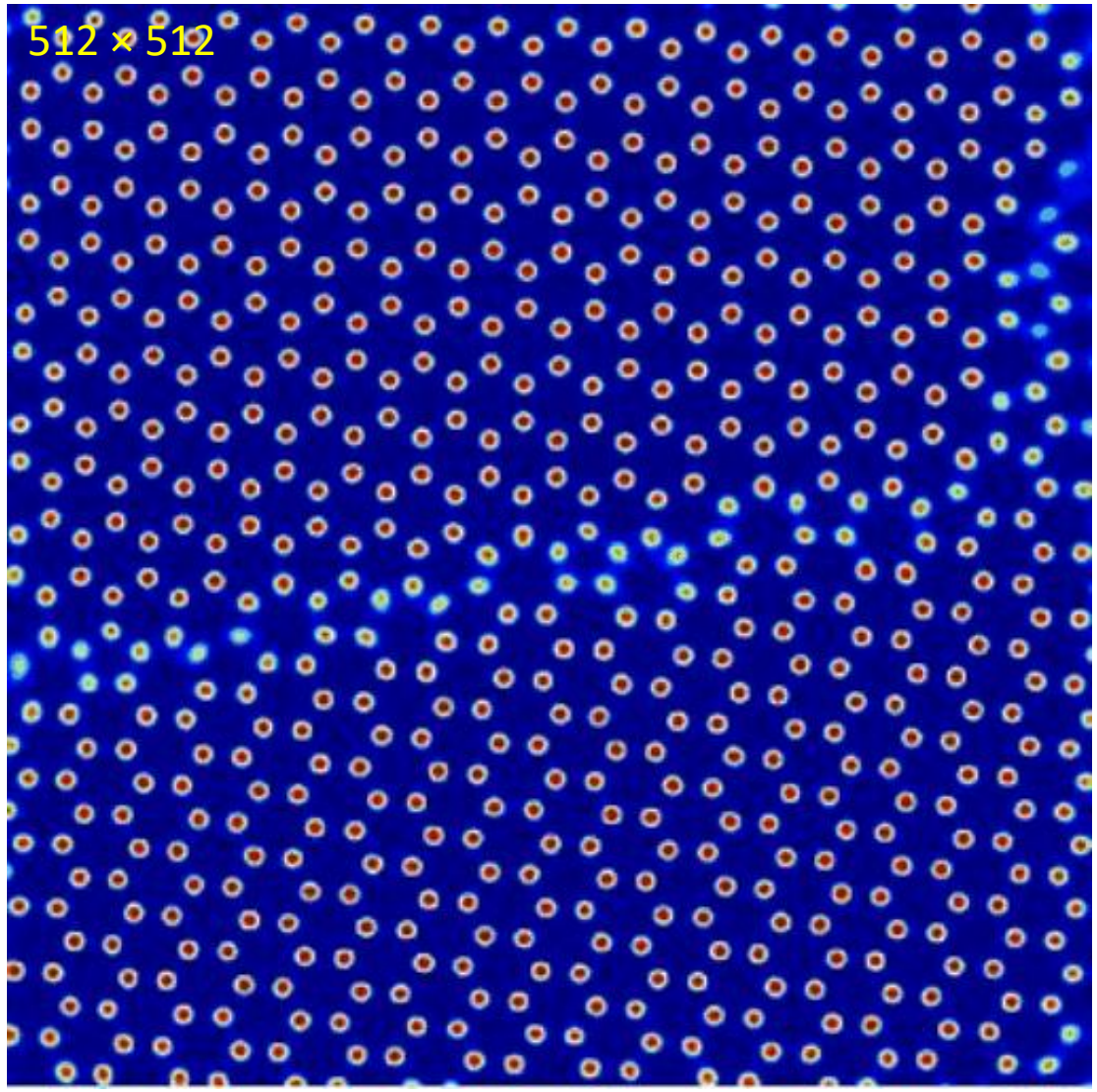
Slide by M. Ziatdinov

Application to electron microscopy data

Experiment



Neural Network Output



Slide by M. Ziatdinov

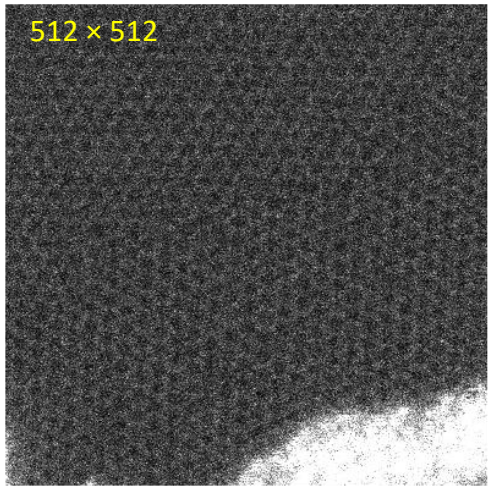
Application to electron microscopy data

Avoids surface "junk"

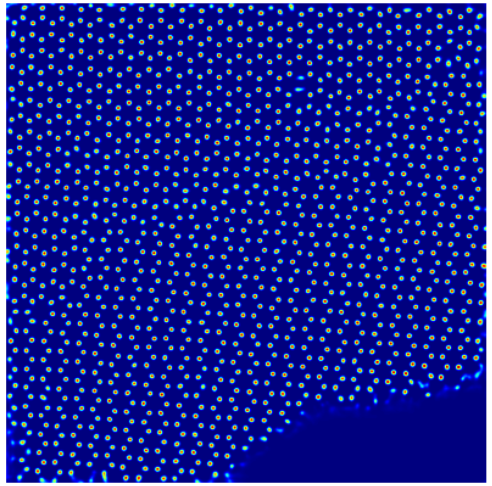
Distinguishes between point impurities and amorphous

Robust to changes in image dimensions

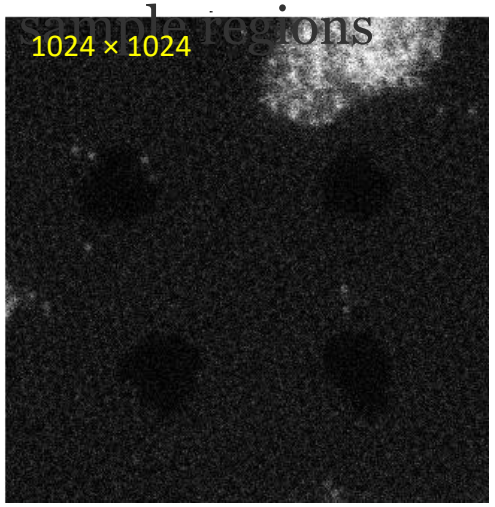
Experimental



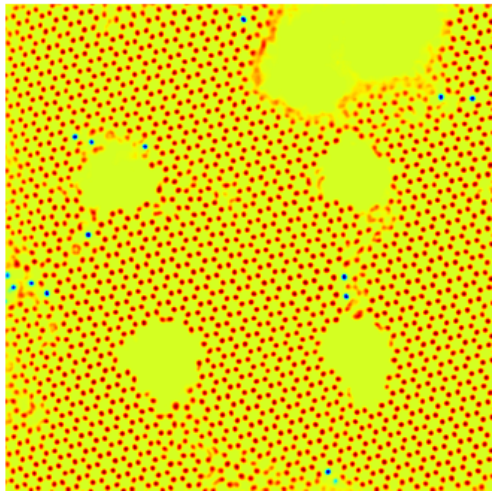
DL output



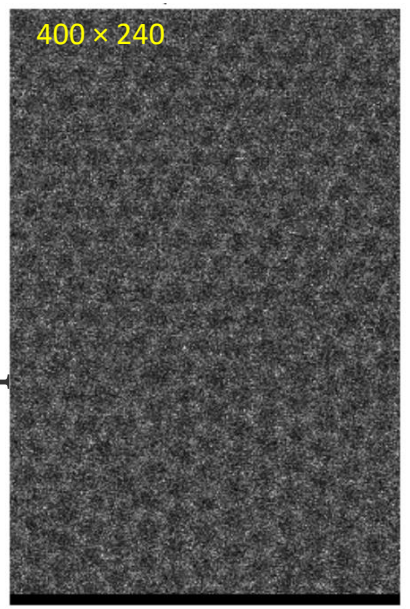
Experimental



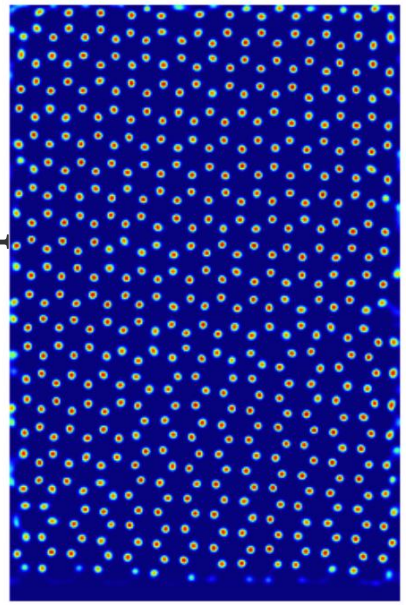
DL output



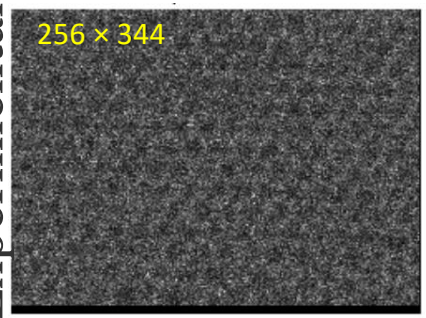
Experimental



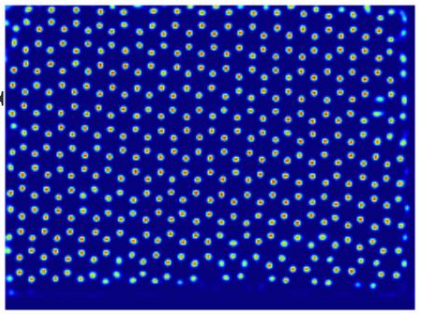
DL output



Experimental

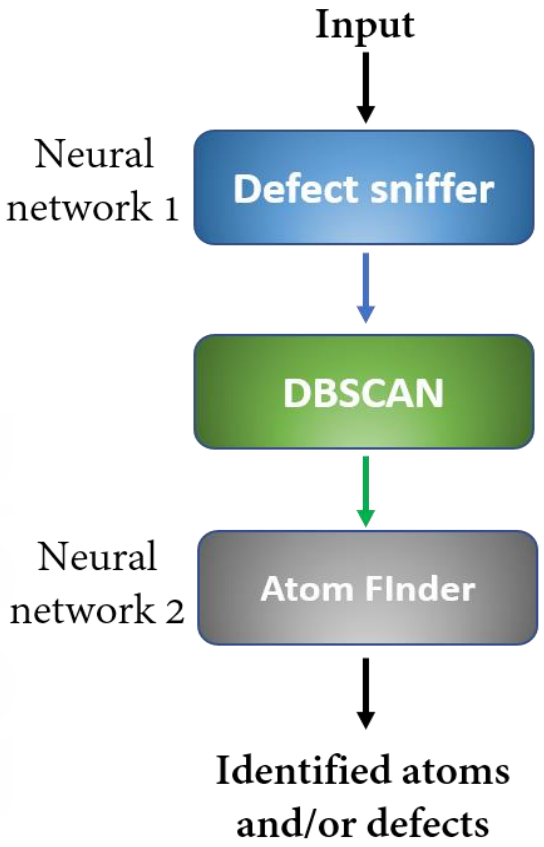


DL output



Slide by M. Ziatdinov

Practically: pipelines of simpler NNs



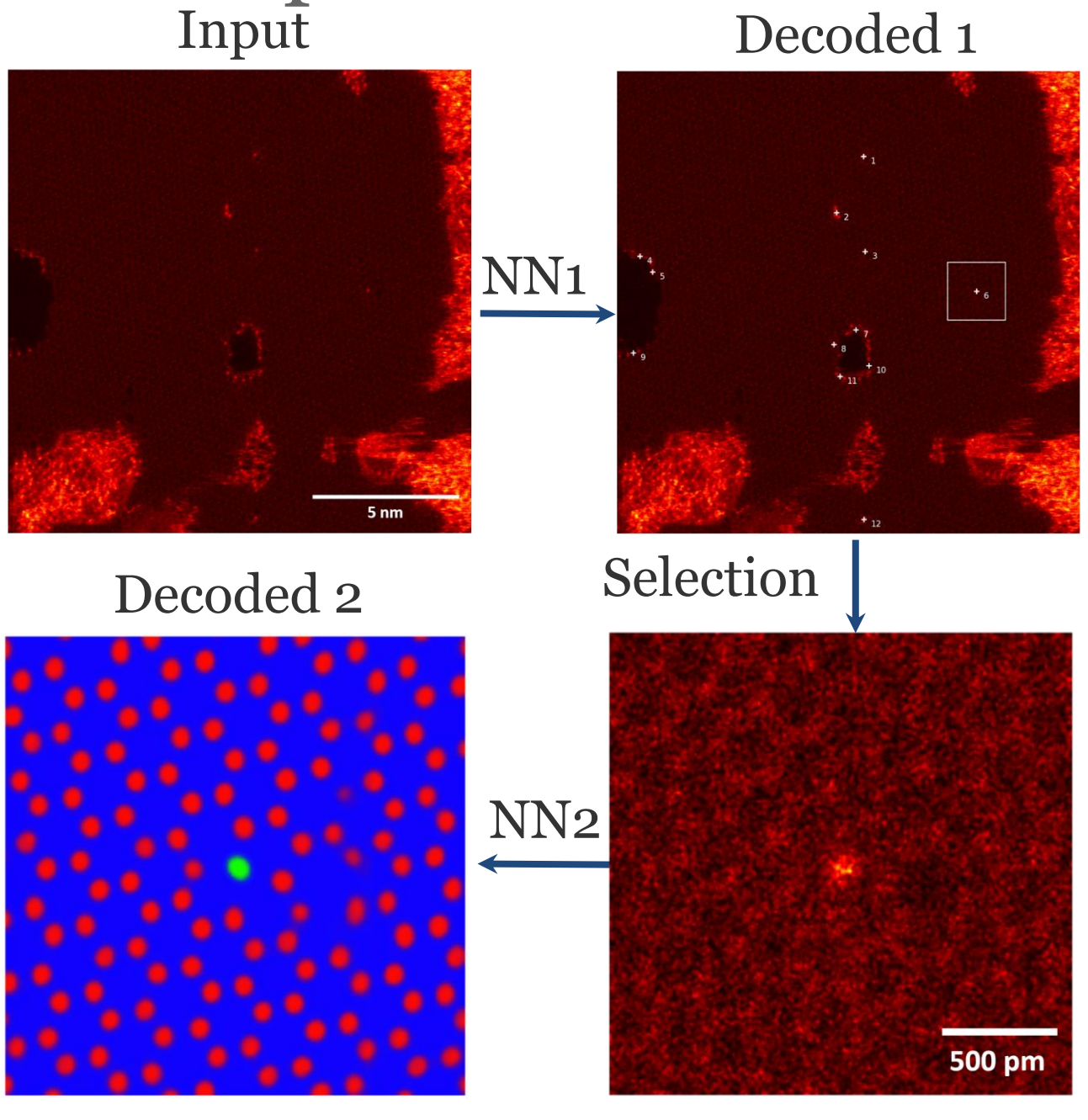
Finding needle in a haystack

We replace categorical cross entropy (CE) with focal loss (FL) function

$$CE(p_t) = -\log(p_t)$$

$$FL(p_t) = -(1-p_t)^{\gamma}\log(p_t)$$

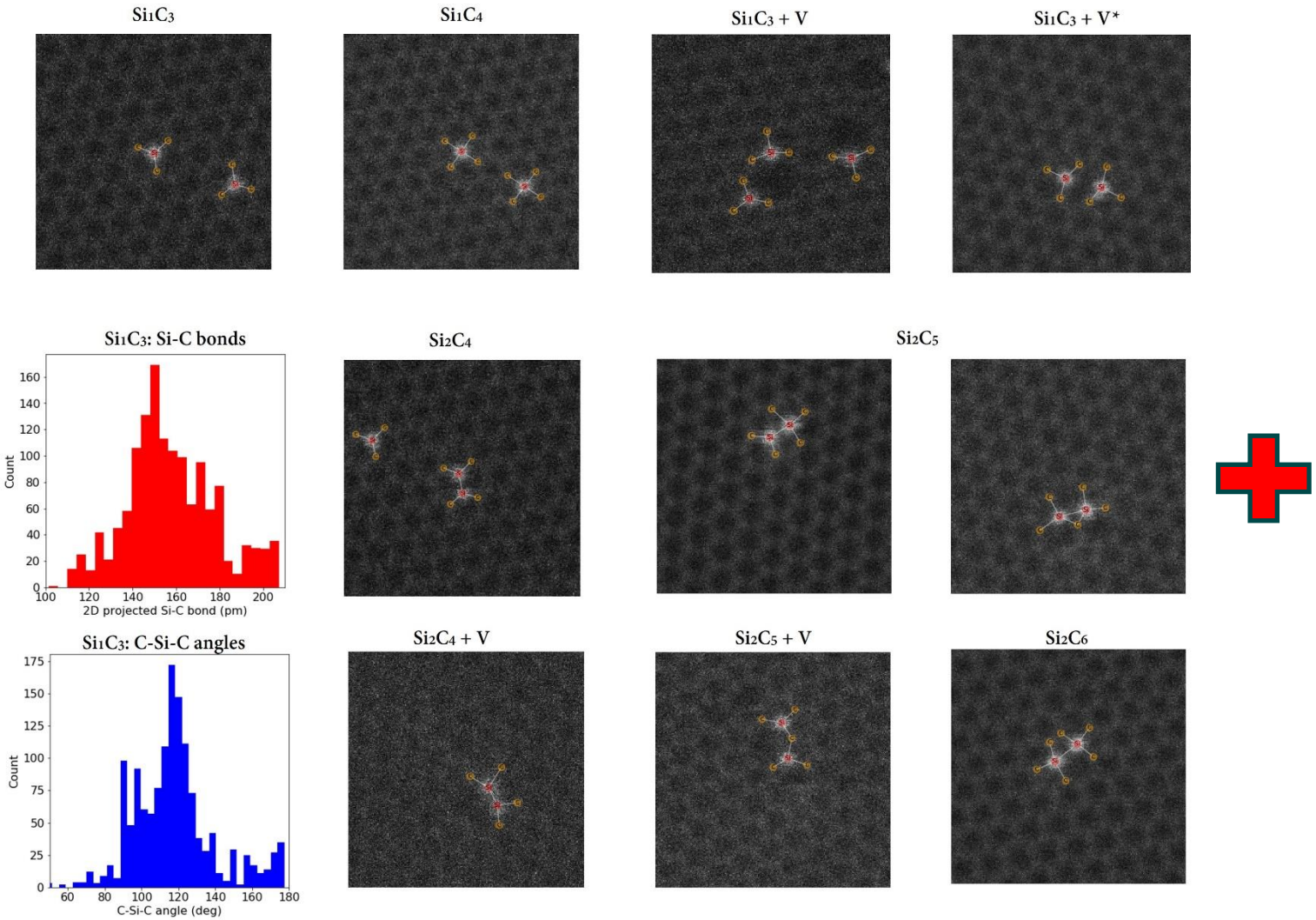
T.-Y. Lin et al., arXiv 1708.02002 (2018)



Opportunities: Building Libraries of Defects

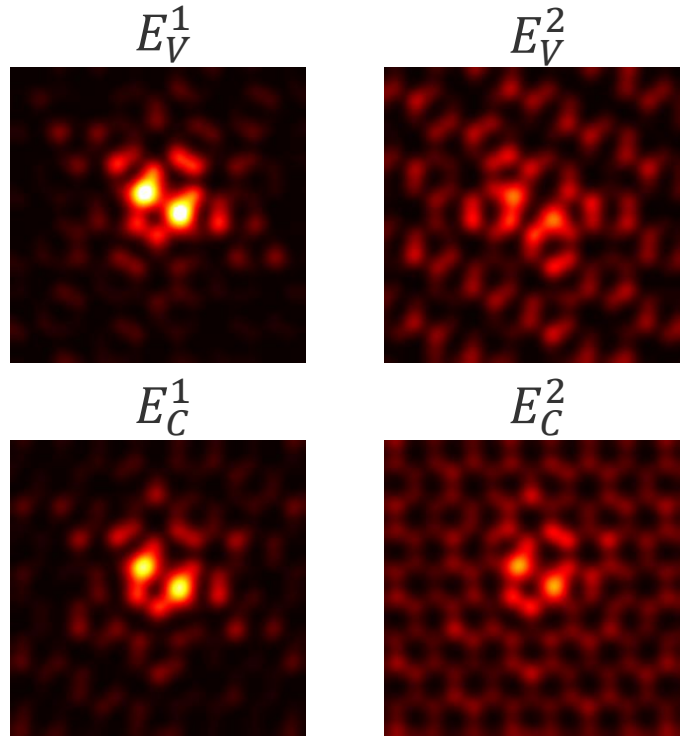
Total number of images analyzed: ~600 (ranging from 256 × 256 to 2048 × 2048 and from 2 nm × 2 nm to 16 nm × 16 nm)

Creation of Si-vacancy libraries from STEM data



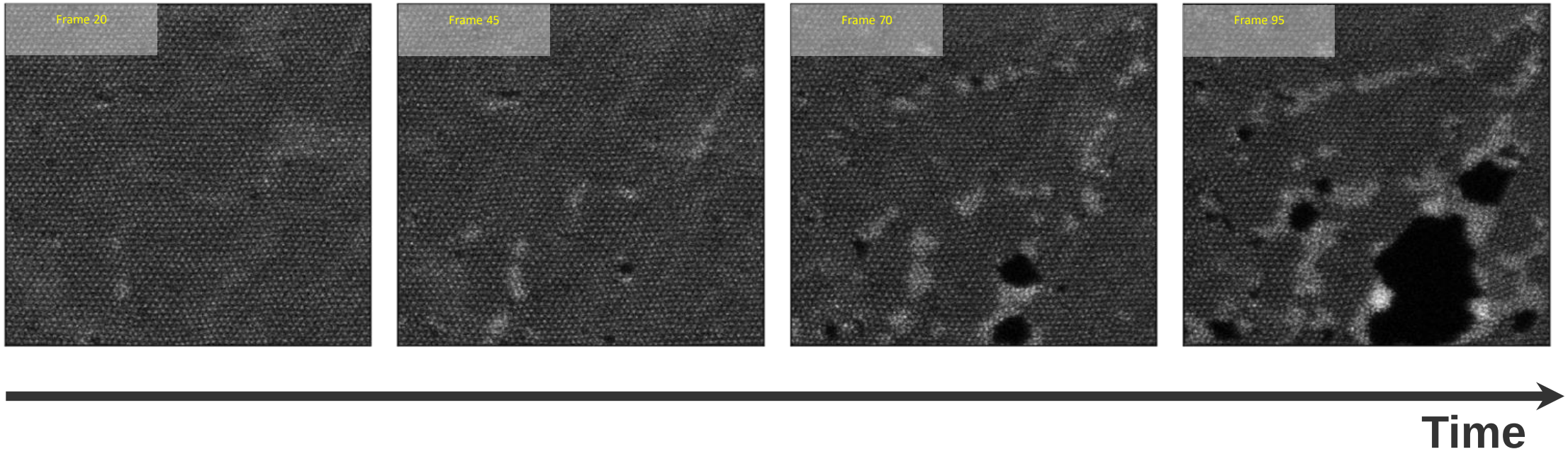
Adding electronic structure calculations

Si2C6



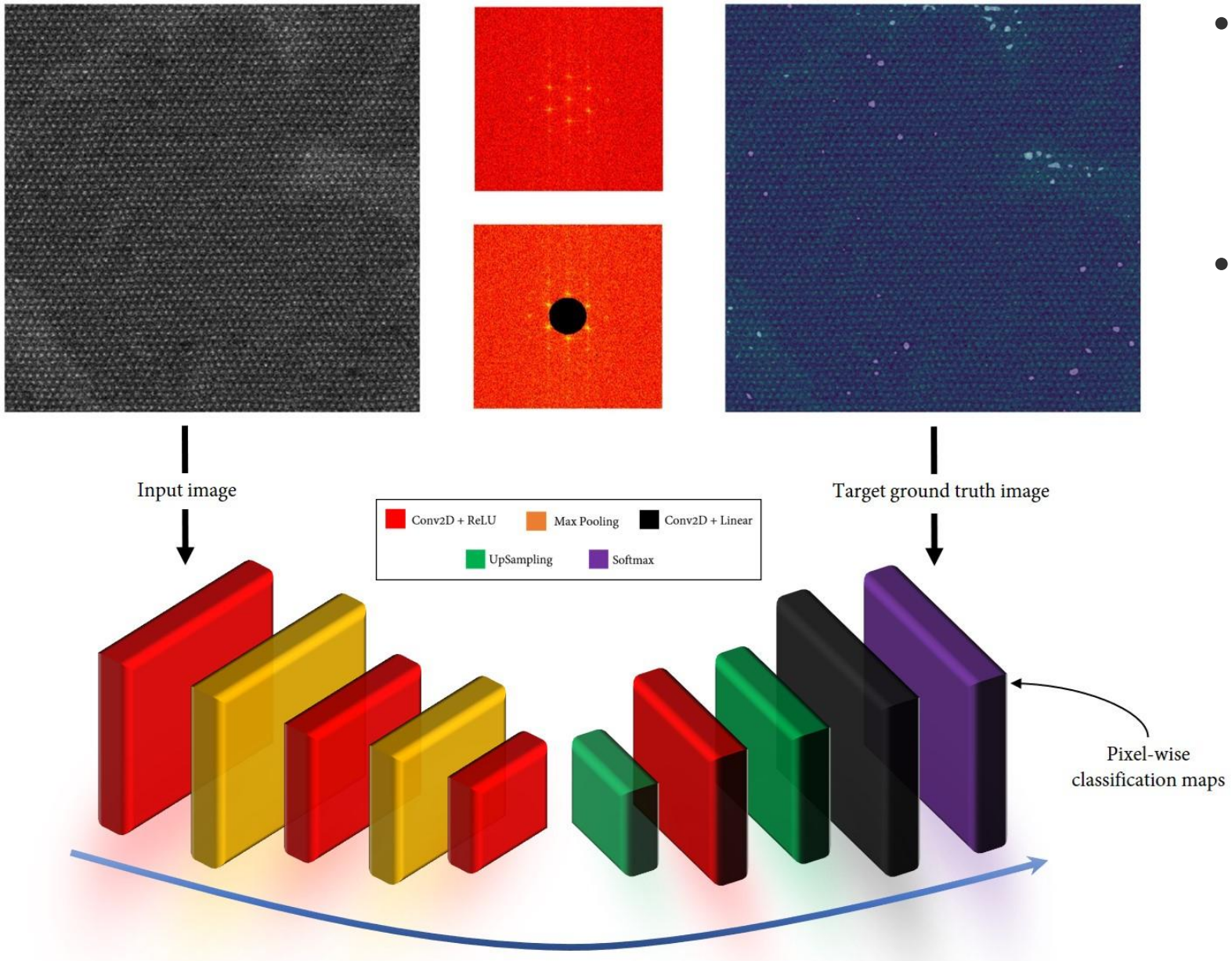
M. Ziatdinov *et al.* ArXiv:1809.04256 (2018).

Exploring defect dynamics: ws_2



- To train a deep learning model, we exploit the fact that each defect is associated with violation of ideal periodicity of the lattice
- We train a model using a single image at the early stage of the beam-induced transformation, when macroscopic periodicity is still maintained, and each defect can be readily discovered, providing the "ground truth" for network training.
- The extracted defect structures are then classified using unsupervised clustering and unmixing techniques

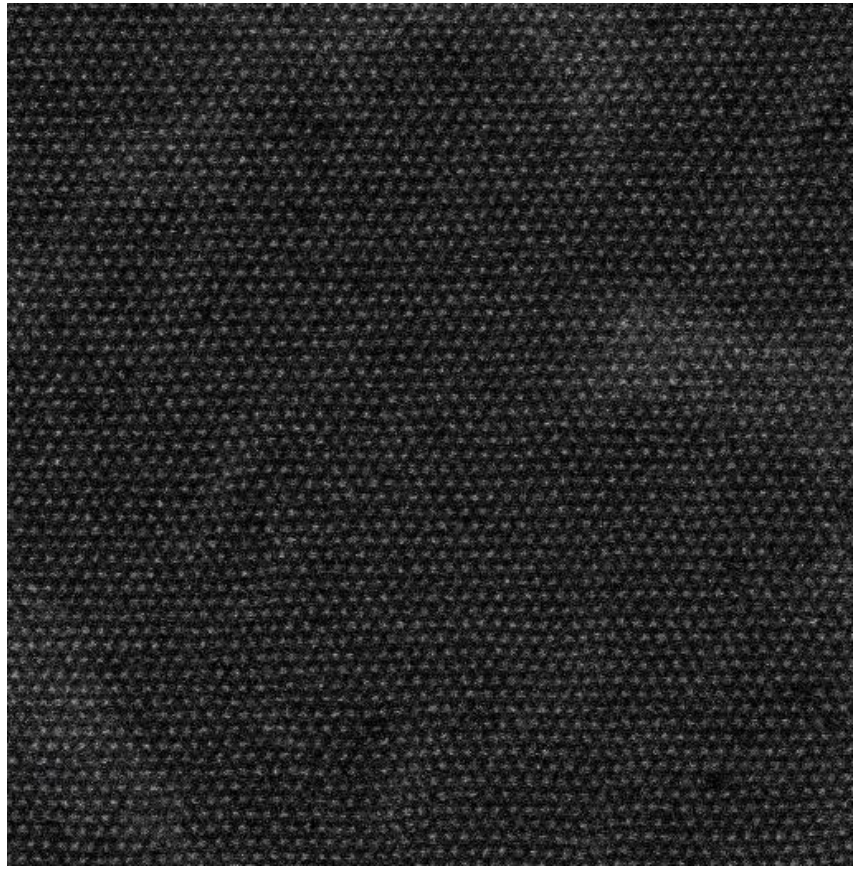
Exploring defect dynamics



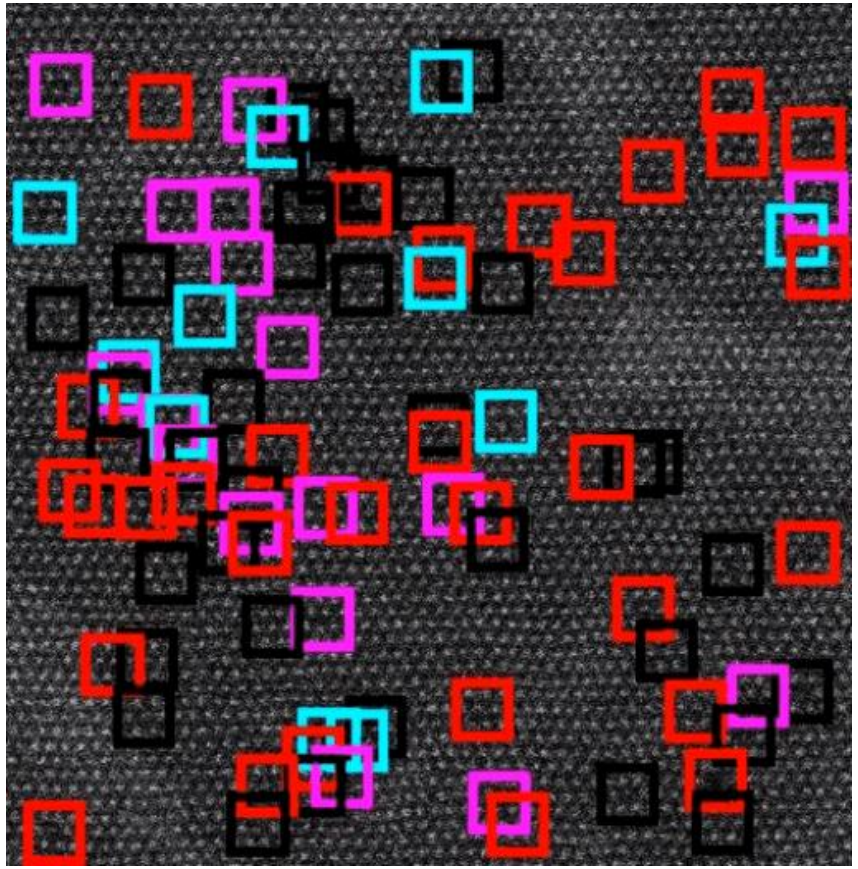
- Training a CNN model (defect identification)
- Gaussian mixture model (defect classification)

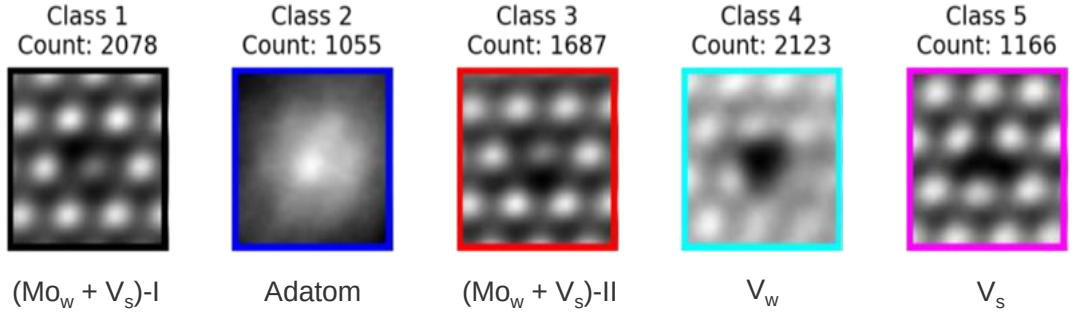
Exploring defect dynamics

Experimental



Decoded

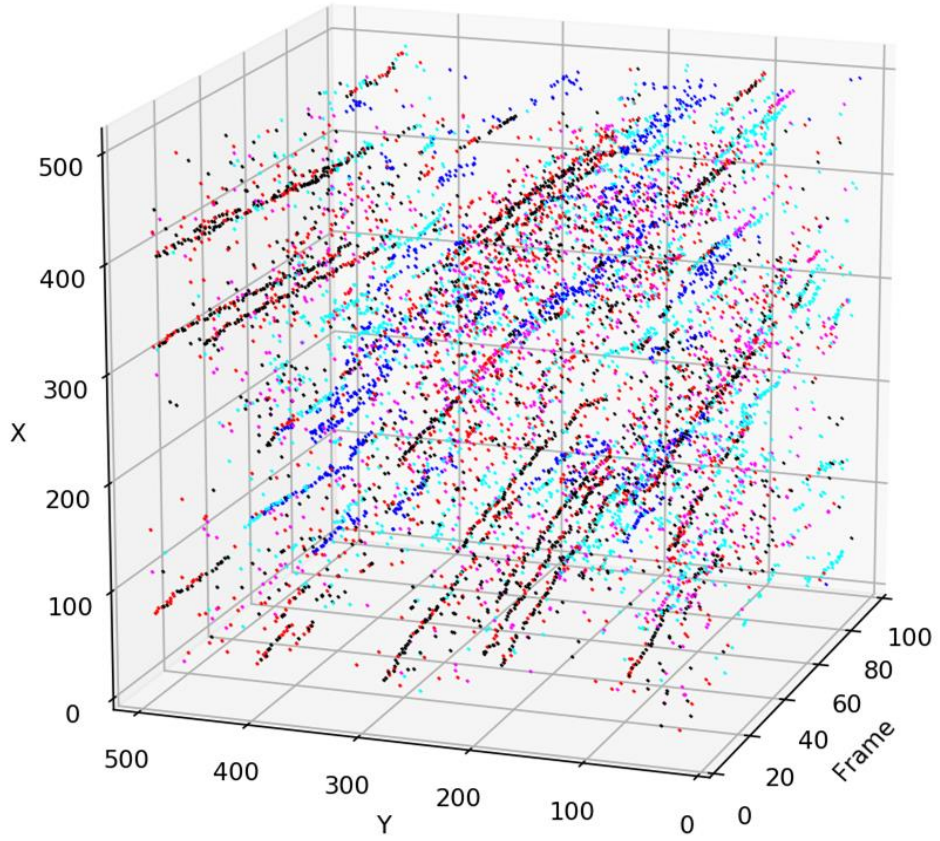




Sample: WS2
E-beam energy: 60 kV
Data by Ondrej Dyck (CNMS/ORNL)

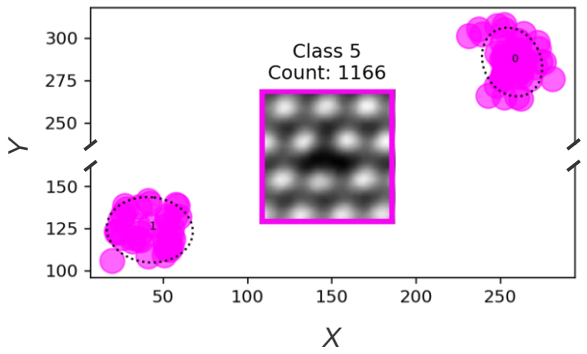
Learning Physics of Defects

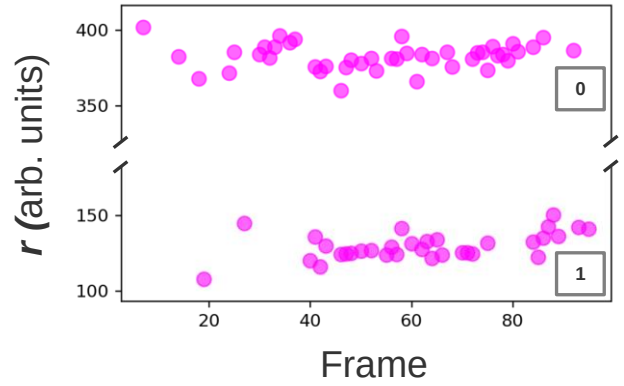
Spatio-temporal trajectories



Maksov *et al.*, npj Computational Materials 5, 12 (2019)

Diffusion parameters



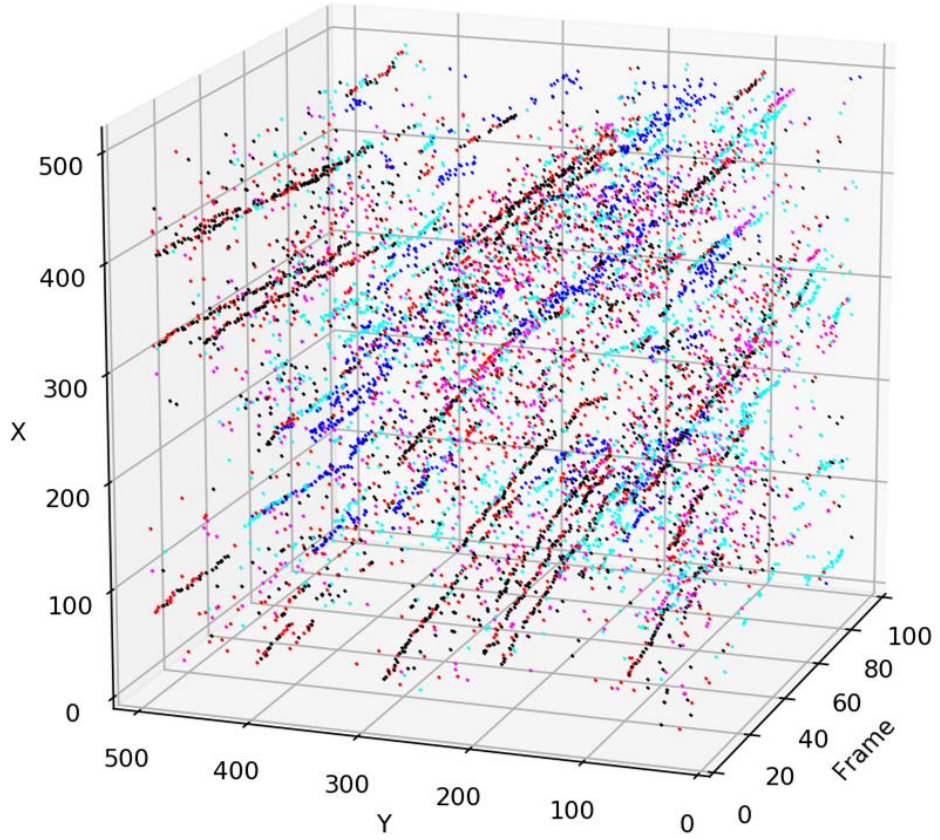


Diffusion coefficient: (3-6)×10^{-4} nm^2/s within 2D random walk approximation

- Identification of dominant point defects and their characteristic statistical behaviors
- Analysis of diffusion parameters for the selected defect species
- Study of transformation pathways and transition probabilities for composite defects

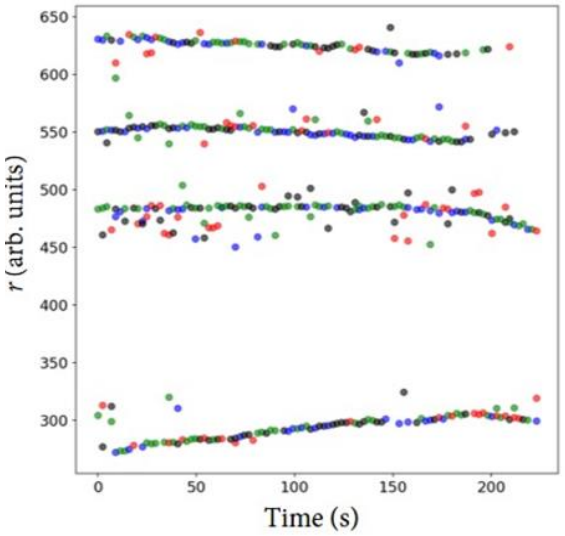
Learning Physics of Defects

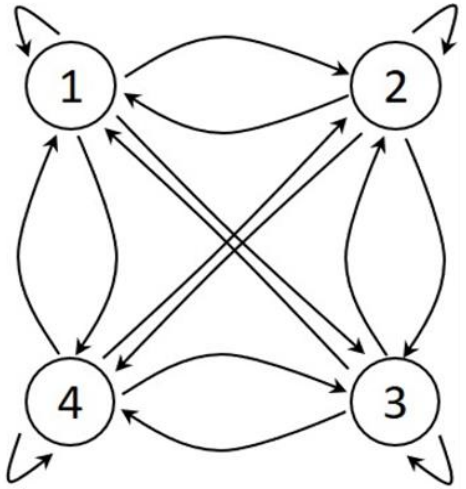
Spatio-temporal trajectories

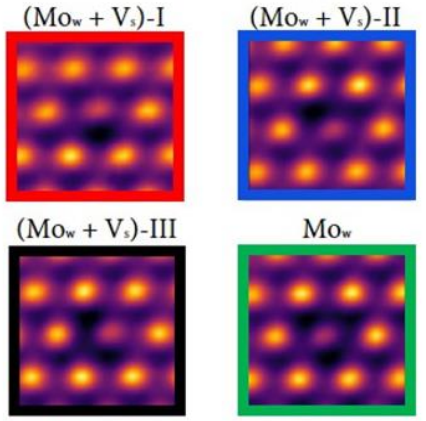


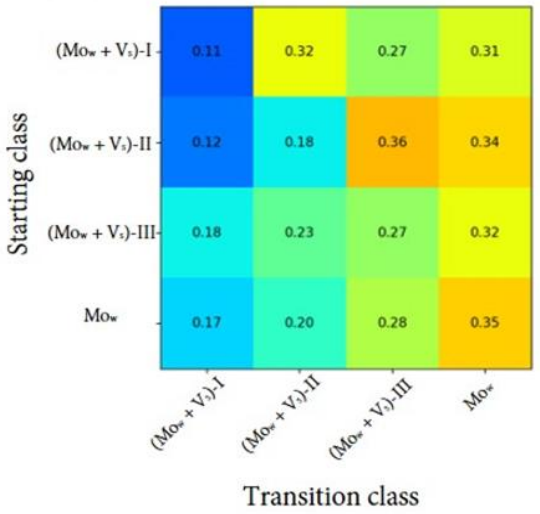
Maksov *et al.*, npj Computational Materials 5, 12 (2019)

Evolution of defects



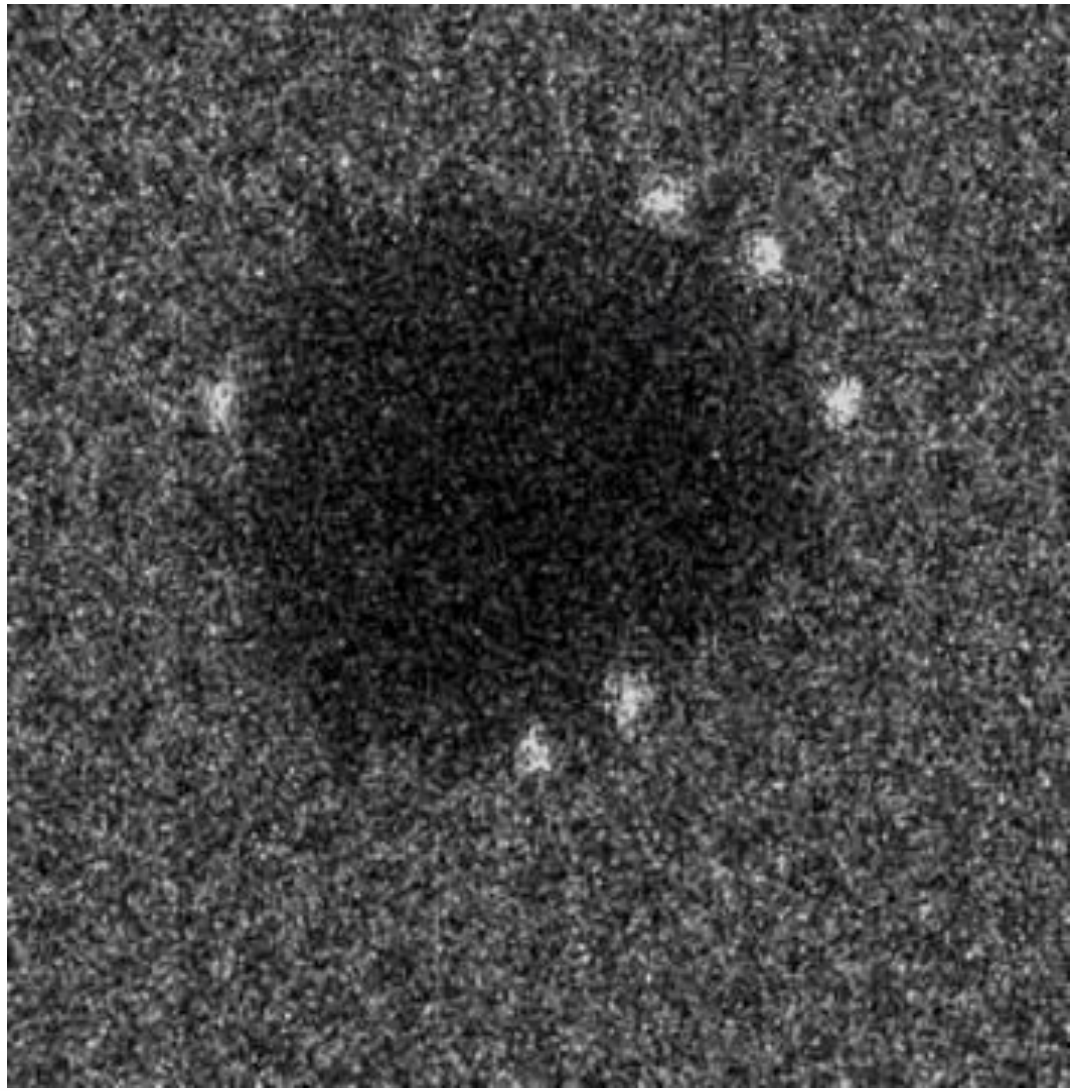




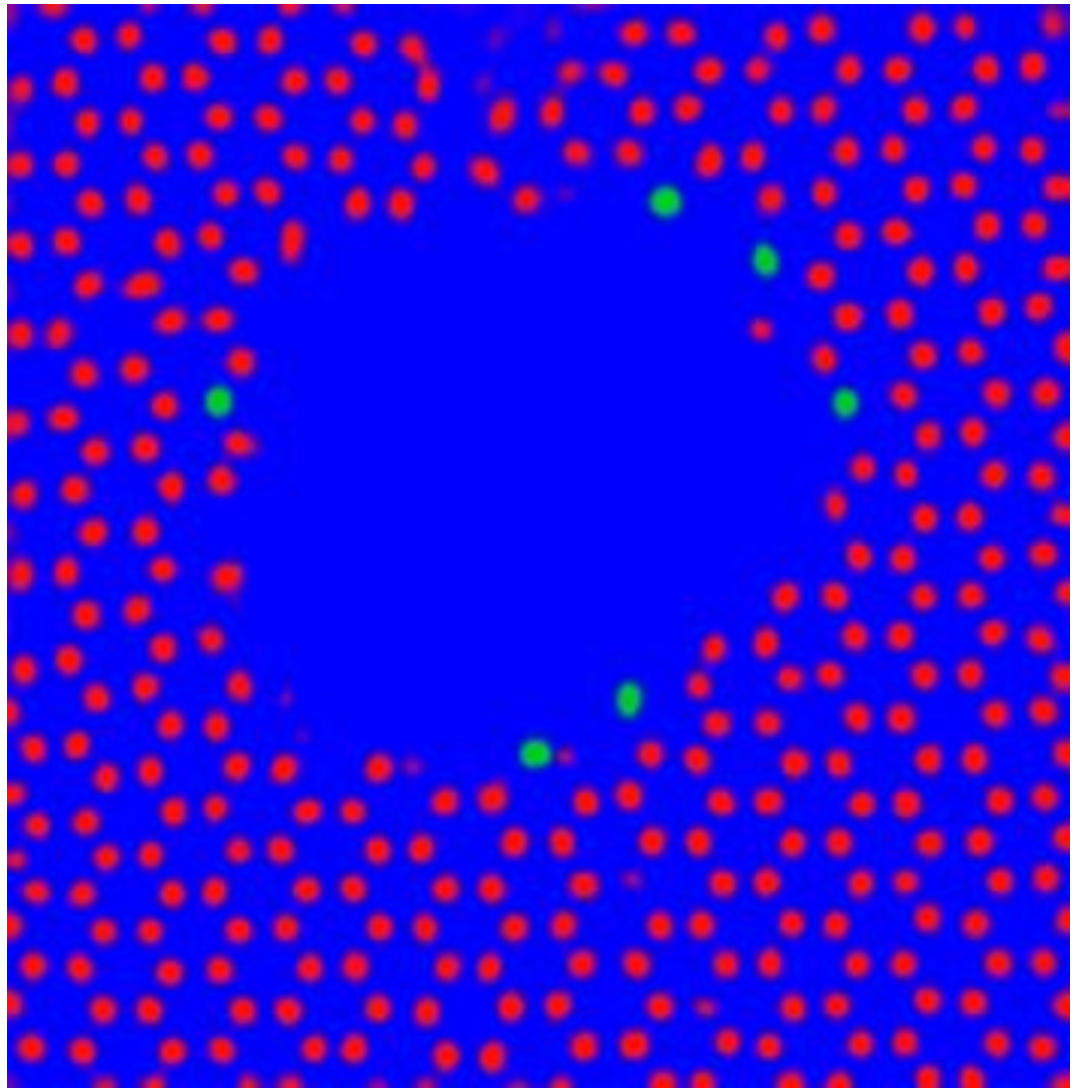


Si impurities on the graphene edge

Experimental data



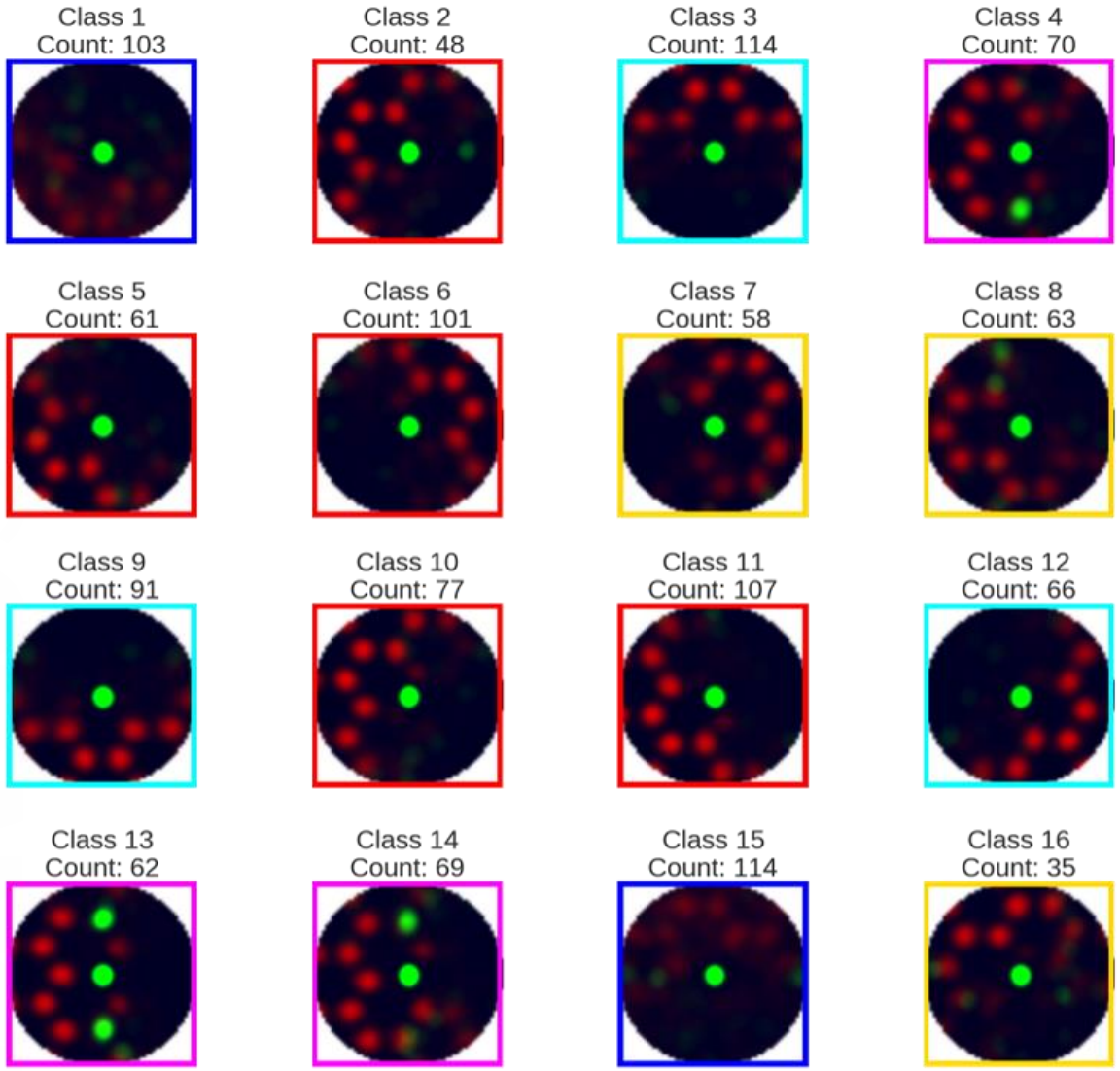
Network's output



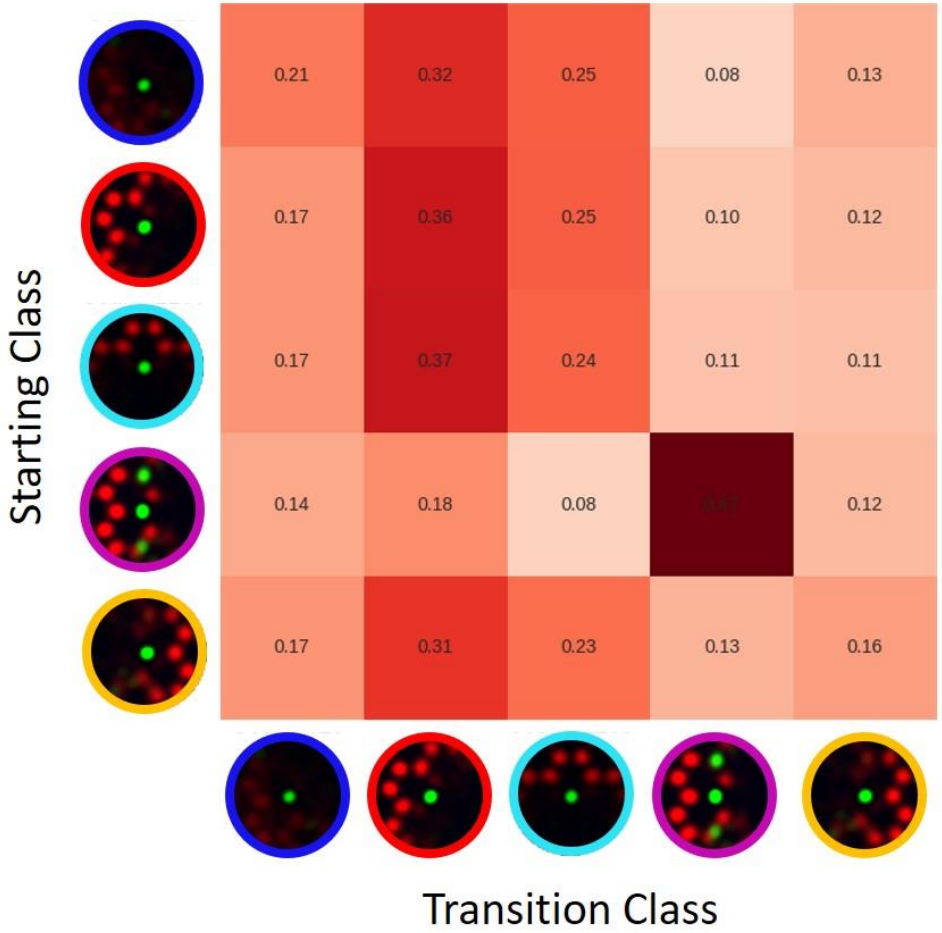
Ziatdinov *et al.*, ArXiv:1901.09322

Classification of edge states

Derived classes of Si-C edge configurations



Transition probabilities matrix



- Gaussian mixture model
- Discrete rotation symmetry
- Markov state analysis

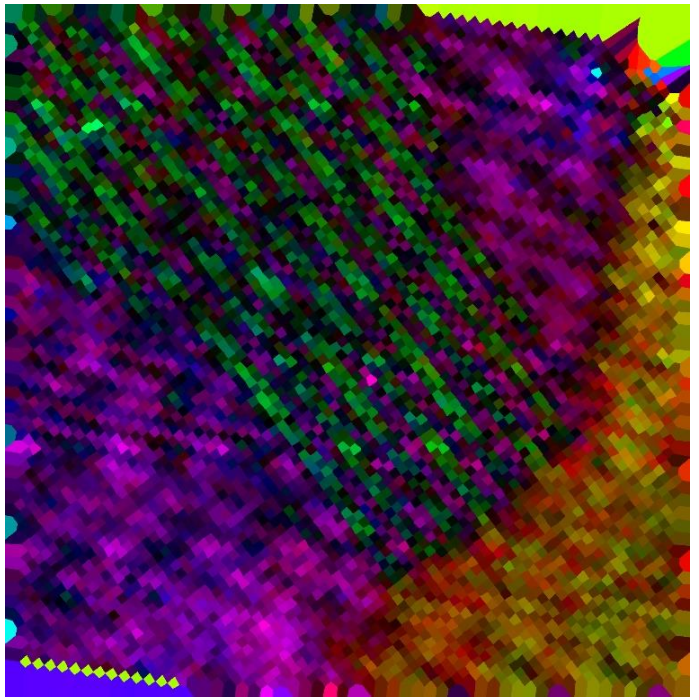
Ziatdinov *et al.*, ArXiv:1901.09322 (2019)

Crystalline materials

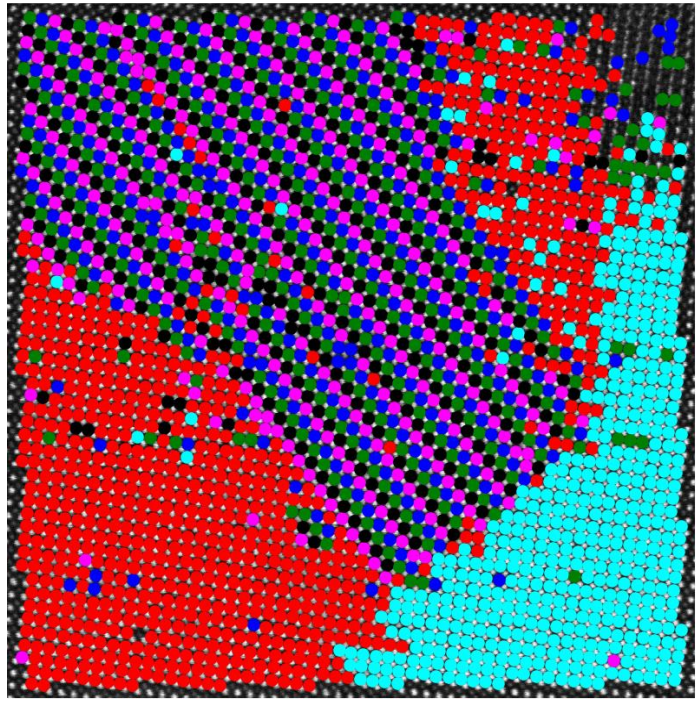
Experimental data (Ferroelectric LBFO)

Data by C. Nelson (MSTD ORNL)

Domain expert analysis



DCNN + k-means (~ 1-3 minutes)



CNN models work with patch by patch and sliding window approaches to image analysis. This allowed us to analyze high resolution images. CNN models were trained using Multislice simulations *and* simulations when atoms are modelled simply as 2D gaussians.

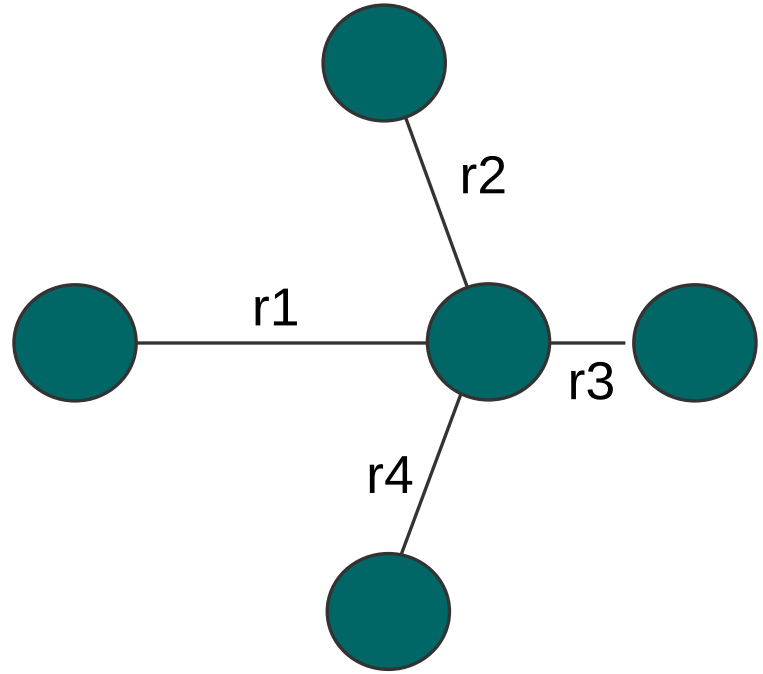
Local crystallography

For each atom, define nearest neighbors and generate array of the corresponding radius-vectors of the form

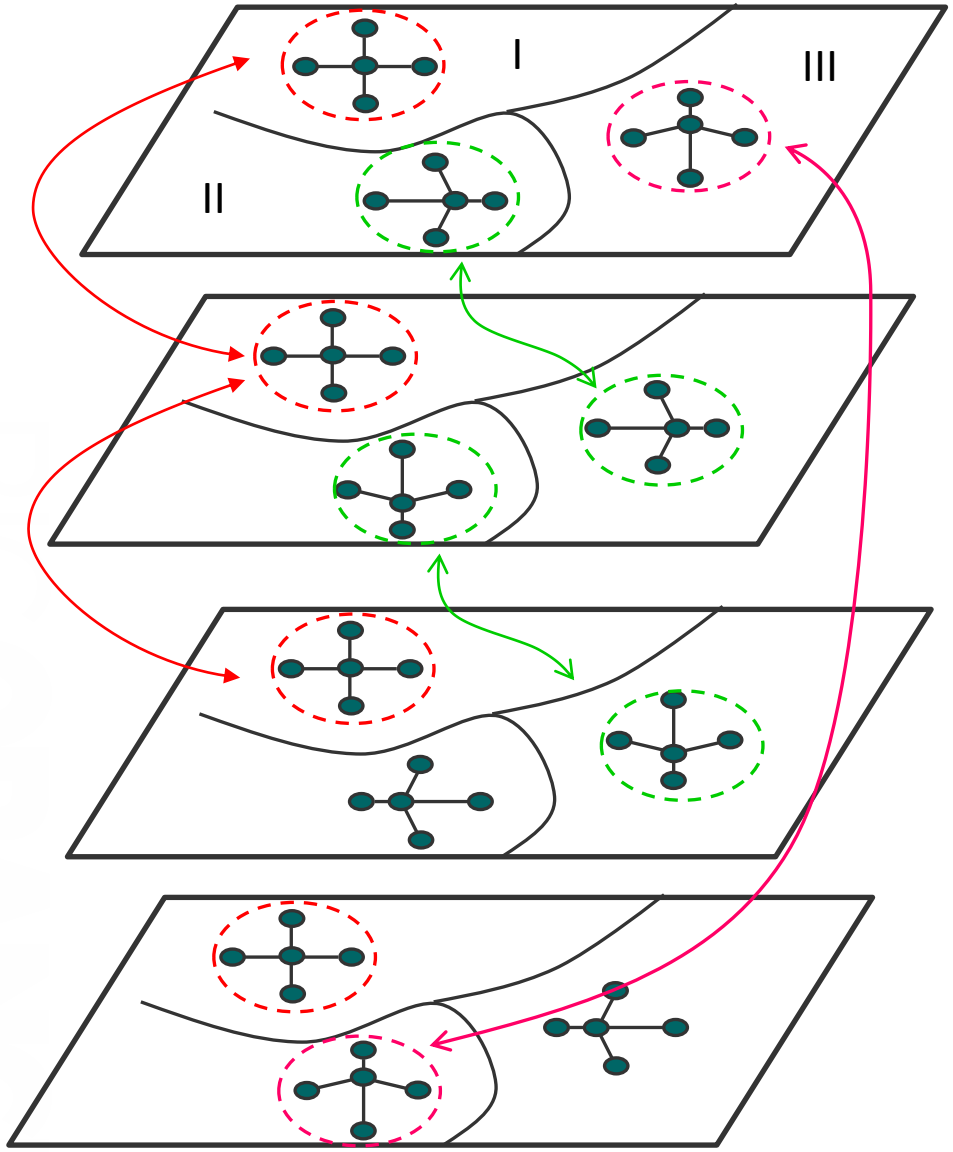
$NA_{ij} = (rx_1, ry_1, rx_2, ry_2, rx_3, ry_3, rx_4, ry_4)_{ij}$

Indexes 1,2,3,4 are chosen in the same sense for all atoms (generalization for different lattice and/or next coordination sphere obvious)

Then, phase/ferroic variant identification problem can be reduced to finding equivalent (in statistical sense) groups of nearest neighbors (for limited sense, we use point groups, for general sense, we use the spatial group and add translation symmetry operations, i.e. i -> i+1 and j -> j+1 for lattice doubling)



Local crystallography



A

A’

A’’

Same cluster in all replicas:
Non-ferroic phase

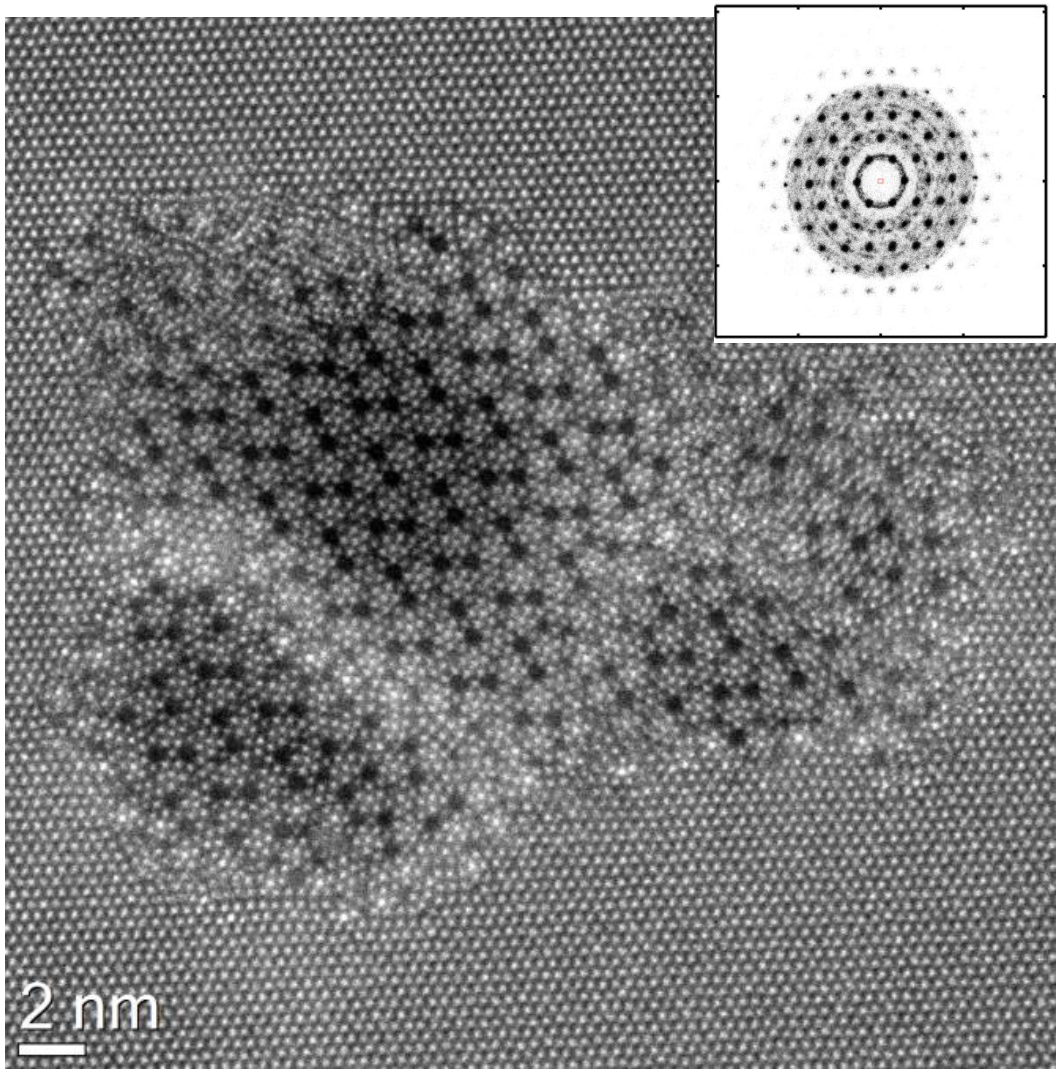
Form cluster with different regions
in different replicas:
Ferroic phase

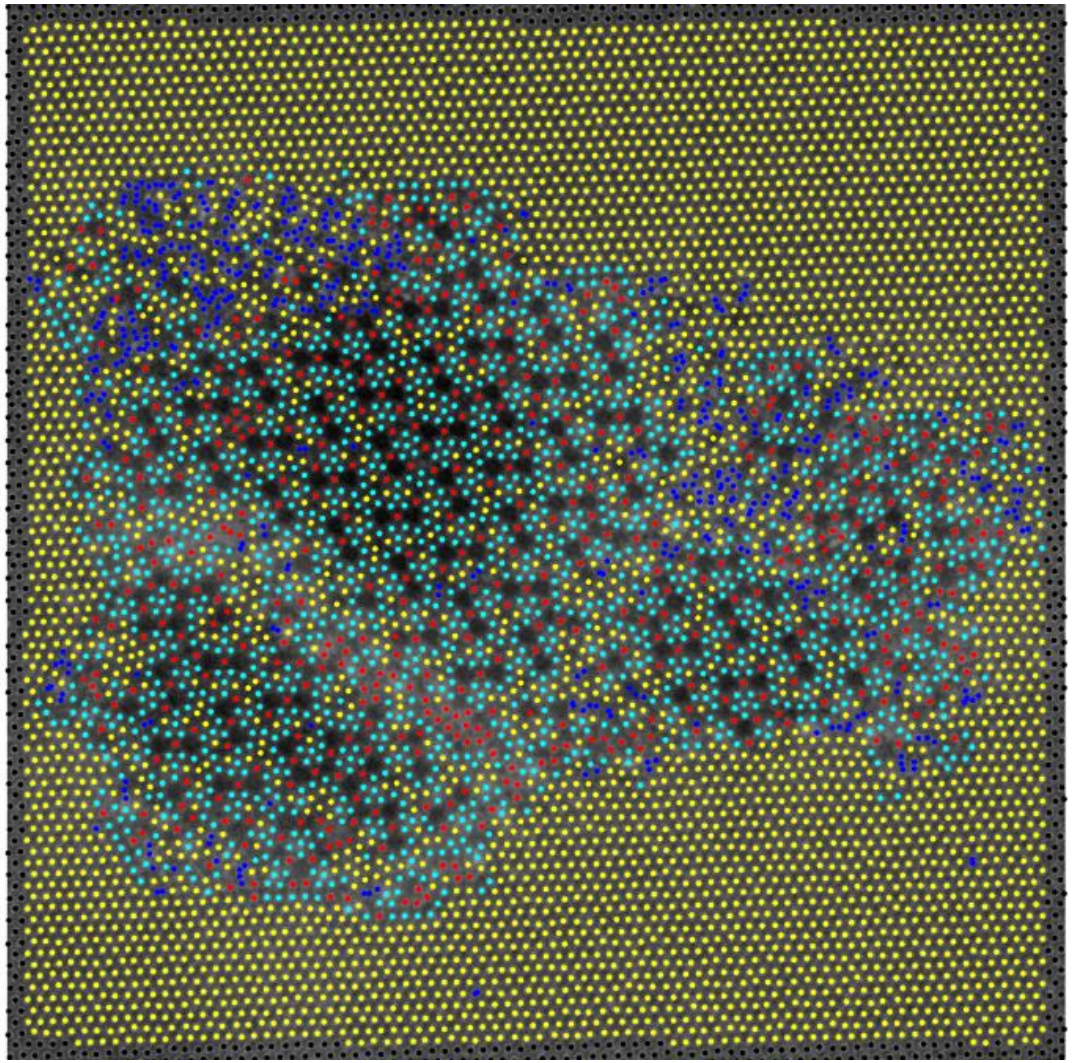
Only some of the correspondences
are shown (but these are obvious)

A. Belianinov, Q. He, M. Kravchenko, S. Jesse, A. Borisevich, and S.V. Kalinin, *Identification of phases, symmetries, and defects through local crystallography*, Nat. Comm. **6**, 7801 (2015).

W. Lin, Q. Li, A. Belianinov, B.C. Sales, A. Sefat, Z. Gai, A.P. Baddorf, M. Pan, S. Jesse, and S.V. Kalinin, *Local crystallography analysis for atomically resolved scanning tunneling microscopy images*, Nanotechnology **24**, 415707 (2013).

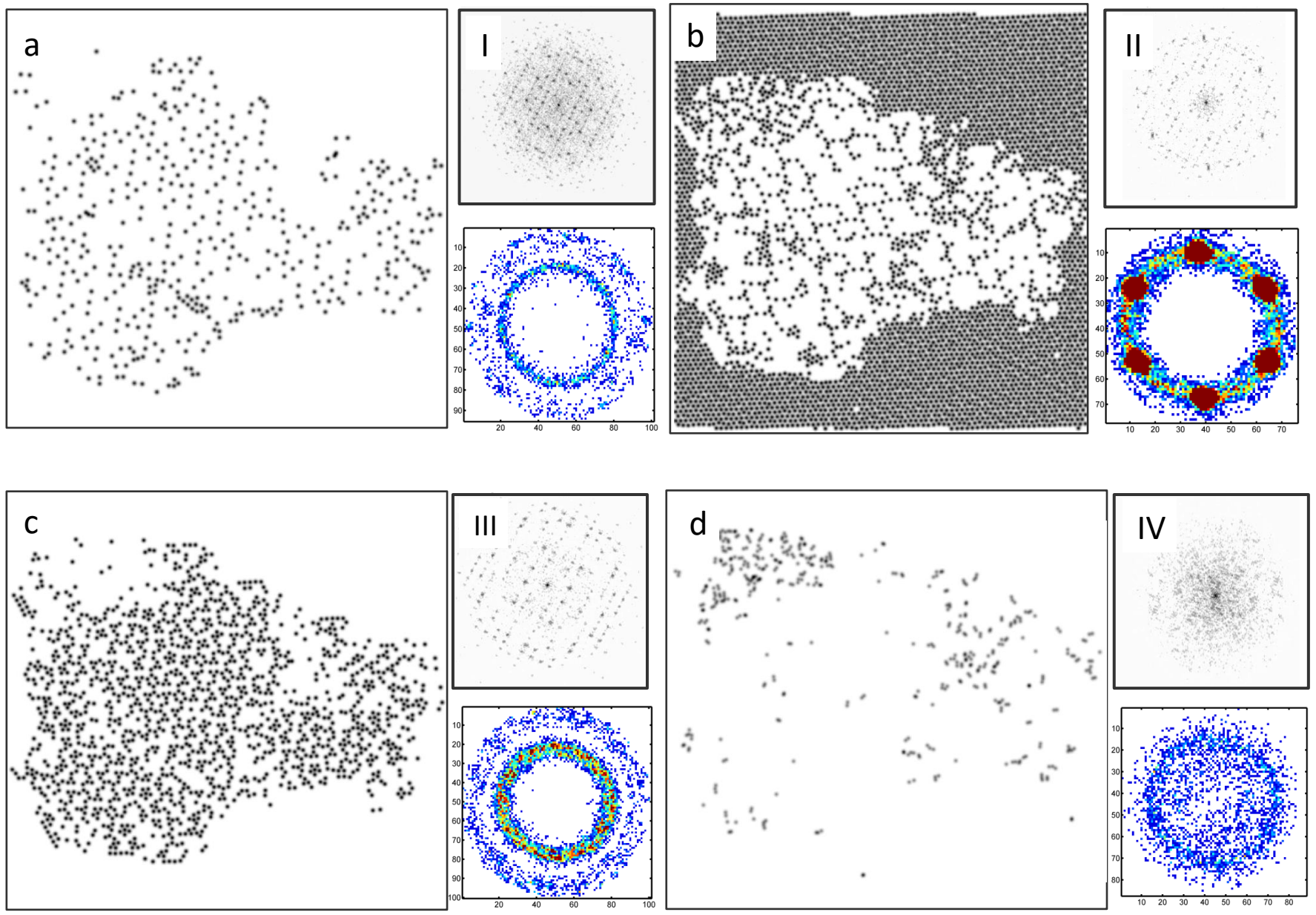
Local crystallography: k-means



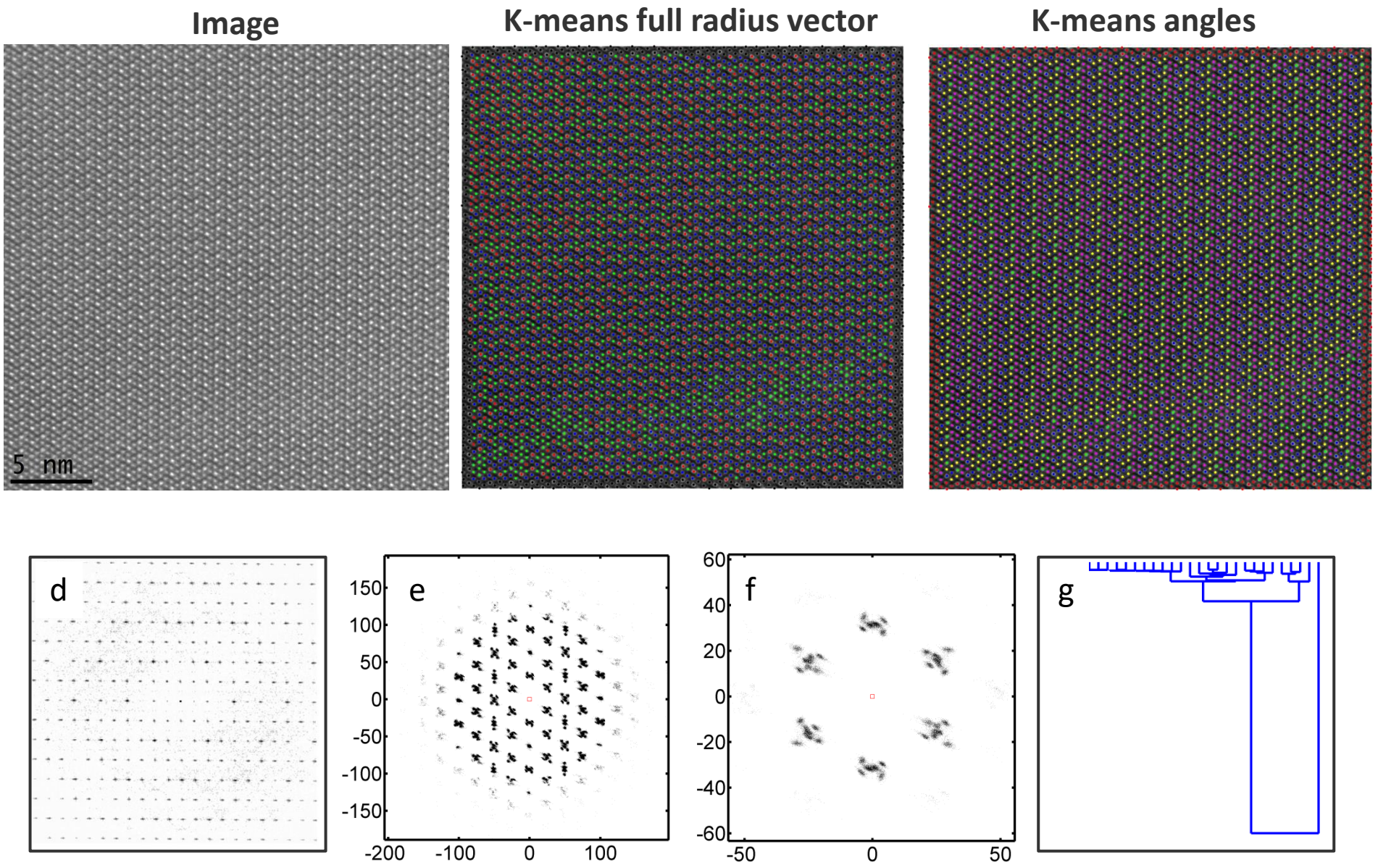


A. Belianinov, Q. He, M. Kravchenko, S. Jesse, A. Borisevich, and S.V. Kalinin, *Identification of phases, symmetries, and defects through local crystallography*, Nat. Comm. **6**, 7801 (2015).

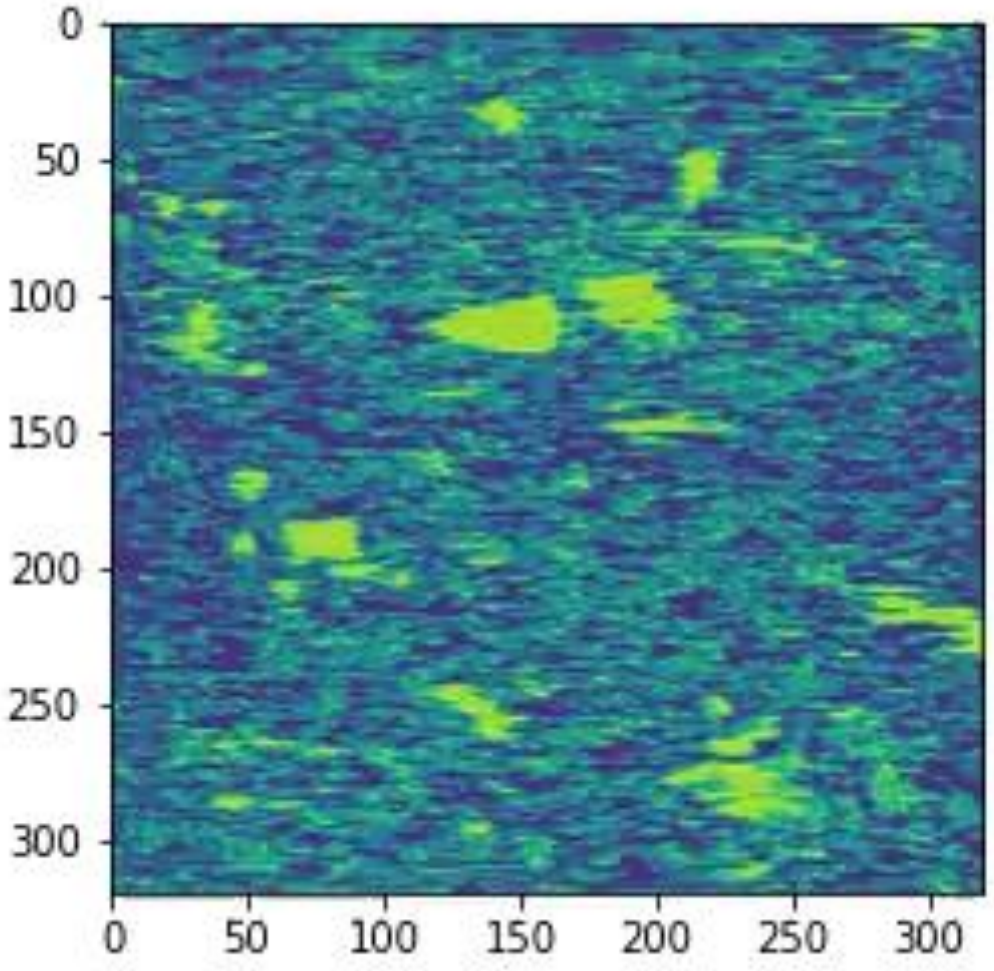
Local crystallography

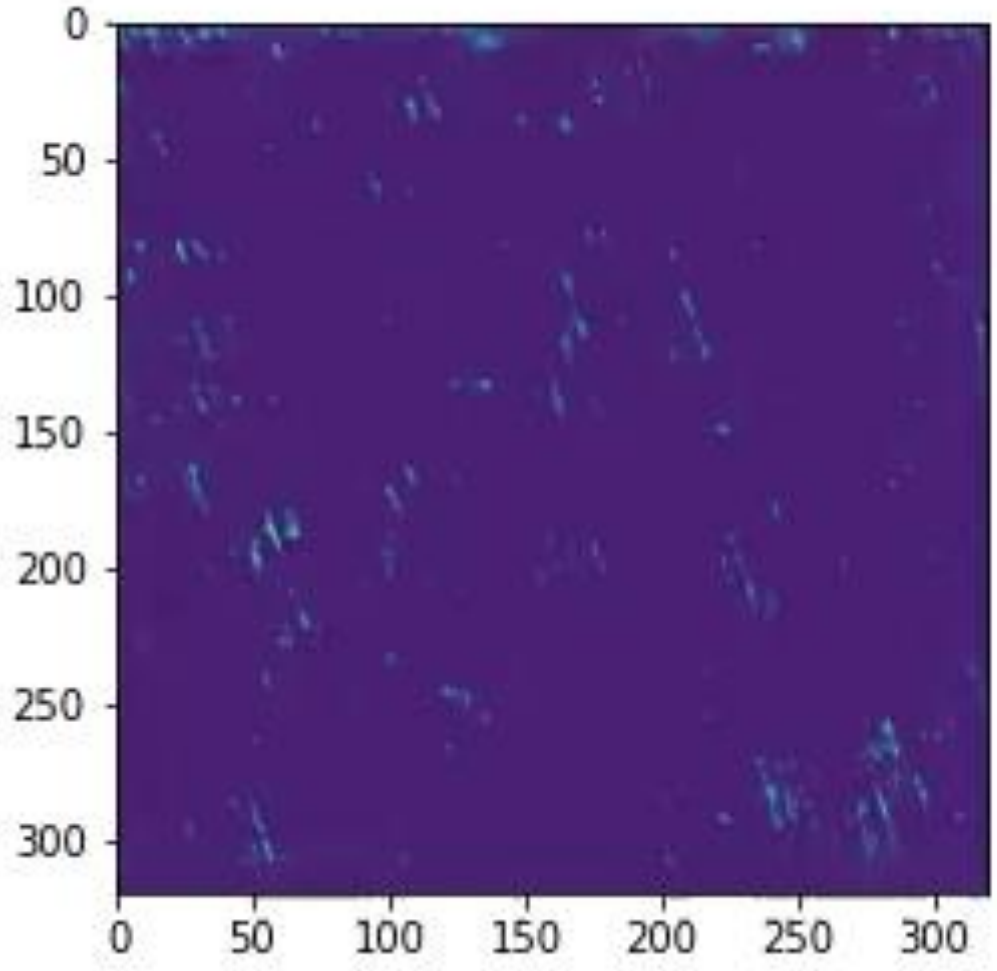


Local crystallography



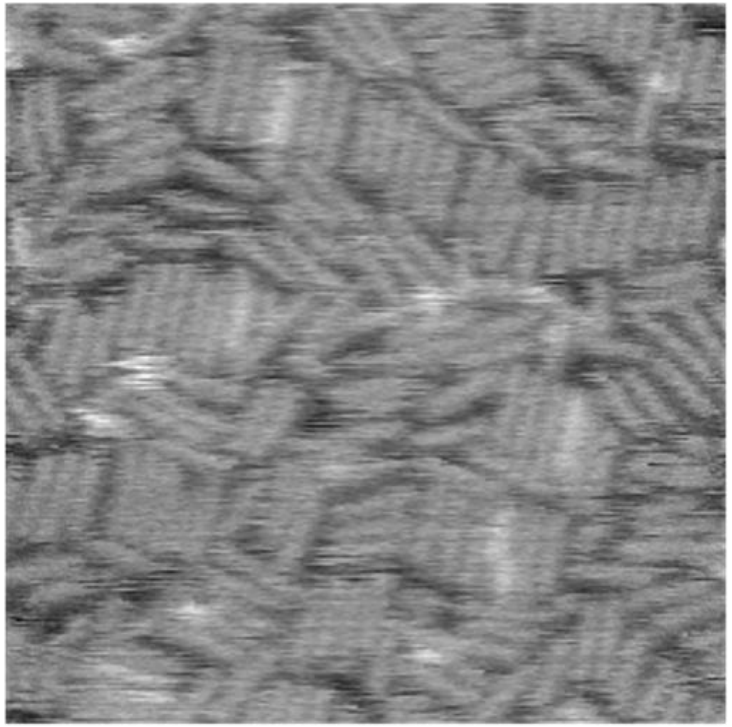
Protein assembly



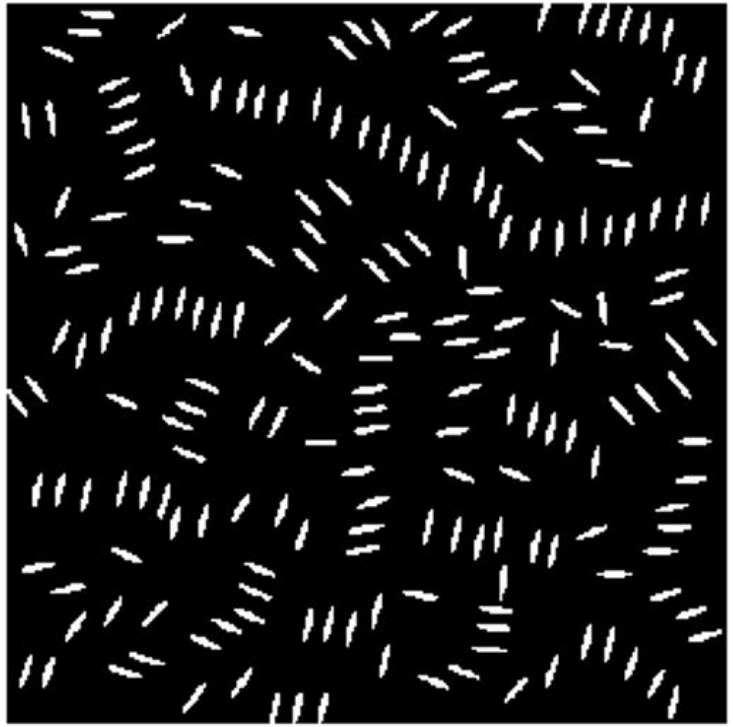


Protein assembly

Experimental image

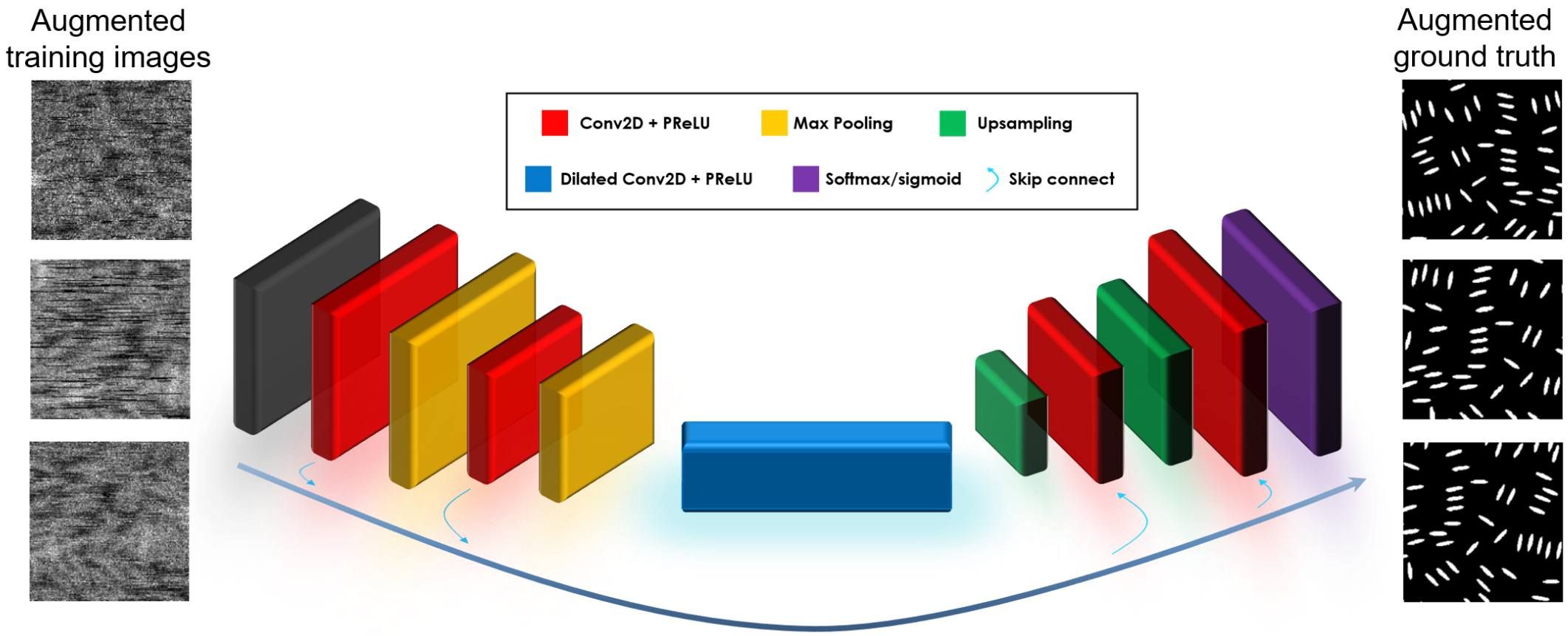


Ground truth



- Select an image where the position and orientation of all particles is well-defined (usually the last frame of an AFM movie)
- Perform manual labelling. The trained network should allow extraction of position and orientation of each particle.
- Generate a training set on-the-fly via data augmentation by random cropping, adding Gaussian and Poisson noises, artificial scan “scars”, zooming and resizing.

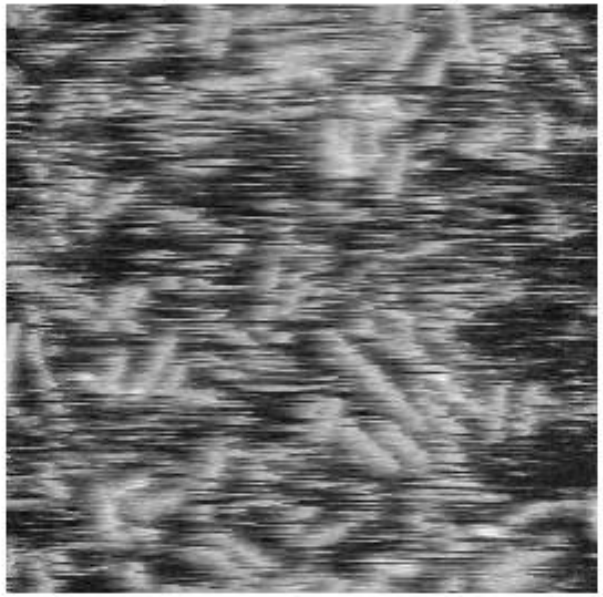
Protein assembly



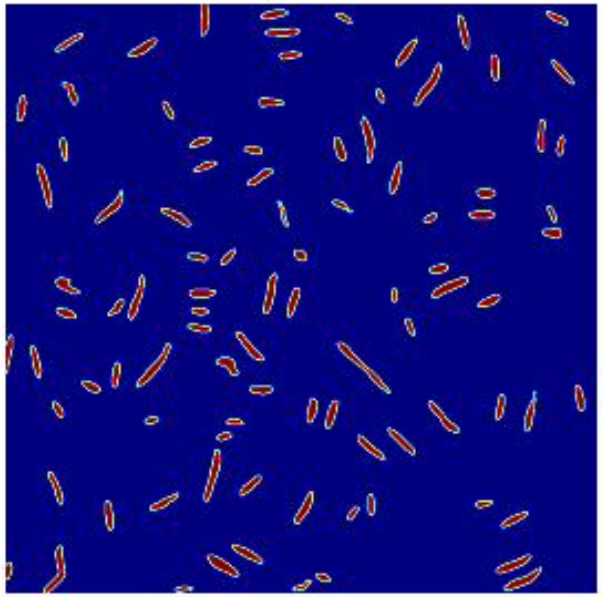
- Encoder-decoder type of architecture with spatial pyramid of dilated convolutions in the bottleneck layer
- Skip connections between encoder and decoder parts for mixing global and local information

Protein assembly

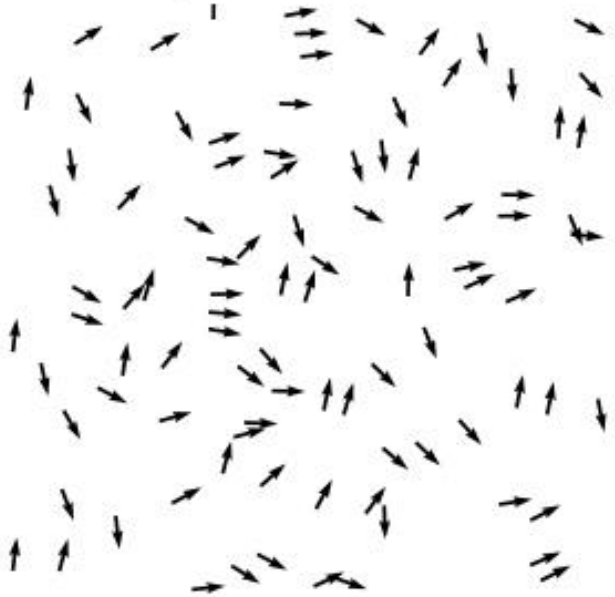
Experimental image



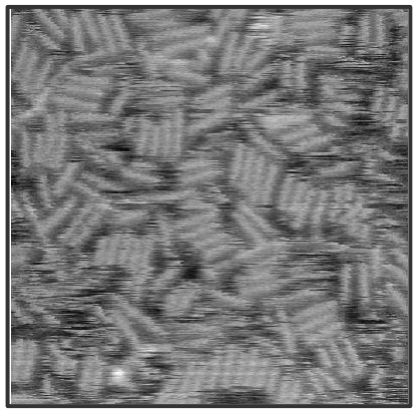
Model prediction

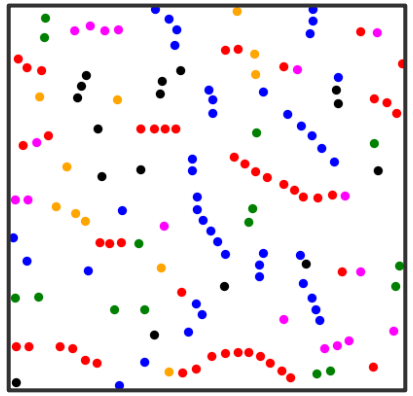


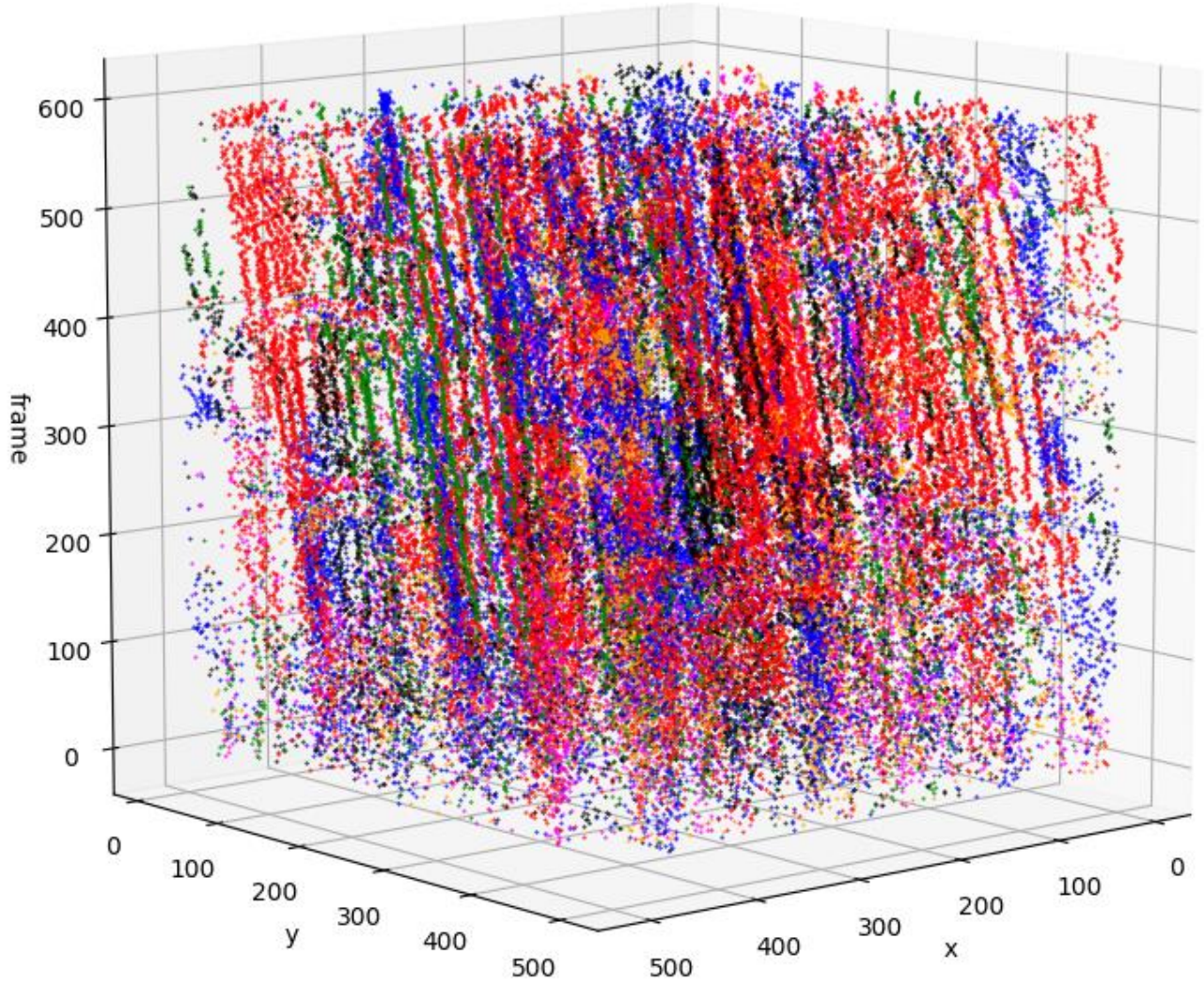
Proteins position and orientation



Protein assembly







Reconstructing spatio-temporal trajectories for all the detected particles in the movie, which show the evolution of different domains in time.