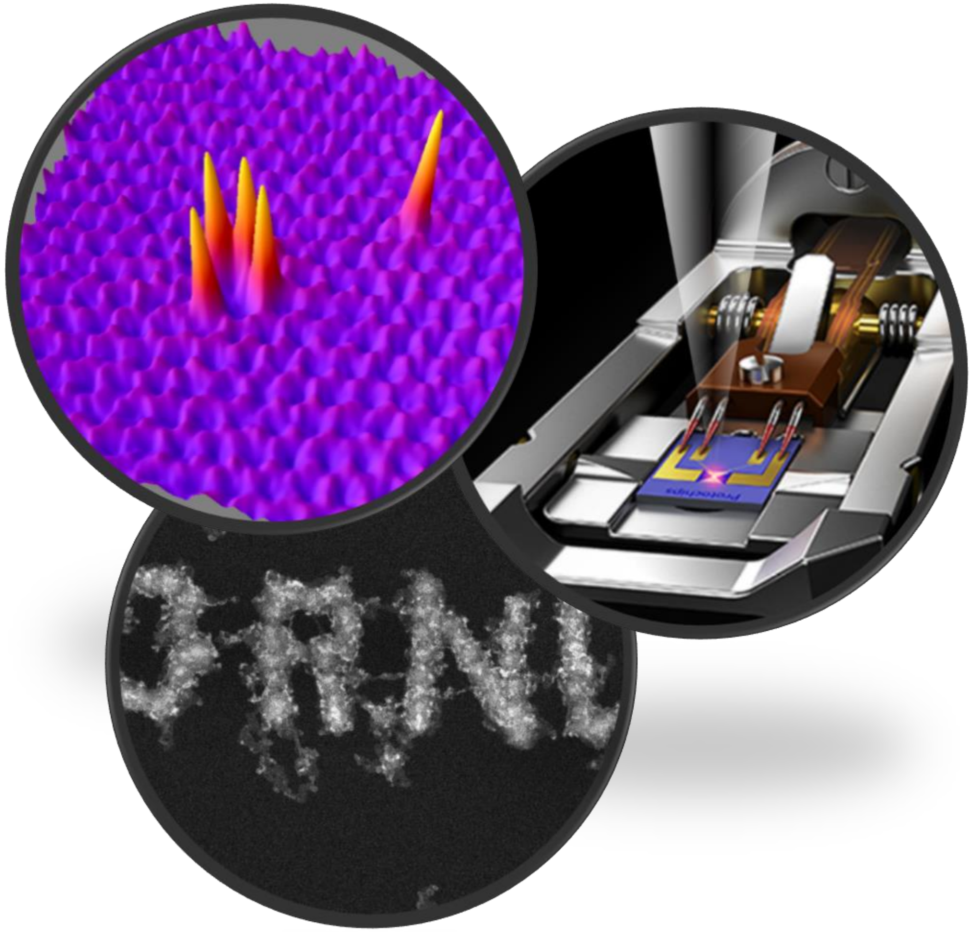
Automated experiments: Workflow design and optimization

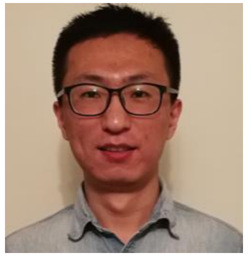


Sergei V. Kalinin

University of Tennessee, Knoxville

January 24, 2023

Mahshid Ahmadi



Yongtao Liu

Rama Vasudevan

Maxim Ziatdinov

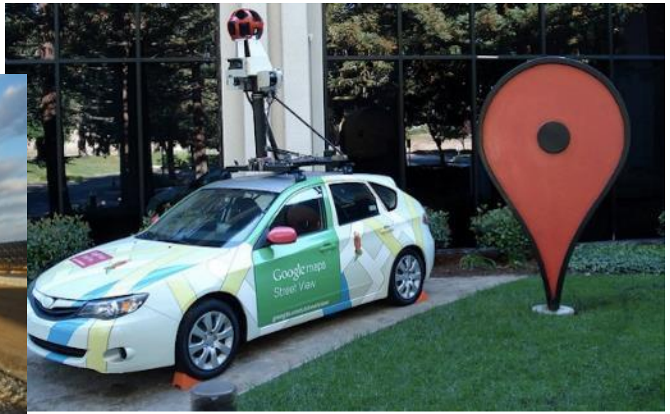
Kyle Kelley

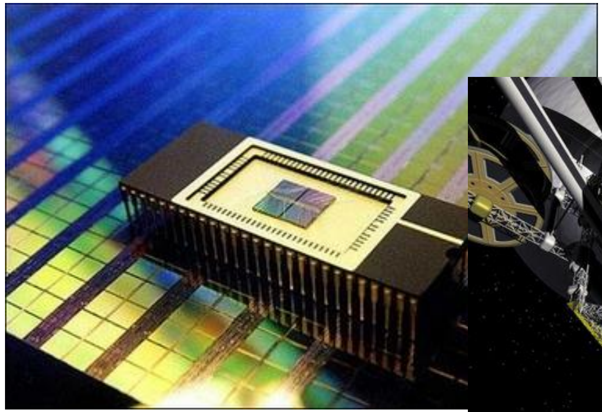
Ondrej Dyck

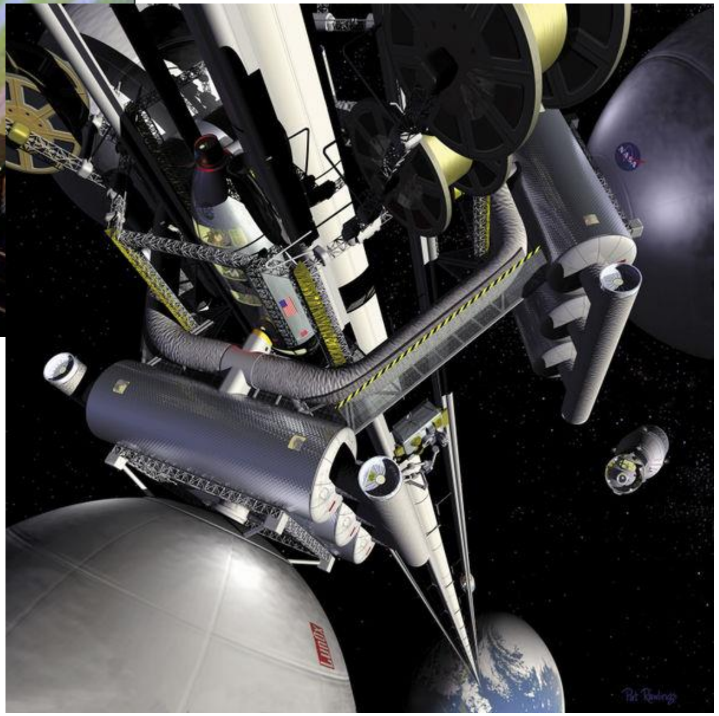
Kevin Roccapriore

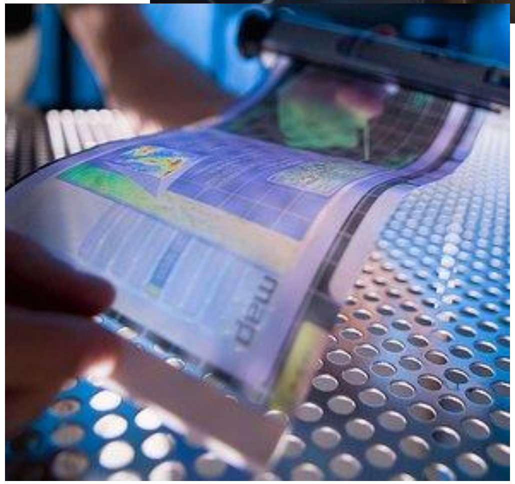
2002 - 2022
Since 2023
2022 -2023
amazon
SCANNING PROBE MICROSCOPY
Electrical and Electromechanical Phenomena at the Nanoscale
Sergei Kalinin
Alexei Gruverman
EDITORS
VOLUME 1
Springer
Sergei V. Kalinin
Alexei Gruverman
Editors
Scanning Probe Microscopy of Functional Materials
Springer
World Scientific Series in Nanoscience and Nanotechnology
7
Scanning Probe Microscopy for Energy Research
Dawn A Bonnell • Sergei V Kalinin
World Scientific
World Scientific Series on Emerging Technologies – Volume 1
Handbook on Big Data and Machine Learning in the Physical Sciences
In 2 Volumes
Volume 1
Big Data Methods in Experimental Materials Discovery
Editors
Surya R. Kalidindi • Sergei V. Kalinin
Turab Lookman
Editors-in-Chief
Sergei V. Kalinin • Ian Foster
World Scientific

The World is the Material Opportunity









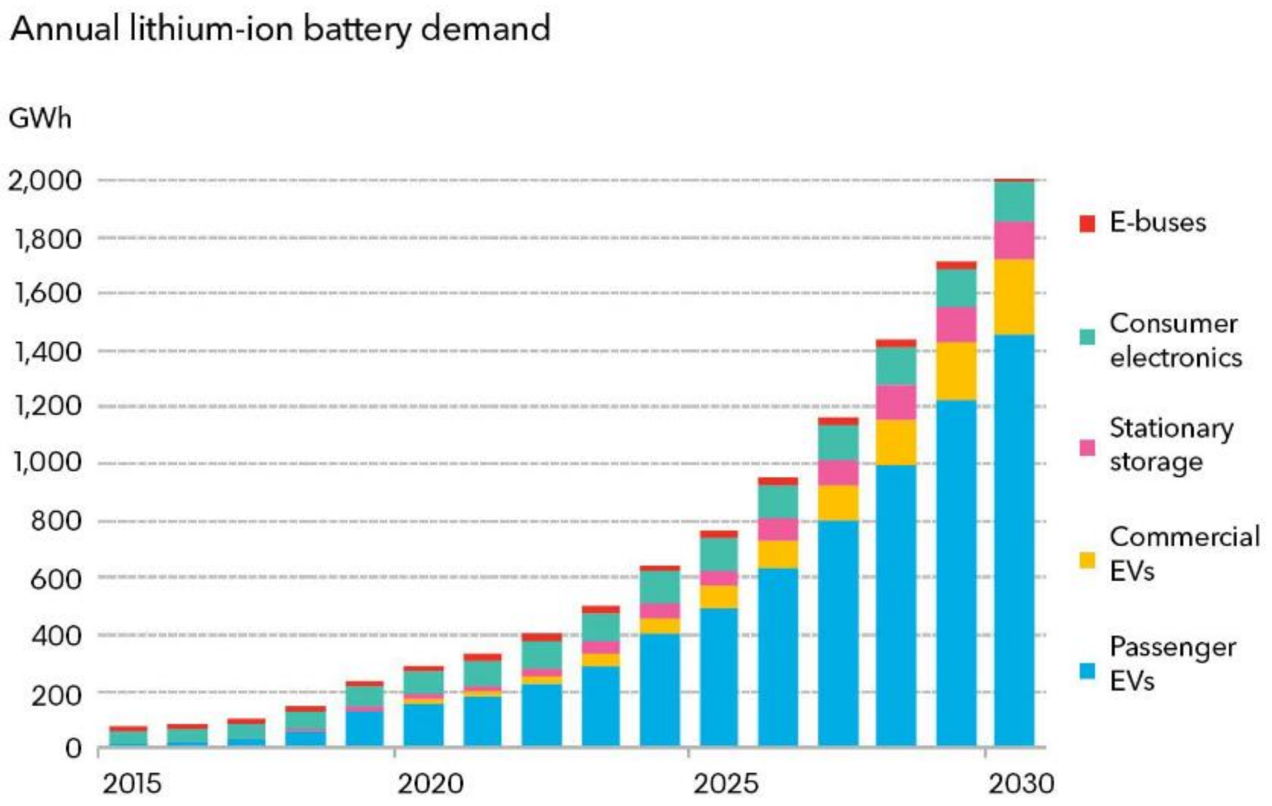
Predicting crystal structure by merging data mining with quantum mechanics

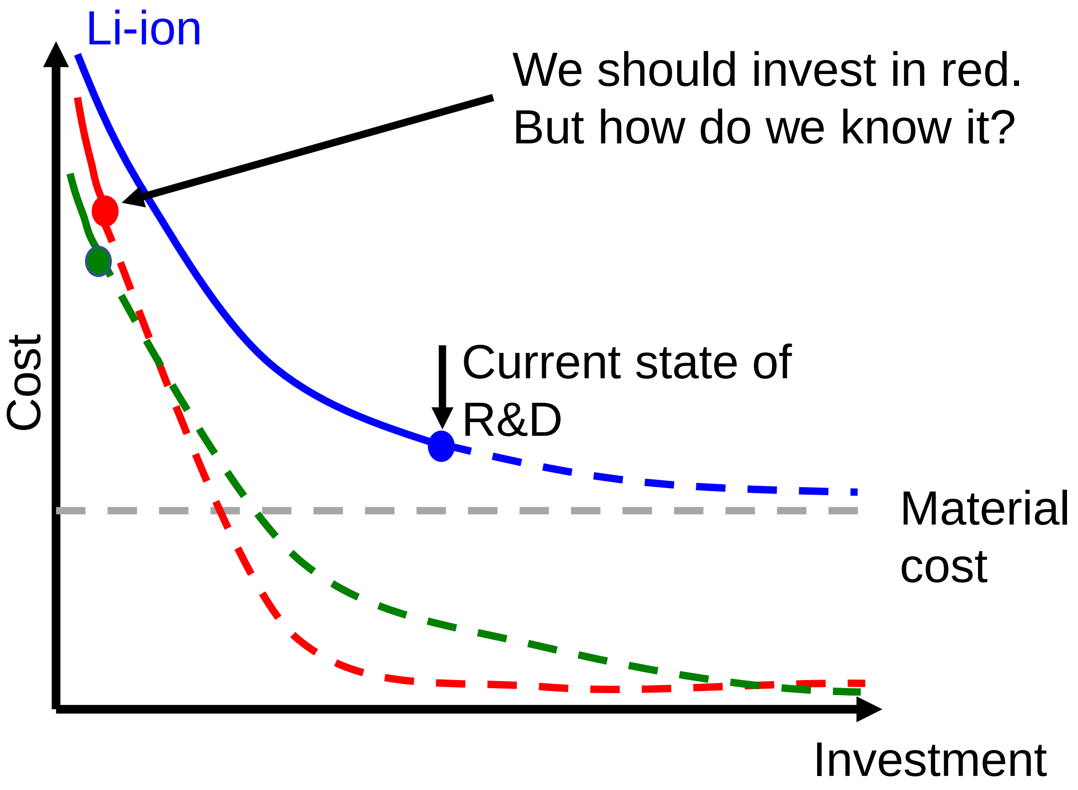
CHRISTOPHER C. FISCHER[1], KEVIN J. TIBBETTS[1], DANE MORGAN[2] AND GERBRAND CEDER[1]*
[1]Department of Materials Science and Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA
[2]Department of Materials Science and Engineering, University of Wisconsin, Madison, Wisconsin 53706, USA
*e-mail: gceder@mit.edu

- **“Improve”:** Renewable energy, self-driving cars, transparent displays, new memory technologies
- **“Discover”:** Room temperature superconductivity, high mechanical stress materials
- **“Engineer”:** Quantum computing, single-atom catalysts, biomolecules

Functionality, manufacturability, cost

Batteries: Li-ion and Beyond

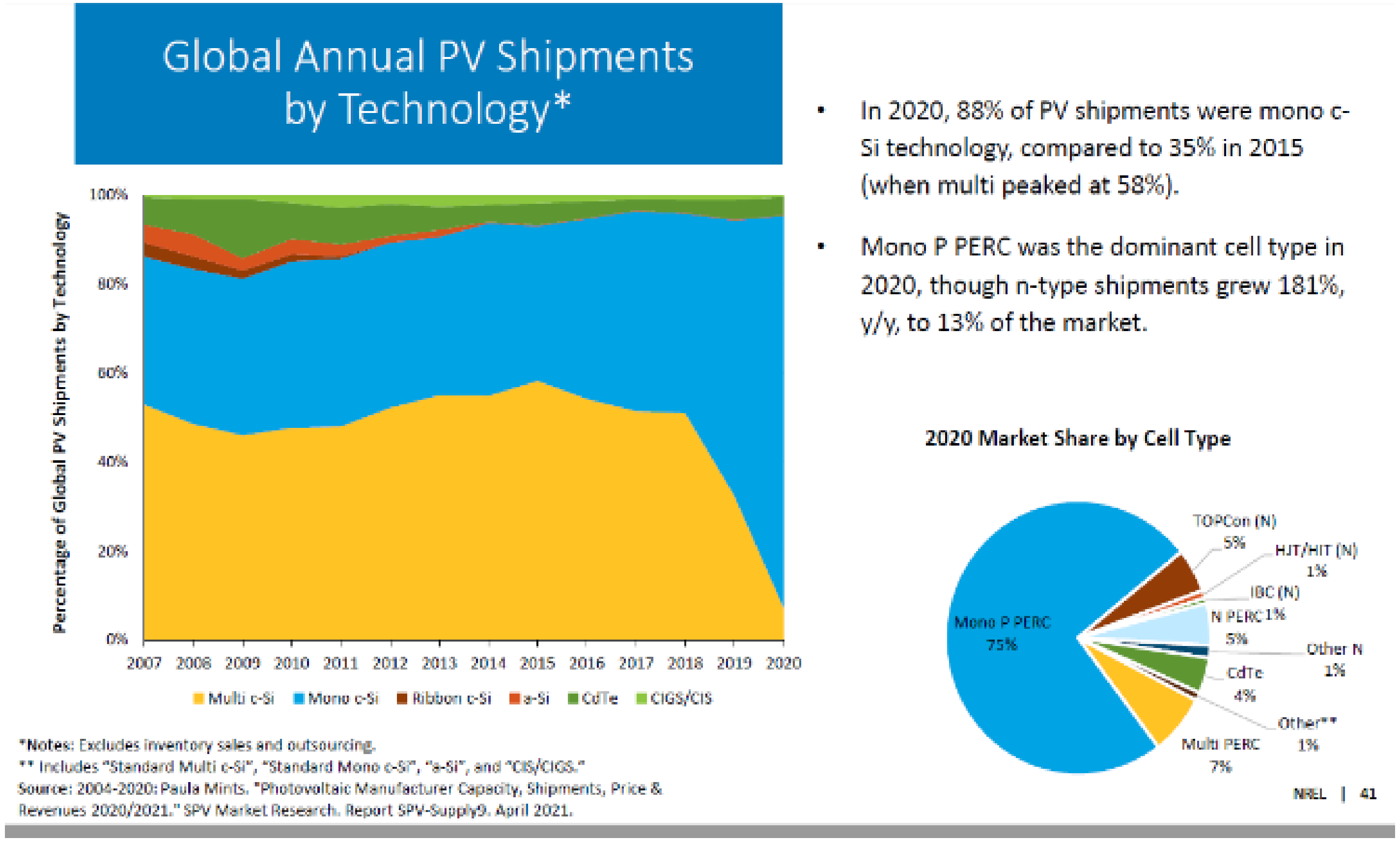


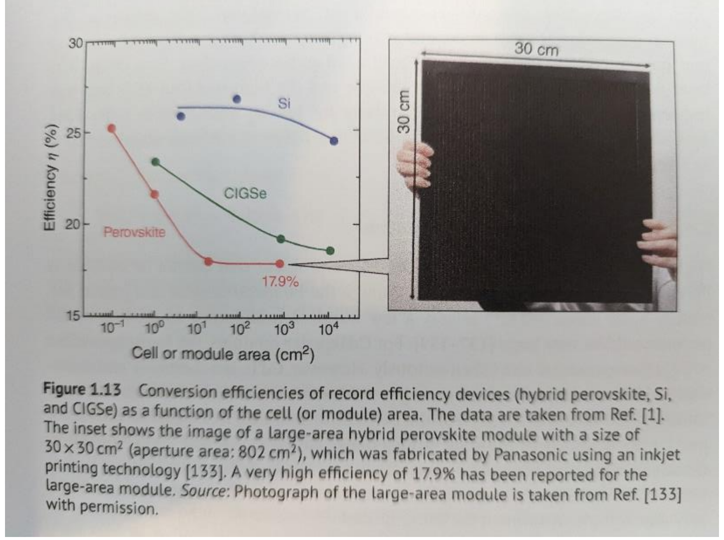


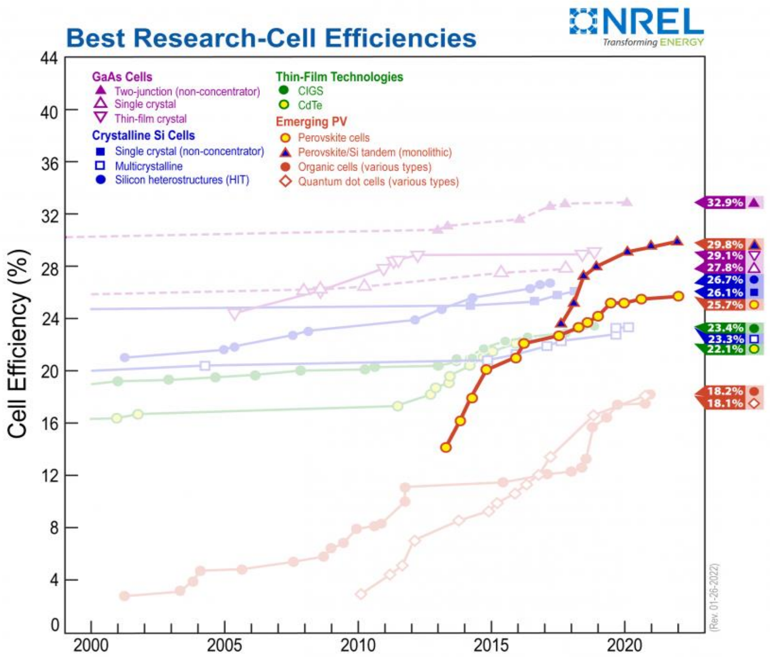
- Batteries are required element of energy transition (EVs, ESS, mobile devices)
- Currently Li-ion is the primary technology
- Optimization of Li-ion batteries takes years (even with same process on new Gigafactory)
- However, it is far from Goldilock zone for ESS or energy transport

- How can we optimize usage and safety for Li-ion batteries in EVs?
- How do we select beyond Li technologies for ESS?

Solar Photovoltaics: Will Silicon Ever Reign?

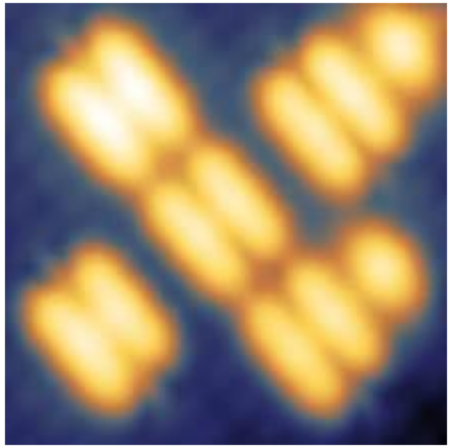




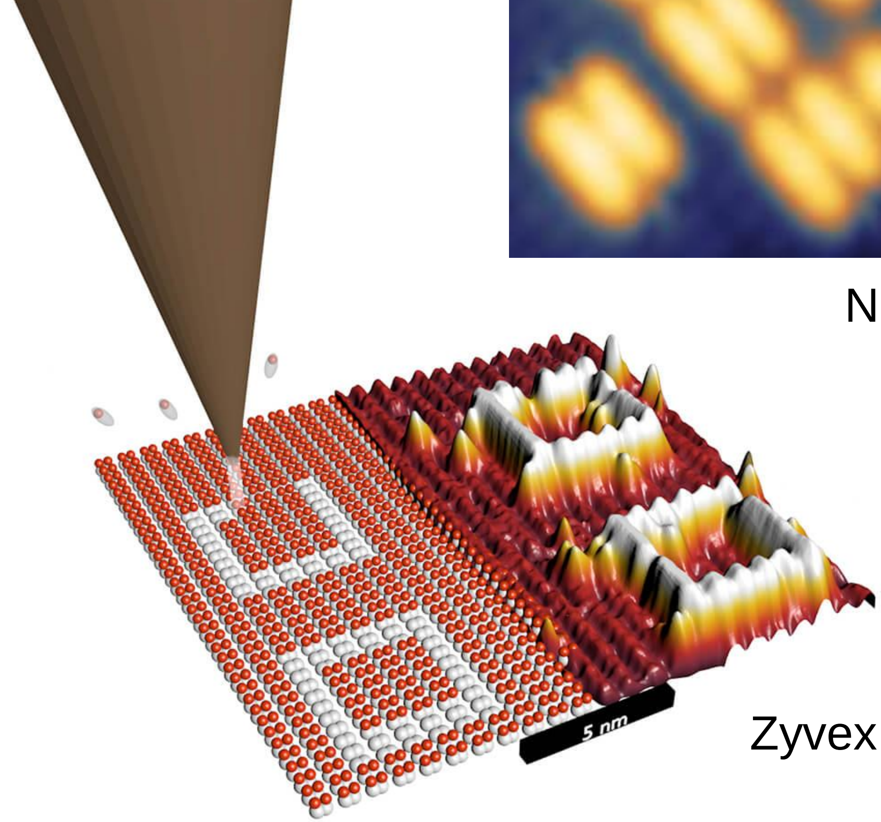


- Solar energy is the fastest growing energy sector
- Si is now reining material – however, it is really not the optimal material for PV (heavy, expensive)!
- Hybrid perovskites can be used as ideal PV materials – if we can make them stable and scale manufacturing!

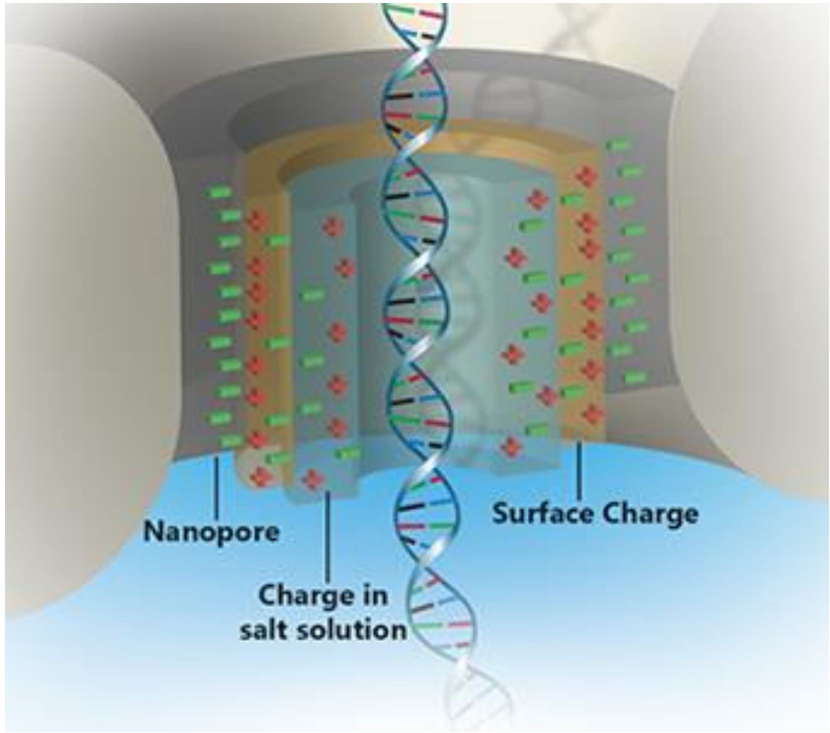
Quantum Computing and Single-Molecule Biology



NIST



Zyvex Labs



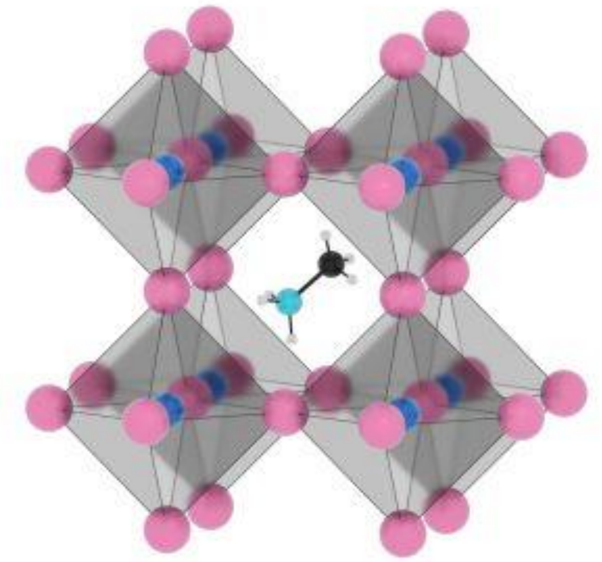
Oxford Nanopore

- Direct atomic fabrication: quantum communications and quantum computing, environmental sensing
- Single-molecule biological devices

- Success story 1: cryo-electron microscopy
- Success story 2: nanoelectron diffraction

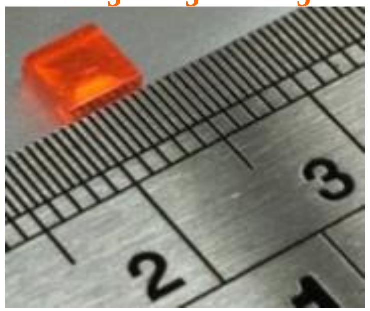
Why Metal Halide Perovskites

ABX_3

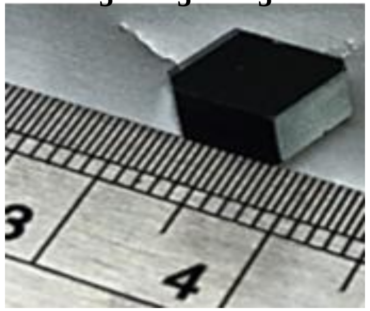


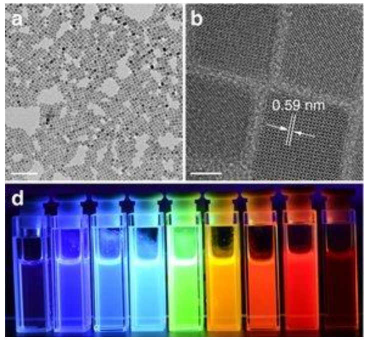
Solution processable thin films and single crystal growth

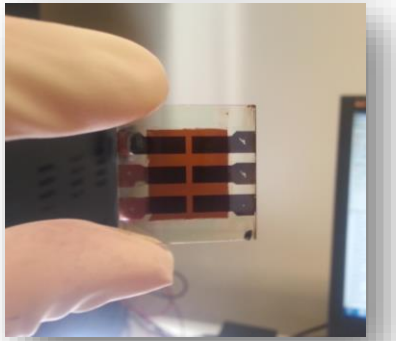
$CH_3NH_3PbBr_3$



$CH_3NH_3PbI_3$

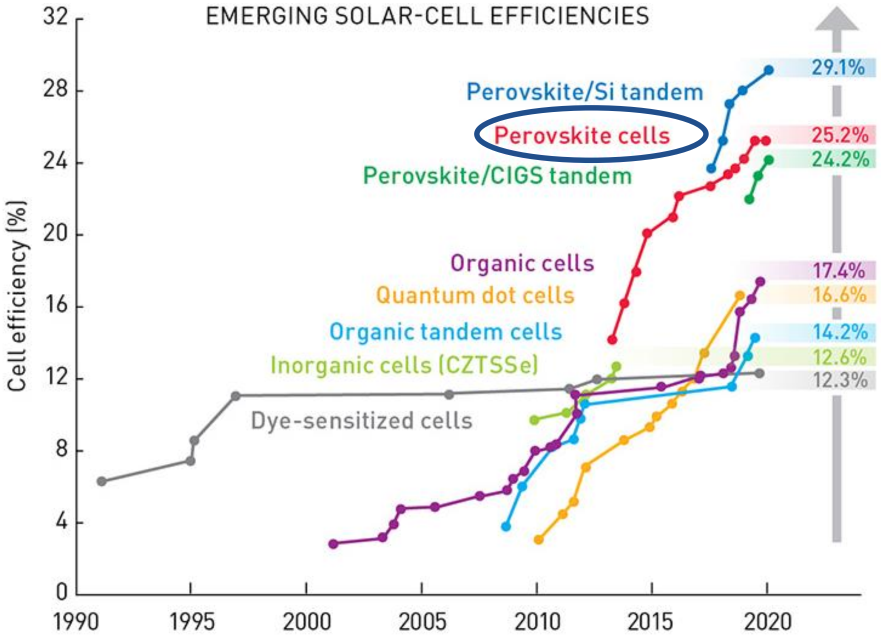






Properties

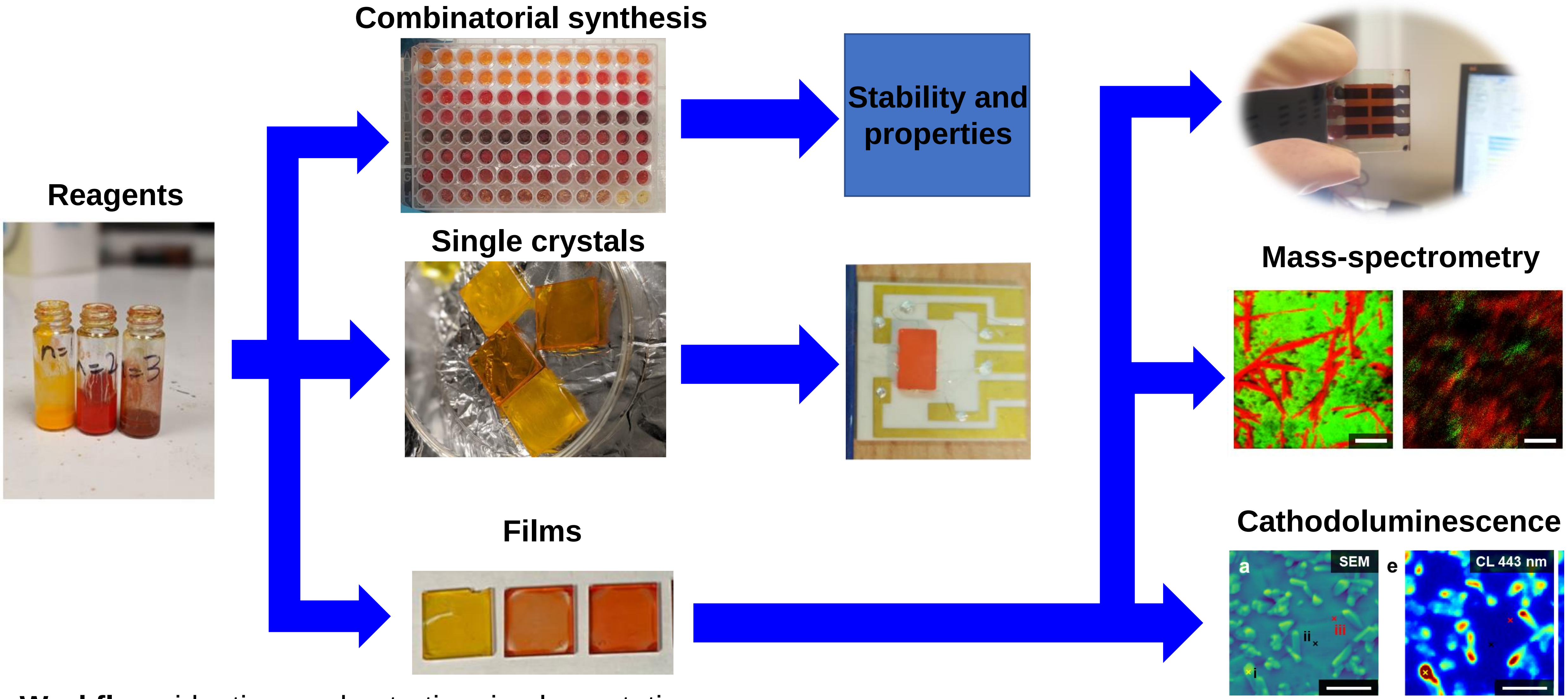
- ✓ Mixed ionic, electronic conductivity
- ✓ Low defect density (Defect tolerance)
- ✓ High optical absorption
- ✓ Moderate mobility
- ✓ Long carrier diffusion length



Applications

- ❑ Solar Cells
- ❑ Photodetectors
- ❑ Light Emitting Diodes
- ❑ Ionizing Radiation Sensors
- ❑ Memristors

What is a workflow?

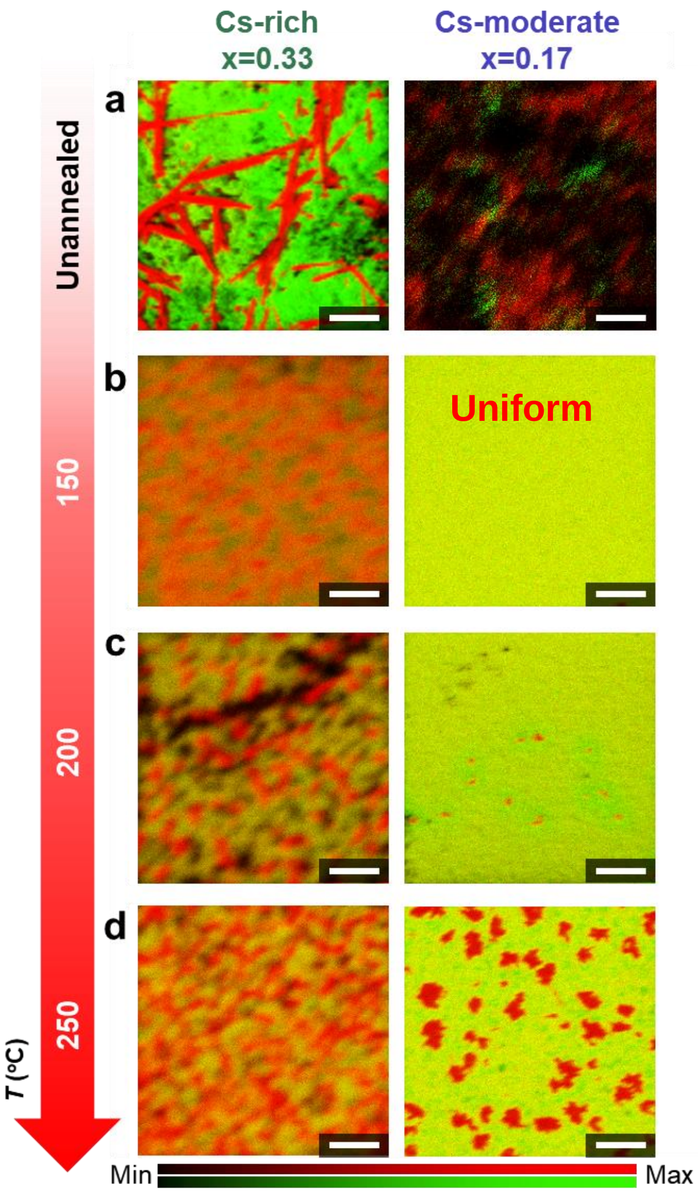


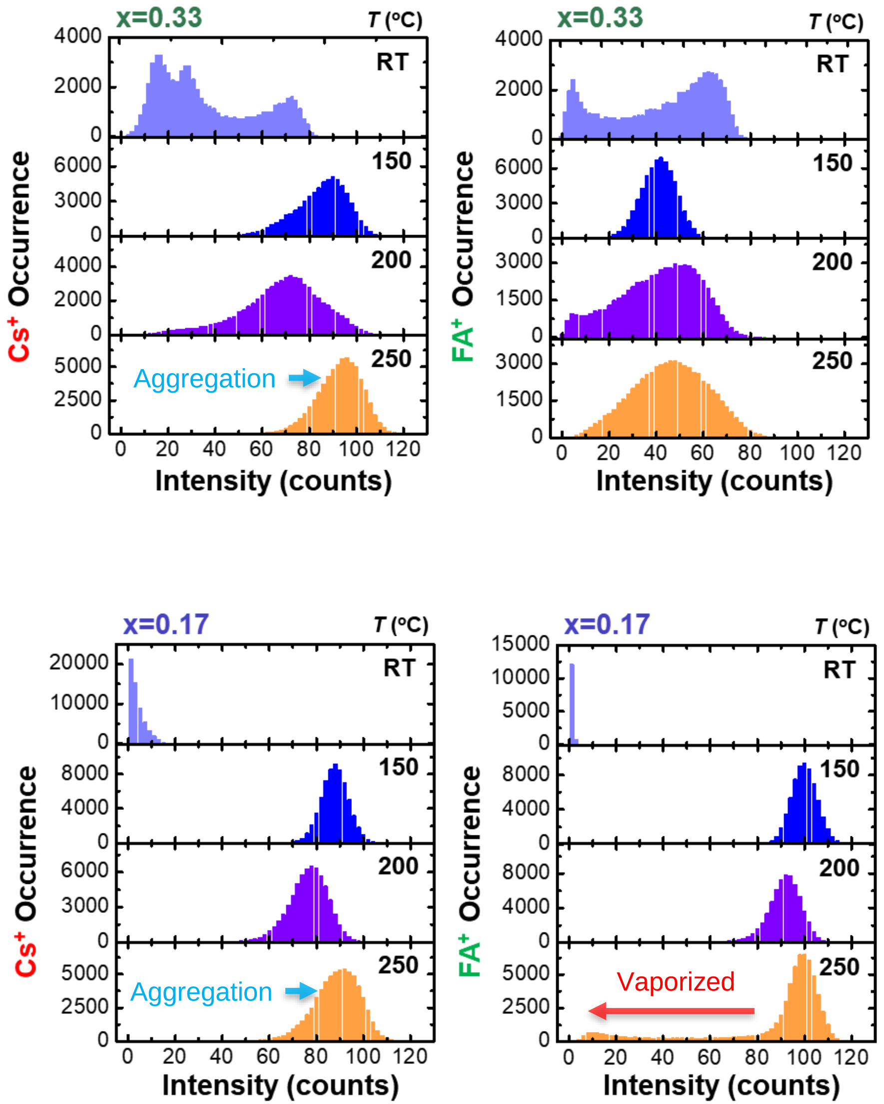
- **Workflow:** ideation, orchestration, implementation
- Domain specific language
- Dynamic planning: latencies and costs
- Reward and value functions

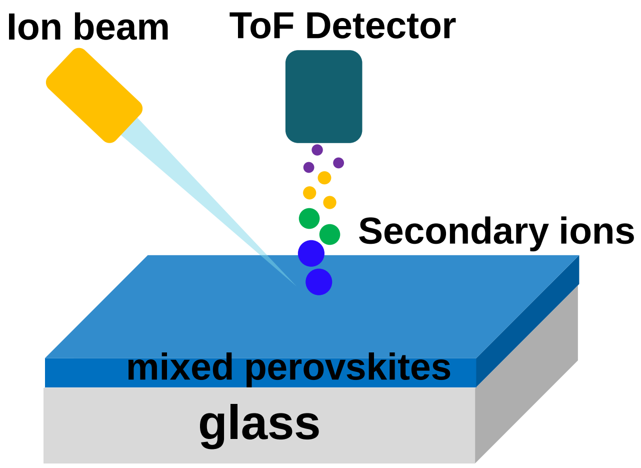
Workflows are often designed in academic labs and adopted by industry.

Chemical Exploration of MHP: ToF-SIMS

Spatial Distribution & Histogram of A-cations: $Cs_xFA_{1-x}PbI_3$ MHPs







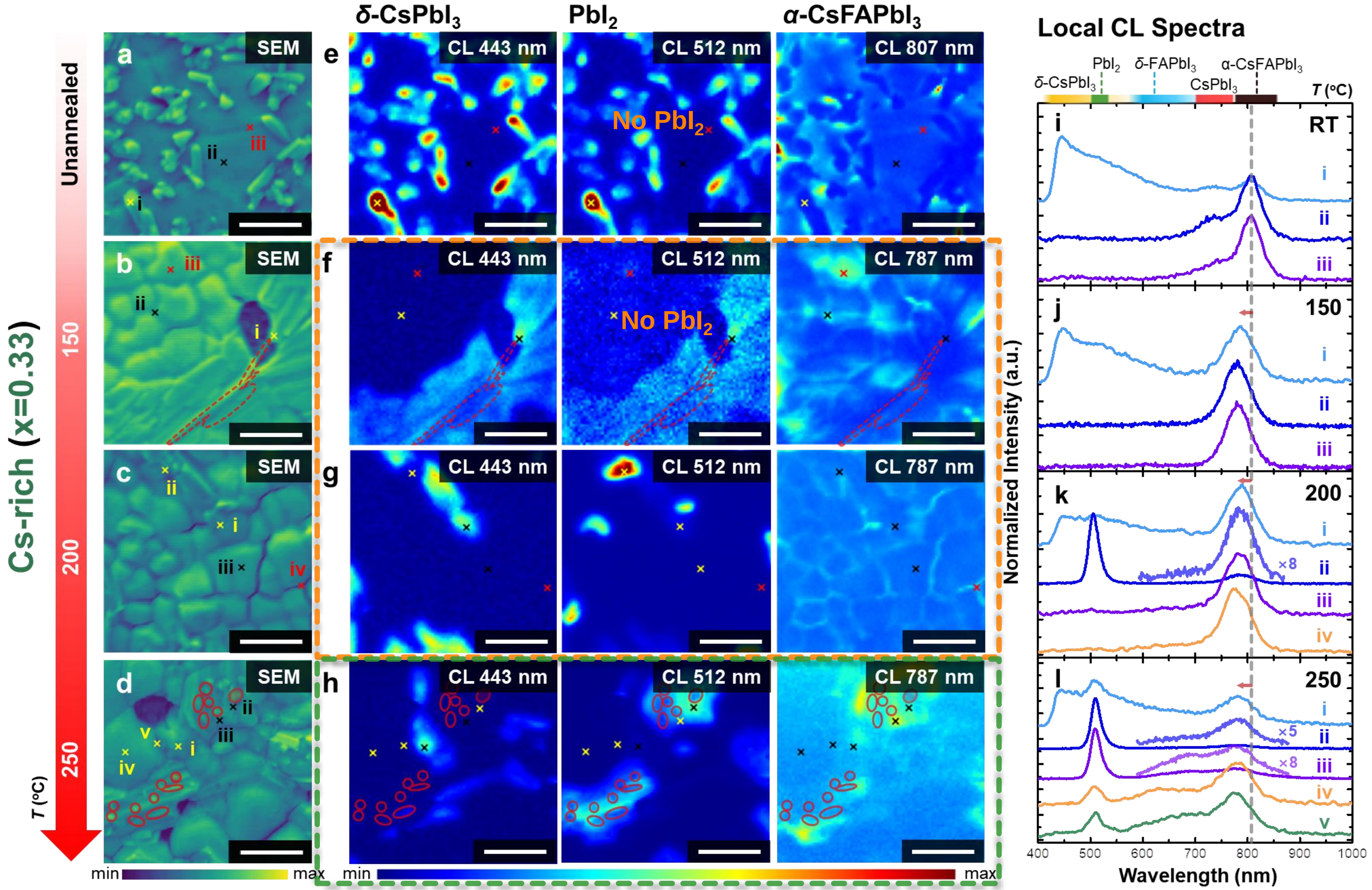
Key observations

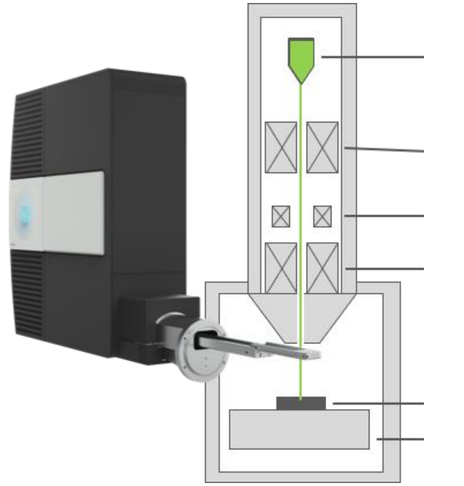
1. Highly disordered MHP structure w/o thermal annealing, mainly responsible for **CsI**
2. Thermal annealing promotes the reorganization of A-cations
3. FA evaporate at high *T*
4. Appropriate **Cs** can regulate the local inhomogeneities

Physical Exploration of MHP: Cathodoluminescence

Identification of Local Phase Inhomogeneities

*Pixel width: 40 nm, Scale bar: 1 μm





*Red-marked region: CL-inactive (non-emissive), Assigned as CsI

1. **CsI → δ-$CsPbI_3$ → Alloyed α-$CsFAPbI_3$**
2. **δ-$CsPbI_3$ → Alloyed α-$CsFAPbI_3$ → +PbI_2 (allowed)**
3. ***33% of Cs is inferior to formation of*** **α-$CsFAPbI_3$ due to FA evaporation**

What do we hope to achieve?

Microscopy today:
- Primary component of research in materials science, physics and biology
- 1000s of high-end (S)TEM platforms, ~10,000 overall
- 1000s of high-end UHV SPMs, >50,000 ambient
- Chemical and mass-spectrometric imaging

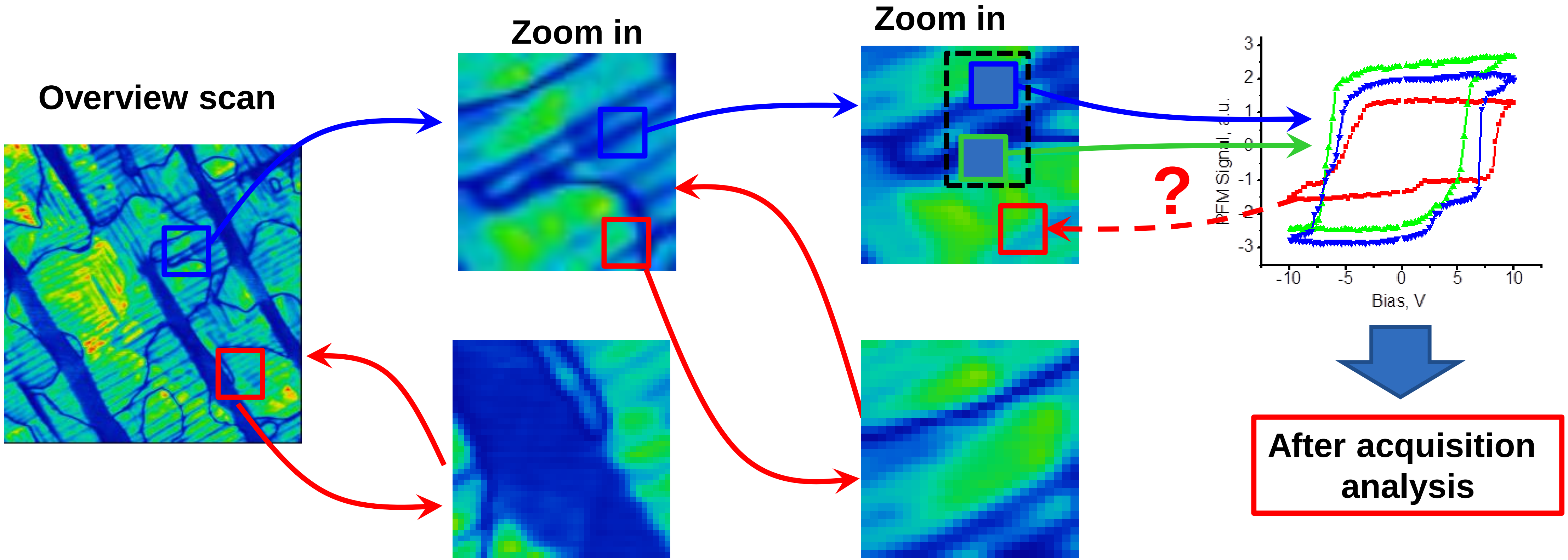
What do microscopists do?
- Most of the time - sit alone in the dark room and turn knobs 😏
- Limited amount of collected data
- Case for automation: CryoEM

Unsurprisingly, inspired by autonomous cars, etc. – multiple proposals to make automated microscopes!

July 2019

Workflows in Scanning Probe Microscopy

Workflow plane

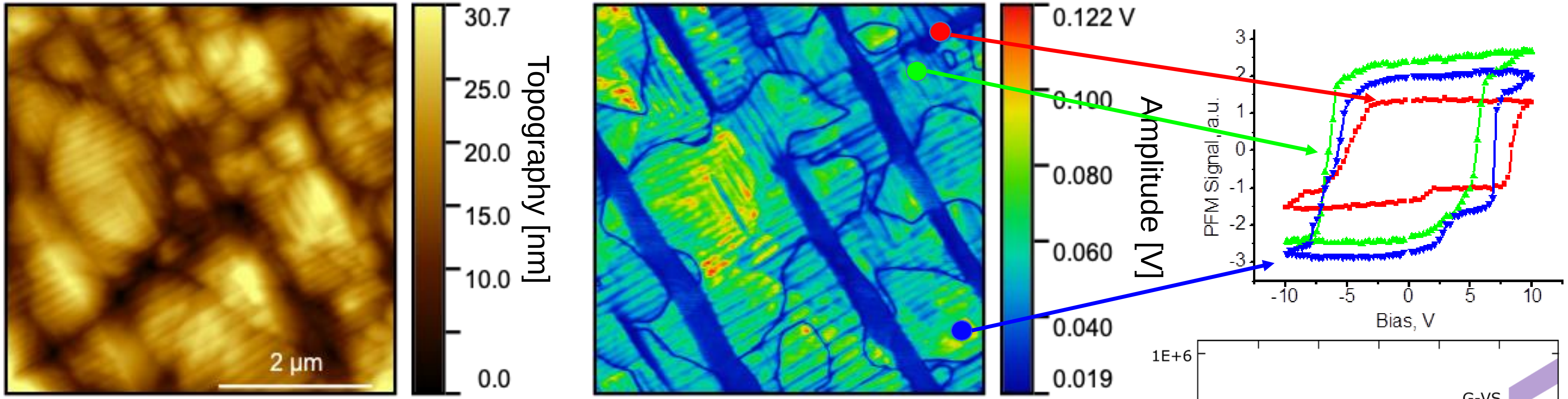


Instrument plane

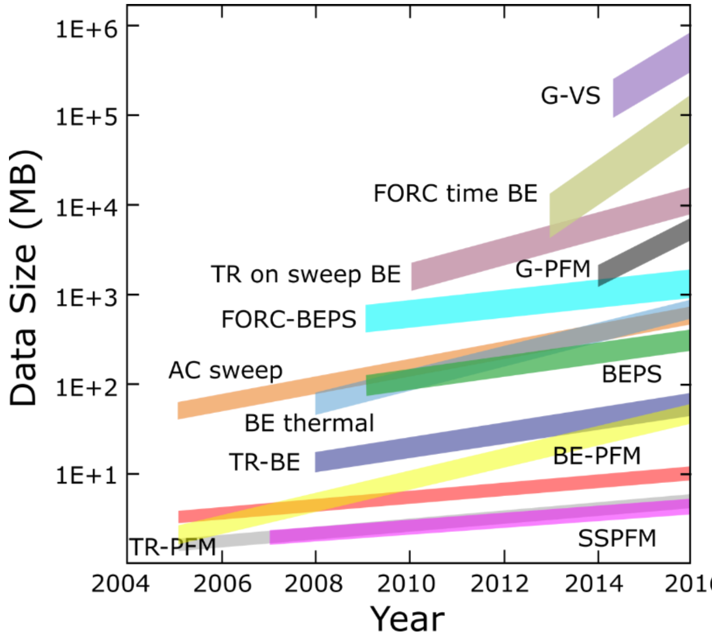
Minimal instruction set control language

Tune microscope	Initiate scan (parameters)	Position probe (x,y,t)	Initiate spectrum (parameters)	Stream data

Why Automated Experiment?

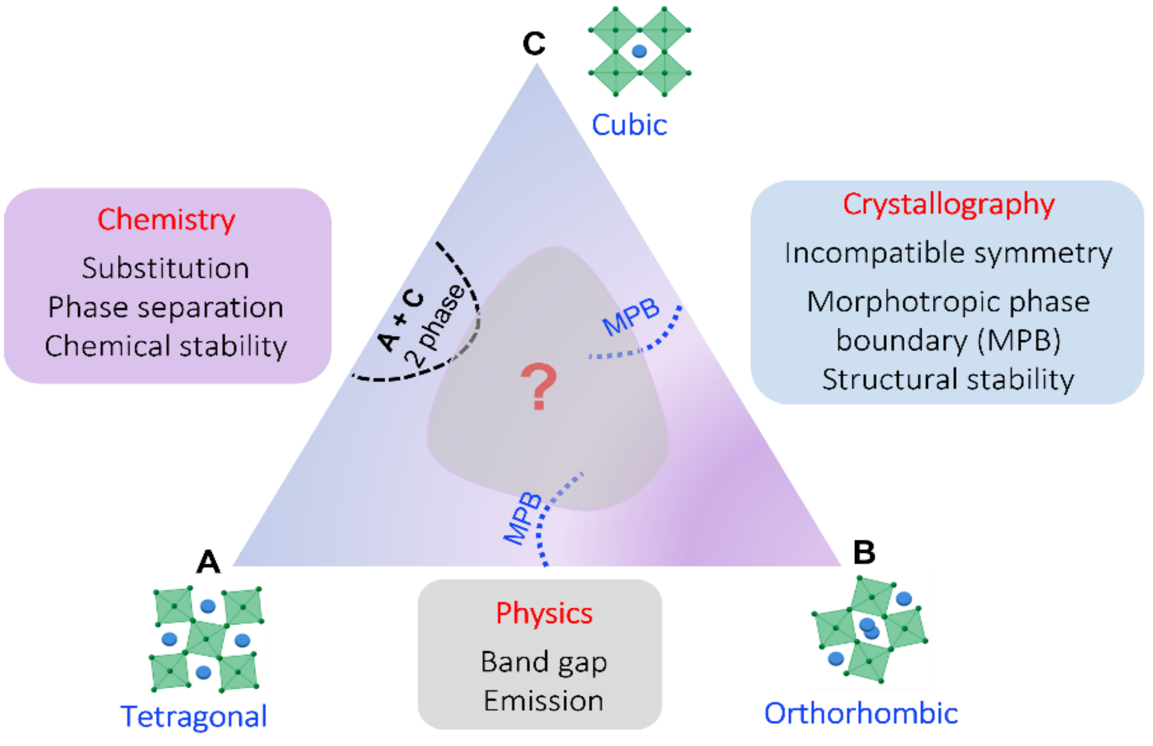


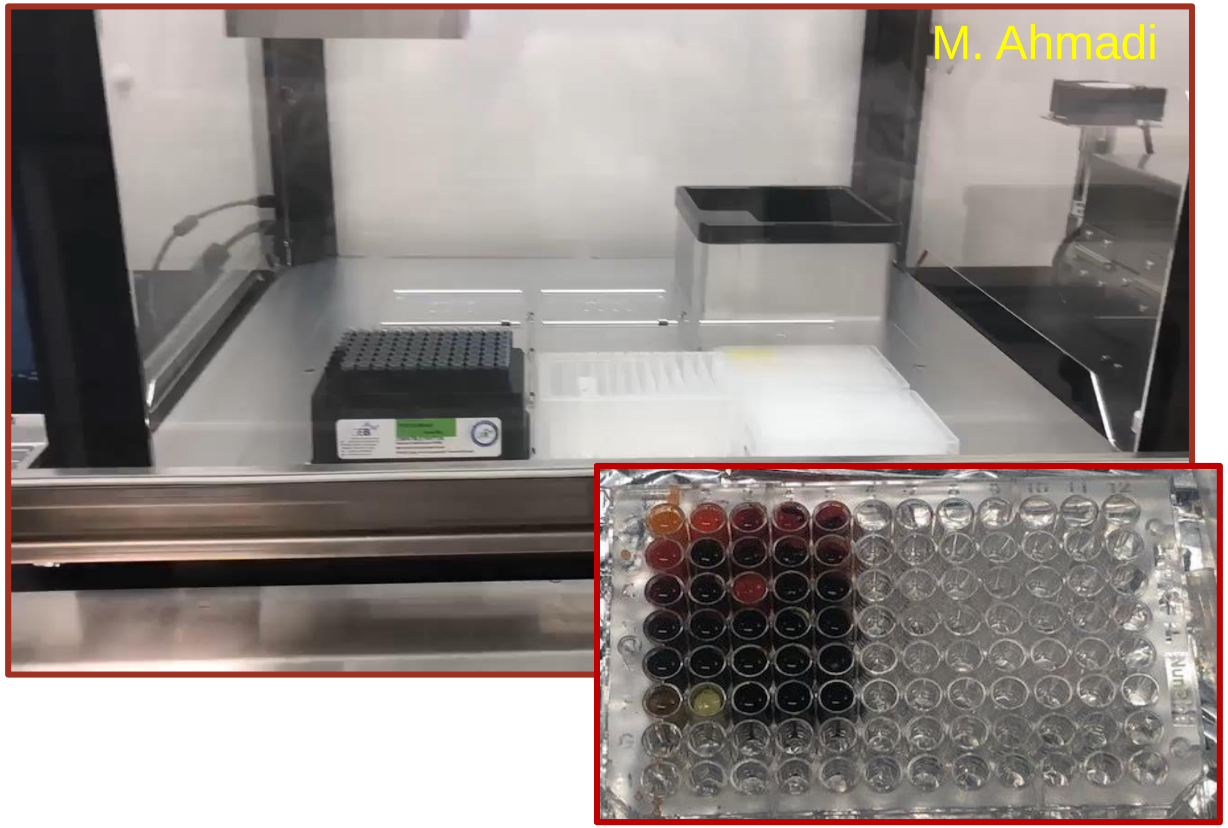
- Interesting functionalities are expected at the certain elements of domain structure
- We can guess some; we have to discover others



- **Can we run experiment so that we either explore**
  - **only selected regions,**
  - **discover new functionalities, or**
  - **seek functionalities we want?**

Why synthesis (or theory)?

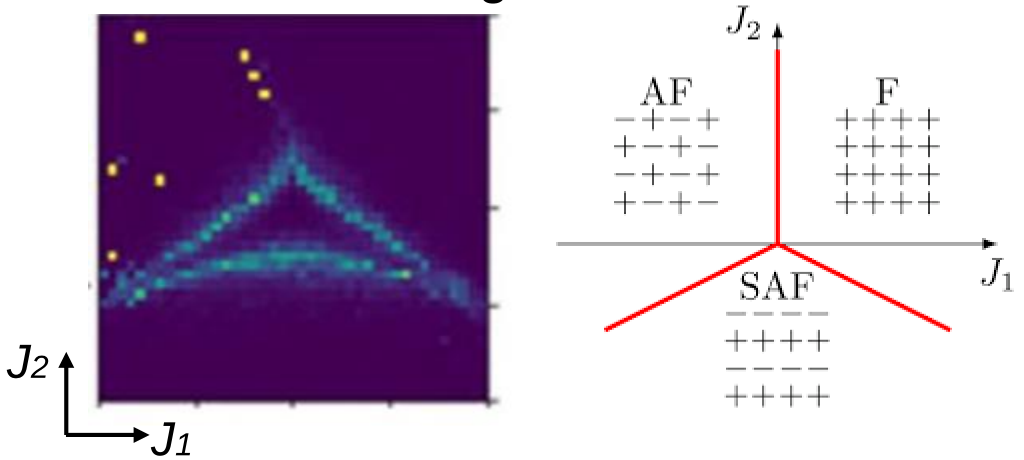




- Automated synthesis in its simplest form requires some way to navigate phase diagrams
- In more complex form, processing space.
- Ideally, incorporate physical knowledge

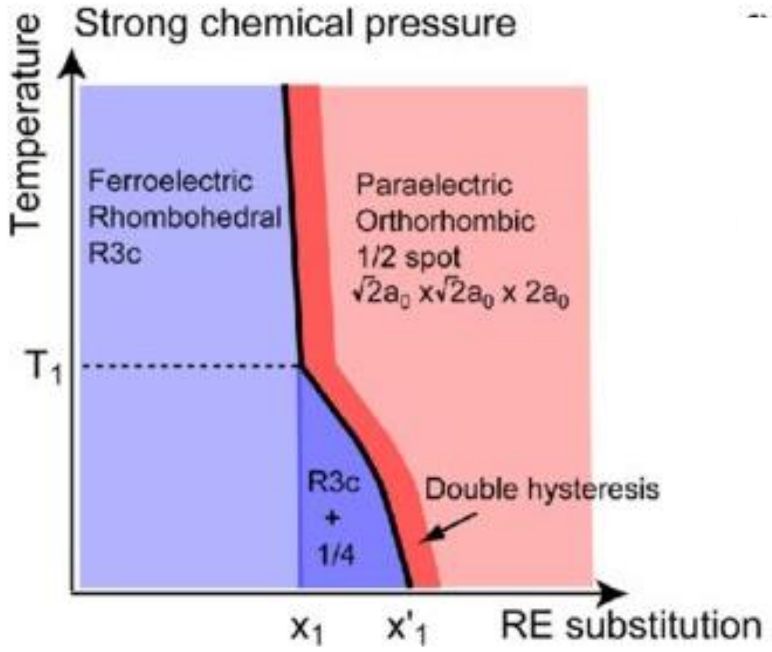
- Similar problem - theory

Ising model

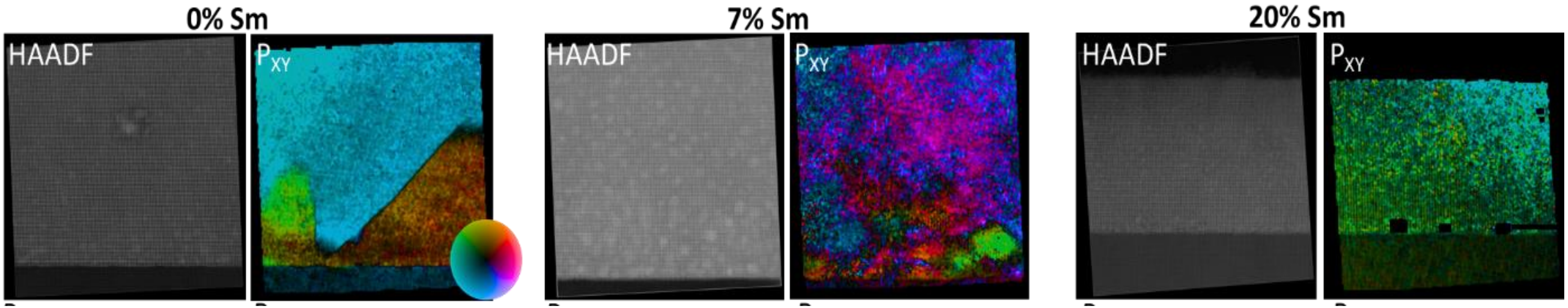


Combinatorial library

Sample by I. Takeuchi, UMD
Phase diagram by N. Valanoor et al.



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



arXiv:2004.11817

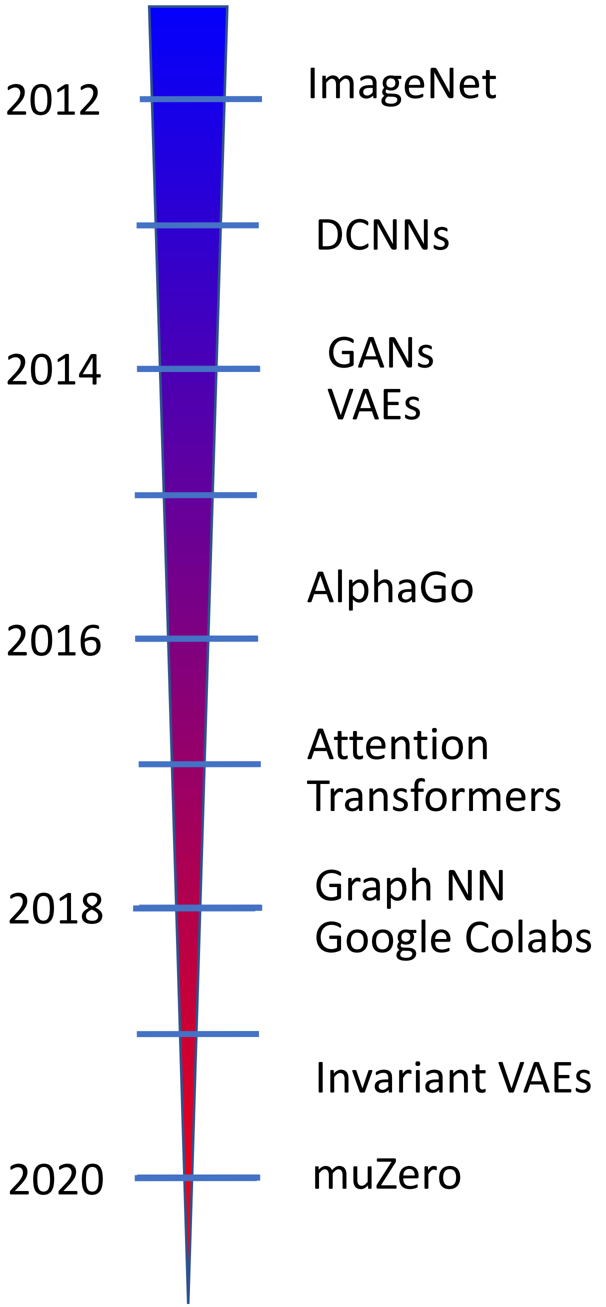
Why Machine Learning?

- Last decade has experienced an explosive growth of machine learning and artificial intelligence applications
- These developments have spanned areas from computer vision to medicine to autonomous systems and games
- However, the progress and impact as applied to experimental physical sciences has been minimal….

Why is it difficult?

- Requires domain expertise and domain-specific goals
- Deeply causal and hypothesis drive nature of domain sciences
- No single answer: culture, not a method
- Infrastructure, open code, open data
- **Most important:** active nature of scientific proces

Microsoft: GitHub
Meta: Open Catalyst,
Meta: Papers with Code
Toyota: TRI
Google: AlphaFold
NVIDIA: protein folding

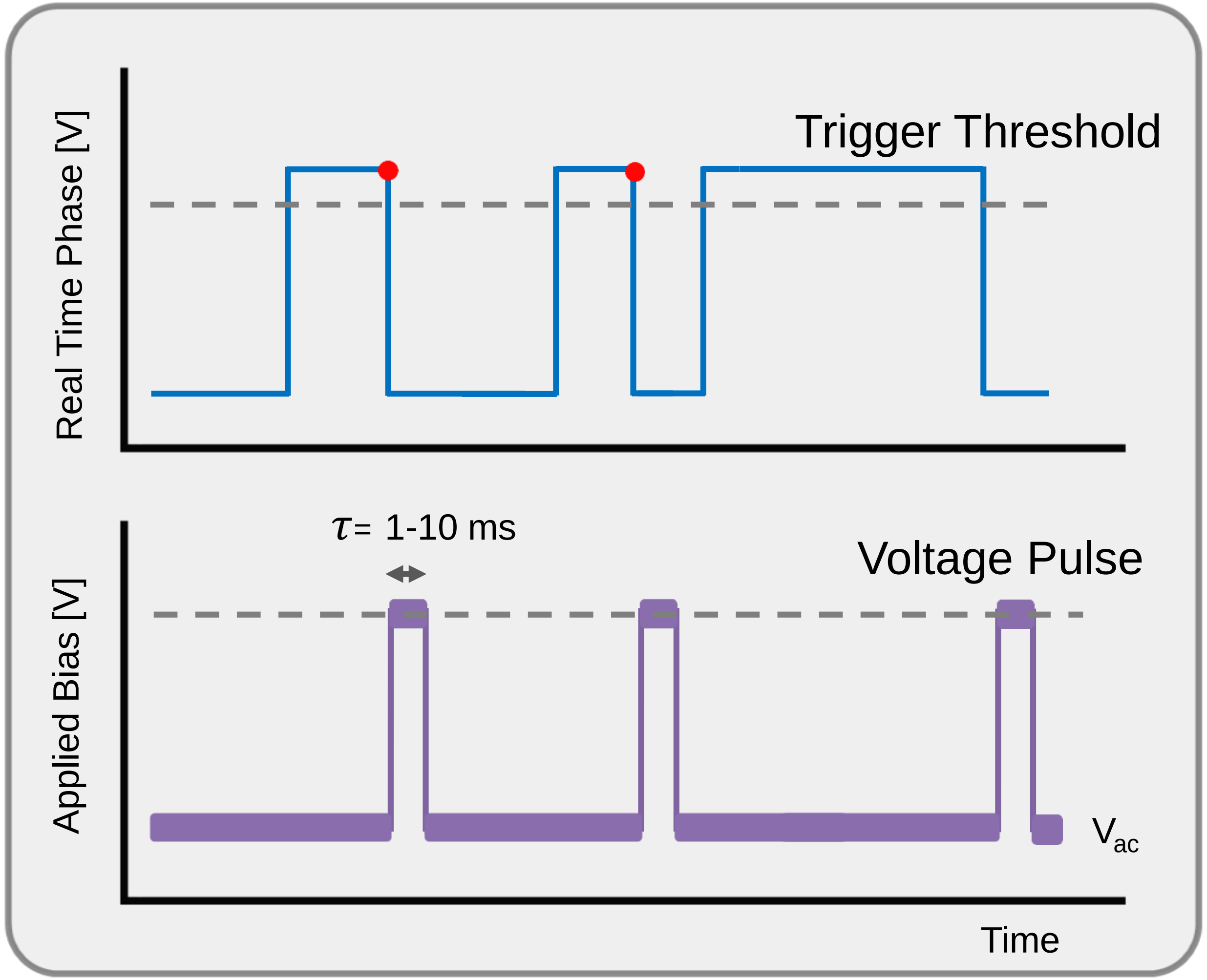


Automated Experiment:

almost easy…. If you know what you are looking for

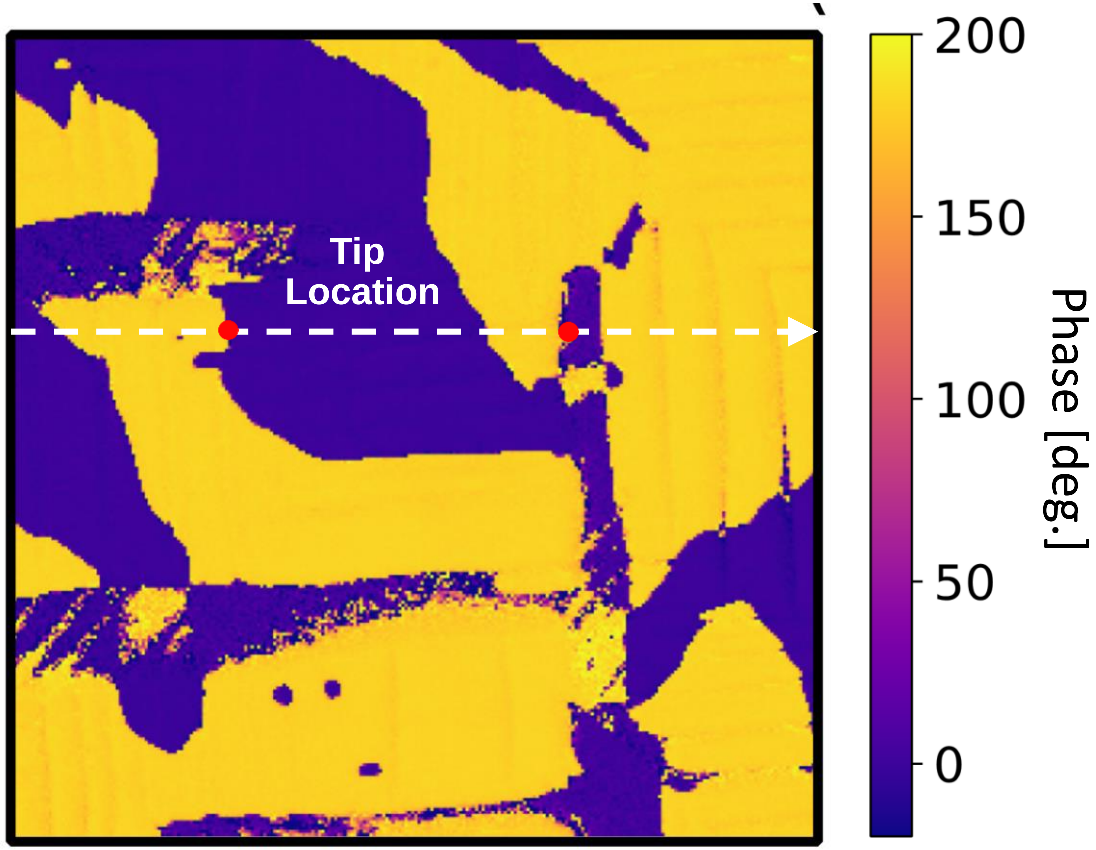
FerroBOT: Image-based feedback

Real-time Feedback



Example PFM Phase

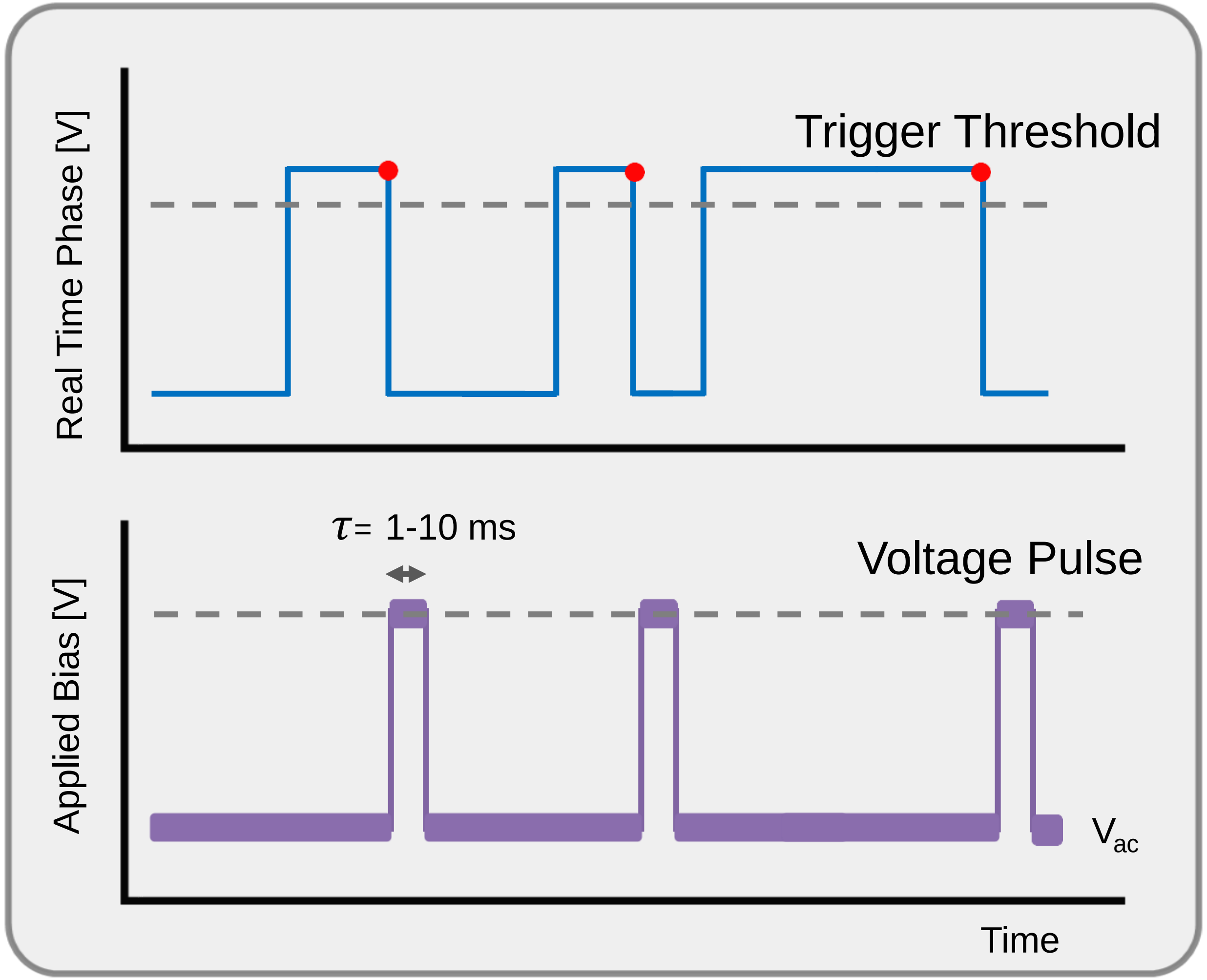
- FPGA – Labview – Matlab framework
  - Driving with AFM



Reduction of "purple" domain

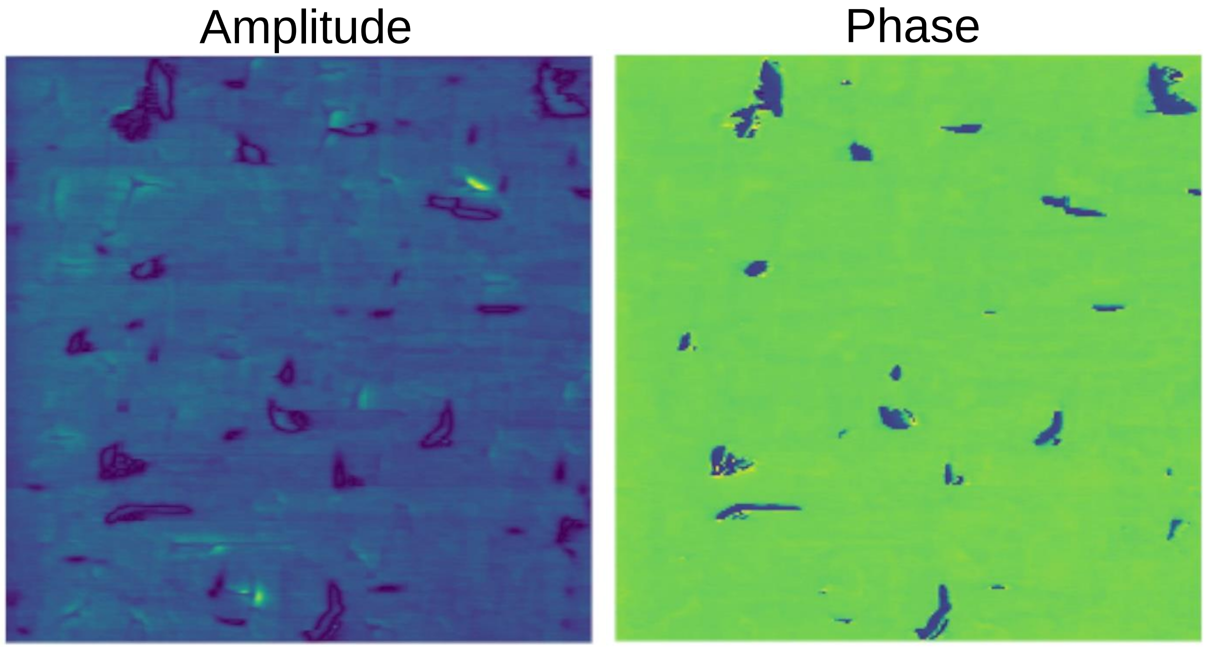
FerroBOT: single action table

Real-time Feedback



Real-time Manipulation

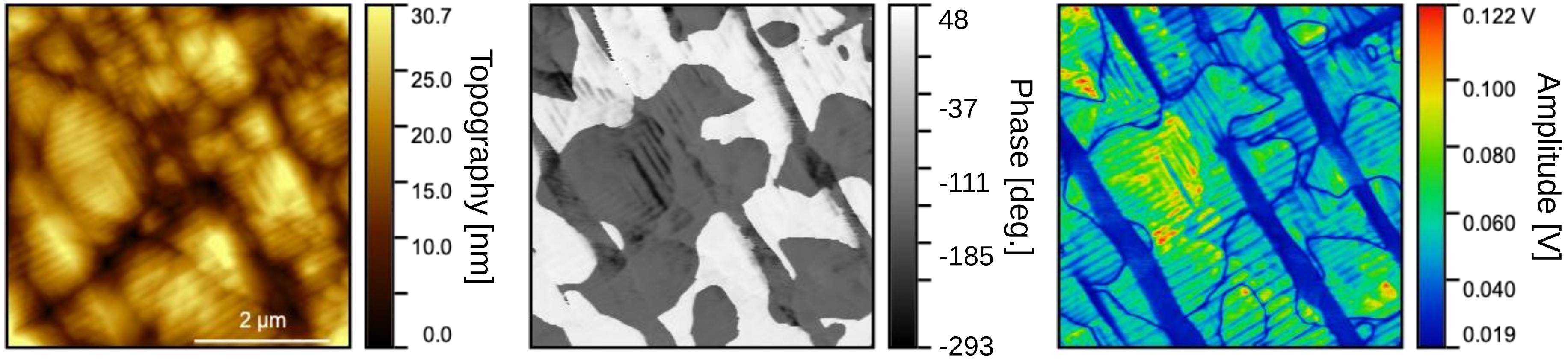
- FPGA – Labview – Matlab framework
  - Driving with AFM
- $BiFeO_3$: map DW energy landscape
  - Pulsing 5 ms, 1.5V, vacuum, domain wall growth



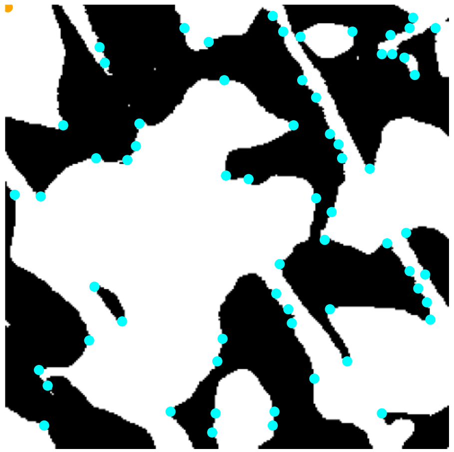
Realtime Feature Finding: $PbTiO_3$

- Cypher microscope
  - Ethernet connection for transmitting locations
- Skimage corner finding
  - Thresholded image feed
- Python – LabView framework
- Outlook
  - IV curves on domain walls
  - Ferroelastic domain wall probing

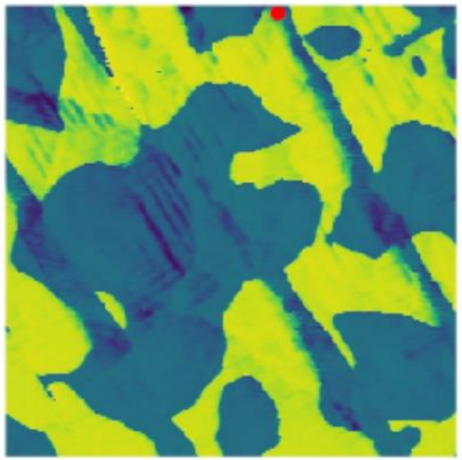
Initial State via Vertical Piezoresponse Force Microscopy



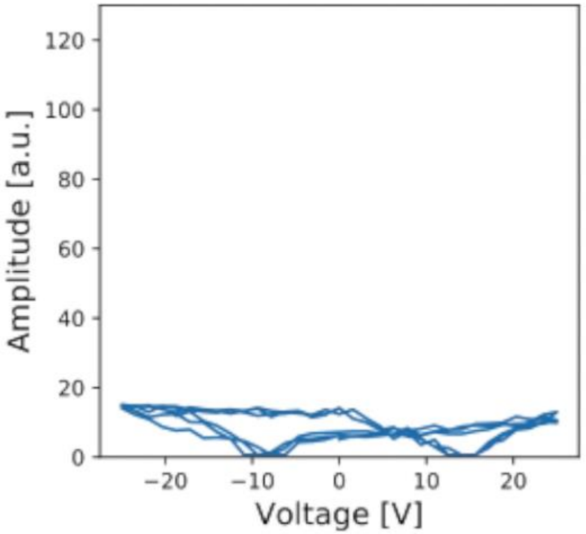
Feature Recognition



Tip Location



BEPS at Features



Here, we use simple computer vision algorithms to explore polarization switching in the regions predefined by operator based on prior knowledge

File Edit View Insert Cell Kernel Widgets Help

Trusted gpuenv

Run Code

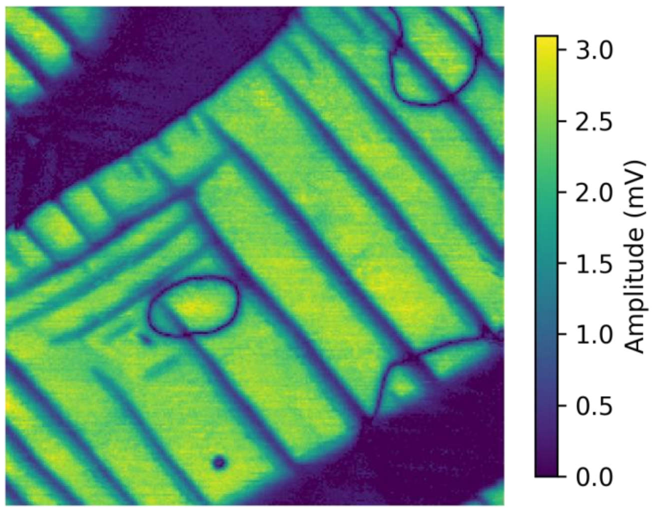
```
move_(-volt*2-(offsetvx), 0, 0-offsetvy, 0, move_speed)
```

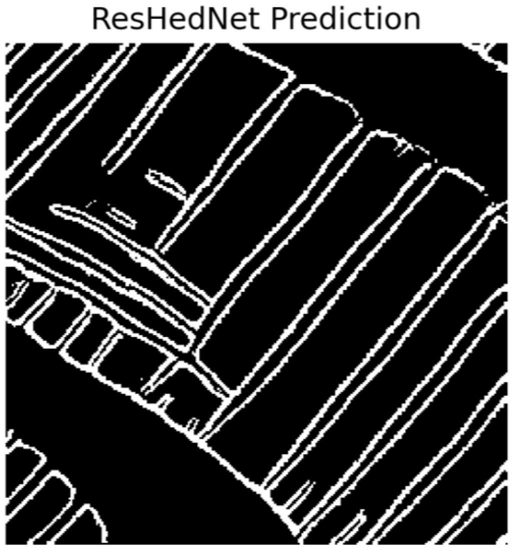

```
scanning line #56
```

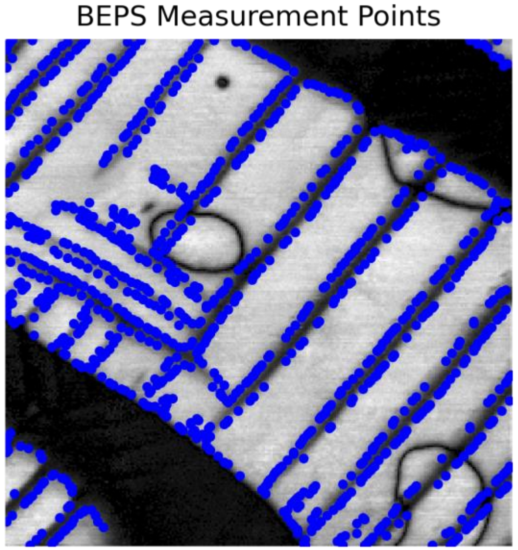
In [ ]:

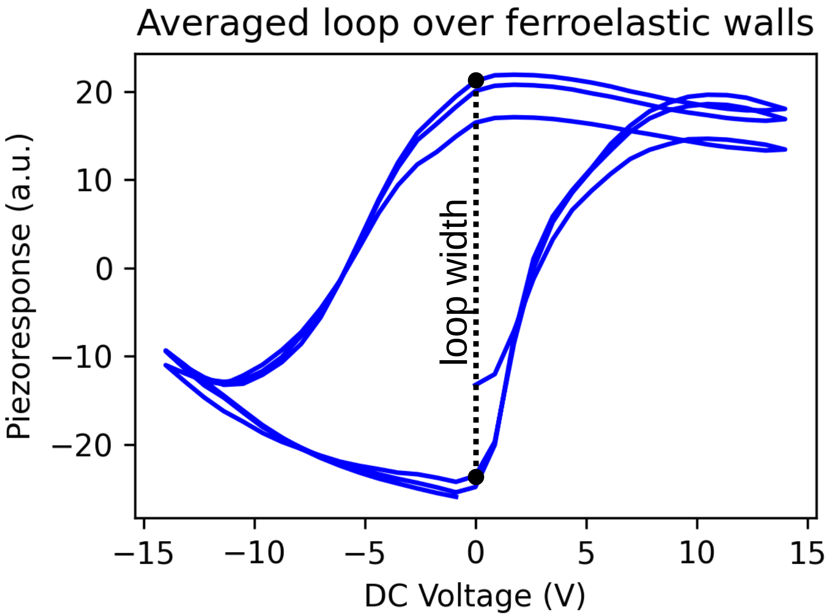
In [ ]:

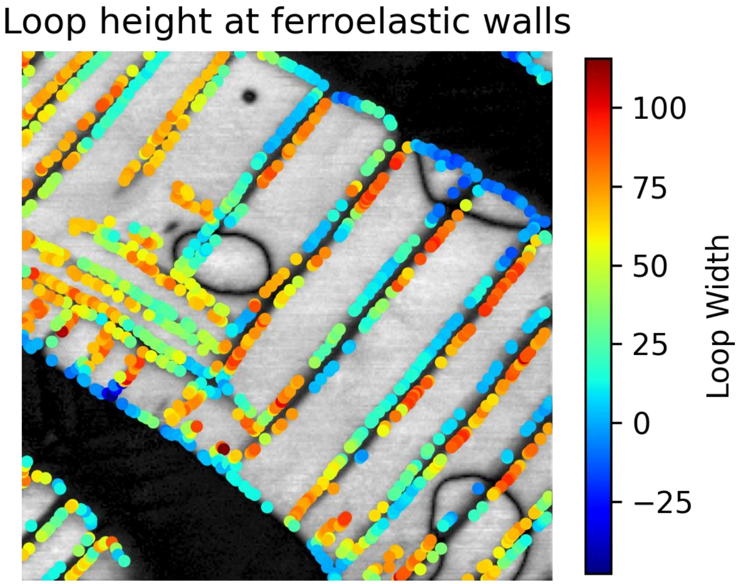
Automated SPM #1

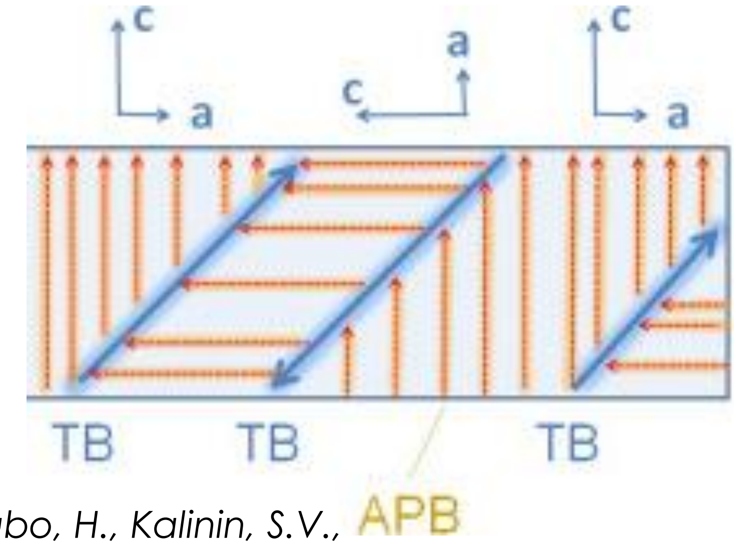






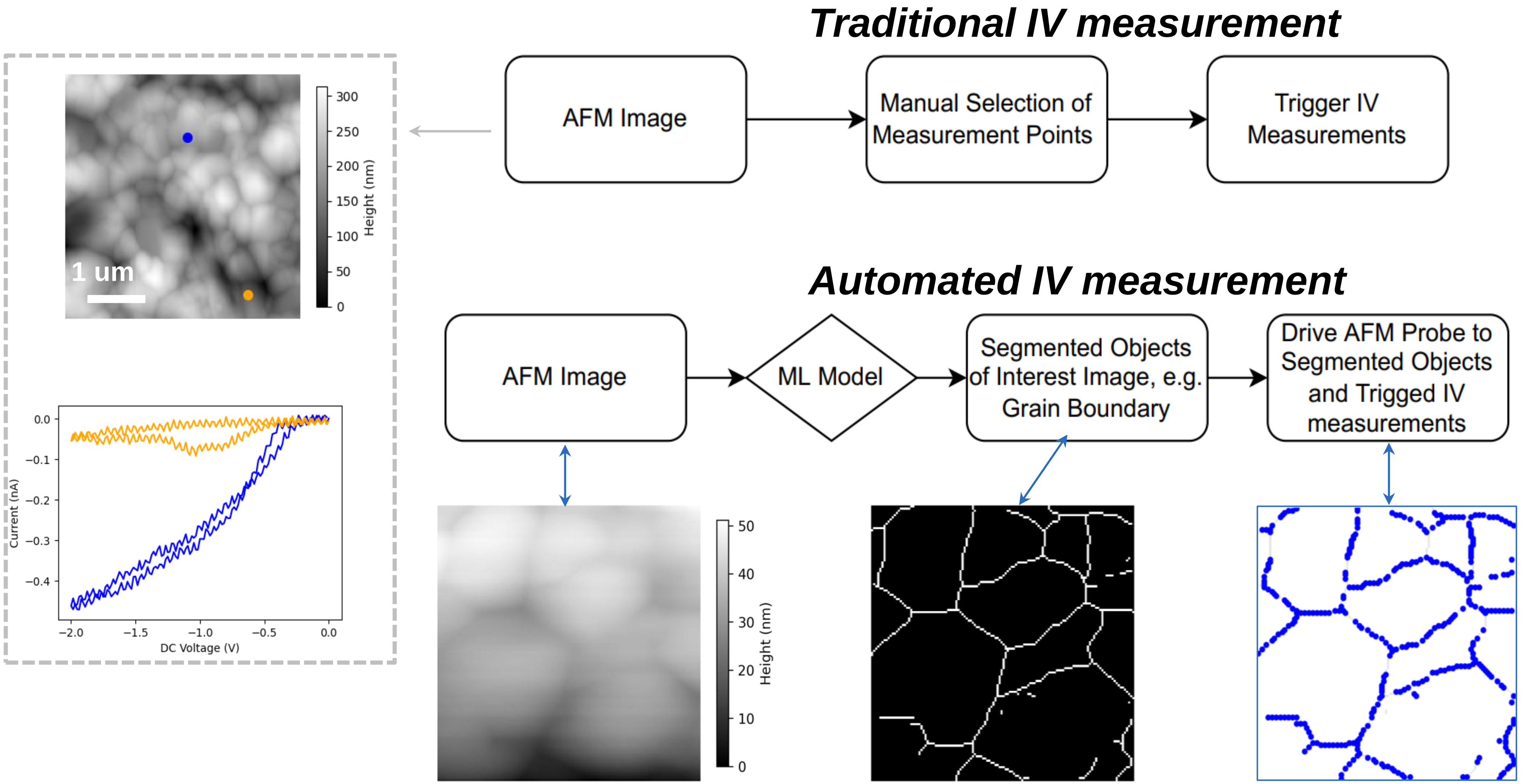




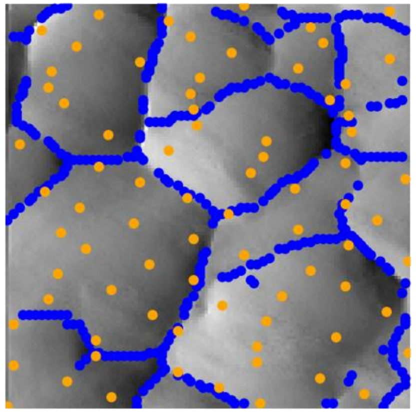


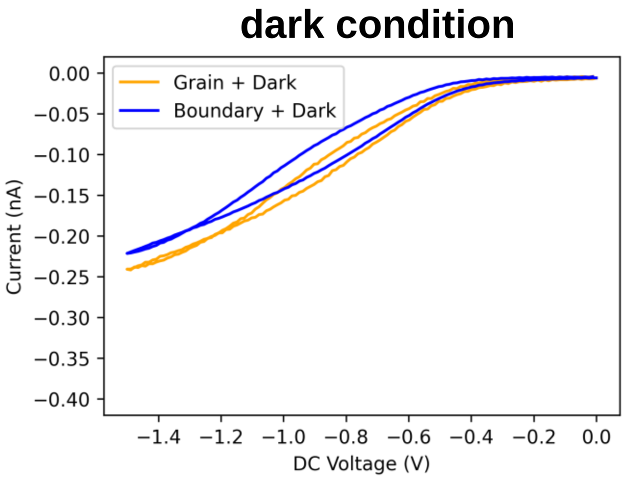
Liu, Y., Kelley, K.P., Funakubo, H., Kalinin, S.V., Ziatdinov, M., Arxiv, submitted

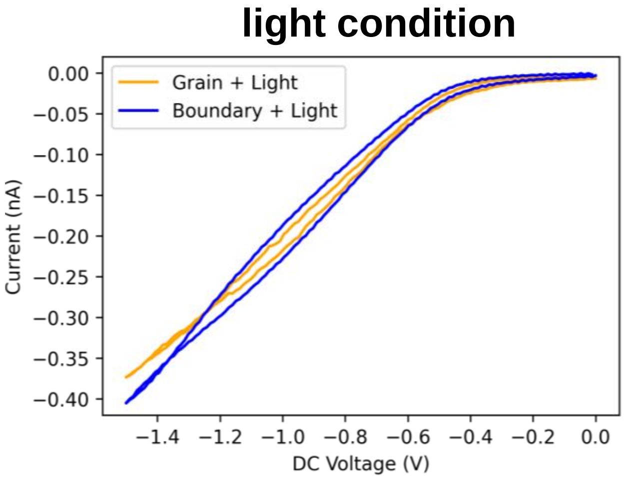
Automated Experiment in cAFM

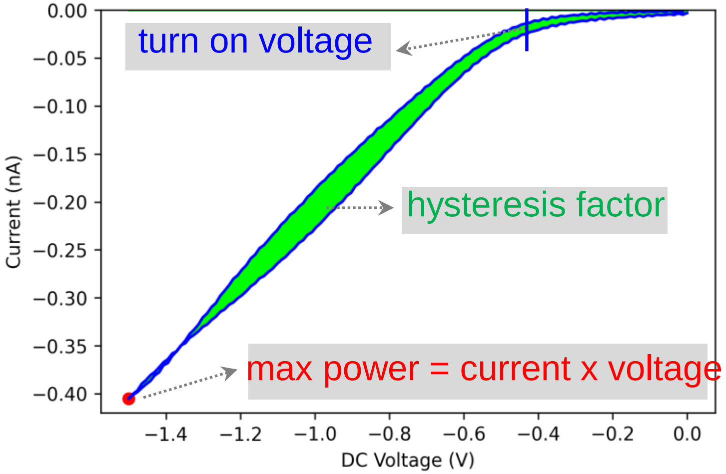


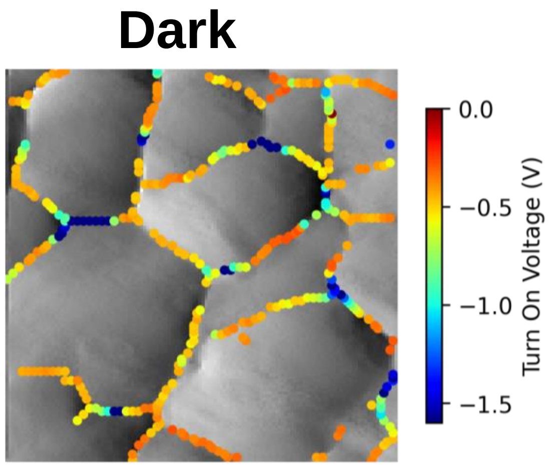
Automated Experiment in IV

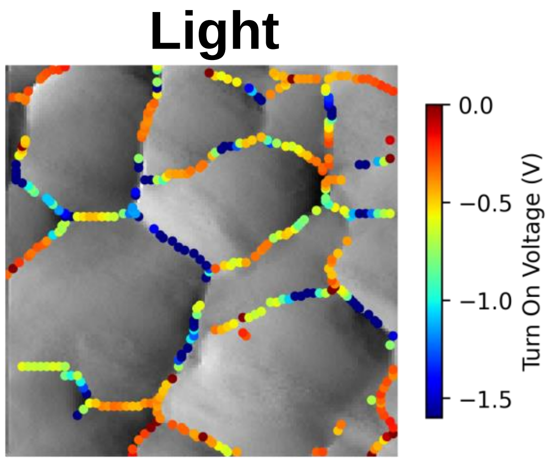






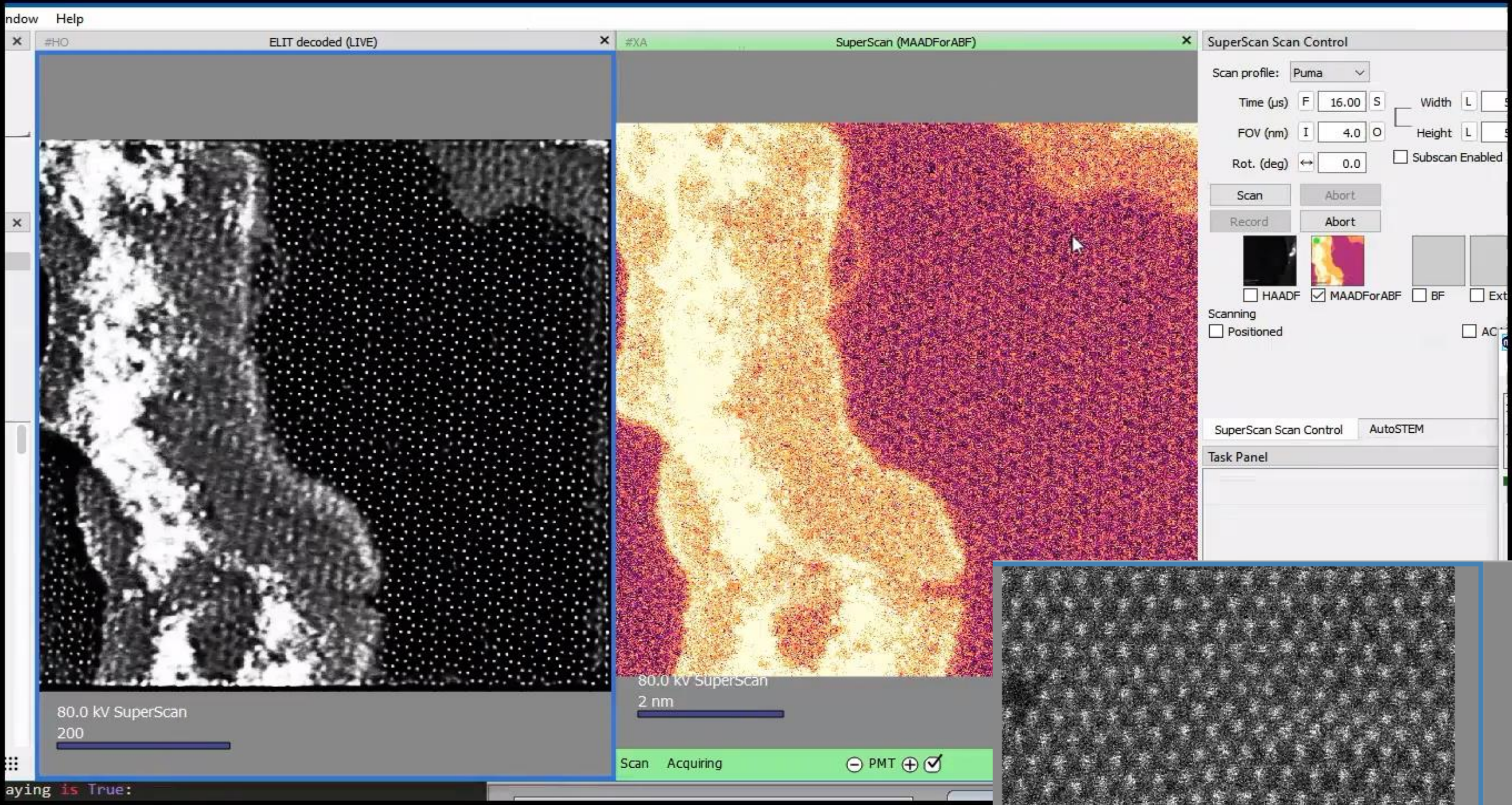


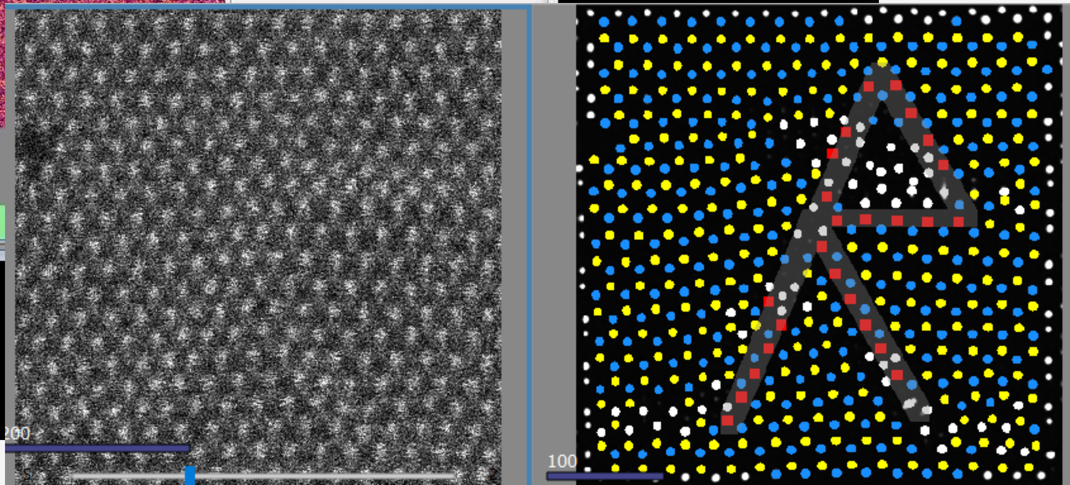




Decision making during the experiment: real time analysis

Implementation: Kevin Roccapriore, Ayana Ghosh, Sergei V. Kalinin & Maxim Ziatdinov





Automated control

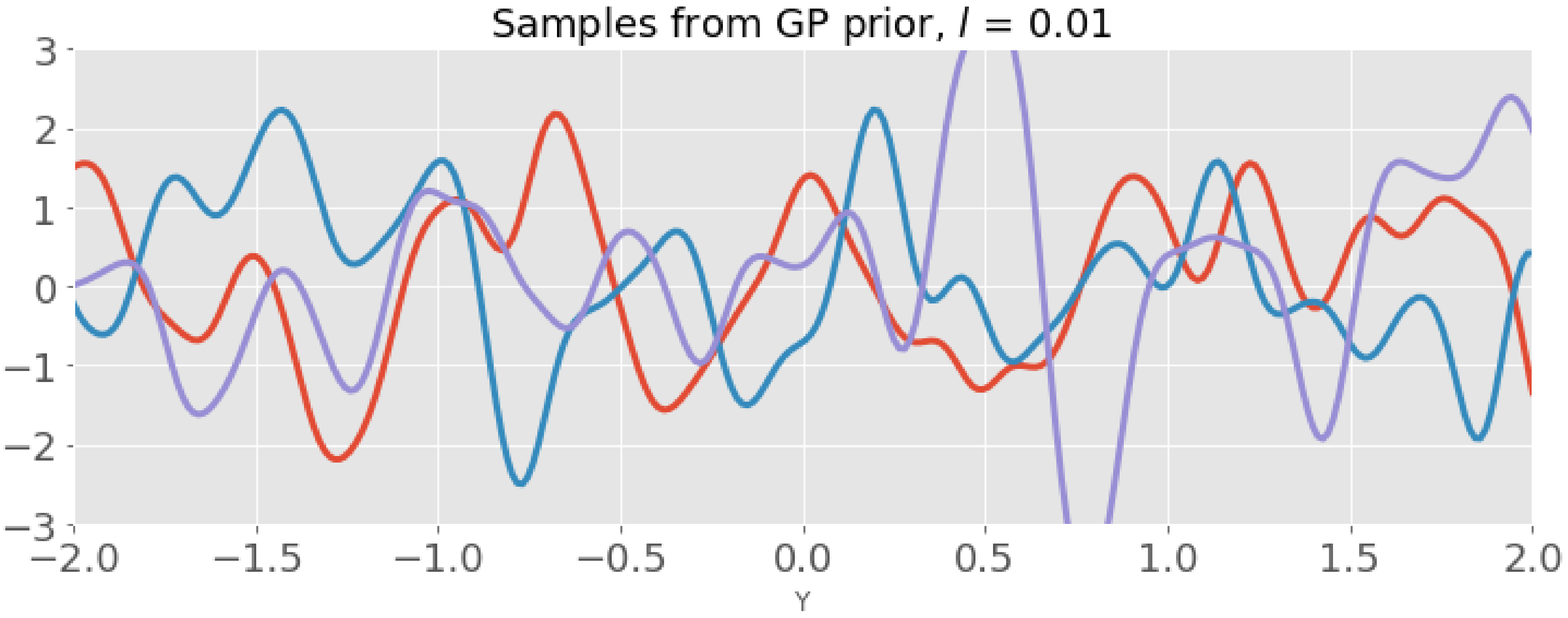
Automated Experiment:

… with John Snow priors…

Gaussian Process Regression

- Covariance matrix determines what type of functions we will allow.

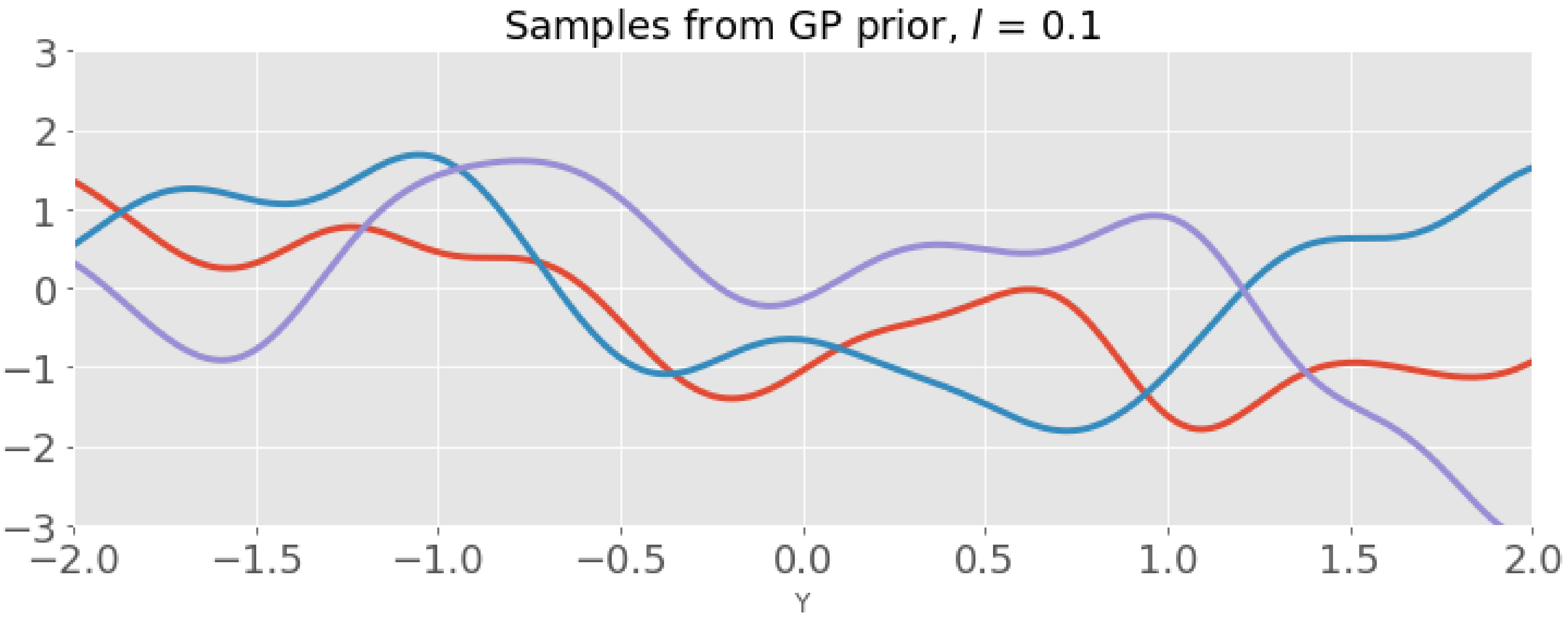
$$k(x, x') = \exp\left(-\frac{1}{2l}(x - x')^2\right)$$



Gaussian Process Regression

- Covariance matrix determines what type of functions we will allow.

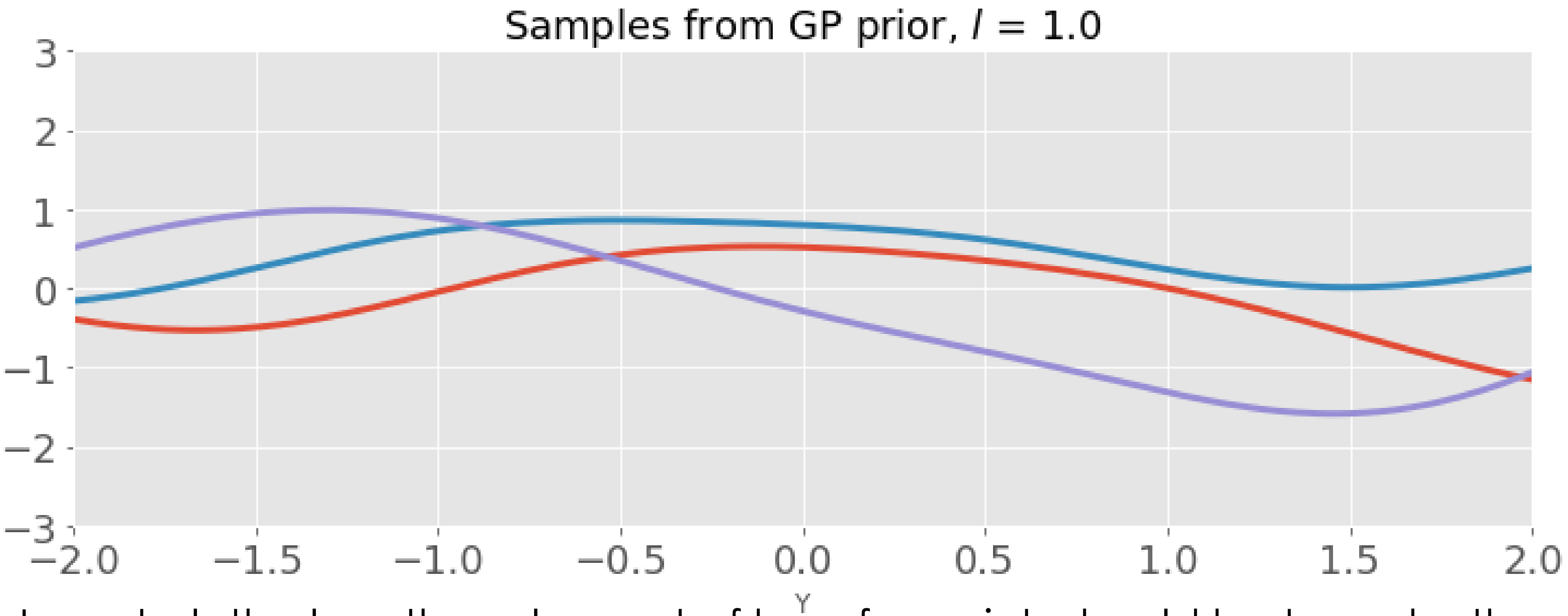
$$\mathrm{k}(\mathrm{x},\mathrm{x}') = \exp\left(-\frac{1}{2l}(x-x')^2\right)$$



Gaussian Process Regression

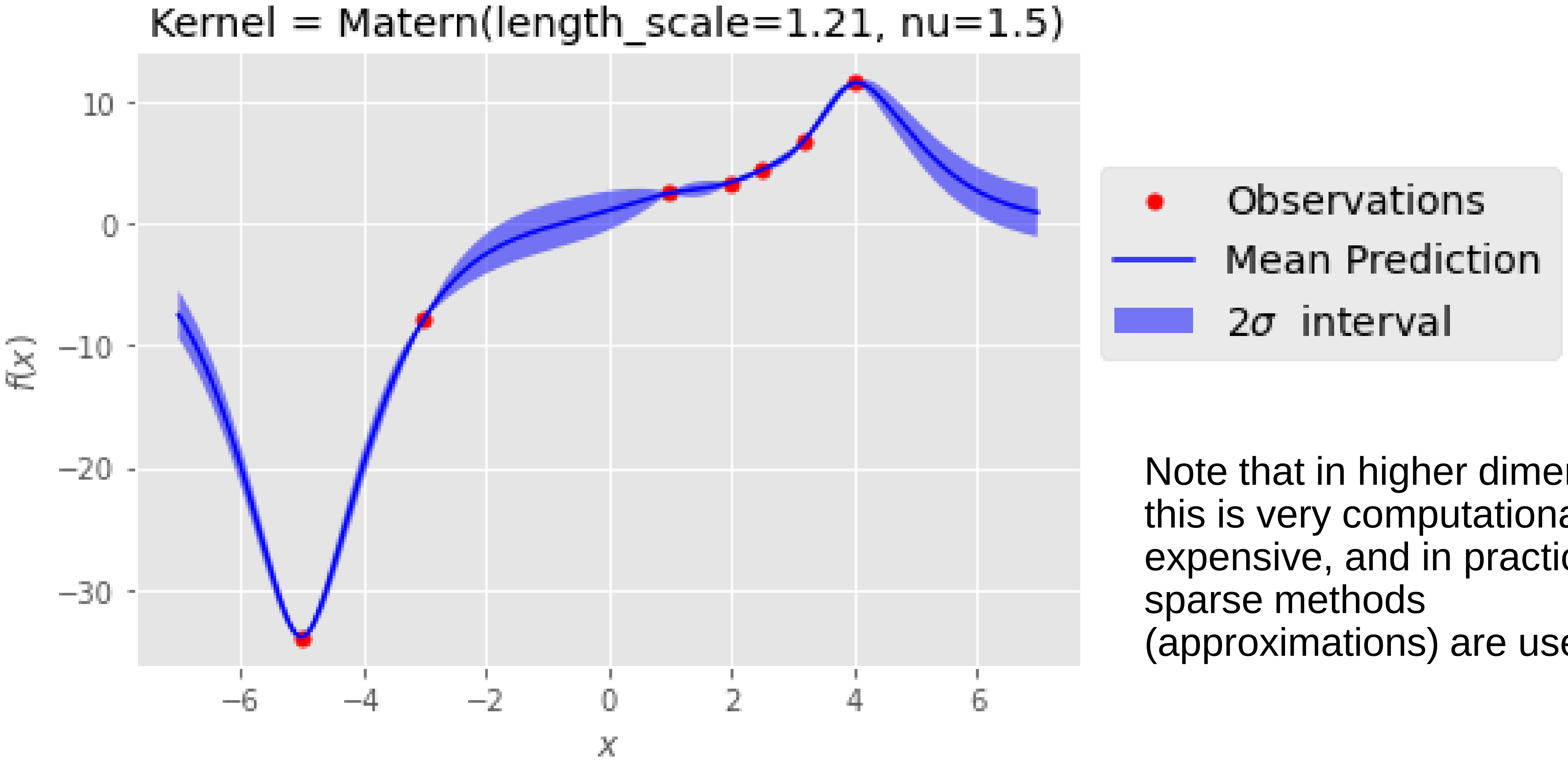
- Covariance matrix (kernel) determines what type of functions we will allow.

$$k(x, x') = \exp\left(-\frac{1}{2l}(x - x')^2\right)$$



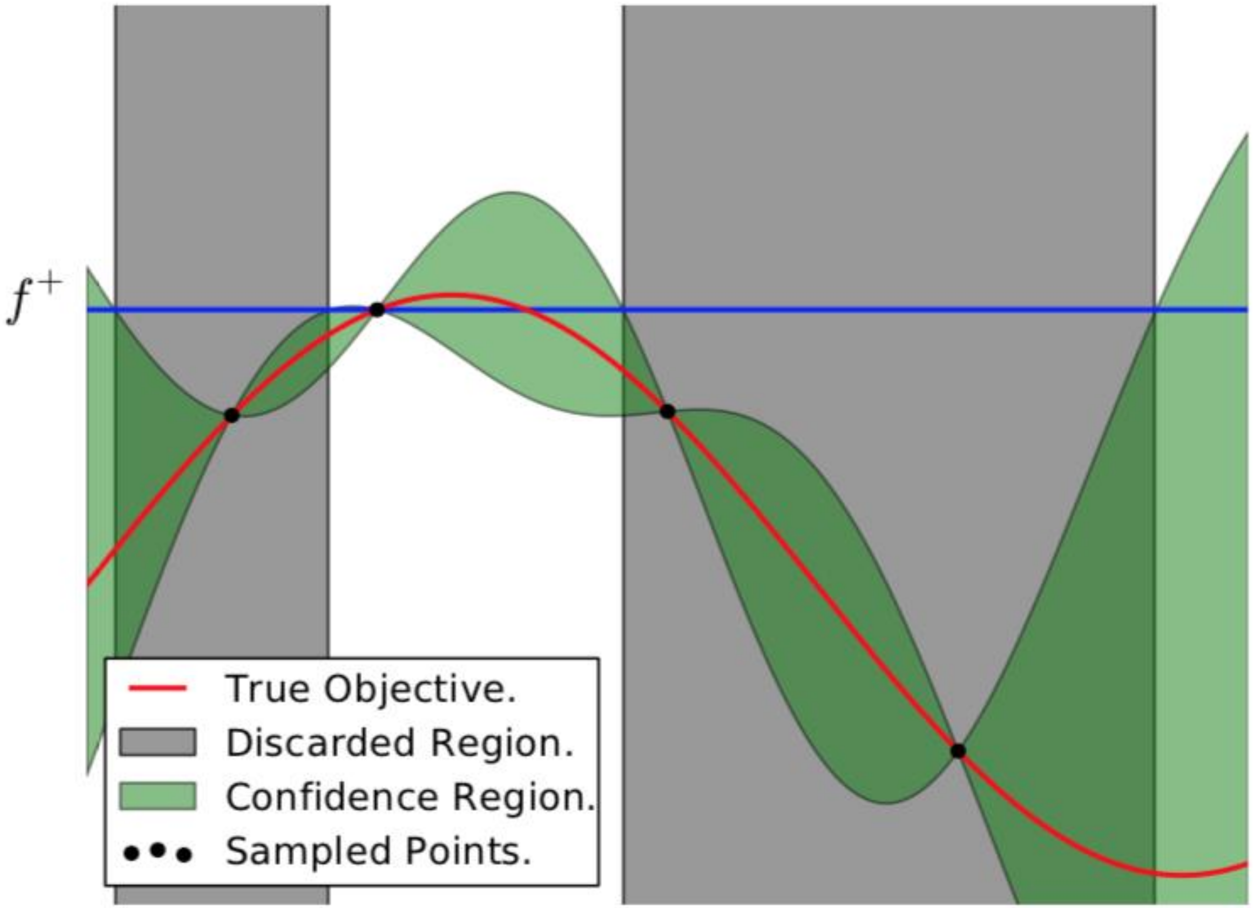
L controls the length scale – sort of how far points should be to make them independent of each other.

Gaussian Process Regression



Note that in higher dimensions this is very computationally expensive, and in practice sparse methods (approximations) are used.

Bayesian Optimization

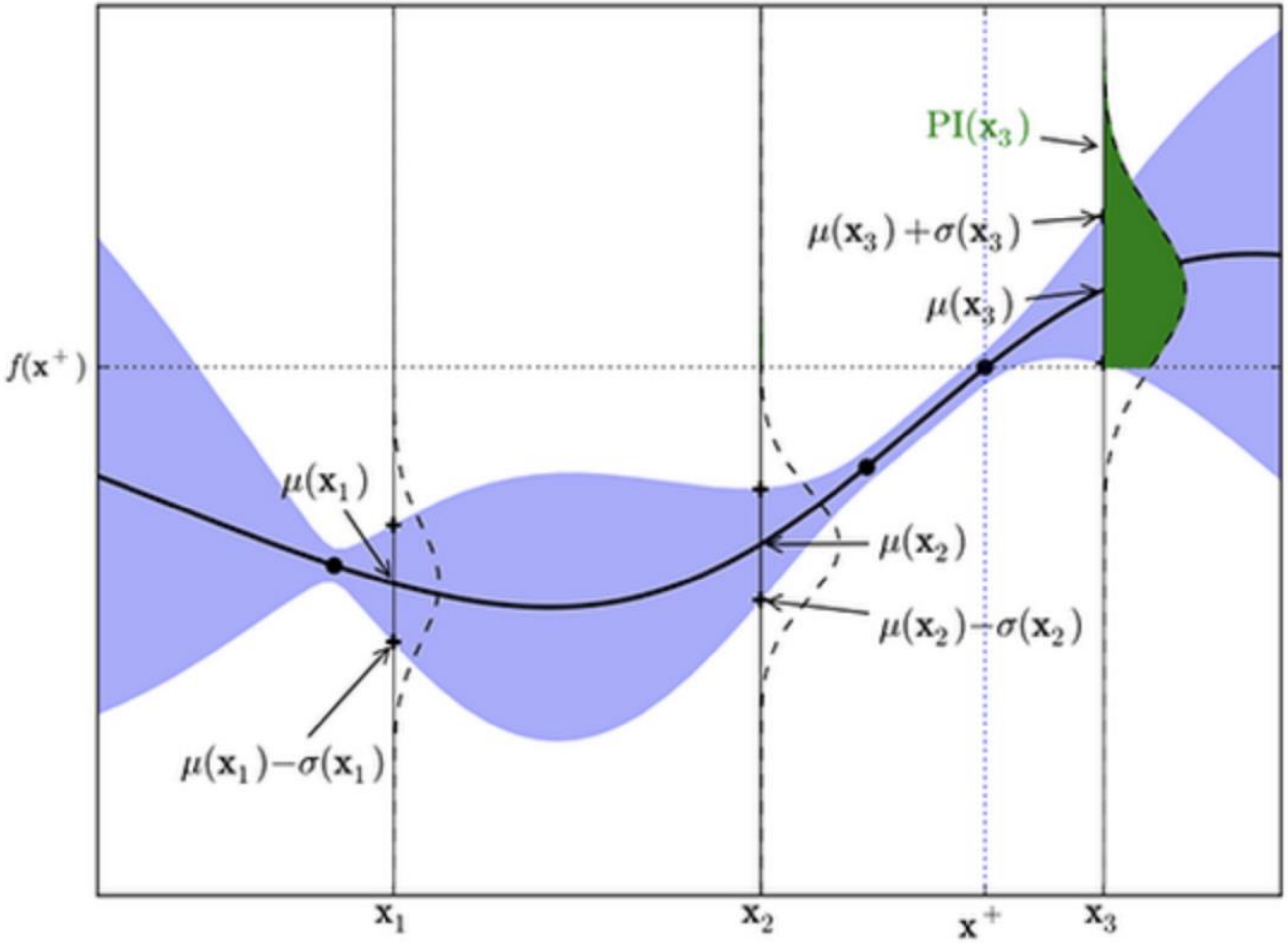


N. de Freitas et al., Taking the Human Out of the Loop: A Review of Bayesian Optimization , *Proceedings of the IEEE* **104**, 148 (2015)

- We have some measurements in space X, and we want to maximize some property f(X).
- How can we decide what point to measure next to best maximize f?
- We need to balance the exploration of the space with exploitation of regions near we have already know

Acquisition Functions

Probability of Improvement
Acquisition Function



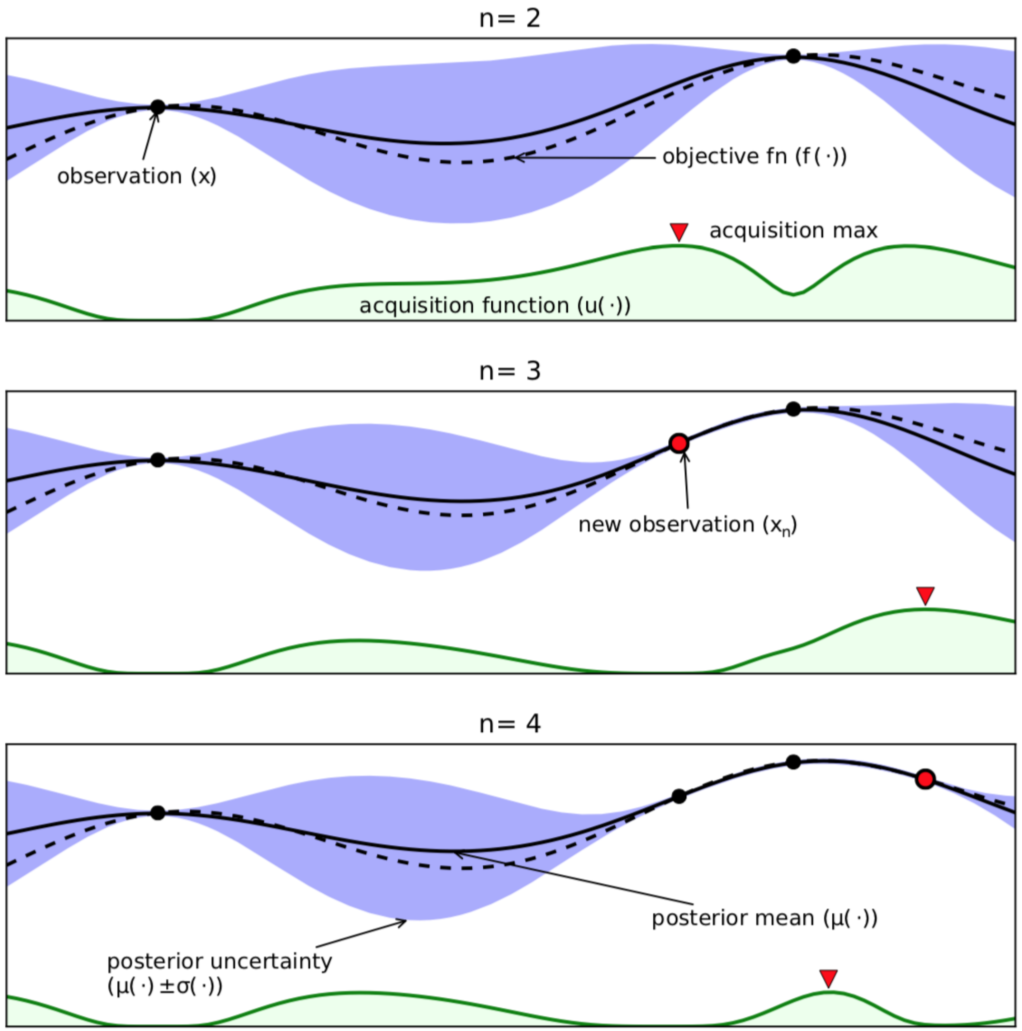
1. **Confidence bound:** simplest possible - just take the upper confidence bound from the prediction

2. **Probability of Improvement:** Integral from current functional maximum to upper limit of distribution as test point

3. **Expected Improvement:** Instead of probability of improvement, we want to maximize the expected increase in the function value

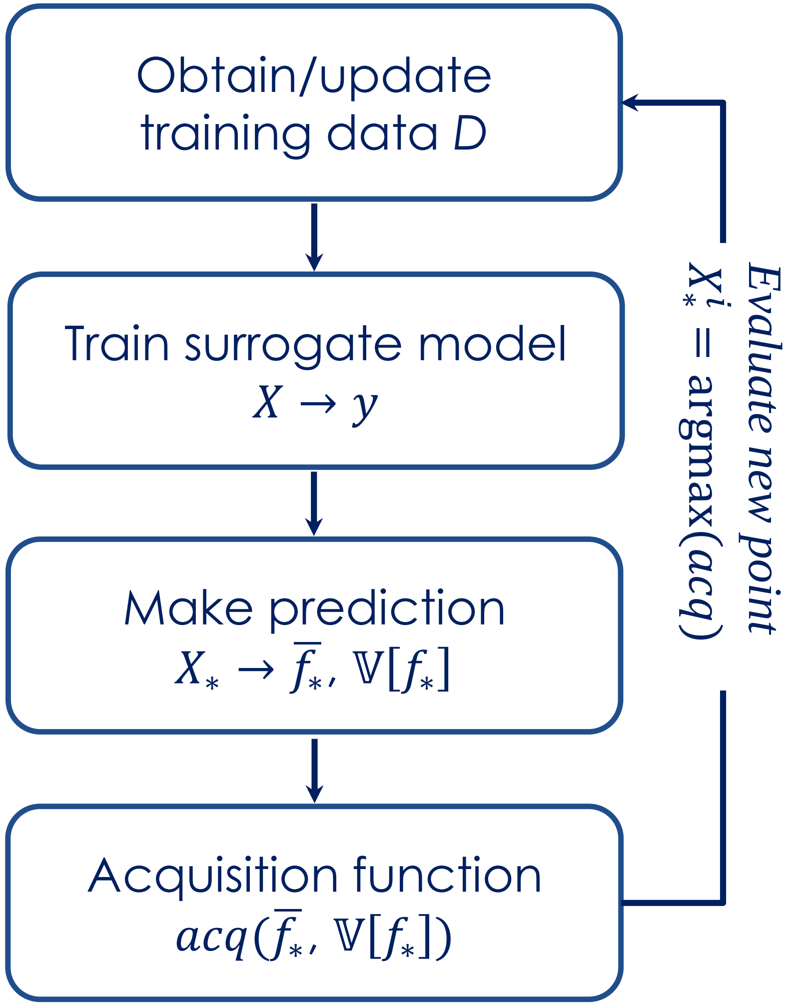
4. **There are (always) more…**

The basics: Bayesian Optimization



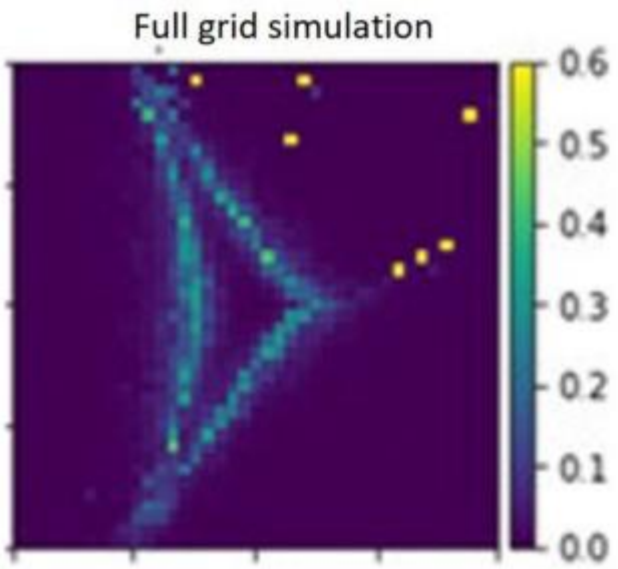
X, y: (sparse) Training data

X_*: New (not yet evaluated) points



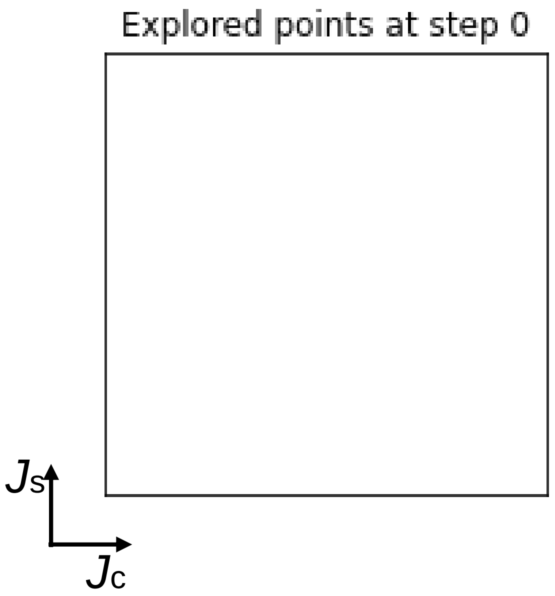
N. de Freitas et al., Taking the Human Out of the Loop: A Review of Bayesian Optimization , *Proceedings of the IEEE* **104**, 148 (2015)

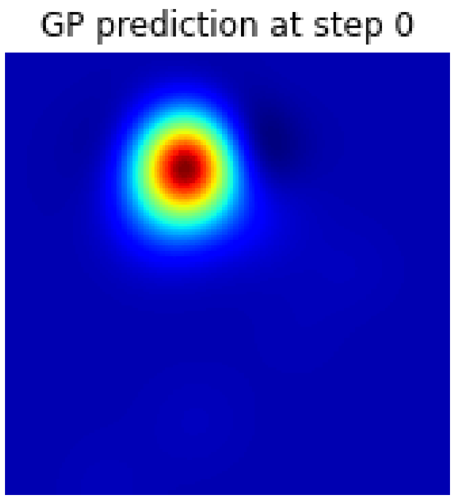
Bayesian Optimization for physical discovery



- Started to work in August 2019
- First, we planned to apply BO for autonomous experimentation at CNMS. Then, COVID happened…
- “So what if we use BO to explore the parameter spaces of theoretical models?”

Discovering regions where heat capacity is maximized in NNN Ising model





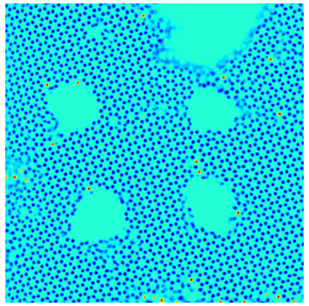
Implementation in GPim

```
def acquisition_function(gpmodel, X_full, X_sparse):
    mean, sd = gpmodel.predict(X_full, verbose=0) # leave it as is
    acq = 5 * sd + 10 * mean # Good for heat capacity
    return acq, (mean, sd) # leave it as is
```

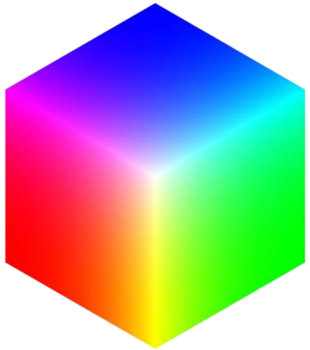
```
# Initialize Bayesian optimizer with a custom acquisition function
boptim = gpim.boptimizer(
    X_sparse, Z_sparse, X_full,
    J2_to_S_func, acquisition_function=acquisition_function,
    batch_size=1000, lengthscale=[1., 40.], dscale=4,  exit_strategy=1,
    exploration_steps=650, use_gpu=True, verbose=1, save_checkpoints=True,
    filename='/content/drive/My Drive/research/Ising_BO/bo_ising_heat_acq1_e')
# Run Bayesian optimization
boptim.run()
```

Automated experiment workflows

SPM or STEM image



EELS or SPM datacube



- **Sliding window/linear transform**
- **Keras DCNN**
- rVAE (rotational invariance)
- rcVAE (plus classification)

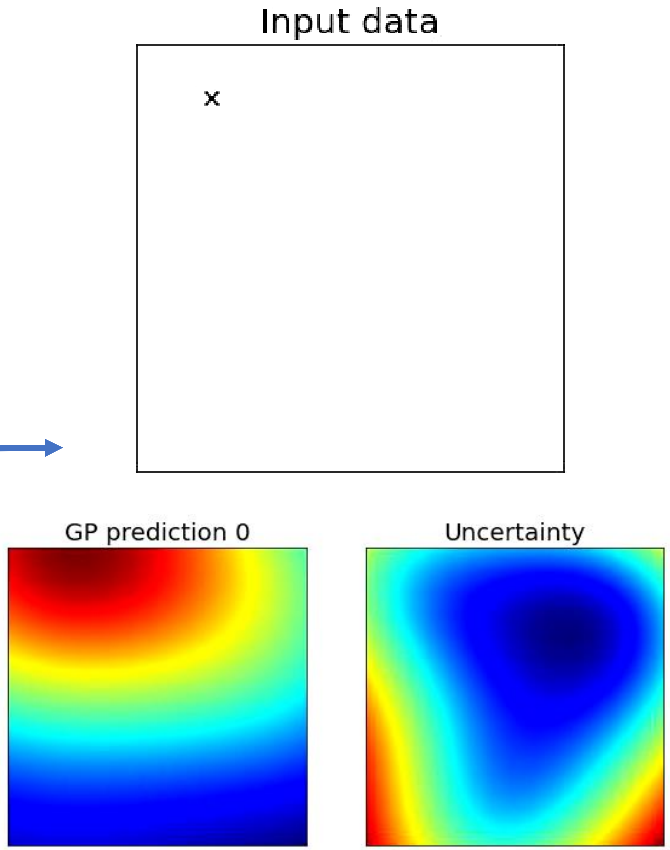
Descriptor

Gaussian processing

- **Integrated intensity**
- **Keras DCNN**
- Spec2im autoencoder
- (im,spec)2(spec,im)
- CycleGAN

GPim library (M. Ziatdinov)

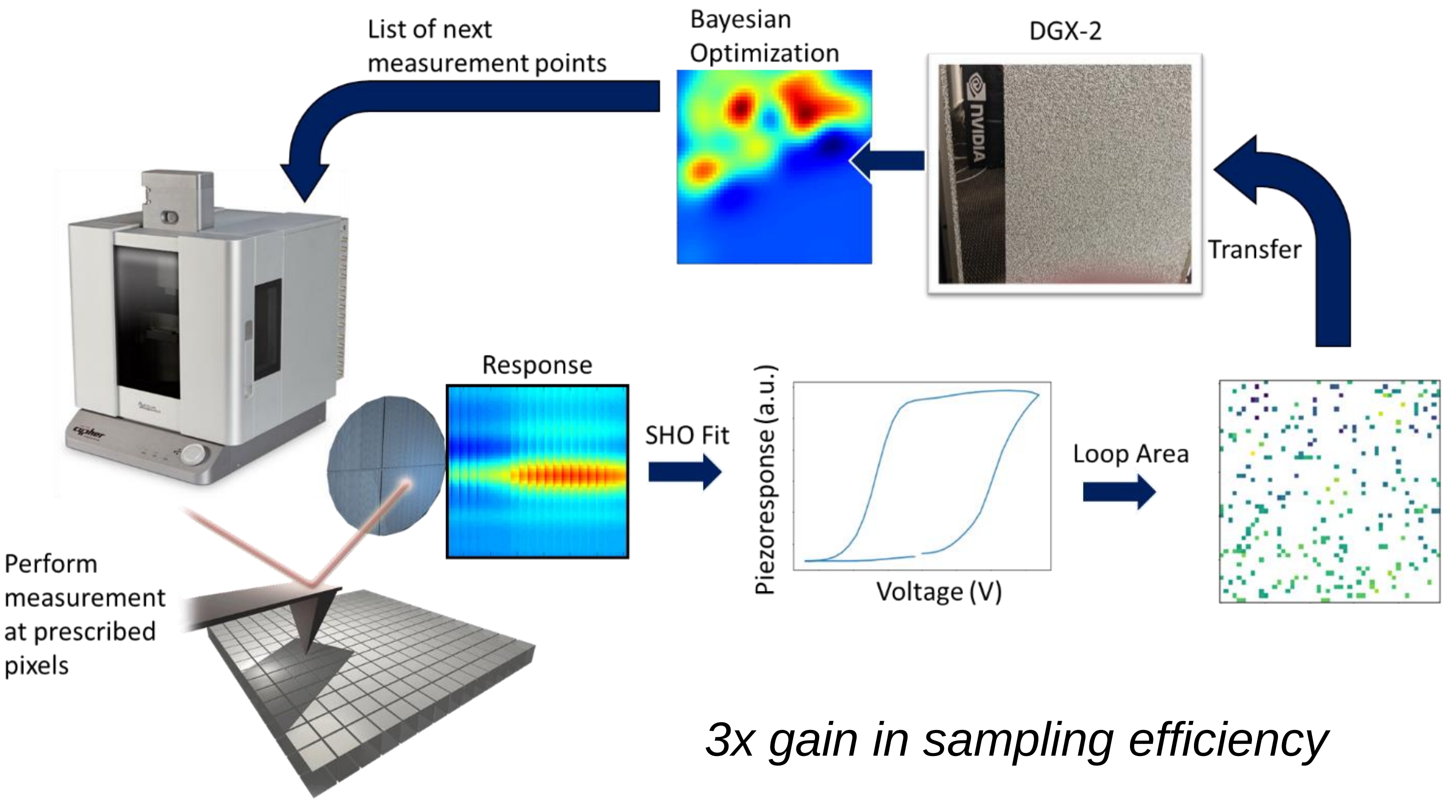
- **Acquisition functions**
- **Pathfinder functions**
- **Kernel control**



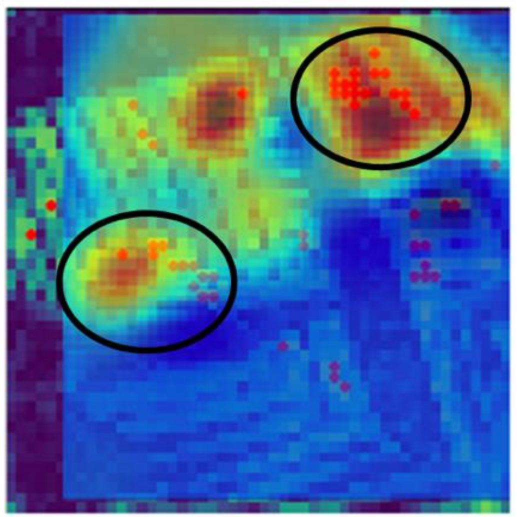
- **AE based on structural analysis for STEM data**
- **AE based on spectral data in PFM**
- **AE based on DL for EELS data**
- Feature of interest finding for mesoscopic images

Bayesian Optimization for Self-Driving Microscopy

Then, COVID restrictions got relaxed and Rama Vasudevan realized a “self-driving” PFM



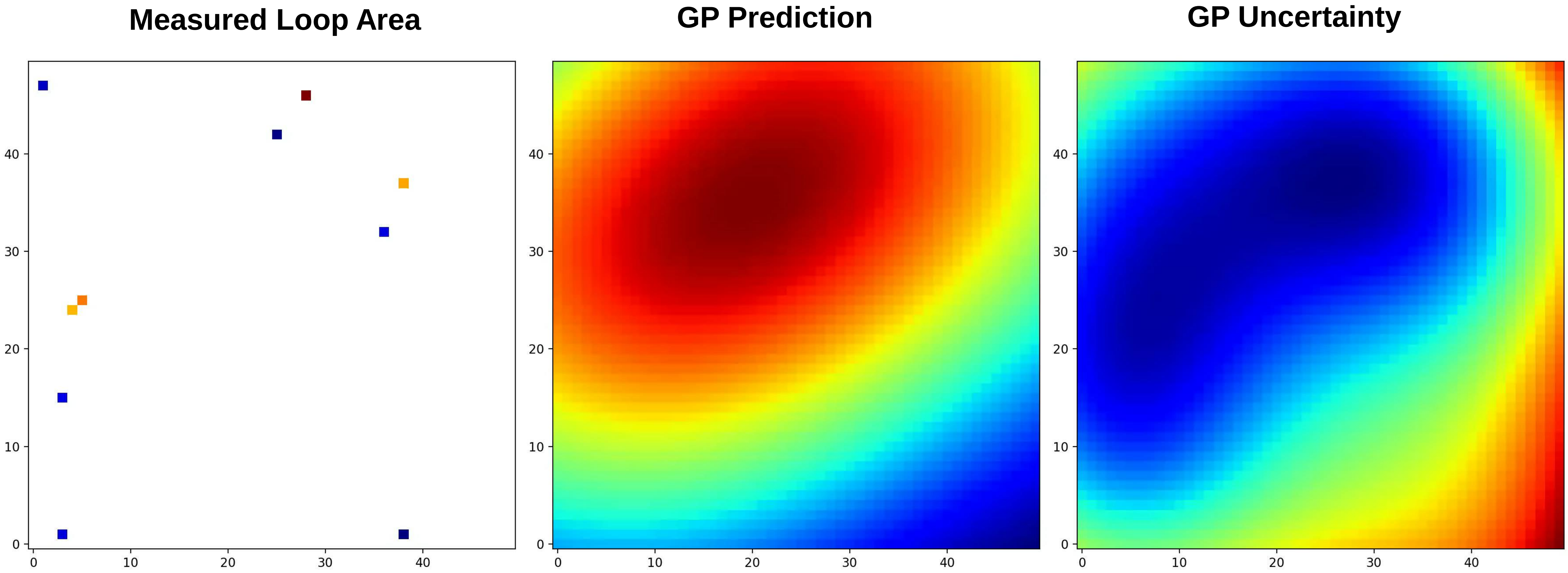
Comparison with “ground truth”



3x gain in sampling efficiency

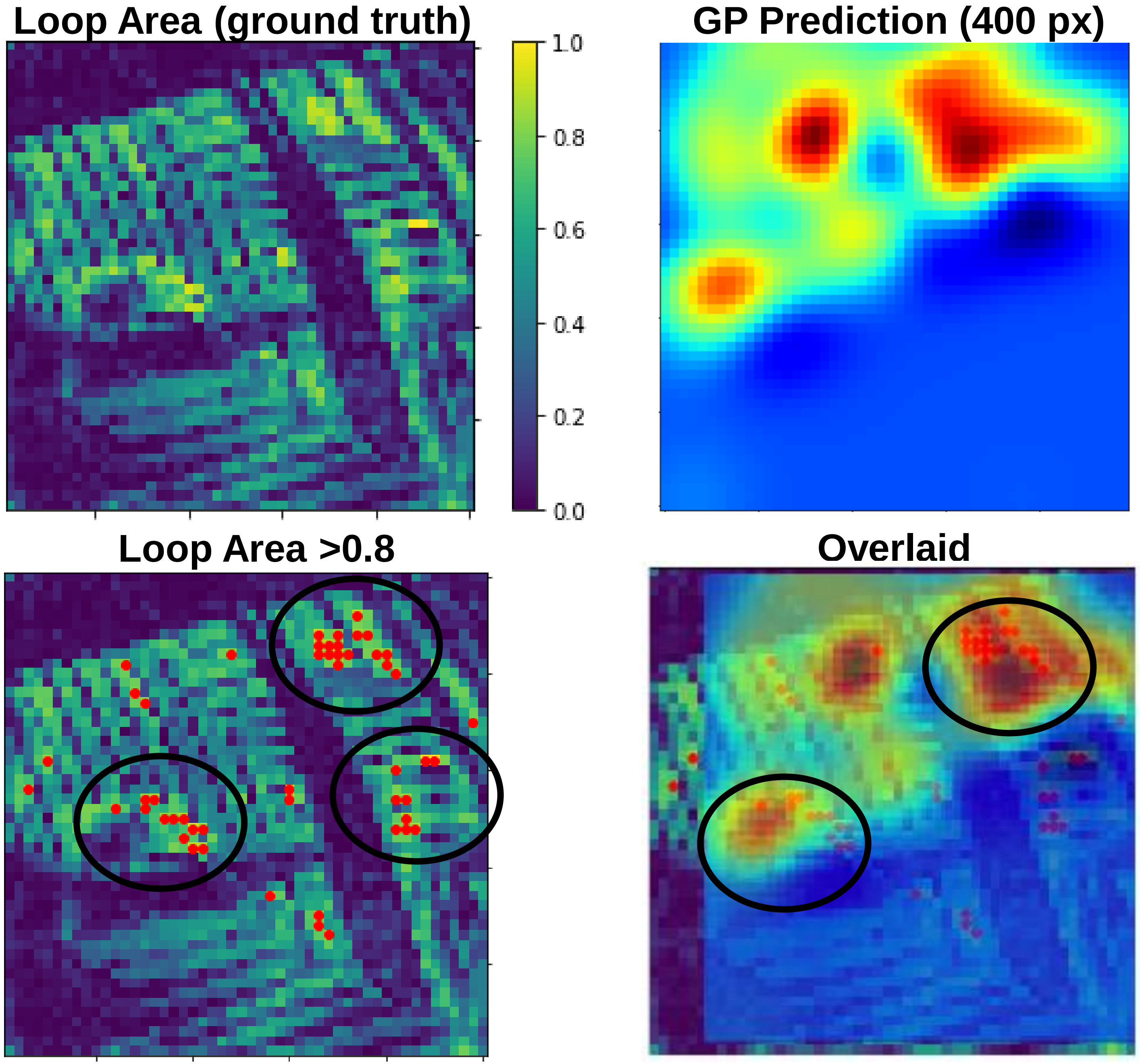
R. K. Vasudevan, K. Kelley, H. Funakubo, S. Jesse, S. V. Kalinin, M. Ziatdinov, ***ACS Nano*** *(2021) https://doi.org/10.1021/acsnano.0c10239*

Putting it together: GP optimized experiments



But what if we do not know a priori what elements of domain structure are we interested in?

- First step – Gaussian Processing towards exploration of specific behavior
- Here, we explore regions with maximal area under hysteresis loop
- For *N* measured points, the GP reconstructs the loop area map and uncertainties of the reconstruction
- Based on these, next locations for measurements are selected



arXiv:2103.12165

arXiv:2011.13050

Next steps:

- Incorporate prior knowledge of domain structure
- Factor in generative physics of ferroelectric domain structures
- These are complex ML problems

But: the bridge is built!

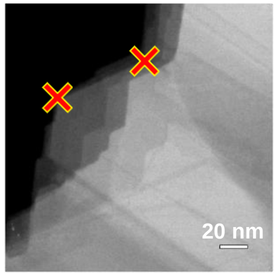
Automated Experiment:

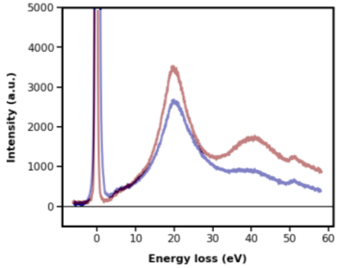
… as a scientist…

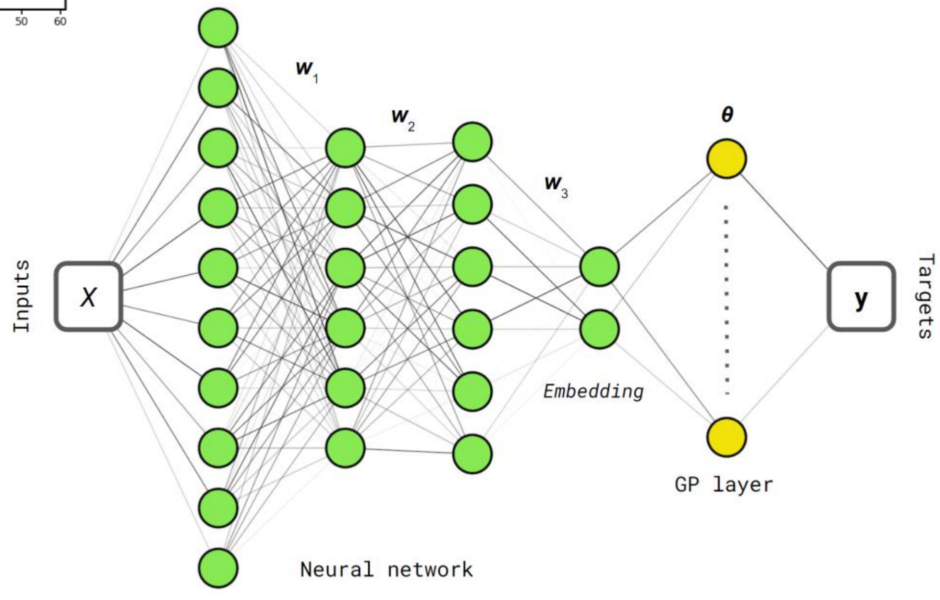
Physics-based feature engineering:
Deep kernel learning – Bayesian optimization

Specify physics criteria

Active learning







Acquire **structural data**

↓

Measure a **spectrum**

↓

Train DKL model with new data

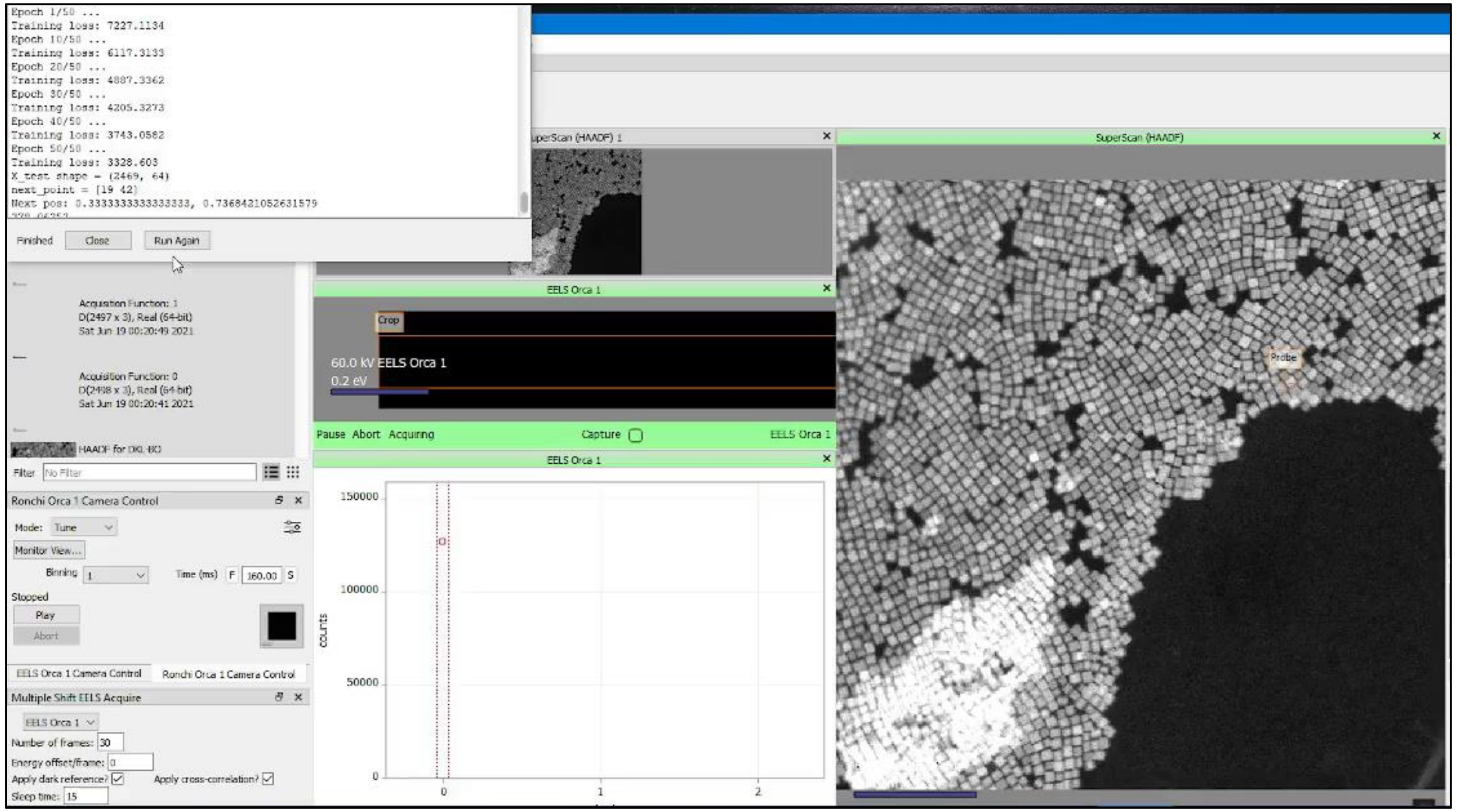
↓

Decide next position (optimize physics criteria)

Allows navigation of the system to search for physics

Physics-based feature engineering:
Microscope Operation

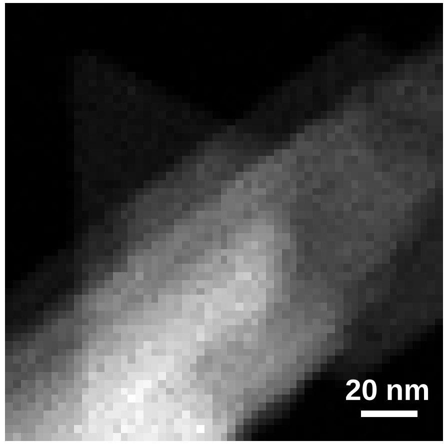
Structural image + Set of spectra



Opportunity alert

- Between measurements, the beam is optionally blanked or placed in a safe position.
- This is an **excellent** opportunity for "smart EELS" with beam sensitive materials

Discovering region with interesting physics

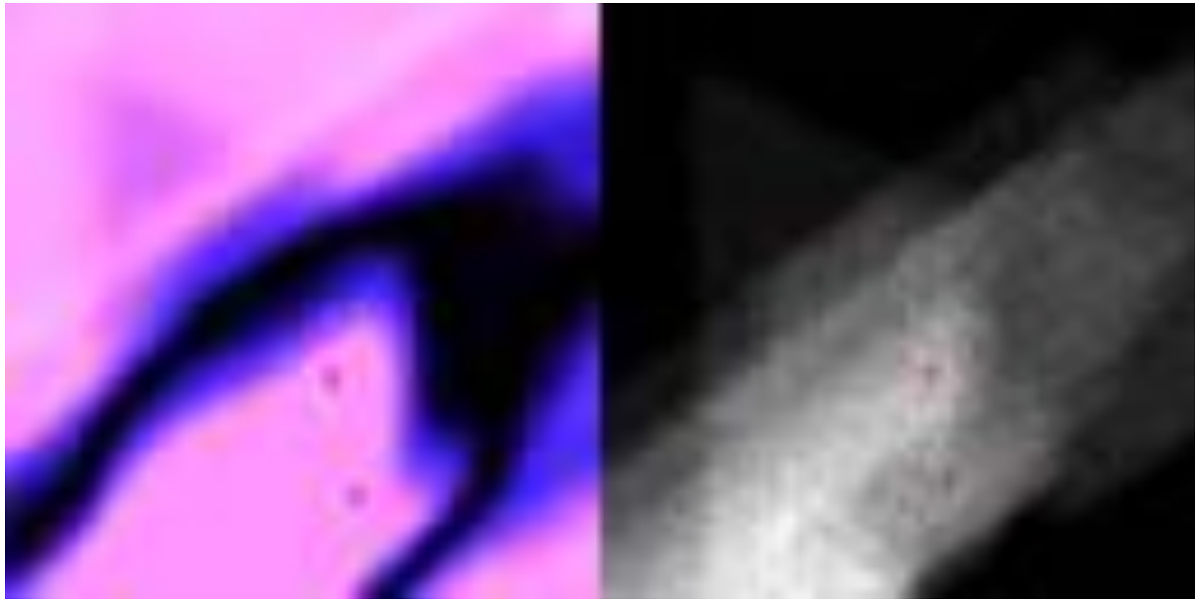


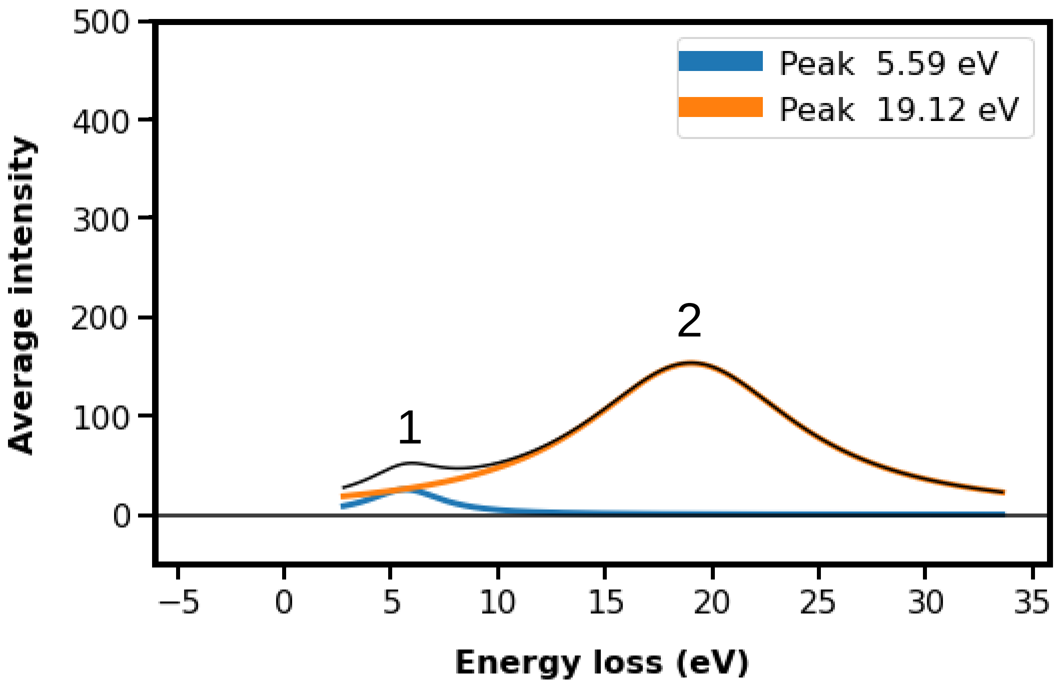
- **Discovering physics in a "new" material $MnPS_3$**
- **Curve fitting** to help enforce physical processes

Physics search criteria:

$$Ratio = Peak\,1 \,/\, peak\,2$$

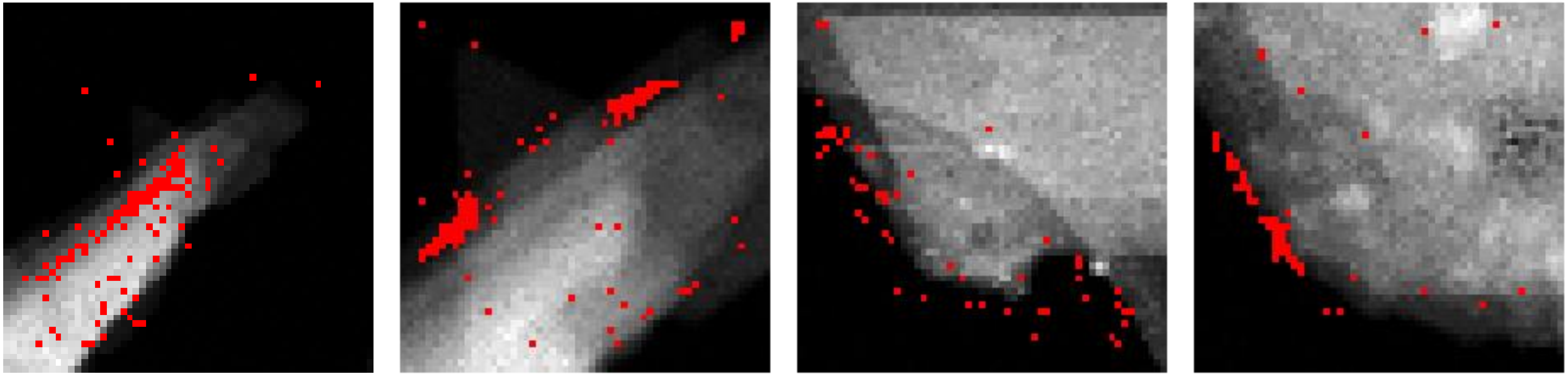
"Acquisition function" HAADF-STEM





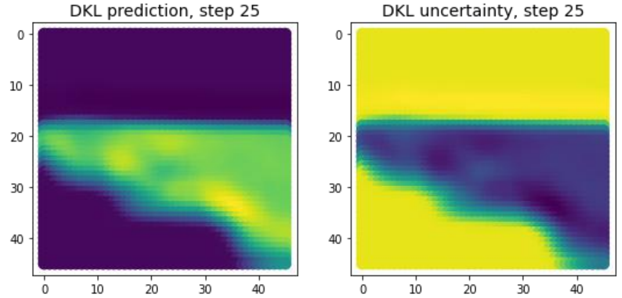
More examples of physics discovery

- Very similar behavior when searching for the same criteria!
- Success!



Discovery pathway depends on the reward structure (scalarizer that defines signature of physics we want to discover)!

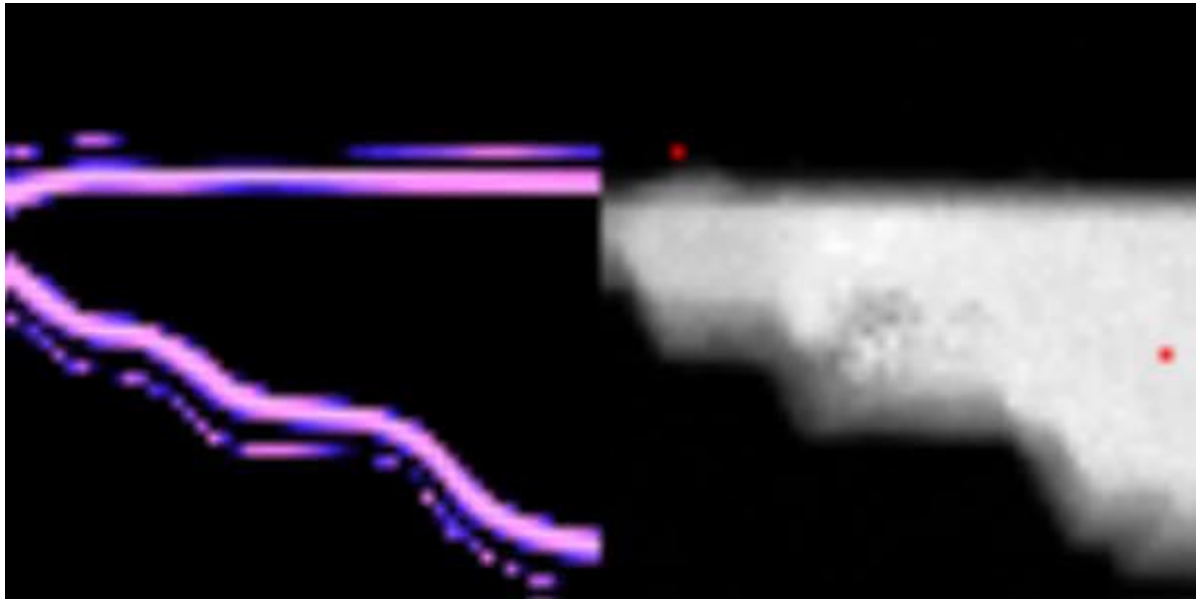
Changing the criterion



- (Same region) **Simple physics search:** peak max in selected region

“Acquisition function”

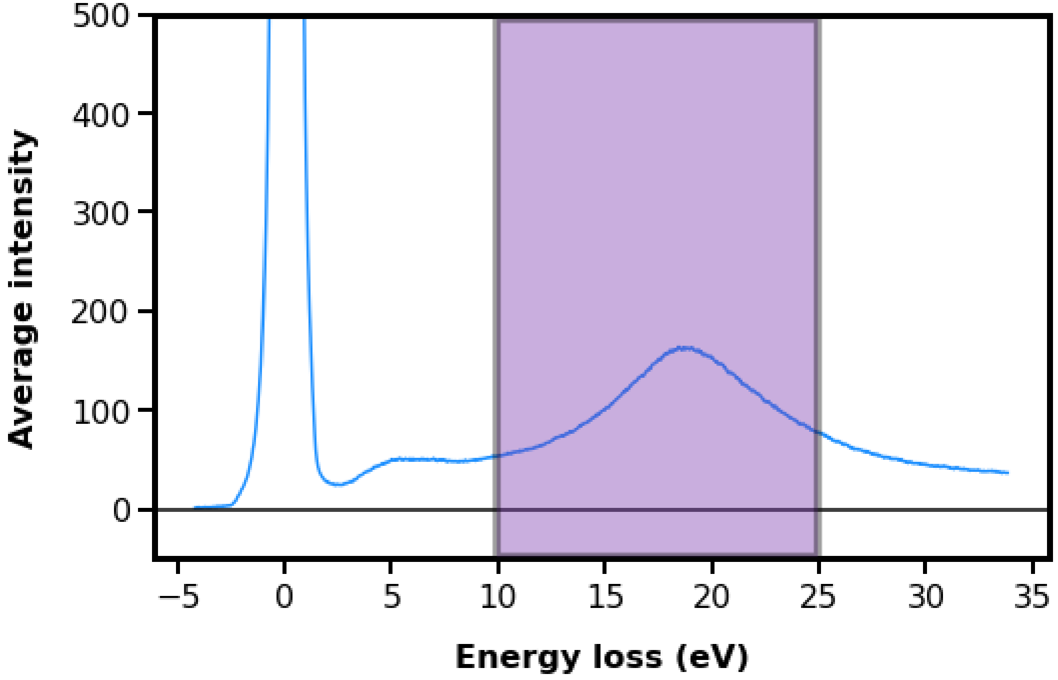
HAADF-STEM
+ points visited



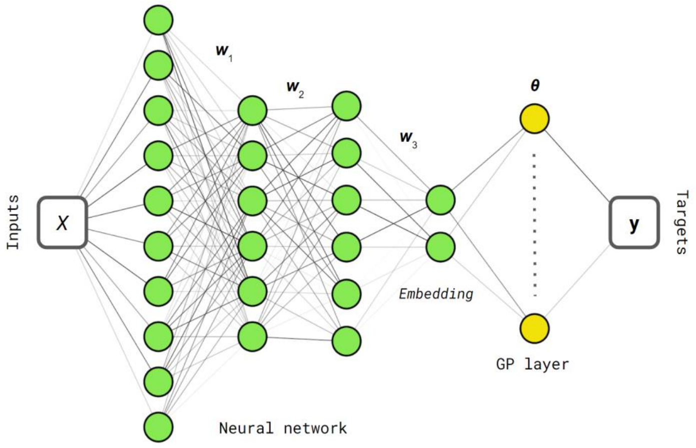
Physics search criteria:

$$\boldsymbol{Maximize(f)}$$

(Specific peak intensity)



Automated Experiments in 4D STEM

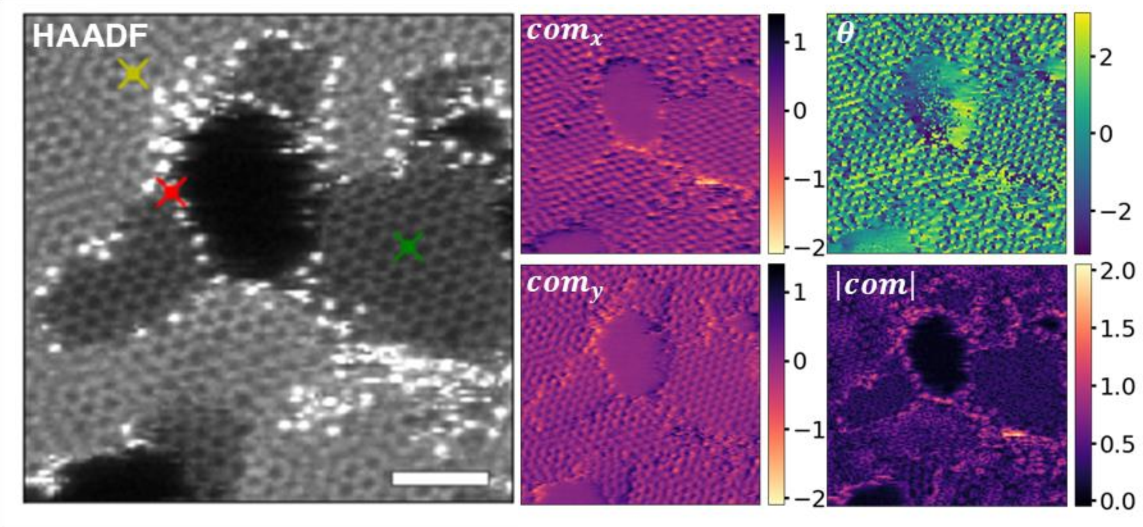


Quantities to explore

- ➢ Electric field
- ➢ Potential
- ➢ Charge density
- ➢ Strain

Choose **explorable quantity**

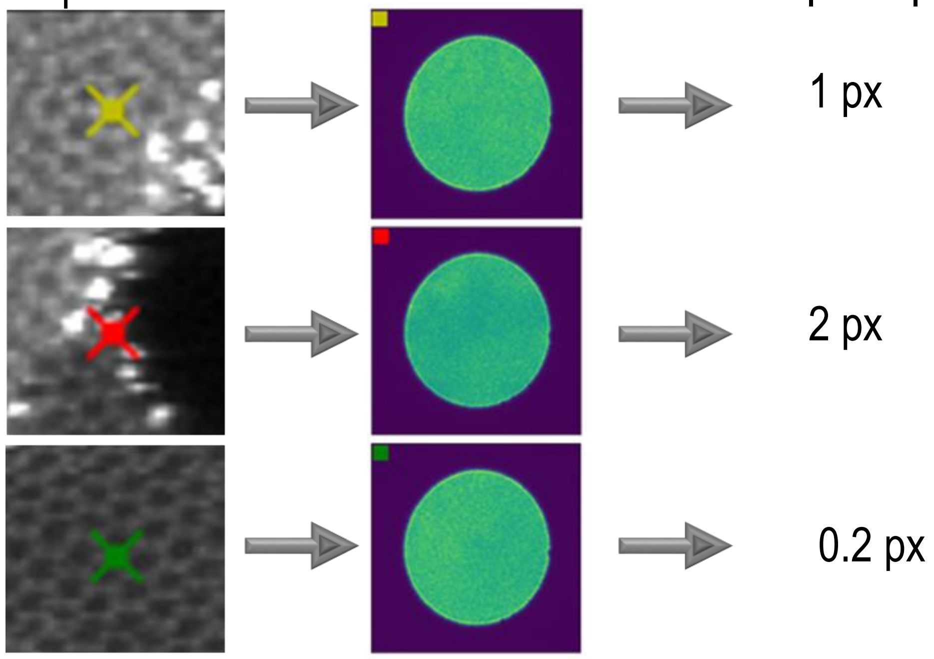
Traditional 4D STEM: graphene "DPC"



Local image patch

D.P.

Scalarize $|com|$



Automated Experiments in 4D STEM DPC example

Choice of scalarizer (pre-acquired data)

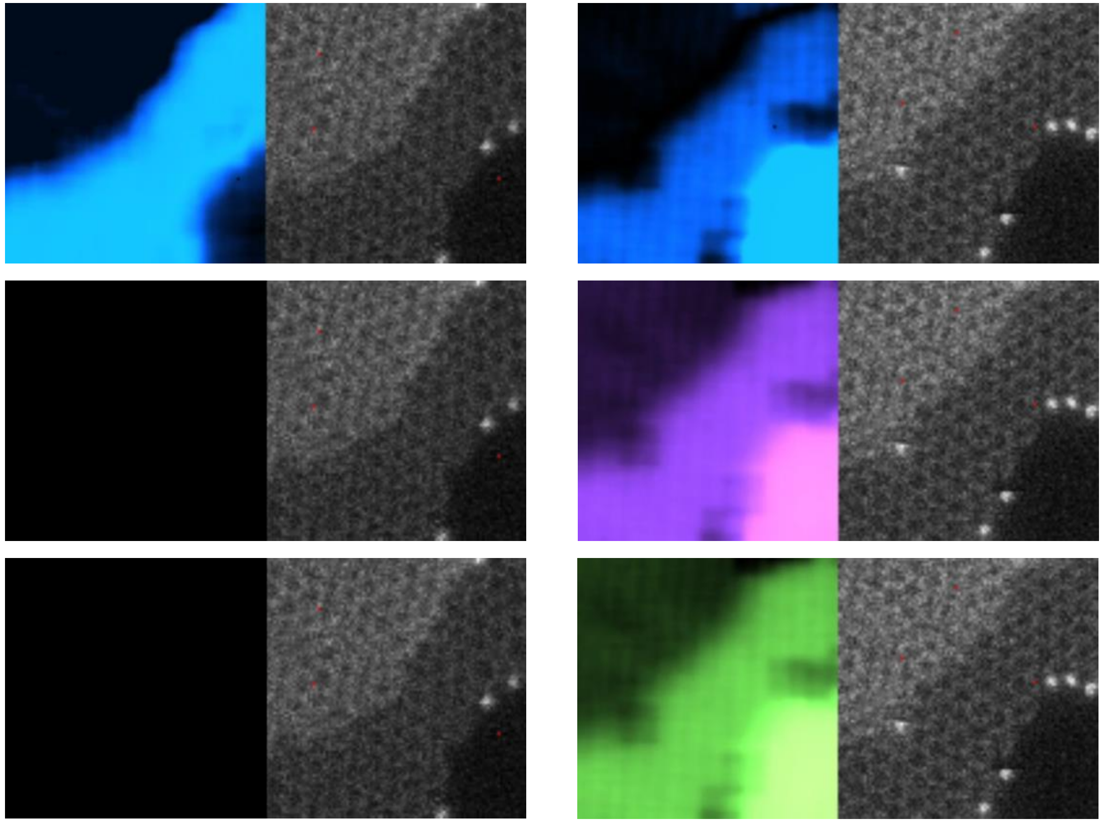
CoM angle

CoM magnitude

Acquisition function

Prediction map

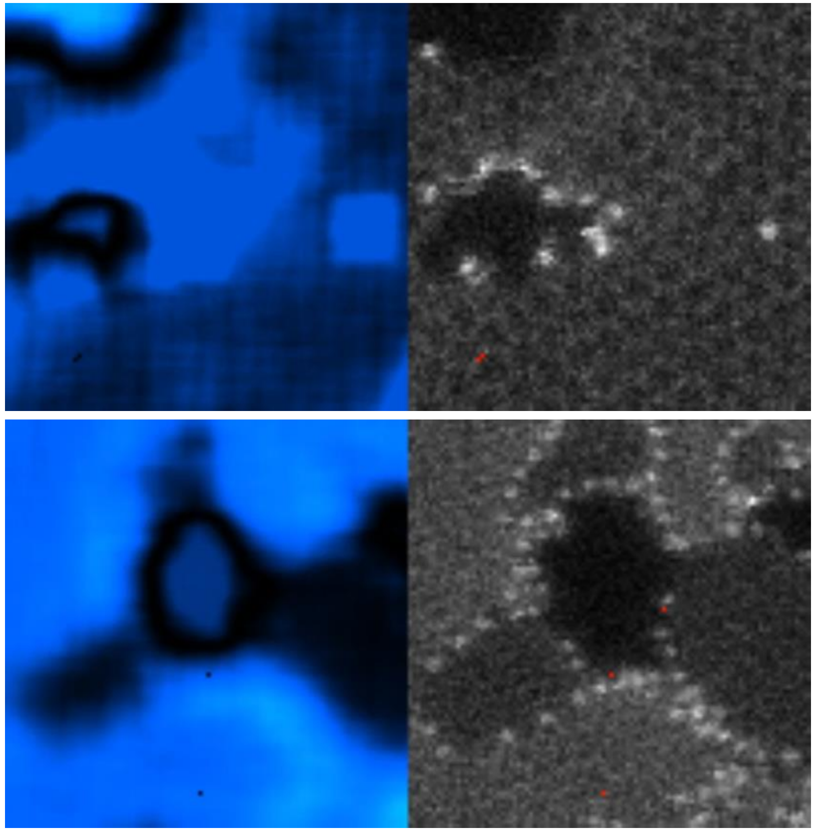
Uncertainty map



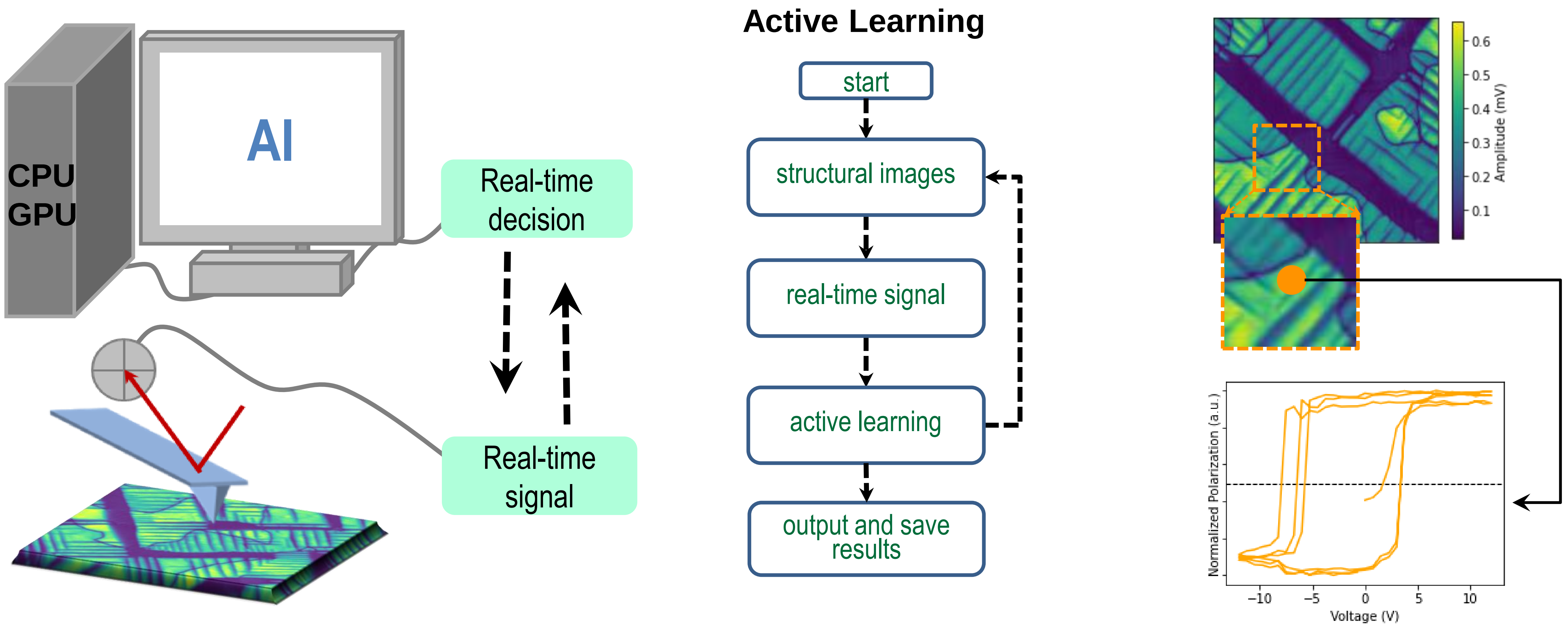
Real experiment

Next measurement map

Visited points

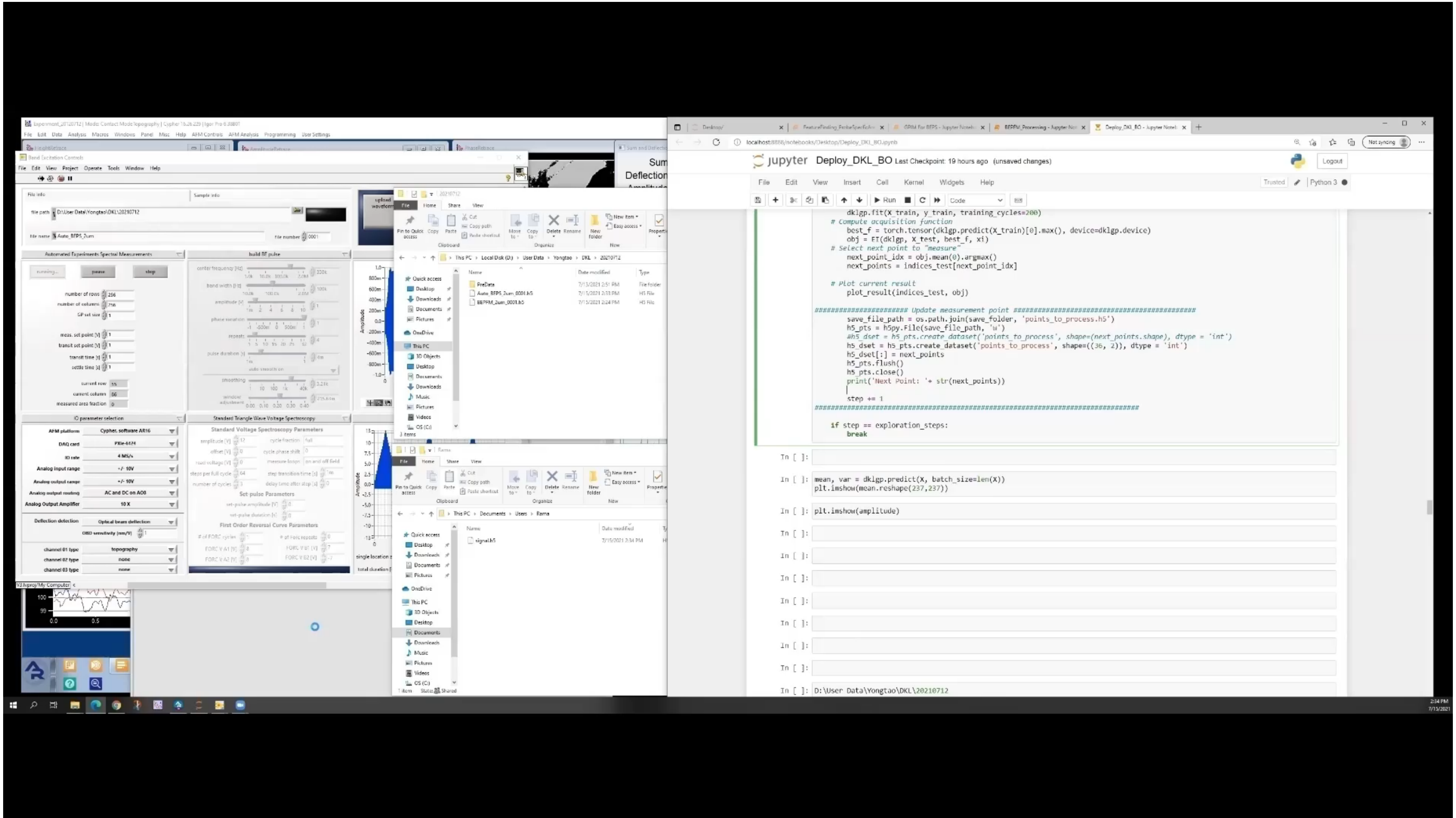


Deep Kernel Learning for PFM



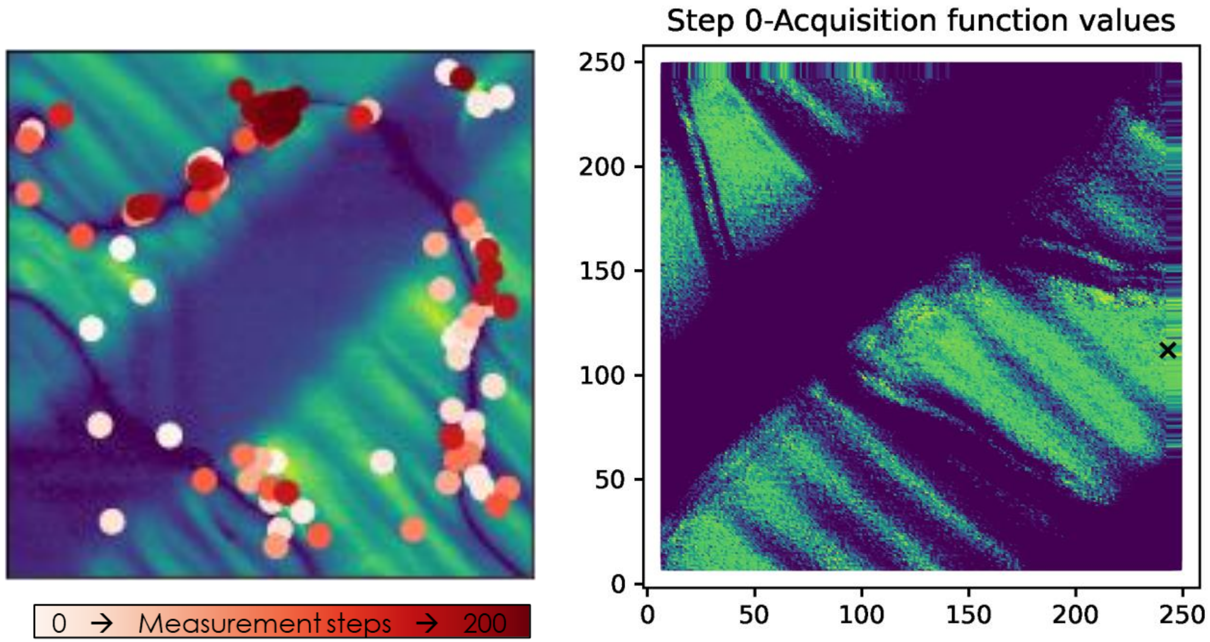
➢ (Deep Kernel Learning) Active learning of structure-property correlation.

Deep Kernel Learning AE

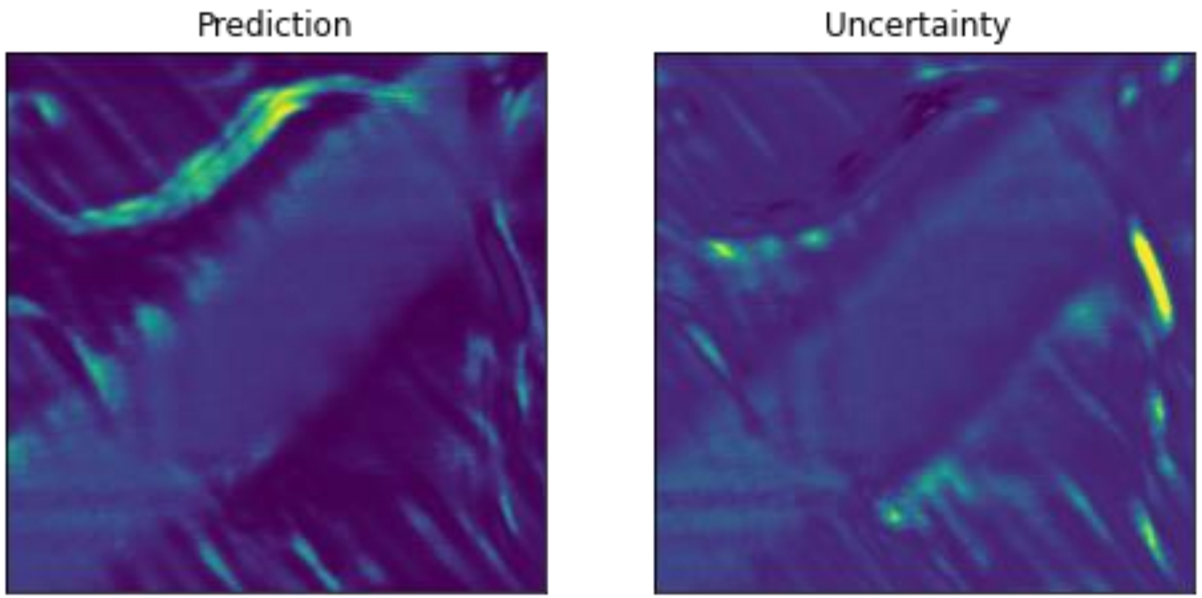


DKL AE

Guided by: Off field loop area



Results: DKL predicted loop area map



- ➢ Large loop opening corresponding 180º domain walls probably due to the large polarization mobility of 180º walls.
- ➢ Future work: DKL-nonlinearity study of HZO; CIPS domain walls.

Automated Experiment:

… as a scientist…

Bayesian optimization:

1. Works only in low-dimensional spaces
2. The correlations are defined by the kernel function (very limiting)
3. We do not use any knowledge about physics of the system
4. We do not use cheap information available during the experiment (proxies)

GP Augmented with Structural model

Define a probabilistic model:

$$\mathbf{y} \sim MVNormal(\mathbf{m}, \mathbf{K})$$

$$K_{ij} = \sigma^2 \exp(0.5(x_i - x_j)^2 / l^2$$

$$\sigma \sim LogNormal(0, s_1)$$

$$l \sim LogNormal(0, s_2)$$

- We substitute a constant GP prior mean function **m** with a structured probabilistic model of the expected system's behavior.
- This probabilistic model reflects our prior knowledge about the system, but it does not have to be precise.
- The model parameters are inferred together with the kernel parameters via the Hamiltonian Monte Carlo.
- The fully Bayesian treatment of the model allows additional control over the optimization via the selection of priors for the model parameters.

Prediction on new data X_*:

$$\mathbf{f}_*^i \sim MVNormal\left(\mu_{\boldsymbol{\theta}^i}^{\text{post}}, \Sigma_{\boldsymbol{\theta}^i}^{\text{post}}\right)$$

replaced with

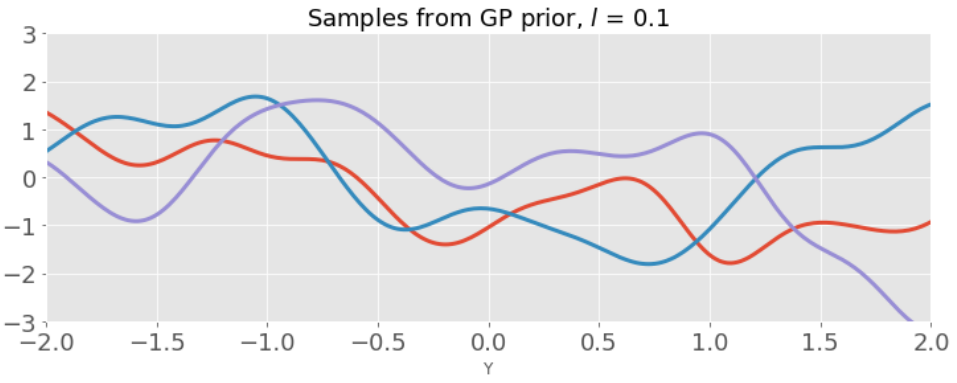
$$\mu_{\boldsymbol{\theta}^i}^{\text{post}} = \mathbf{m}(X_*) + \mathbf{K}(X_*, X|\boldsymbol{\theta}^i)\mathbf{K}(X, X|\boldsymbol{\theta}^i)^{-1}(\mathbf{y} - \mathbf{m}(X)) \rightarrow \mu_{\boldsymbol{\Omega}^i}^{\text{post}} = \mathbf{m}(X_*|\phi^i) + \mathbf{K}(X_*, X|\boldsymbol{\theta}^i)\mathbf{K}(X, X|\boldsymbol{\theta}^i)^{-1}(\mathbf{y} - \mathbf{m}(X|\phi^i))$$

$$\Sigma_{\boldsymbol{\theta}^i}^{\text{post}} = \mathbf{K}(X_*, X_*|\boldsymbol{\theta}^i) - \mathbf{K}(X_*, X|\boldsymbol{\theta}^i)\mathbf{K}(X, X|\boldsymbol{\theta}^i)^{-1}\mathbf{K}(X, X_*|\boldsymbol{\theta}^i)$$

$\boldsymbol{\Omega}^{\boldsymbol{i}} = \{\phi^i, \boldsymbol{\theta}^i\}$ is a single HMC posterior sample with the kernel and prob model parameters

GP Augmented with Structural model

Prior predictive distribution: GP



Probabilistic model

$$m = y_0 - \sum_{n=1}^{N} L_n \quad (N{=}2)$$

$$y_0 \sim Uniform(-10, 10)$$

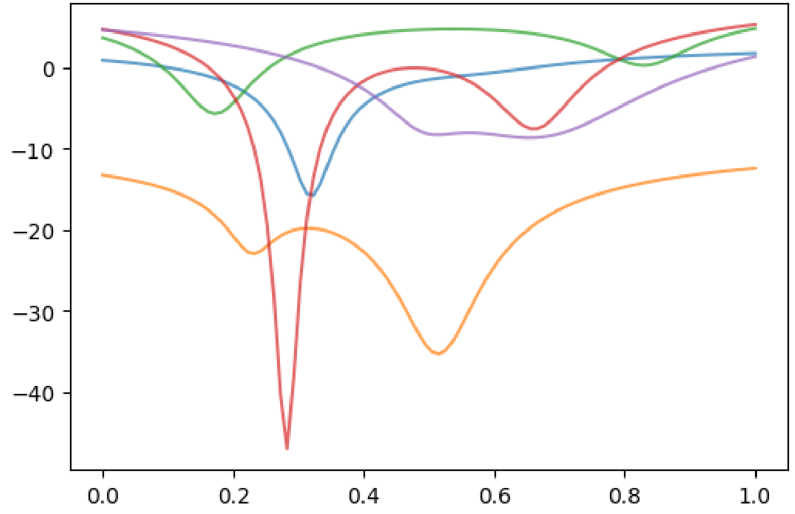
$$L_n \sim \frac{A_n}{\sqrt{(x - x_n^0)^2 + w_n^2}}$$

$$A_n \sim LogNormal(0, 1)$$

$$w_n \sim HalfNormal(.1)$$

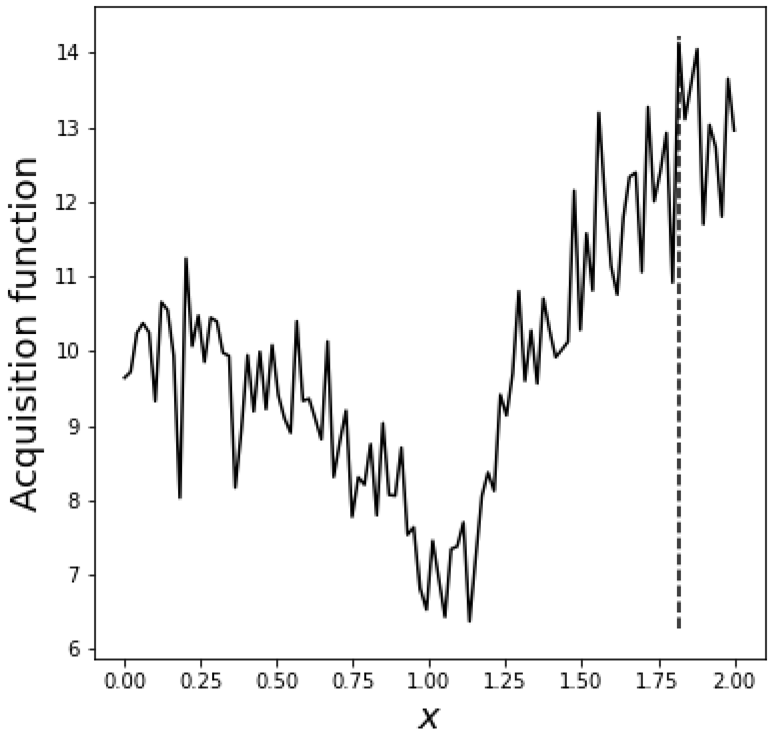
$$x_n^0 \sim Uniform(0, 1)$$

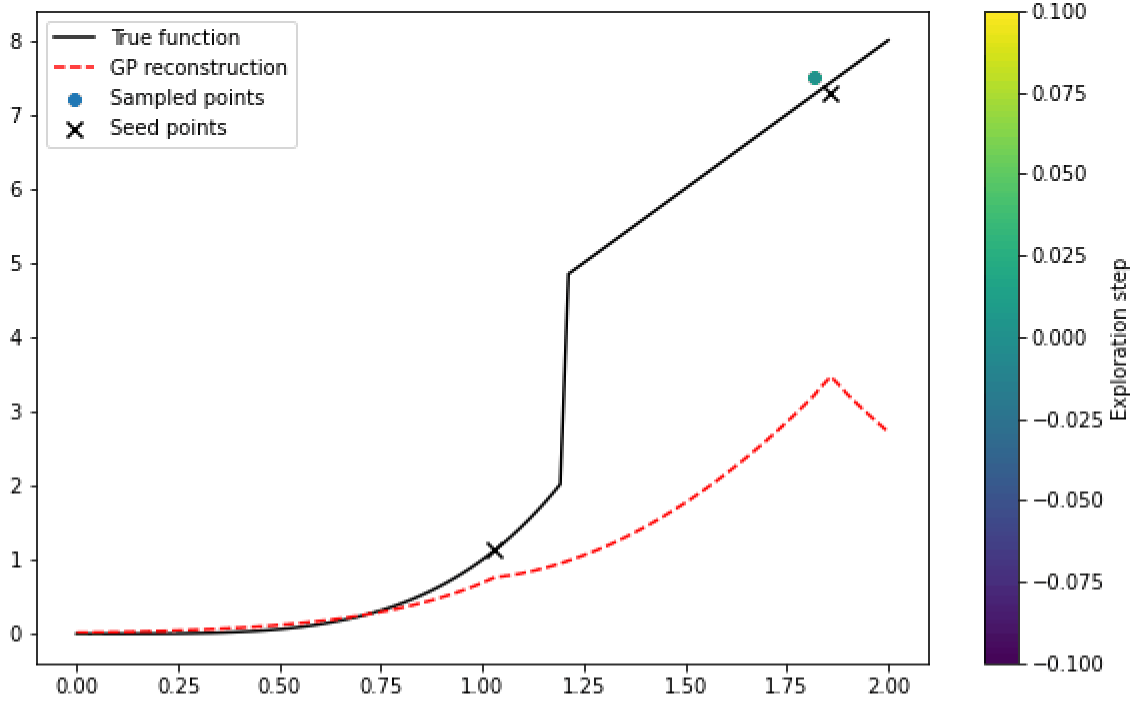
Prior predictive distribution: sGP



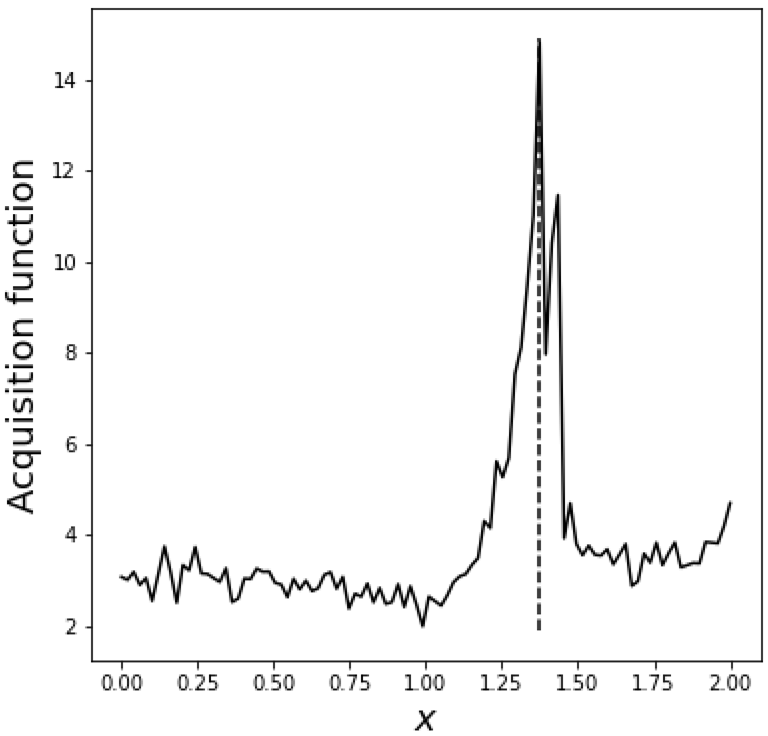
This model simply tells us that there are two minima in our data but does not assume to have any prior knowledge about their relative depth, width, or distance

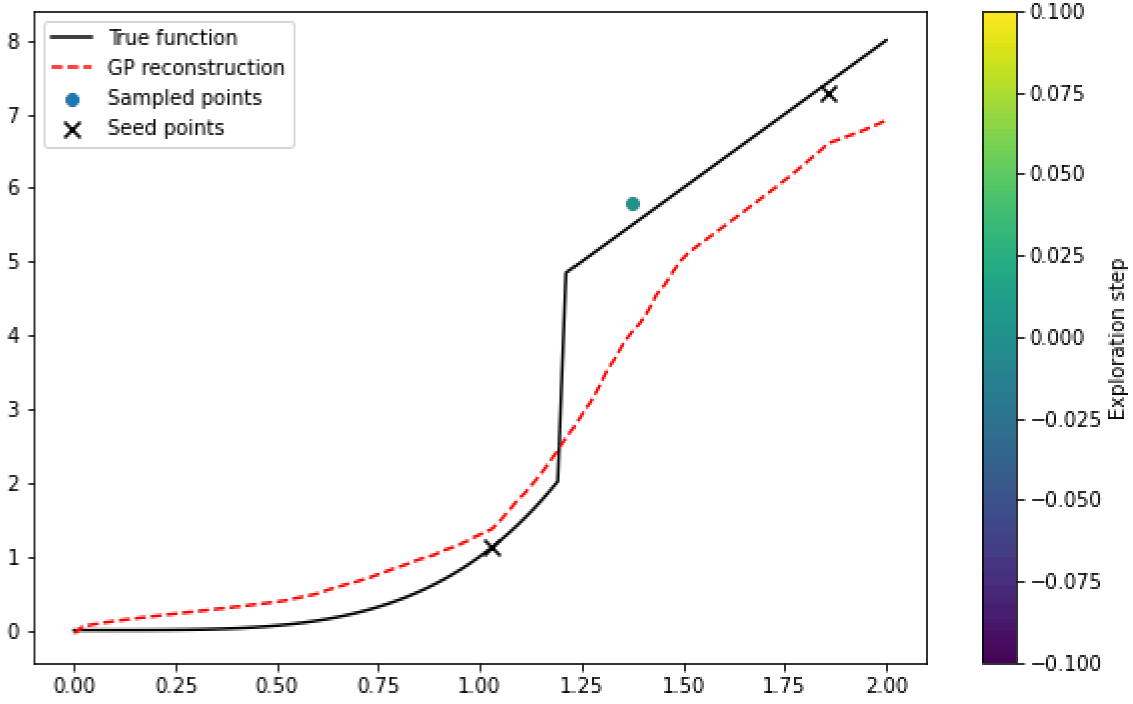
Simple GP search





Structured GP search



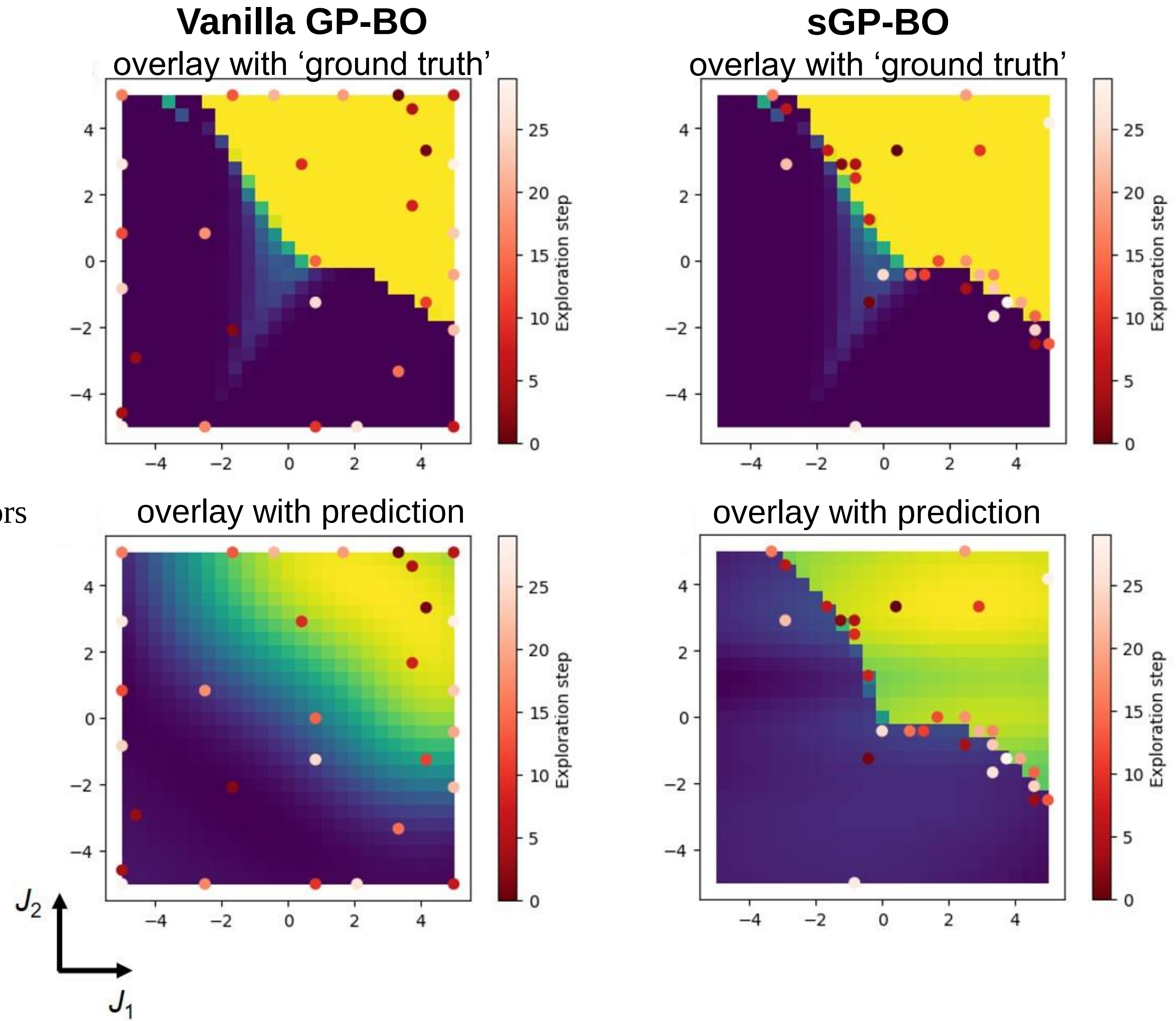


Application to Ising model

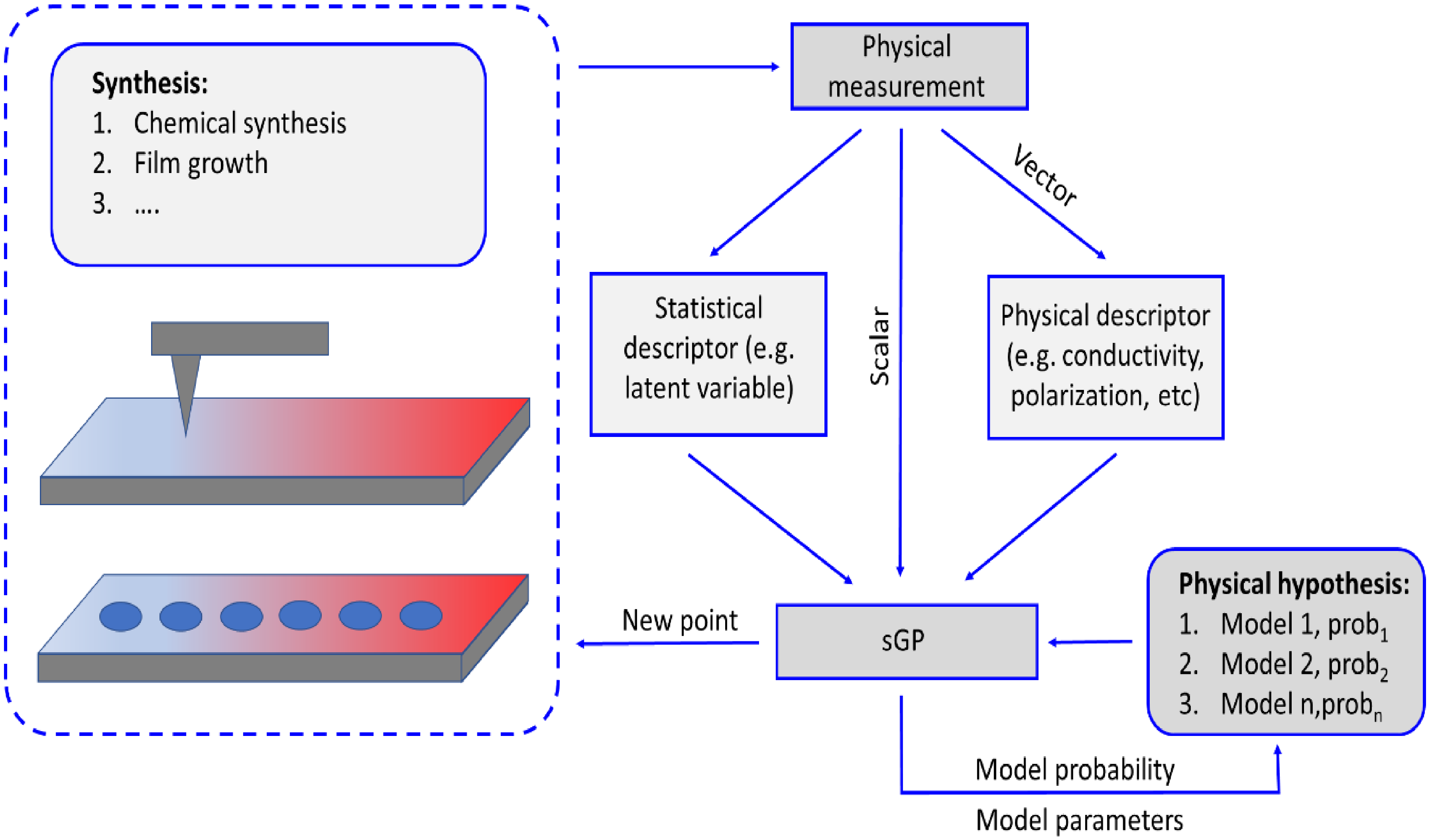
Probabilistic model

$$A/\tanh\left(\frac{f(J_1)+f(J_2)}{w}\right)$$

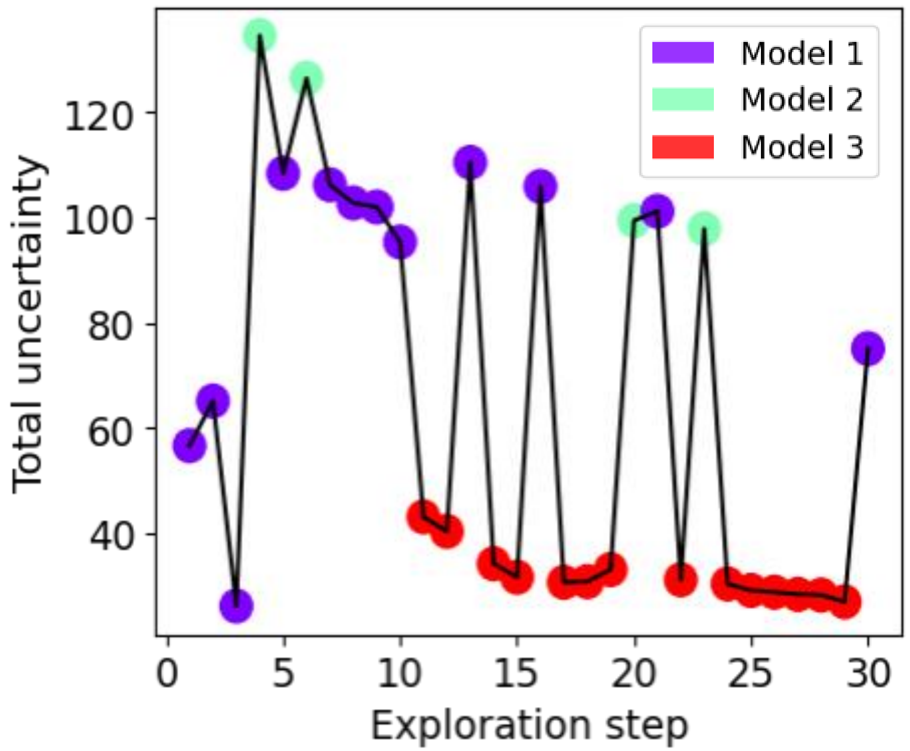
where $f(J)$ is a third-degree polynomial with normal priors on its parameters

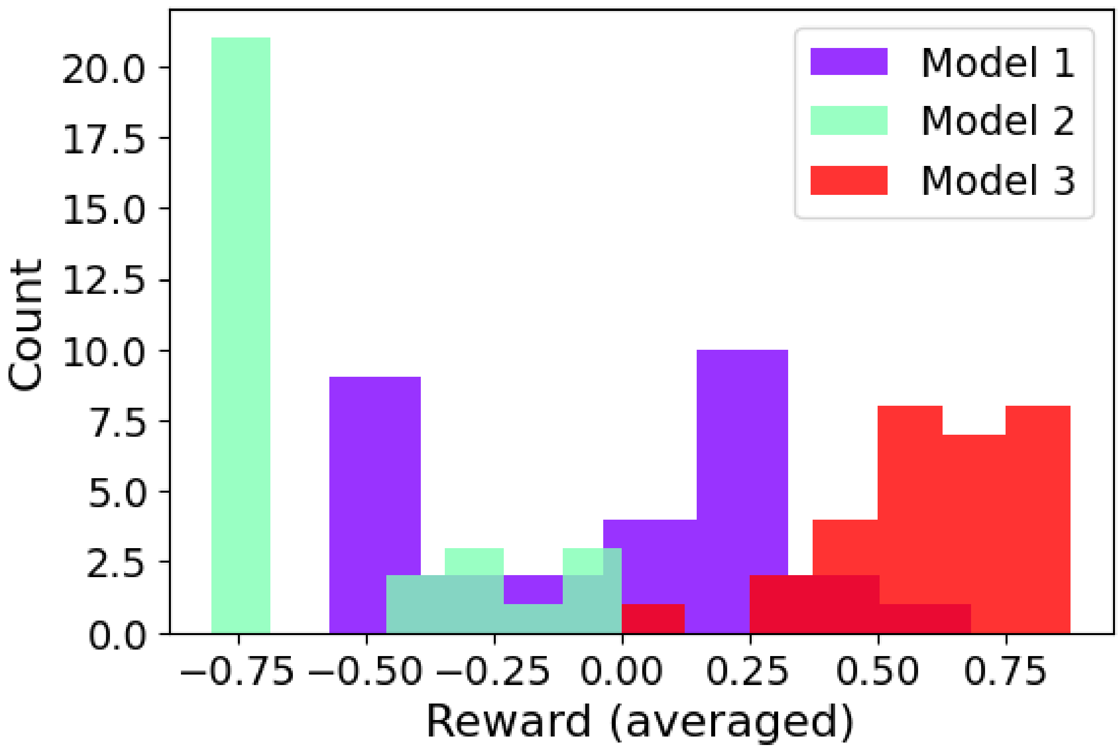


Hypothesis active learning: hypoAL

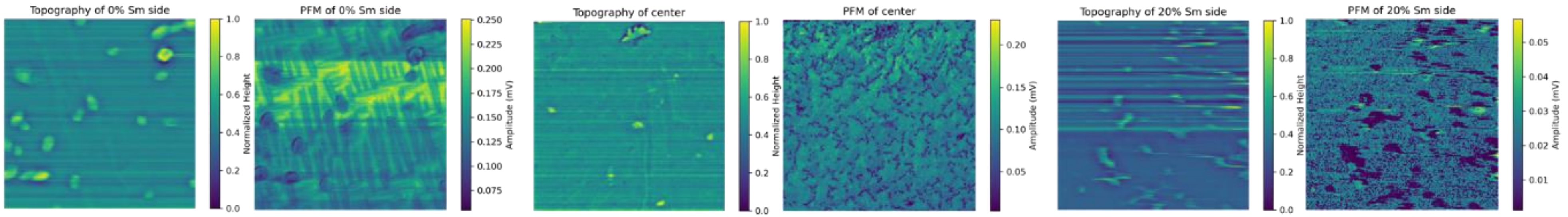


Next step: active model selection



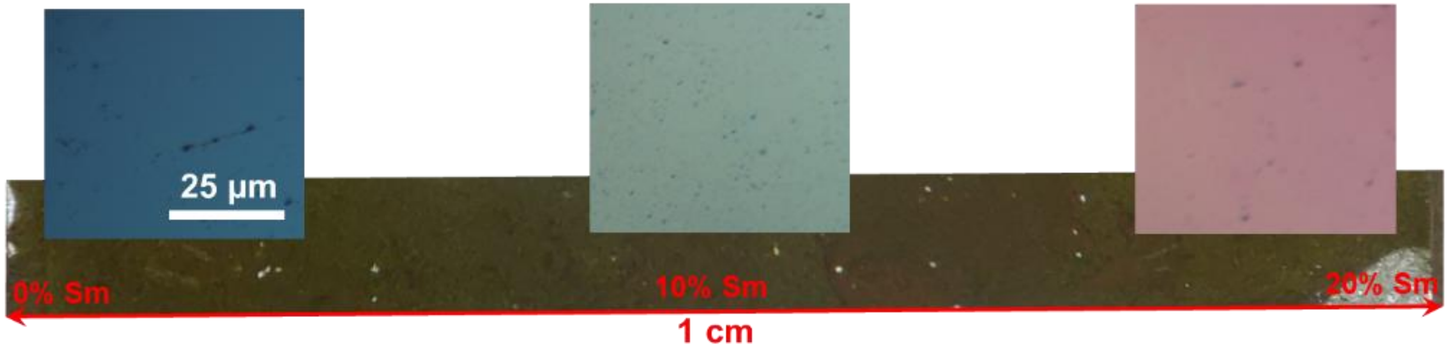


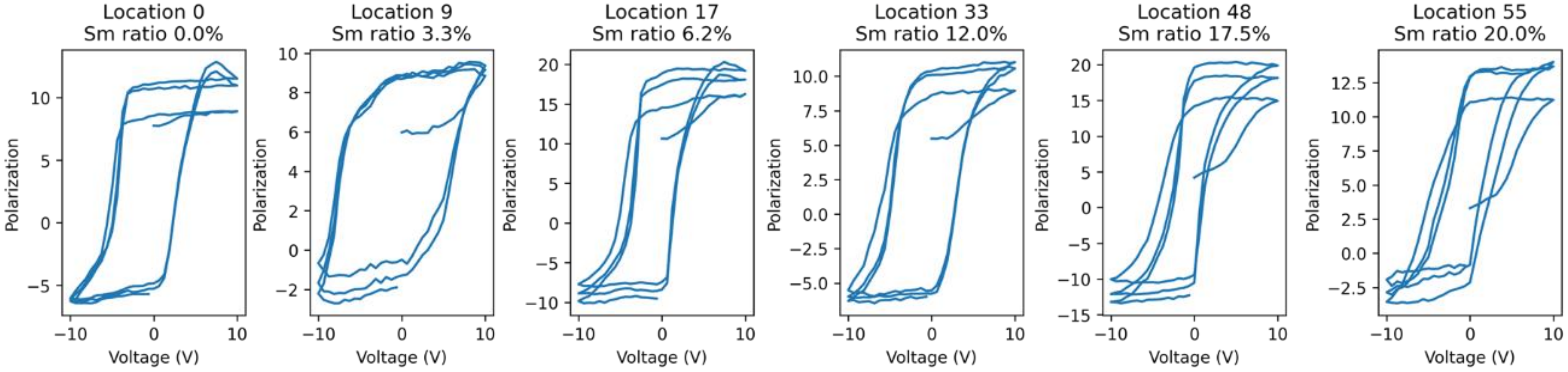
Back to combinatorial libraries:



optical images

sample picture





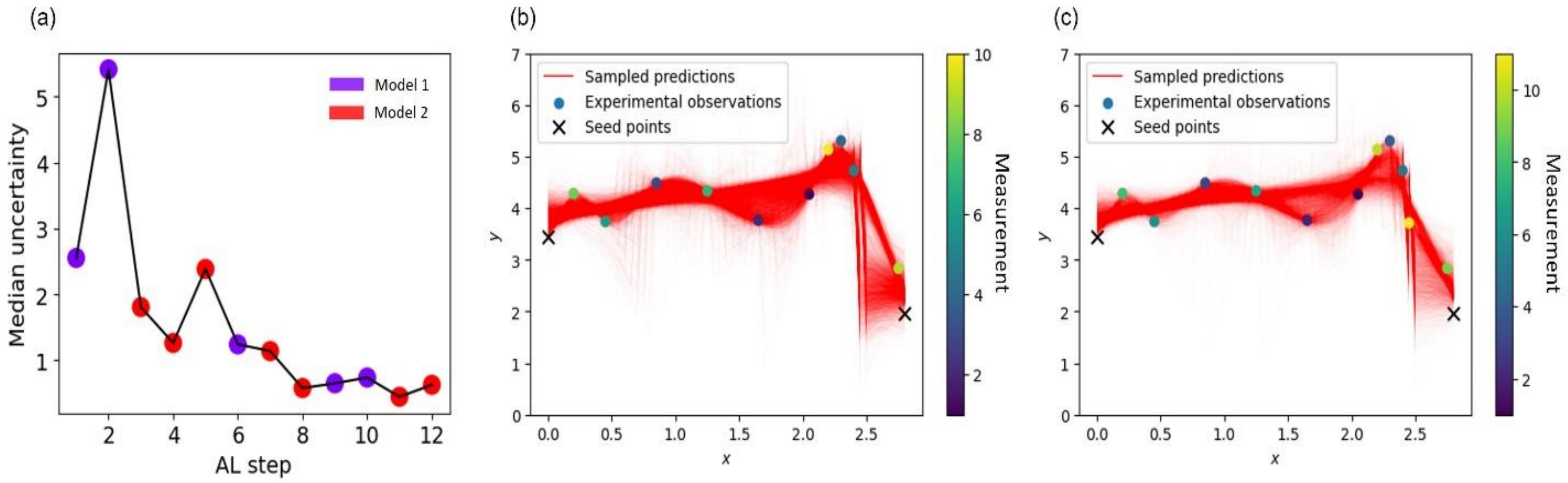
Hypothesis selection for ferroelectric response

Model 1 (second order phase transition):

$$S = \begin{cases} S_0\left(1 - \dfrac{x}{x_0}\right)^2 + C, & x \le x_c, \\ C, & x > x_c \end{cases}$$

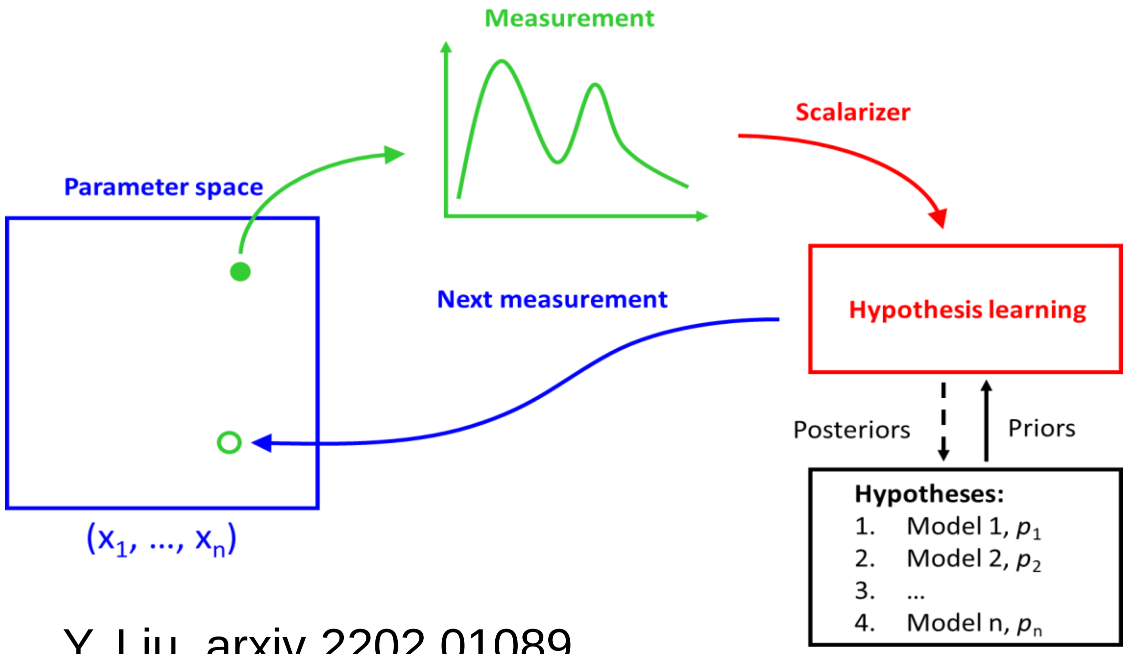
Model 2 (first order phase transition):

$$S = \begin{cases} S_0\left(1 - \dfrac{x}{x_0}\right)^{\frac{5}{4}} + C_0, & x \le x_c, \\ C_1, & x > x_c \end{cases}$$

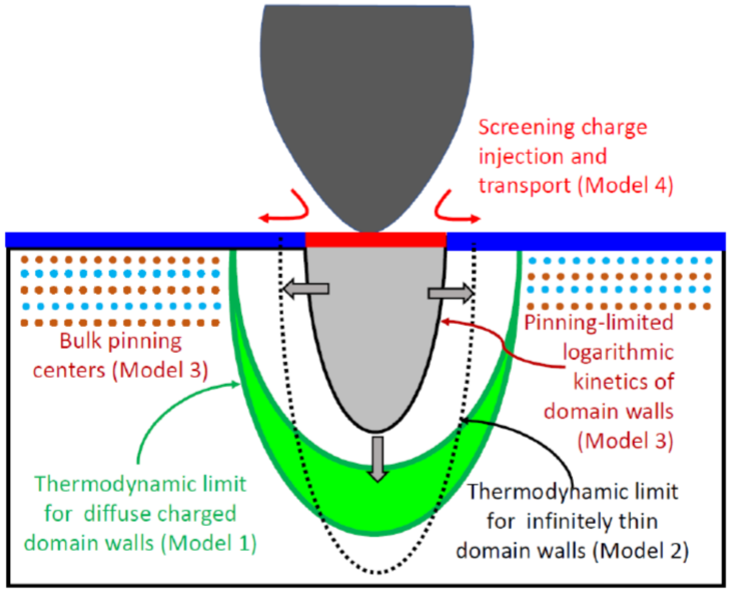


Hypothesis Learning

- Can ML algorithm think like a scientist?
- Yes – automated experiment can pursue hypothesis-driven science!

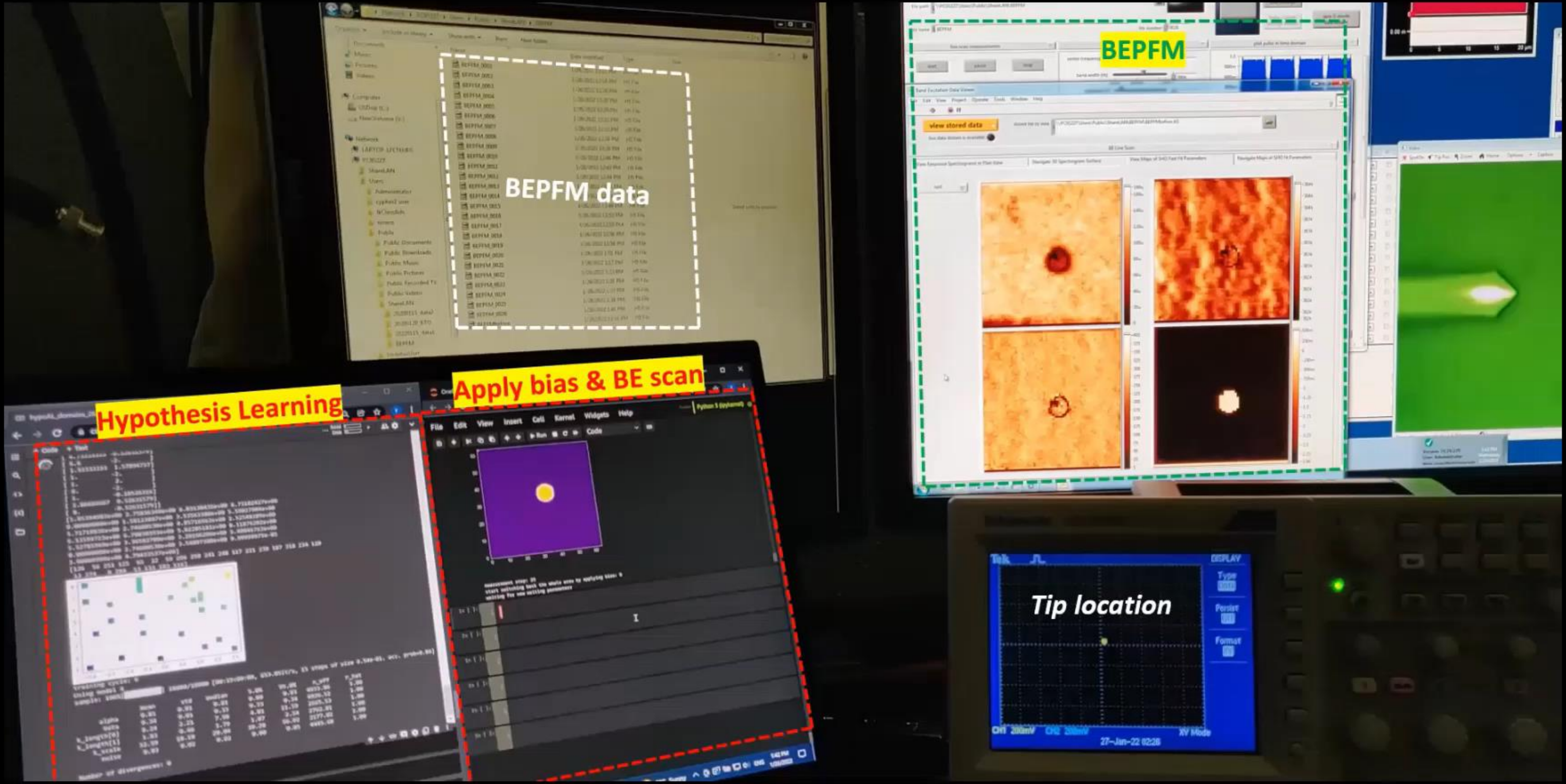


Y. Liu, arxiv 2202.01089
Y. Liu, arxiv 2112.06649

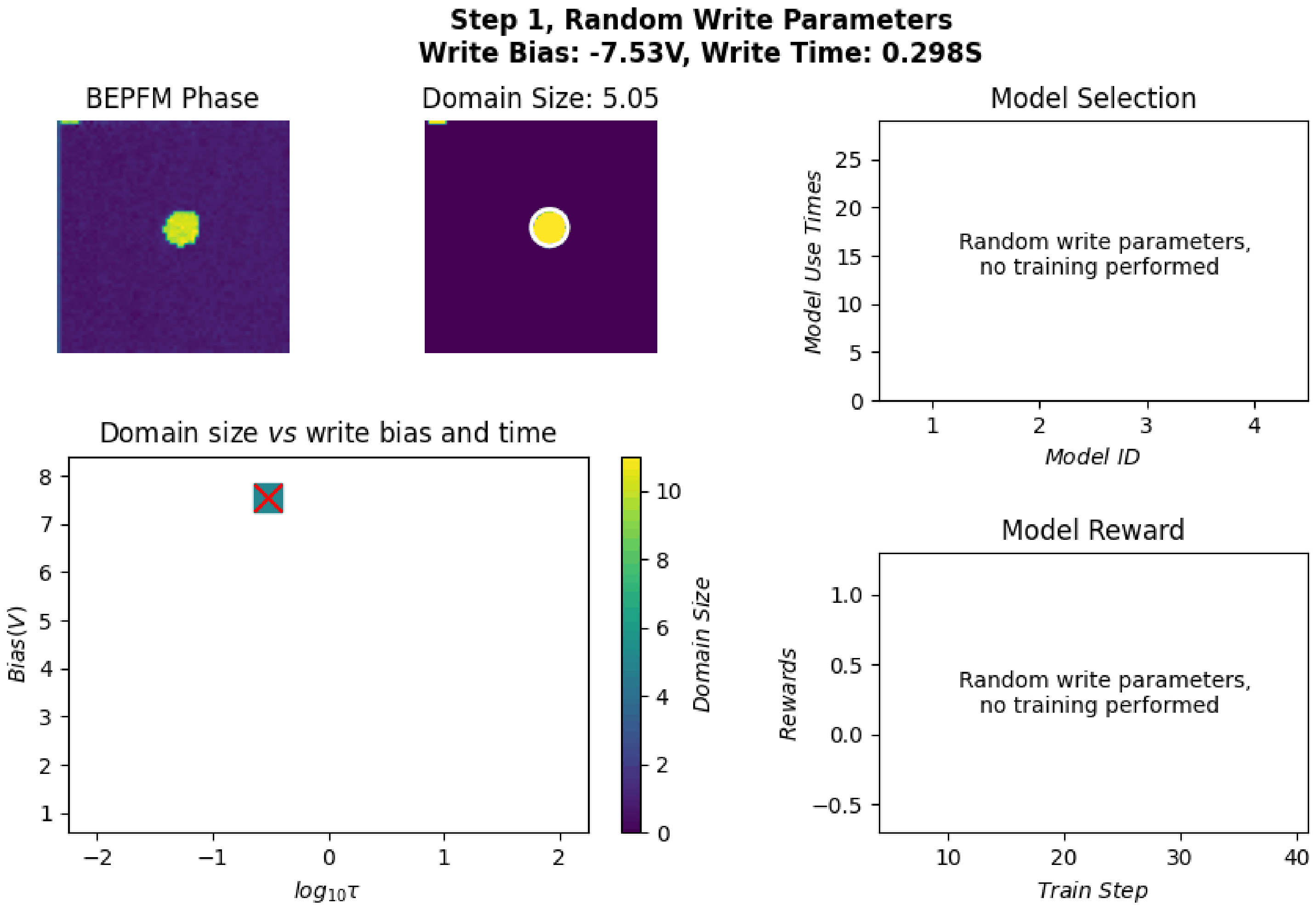


	Model	Equation
Thermodynamic 1	**Model I**	$r(V) = r_{cr} + r_0\sqrt{\left(\frac{V}{V_c}\right)^{2/3} - 1}$
Thermodynamic 2	**Model II**	$r(V) = r_{cr} + r_0\sqrt[3]{\left(\frac{V}{V_c}\right)^{2} - 1}$
Wall pinning	**Model III**	$r(V,t) = V^{\alpha}\log\tau$
Charge injection	**Model IV**	$r(V,t) = V^{\alpha}\tau^{\beta}$

Automated Experiment #3—Automated Hypo-learning



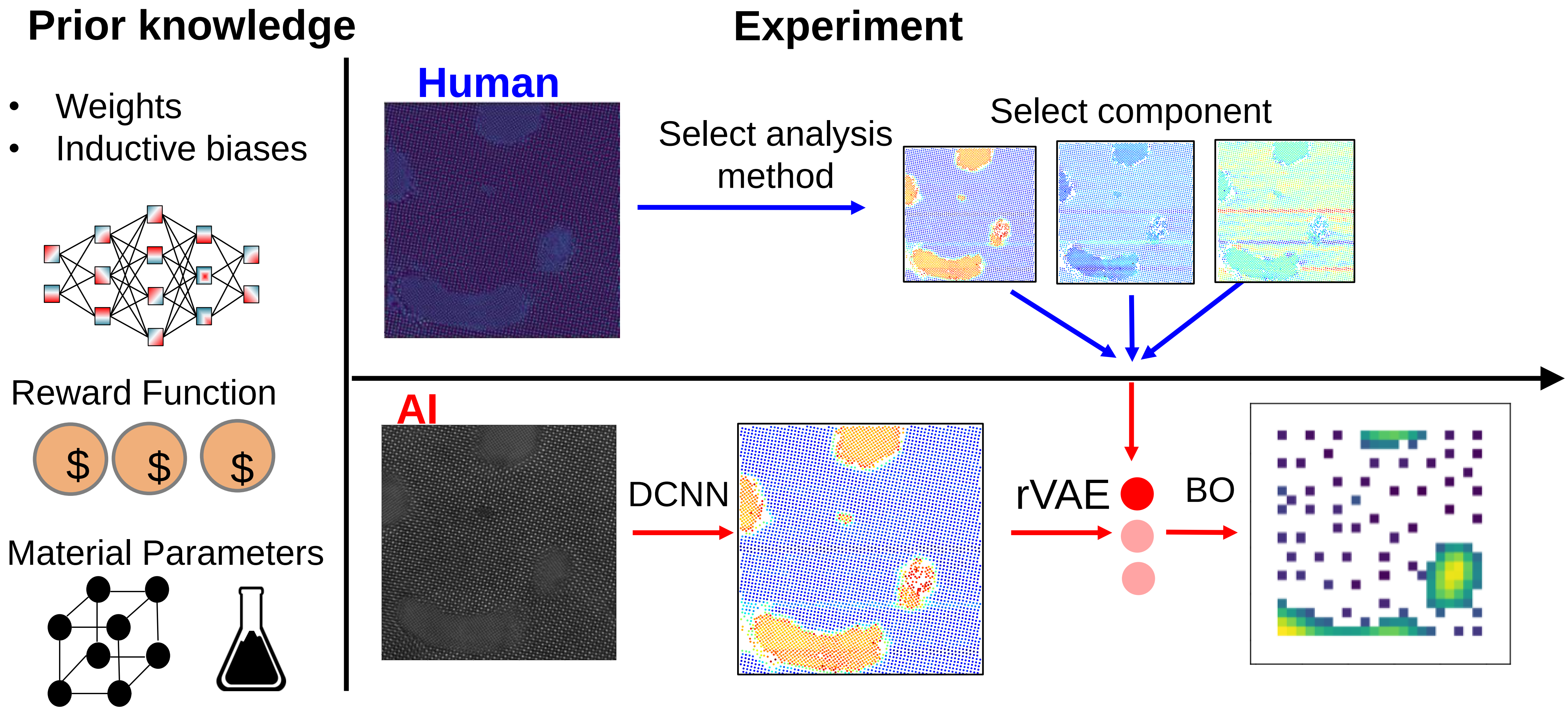
Hypothesis learning in action



- ML algorithm has 4 competing hypothesis on domain switching mechanisms
- These hypothesis represent full set of possibilities for this system
- The microscope chooses experimental parameters in such a way as to establish which hypothesis is correct fastest
- Important: the same approach can be implemented in synthesis and electrical characterization
- Machine learning meets hypothesis-driven scientific discovery!

Y. Liu, arxiv 2202.01089
Y. Liu, arxiv 2112.06649

More then ML: Human in the Loop



Discovery pathway depends on the reward structure!

The World is Bayesian: Physics from Observations

Hypothesis driven science:
What we want to learn

Forward model: Theory

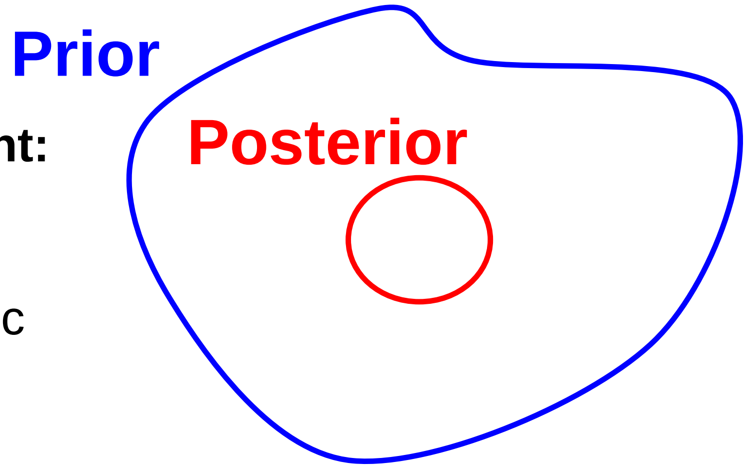
Domain expertise:

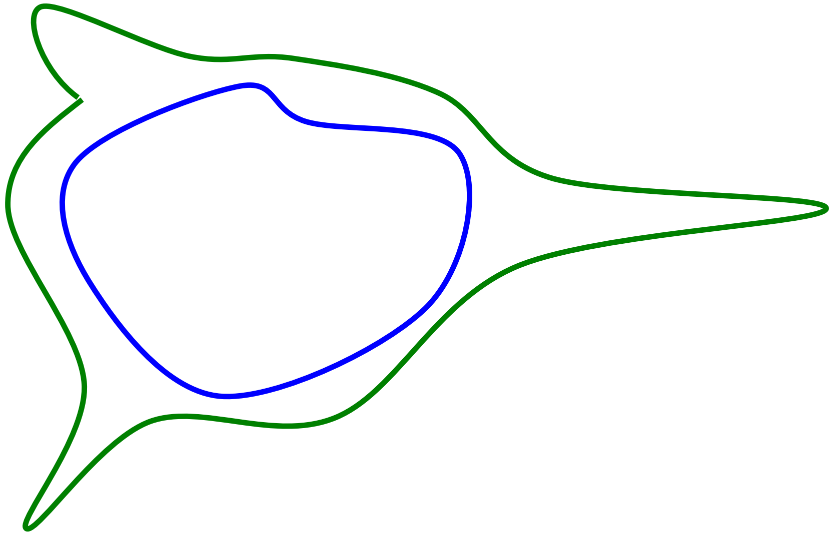
$$P(Theory|Data) = \frac{P(Data\,|Theory)\,P(Theory)}{P(Data)}$$

High Performance Computing

- Experimentalists know the priors. Albeit they do not know that they know it, or how to convert them to algorithmic form
- **However, how do we make guesses about the unknown?**

Refinement: Can be defined as probabilistic model

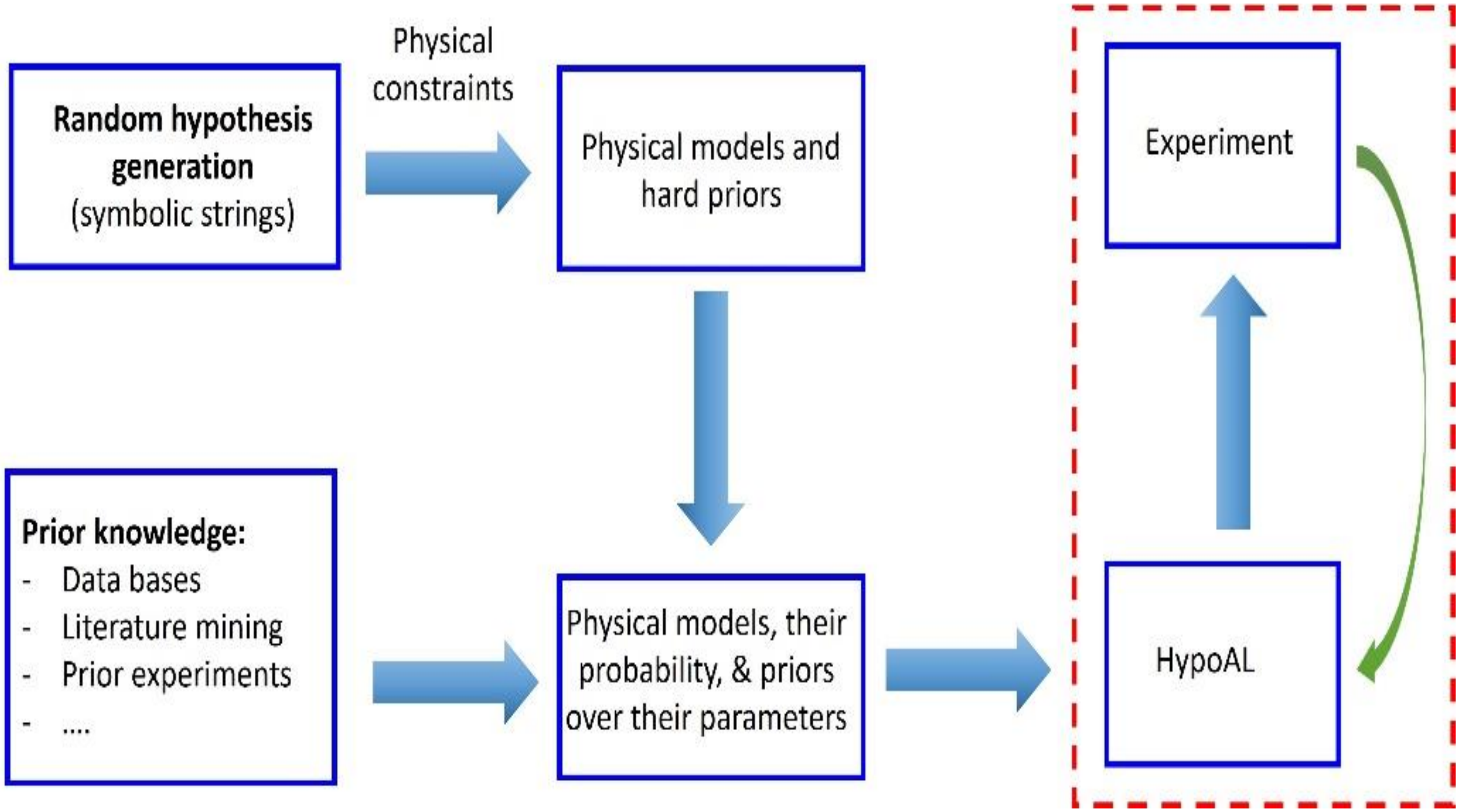




Hypothesis formation: How can we do it?

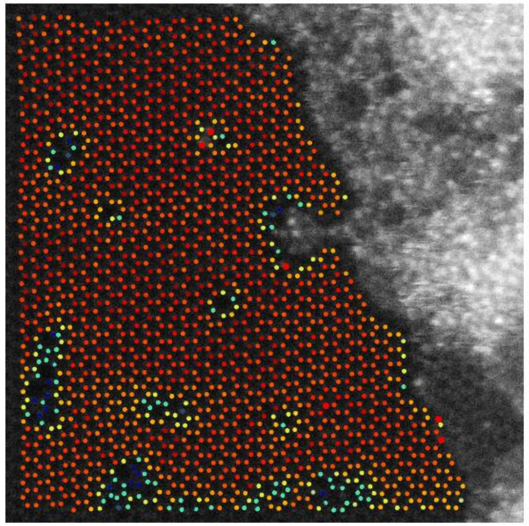
Towards hypothesis learning

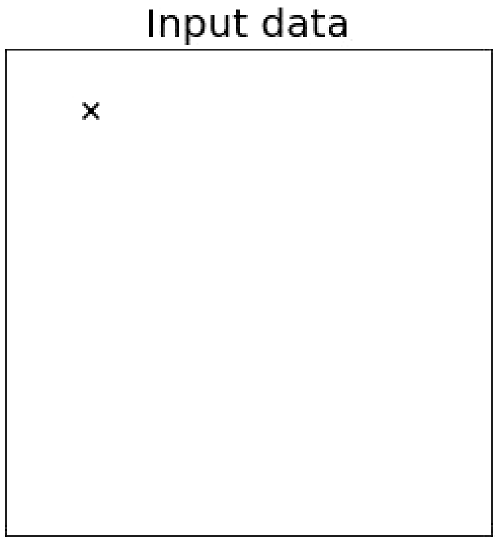
- Data already exists: Eureka, SISSO, SinDY, etc.
- What if we make hypothesis learning a part of active experiment?
- Need policies for hypothesis generation



What if I want to do it myself?

- **AtomAI:** comprehensive toolbox for DCNN-based supervised exploration of STEM and SPM Data:
- **PyroVED:** building structure-property correlations and unsupervised and semi-supervised physical discovery
- **GPim:** Gaussian processing toolbox for image analytics and automated experiment
- **gpax:** hypothesis-driven structured Gaussian Processes
- **PyCroscopy:** General data formats, workflows, and image analytics





YouTube: M*N: Microscopy, Machine Learning, Materials
Medium: https://ziatdinovmax.medium.com/

State of microscopy now

- Ideas
- Hypotheses
- Contacts from people
- Us reaching out

Materials:
- Single crystals
- Films
- Powders

STEM samples

Imaging and spectroscopy data

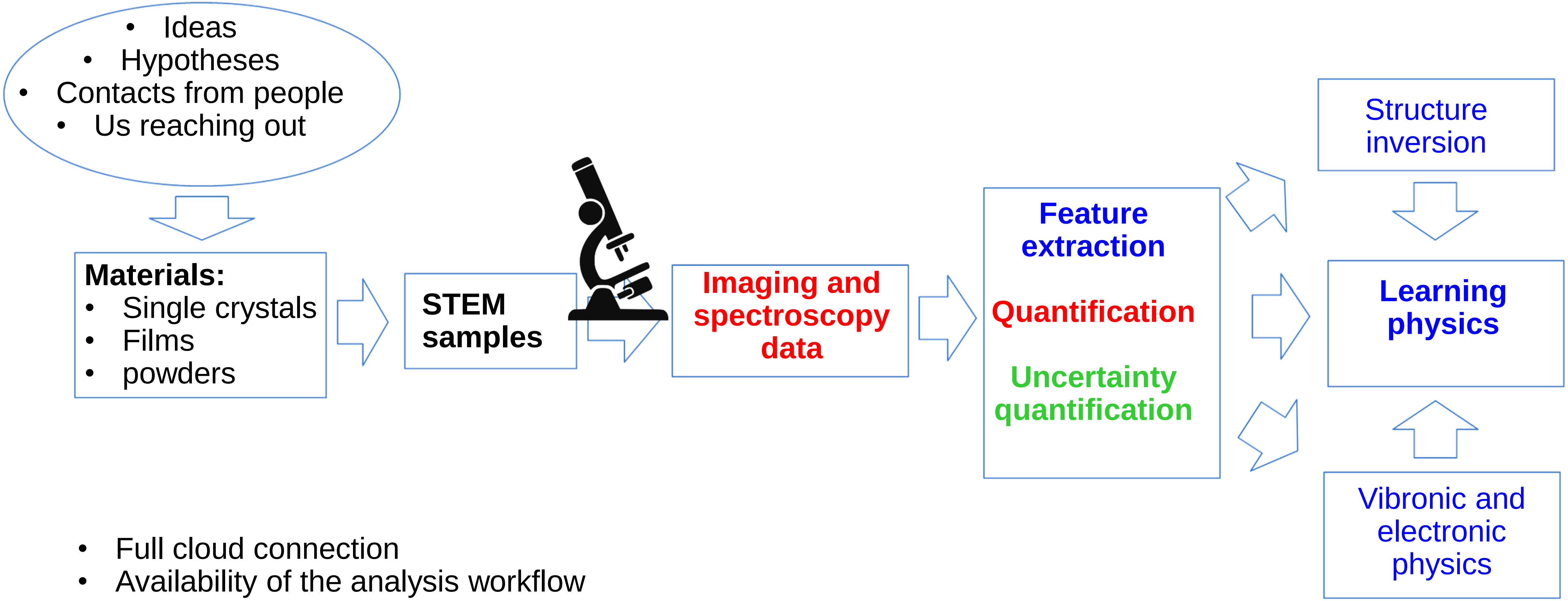
"Qualitative" image you can put in the paper

Paper on how microscope works/technique advancement

Theory-experiment comparison based on "simple matching"

Questions:
- Where is the physics?
- Is this a fishing expedition?
- hypothesis driven science?

State of microscopy in the future

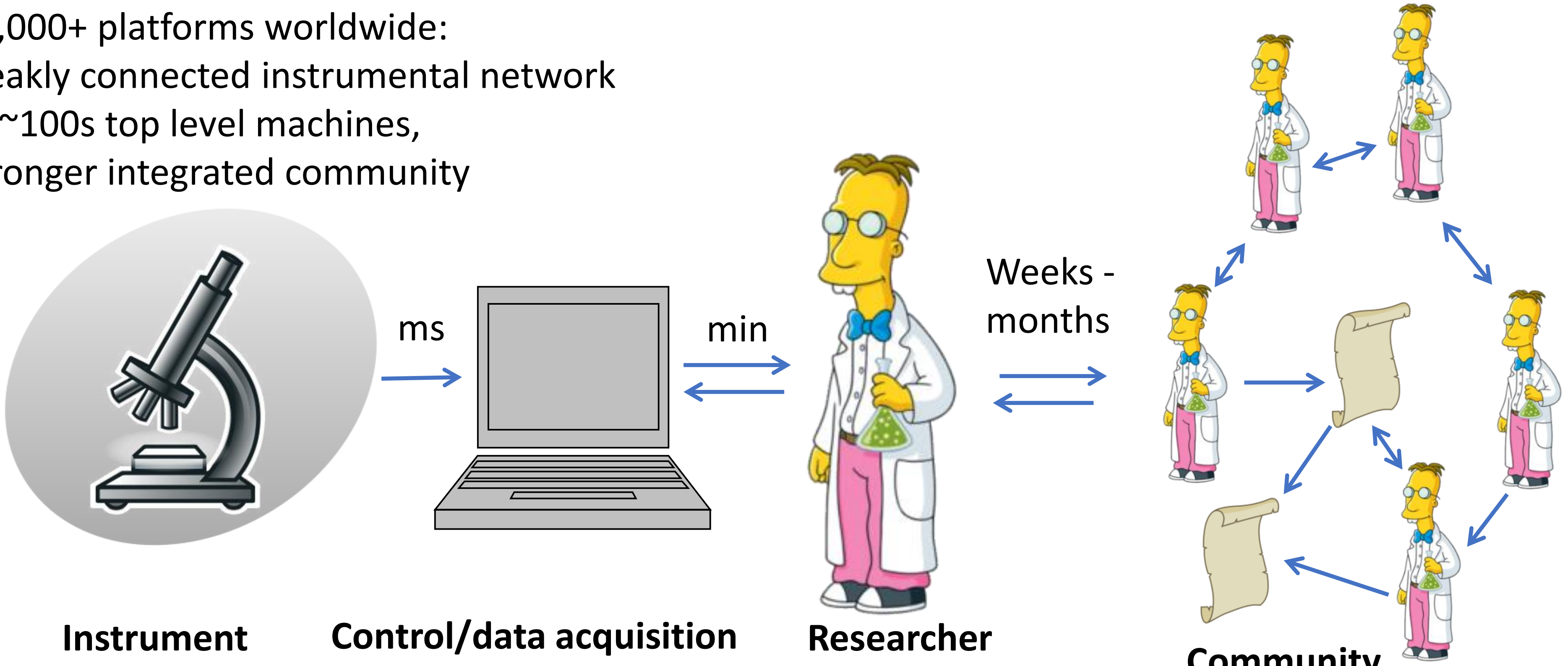


- Full cloud connection
- Availability of the analysis workflow
- Current status: ad-hoc local solutions/no scalability

- **Chicken and egg problem:** just cloud connection and data infrastructure is not enough if there is no way to analyze data

Classical Instrument research paradigm

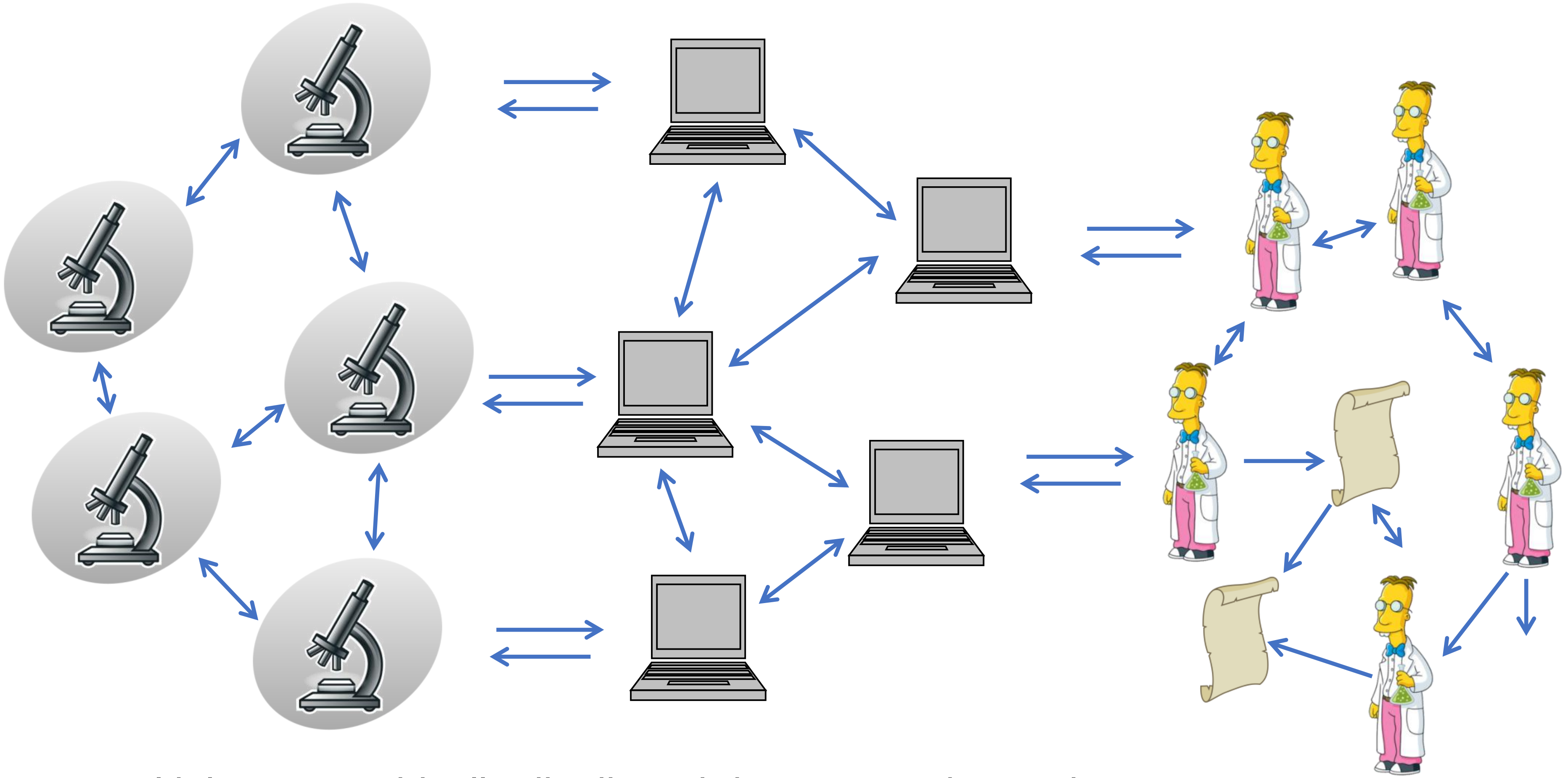
SPM: 30,000+ platforms worldwide:
Large weakly connected instrumental network
(S)TEM: ~100s top level machines,
much stronger integrated community



- Social networking/education
- Publications/citations

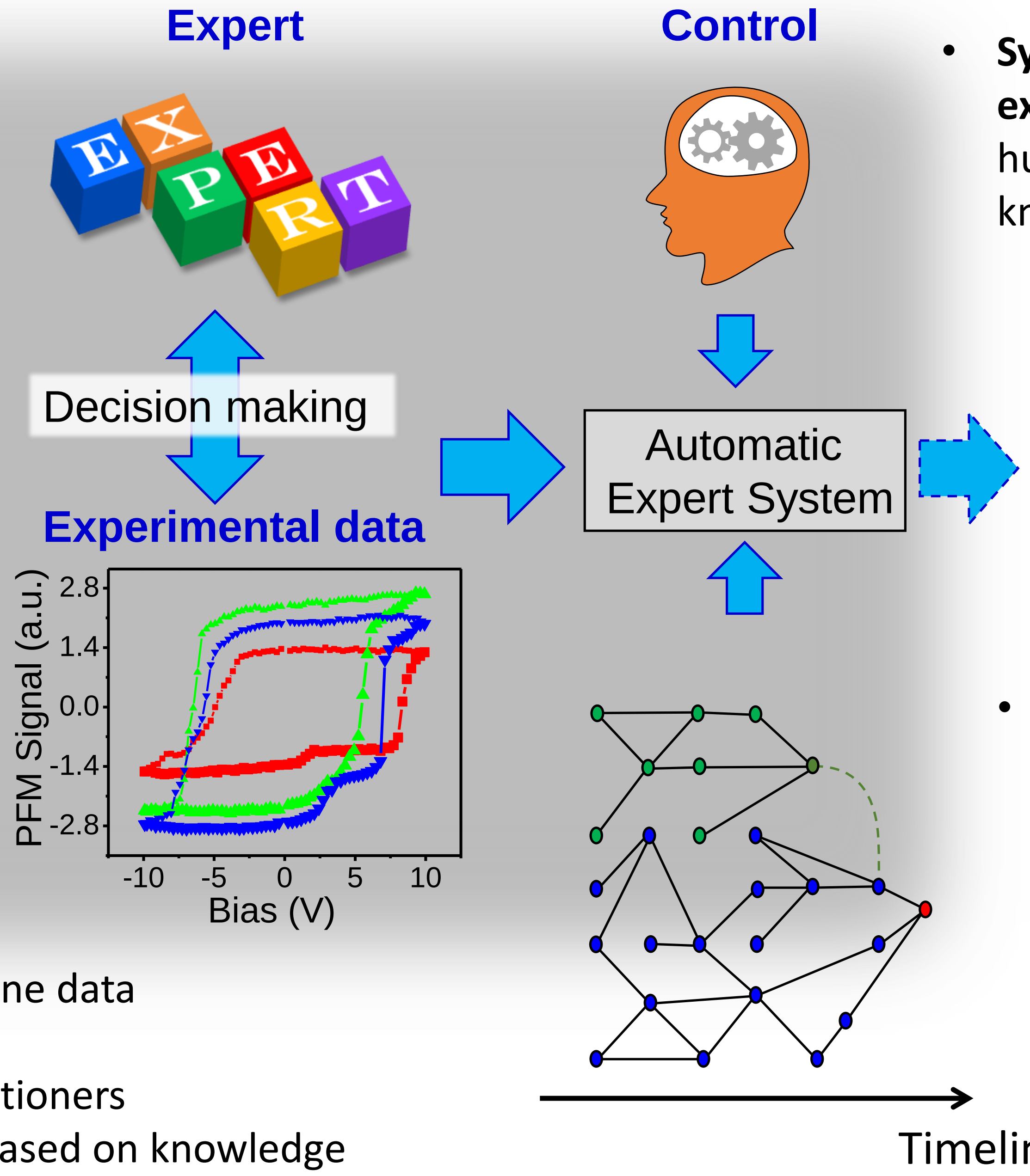
1. Only small fraction of data stream from the instrumentation is captured
2. Only small fraction of captured data is analyzed, interpreted, and put in the context
3. Human-machine interaction during acquisition is often slow and can be non-optimal
4. Human interpretation of data is limited: bias and ignoring serendipity
5. Information propagation and concept evolution in scientific community is slow and affected by non-scientific factors

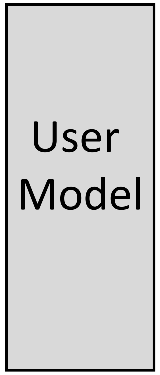
Imaging in the Cloud



1. Multiple geographically-distributed data generation node
2. Full capture of instrumental data stream /compression/curation
3. Coordination of protocols and data/metadata across the cloud
4. Cloud-based processing and dimensionality reduction
5. Community-wide analytics

Opportunities: from single human expert to augmented systems



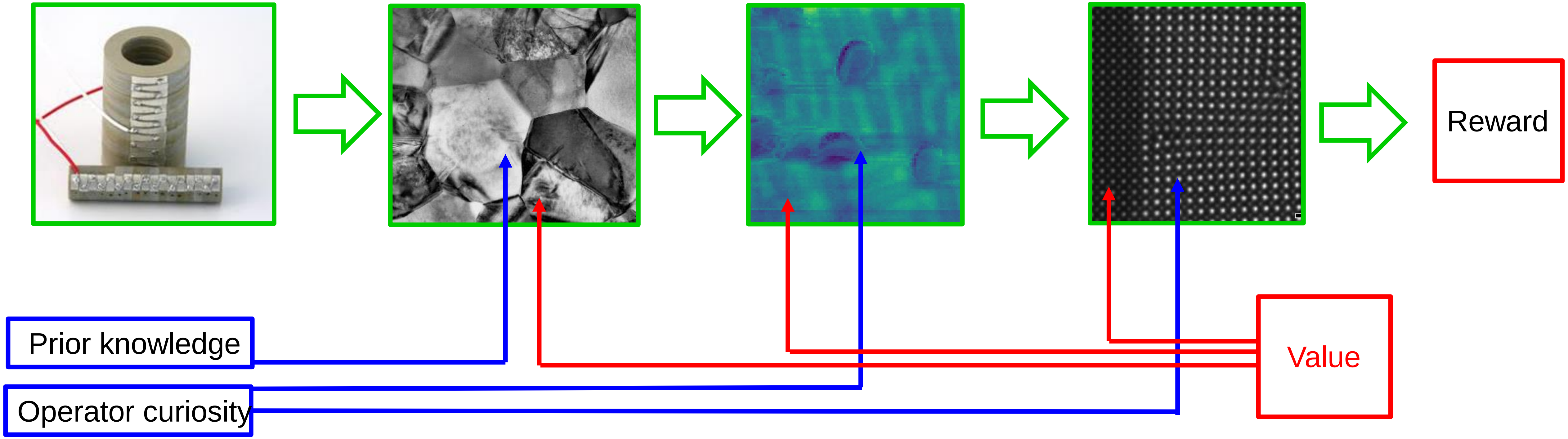


- **Synthesis of expertise:** factor in human expert knowledge
- **Context search:** published results data mining/social networks

Future:

- Automated analysis of routine data
- Identification of anomalies
- Initial training of new practitioners
- Data centers: information based on knowledge

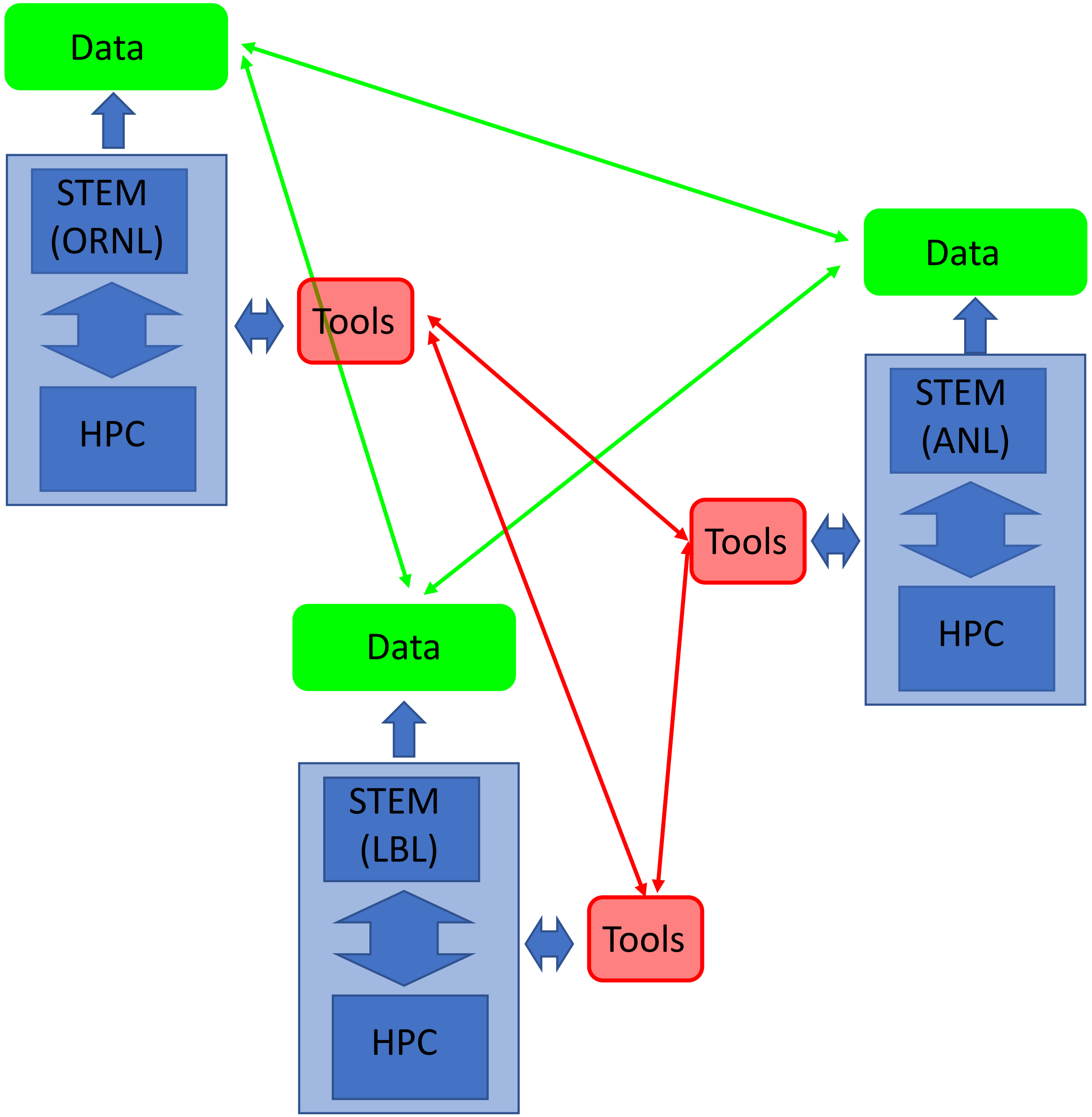
Beyond single tool: workflow design



Traditional experiment:
1. Always based on workflows
2. Ideated, orchestrated, and implemented by humans
3. The "gain of value" during the workflow implementation is uncertain

Value of the step is key element:
- Either based on prior knowledge
- Or defined in a sense of the reinforcement learning Q-function

Need: cloud infrastructure!



Microscope control

- Fast data analytics
- On the fly feedback

Internal data infrastructure

- Full information capture
- Searchability and discoverability
- HPC assisted data analysis workflows

Tool sharing: open source

- PyCroscopy (data structures, image analysis)
- HyperSpy
- LieberSTEM

Data sharing

- CITRIN
- Materials Project

Cloud labs: facilities of the future

Emerald Cloud Lab,
SF and CMU

1. Combined human-machine workflow implementation
2. Computer orchestrating agent
3. How would beyond human workflows be ideated?

Key challenge: reward-based workflow design

1. Development of the labs capable of orchestrating predefined workflows based on human and robotic agents.

2. Workflow design based on AI and human decision making, meaning specific series of synthesis and characterization steps described via executable hyperlanguage.

3. Defining domain-specific reward functions. Why are we running experiments? Ultimately, we need to quantify (in the style of Bell's equation) what is the benefit of the specific step in the workflow, and how does it accomplish or affects exploration and exploitation goals.

4. Integration of reward functions from dissimilar domains. For example, how does better microscope help us learn physics of specific material? Why would the specific DFT calculation help us understand experimental data?

5. Creating experimentally falsifiable hypothesis from the domain specific body of knowledge that can be incorporated in the exploratory part of automated workflows.

6. Hypothesis generation beyond human. I suspect that this part gets close to AGI and is non-solvable in general, since it is extrapolation/intuition. Makes it interesting for ML!

And technology

“New directions in science are launched by new tools much more often than by new concepts. The effect of a concept-driven revolution is to explain old things in new ways. The effect of a tool-driven revolution is to discover new things that have to be explained.”

Freeman Dyson

- Cloud connection and data infrastructure is necessary, but not enough
- Can we offer comprehensive workflow design, building, and optimization solutions?
- Potential to change way R&D is performed: academia, start-ups, industry

Concluding:

- **Machine learning is great, but**
  - Requires domain expertise
  - Ease of use for deployment
  - Some ML knowledge

- **Microscope is a laboratory:**
  - Engineering controls
  - Algorithms
  - Connection to domain expertise

Sergei V. Kalinin
sergei2@utk.edu

Connect!

YouTube

Twitter: Sergei_imaging