THE UNIVERSITY of TENNESSEE KNOXVILLE

# Lecture 7: Human in the loop automated experiment and LLM-based co-scientists

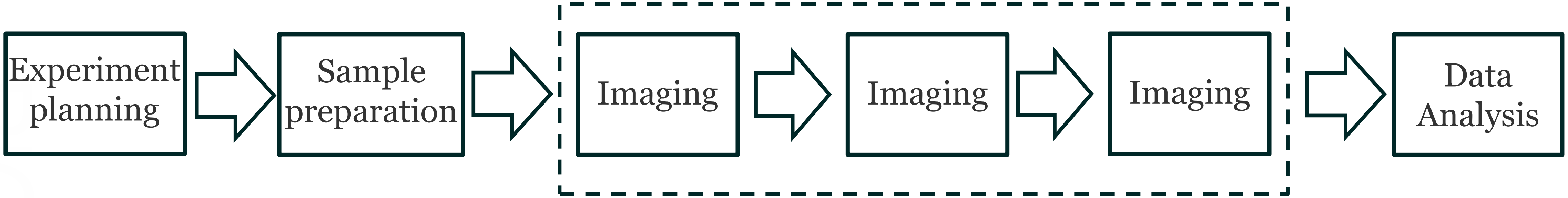
Sergei V. Kalinin

University of Tennessee, Knoxville, and
Pacific Northwest National Laboratory

Pacific Northwest
NATIONAL LABORATORY

# What is your goal?

**Level 5:** Downstream Use of Microscopy Data: Incorporates microscopy data into theory analysi pipelines, closing the synthesis-characterization-discovery loop.




**Level 4:** Upstream Task Planning:
ML is used for planning experiments, including sample selection and integrating microscopy with materials synthesis.

**Level 2:** Real-Time Analytics

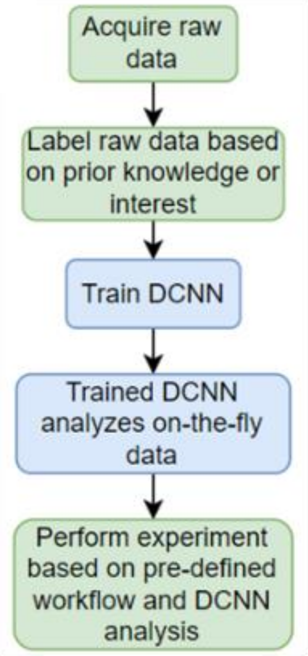
- ML helps represent data in a form that is more understandable to humans.
- Decisions are still made and orchestrated by humans.

**Level 1:** Post-Acquisition Data Analysis

**Level 3:** ML Agent Introducing Decisions: automated microscopy

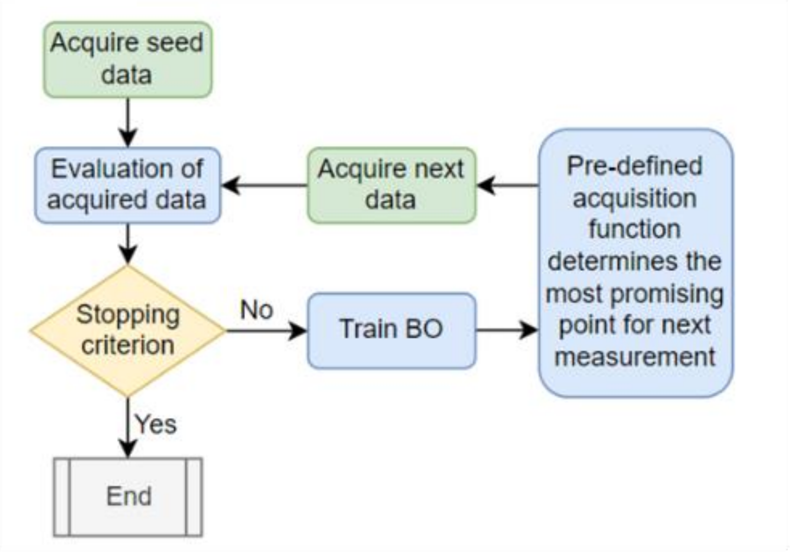
# Types of automated experiment

**Direct**



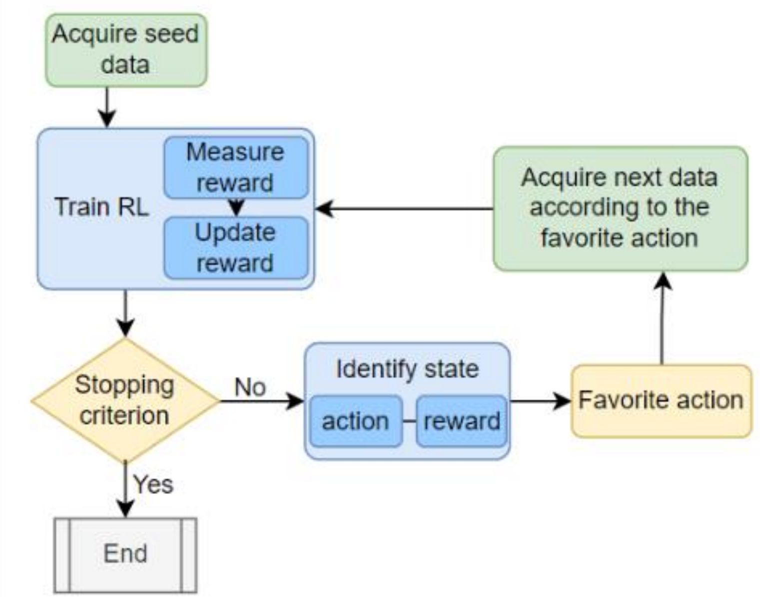

- Fixed policies
- Need DCNNs stable wrt. out of distribution shift

**Myopic discovery**



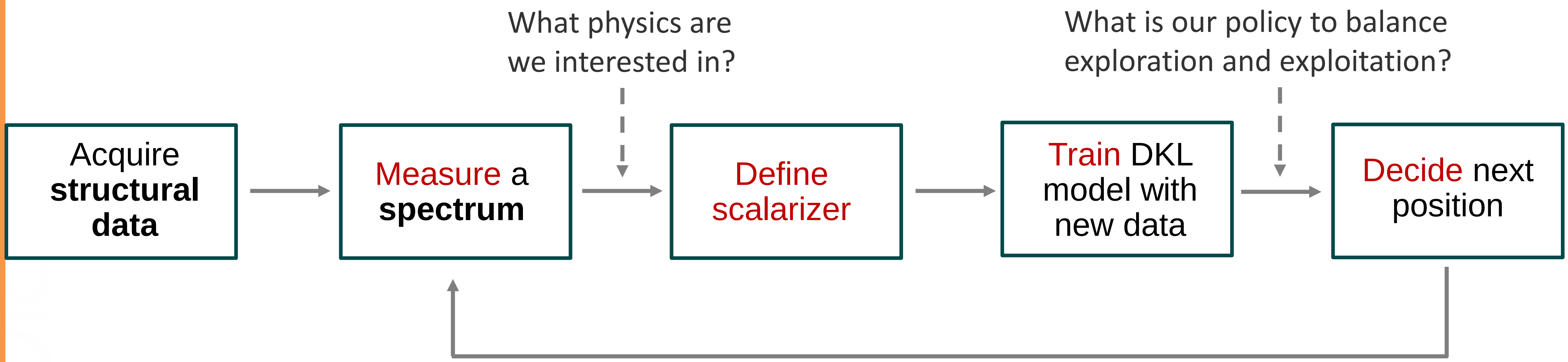

- Adjustable policies
- One step planning
- Can be implemented via Bayesian workflows
- Can be human in the loop

**Multistage discovery**




- Adjustable policies
- Multi-step planning
- Requires heuristic to start
- Requires **reward function**

# Deep Kernel Learning based BO

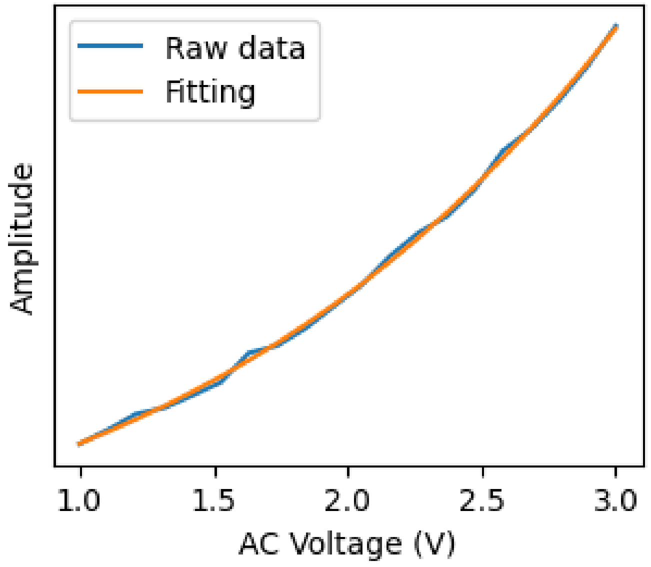


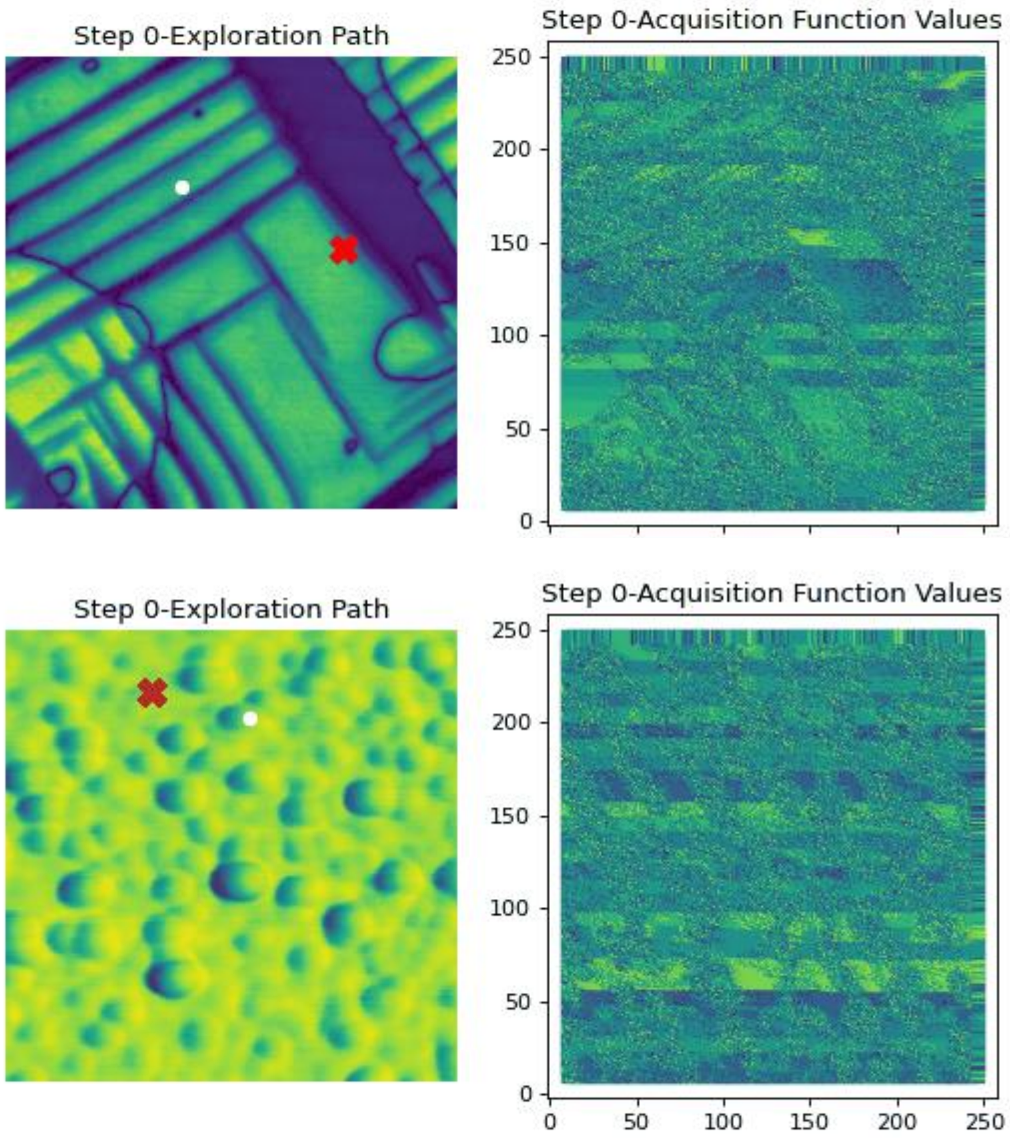

**Key concepts:**

- **Scalarizer:** (any) function that transforms spectrum into measure of interest. Can be integration over interval, parameters of a peak fit, ration of peaks, or more complex analysis

- **Experimental trace:** collection of image patches and associated spectra acquired during experiment. Note that we collect spectra, not only scalarizers

# 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!**

- In conventional microscopy experiment, human runs everything directly – defines scan, positions the probe, defines measurement parameters.

- In AE SPM, the policies are defined before the experiment and do not change. Sometimes it works – but not always.

- How would we:
  - (a) explain the AE progression after the experiment and
  - (b) control it during the experiment ?

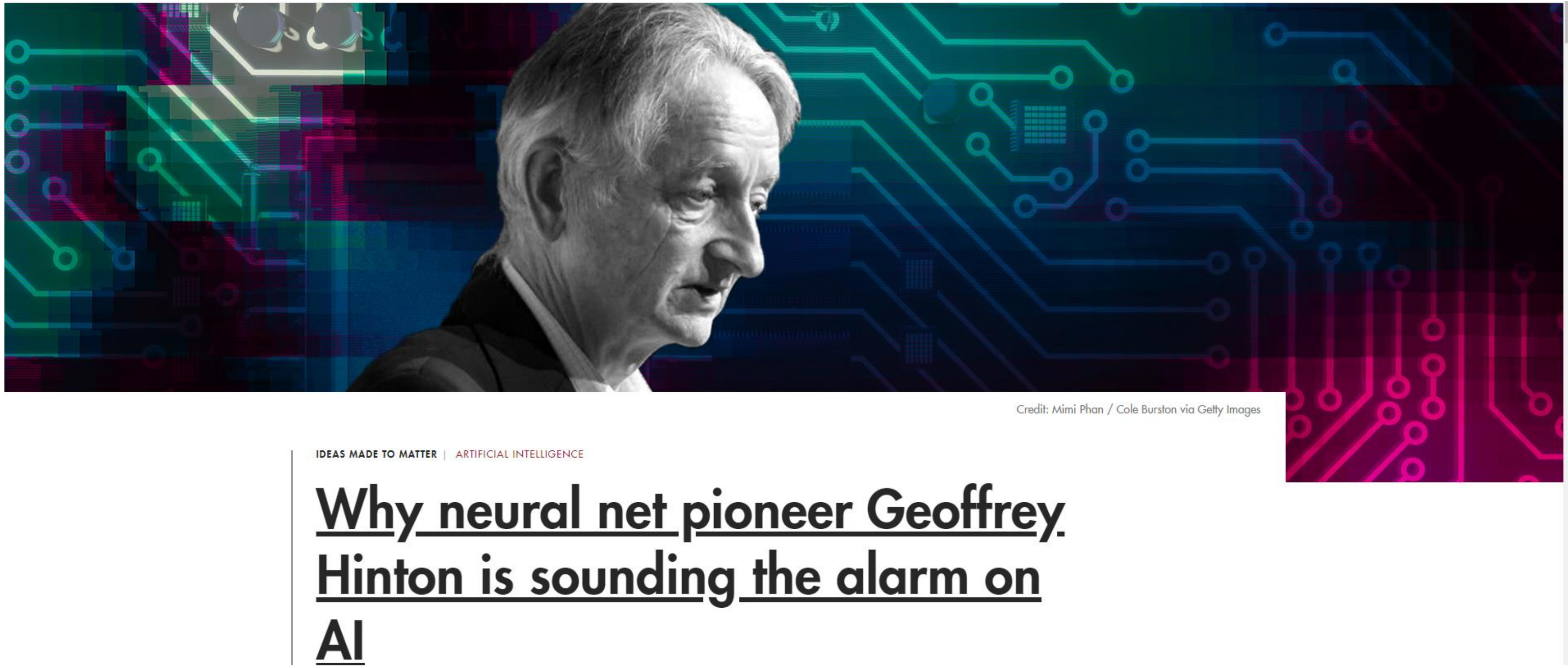
# Taking the Human Out of the Loop: A Review of Bayesian Optimization

## Citation

Shahriari, Bobak, Kevin Swersky, Ziyu Wang, Ryan P. Adams, and Nando de Freitas. 2016. "Taking the Human Out of the Loop: A Review of Bayesian Optimization." Proc. IEEE 104 (1) (January): 148–175. doi:10.1109/jproc.2015.2494218.

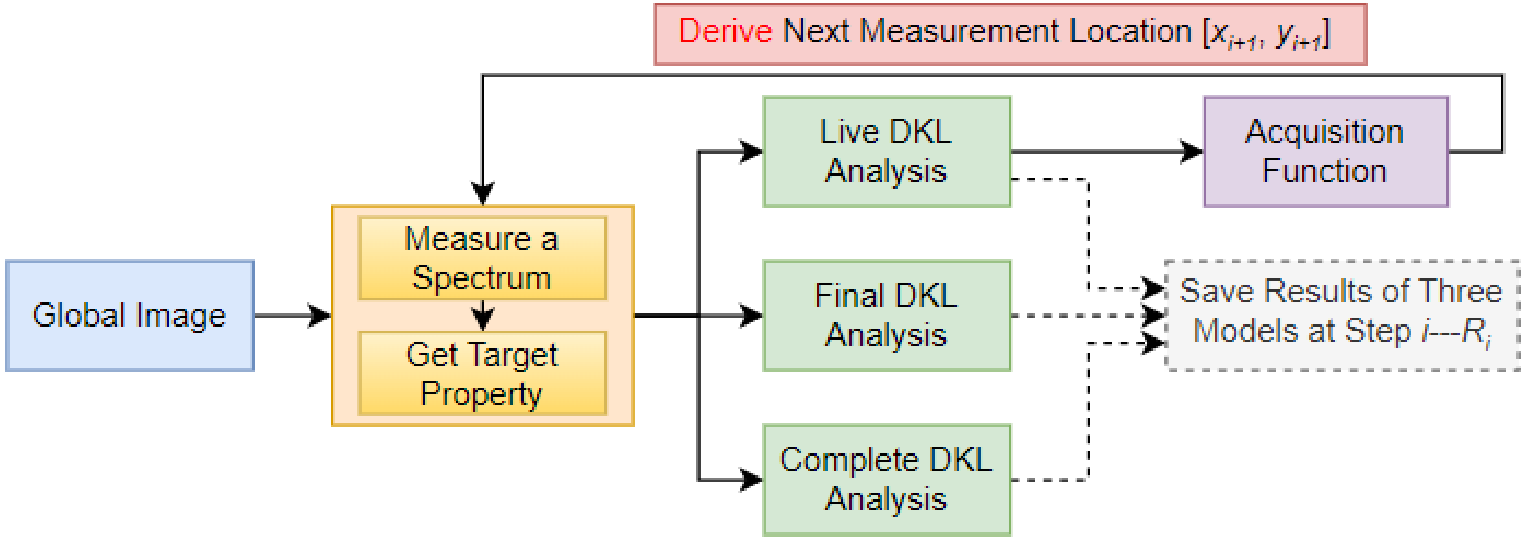
## Published Version

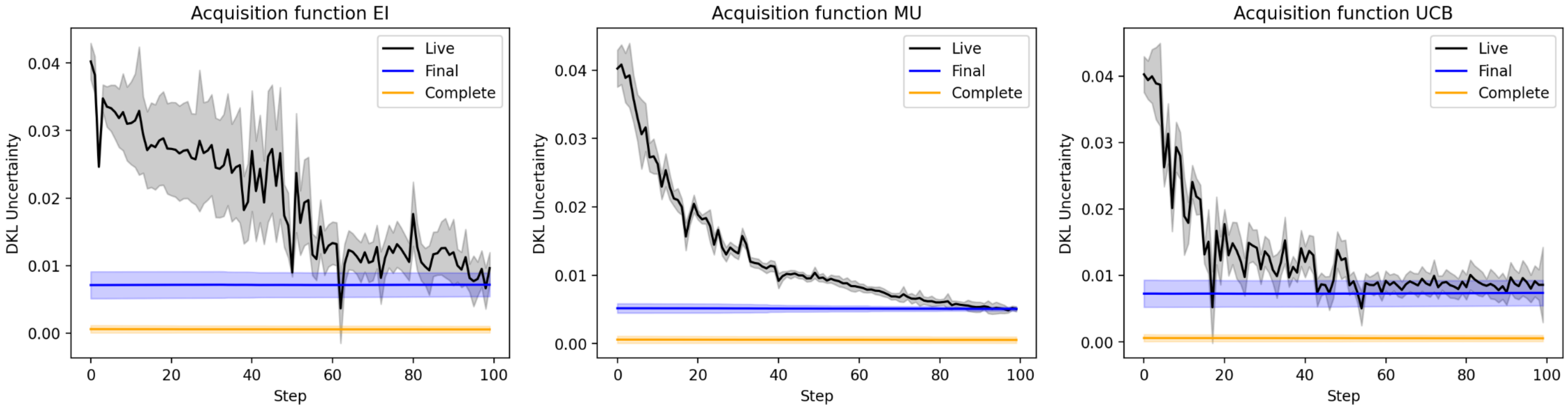
doi:10.1109/JPROC.2015.2494218



# Explainable AE

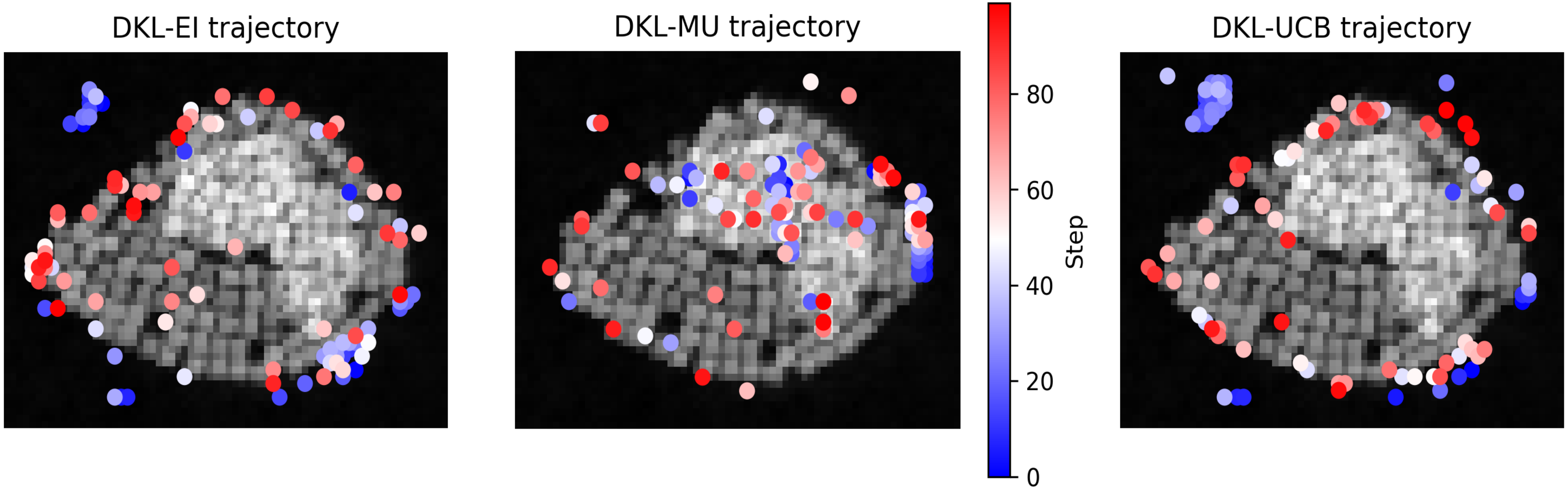
- During the AE, model learns structure-property relationships.
- What if we retrace the experimental steps – using the fully trained model?




U. Pratiush, K.M. Roccapriore, Y. Liu, G. Duscher, M. Ziatdinov, S.V. Kalinin, *Building Workflows for Interactive Human in the Loop Automated Experiment (hAE) in STEM-EELS*, arXiv:2404.07381

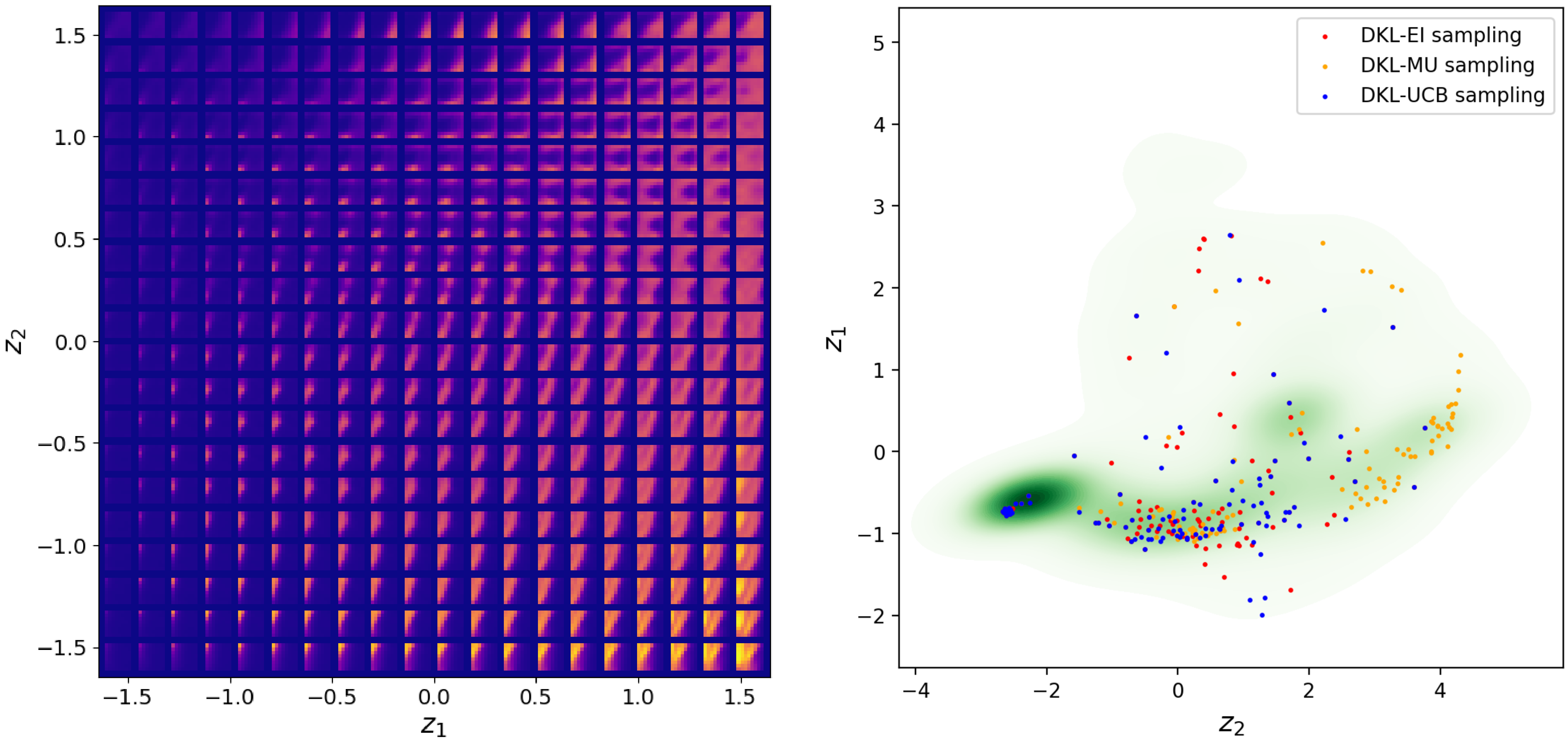
# Monitoring the AE




- Different acquisition functions (policies) give different experimental paths for AE
- Can we analyze what is special about points visited?

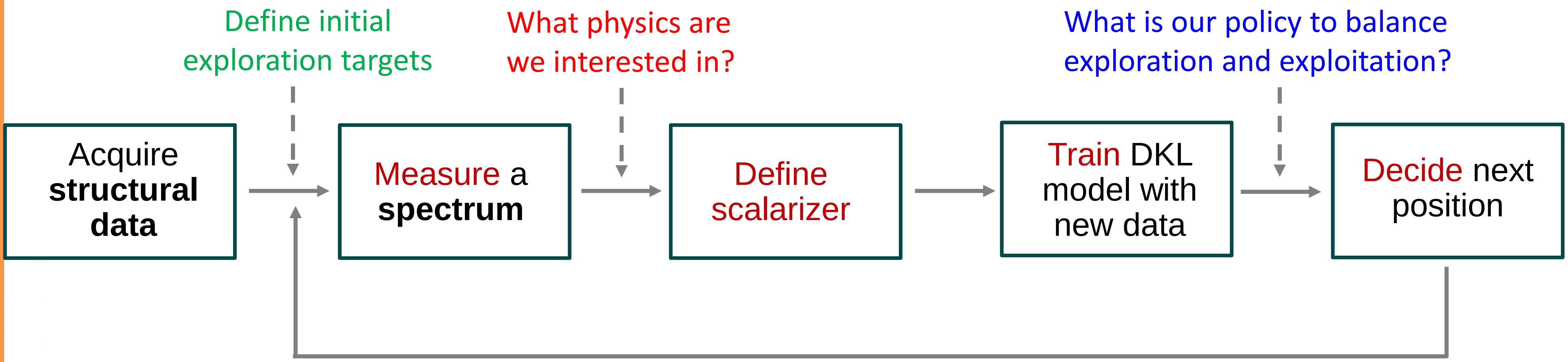
U. Pratiush, K.M. Roccapriore, Y. Liu, G. Duscher, M. Ziatdinov, S.V. Kalinin, *Building Workflows for Interactive Human in the Loop Automated Experiment (hAE) in STEM-EELS*, arXiv:2404.07381

# VAE approach: full feature space




U. Pratiush, K.M. Roccapriore, Y. Liu, G. Duscher, M. Ziatdinov, S.V. Kalinin, *Building Workflows for Interactive Human in the Loop Automated Experiment (hAE) in STEM-EELS*, arXiv:2404.07381

# Bringing Human into the Loop

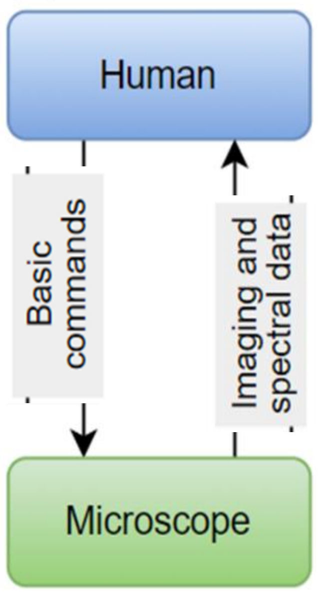


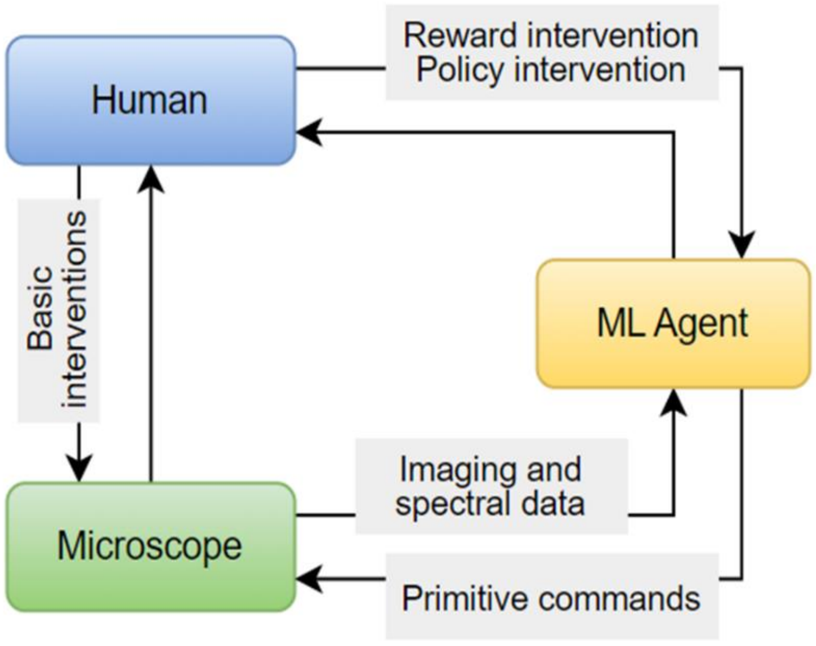

**We can intervene on:**

- **Policies** (acquisition functions): type and parameters
- **Scalarizers**: what physics are we interested in - type and parameters
- **Knowledge injection:** what microstructures are we interested in?
- **Cost and latencies:** trivial via acquisition functions

U. Pratiush, K.M. Roccapriore, Y. Liu, G. Duscher, M. Ziatdinov, S.V. Kalinin, *Building Workflows for Interactive Human in the Loop Automated Experiment (hAE) in STEM-EELS*, arXiv:2404.07381

# Human in the loop AE



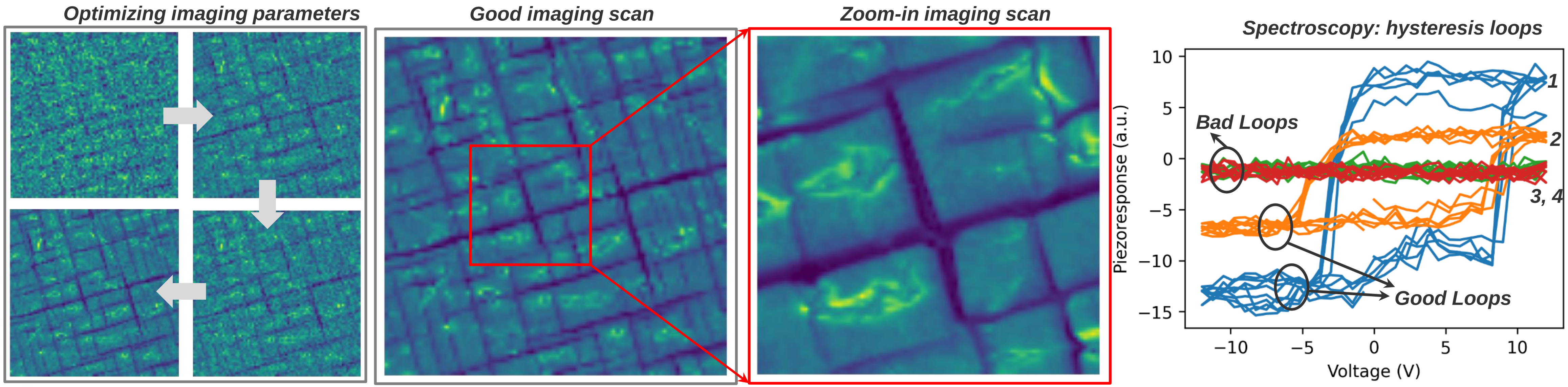



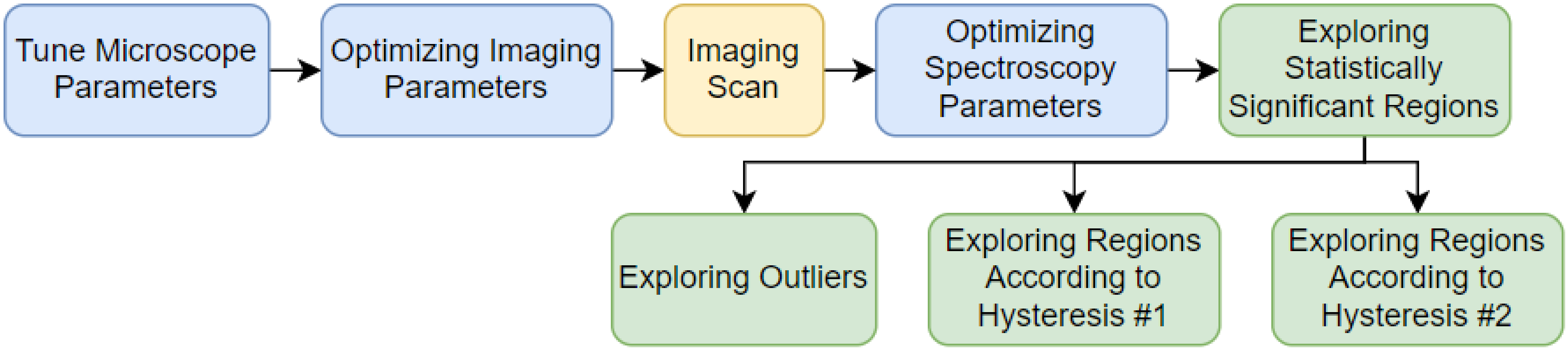


**We can intervene on:**

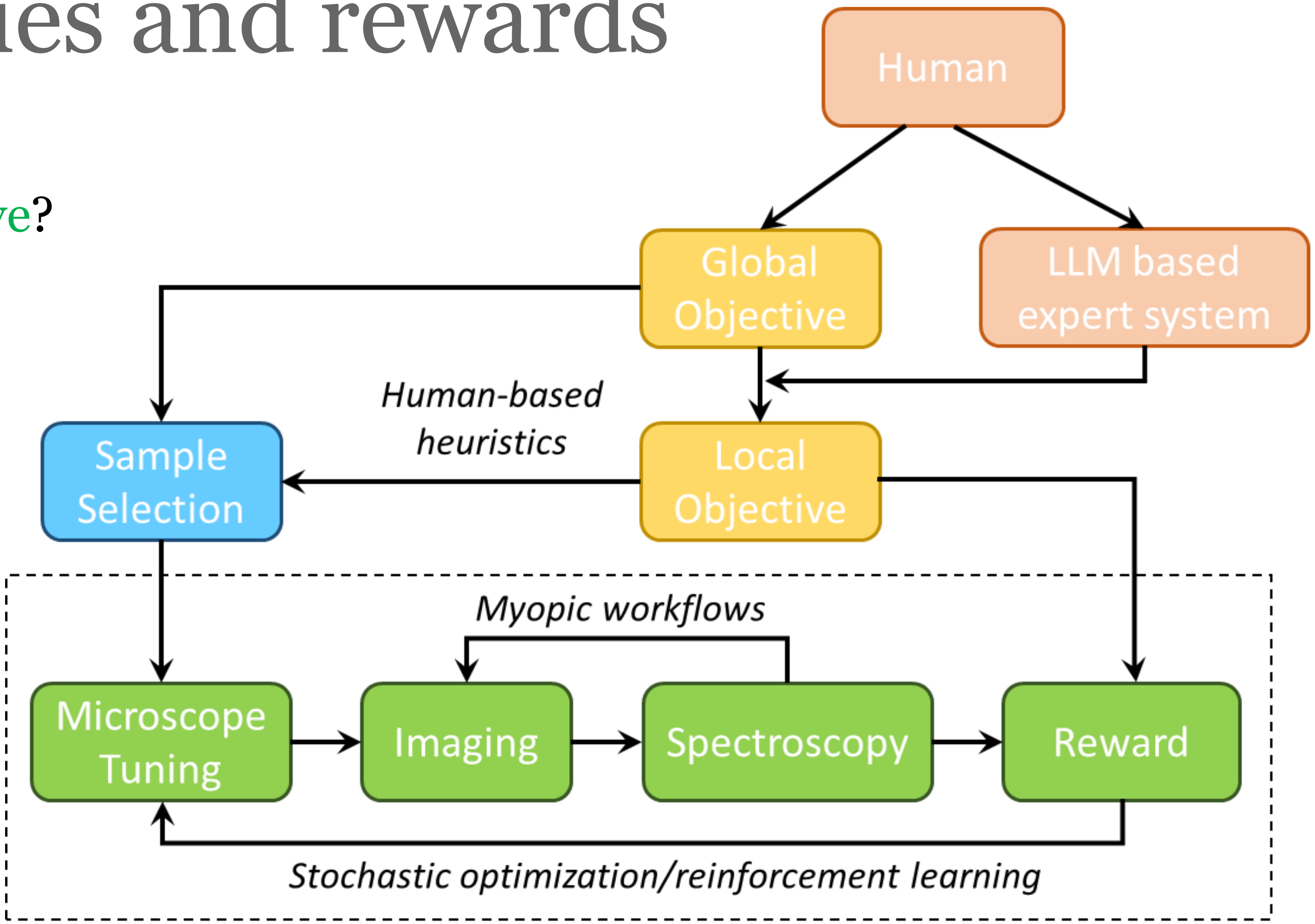
- Policies (acquisition functions): type and parameters
- Scalarizers (physics descriptors): type and parameters
- Knowledge injection
- Direct operation

# Future: full workflow optimization



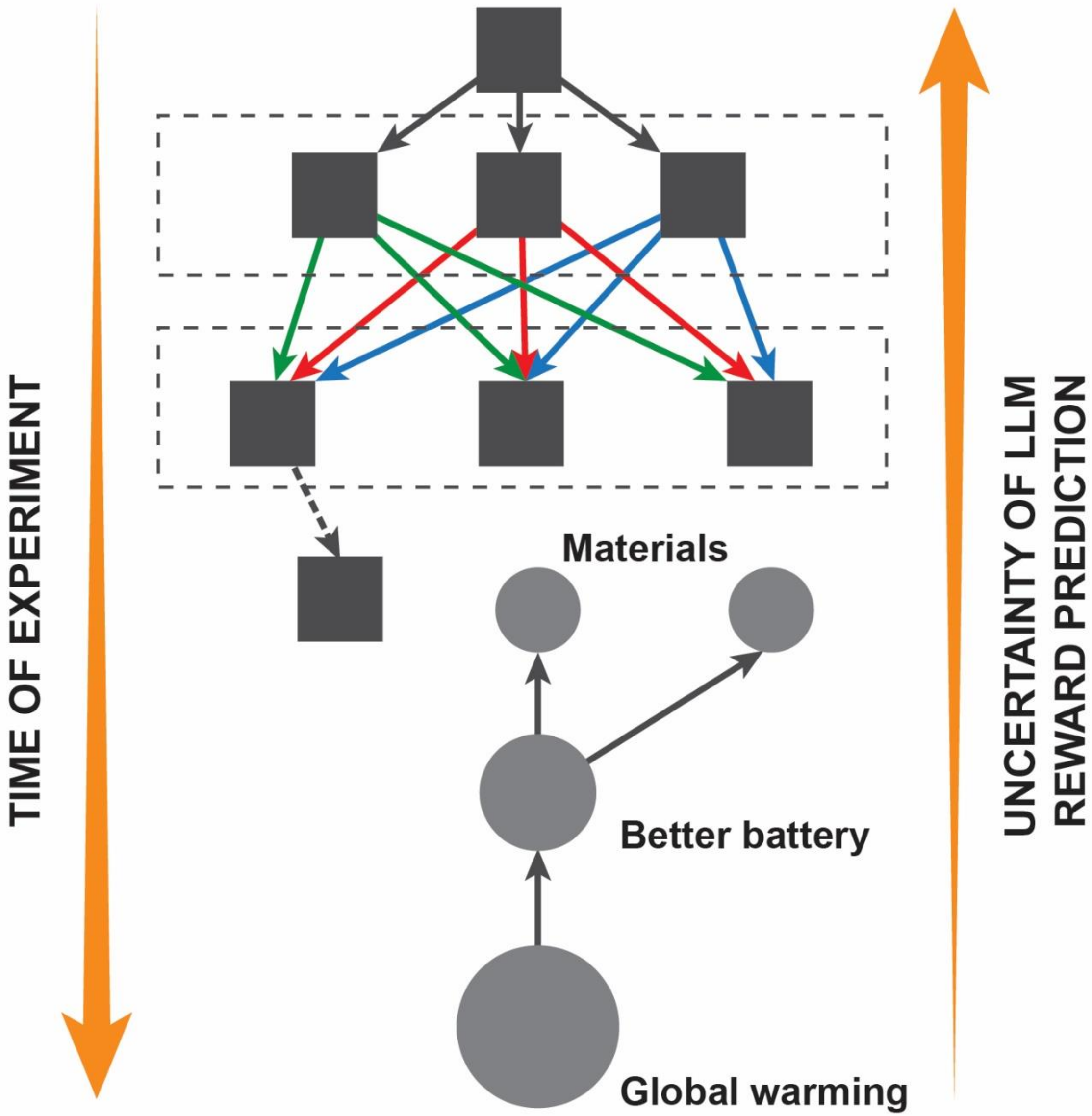
# The dance of policies and rewards




**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?




- **Big data** and associated infrastructure is useful and sometimes sufficient in well established fields with plenty of real (or perceived) downstream applications

- Automated discovery research necessitates creation of the **reward functions**, that maps discovery onto established optimization frameworks

- Large Language Models may be a viable strategy to discover probabilistic reward functions that can be refined experimentally arXiv > physics > arXiv:2304.05332

Search...

Help |

Physics > Chemical Physics

[Submitted on 11 Apr 2023]

## Emergent autonomous scientific research capabilities of large language models

Daniil A. Boiko, Robert MacKnight, Gabe Gomes

Transformer-based large language models are rapidly advancing in the field of machine learning research, with applications spanning natural language, biology, chemistry, and computer programming. Extreme scaling and reinforcement learning from human feedback have significantly improved the quality of generated text, enabling these models to perform various tasks and reason about their choices. In this paper, we present an Intelligent Agent system that combines multiple large language models for autonomous design, planning, and execution of scientific experiments. We showcase the Agent's scientific research capabilities with three distinct examples, with the most complex being the successful performance of catalyzed cross-coupling reactions. Finally, we discuss the safety implications of such systems and propose measures to prevent their misuse.

Comments: Version 1, April 11, 2023. 48 pages
Subjects: **Chemical Physics (physics.chem-ph)**; Computation and Language (cs.CL)
Cite as: arXiv:2304.05332 **[physics.chem-ph]**
(or arXiv:2304.05332v1 **[physics.chem-ph]** for this version)
https://doi.org/10.48550/arXiv.2304.05332

Computer Science > Human-Computer Interaction

[Submitted on 24 Jan 2024]

## Synergizing Human Expertise and AI Efficiency with Language Model for Microscopy Operation and Automated Experiment Design

Yongtao Liu, Marti Checa, Rama K. Vasudevan

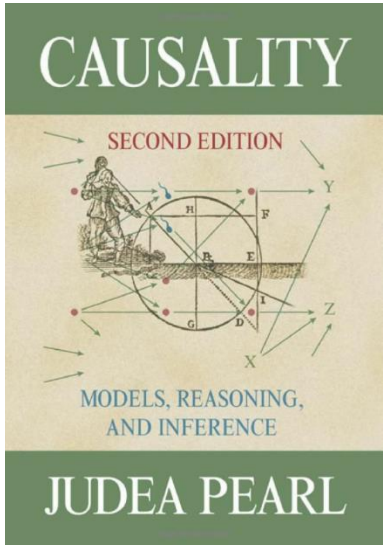
With the advent of large language models (LLMs), in both the open source and proprietary domains, attention is turning to how to exploit such artificial intelligence (AI) systems in assisting complex scientific tasks, such as material synthesis, characterization, analysis and discovery. Here, we explore the utility of LLM, particularly ChatGPT4, in combination with application program interfaces (APIs) in tasks of experimental design, programming workflows, and data analysis in scanning probe microscopy, using both in-house developed API and API given by a commercial vendor for instrument control. We find that the LLM can be especially useful in converting ideations of experimental workflows to executable code on microscope APIs. Beyond code generation, we find that the GPT4 is capable of analyzing microscopy images in a generic sense. At the same time, we find that GPT4 suffers from inability to extend beyond basic analyses or more in-depth technical experimental design. We argue that a LLM specifically fine-tuned for individual scientific domains can potentially be a better language interface for converting scientific ideations from human experts to executable workflows, such a synergy between human expertise and LLM efficiency in experimentation can open new door for accelerating scientific research, enabling effective experimental protocols archive and sharing in scientific community.

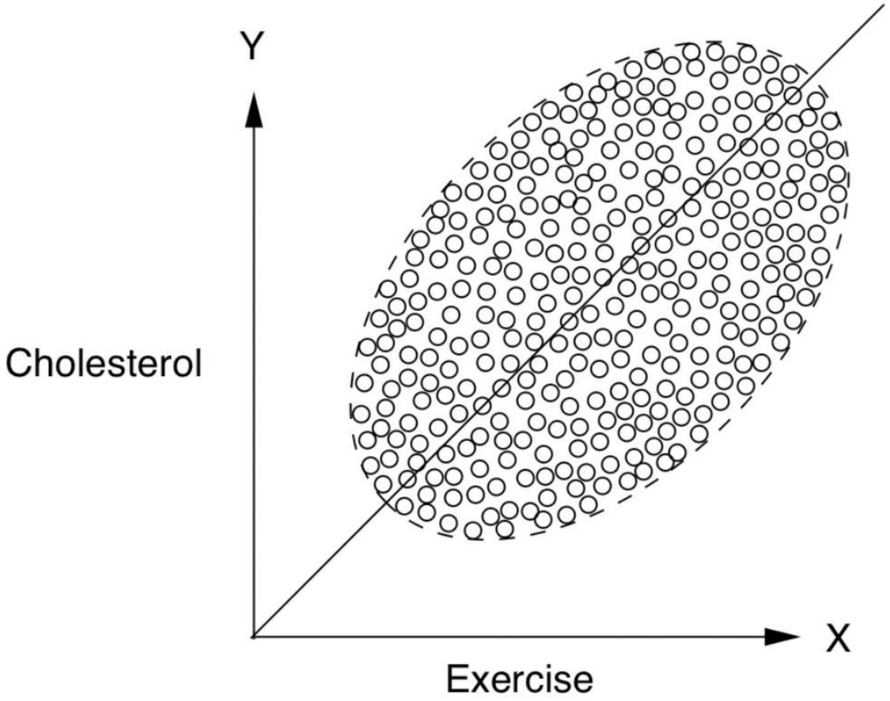
Comments: 16 pages; 7 figures
Subjects: **Human-Computer Interaction (cs.HC)**; Materials Science (cond-mat.mtrl-sci)
Cite as: arXiv:2401.13803 **[cs.HC]**
(or arXiv:2401.13803v1 **[cs.HC]** for this version)
https://doi.org/10.48550/arXiv.2401.13803

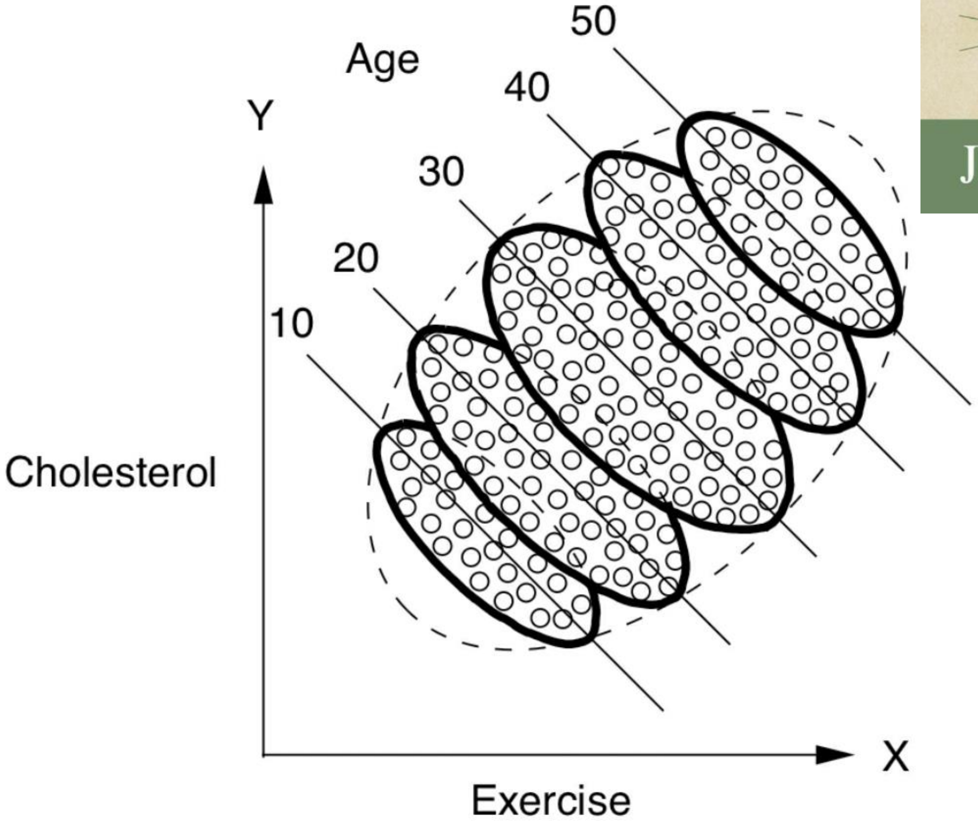
# What’s wrong with correlation?

**Simpson paradox:**

- Exercise is bad for cholesterol
- Some drugs are bad for men and women, but good for people in general
- … and many more




- Exercise is good for cholesterol
- People try to exercise more to reduce it
- But it is not enough…

# Causal knowledge is crucial!

**If the causal link is well known, ML is the tool:**

- Atom finding via U-Nets
- All machine learning applications in theory

**If the causal link is known and confounders are "frozen"**

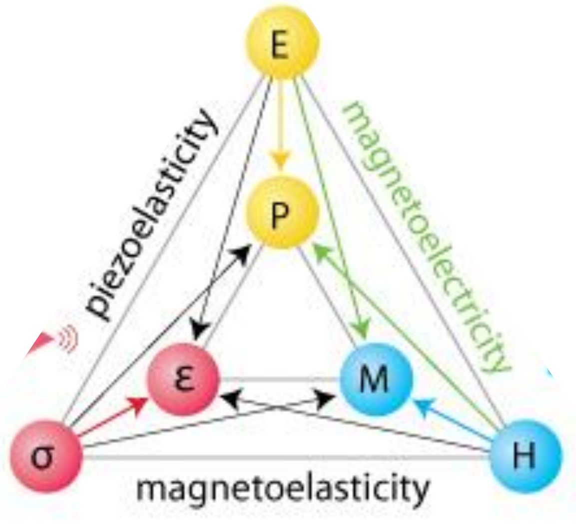
- Materials synthesis-property relations
- Some experimental based materials predictions

**If the causal links are multiple and unknown**

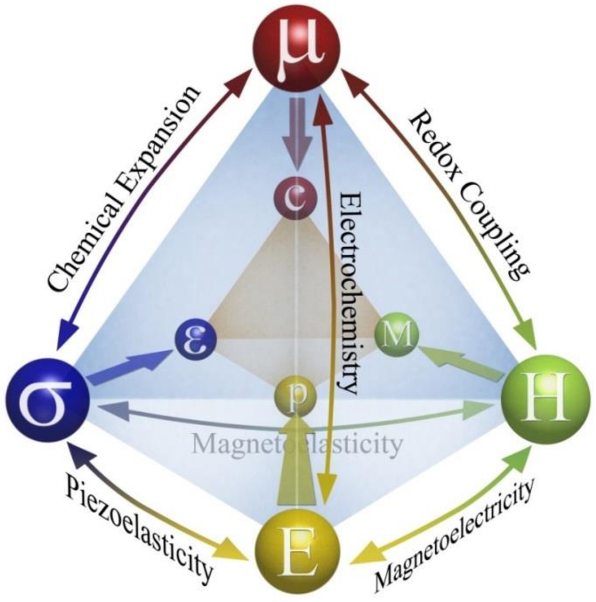
- There be dragons


https://www.gpsworld.com/here-there-be-dragons-gis-explores-the-unknown/

# Materials with Coupled Functionalities




Spaldin & Fiebig,
*Science* **309**:391 (2005).




Chemistry is confounder!

Jesse et al, *MRS Bull*. 2012
Kalinin and Spaldin, *Science* 2013

1. **Oxidation states:** induce metal insulator transition, control charge compensation
2. **Molar volume:** effect similar to chemical pressure in bulk phase diagram
3. **Molar volume:** strain compensation at defects
4. **Crystal field effects:** changing environment of cation

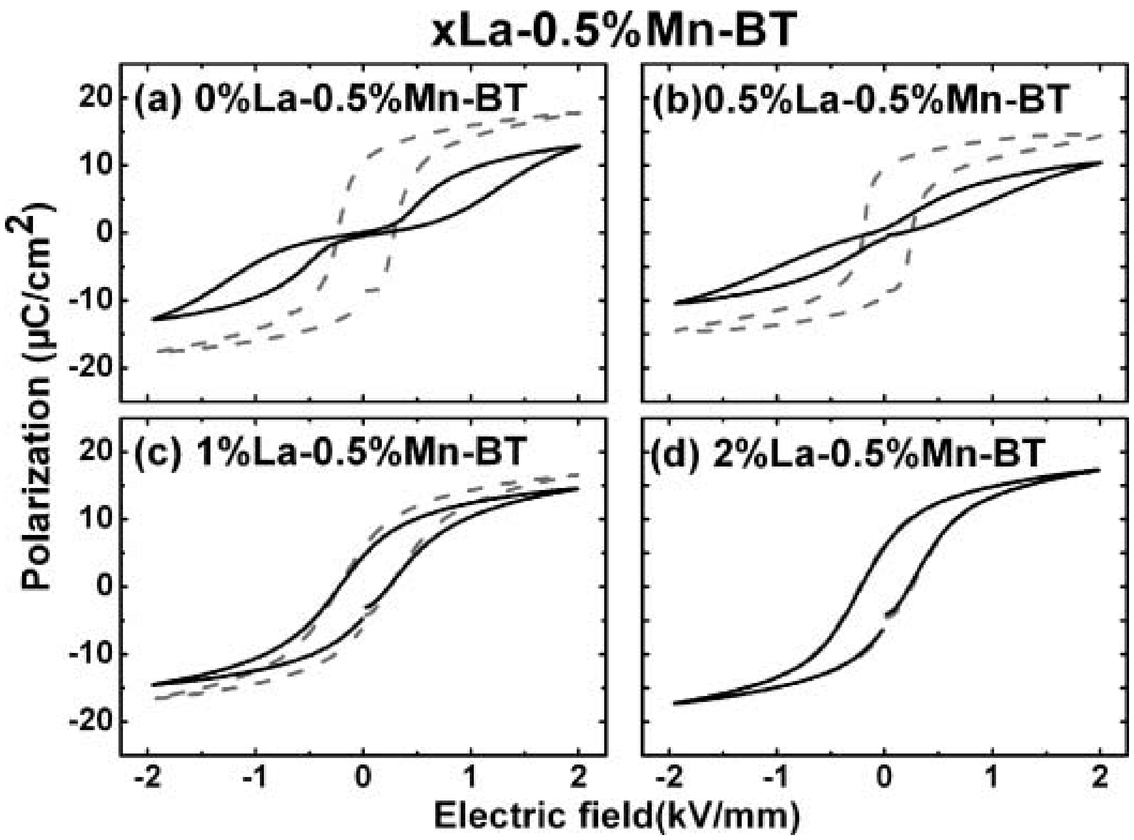
# Cause and Effect in Ferroelectrics

For ferroelectric, we generally assume that cationic order is frozen at the state of material formation, and then polarization field evolves to accommodate average polarization instability and local pinning.

However, we know that ions can move to compensate polarization – segregation at the domain walls, memory effects, etc.

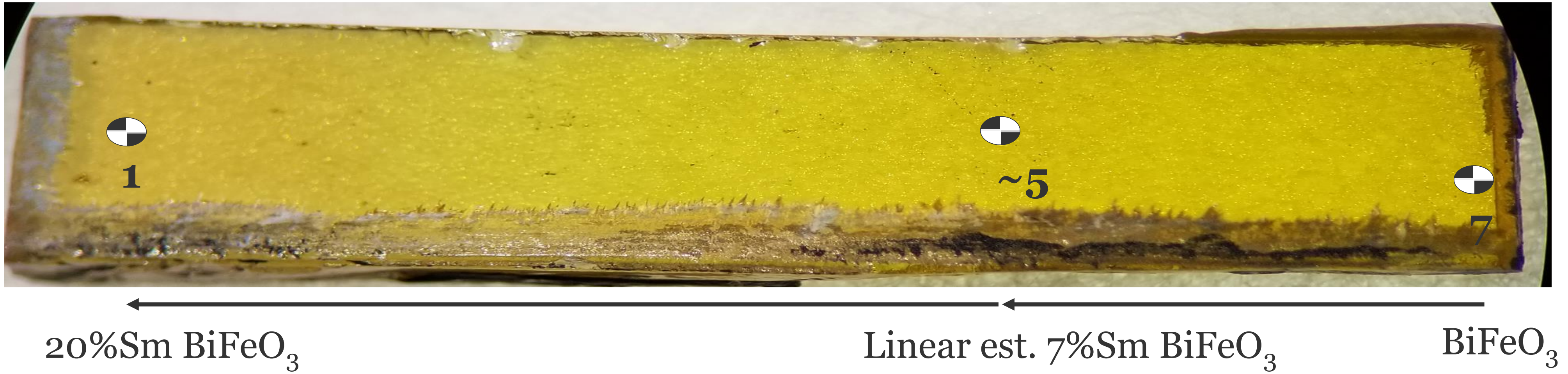
For real material, can we establish what is the cause and what is the effect:

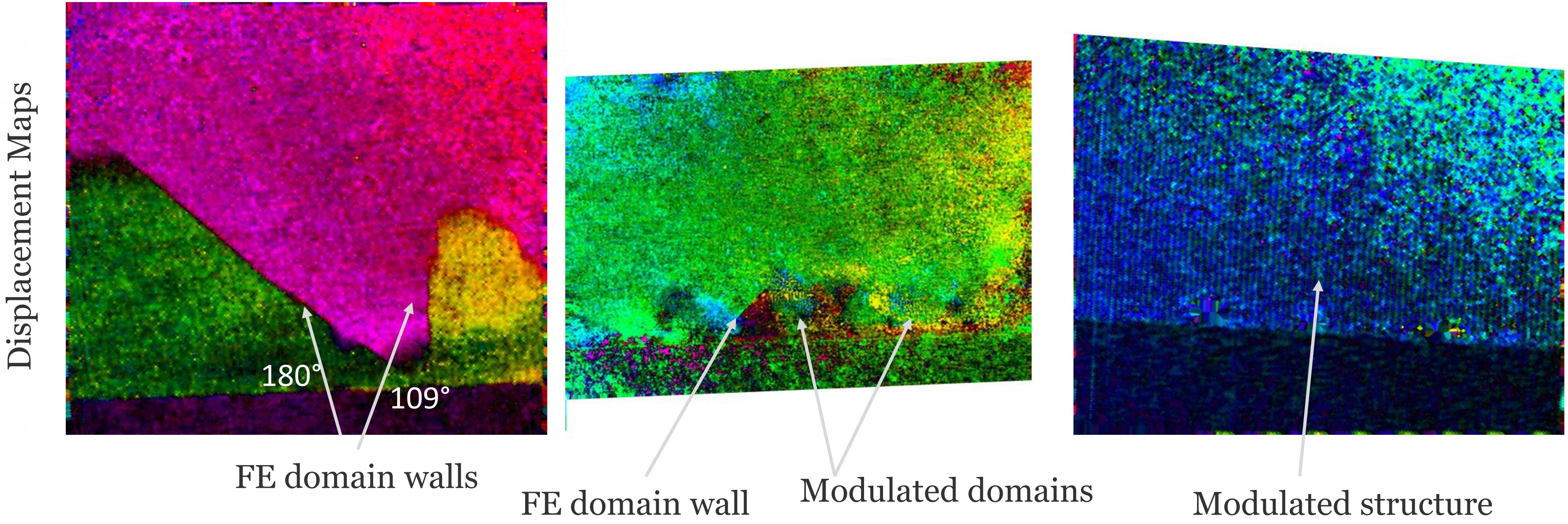
- Does polarization align to the cationic disorder
- Or does polarization instability drive cationic disorder?

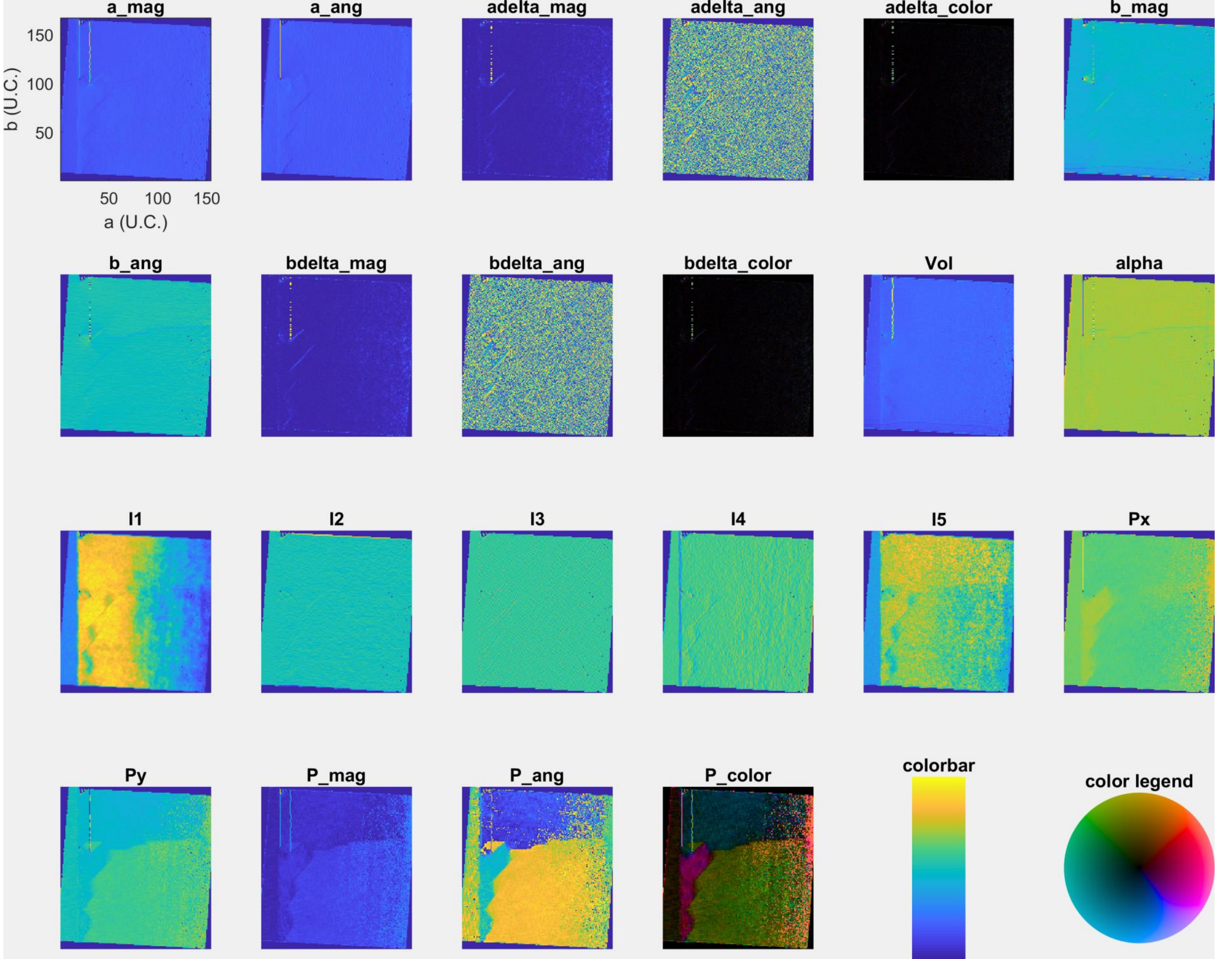



https://www.researchgate.net/figure/Control-of-ferroelectric-aging-in-La-Mn-hybrid-doped-BaTiO-3-polycrystals-a-b-c_fig1_233237569

# STEM of Combinatorial Libraries

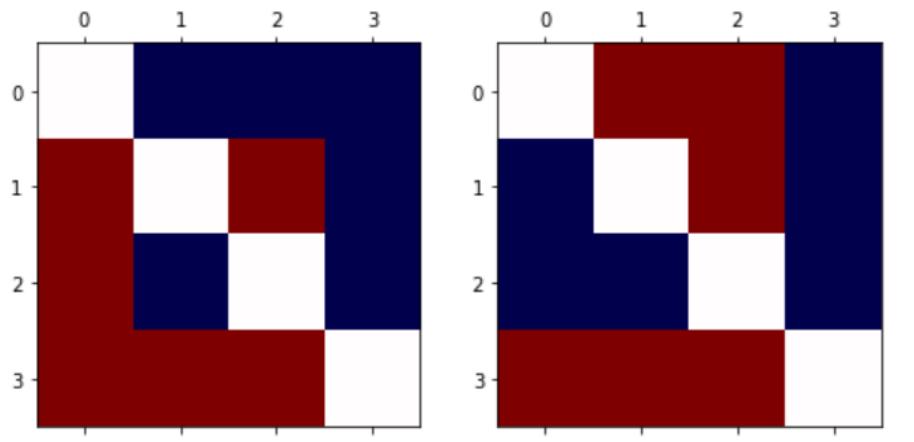


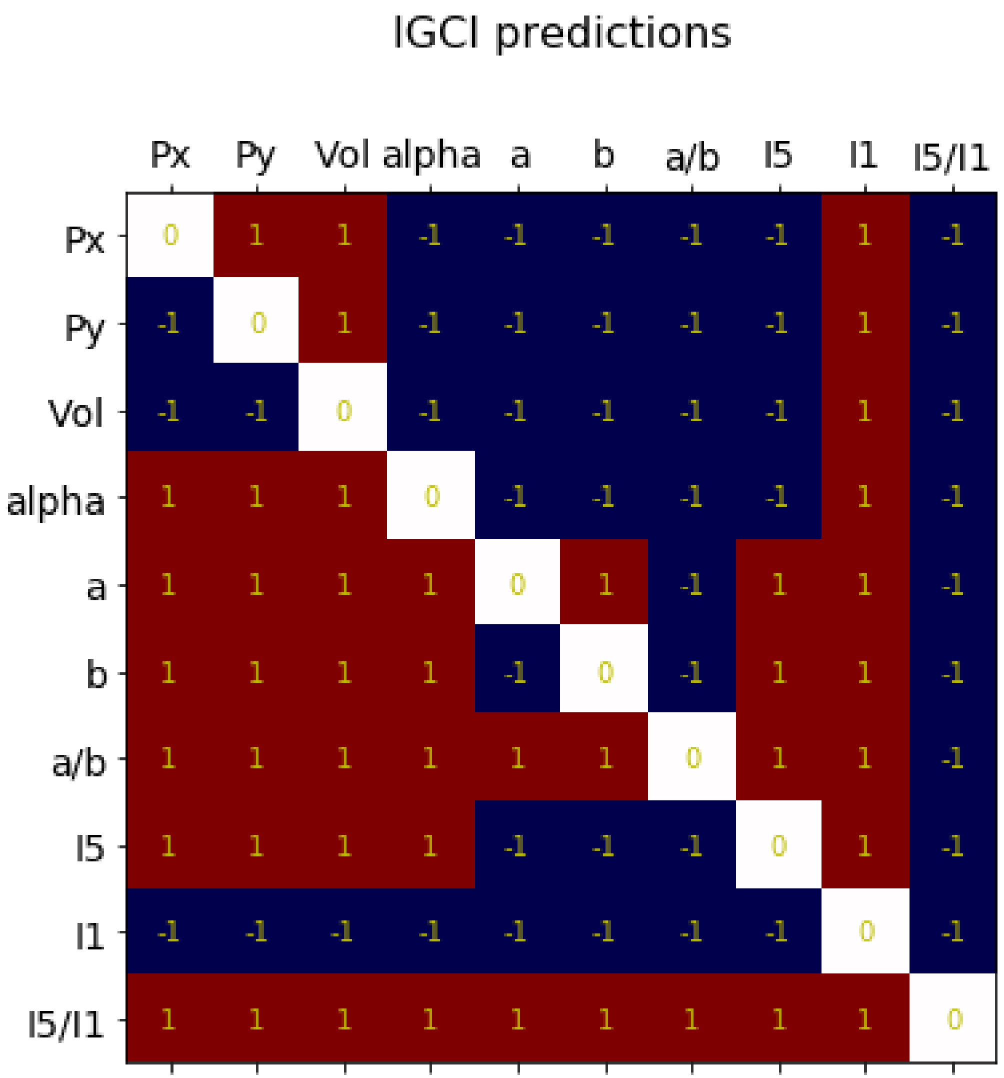



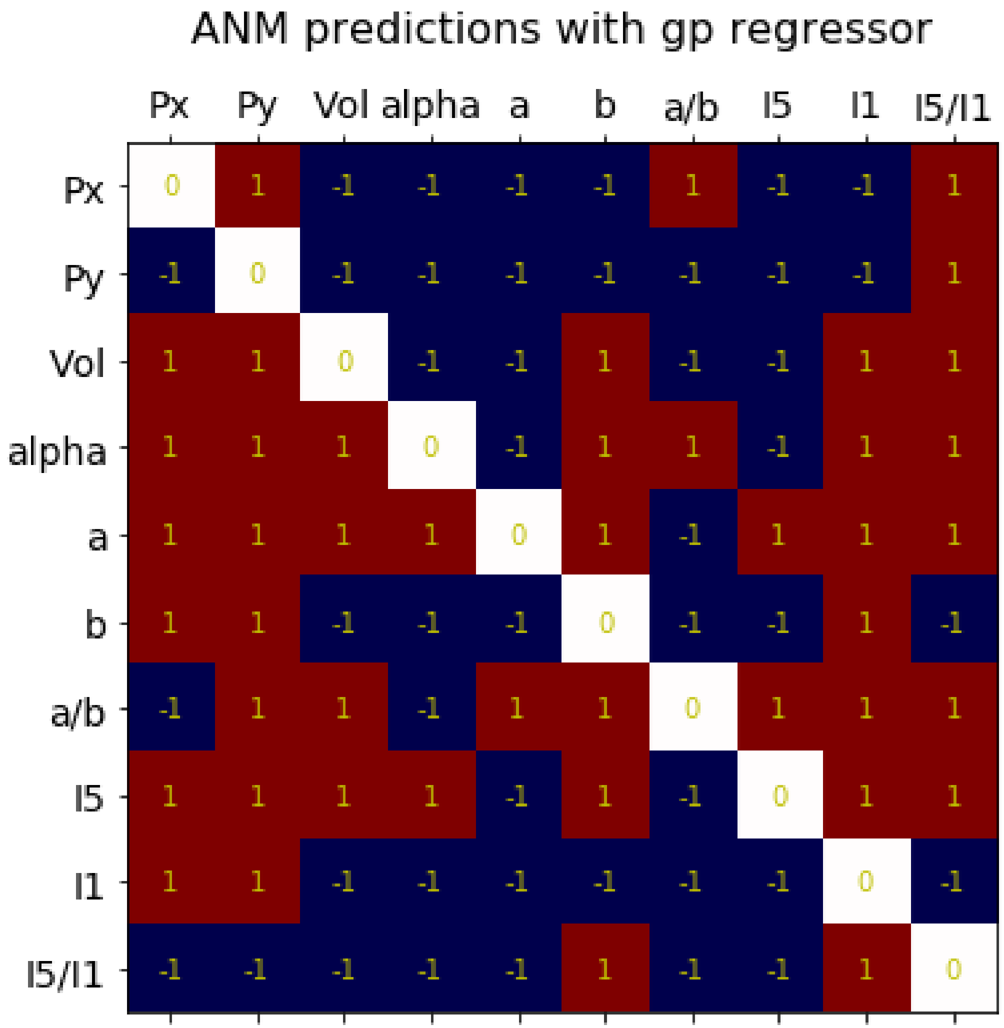

a_mag
a_ang
adelta_mag
adelta_ang
adelta_color
b_mag
150
100
50
b (U.C.)
50
100
150
a (U.C.)
b_ang
bdelta_mag
bdelta_ang
bdelta_color
Vol
alpha
I1
I2
I3
I4
I5
Px
Py
P_mag
P_ang
P_color
colorbar
color legend

ferroic_causal.ipynb

```python
# IGCI model
caus_mat1[caus_mat1>0] = 1 # y --> x
caus_mat1[caus_mat1<0] = -1 # x --> y
_, ax = plt.subplots(1, 2, figsize=(8, 5))
ax[0].matshow(caus_mat1, cmap='seismic')
# RESIT model
caus_mat2[caus_mat2>0] = 1 # y --> x
caus_mat2[caus_mat2<0] = -1 # x --> y
ax[1].matshow(caus_mat2, cmap='seismic')
plt.show()
```





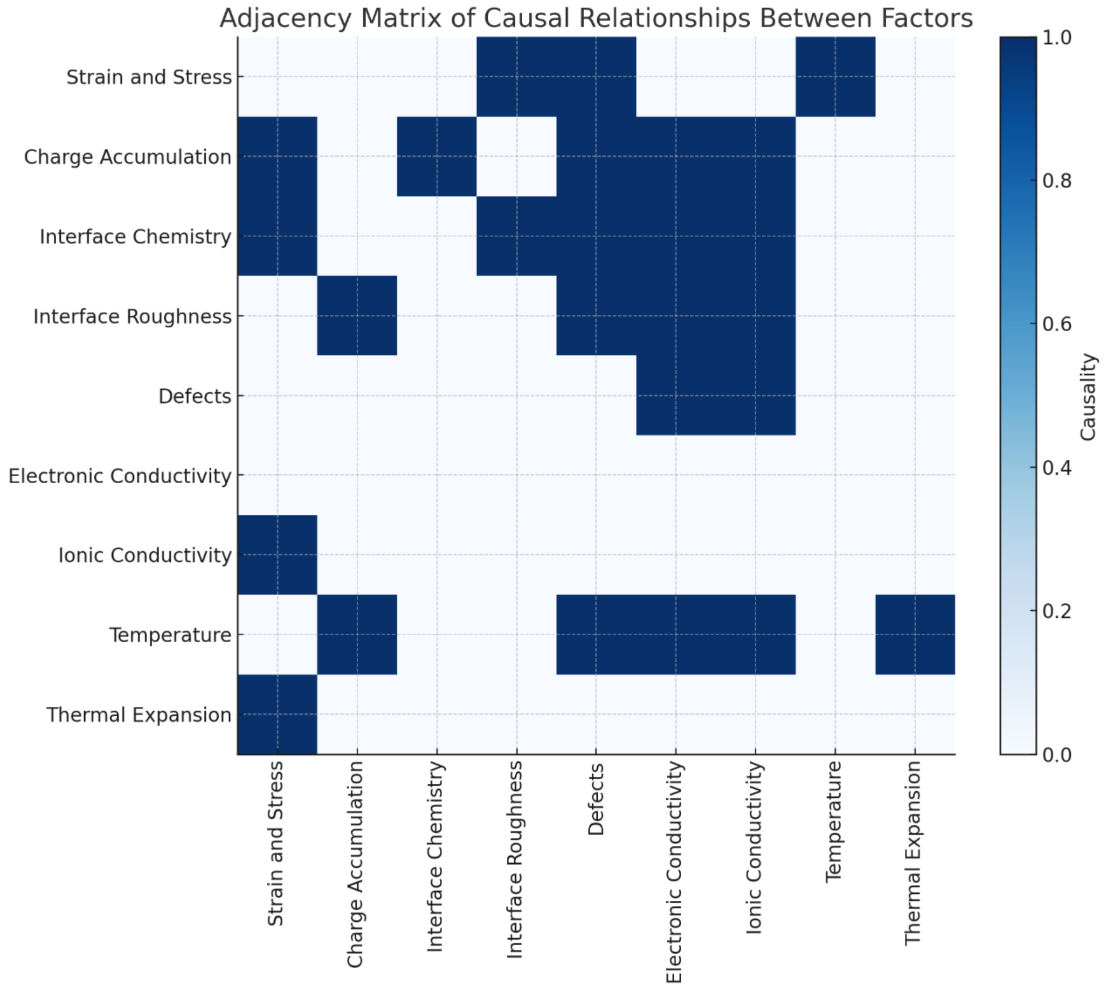




```
['I1' 'Vol' 'Py' 'Px' 'alpha' 'I5' 'b' 'a' 'a/b' 'I5/I1']
['Py' 'I1' 'I5/I1' 'Px' 'b' 'Vol' 'alpha' 'a/b' 'I5' 'a']
```

# LLM Co-Scientists: Causal Discovery

```python
llm = ChatOpenAI(temperature=0, model='gpt-3.5-turbo')
tools = load_tools(["arxiv"], llm=llm)
agent = initialize_agent(tools, llm, agent=AgentType.CHAT_ZERO_SHOT_REACT_DESCRIPTION,
    handle_parsing_errors=True, verbose=False)
```

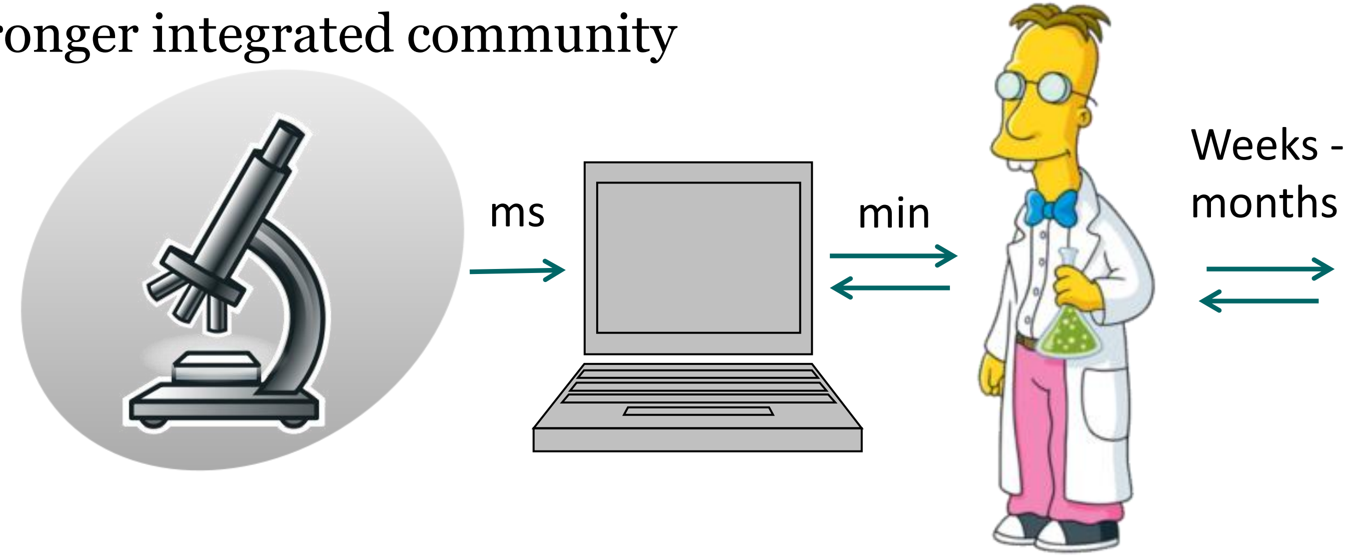
```python
def get_llm_info(llm, agent, var_1, var_2):

    out = agent(f"Does {var_1} cause {var_2} or the other way around?\
    We assume the following definition of causation:\
    if we change A, B will also change.\
    The relationship does not have to be linear or monotonic.\
    We are interested in all types of causal relationships, including\
    partial and indirect relationships, given that our definition holds.\
    ")

    print(out)

    pred = llm.predict(f'We assume the following definition of causation:\
    if we change A, B will also change.\
    Based on the following information: {out["output"]},\
    print (0,1) if {var_1} causes {var_2},\
    print (1, 0) if {var_2} causes {var_1}, print (0,0)\
    if there is no causal relationship between {var_1} and {var_2}.\
    Finally, print (-1, -1) if you don\'t know. Importantly, don\'t try to\
    make up an answer if you don\'t know.')

    print(pred)

    return pred
```

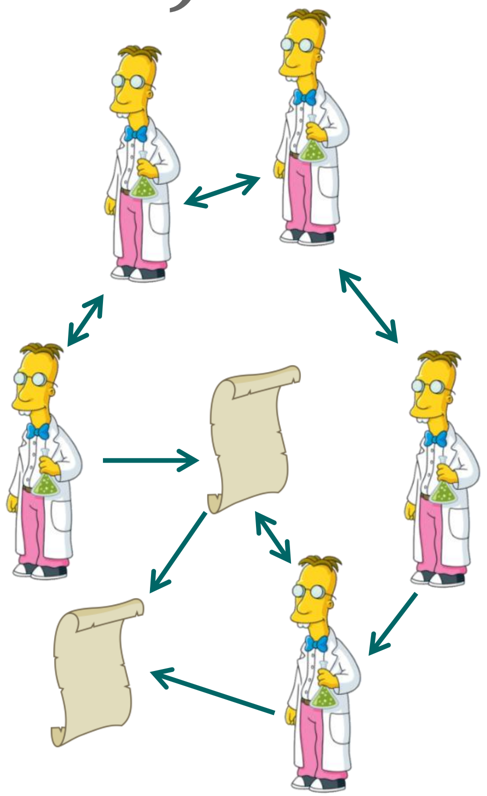



Large Language Models allow exploring body of literature via RAG to form objects (here, prior causal knowledge) that can be used to complement the data

# Classical Instrumental Research (2016)

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



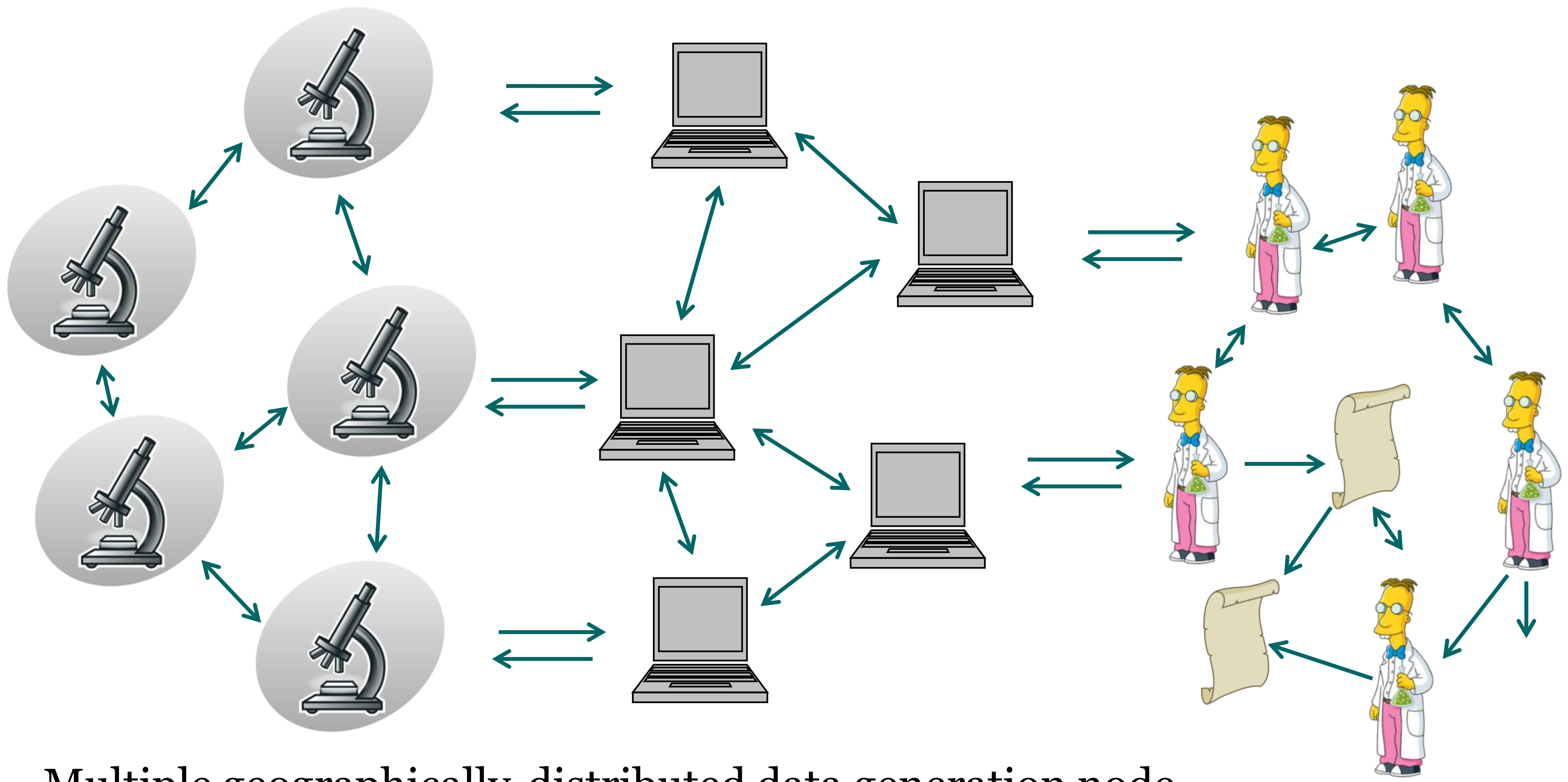



**Instrument** **Control/data acquisition** **Researcher** **Community**

- Social networking/education
- Publications/citations

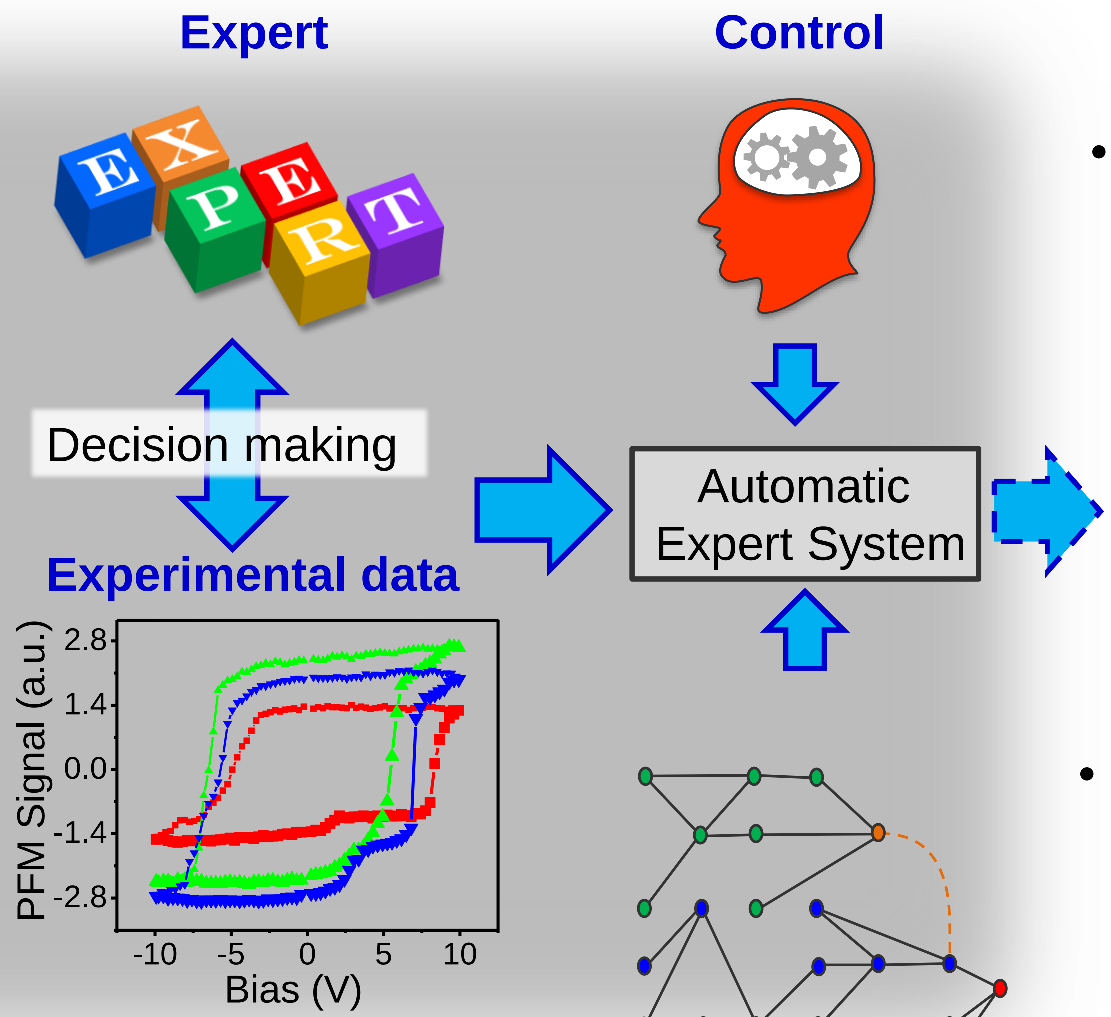
1. Only small fraction of data stream from the instrumentation is captured
2. Only small fraction of captured data is analyzed, interpreted, and put in the context
3. Human-machine interaction during acquisition is often slow and can be non-optimal
4. Human interpretation of data is limited: bias and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

# Step 1: Cloud Integration (2016)

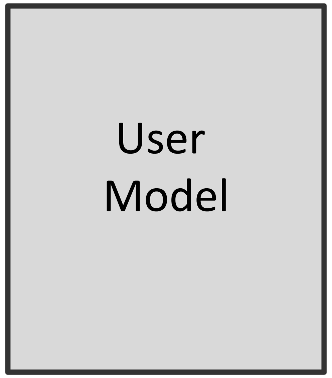


1. Multiple geographically-distributed data generation node
2. Full capture of instrumental data stream /compression/curation
3. Coordination of protocols and data/metadata across the cloud
4. Cloud-based processing and dimensionality reduction
5. Community-wide analytics

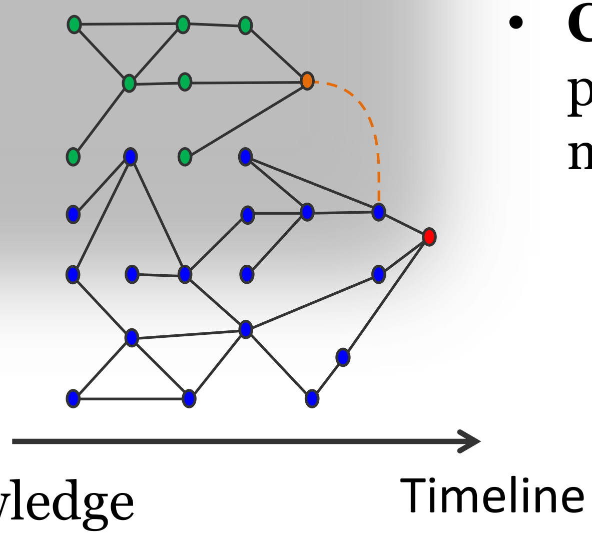
# Step 2: Cloud Analytics (2016)




- **Synthesis of expertise:** factor in human expert knowledge



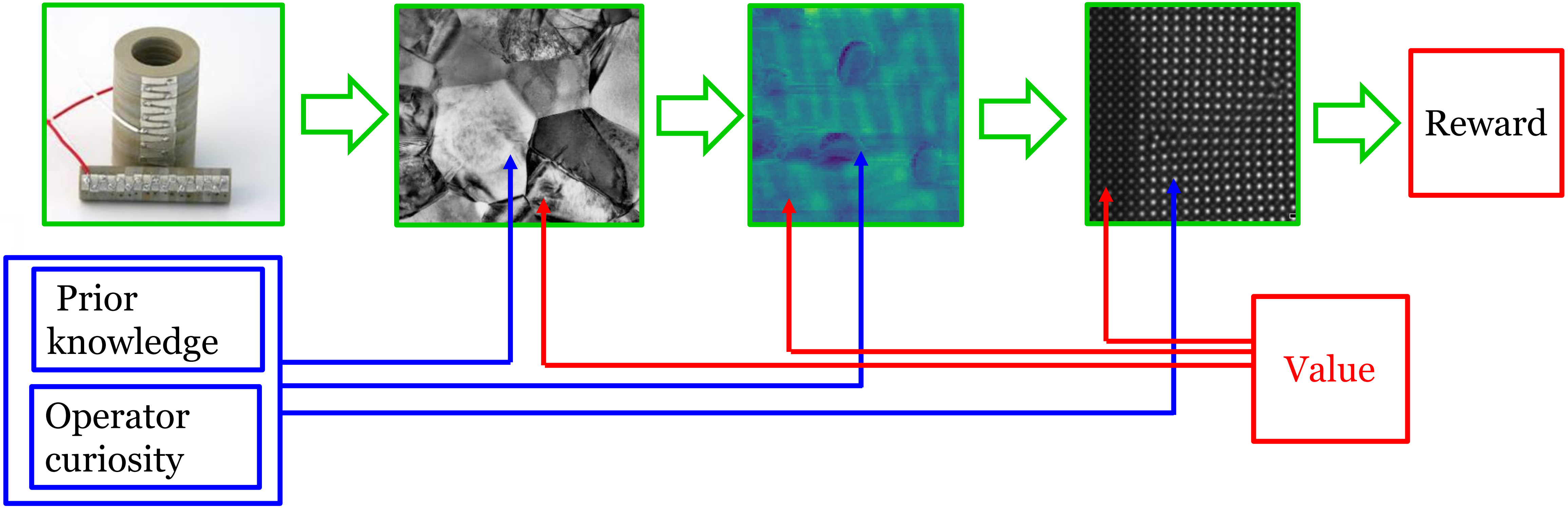

- **Context search:** published results data mining/social networks




**Future:**

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

# Step 3: Workflow Design (2022)




**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

# Reward Driven 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 (an AGI question).