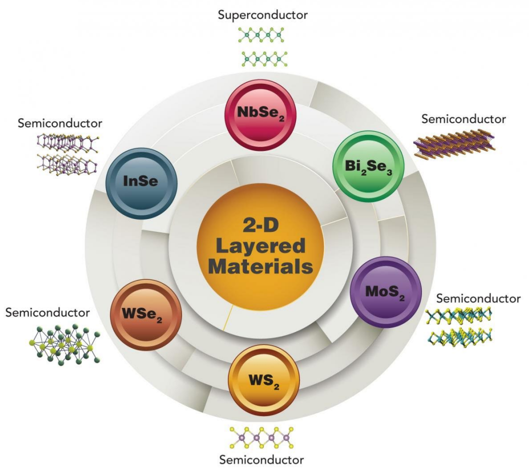
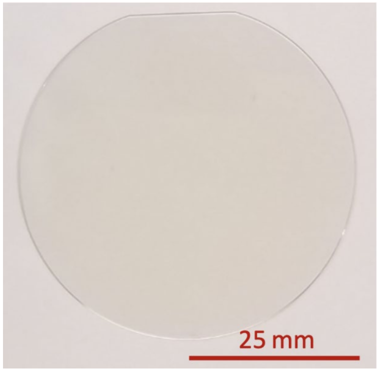
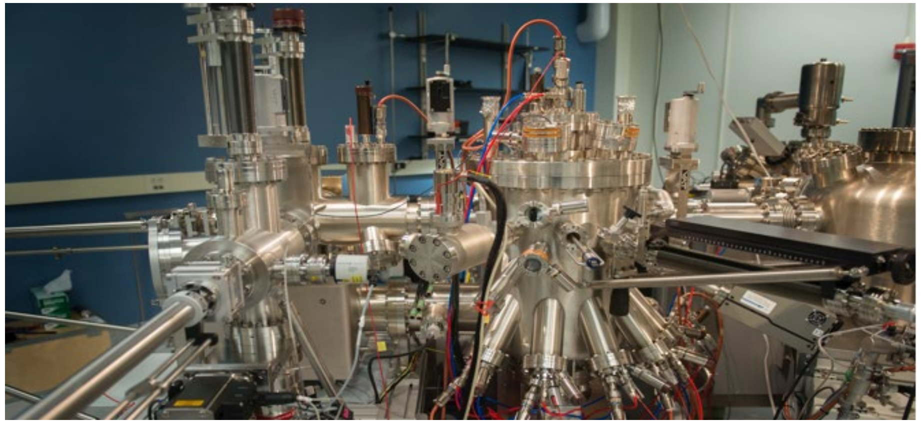


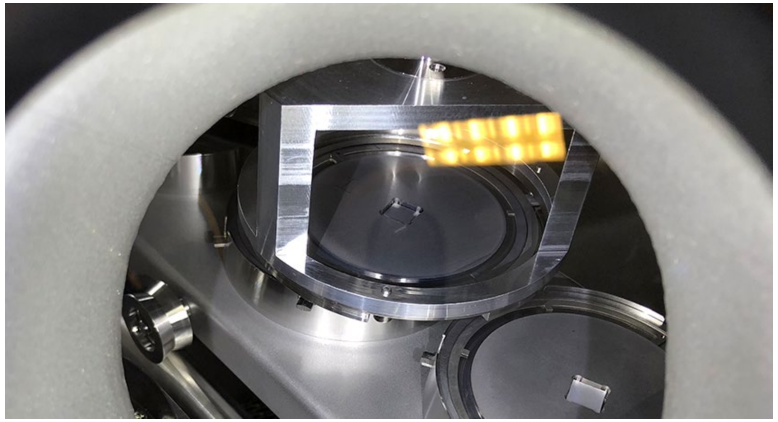
Data Science Activities in the 2DCC-MIP

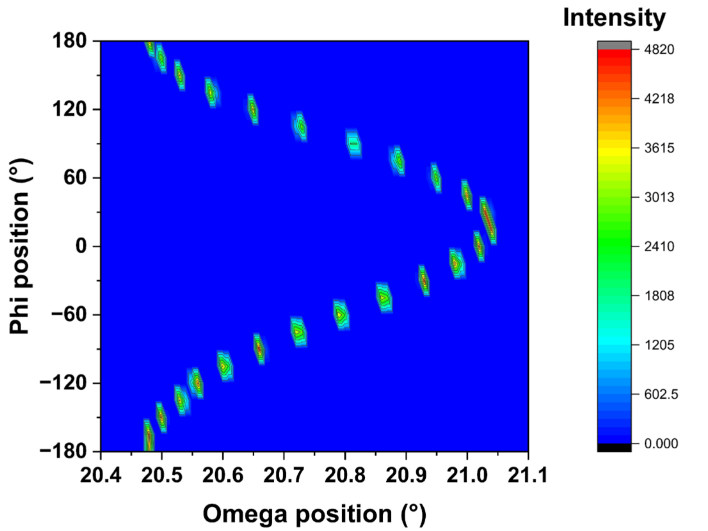


Joan Redwing, Anthony Richardella, Konrad Hilse, Kevin Dressler, Wesley Reinhart, Nitin Samarth, Vincent Crespi

2D Crystal Consortium (2DCC), Pennsylvania State University







A national user facility focused on advancing the synthesis of 2D layered chalcogenides

2D Crystal Consortium
NSF Materials Innovation Platform

User Facilities

Thin Films and *in situ* Characterization

- MBE, MOCVD and CVD/epi graphene

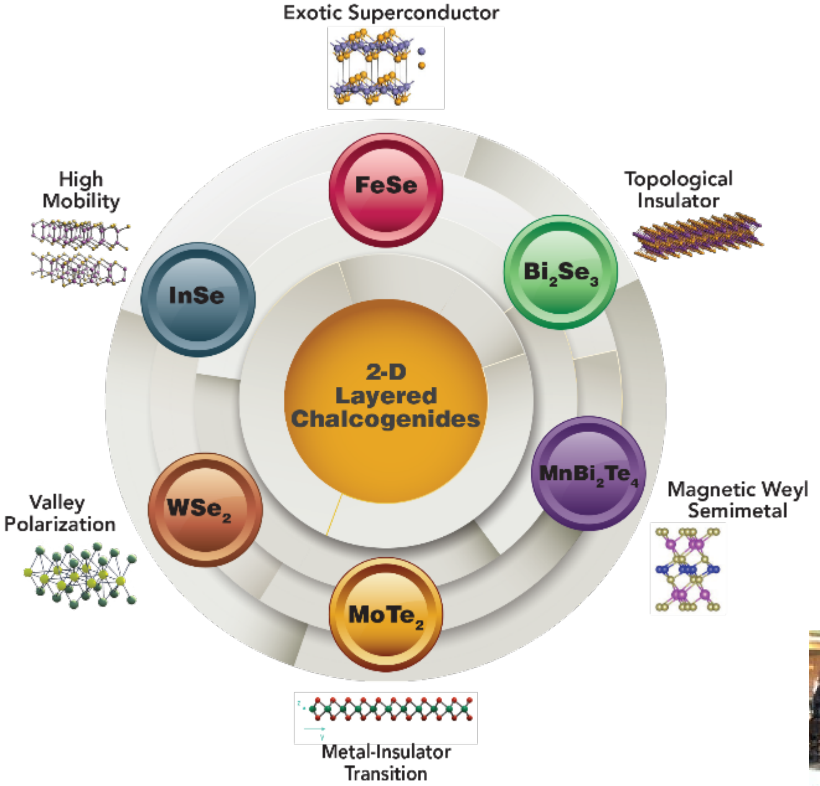
Bulk Crystal Growth

- CVT, Flux growth, Bridgman, Float Zone

Theory, Simulation and Data Science

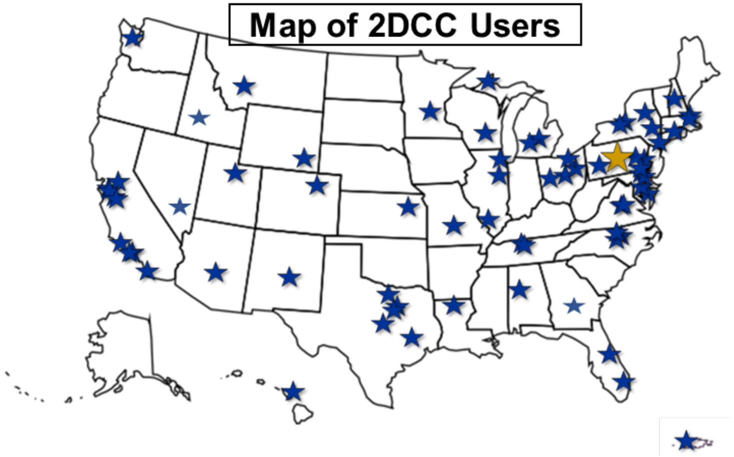
- DFT, ReaxFF-based MD, Phase-field
- LiST database

In-House Expertise



For more information: www.2dccmip.org

User Program

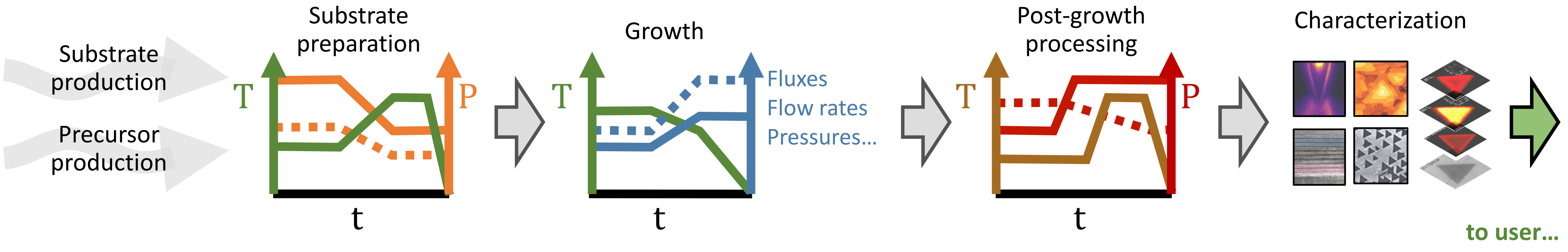


Workshops, Webinars and Hands-on Training

Funded in 2016, renewed in 2021

In synthesis, every sample has a history…

- Series of processes in time with parameters and uncertainties at every step.



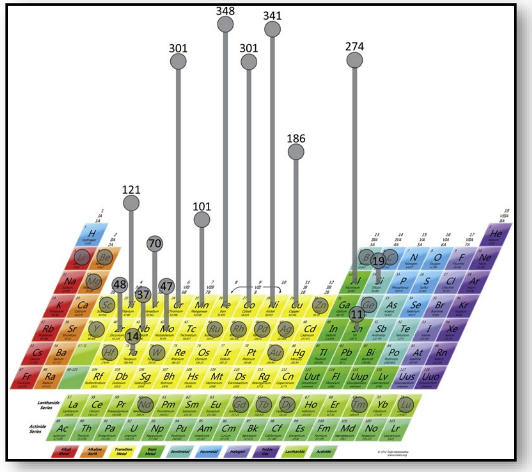
<u>Theorist</u>: "Every sample is identical" **<u>Experimentalist</u>: "Every sample is unique"**

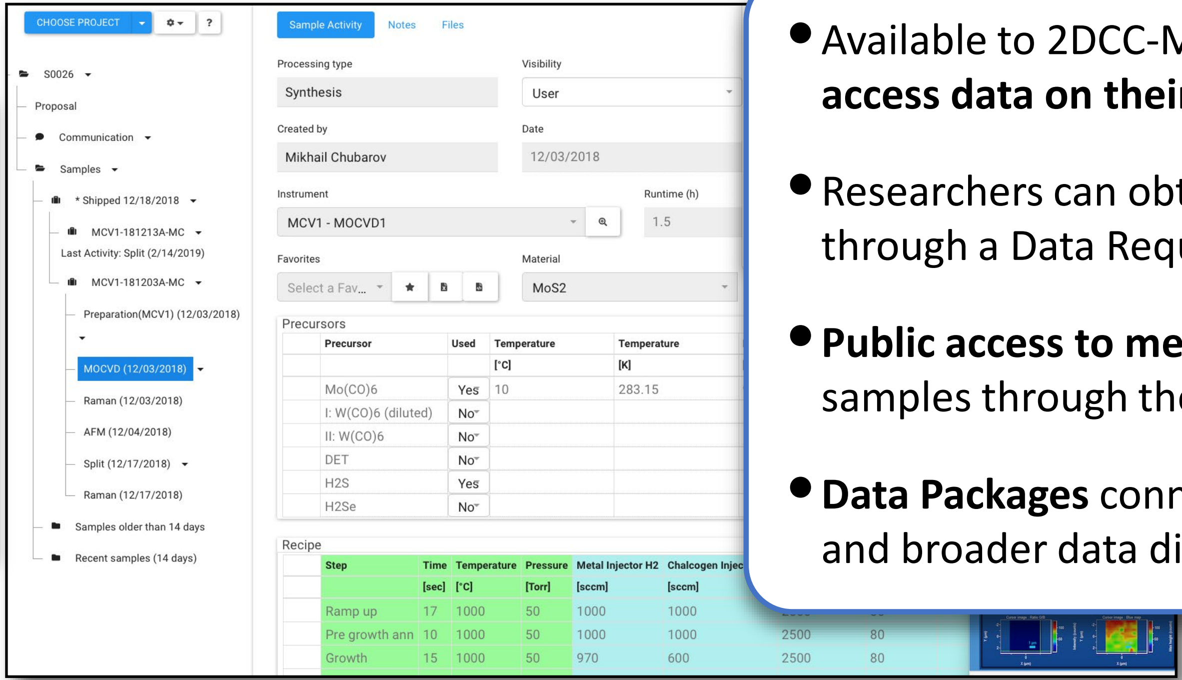
- Important to capture the histories, especially the differences, across samples as well as "failed" experiments.

Lifetime Sample Tracking - LiST

2D Crystal Consortium
NSF Materials Innovation Platform

- Infrastructure for the capture, curation, and analysis of materials growth and characterization data
  - Sample growth and characterization data > 16,000 bulk and thin film samples
  - Growth recipes/parameters imported directly from the synthesis tool
  - Automated pipelines to collect data wherever possible



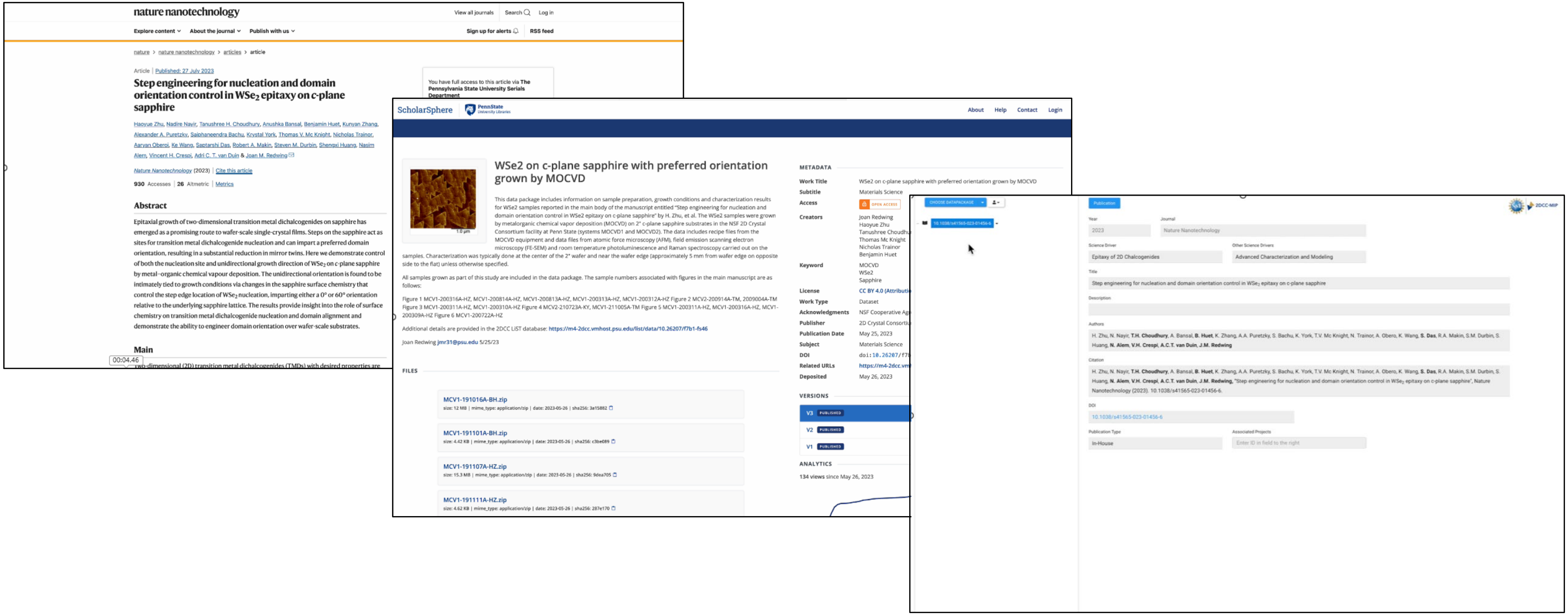


- Available to 2DCC-MIP users to **view their own projects and access data on their samples**.
- Researchers can obtain **direct access** to LiST data via **API**, through a Data Request.
- **Public access to metadata** on all (non-proprietary) LiST samples through the 2DCC website.
- **Data Packages** connected to **data DOIs** support publication and broader data dissemination.

Data Packages

2D Crystal Consortium
NSF Materials Innovation Platform

- Paper links to data archived in Penn State ScholarSphere, which links to LiST



CHOOSE DATAPACKAGE

Public Data | Publications | Projects

2DCC-MIP

Materials

Available Data: Any

Synthesis Techniques: Any

Available Characterization Data: Any

RESET SEARCH CRITERIA | SEARCH

Search in Title, DOI and Description

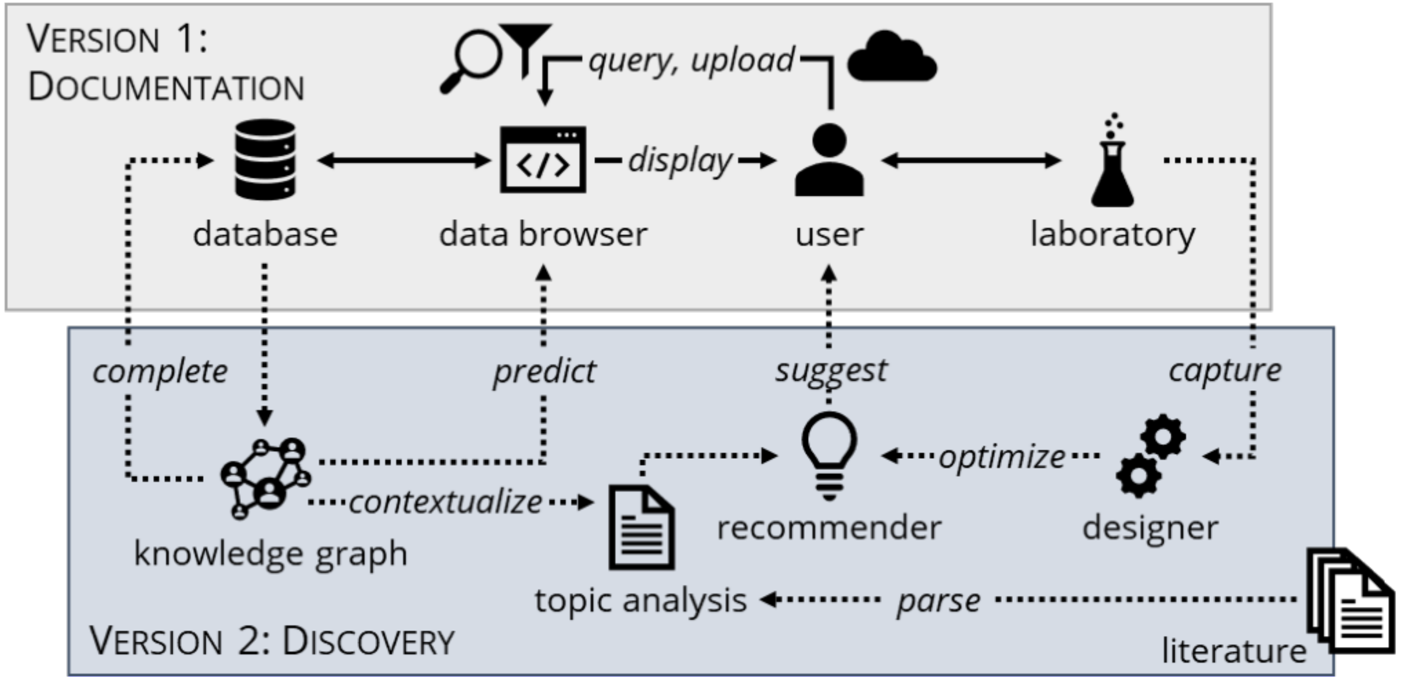
Show 10 entries

Label	2DCC Data DOI	Project	Publication DOI	Available Data	Type
Young Modulus		I_ReaxFF_2D_Toolbox		Theory Data	Snapshot
WSe2 on c-plane sapphire with preferred orientation grown by MOCVD	10.26207/f7b1-fs46		10.1038/s41565-023-01456-6	87 Samples	Snapshot - Published
Wafer-scale MOCVD TMD films used for 3D Monolithic Integration	10.26207/khwb-rr73			38 Samples	Snapshot - Published
Structures of defect pairs		I_SIM_BN_MoS2_1	10.1103/PhysRevB.99.155430	Theory Data	Snapshot
Potatoe Stamp Concept		I_ReaxFF_2D_Toolbox	10.1021/acs.jpcc.8b02991	Theory Data	Snapshot
MoSe2-WSe2 in-plane heterostructures for exciton confinement				21 Samples	Snapshot - Published
InSe on 3 Inch Si(111) Wafer-Scale Synthesis with Combinatorial Growth Approach by MBE	10.26207/qakv-p610			13 Samples	Snapshot - Published

Showing 1 to 7 of 7 entries

Previous 1 Next

LiST 2.0: From Data Store to Knowledge Graph



1.0: Documentation

⬇

2.0: Discovery

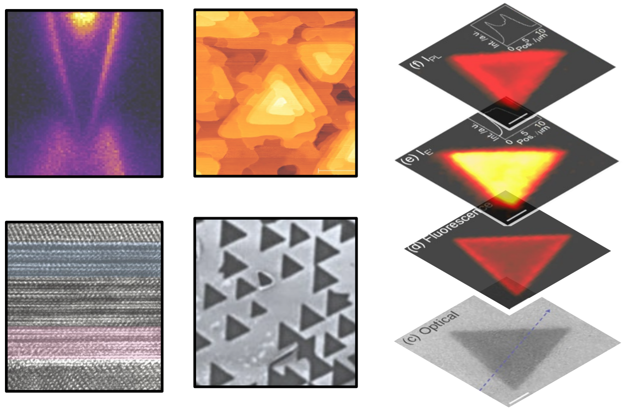
LiST 1.0 is a **traditional database** whose consistent schema are straightforward to query, synthesize, and maintain, but it can only map explicit relationships between data.

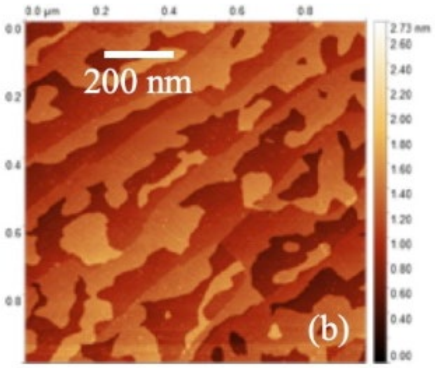
List 2.0's **knowledge graph** will contain both a data store and an ontology describing the relationships between entities and will be able to infer relationships not explicitly provided.

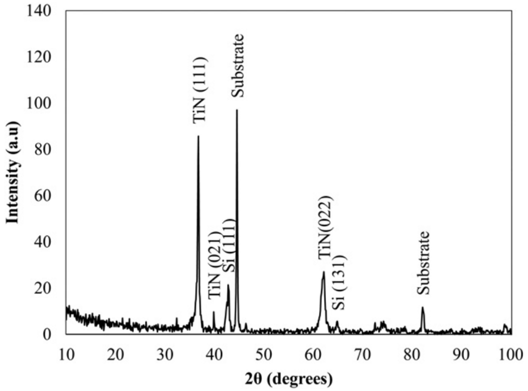
Computer vision tools for LiST 2.0

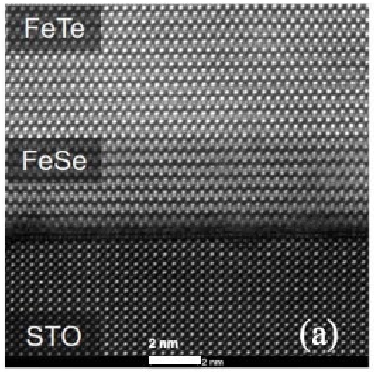
Wes Reinhart
Data Science Lead

Why should materials researchers care about computer vision?









- In the data science sense, “images” are any data with regular spatial structure. Images thus are an abundant data modality for 2DCC.
- Each image contains more data than other common modalities like text or scalars.
- Humans under-utilize image data (in science) as we can’t study them pixel-by-pixel.

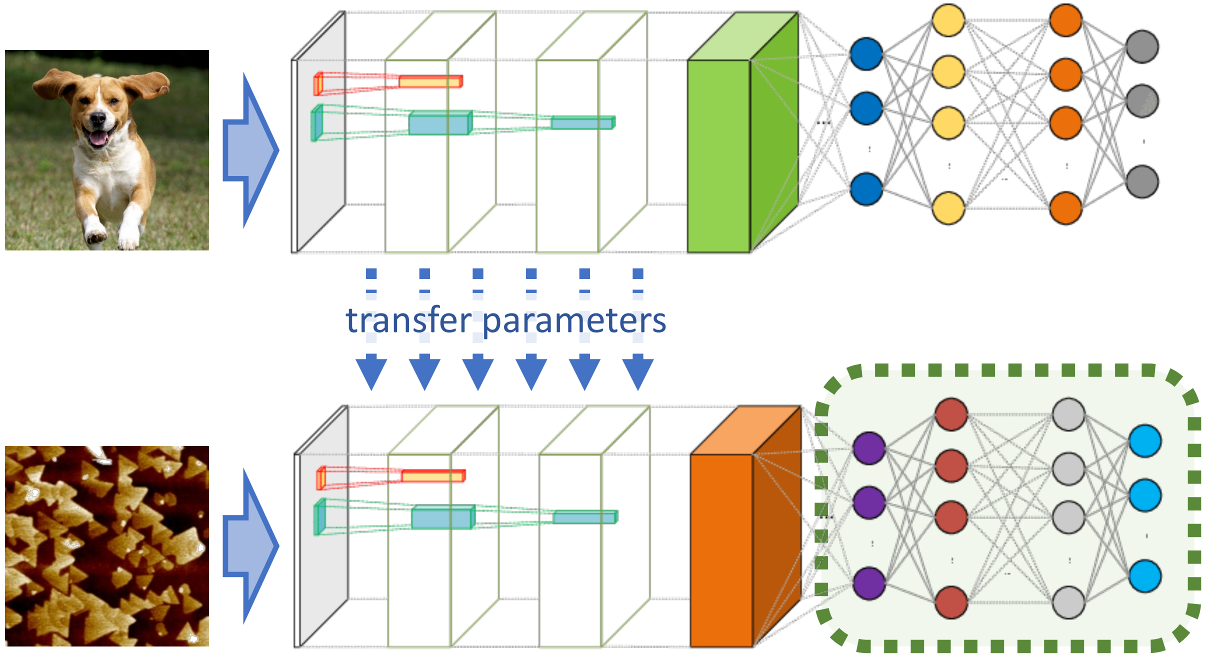
Thus, computer vision will be a key component of materials informatics.

Pretrained Convolutional Neural Networks

Convolutional Neural Networks learn abstractions for patterns in large image collections, such as classifying "dog" versus "cat". They can have billions of learned parameters and leverage the internet for data.

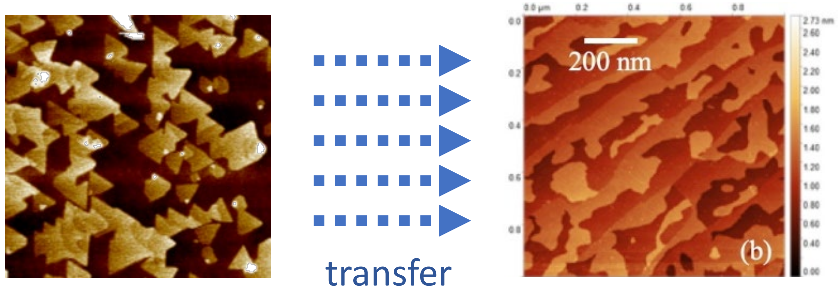
Transfer learning can retrain just a few of these weights to **specialize this highly trained classifier to a new domain** that has a more modest image collection.

More recently, we have also been leveraging "feature extraction" with pretrained networks to evaluate patterns in micrographs without any network retraining.



Lemley, Bazrafkan, and Corcoran (2017) MAICS

Transfer learning can leverage pattern recognition developed in one (non-materials) context into new (materials research) contexts, and also between different materials datasets (e.g. MOCVD→MBE).



Leveraging computer vision advances for materials

Recent examples of transfer learning:

Monolayer coverage of WSe_2

We perform systematic testing of different ML methods for estimating a scalar quantity from 221 AFM. Top-performing models have $R^2 = 0.98$.

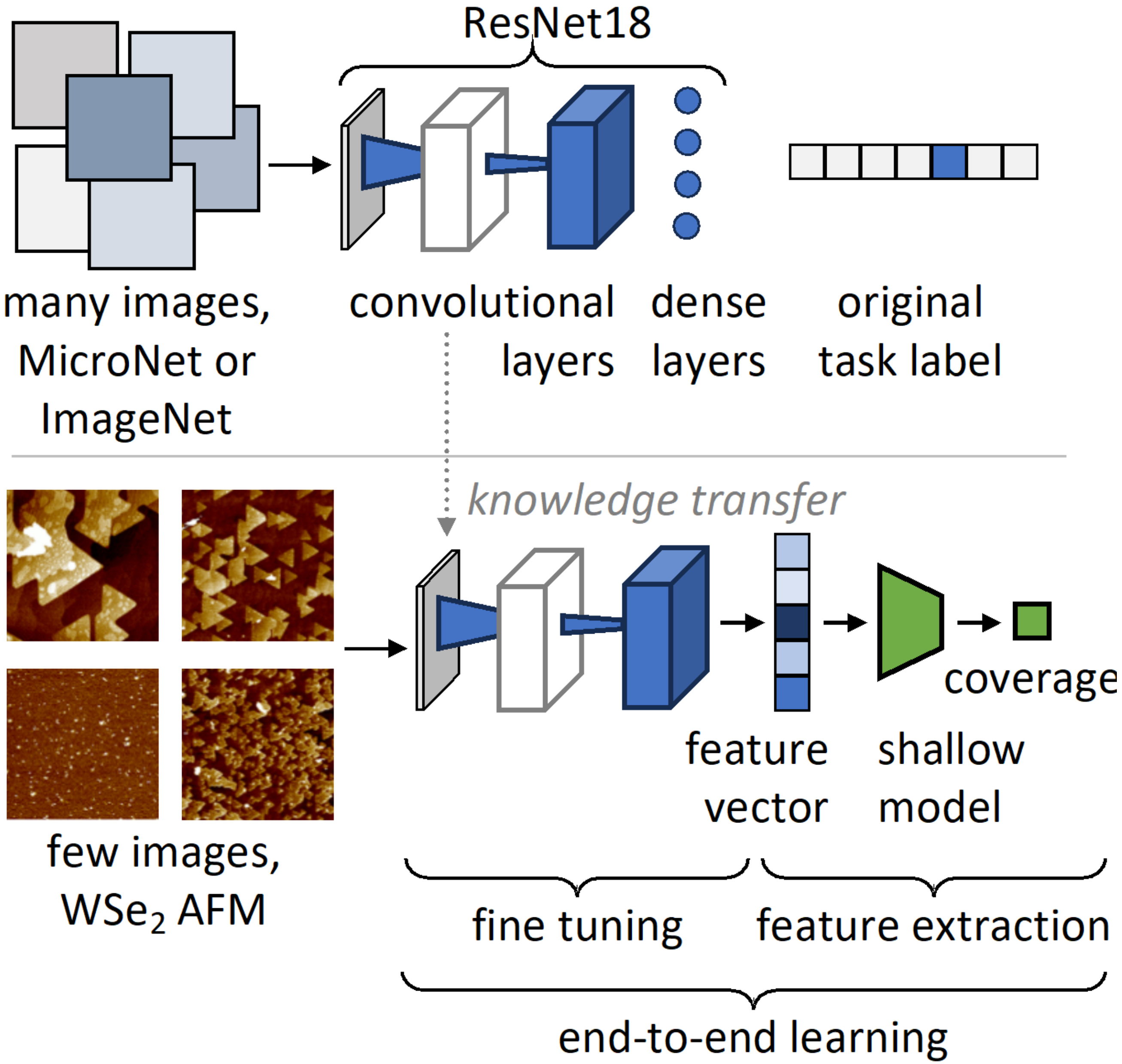
I.A. Moses, C. Wu, W.F. Reinhart, Materials Today Advances, 2024, 22, 100438.

Classification of MoS_2 growth conditions

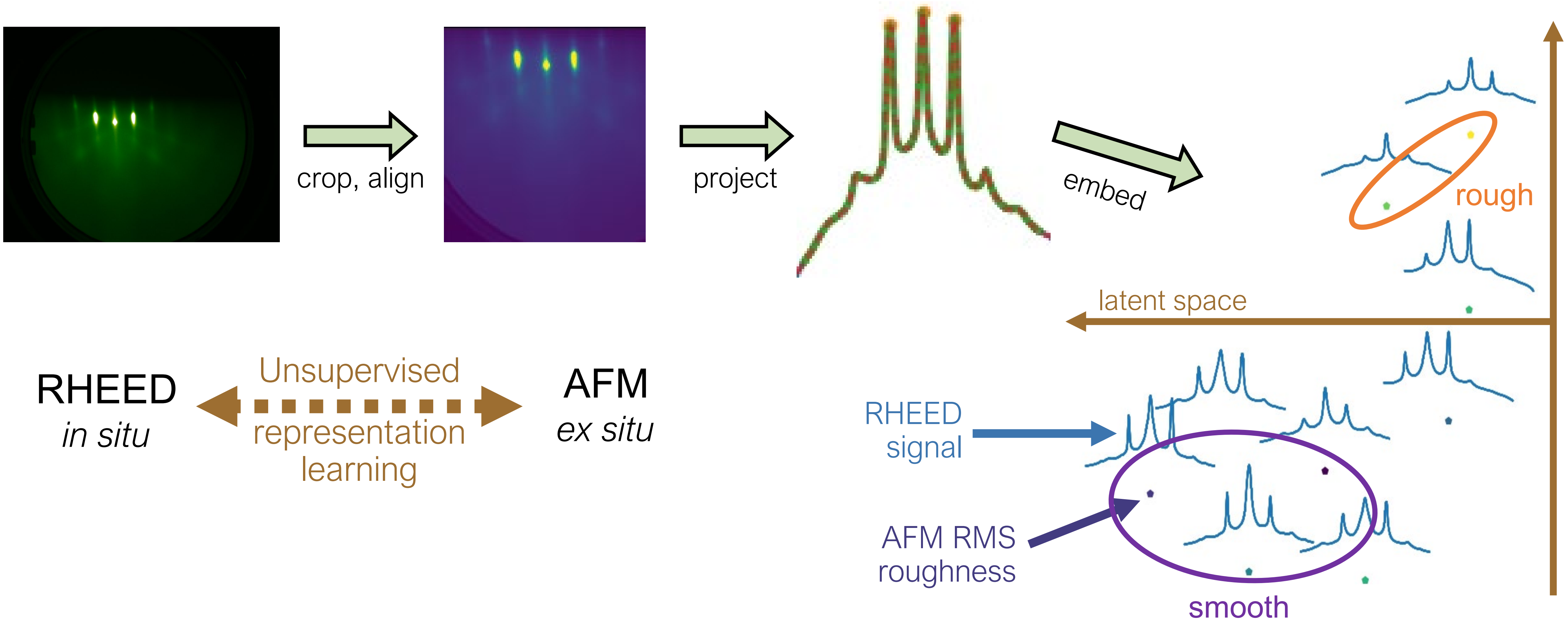
We investigate many different classification methods to retroactively "predict" the growth conditions from 262 AFM. Best $R^2 = 0.70$.

I.A. Moses, W.F. Reinhart, Materials Characterization, 2024, 209, 113701.

This appears to be an effective strategy for producing accurate models with few data.



Reflection high-energy electron diffraction (RHEED)

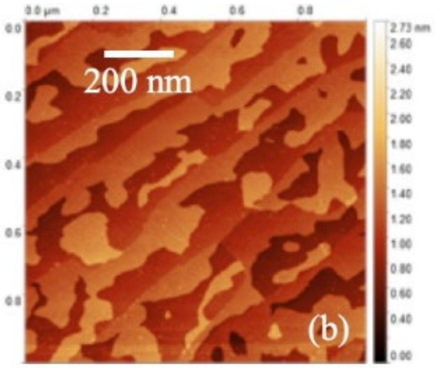


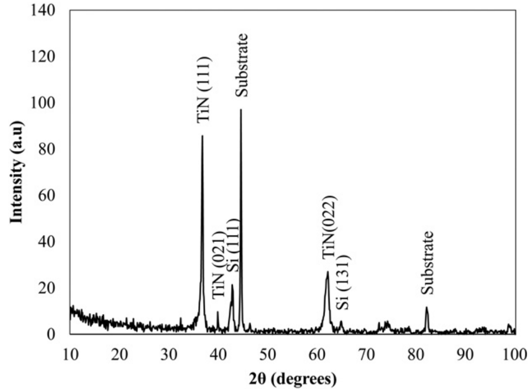
Correlating signals between instruments unlocks new predictive capabilities, especially if *in situ* measurements acquired during growth can predict *ex situ* ones measuring growth outcomes.

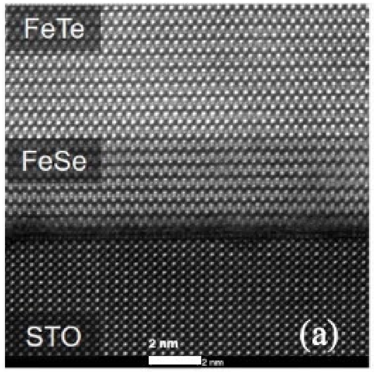
2DCC-MIP

Vector database for multi-modal data

Big, multi-modal data source (i.e., LiST) – how to compare different modalities?

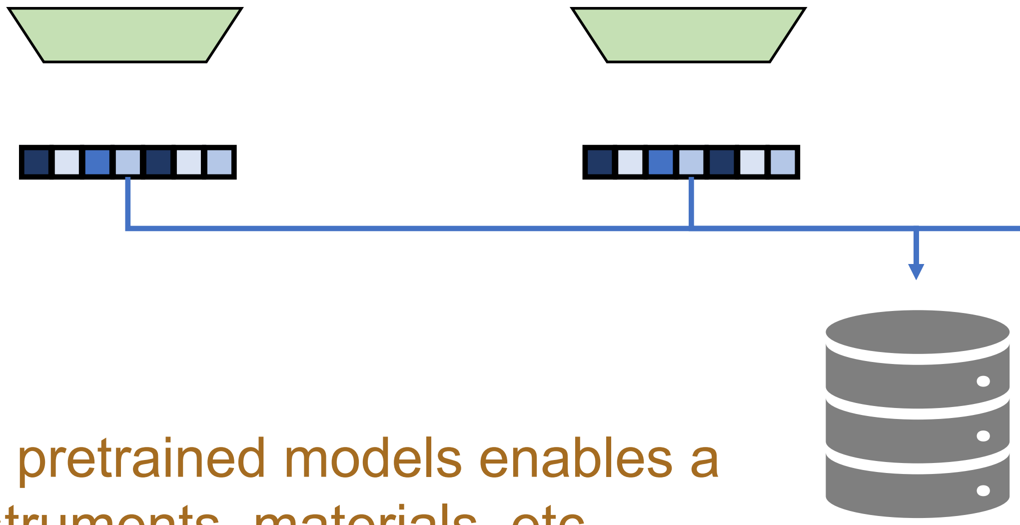






Pretrained computer vision models

Everything is a vector!



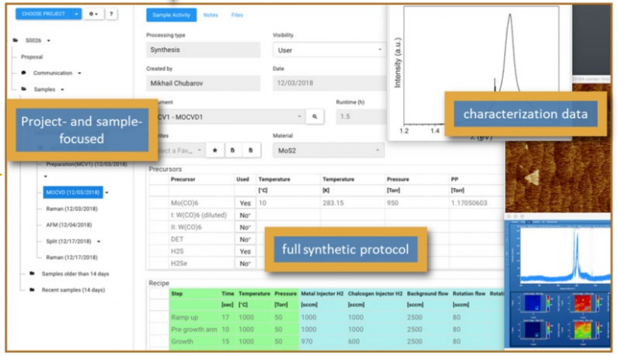
Vector database

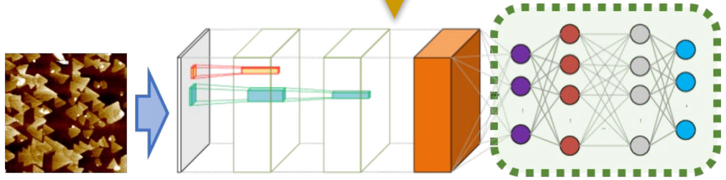
Leveraging feature extraction from pretrained models enables a common representation across instruments, materials, etc. Once in vector form, standard analyses can be performed.

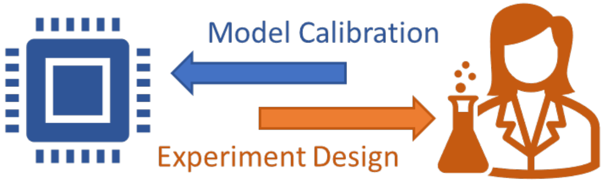
Advancing 2D materials synthesis through data science

In situ and post-growth characterization data

Source information
Sample prep
Process recipes
Lab sensor data
User notes







AI assistant to improve human decision making and experimental design

LiST 1.0

- Automated data capture and organization
- Facilitates **storage and sharing**

LiST 2.0

- Computer vision tools and vectorization
- Knowledge graph generation
- Facilitates **analysis & correlations**

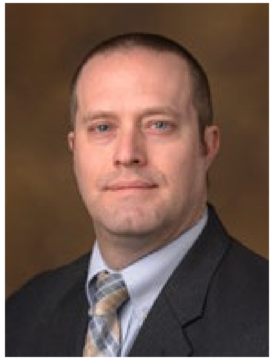
LiST 3.0

- Bayesian design of experiments
- Active learning
- Facilitates **human-AI collaboration**

2DCC Data Science Team

LiST 1.0

Konrad Hilse
Full Stack Dev



Kevin Dressler
Operations Manager

Vin Crespi
Theory, Simulation, Data Science Facility Lead

LiST 2.0

Wes Reinhart
Data Science Lead

Isaiah Moses
Postdoctoral Scholar

Tool Integration

Anthony Richardella
MBE Research Prof