THE UNIVERSITY of TENNESSEE KNOXVILLE

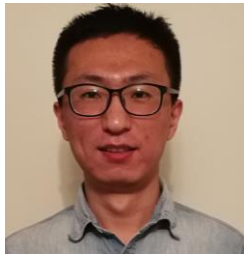
Automated Experiment in Materials Synthesis and Characterization: from Optimization to Reward-Driven Workflows

Sergei V. Kalinin

University of Tennessee, Knoxville and
Pacific Northwest National Laboratory

Maxim Ziatdinov

Mahshid Ahmadi



Yongtao Liu

Rama Vasudevan

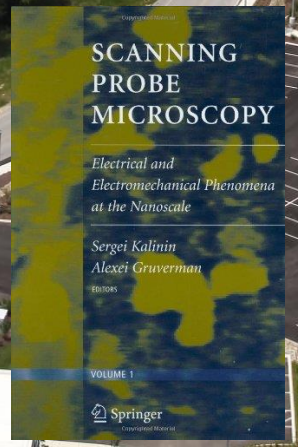
Kevin Roccapriore

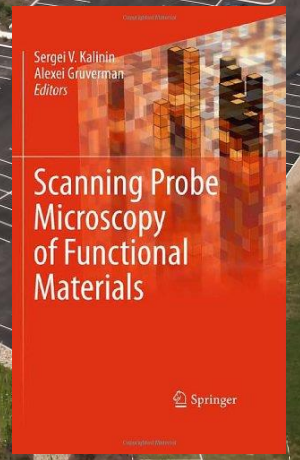
2002 - 2022

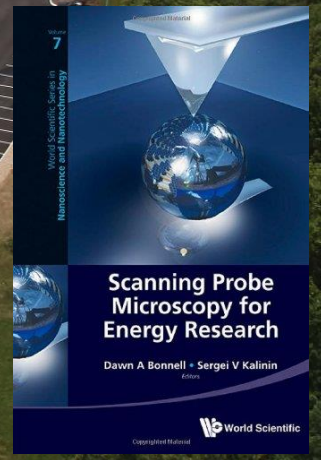
Since 2022

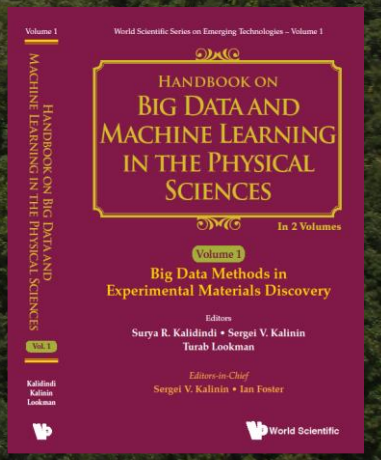
2022 -2023

amazon







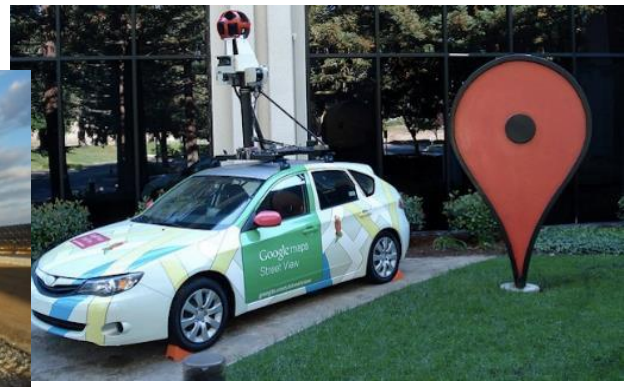


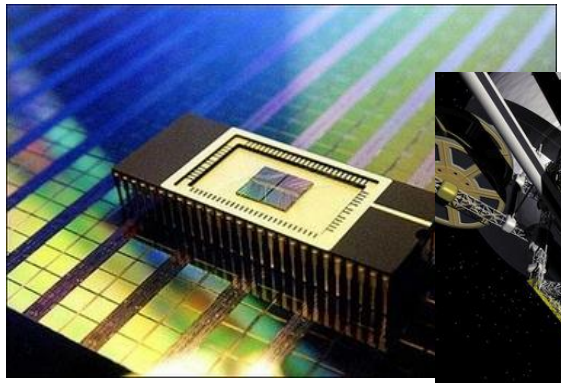
This course

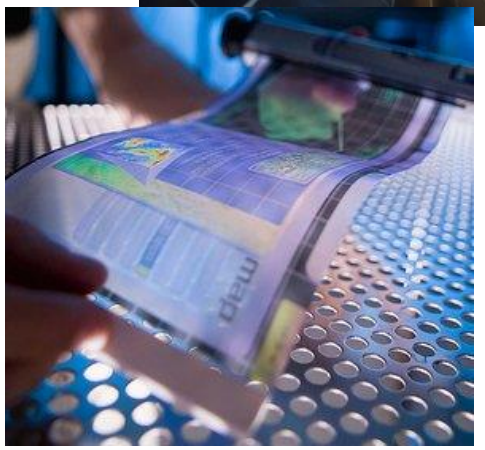
1. Introduction: rewards, policies, and stochastic optimization
    a. Current opportunities for automated experiment in synthesis and characterization: from single tools to cloud laboratories
    b. Objectives and rewards for automated experiment
    c. Reward driven workflow design, orchestration, and execution
    d. Bayesian world view: priors, likelihood, evidence, and posteriors

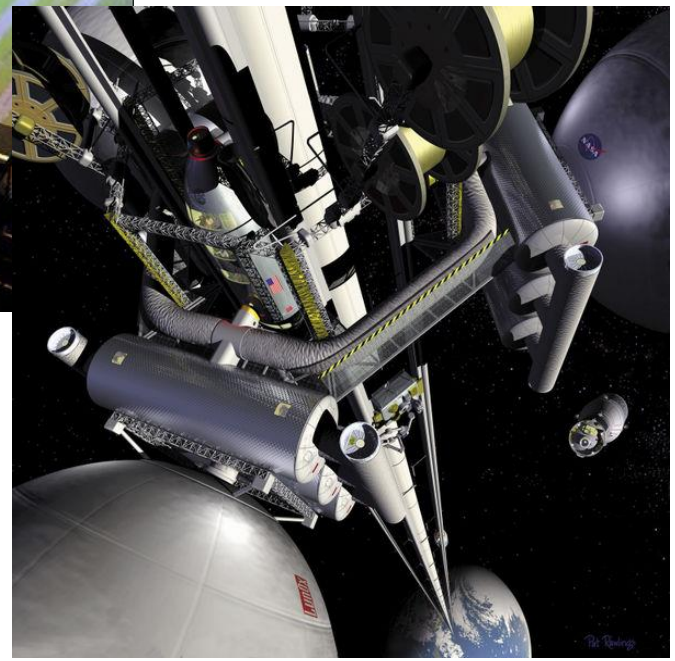
2. Gaussian processes and Bayesian Optimization: from concept to real world
3. Bayesian Inference and structured Gaussian Processes
4. Hypothesis learning
5. Manifold learning, variational autoencoders, and encoders-decoders,
6. Deep Kernel Learning and structure-property discovery
7. DKL, explainable automated experiments, and human in the loop AE
8. Multifidelity structured GP for co-orchestration of multiple tools
9. Future perspectives

The World is a Material Opportunity









Predicting crystal structure by merging data mining with quantum mechanics

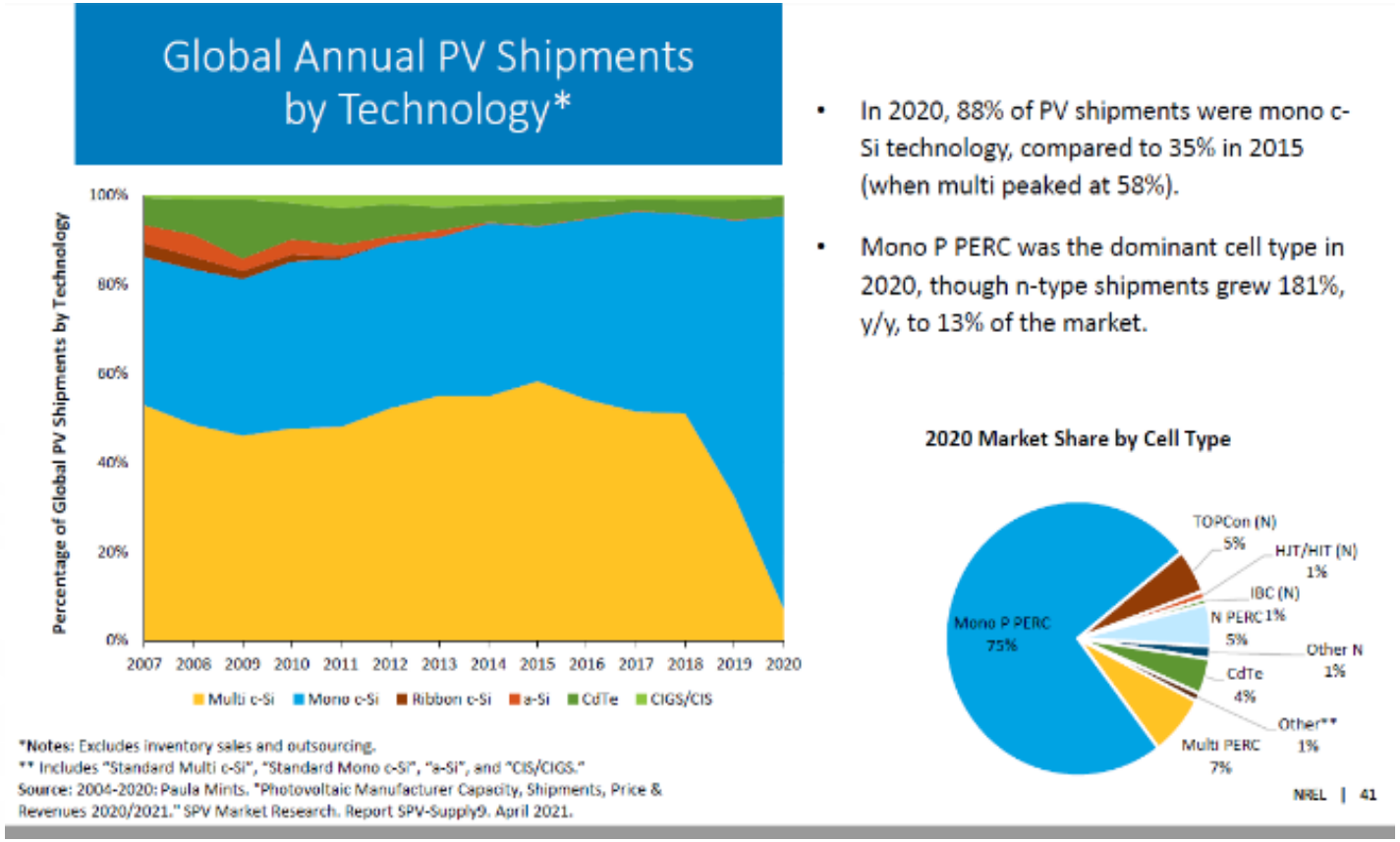
CHRISTOPHER C. FISCHER[1], KEVIN J. TIBBETTS[1], DANE MORGAN[2] AND GERBRAND CEDER[1]*
[1]Department of Materials Science and Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA
[2]Department of Materials Science and Engineering, University of Wisconsin, Madison, Wisconsin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

- **"Improve":** Renewable energy, structural materials
- **"Discover":** RT superconductivity, high mechanical stress materials
- **"Engineer":** Quantum computing, single-atom catalysts, biomolecules

Functionality, manufacturability, cost

The theory can only get you so far!

Solar Energy: Will Silicon Ever Reign?



- In 2020, 88% of PV shipments were mono c-Si technology, compared to 35% in 2015 (when multi peaked at 58%).
- Mono P PERC was the dominant cell type in 2020, though n-type shipments grew 181%, y/y, to 13% of the market.

NREL | 41

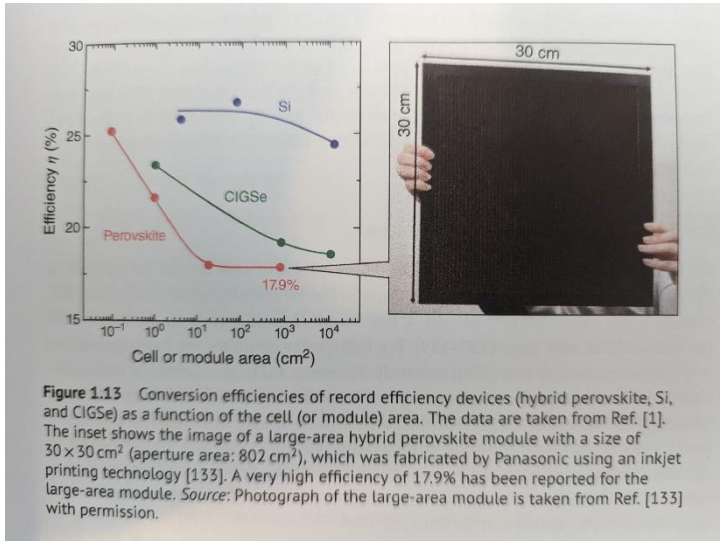
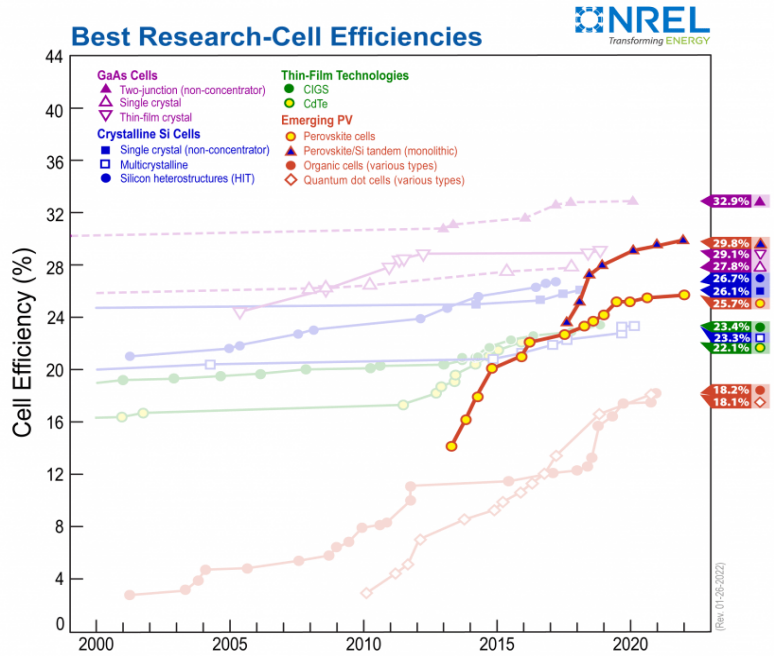
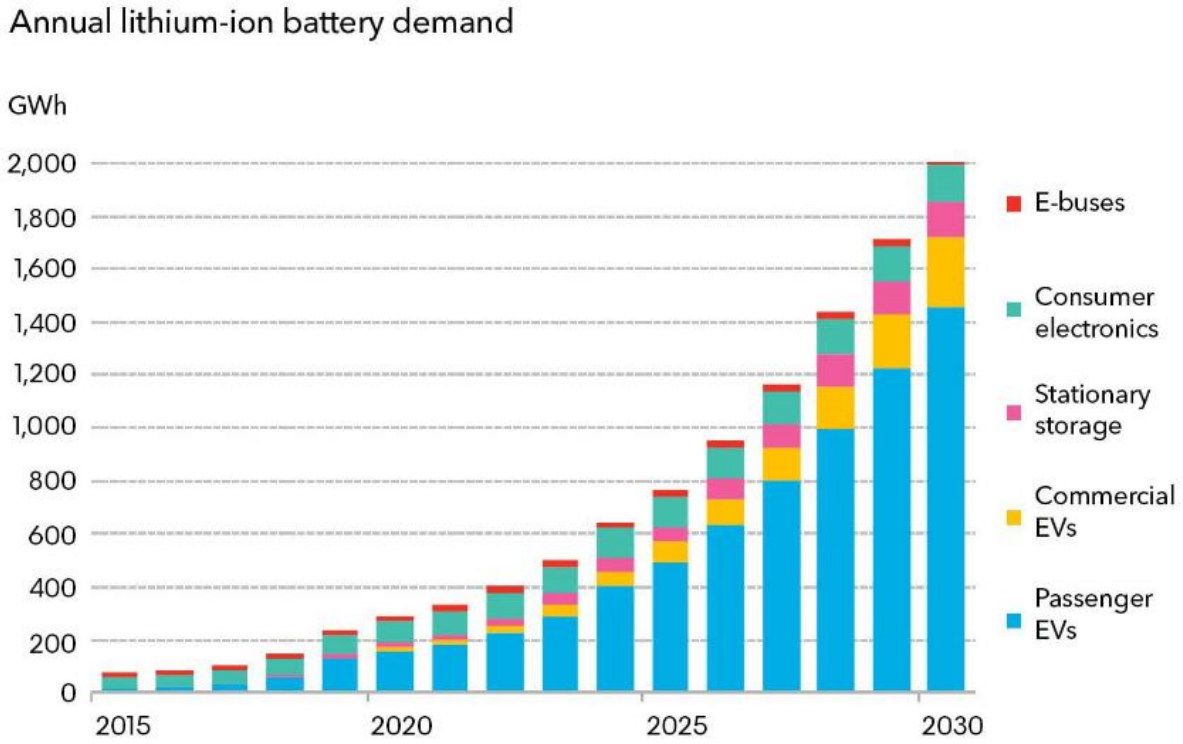


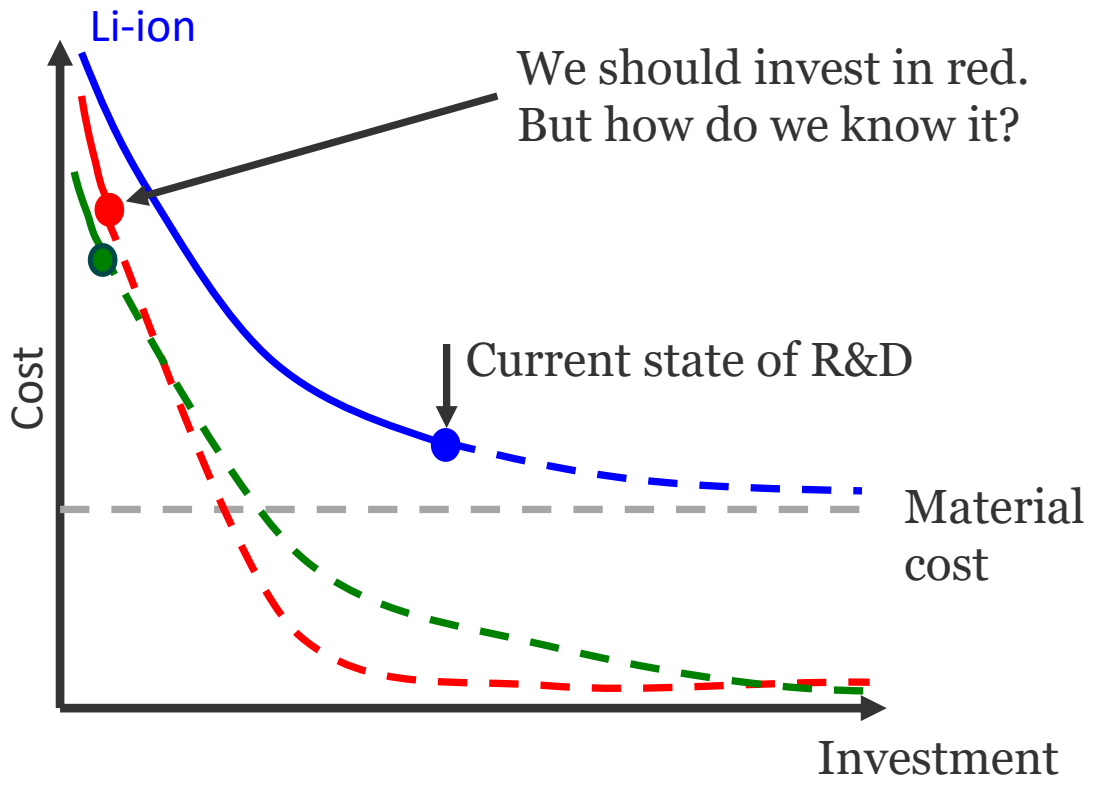
Figure 1.13 Conversion efficiencies of record efficiency devices (hybrid perovskite, Si, and CIGSe) as a function of the cell (or module) area. The data are taken from Ref. [1]. The inset shows the image of a large-area hybrid perovskite module with a size of $30 \times 30\ cm^2$ (aperture area: $802\ cm^2$), which was fabricated by Panasonic using an inkjet printing technology [133]. A very high efficiency of 17.9% has been reported for the large-area module. *Source*: Photograph of the large-area module is taken from Ref. [133] with permission.



- 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!

Batteries: Li-ion and Beyond

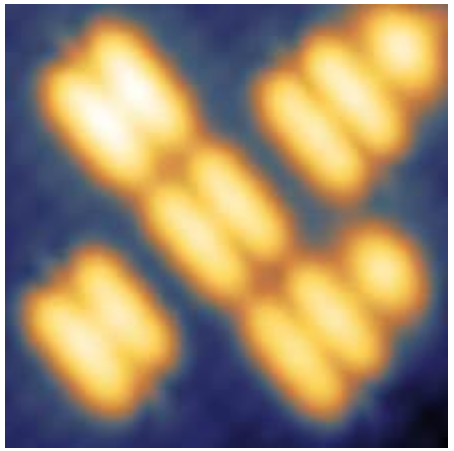




- Batteries are required element of energy transition (EVs, ESS, mobile devices)
- Currently Li-ion is the primary technology
- Optimization of Li-ion batteries takes years (even with same process on new Gigafactory)
- However, it is far from Goldilock zone for ESS or energy transport

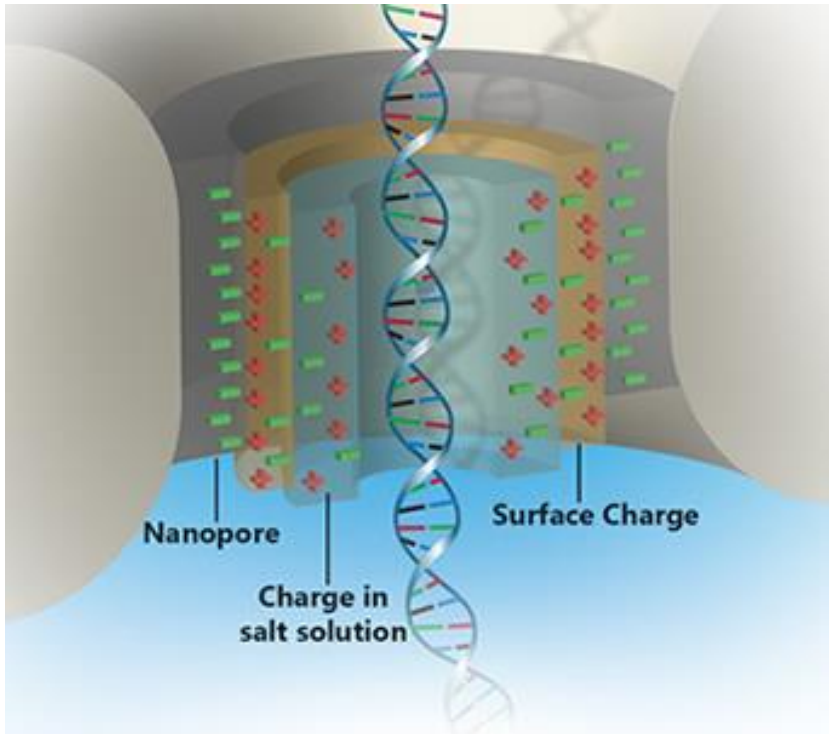
- How can we optimize usage and safety for Li-ion batteries in EVs?
- How do we select beyond Li technologies for ESS?

Quantum Computing and Single Molecule Bio



NIST

Zyvex Labs

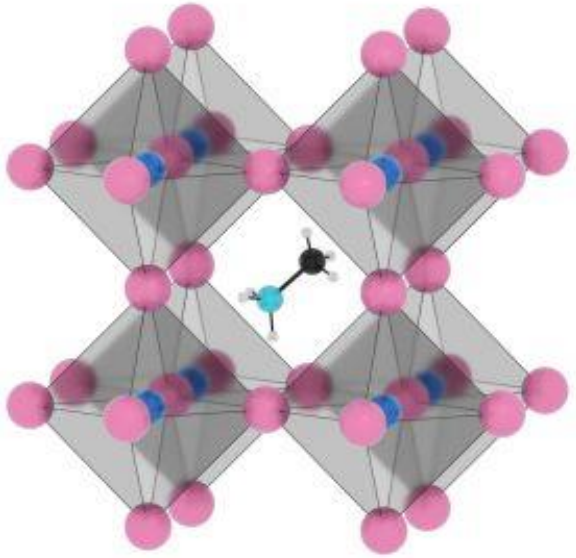


Oxford Nanopore

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

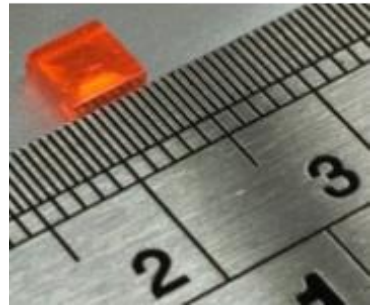
Metal Halide Perovskites

ABX_3

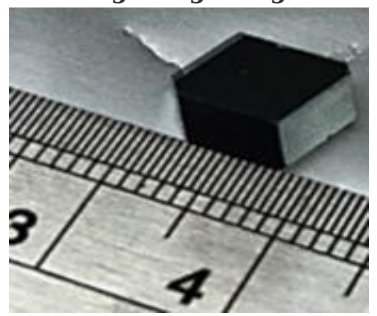


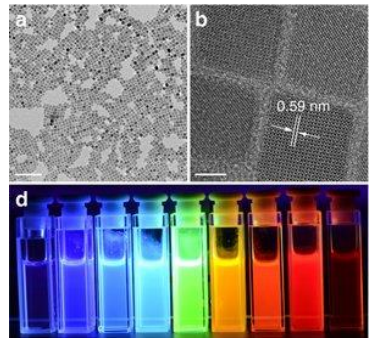
Solution processable thin films and single crystal growth

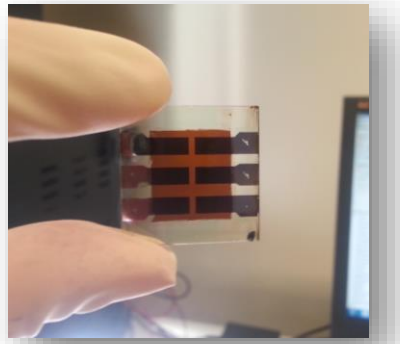
$CH_3NH_3PbBr_3$



$CH_3NH_3PbI_3$

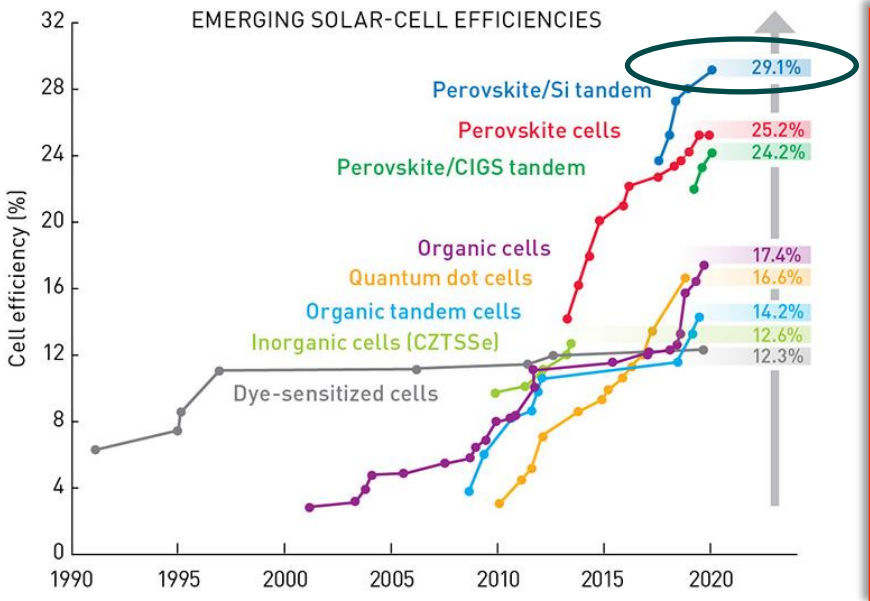






Properties

- ✓ 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

… there are a lot of them

A

13 protonated amines, A^+

$[NH_4]^+$
$[NH_3OH]^+$
$[CH_3NH_3]^+$
$[(CH_2)_3NH_2]^+$
$[CH(NH_2)_2]^+$
$[C_3N_2H_5]^+$
$[(CH_3)_2NH_2]^+$
$[NC_4H_8]^+$
$[CH_3CH_2NH_3]^+$
$[(NH_2)_3C]^+$
$[(CH_3)_4N]^+$
$[HN(CH_3)_3S]^+$
$[C_7H_7]^+$

4 alkali metals

Cs^+
Rb^+
K^+
Na^+

B

22 divalent metals, B^{2+}

Be^{2+} Cu^{2+}
Mg^{2+} Zn^{2+}
Ca^{2+} Cd^{2+}
Sr^{2+} Hg^{2+}
Ba^{2+} Ge^{2+}
Mn^{2+} Sn^{2+}
Fe^{2+} Pb^{2+}
Co^{2+} Eu^{2+}
Ni^{2+} Tm^{2+}
Pd^{2+} Yb^{2+}
Pt^{2+} V^{2+}

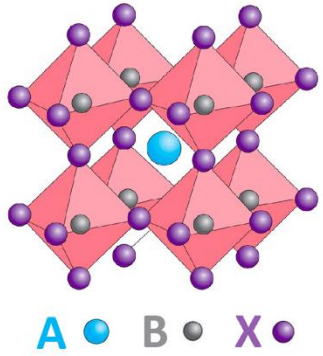
9 B^I or B^{III} metals

Bi^{3+} Au^+
Eu^{3+} Au^{3+}
Sb^{3+} In^{3+}
Ag^+
Na^+
Cu^+

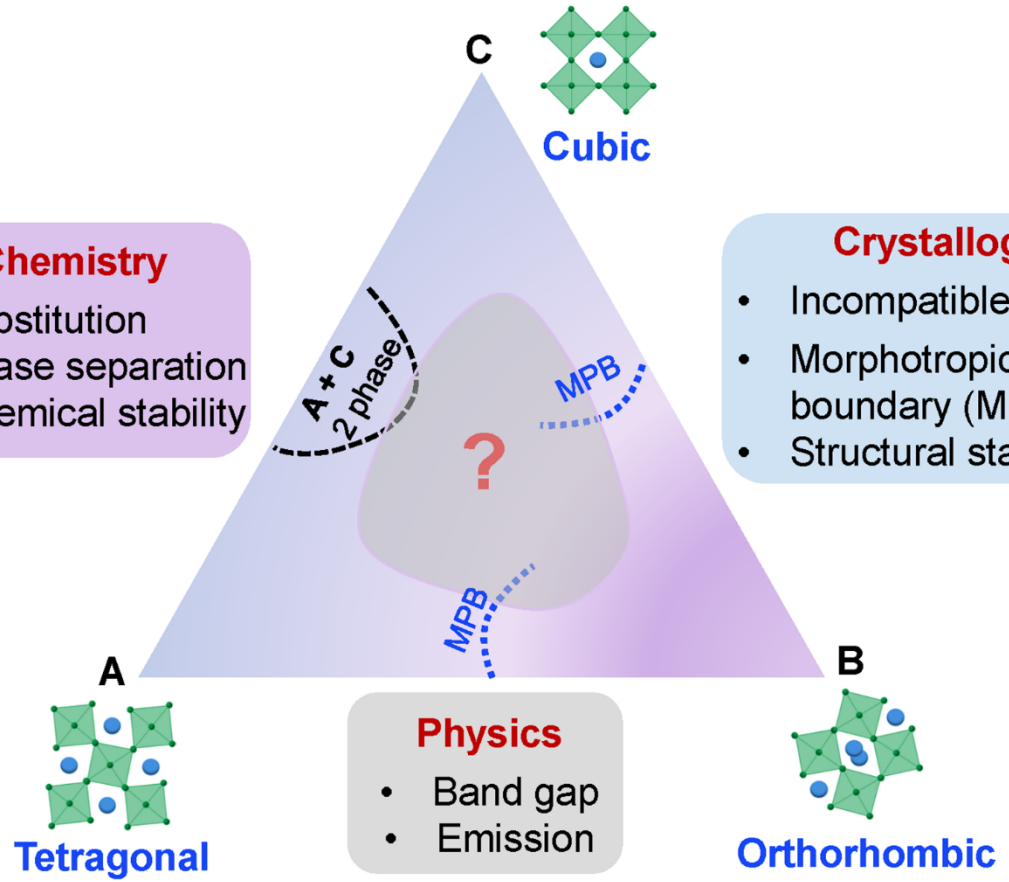
X

9 anionic species, X^-

F^-
Cl^-
Br^-
I^-
$HCOO^-$
CN^-
N_3^-
BH_4^-
SCN^-



Chemistry
- Substitution
- Phase separation
- Chemical stability

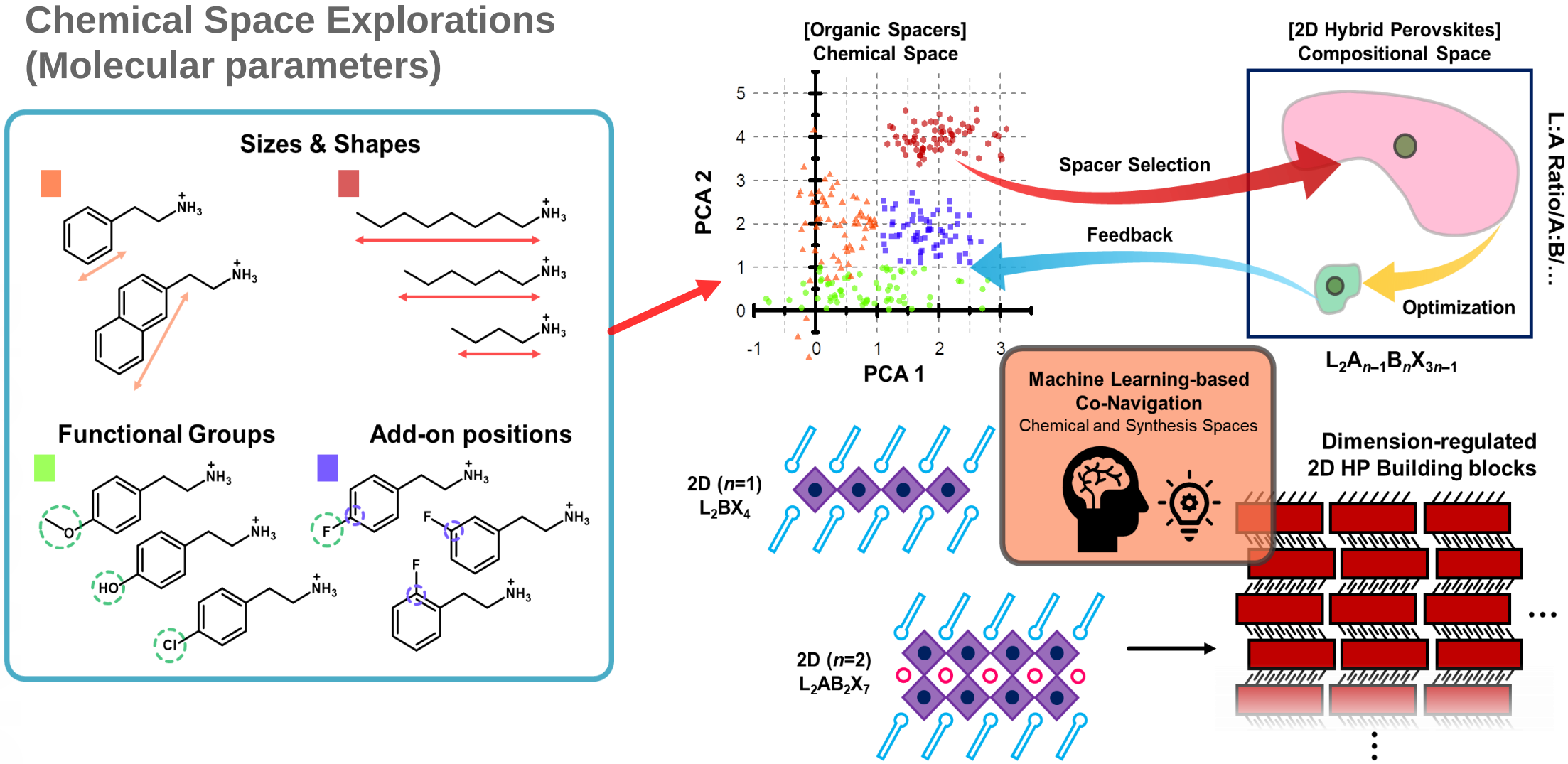


Crystallography
- Incompatible symmetry
- Morphotropic phase boundary (MPB)
- Structural stability

Physics
- Band gap
- Emission

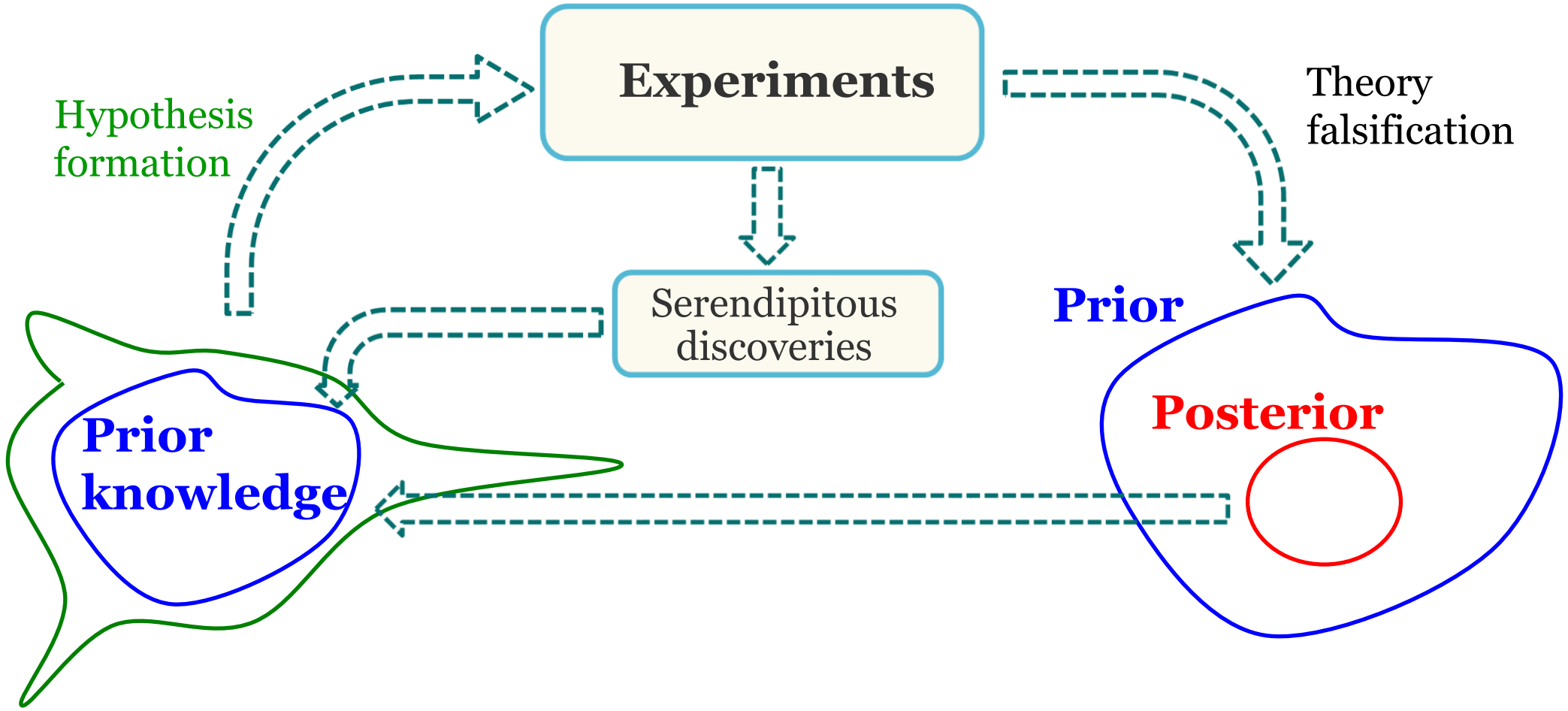
All these parameters are complex function of composition, and are virtually impossible to predict theoretically or from knowledge of binary phase diagrams

Going 2D: Co-Navigating Chemical spaces



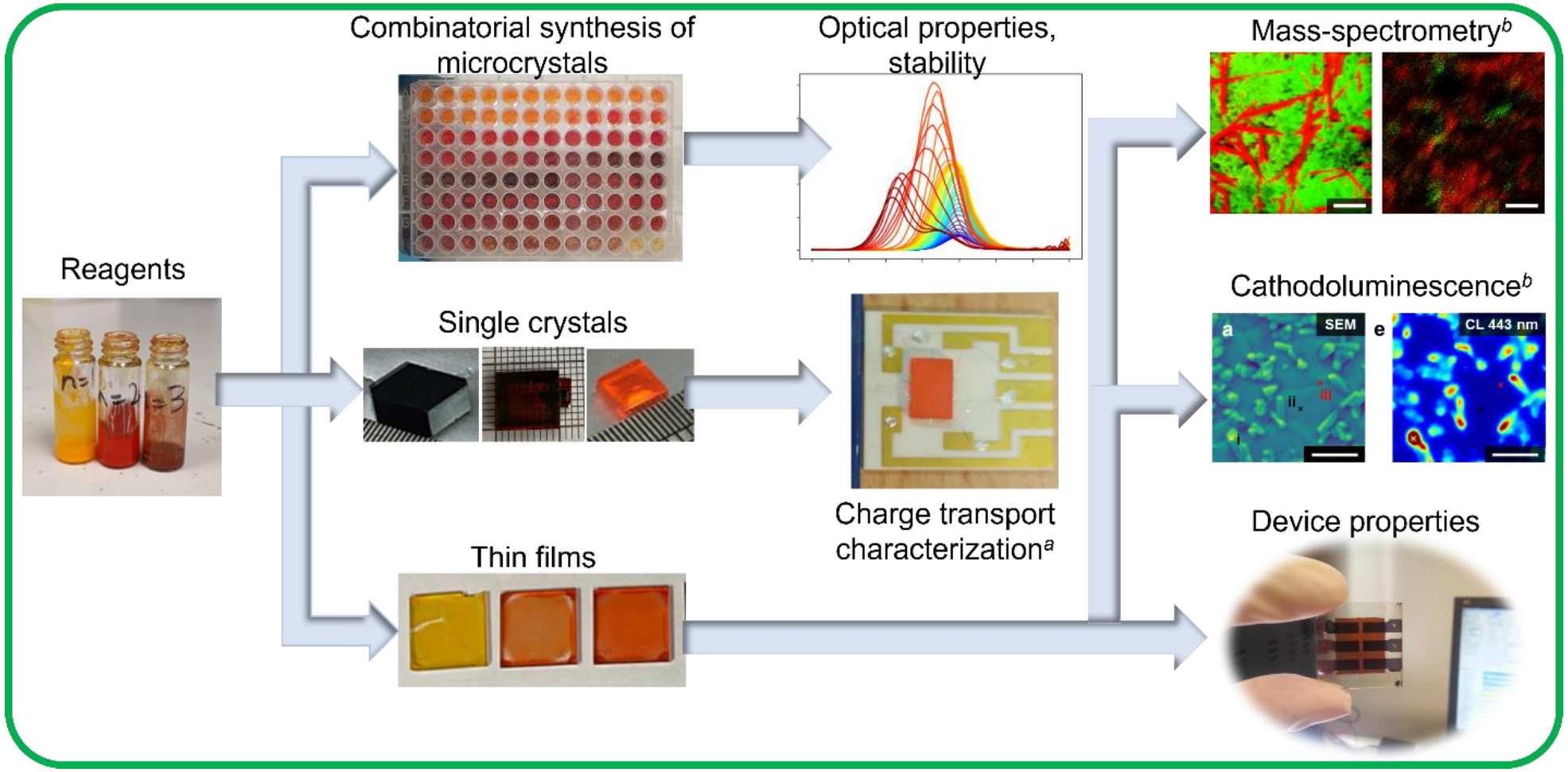
J. Yang *et. al. MRS Bull.* **2023,** 48, 164-172

How does science work



Rewards: Why are we doing science/R&D?
Policies: Exploration-exploitation balance
Instrument development: New tools create new opportunities
Hypothesis making: Extrapolation into the unknown

What is a workflow?



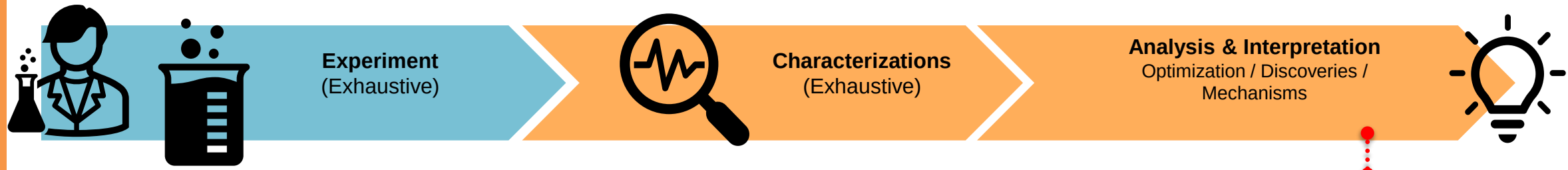
- **Workflow: ideation, orchestration, implementation**
- Domain specific language
- Dynamic planning: latencies and costs
- Reward and value functions

Designed in academia and adopted by industry

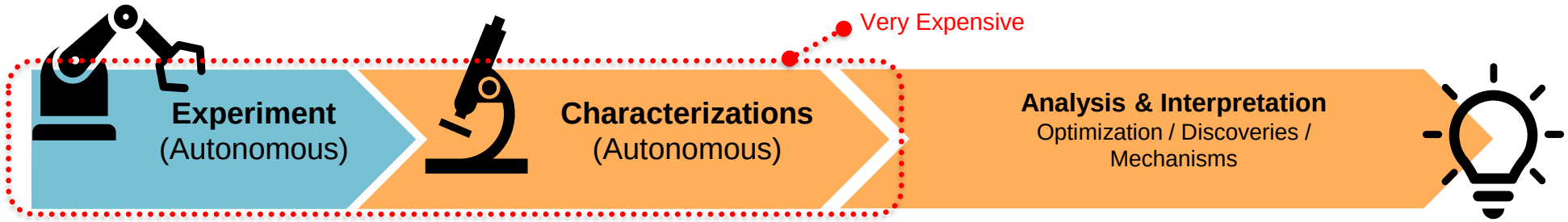
- Are they optimal?
- Can we design them better?
- Can they be changed dynamically?

Implementation of research workflows

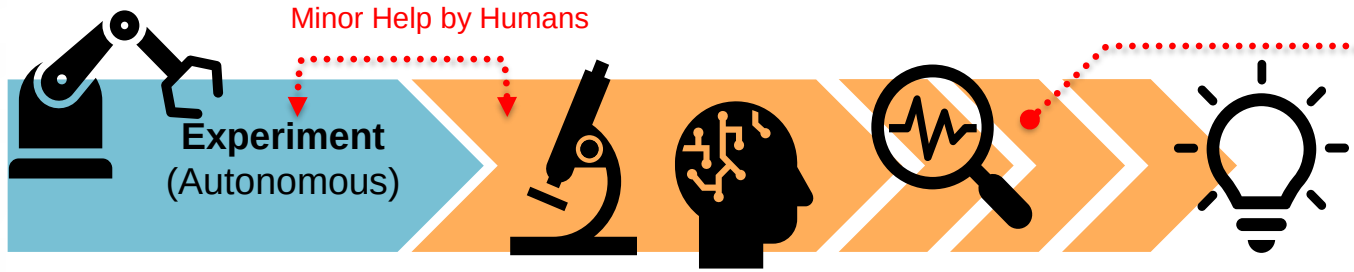
a. Classical Manual High-Throughput Research (Slow)



b. Fully-Automated Self-Driving Lab (Intermediate)



c. Automated Experiment-Characterization-Analysis Workflow (Fast)



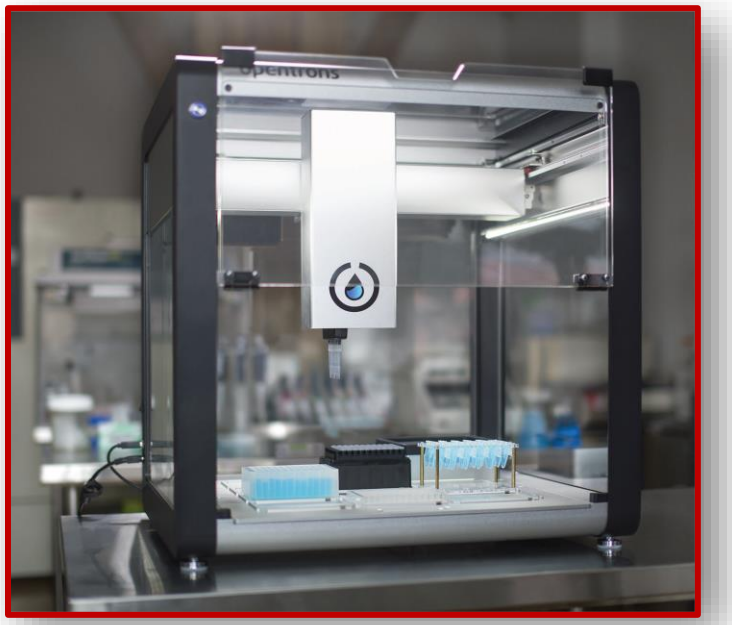
+ Integrated with Machine Learning (ML) / Automated Data Analysis

→ Release the bottleneck

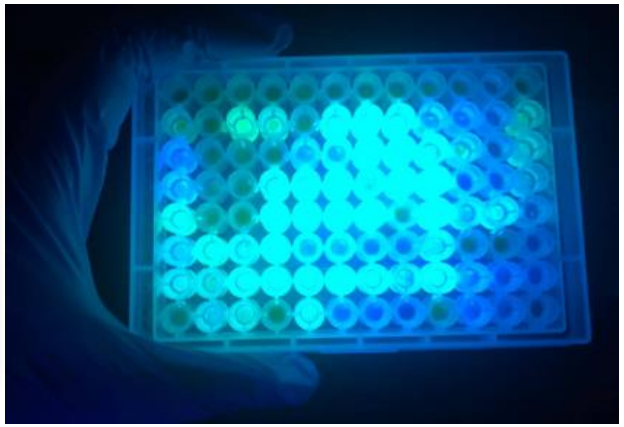
→ Accelerates and boosts research throughput

Laboratory robotics – microcrystals

Micropipette Robot

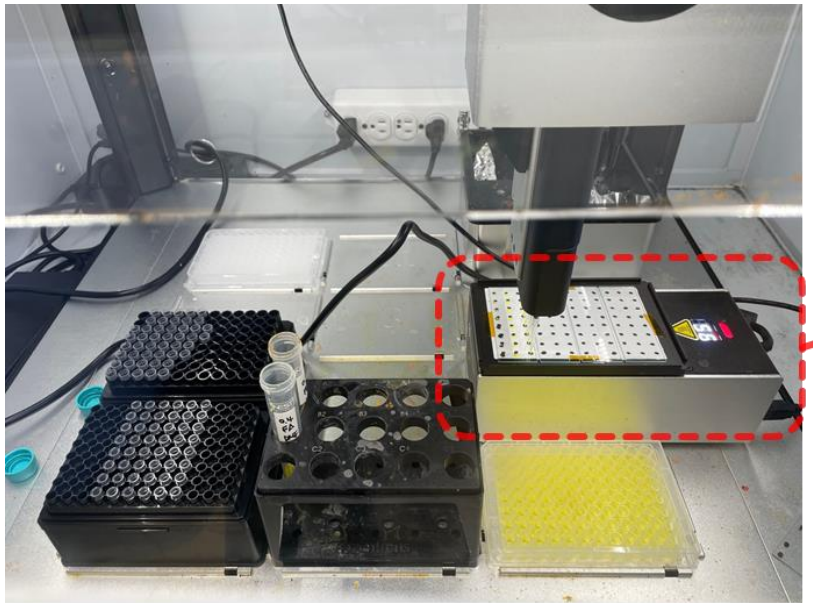


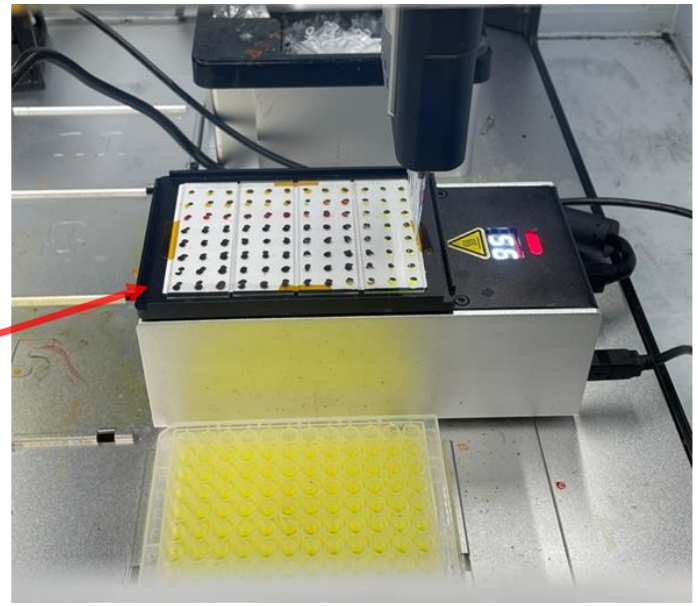
Heating module

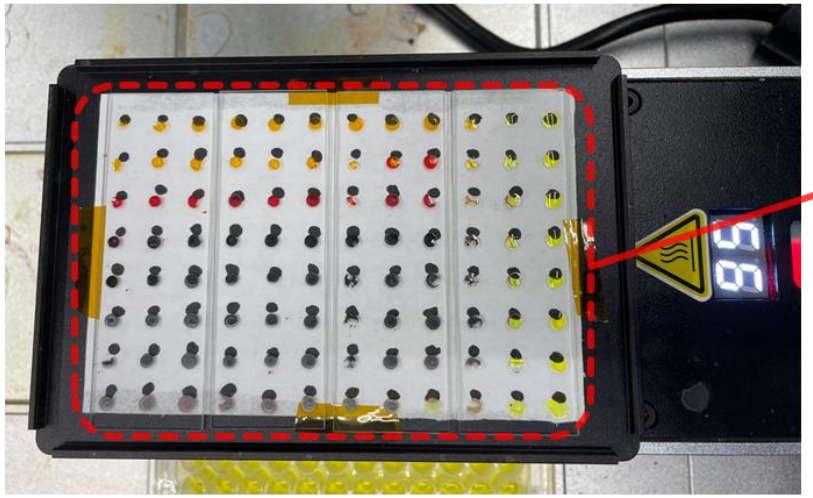


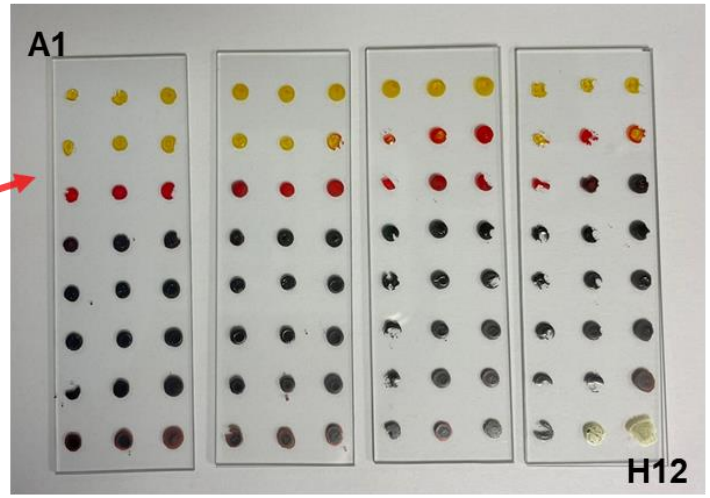
Multi-Mode Spectrometer

Laboratory robotics – drop casting

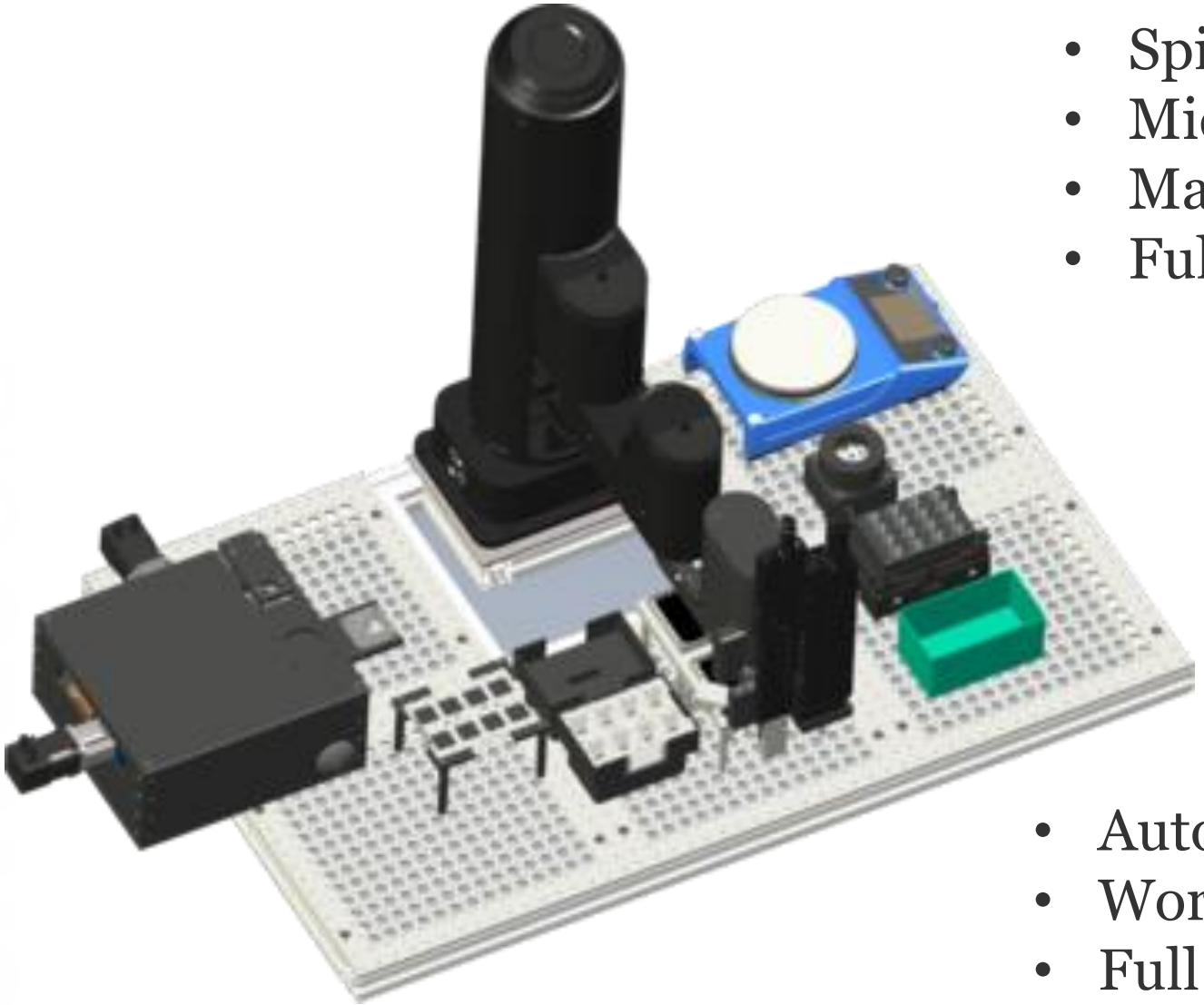








… and there is more!



- Spin coating
- Microfluidics
- Manipulation
- Full chemical robotics solutions

- Automated single operation tools
- Workflow step implementations
- Full solutions

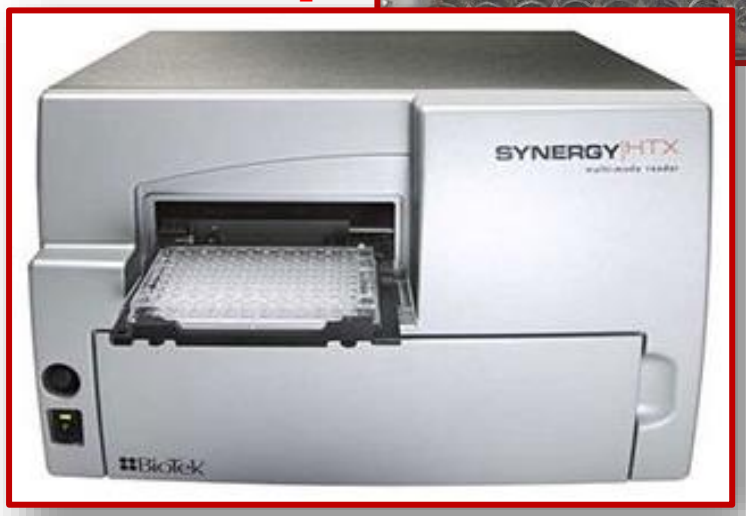
Cloud Labs: Facilities of the Future

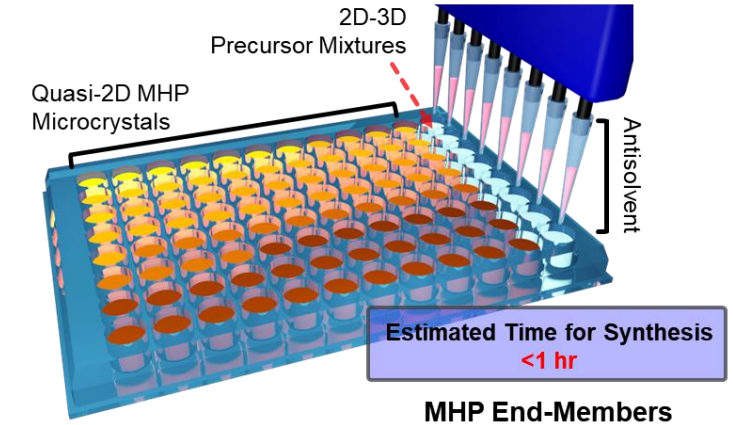
Emerald Cloud Lab, TX and CMU

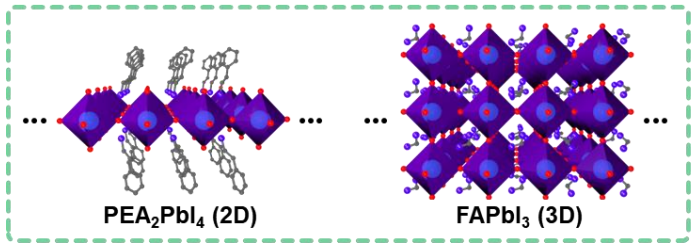
1. Combined human-machine workflow implementation
2. Computer orchestrating agent
3. How would beyond human workflows be ideated?

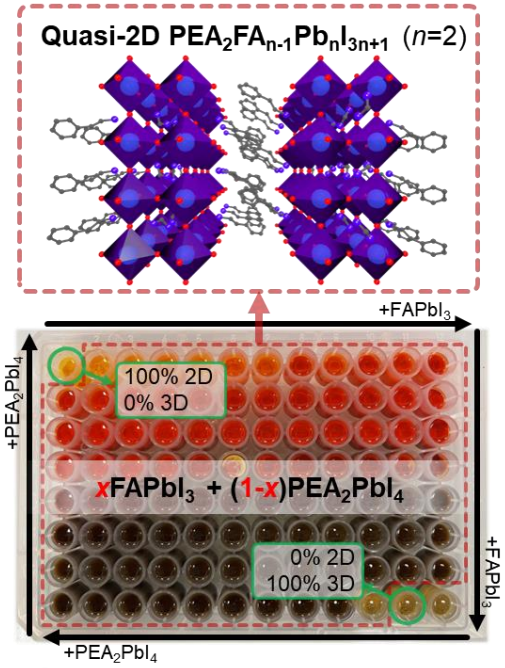
Enter automated synthesis (2019)

Multi-Mode Spectrometer

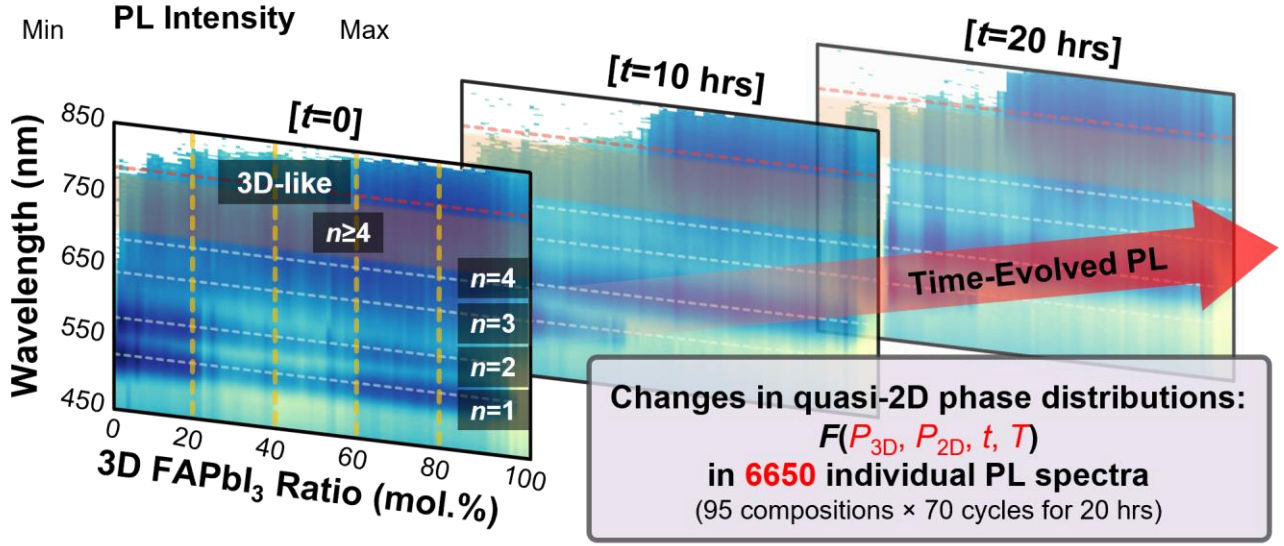








Phase Growth Behaviors



Changes in quasi-2D phase distributions:
$F(P_{3D}, P_{2D}, t, T)$
in 6650 individual PL spectra
(95 compositions × 70 cycles for 20 hrs)

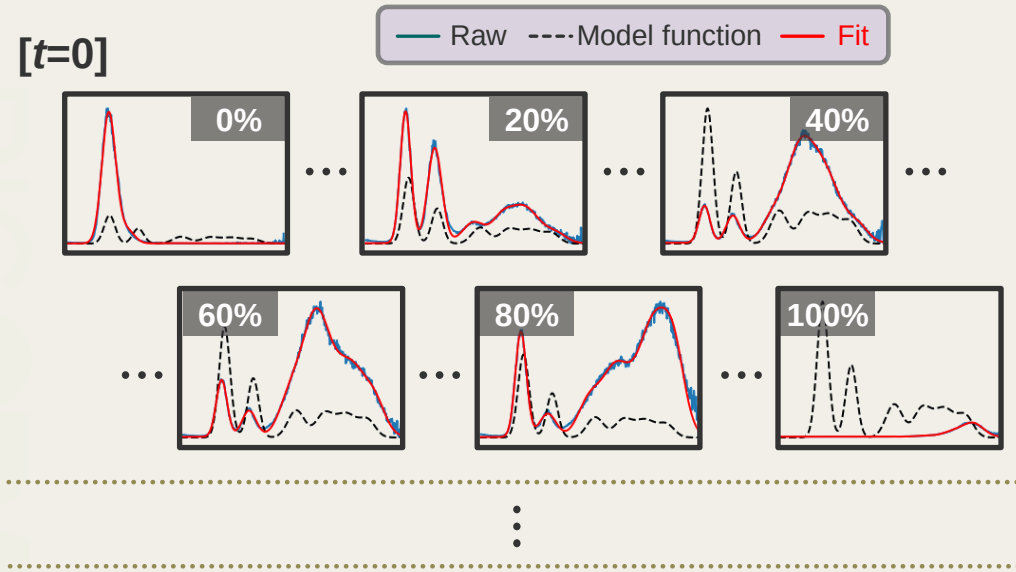
J. Yang *et. al. ChemRxiv* 10.26434/chemrxiv-2023-zcvl0-v2

Analysis workflows: from static to cloud

Automated PL Spectrum Analysis

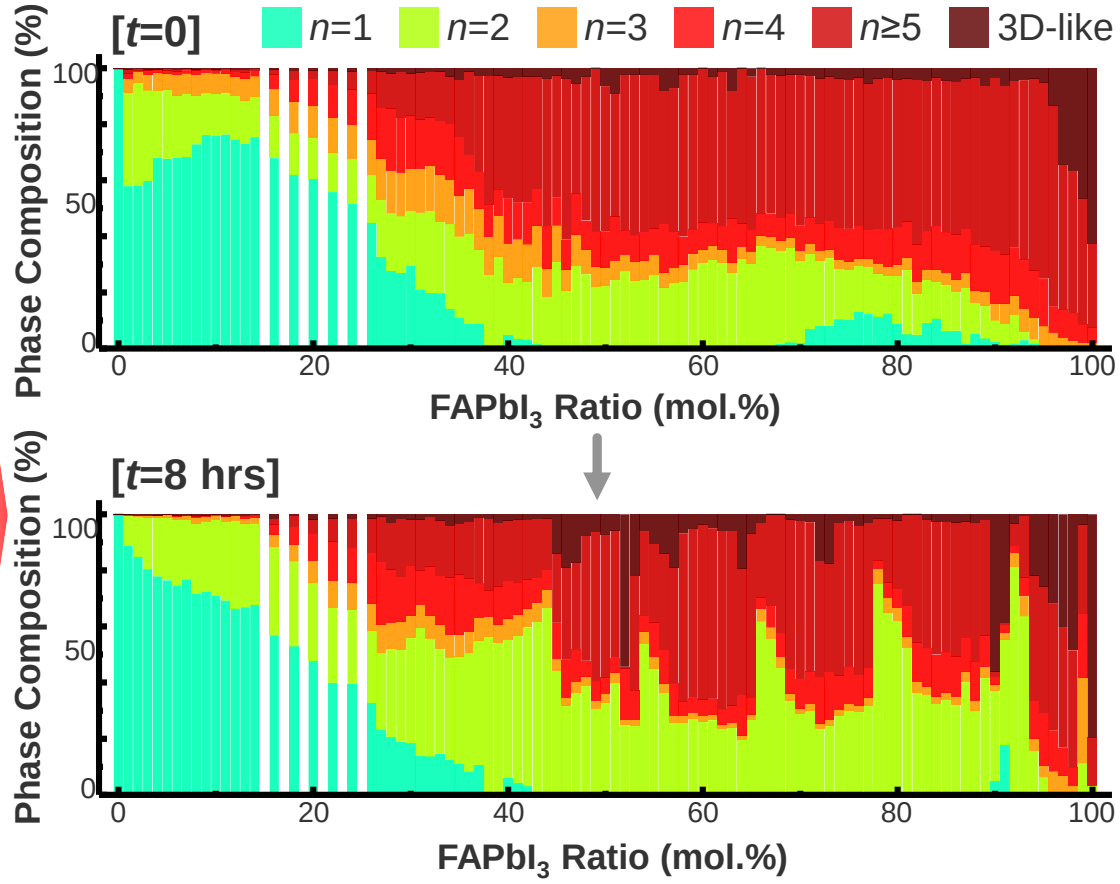
> $Fit = G1_{n=1} + G2_{n=2} + G3_{n=3} + G4_{n=4} + G5_{n\geq5} + G6_{3D}$

(Preset function: each peak defined from known physics)



Automated fitting to
6650 PL spectra
Total time duration: 5 **hrs**

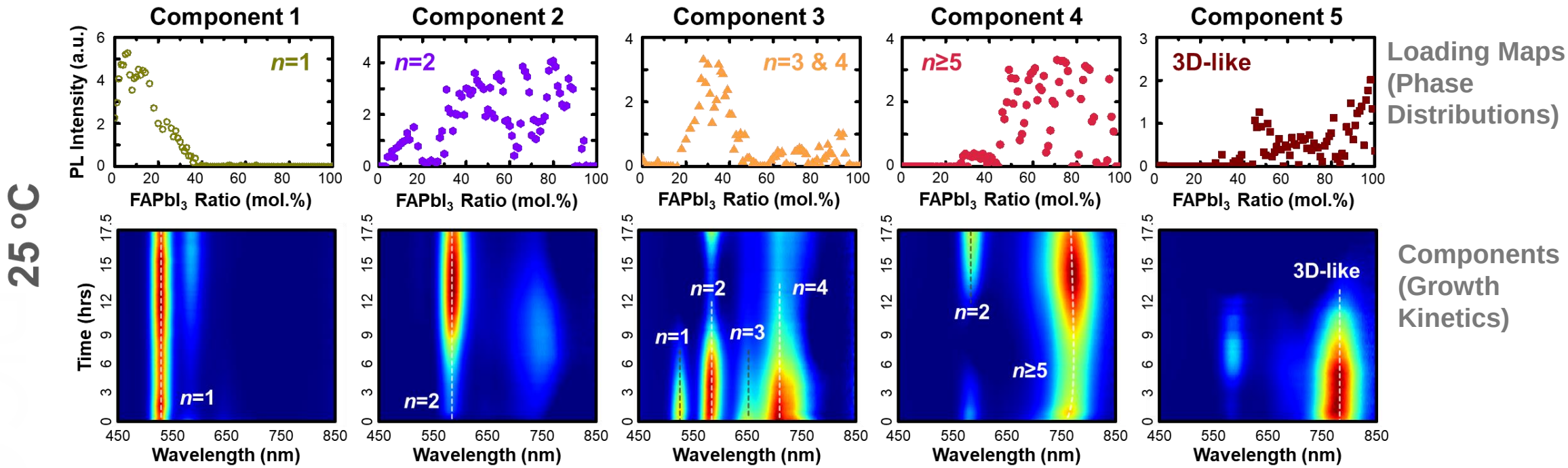
Phase Emergence in the Compositional Space



Visualizing Changes in Quasi-2D MHP Phase Distributions

J. Yang *et. al. ChemRxiv* 10.26434/chemrxiv-2023-zcvl0-v2

Sometimes simple ML is the best ML



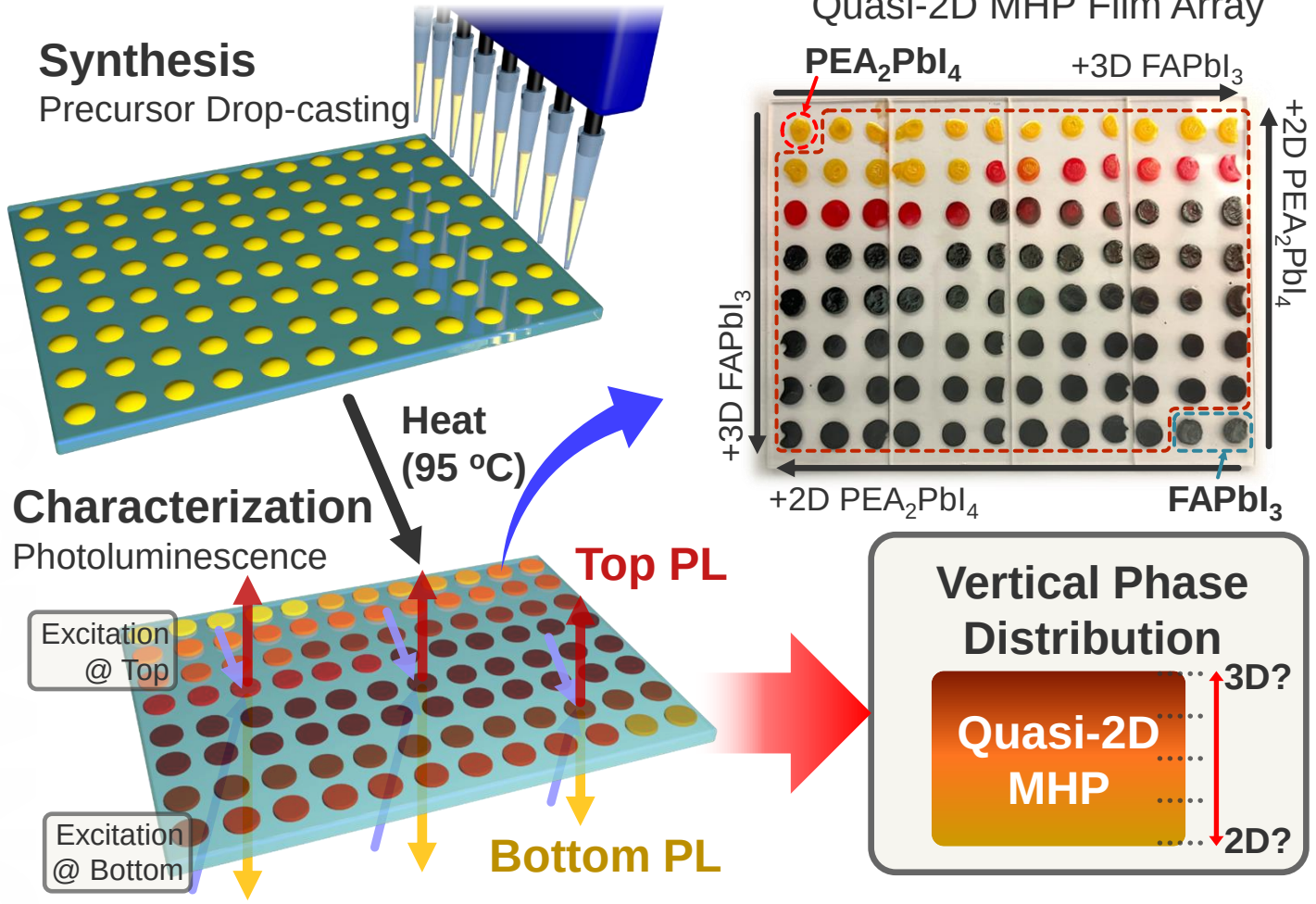
Without any prior knowledge, unsupervised ML can effectively catch the global PL features of the complex materials

Reveal a general trend of growth kinetics and phase distributions across the compositional space

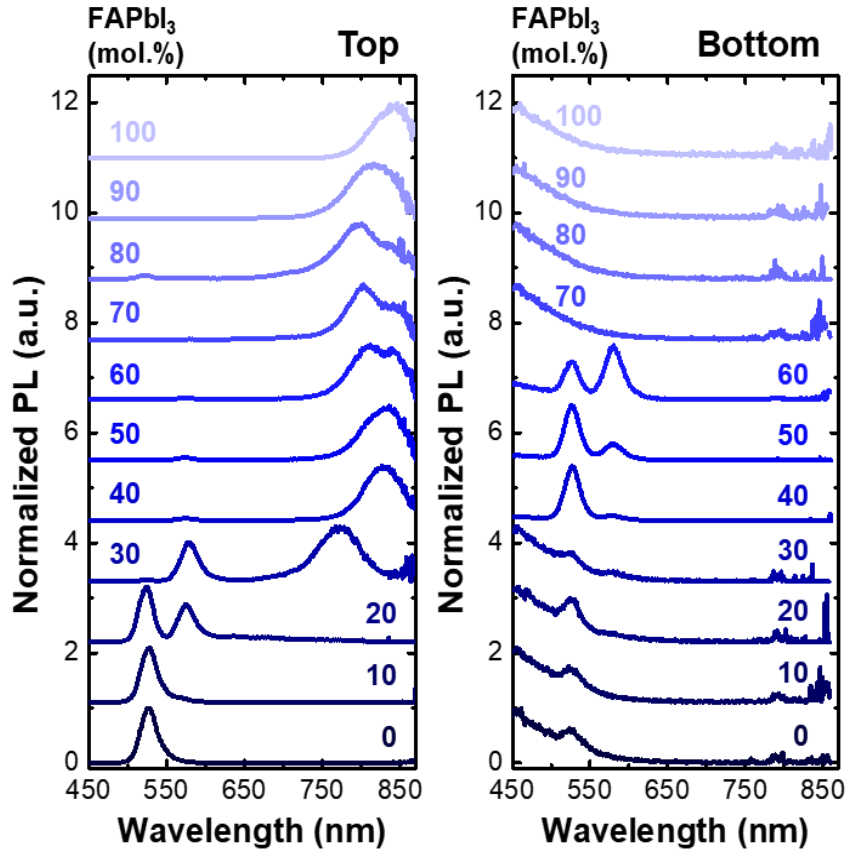
J. Yang *et. al. ChemRxiv* 10.26434/chemrxiv-2023-zcvl0-v2

Up the manufacturing chain

High-Throughput Film Fabrication (Suited form to device applications)



Representative PL



9,880 individual spectra in total

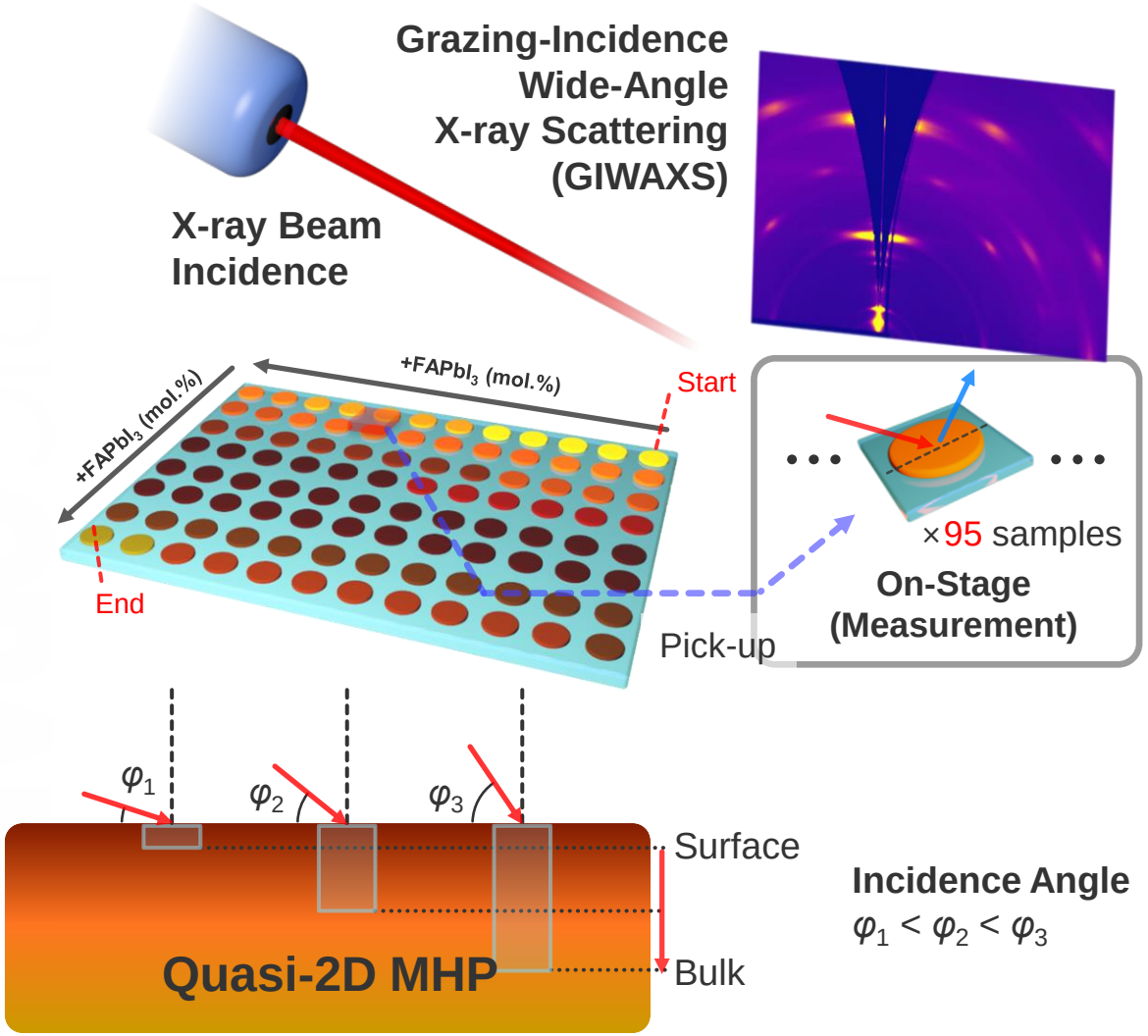
95 compositions × 52 cycles (for 26 hrs)
× 2 (top and bottom)

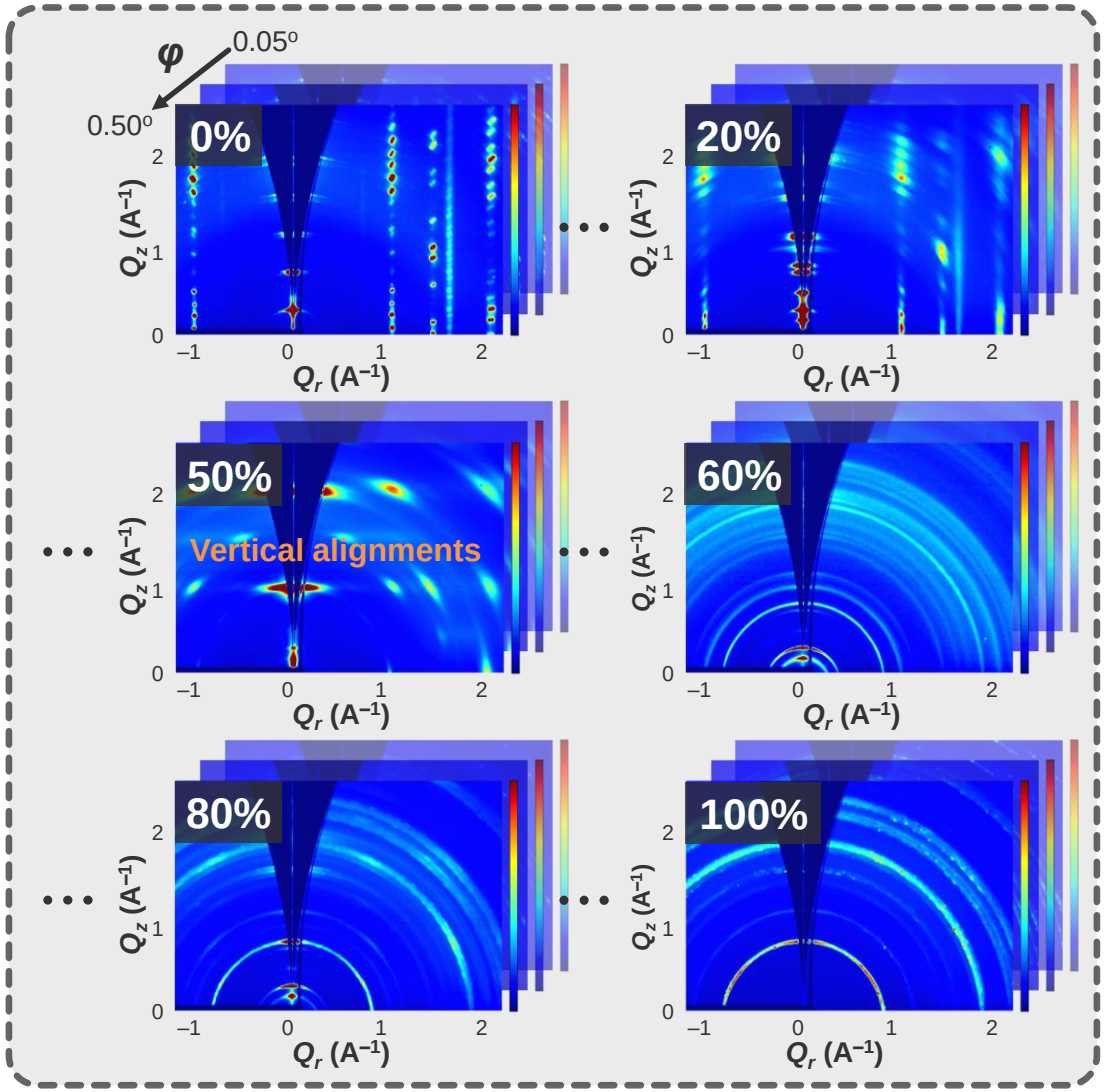
J. Yang *et. al. ChemRxiv* 10.26434/chemrxiv-2023-zcvl0-v2

Synthesis, properties, structure

High-Throughput Grazing-Incidence Wide-Angle X-ray Scattering (HT-GIWAXS)

Collaboration with J. P. Correa-Baena (GATECH)

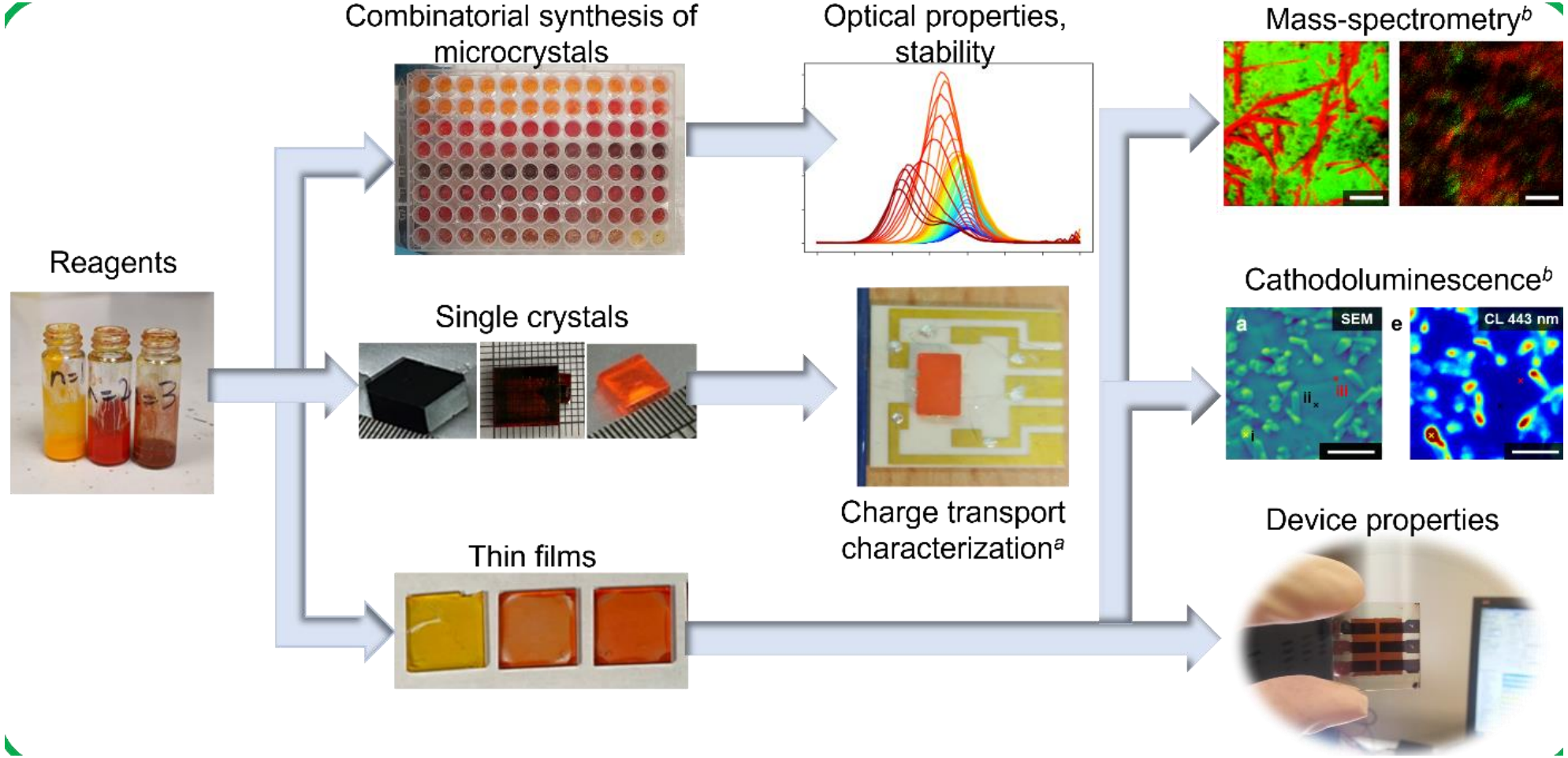




Four eras of automated labs

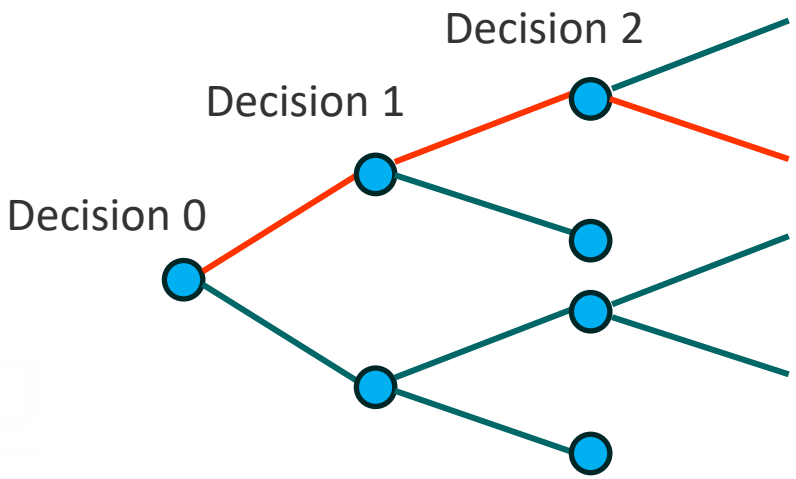
- **Before 2015:** Small number of enthusiasts in academic community, industry efforts
- **2015 - 2020:** Engineering developments
- **2020 – 2025:** Broad adoption of the commercial tools, modification of commercial platforms
- **2025 :** Workflows design?

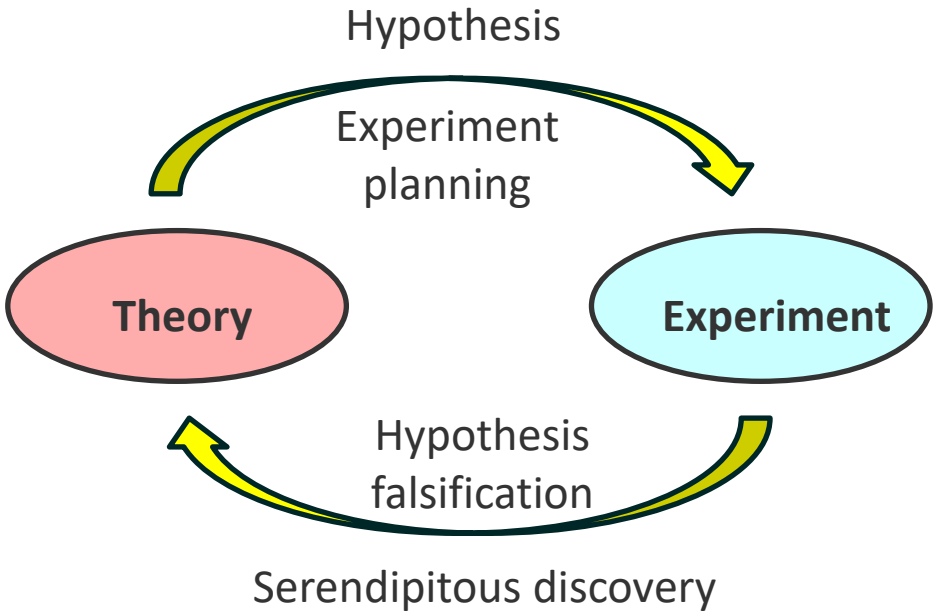
Decision science of workflows



- Multiple levels of decision making based on perceived gain, latencies, and costs
- Iterative cycles between low-cost and expensive measurements
- Learning **basic science/models** as a strategy to minimize cost and answer interventional and counterfactual questions

Decision science of workflows

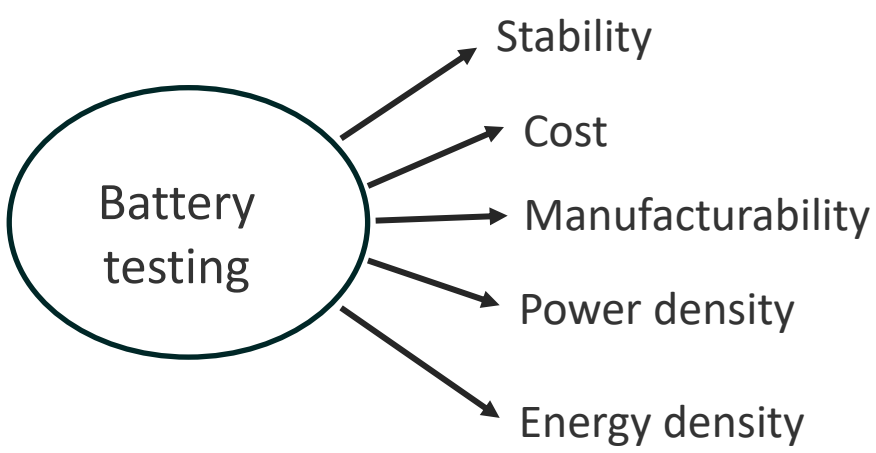


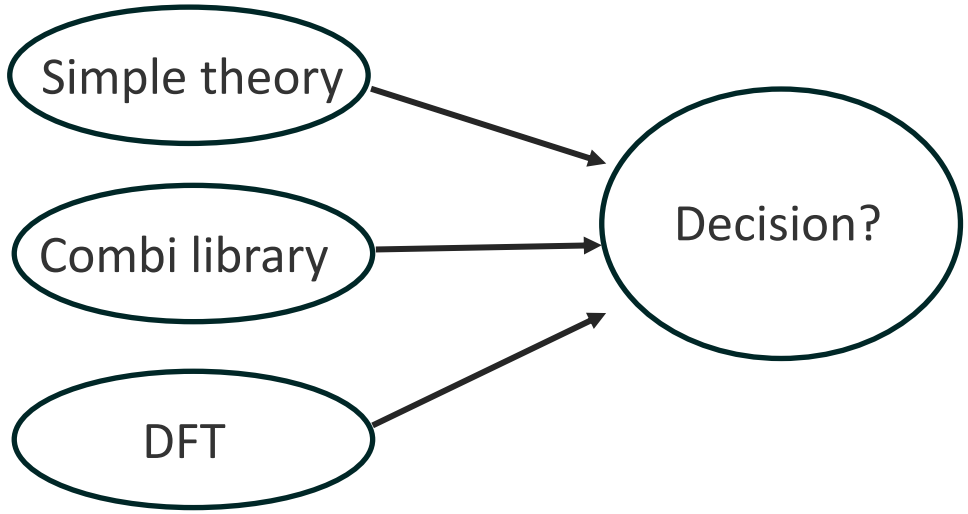


- **Experiment is a combinatorial space of opportunities:**
  → Investing only in scaling of throughput is only a linear improvement
- **Knowledge of physics often allows to reduce complexity: combinatorial to linear**:
  → Basic science pays off (with time)!
- **Science is a cycle between theory-driven hypothesis generation and experiment**:
  → We need to accelerate all elements of the cycle
- **Experimental and computational tool development:**
  → Currently constrained by human paradigm

If the part of a workflow is automated, our autonomous decision-making ability should match the level of autonomy!

Decision science of workflows

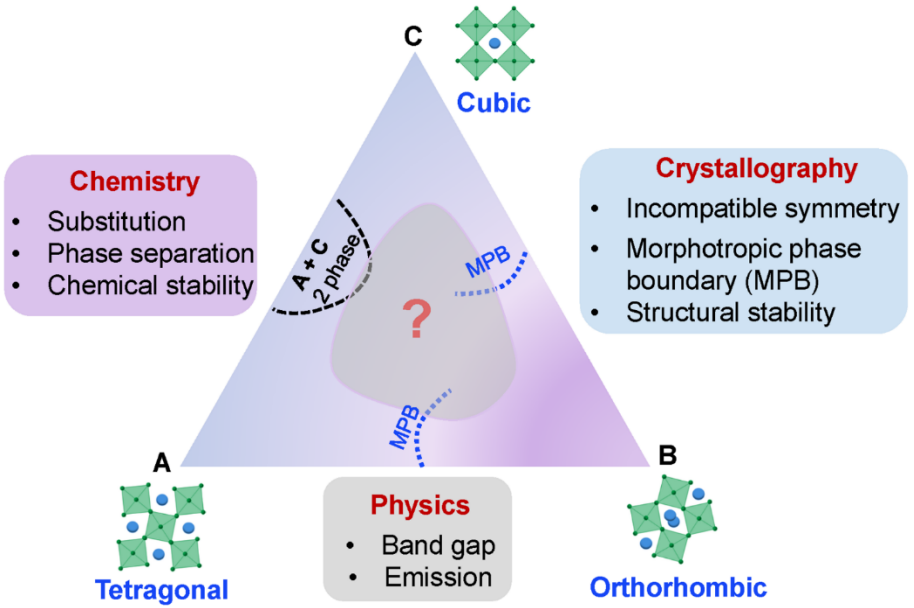




1. We need to balance multiple functionalities
2. Integrate multiple sources of data
3. Make decisions considering costs, latencies, physical inferential biases, and beliefs

Key consideration: reward function

1. Pure physical discovery (symbolic laws)
2. Data-driven exploration
3. Materials optimization
4. …

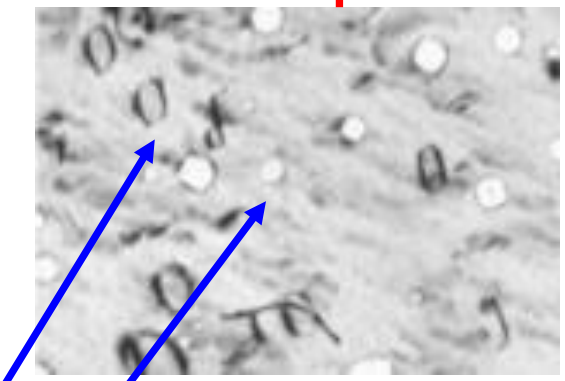


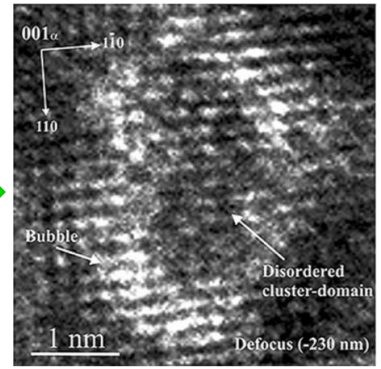
Workflows for Nuclear Materials Design

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

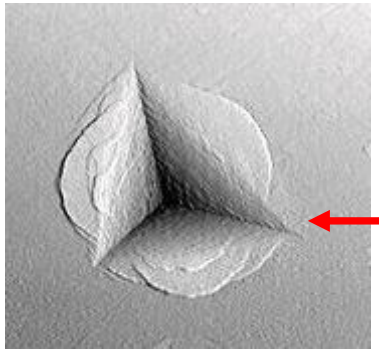
Materials selection





Prior knowledge

Operator curiosity

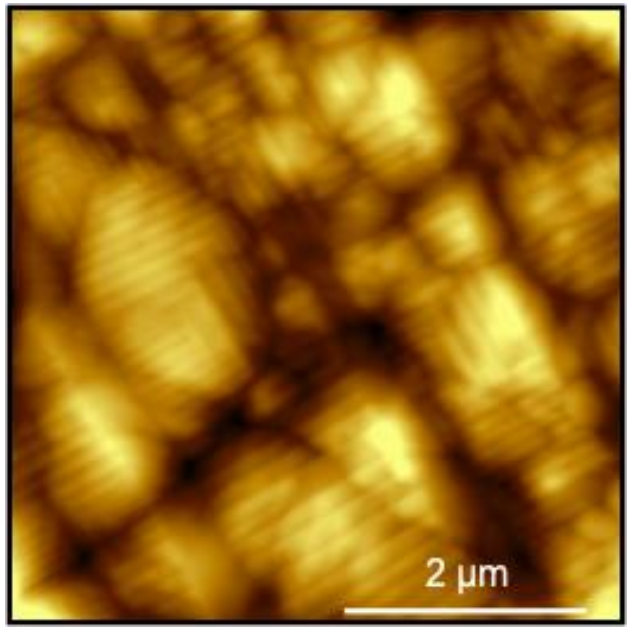


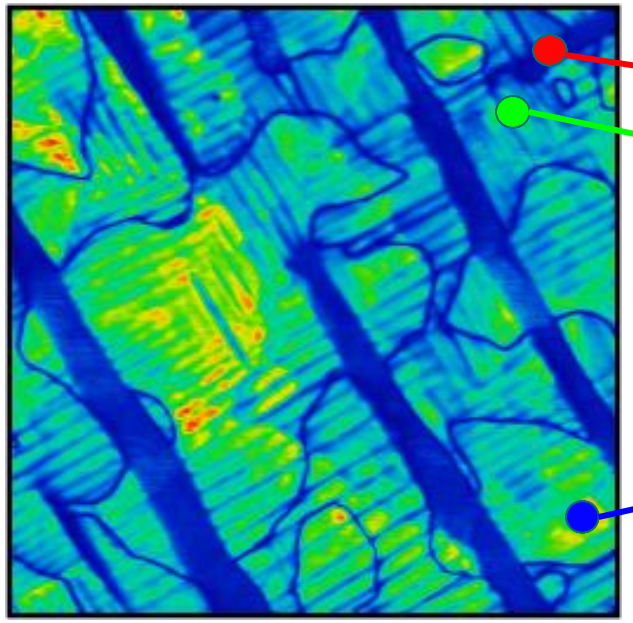
Value

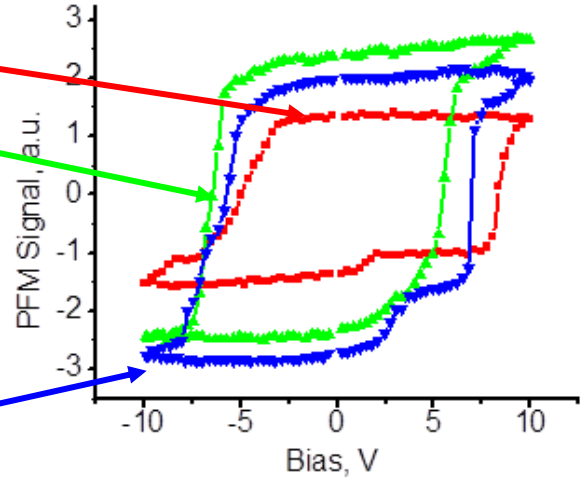
Value of the step is key element:

- Either based on prior knowledge
- Or defined in a sense of the reinforcement learning Q-function

Decision Making in SPM

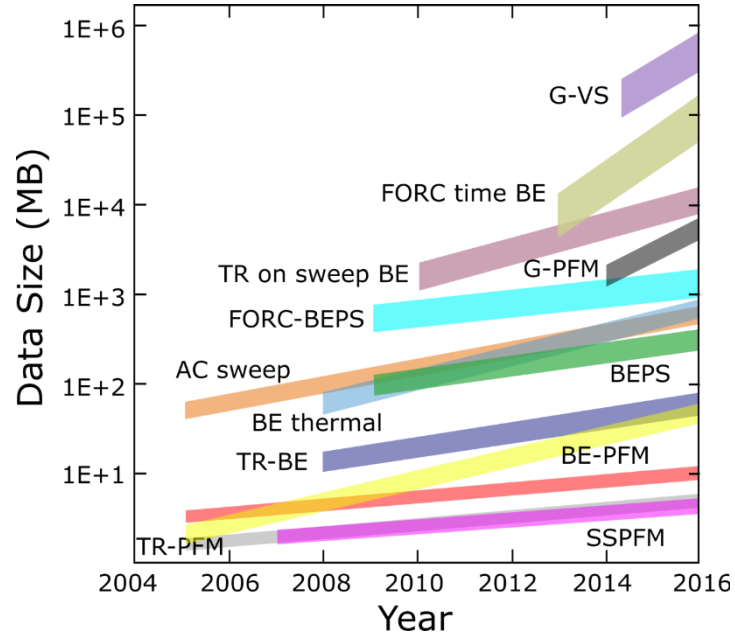






- Interesting functionalities are expected at the certain elements of domain structure
- We can guess some; we have to discover others

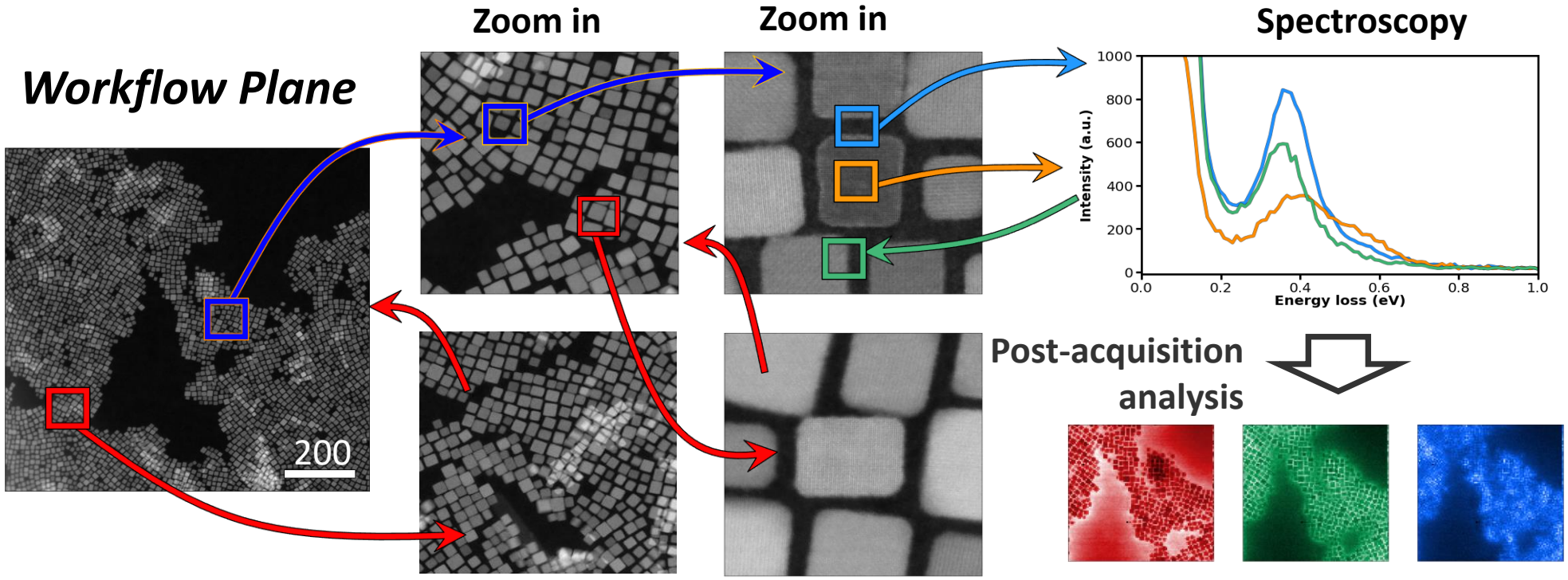
- **Experimental objectives → ML Rewards**
  - Microscope optimization
  - Properties of a priori known regions of interest
  - Discovery of regions with interesting properties
  - Physical theory falsification

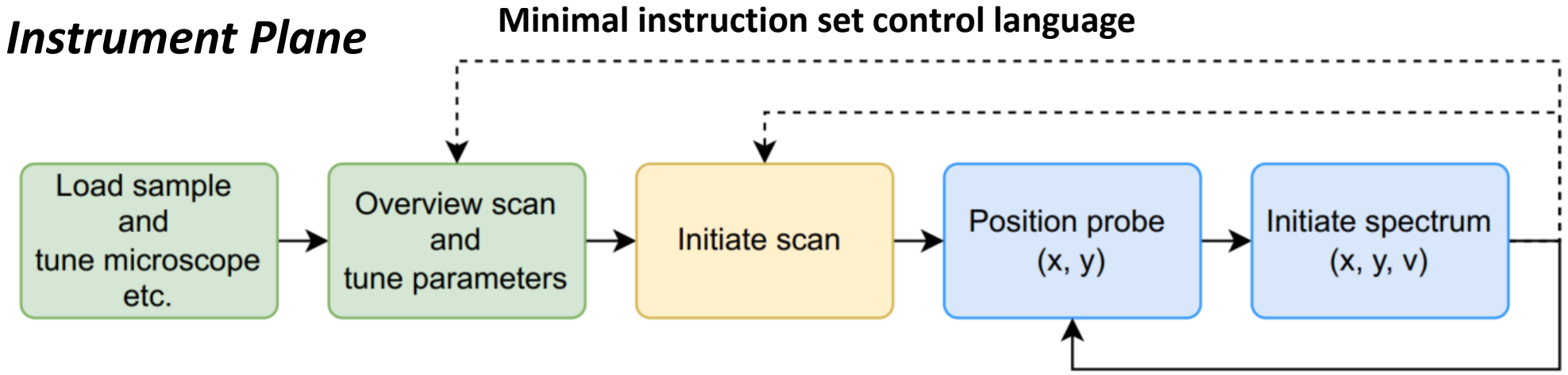


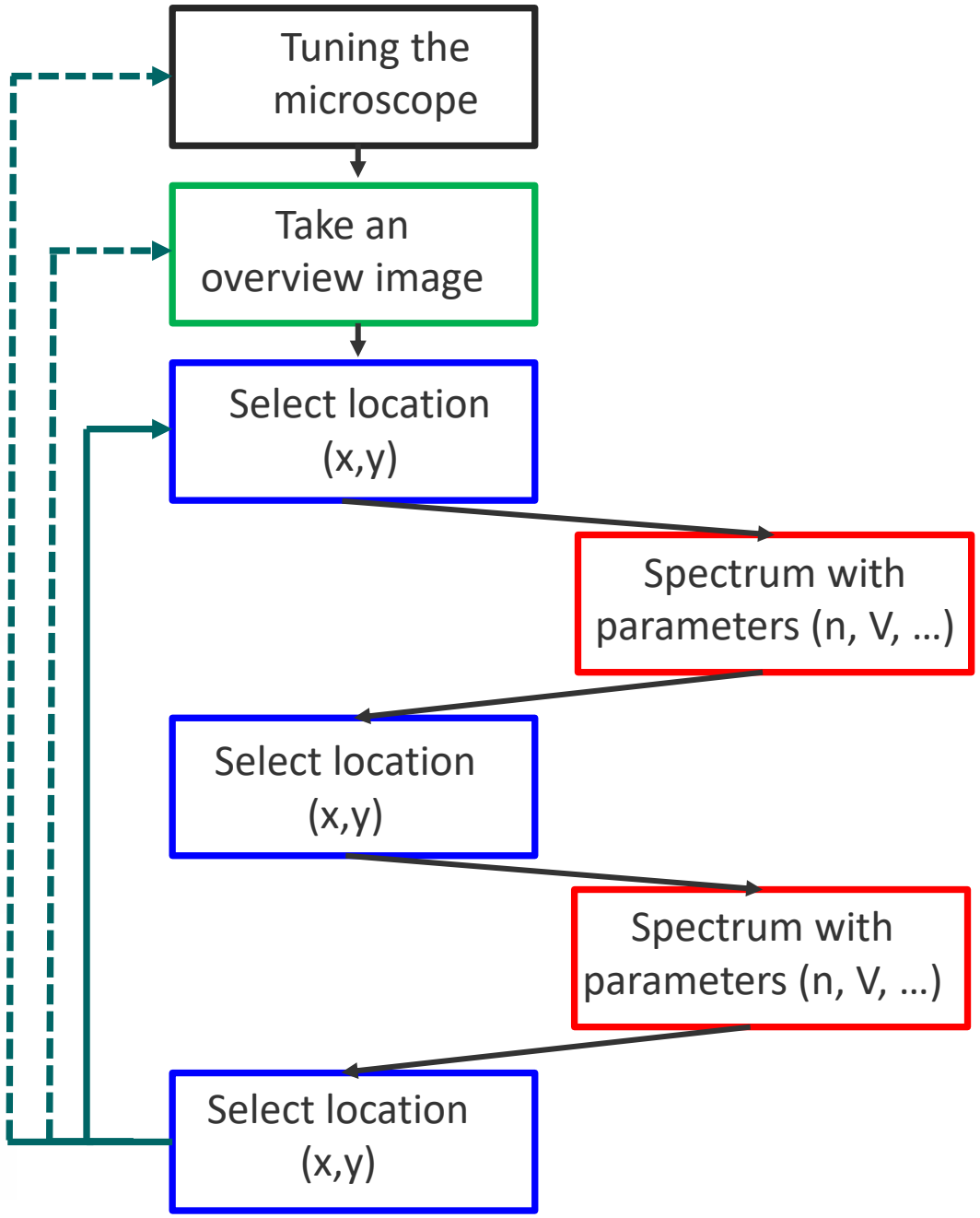
Workflows in STEM

Prior Knowledge

Objective and Reward







To implement the ML workflows, we start from emulating the human operations:

- Well defined and explainable commands
- Extensive domain expertise
- Potentially available data from experiments

Development of ML workflows can give rise to more complex imaging modalities

- Data volumes and dimensionalities above human level
- More complex modes of sampling
- “Guardian angel” modules

However, we always have to think about

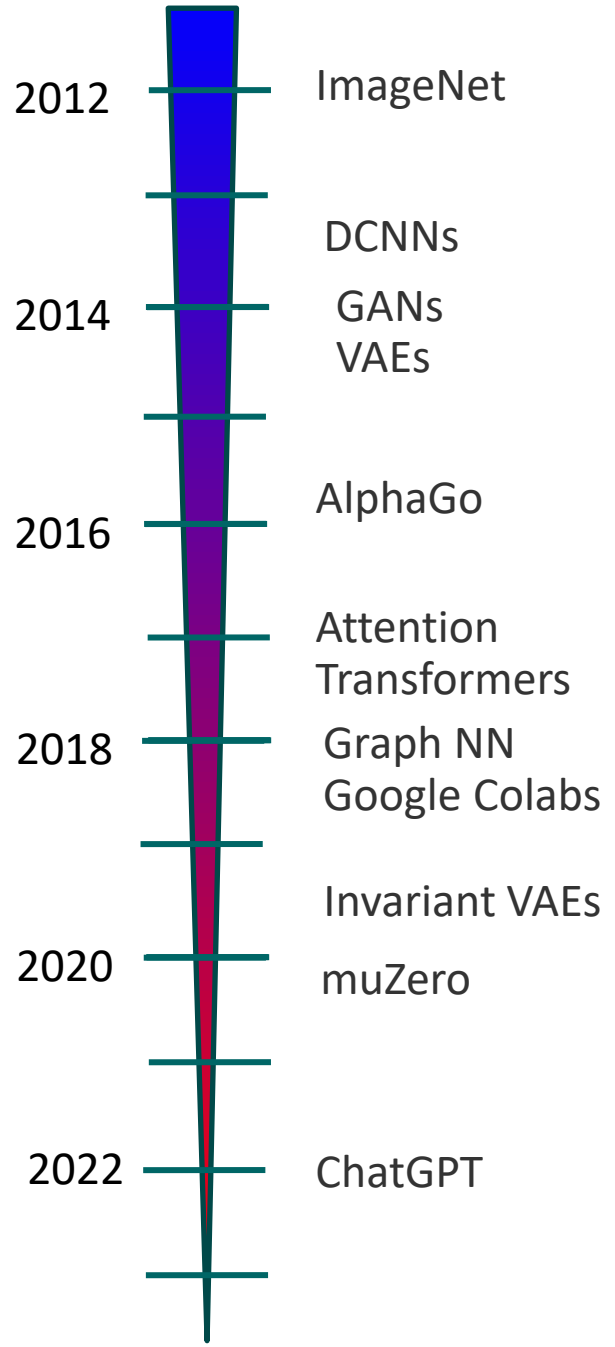
- Reward function(s) for imaging problem
- Reward functions for materials problem
- Overall objective

Why Machine Learning?

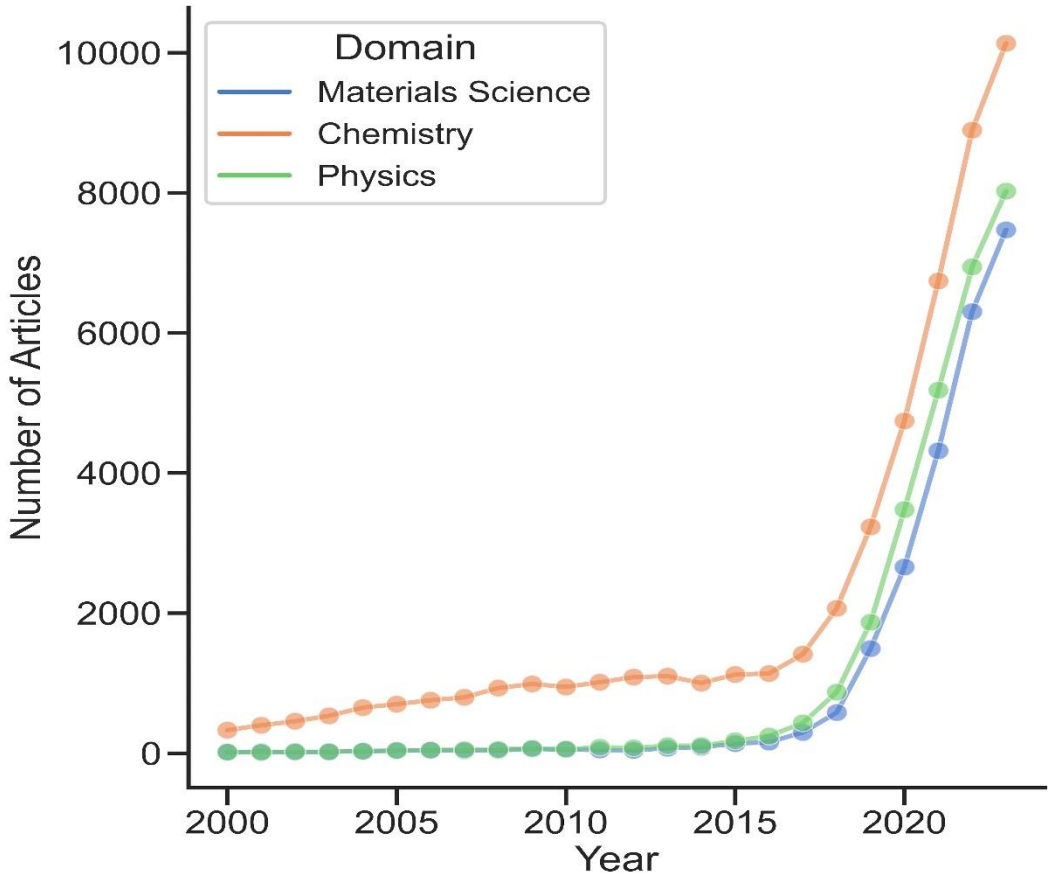
- Last decade has experienced an explosive growth of machine learning and artificial intelligence applications
- These developments have spanned areas from computer vision to medicine to autonomous systems and games
- However, the progress and impact as applied to experimental physical sciences has been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

Why is it difficult?

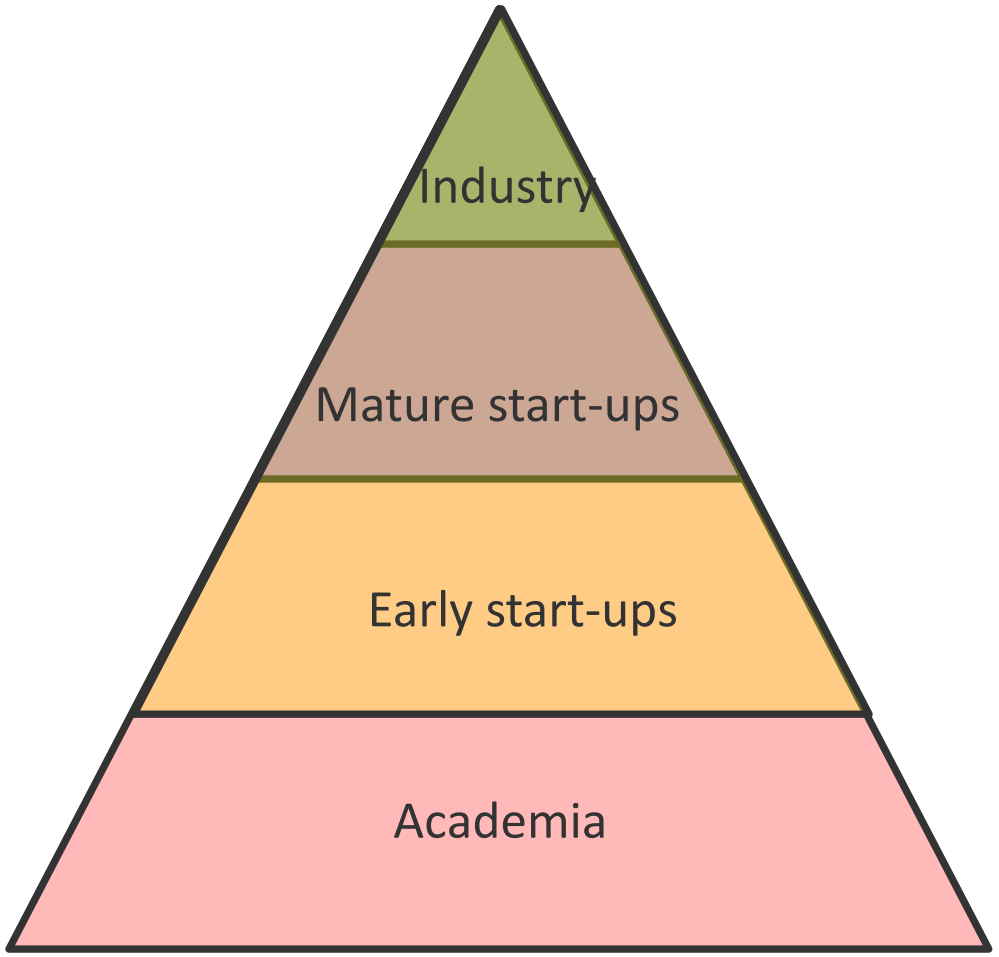
- Requires domain expertise and domain-specific goals
- Deeply causal and hypothesis drive nature of domain sciences
- No single answer: culture, not a method
- Infrastructure, open code, open data
- **Most important:** active nature of scientific process



ML in Domain Sciences



Analysis by B. Blaiszik, Argonne



- The rapid adoption of ML in domain sciences and industrial R&D is a very recent trend
- Technologies and workforce emerge from academia into industry
- We can estimate potential growth rates comparing to cloud computing 15 - 20 years ago

“Eras” of ML in Industry

- **Before 2002:** It’s all about IT (dotcoms, Amazon, etc)
- **2002 - 2012:** It’s all about collecting and searching data (Facebook, Google, Uber)
- **2012 – 2022:** What do we learn from data (correlative era)
- **2022 – now:** Physics is the new data

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

- Classical machine learning is underpinned by the existence of the static data sets – from MNIST to medical, bio, faces, etc.
- Real world problems are associated with the large distribution shifts, small data sets, and presence of uncontrollable exogenous factors
- Real world problems are often active learning: we interrogate the data generation process and provide feedback
- However, we often have extensive prior knowledge of past data, physical laws generalizing them, and strong set of inferential biases

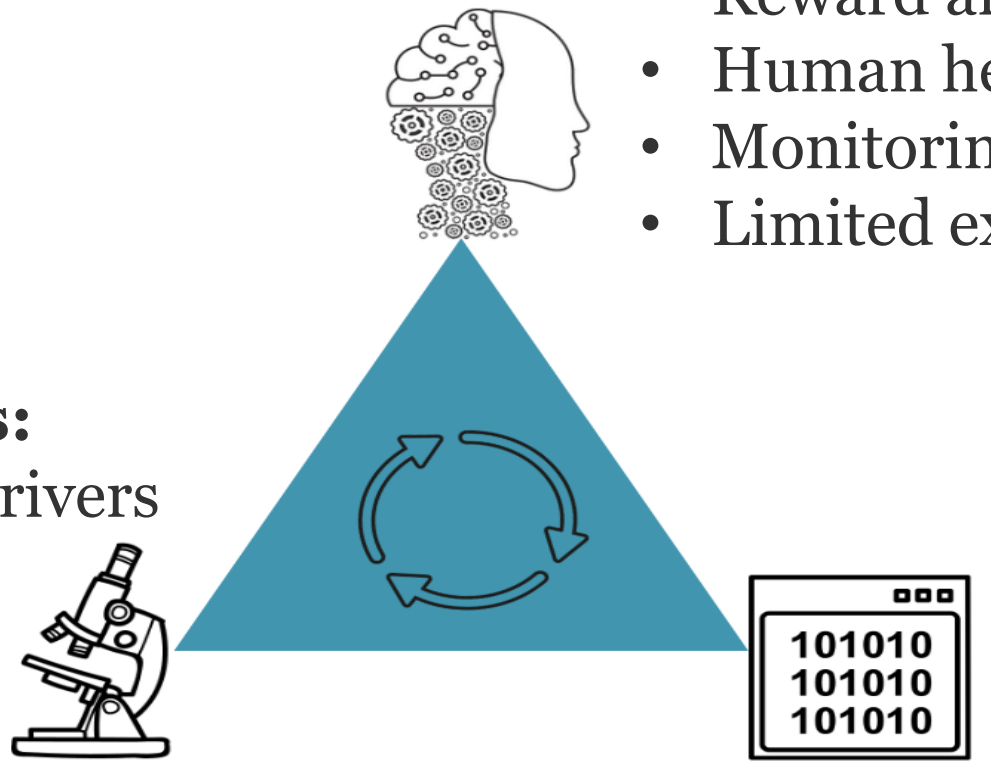
Challenges for the Automated Experiment

Elements of realistic workflow design

- Co-orchestration of multiple measurement modalities
- Building theory in the loop
- Integration between multiple domains

Workflow design:

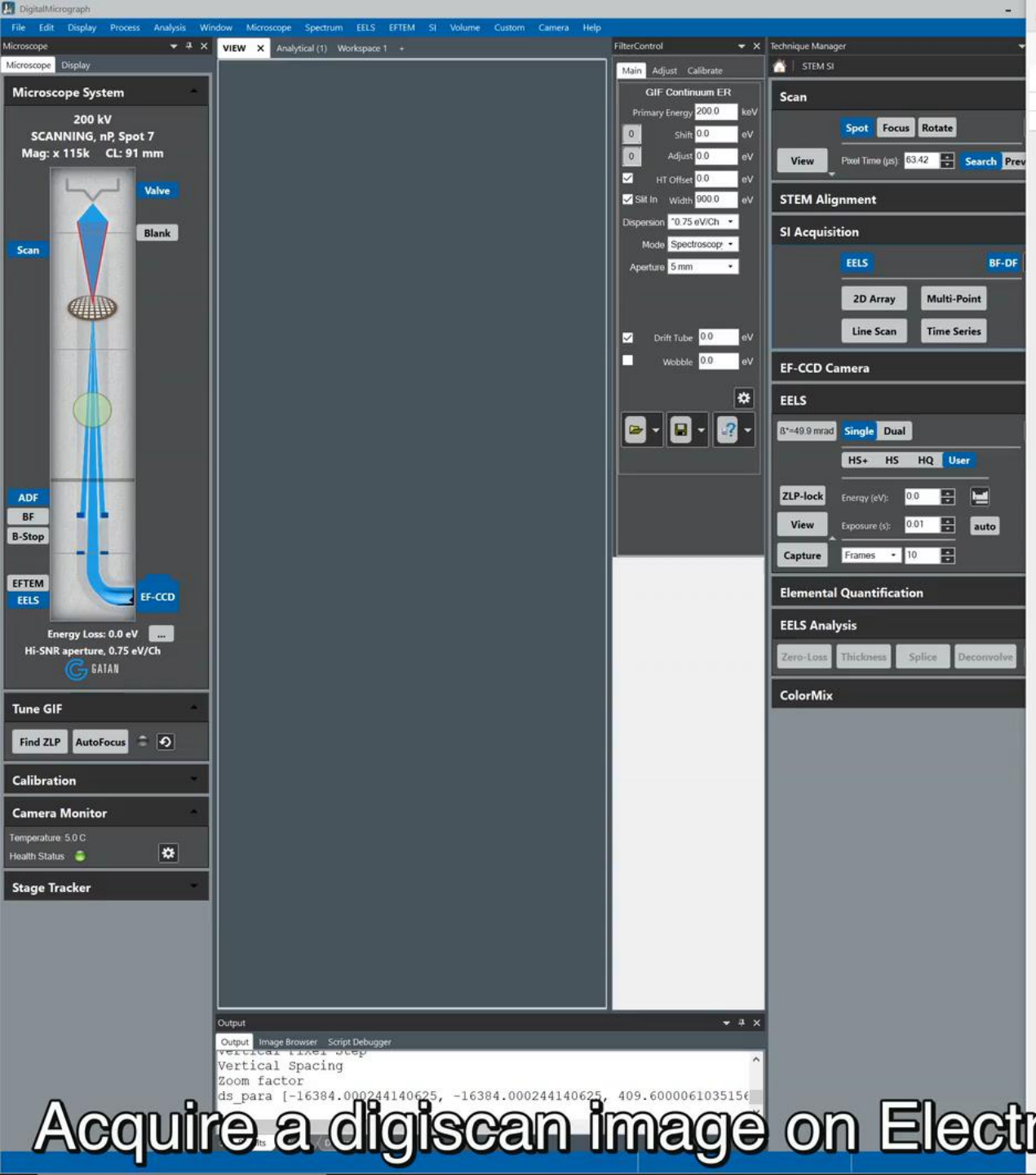
- Reward and value functions
- Human heuristics
- Monitoring
- Limited experimental budget

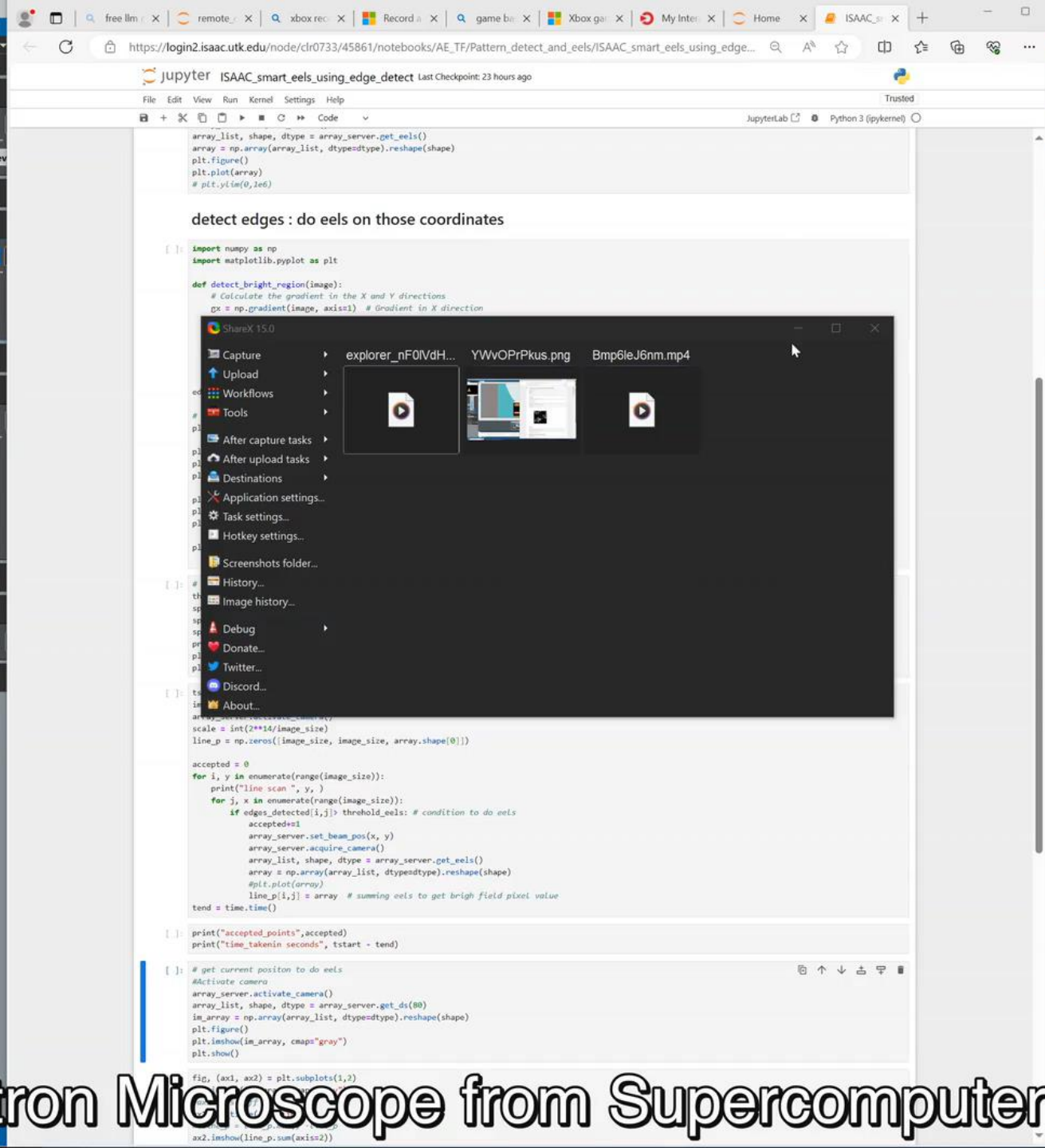


Engineering controls:

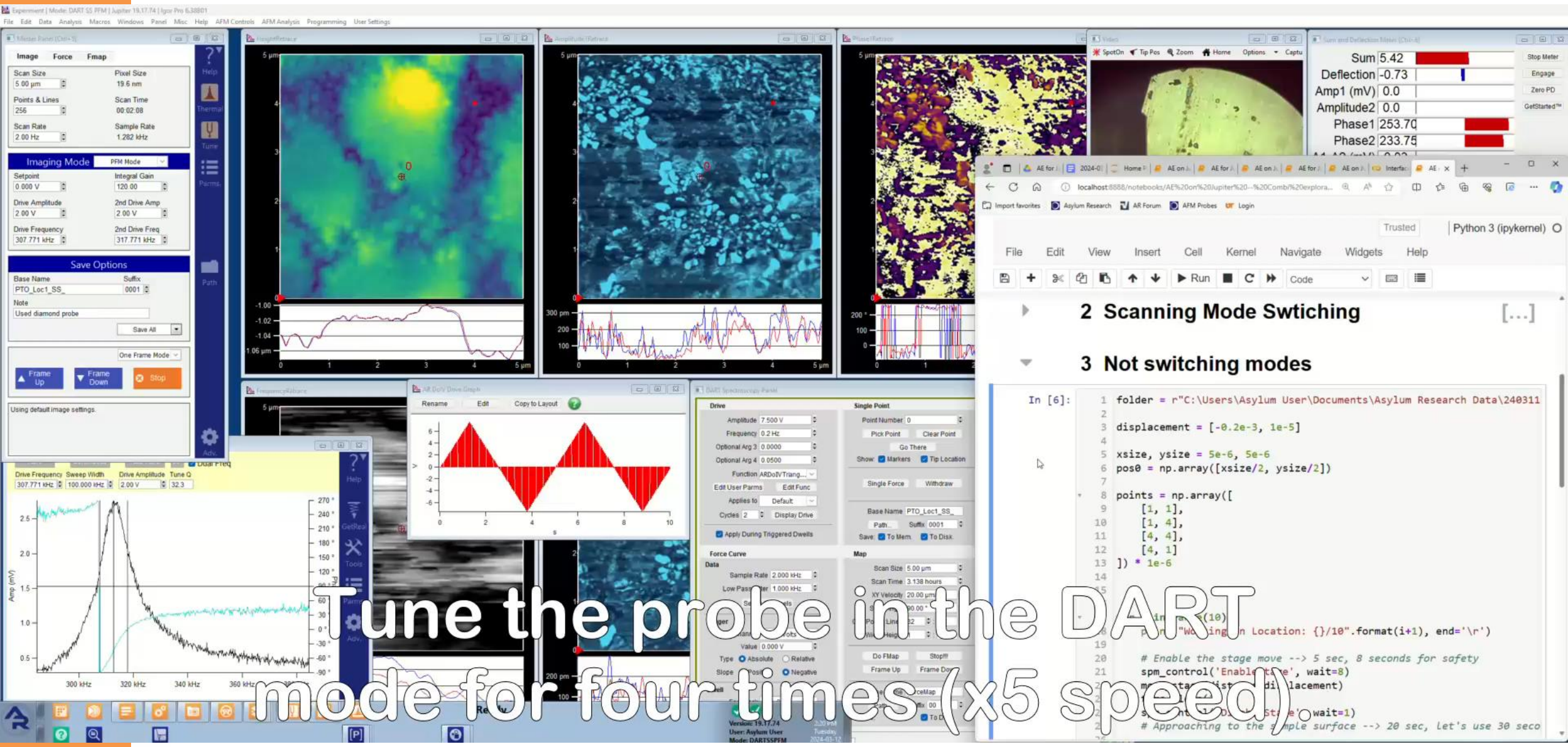
- Instrument-specific drivers
- Hyperlanguage
- Python APIs

ML/AI





Acquire a digiscan image on Electron Microscope from Supercomputer



Tune the probe in the DART mode for four times (x5 speed)

File Edit View Insert Cell Kernel Widgets Help

Trusted gpuenv

Run Code

```
27
28
29
30 move_(-volt*2-(offsetvx), 0, 0-offsetvy, 0, move_speed)
```

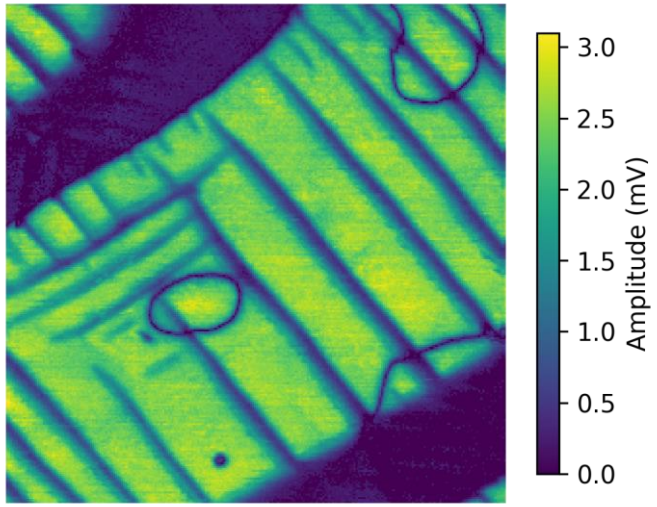
scanning line #56

In [ ]:

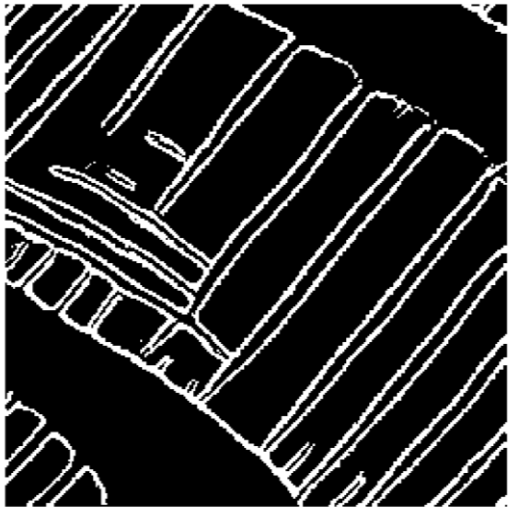
```
1
```

In [ ]:

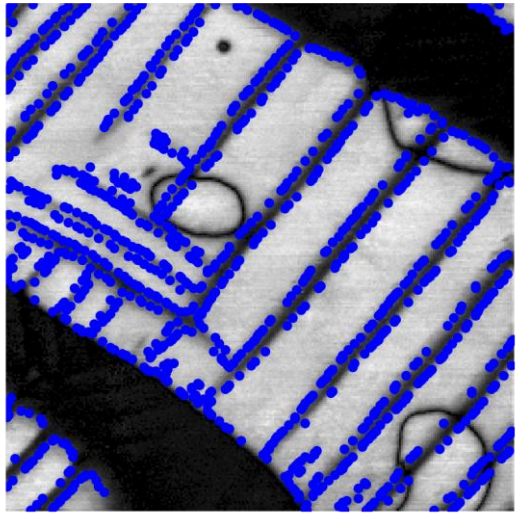
Mapping Activity of Domain Walls



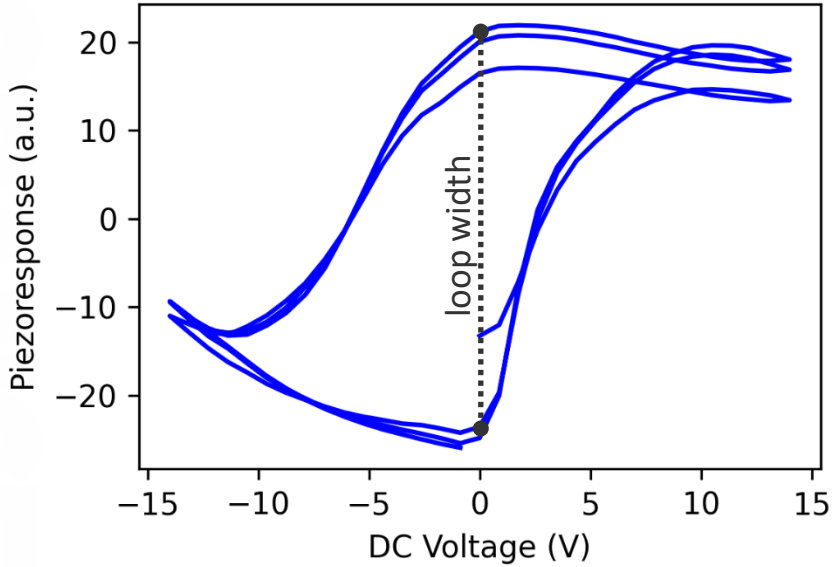
ResHedNet Prediction



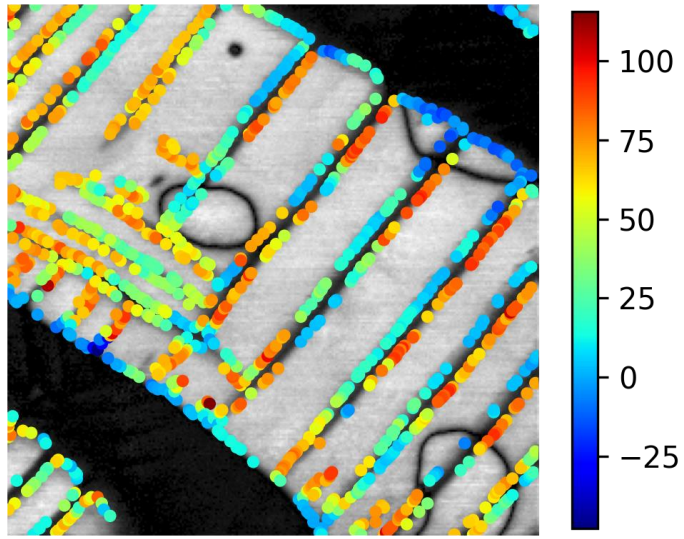
BEPS Measurement Points

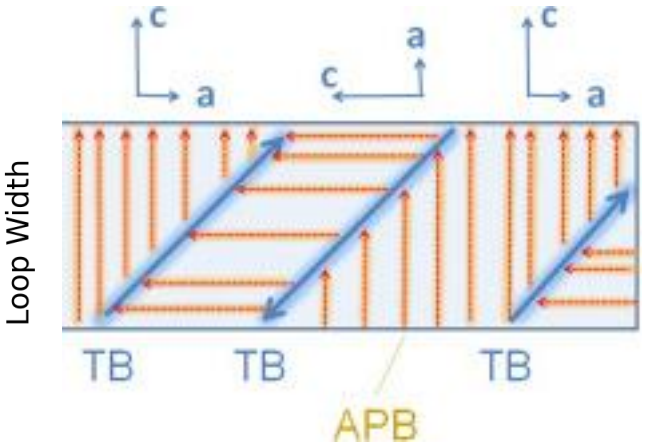


Averaged loop over ferroelastic walls



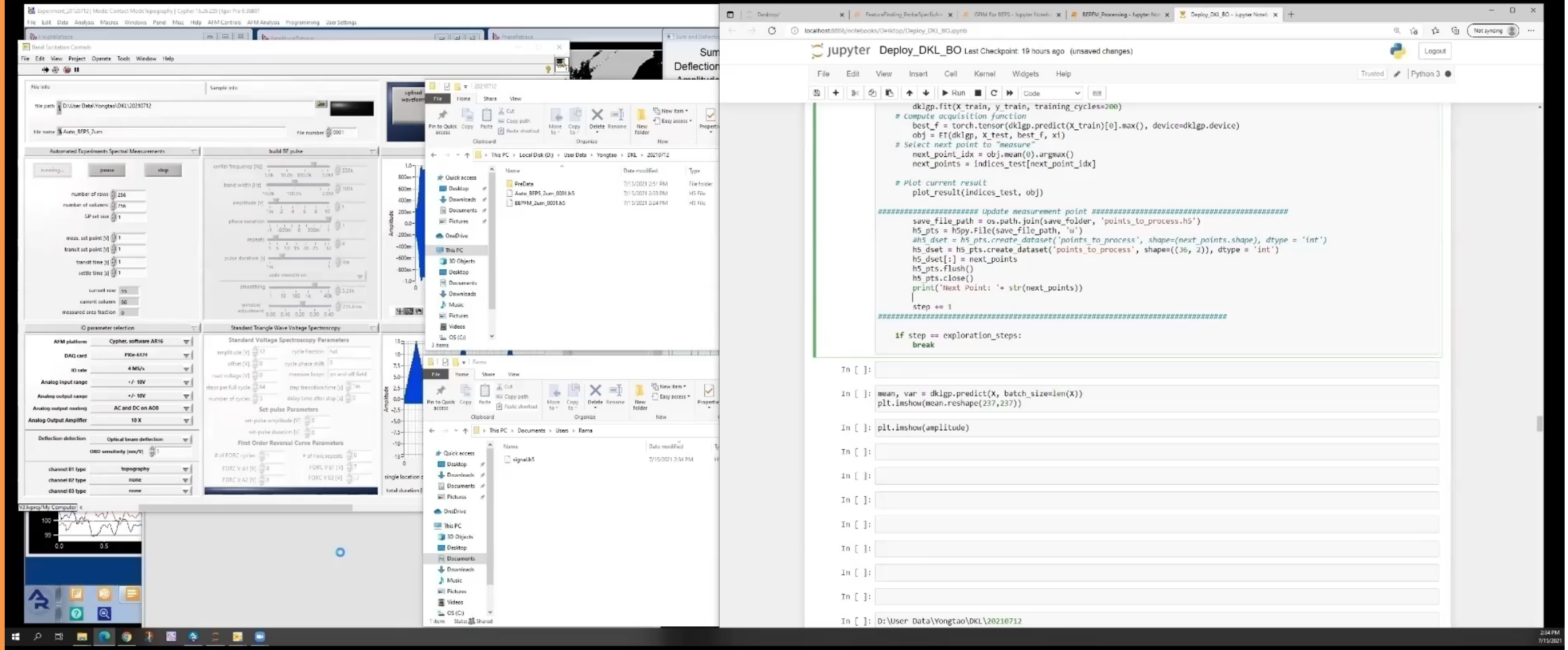
Loop height at ferroelastic walls





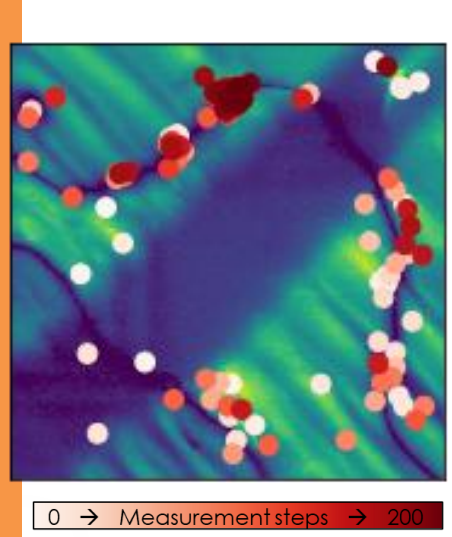
Liu, Y., Kelley, K.P., Funakubo, H., Kalinin, S.V., Ziatdinov, M., Arxiv, submitted

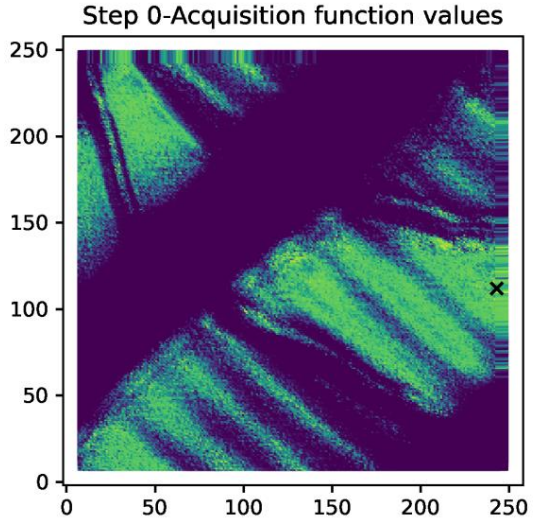
Deep Kernel Learning AE



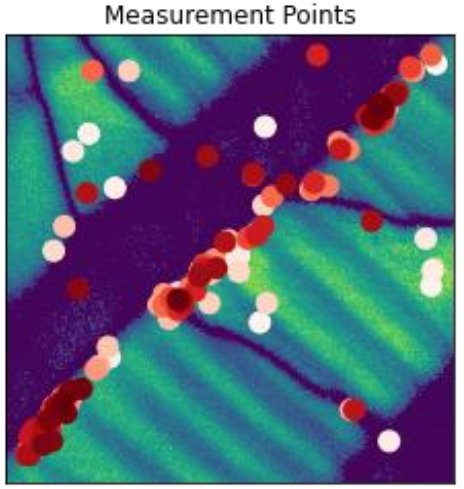
Deep Kernel Learning SPM

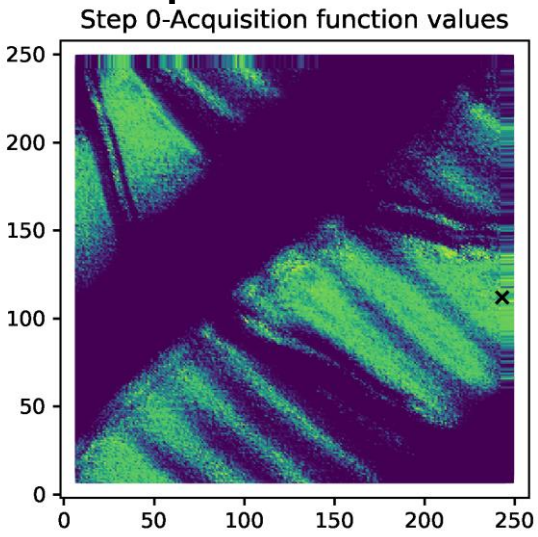
Guided by: On field loop area

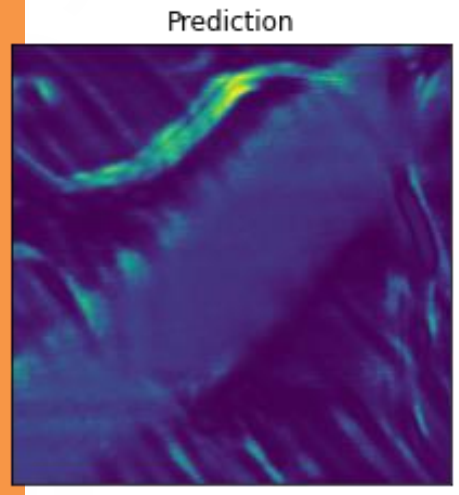


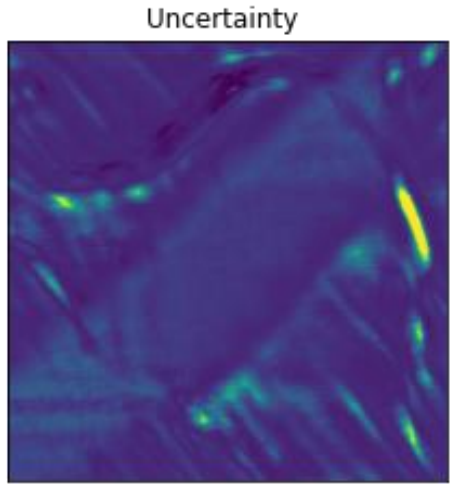


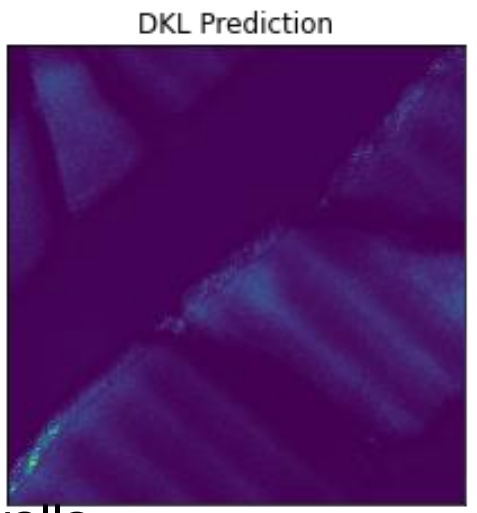
Guided by: Off field loop area

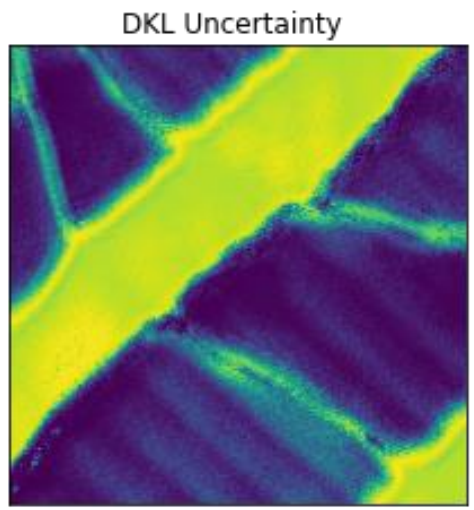








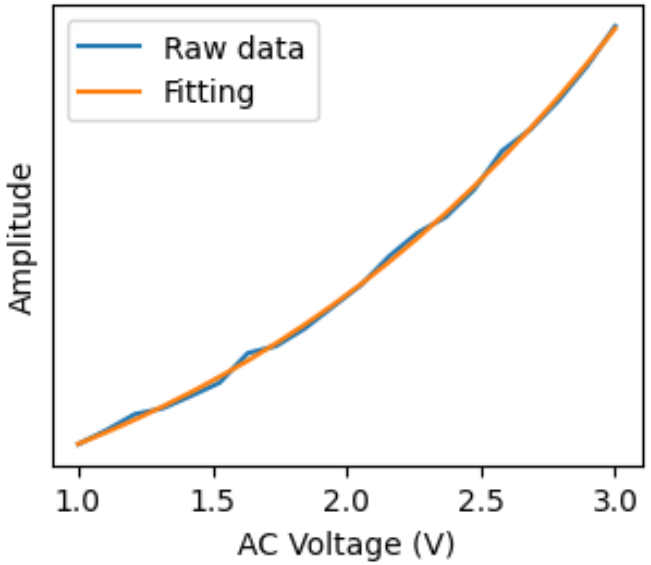


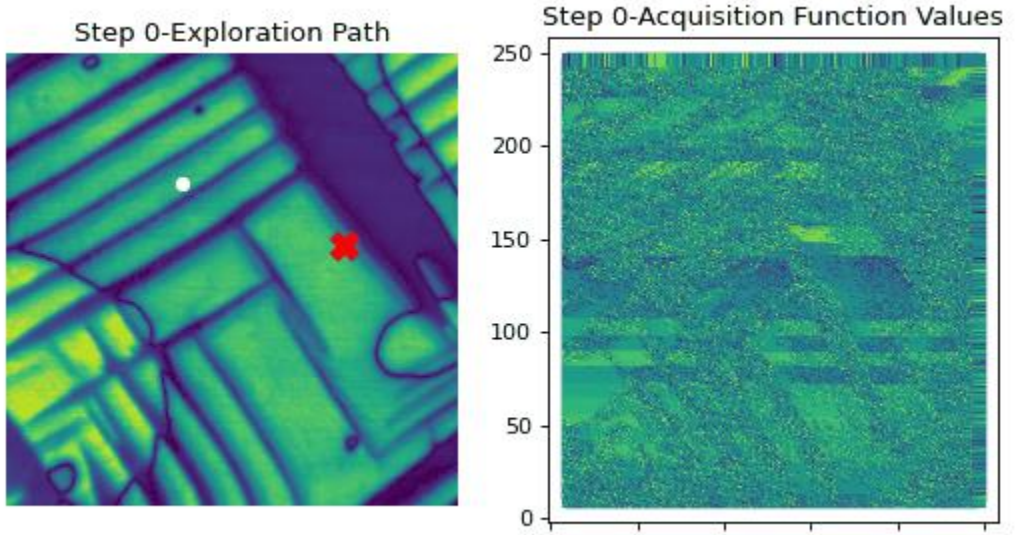


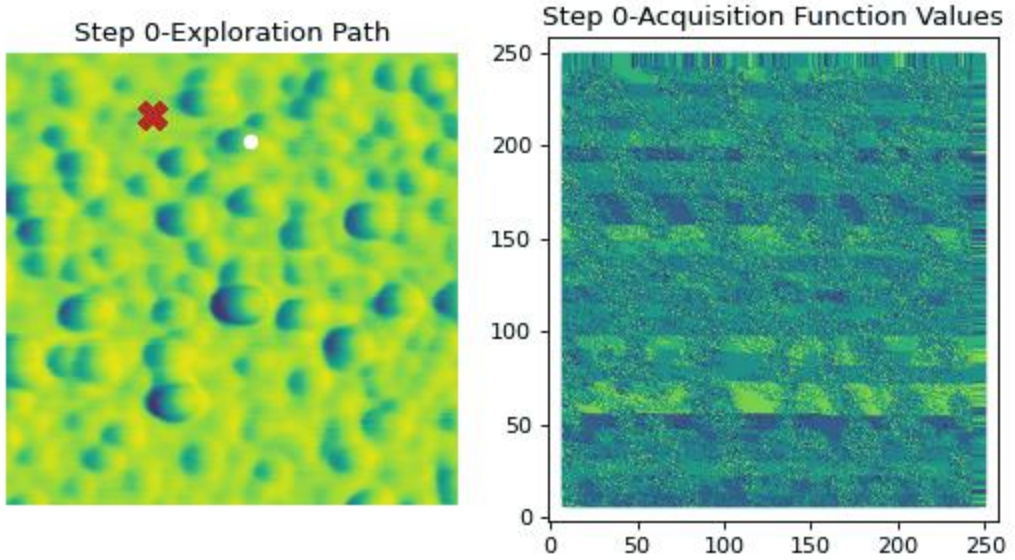
- Large loop opening corresponding 180º domain walls
- This behavior can be attributed to the large polarization mobility of 180º walls

Liu, Yongtao, et al, Nature Machine Intelligence 4, 4 (2022): 341-350.

Why human in the loop?







- 200-step automated experiment
- PFM amplitude was used as structure ima
- V_{AC} sweep curve at each location was fitte
  $y = Ax^3 + Bx^2 + Cx$
- A, B, C, and A/B were used as the target function to guide DKL-V_{AC} measurement.

The methodologies of classical ML (hyperparameter optimization, cross-validation) are rarely applicable for active learning!

The dance of policies and rewards

Rewards and objectives:

- What is our (hierarchical) objective?
- Can we define reward(s)?

Inferential biases:

- What do we know before the experiment?
- What do we (hope to) learn after the experiment?

Experiment planning – policies and values

- How do we plan experiment in advance (policies or values based on rewards)?
- Can we ascribe value to certain steps?
- Do we change our policies during experiment?

Our vocabulary

- **Prior**
- **Posterior**
- **Belief**

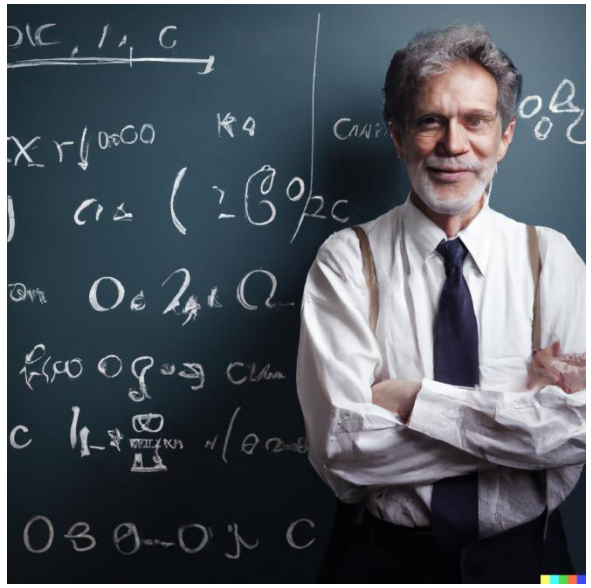
Setting of the problem

- Imagine that you are tossing a coin with your friend to decide who goes for groceries. What is the probability that the coin will land tails?

- Imagine that previously you have tossed this coin several times, and it landed tails 3 times in a row.

- … what if it was 100 times in a row?

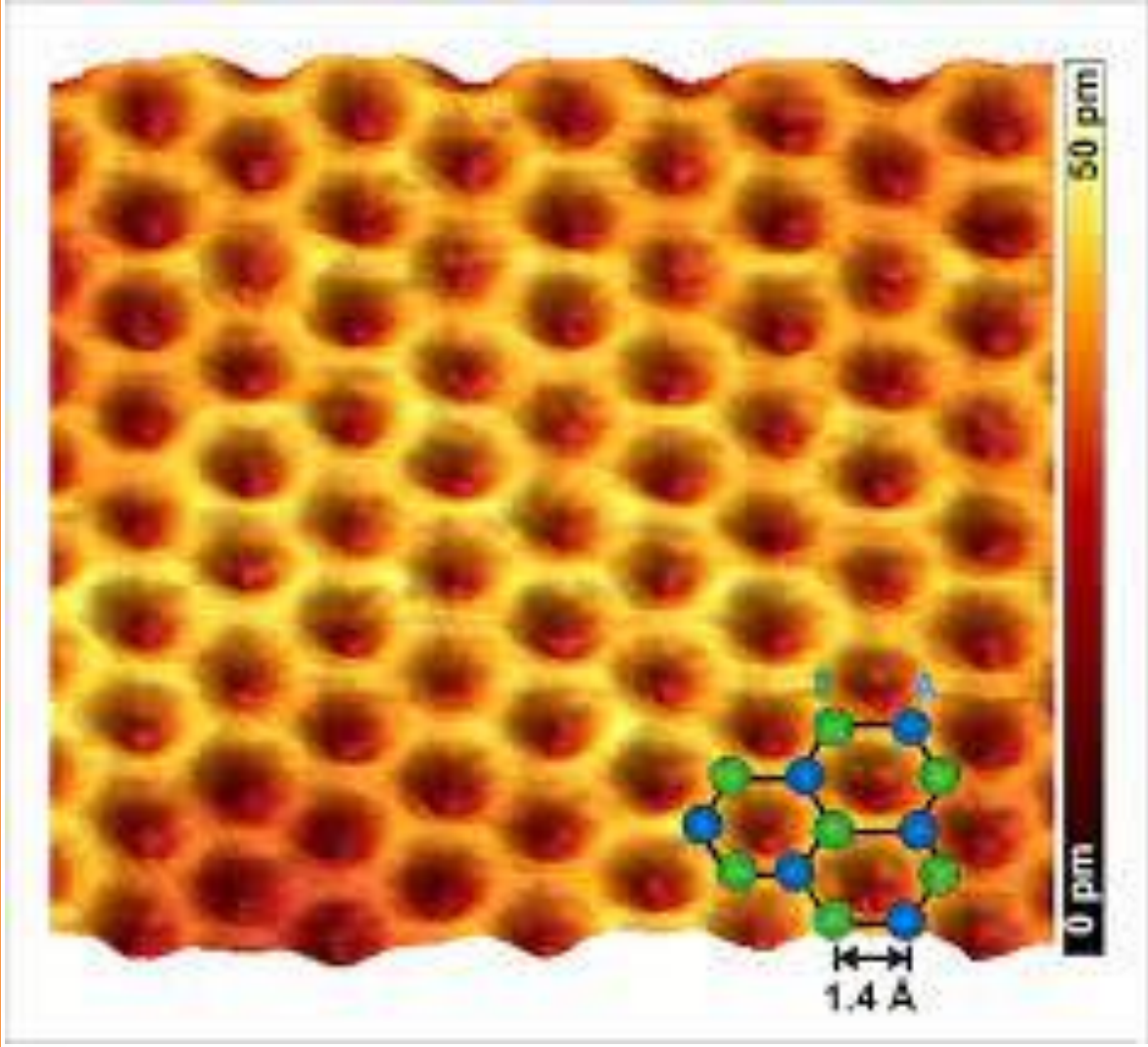
Setting of the problem - 2

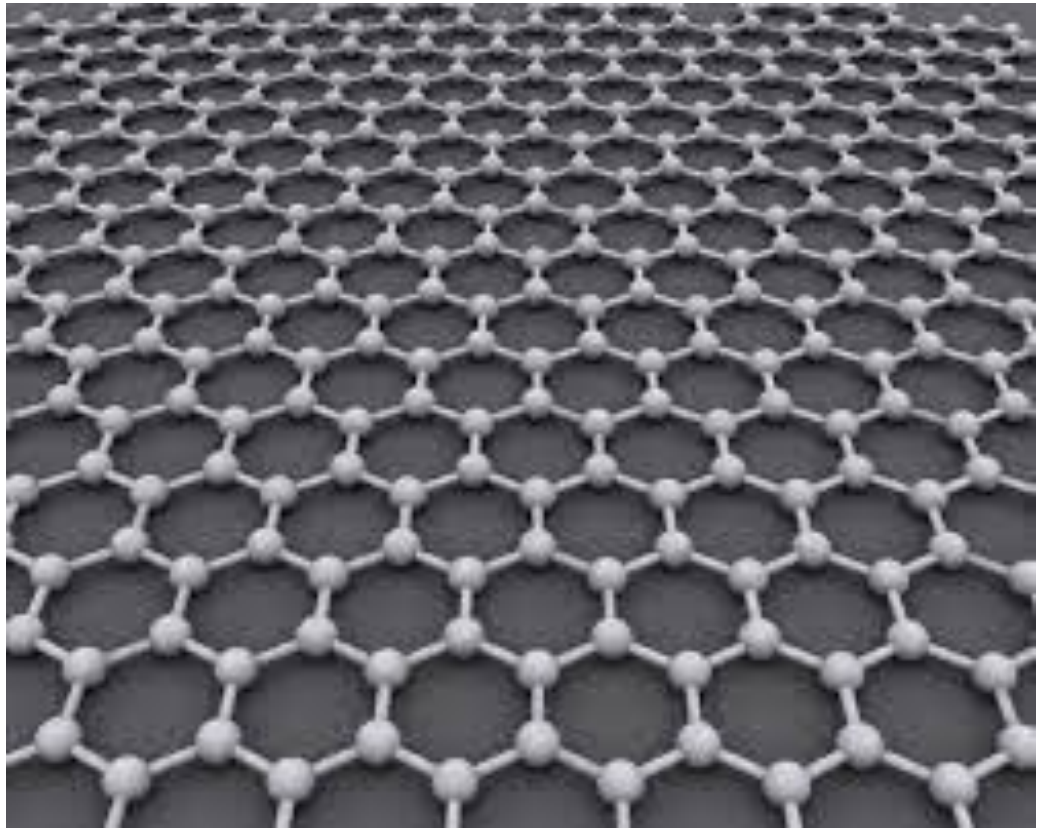


- You have a math exam problem made by Prof. Calculus stating:
  - You have a fair coin.
  - You have tossed it 100 times, and it landed tails.
  - What is the probability that it will land tails during 101 attempt?

- You play the coin toss game with a sketchy person named Joe the Gambler in Las Vegas.
  - He tells you that “We have a fair coin”.
  - You have tossed it 10 times, and it landed tails.
  - What is the probability that it will land tails during 11 attempt?

Off to scientific examples

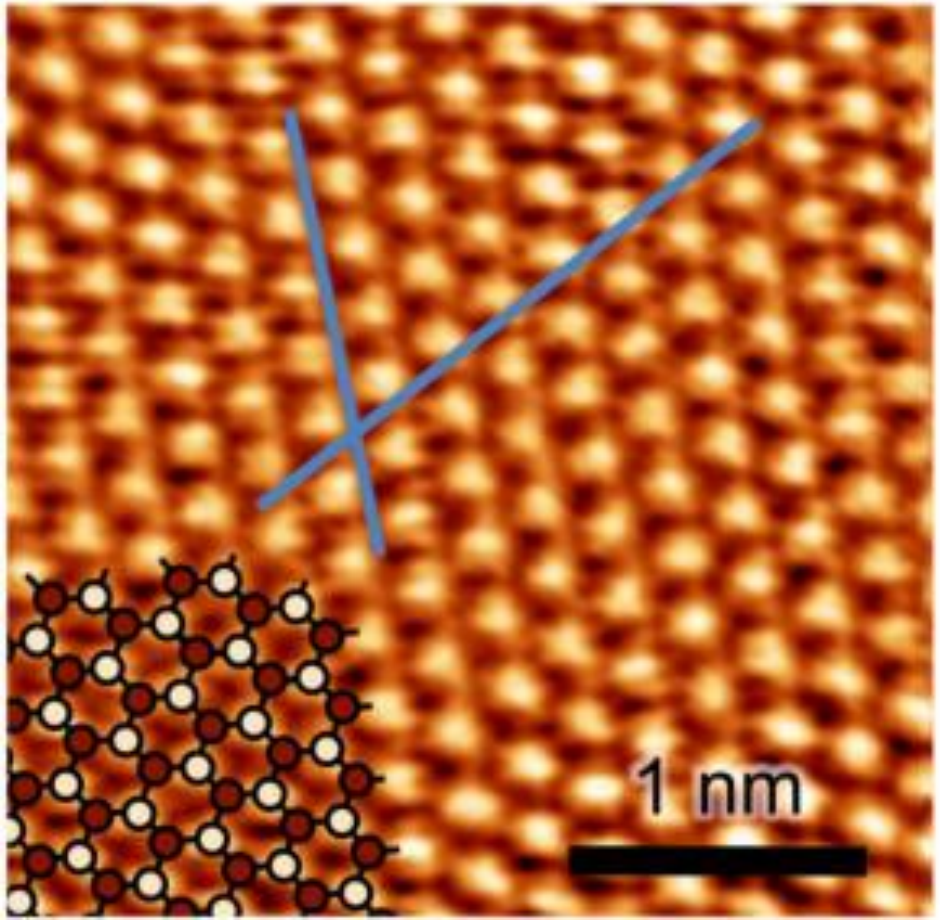


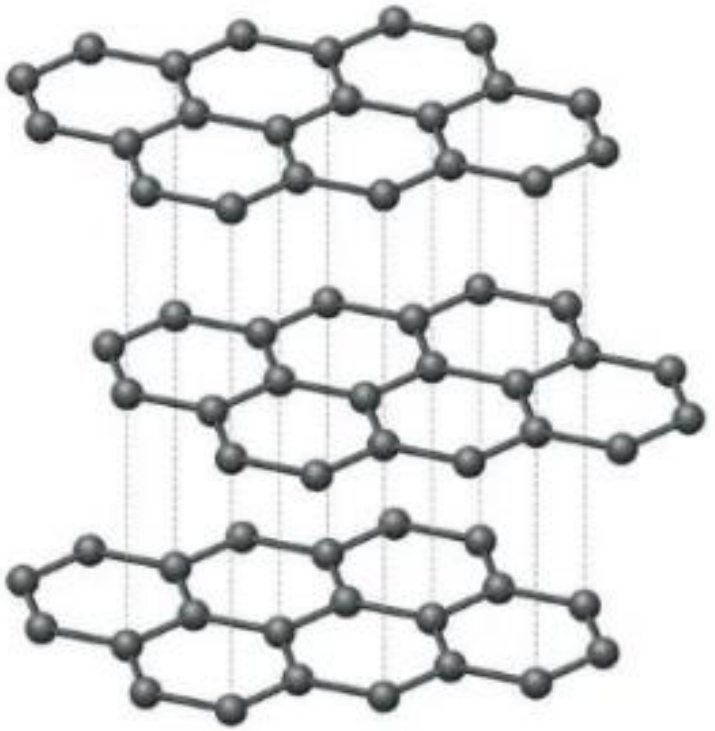


http://www.nanoscience.de/HTML/research/graphene.html

https://en.wikipedia.org/wiki/Graphene

Off to scientific examples





https://unacademy.com/content/jee/study-material/chemistry/structure-of-graphite-and-uses/

Atomically resolved STM image of the graphite surface. A schematic drawing of the hexagonal lattice of graphite is overlaid . Bright and dark filled circles correspond to the two different sublattice sites, and only the white sublattice is observed in STM image.

A. Amend, T. Matsui, H. Sato, and H Fukuyama, STS Studies of Zigzag Graphene Edges Produced by Hydrogen-Plasma Etching, Journal of Surface Science and Nanotechnology 16, 72 (2018) DOI:10.1380/ejssnt.2018.72

Frequentist paradigm

- Defines probability as a long-run frequency independent, identical trials

- Looks at parameters (i.e., the true mean of the population, the true probability of heads) as fixed quantities

This paradigm leads one to specify the null and alternative hypotheses, collect data, calculate the significance probability under the assumption that the null is true, and draw conclusions based on these significance probabilities using size of the observed effects to guide decisions

R. A. Fisher (1890–1962)

https://en.wikipedia.org/wiki/Ronald_Fisher

Bayesian paradigm

- Defines probability as a subjective belief (which must be consistent with all of one's other beliefs)

- Looks at parameters (i.e., the true mean population, the true probability of heads) as random quantities because we can never know them with certainty

This paradigm leads one to the following process:

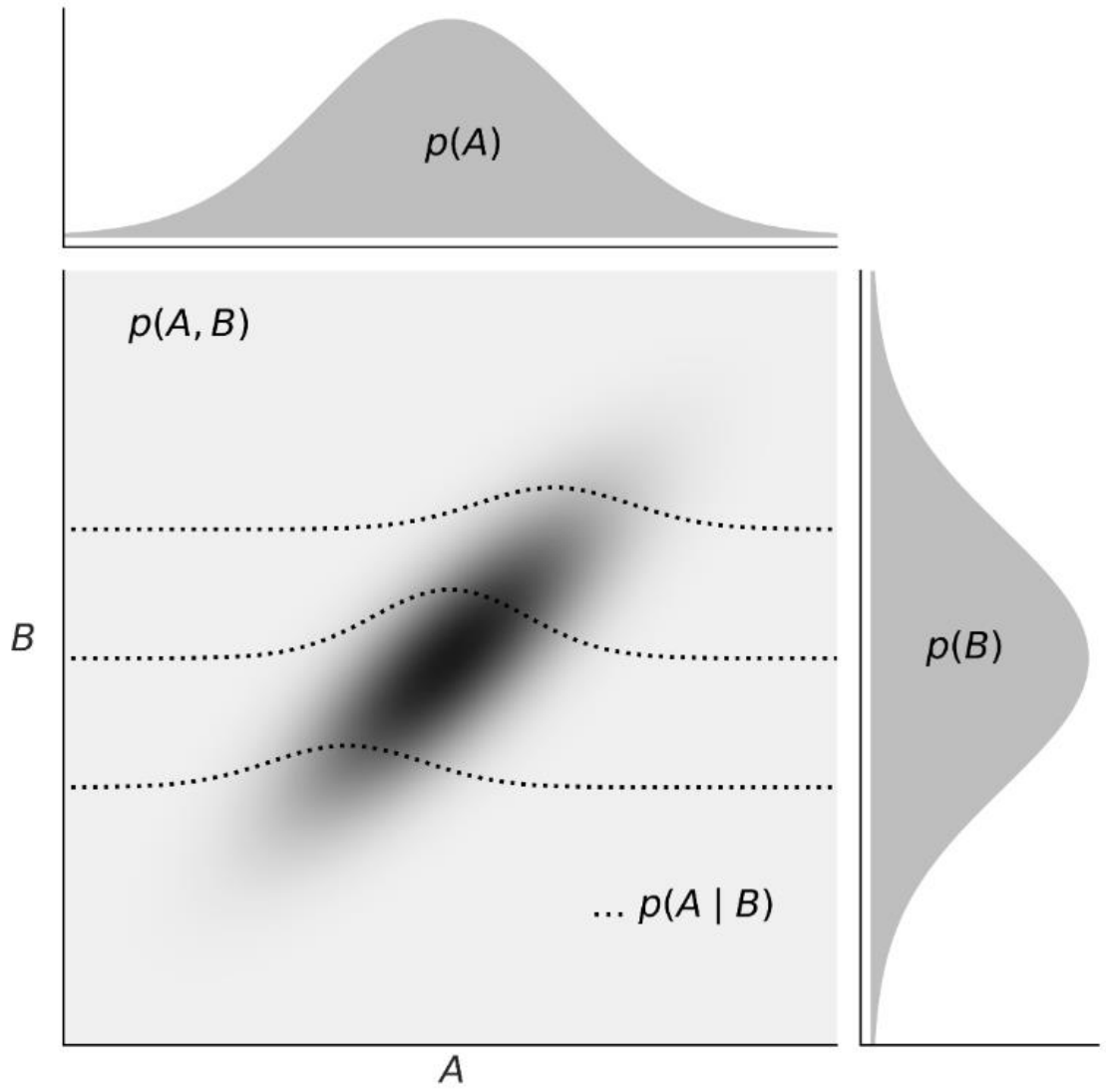
- specify plausible models
- assign a prior probability to each model,
- collect data,
- calculate the probability of the data under each model,
- use Bayes' theorem to calculate the posterior probability of each model,
- make inferences based on these posterior probabilities.

The posterior probabilities enable one to make predictions about future observations and one uses one's loss function to make decisions that minimize the probable loss

Thomas Bayes, 1701 - 1761

https://en.wikipedia.org/wiki/Thomas_Bayes

Probabilities: joint, conditional, marginal



From O. Martin

- **Joint Probability P(A, B):** Probability of two events happening at the same time. For example, the probability that it rains (Event A) and the temperature is below freezing (Event B) on the same day, i.e. intersection of A and B.
- **Conditional Probability, P(A|B):** Probability of an event A happening given that event B has already happened. It's a measure of the probability of one event occurring with some relationship to one or more other events. For instance, the probability that it rains today (Event A) given that it was cloudy in the morning (Event B).
- **Marginal Probability, P(A): P**robability of an event occurring irrespective of the outcome of another event. It can be thought of as the probability of a single event without consideration of another event. For example, the probability that it rains today (Event A) without any regard to the temperature (Event B) or any other conditions.

Bayesian paradigm in science

- Bayes' theorem can be usefully re-written for science as:

Posterior: probability of the model given the data

Likelihood: probability of the data given the model

Prior: probability of the model

$$P(\text{model}|\text{data}) = \frac{P(\text{data}|\text{model})\,P(\text{model})}{P(\text{data})}$$

Evidence [can typically be absorbed into the normalization of the posterior]

The World is Bayesian: Physics from Observations

Hypothesis driven science:
What we want to learn

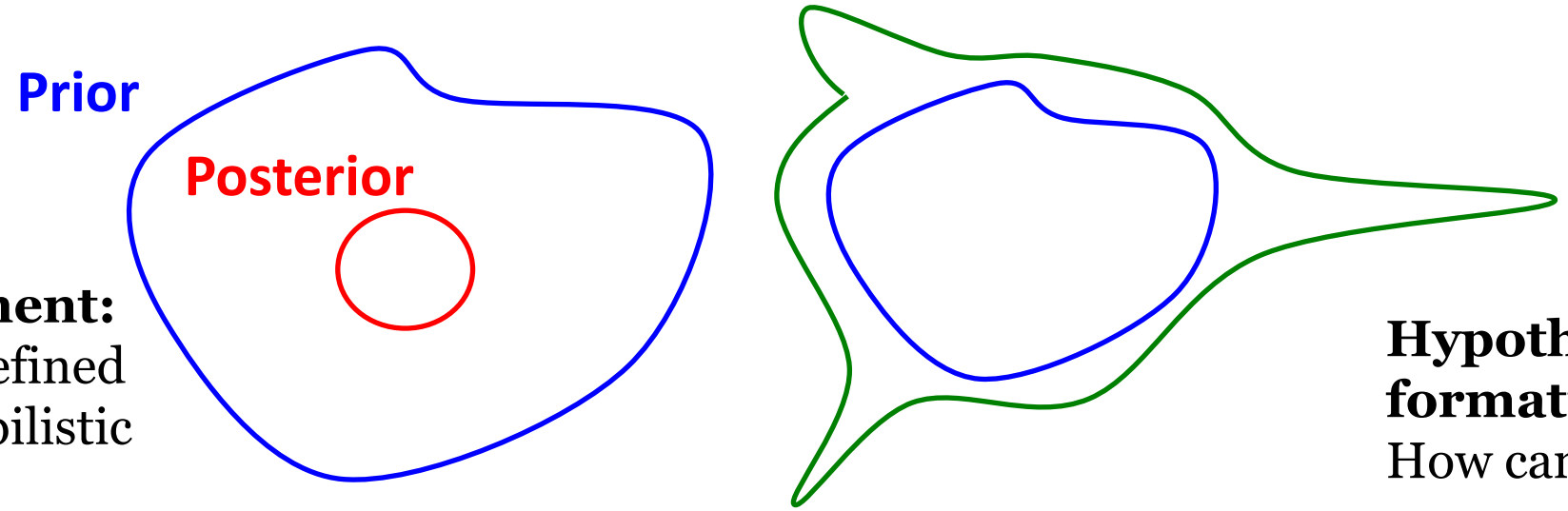
Forward model: Theory

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?

Colab