OAK RIDGE National Laboratory

Towards Theory-in-the-loop for Autonomous Experiments –workflows, ML models and *ab initio* developments leveraging extreme scale computations

P. Ganesh

Distinguished R&D Staff Member and Section Head

Center for Nanophase Materials Sciences

Oak Ridge National Laboratory

Email: ganeshp@ornl.gov

This work was funded by the **INTERSECT-LDRD** (**P. Ganesh**) at ORNL; Work was performed at the **Center for Nanophase Materials Sciences**, which is a US Department of Energy, **Office of Science User Facility at ORNL**.

Acknowledgments – key collaborators

Workflows & AI/ML

Ryan Morelock

Soumendu Bagchi

Victor Fung

Addis Fuhr

MBE

Rob Moore II

Matt Brahlek

STEM

Matt Boebinger

SPM / STM

Rama Vasudevan

RMG DFT

Jerzy Bernholc

QMC

Paul Kent

(….and many more)

Acknowledgements: CNMS—DOE Office of Science User Facility

Enabling users through **user proposals** https://www.ornl.gov/facility/cnms/

QMCPACK **(Center for Predictive Simulation of Functional Materials – DOE CMS @ ORNL)**

DOE HPC platforms we leverage via INCITE/ALCC awards:

U.S. DEPARTMENT OF ENERGY

Office of Science

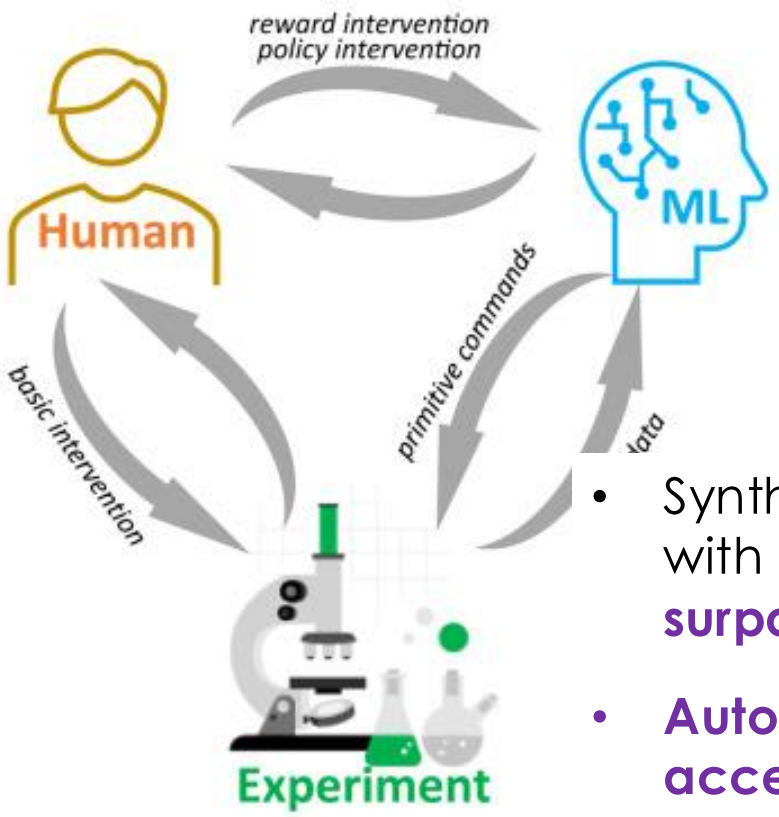
OLCF

NERSC

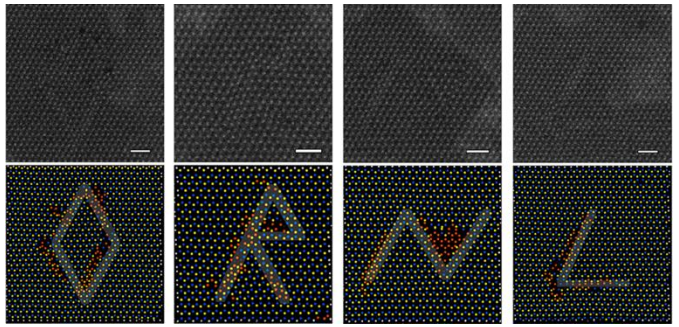
INTERSECT

Building an interconnected science ecosystem

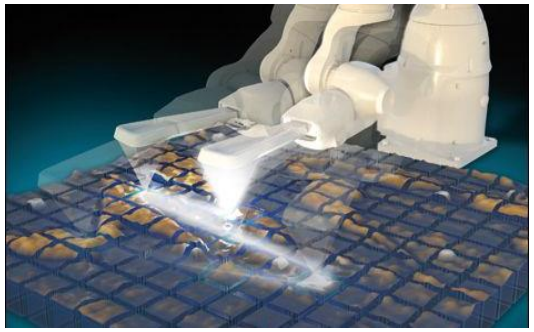
Autonomous synthesis and characterization of materials to accelerate predictions



Automated Experiments

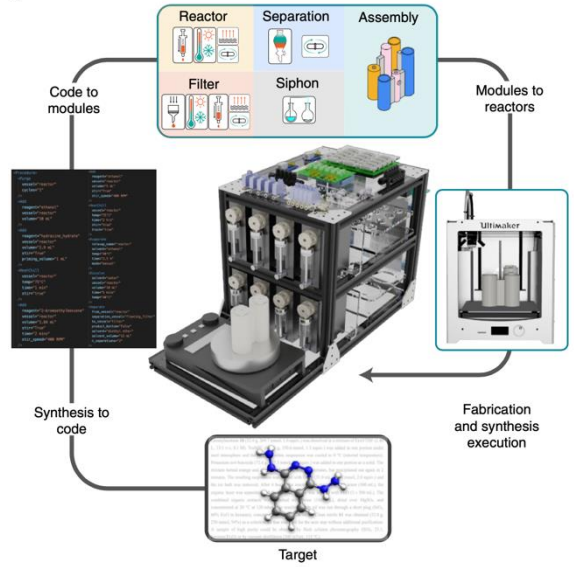


K. Roccapriori, D. Mukherjee, S. Kalinin et al.



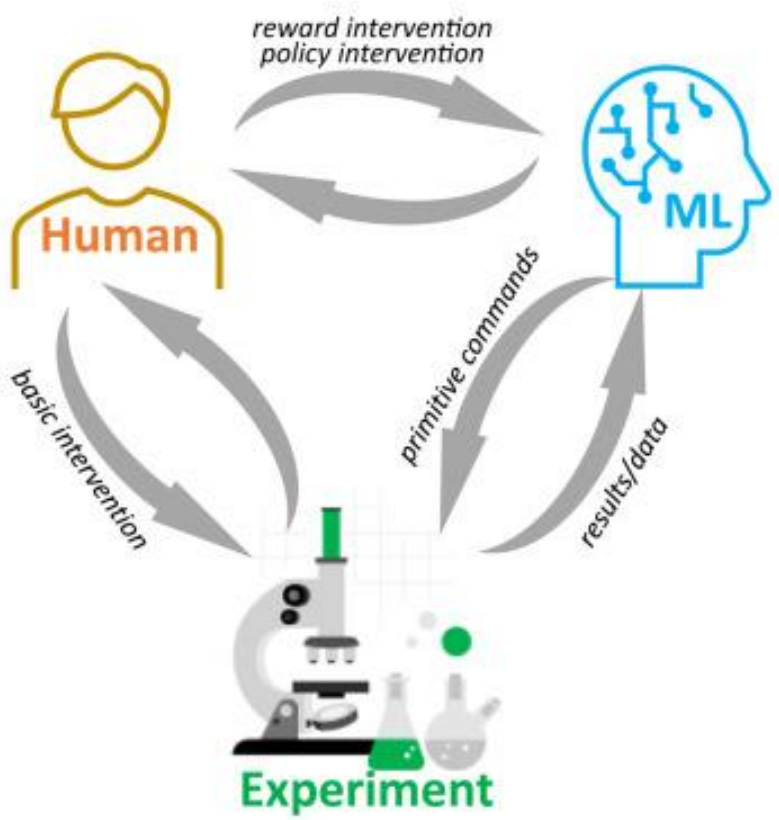
R. Vasudevan et al.

- Synthesizing/Characterizing new materials with targeted functionalities can currently **surpass months/years**.
- **Autonomous Experiments can help with accelerating this process**, can incorporate human- and **ML-algorithms** in-the-loop but ….
- **Algorithms are data-hungry** – not ideal for slow/costly experiments, need ML-algos. for inverse problems (learning physics/structure)



S. Manzano, Nat. Chem. (2022)

Ensemble-based ‘digital twins’ enabled by multiscale theory and AI/ML



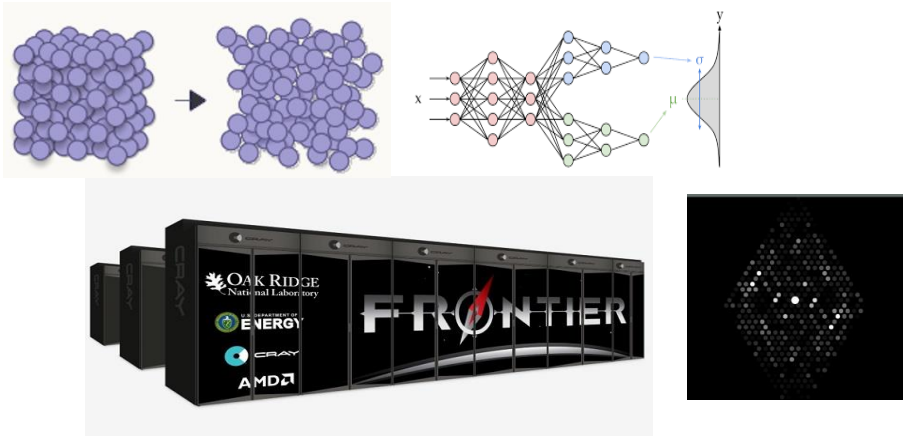
Cheap exploration using digital framework

Priors and ML-models based on DT’s of expt.

train generative-AI models for inverse-design;

measurements encoding synthesis-structure-chemistry-property;

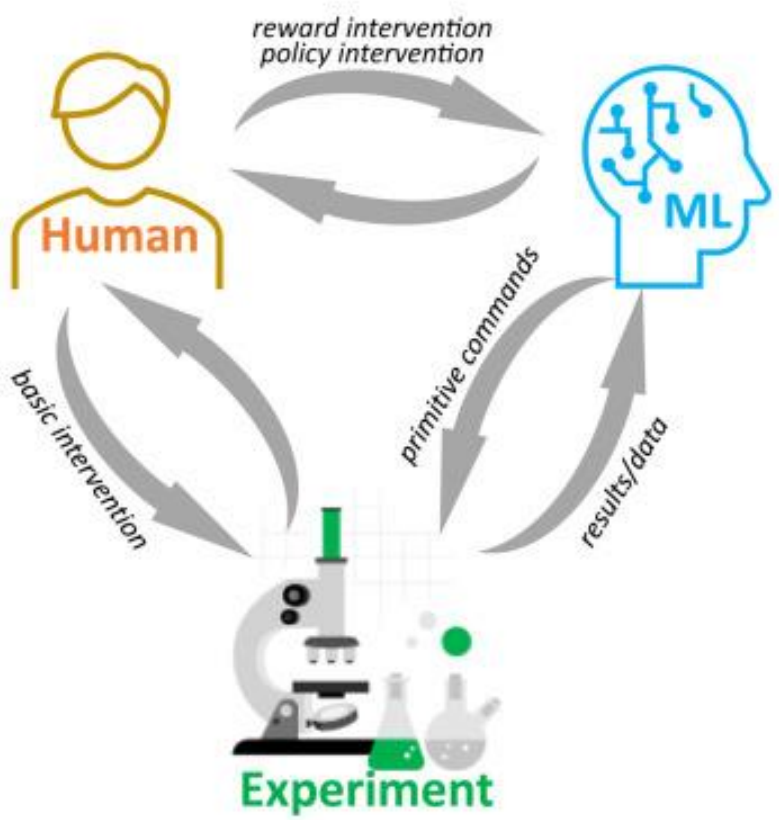
Extract new-physics extending current theories



Synthesis / characterization digital-twins (multi-fidelity and multi-scale HPC simulations of synthesis / characterization)

- While autonomous experiments (AE) could revolutionize such efforts, **significant cost/data boost** is expected from much cheaper **virtual environments**/so called **theory-driven computational platforms**.

Ensemble-based ‘digital twins’ enabled by multiscale theory and AI/ML



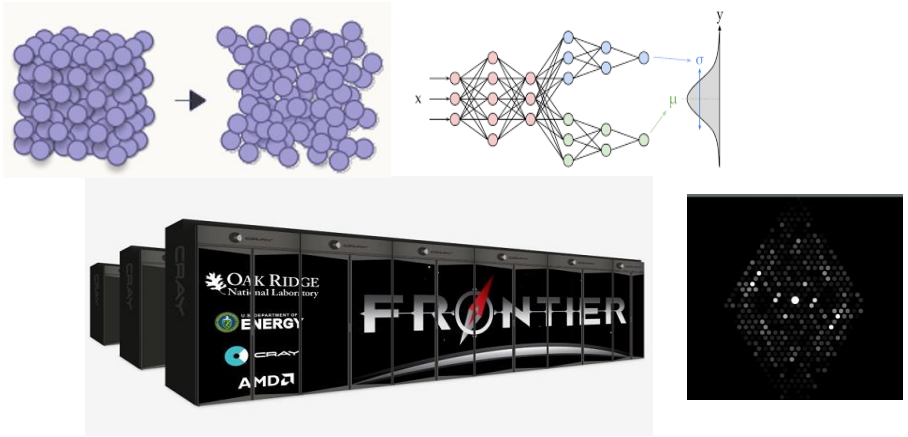
Cheap exploration using digital framework

Priors and ML-models based on DT’s of expt.

measurements encoding synthesis-structure-chemistry-property;

train generative-AI models for inverse-design;

Extract new-physics extending current theories

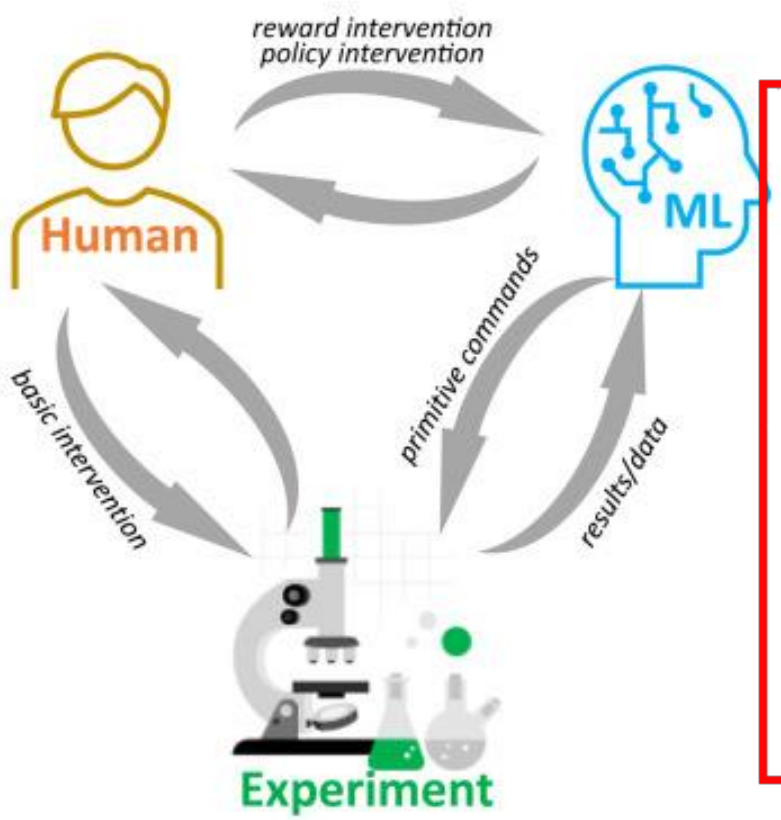


Synthesis / characterization digital-twins (multi-fidelity and multi-scale HPC simulations of synthesis / characterization)

Theory-in-the-loop to achieve true autonomous Expts. accelerating discovery of new materials & science

- While autonomous experiments (AE) could revolutionize such efforts, **significant cost/data boost** is expected from much cheaper **virtual environments**/so called **theory-driven computational platforms**.

Ensemble-based 'digital twins' enabled by multiscale theory and AI/ML



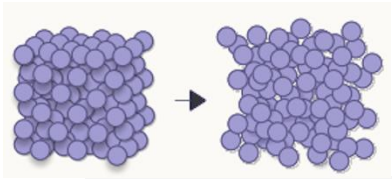
Cheap exploration using digital framework

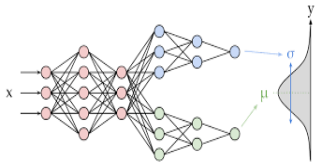
Priors and ML-models based on DT's of expt.

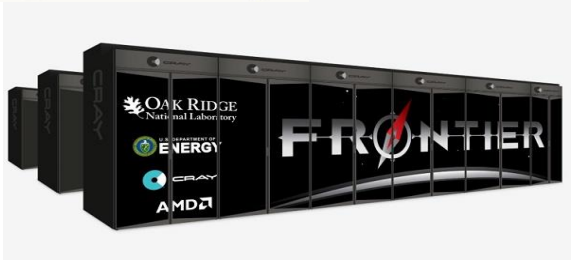
measurements encoding synthesis-structure-chemistry-property;

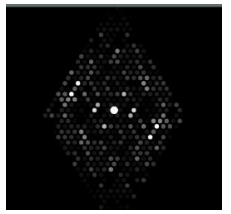
train generative-AI models for inverse-design;

Extract new-physics extending current theories









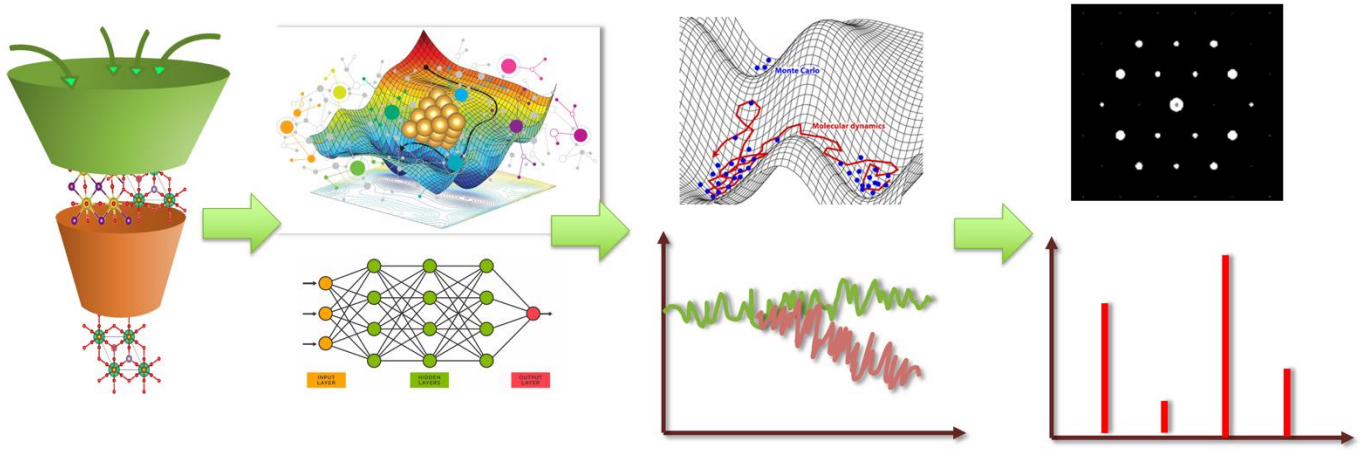
Synthesis / characterization digital-twins (multi-fidelity and multi-scale HPC simulations of synthesis / characterization)

Theory-in-the-loop to achieve true autonomous Expts. accelerating discovery of new materials & science

- Goal: **An ensemble-based active learning framework** incorporating advanced theory, computing, synthesis and characterization via **advanced AI/ML and extreme-scale computing approaches** to allow theory-guided autonomous synthesis and manipulation **with on-the-fly ML-training and smarter experimental feedback**.

Developing AI/ML+HPC Workflows for placing Materials Theory in the loop

Different levels of theory-guided numerical experiments

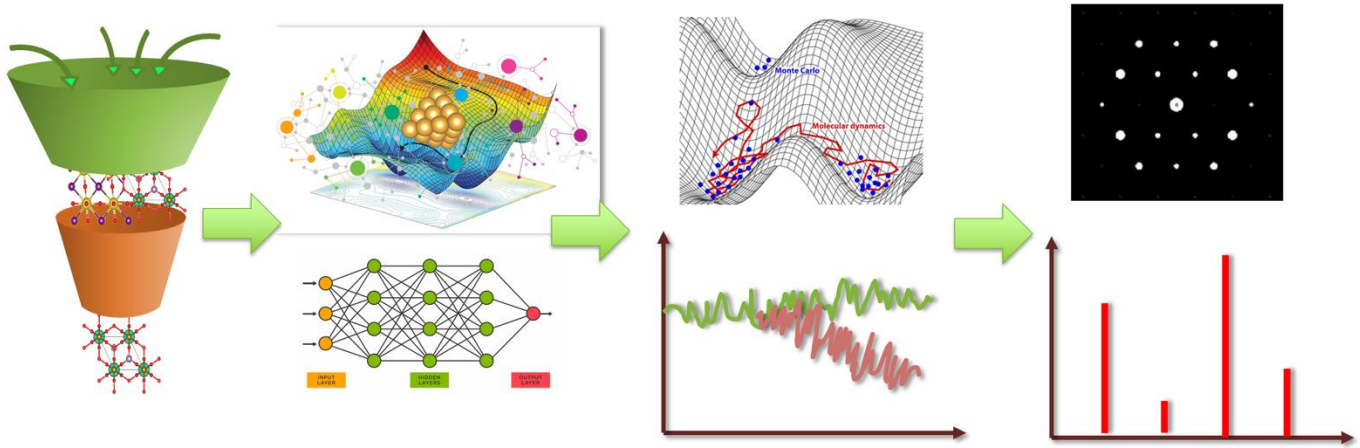


To bridge the gap across newly identified quantum materials to all the way to digital characterization, several intermediate components needs to be realized

- *Scalable-throughput is desired – need concurrency*
- *Heterogenous subworkflows and loop closures – need to be asynchronous*
- Massive and scalable (on the fly?) data processing – *in-memory processing*

Developing AI/ML+HPC Workflows for placing Materials Theory in the loop

Different levels of theory-guided numerical experiments



To bridge the gap across newly identified quantum materials to all the way to digital characterization, several intermediate components needs to be realized

- *Scalable-throughput is desired – need concurrency*
- *Heterogenous subworkflows and loop closures – need to be asynchronous*
- Massive and scalable (on the fly?) data processing – *in-memory processing*

Exploiting HPC (Exascale) for scalable throughput and asynchronous predictions

a. ***One Interface*** *for all different kind of tasks*

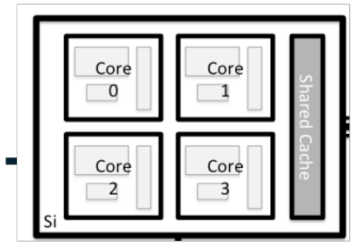
b. ***Bypassing the sub-queueing*** *to make the most out of an Exascale allocation instance*

c. ***On-the-fly processing****/performing digital characterization*

Hierarchical Adaptive Workflows for asynchronous ensemble evaluations

@OLCF

Large allocations



a. ***One Interface*** *for all different kind of tasks*

b. ***Bypassing the sub-queueing*** *to make the most out of an Exascale allocation instance*

c. ***On-the-fly processing****/performing digital characterization*

MatEnsemble

flux

Adaptive scheduling and task management layer through `concurrent.futures` objects

SLURM batch (e.g., 1000 Node-hours)

A Master Flux instance

worker job (4 GPUs)

worker job (4 GPUs)

worker job (1 GPU)

worker job (N GPUs)

PoolExecutors

Sub-worker job (1 GPUs)

Sub-worker job (1 GPUs)

Sub-worker job (2 GPUs)

Sub-worker job (2 GPUs)

Sub-worker job (1 GPUs)

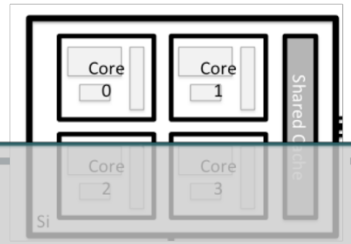
Series of further jobs

https://github.com/Q-CAD/MatEnsemble

Hierarchical Adaptive Workflows for asynchronous ensemble evaluations

@OLCF

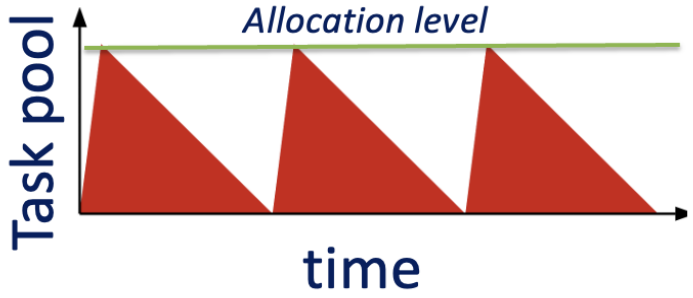
Large allocations



a. ***One Interface*** *for all different kind of tasks*

b. ***Bypassing the sub-queueing*** *to make the most out of an Exascale allocation instance*

c. ***On-the-fly processing***/performing digital characterization



For a similar allocation level, throughput can be **scaled n-folds** by using adaptive task management + scheduling to keep the pool saturated with tasks.

SLURM batch (e.g., 1000 Node-hours)

ter Flux instance

worker job (4 GPUs)

worker job (1 GPU)

worker job (N GPUs)

PoolExecutors

Sub-worker job (2 GPUs)

Sub-worker job (2 GPUs)

Sub-worker job (1 GPUs)

Series of further jobs

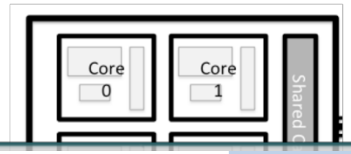
task management layer through `concurrent.futures` objects

https://github.com/Q-CAD/MatEnsemble

Hierarchical Adaptive Workflows for asynchronous ensemble evaluations

@OLCF

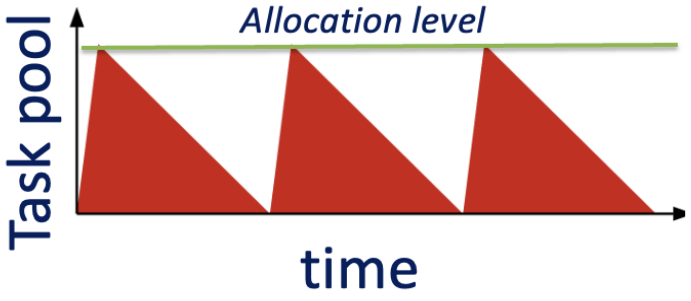
Large allocations



a. ***One Interface*** *for all different kind of tasks*

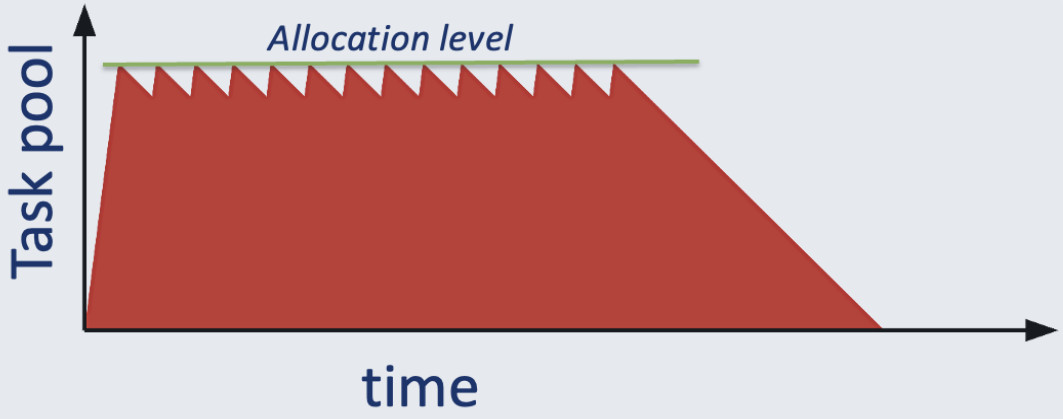
b. ***Bypassing the sub-queueing*** *to make the most out of an Exascale allocation instance*

forming digital



For a similar allocation level, throughput can be **scaled n-folds** by using adaptive task management + scheduling to keep the pool saturated with tasks.

Execution via Matsemble



task management layer through `concurrent.futures` objects

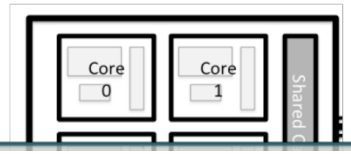
job (1 GPUs)
job (1 GPUs)
job (2 GPUs)
job (2 GPUs)
job (1 GPUs)
Series of further jobs

https://github.com/Q-CAD/MatEnsemble

Hierarchical Adaptive Workflows for asynchronous ensemble evaluations

@OLCF

Large allocations

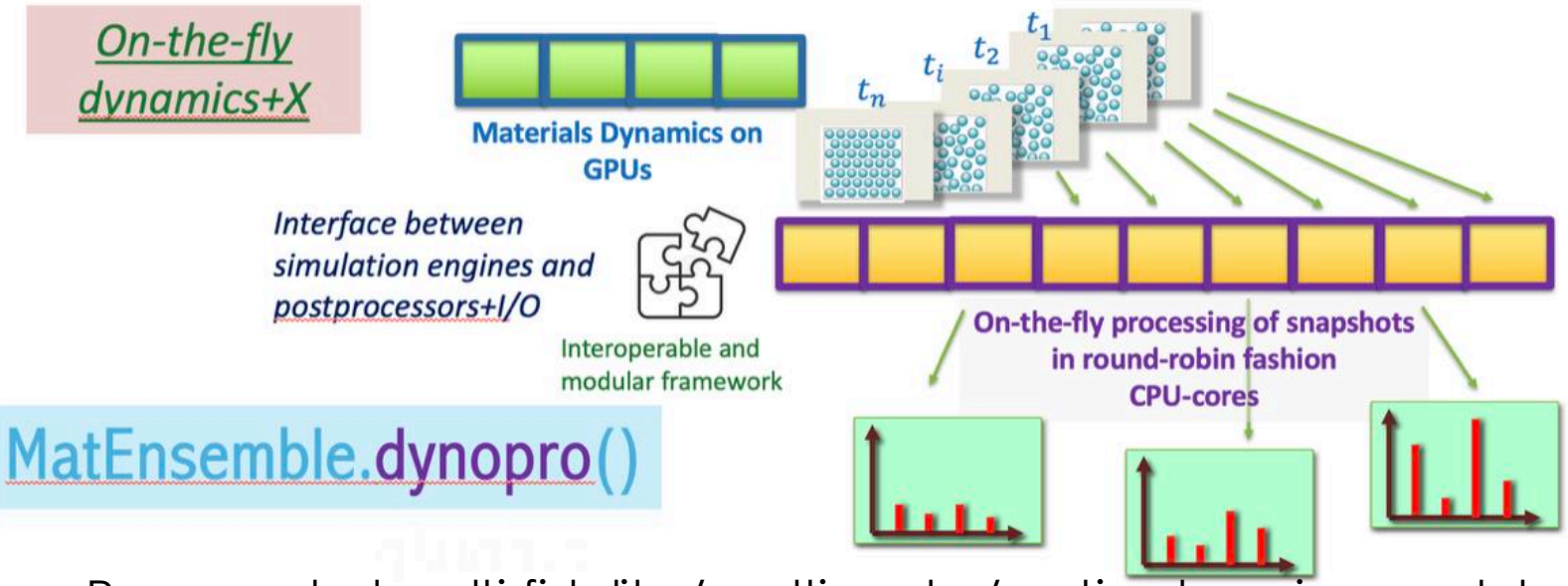


a. ***One Interface*** *for all different kind of tasks*

b. ***Bypassing the sub-queueing*** *to make the most out of an Exascale allocation instance*

forming digital

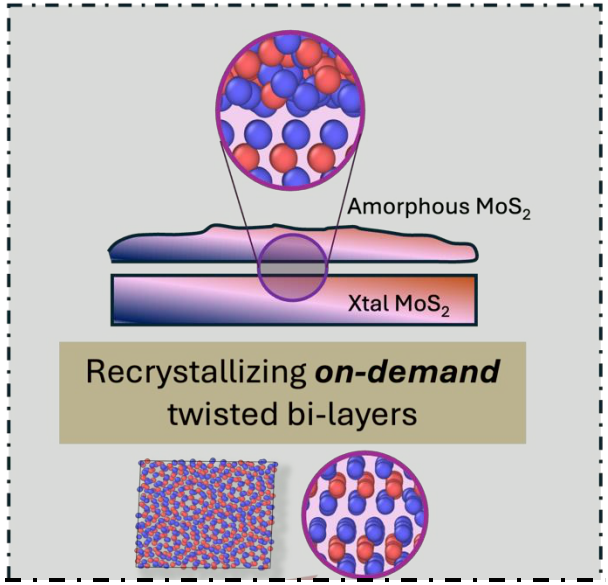
- On-the-fly streaming of materials dynamics with custom analysis algorithms

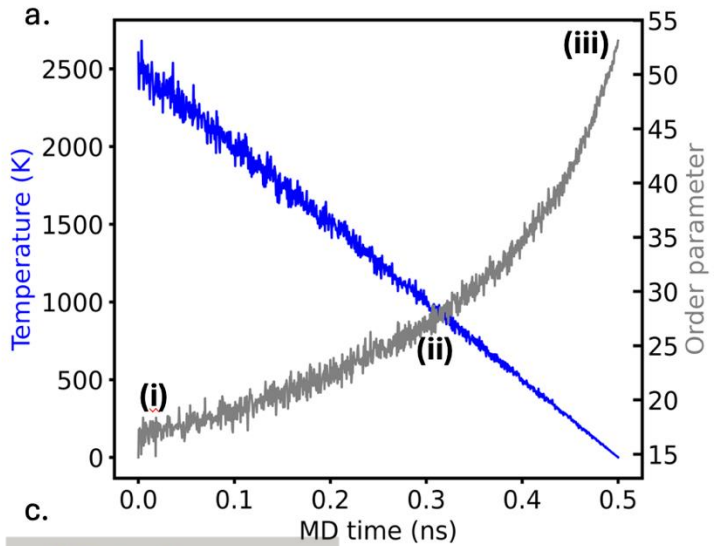


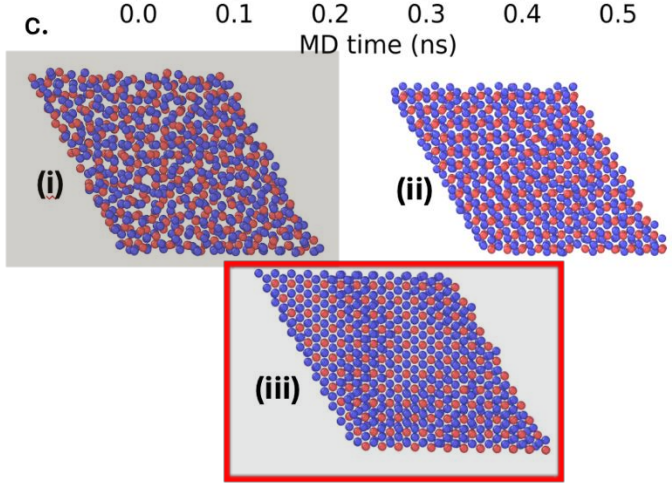
- Run coupled multi-fidelity / multi-scale / active-learning models :
  - e.g. QMC / DMFT / RPA / DFT / MD / KMC / phase-field

https://github.com/Q-CAD/MatEnsemble

Use Case: Controlling *on-demand* Recrystallization Pathways of TMDCs

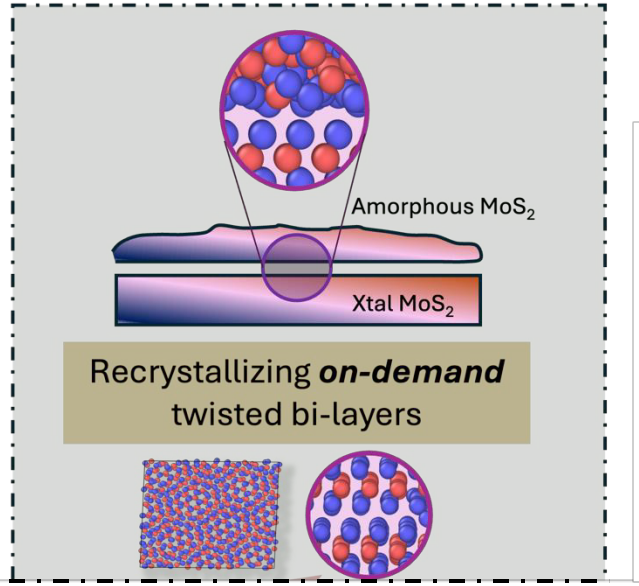






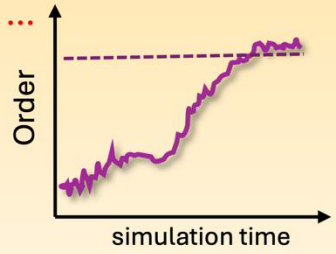
Annealing of MoS_2 amorphous/Xtal bilayers followed by quenching using classical MD (ReaxFF)

Recrystallization Pathways of TMDCs: *Tunable Parameters*

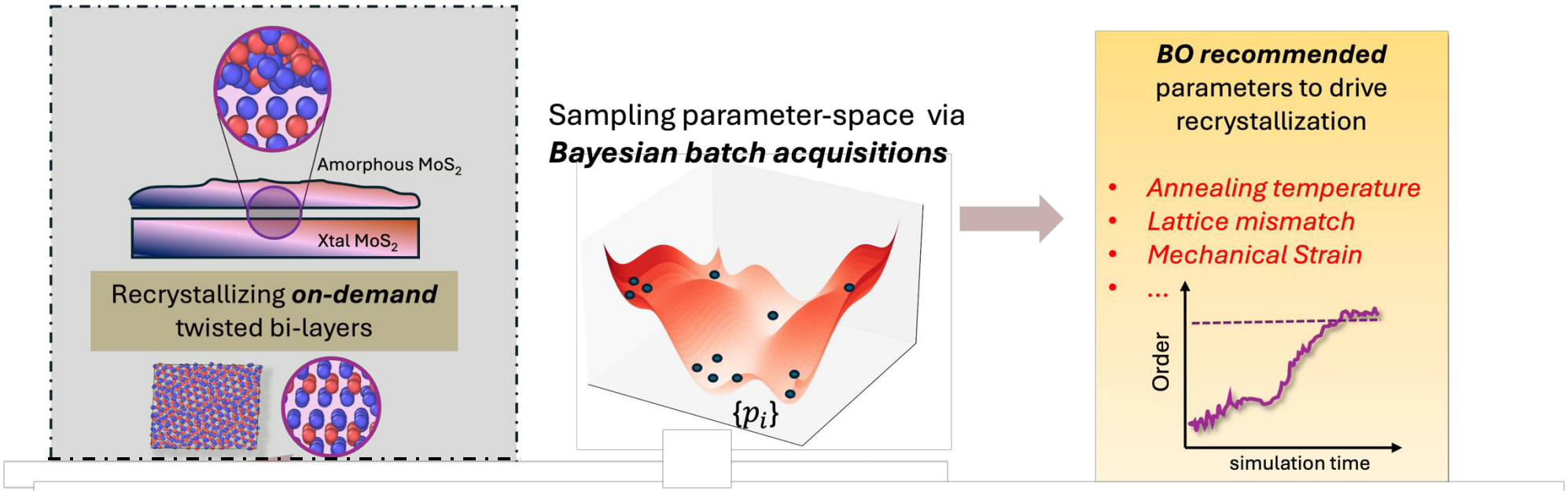


BO recommended parameters to drive recrystallization

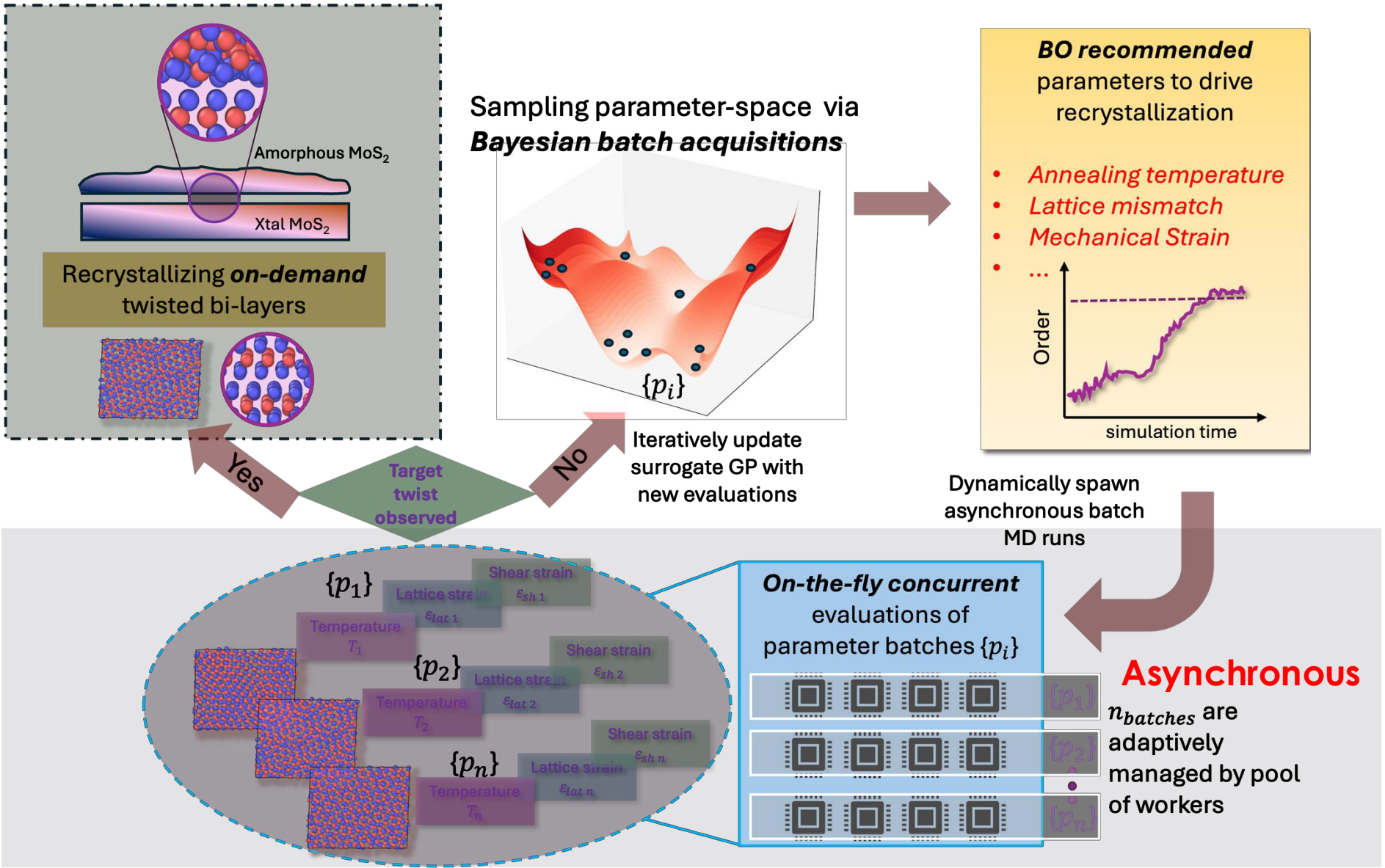
- *Annealing temperature*
- *Lattice mismatch*
- *Mechanical Strain*
- ...



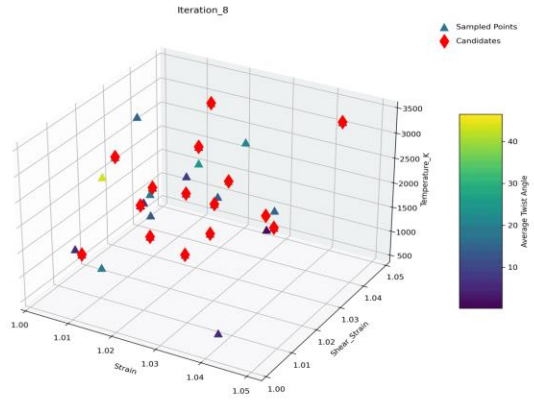
Recrystallization Pathways of TMDCs: *Asynchronous batch BO*



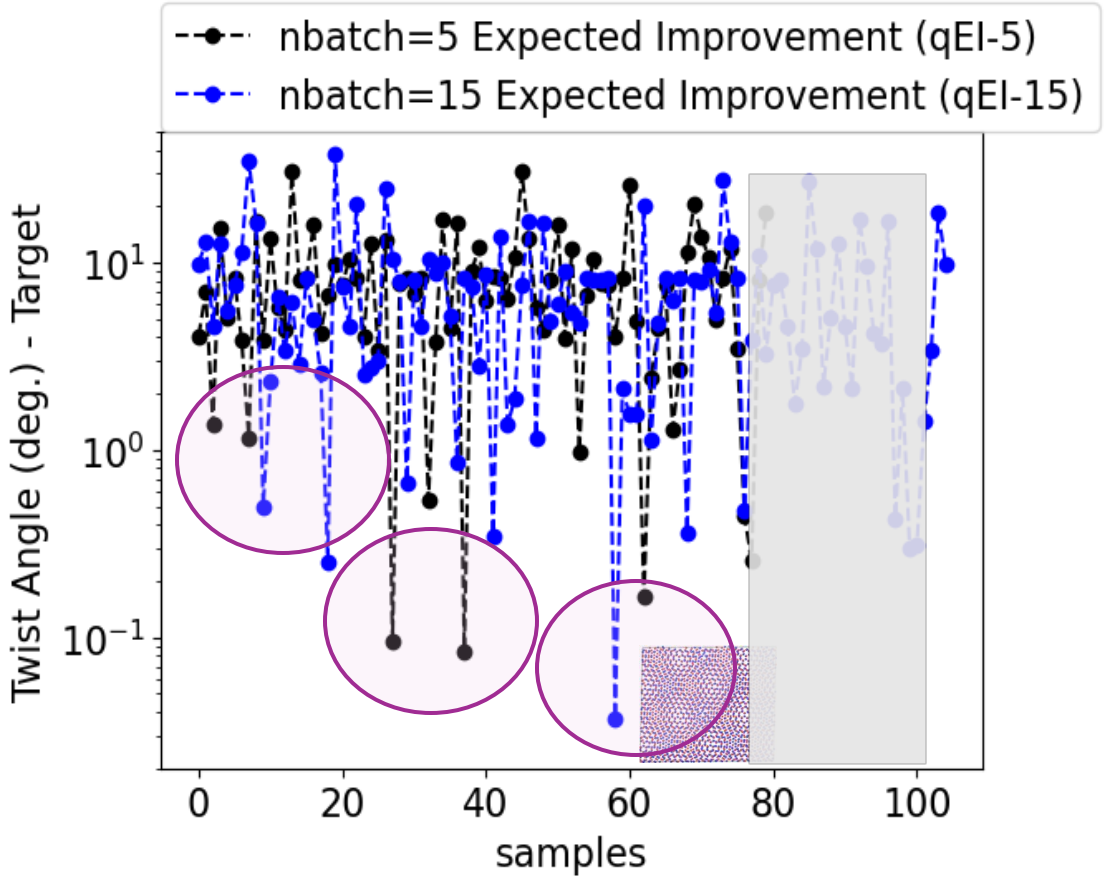
Recrystallization Pathways of TMDCs: *Asynchronous batch BO with adaptive ensemble platform*



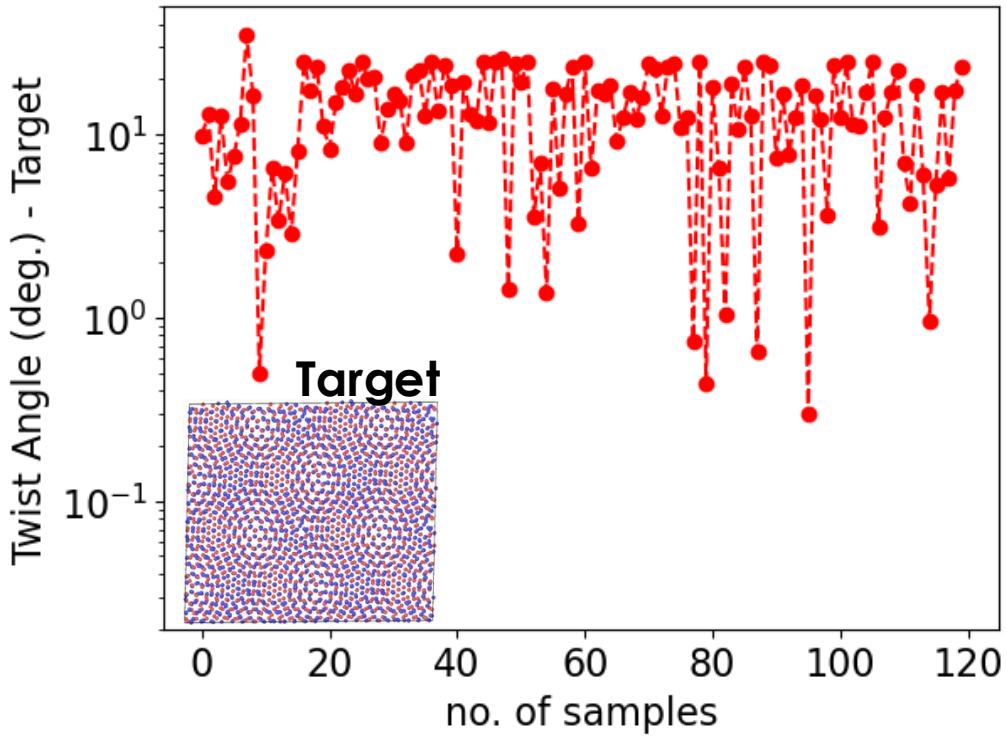
BO advantage over random acquisitions in parallel



Effect of increasing batch size



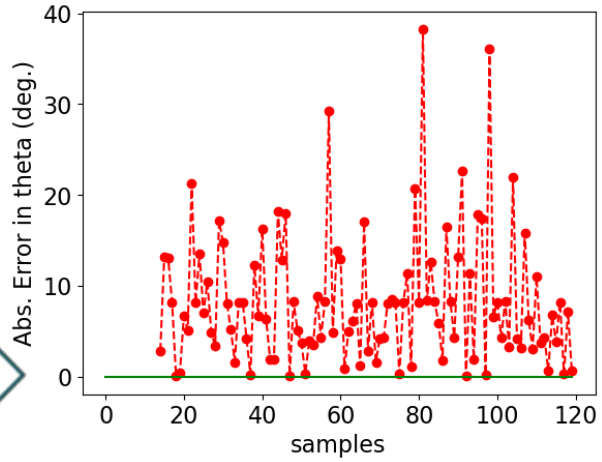
Random sampling (nbatch=15)



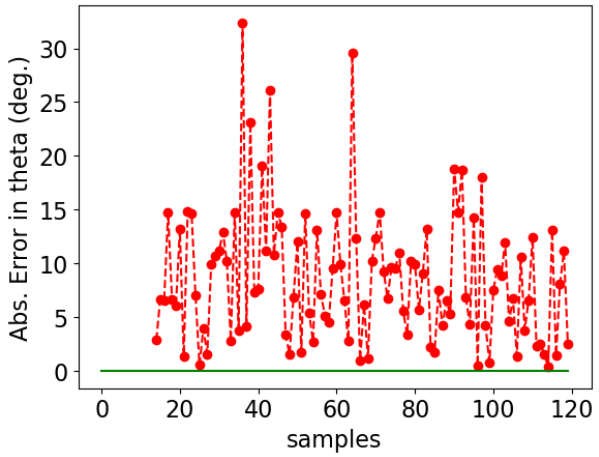
Can we sample "*on-demand*" twisted bi-layers?

Many synthesis-conditions for targeted twist

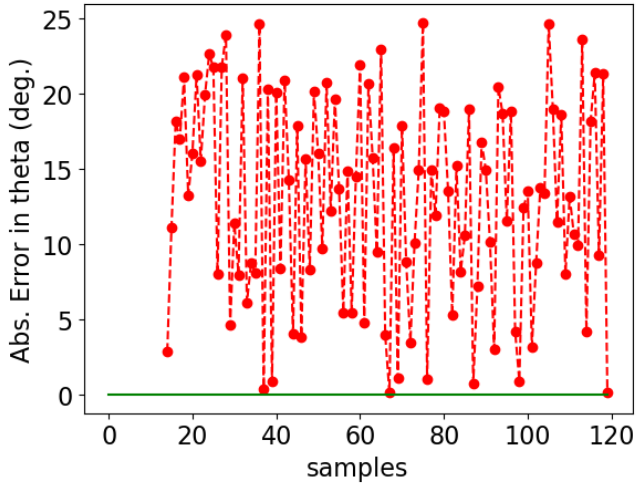
a.



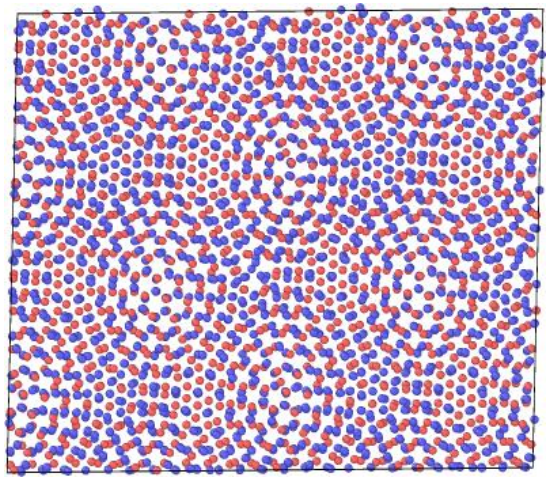
b.



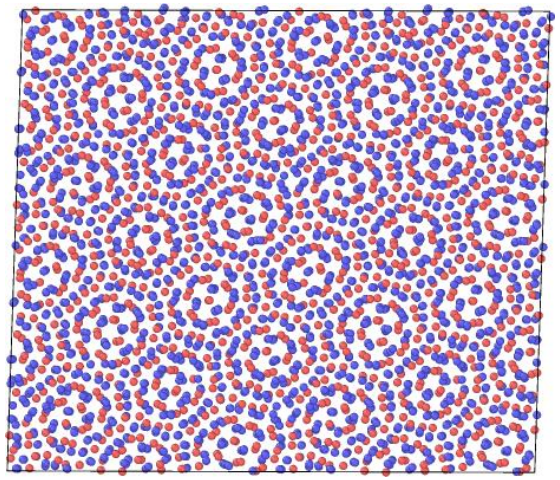
c.



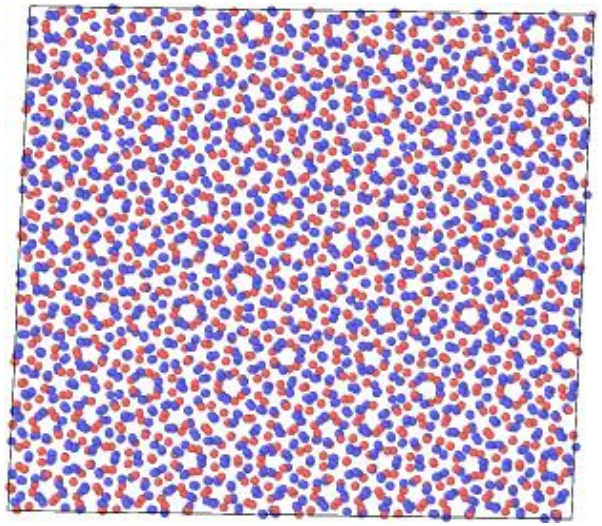
d. 8.5°



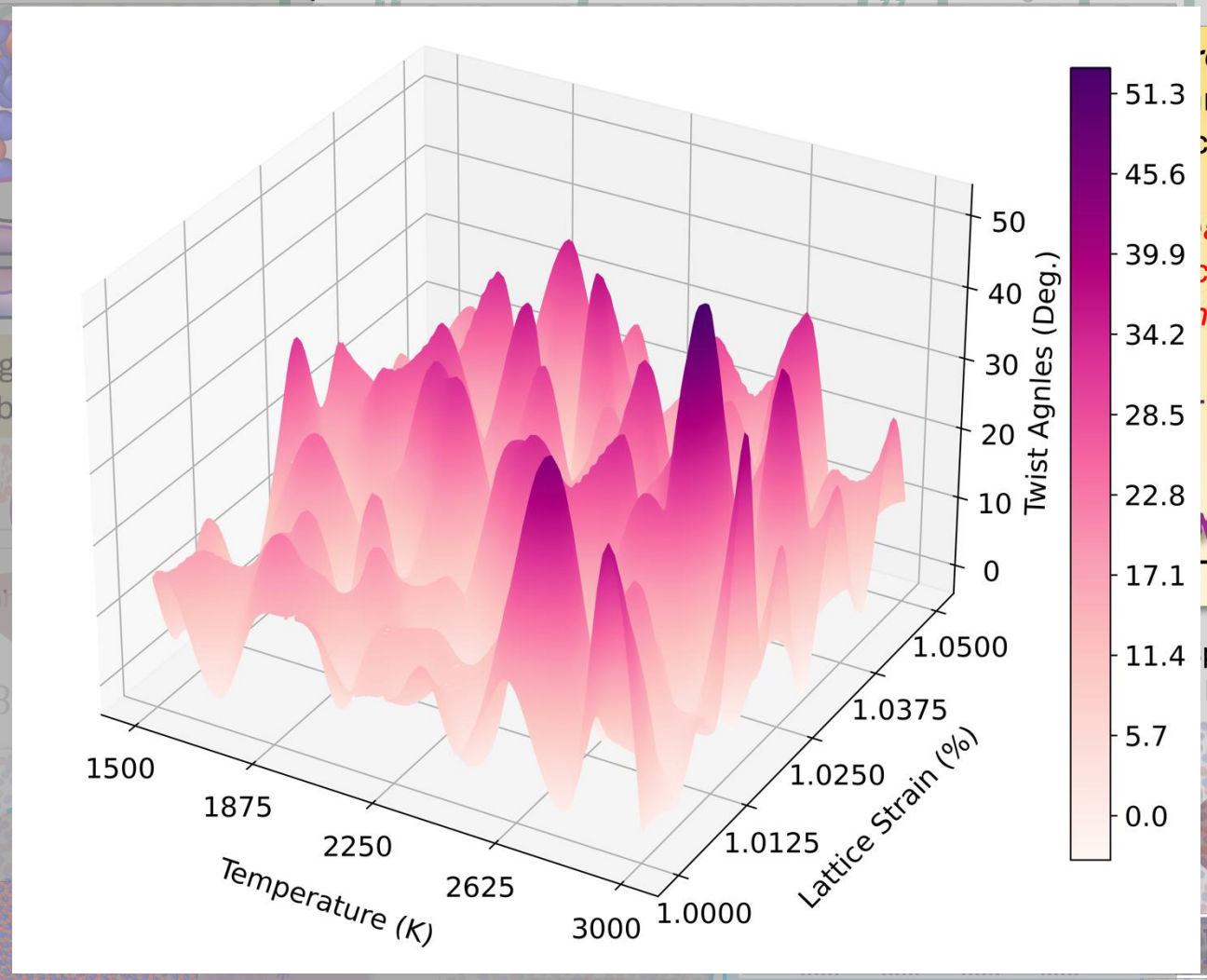
e. 15°



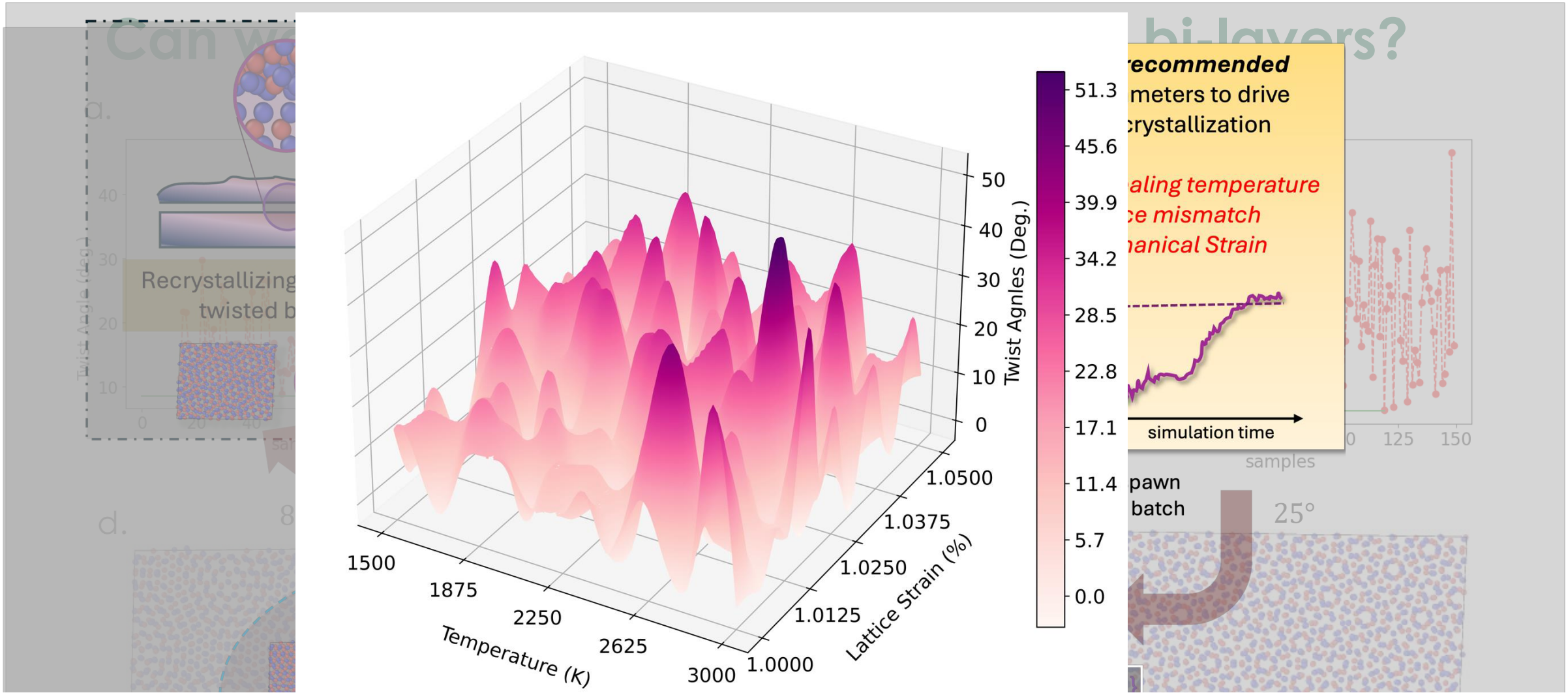
f. 25°



Bagchi S., Biswas A., Ghosh A., P. Ganesh (submitted)

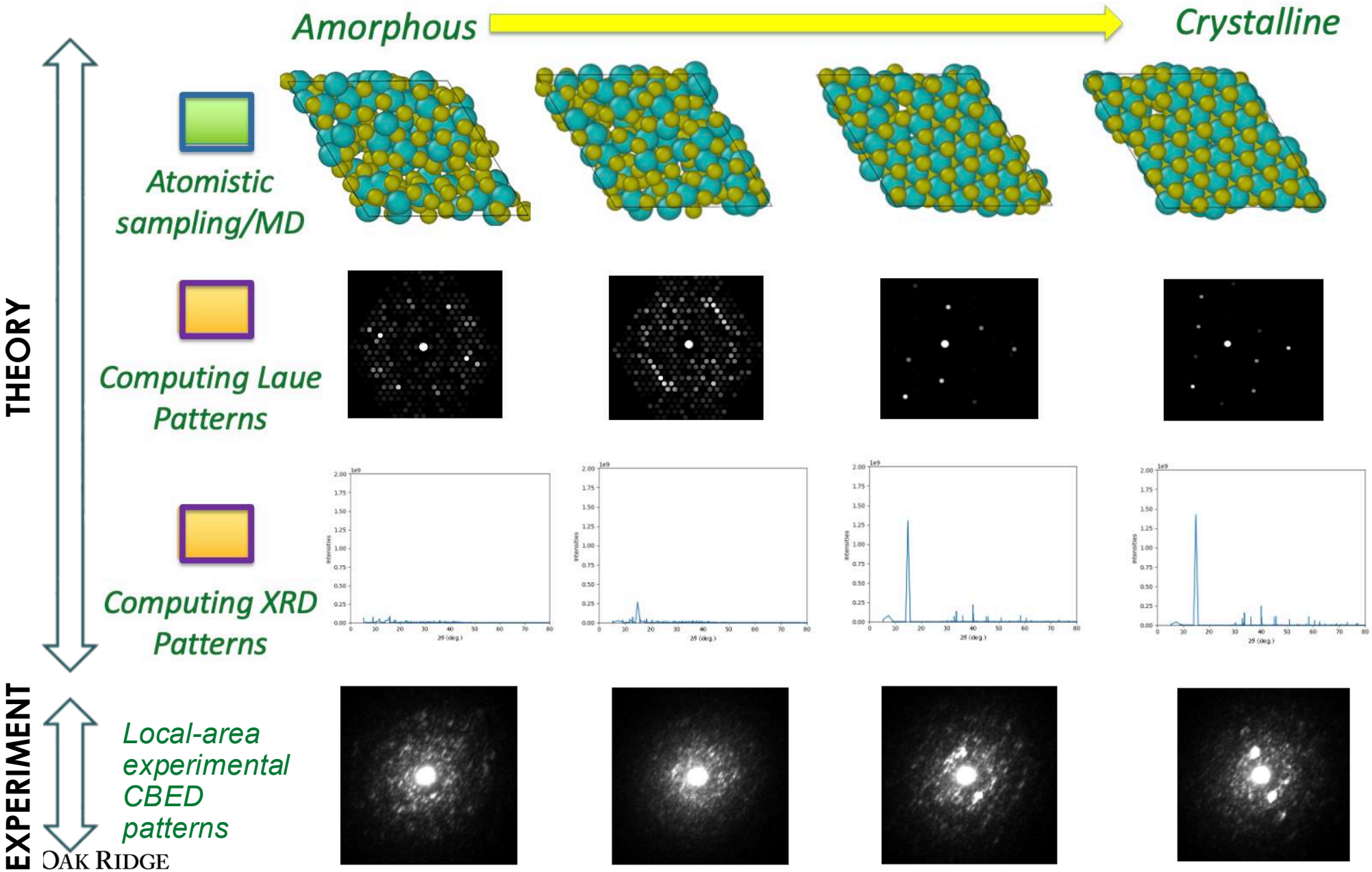


Bagchi S., Biswas A., Ghosh A., P. Ganesh (submitted)



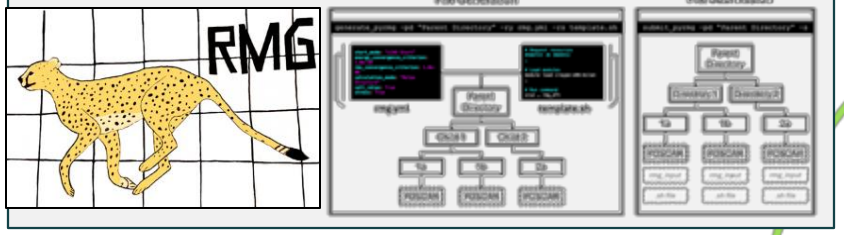
- Intuition alone would be insufficient to navigate this landscape
- Efficient active learning algos. that can learn from minimal explorations outside basins perform well
- Now we can start understanding why certain synthesis conditions lead to specific twisted angles, and effect of 'unknown unknowns' in real physical experiments !!

- Building on-the-fly digital characterization libraries for recrystallization pathways:



Use Case: Ensemble-FF-Fit

High-throughput *ab-initio* with RMG DFT
→ pyRMG



https://github.com/Q-CAD/pyRMG

(RMG @ NCSU)

Optimized ensemble of FFs (JAX-Reax / MACE / SNAP etc.) on Exascale

pyRMG – scalable high-throughput DFT on Exascale

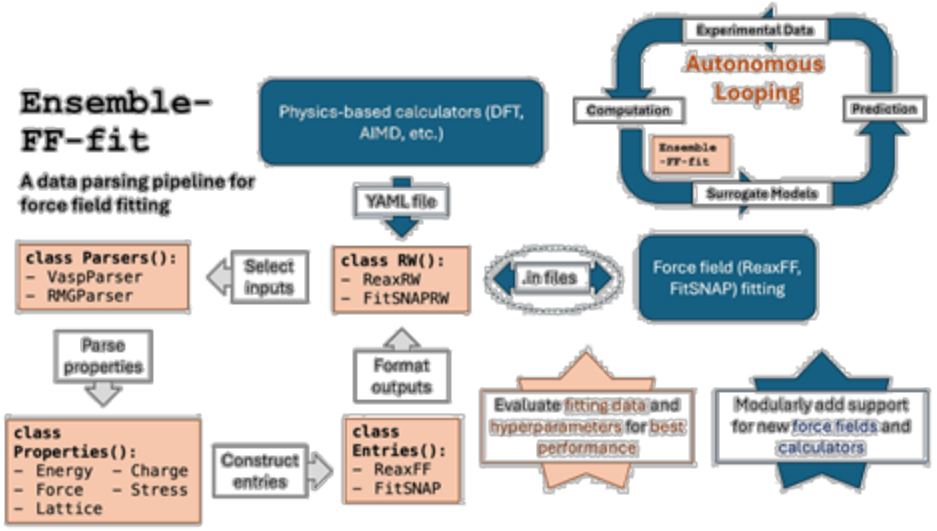
Autonomous orchestration using **MatEnsemble**

Generate 'Digital Twins'

XRD/ RHEED / STEM / STM

Extend training data optimally for non-equilibrium dynamics

Uncertainty Quantification

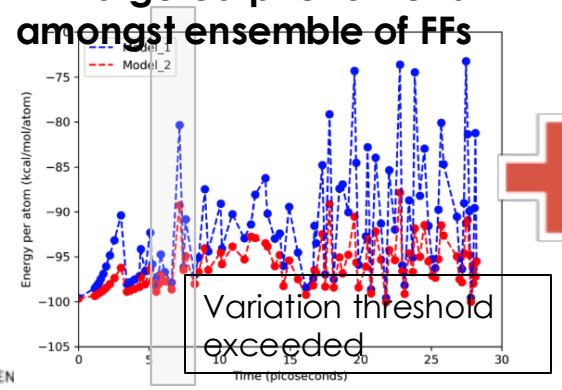


(will be available soon
-- can share a link if anyone wants to test)

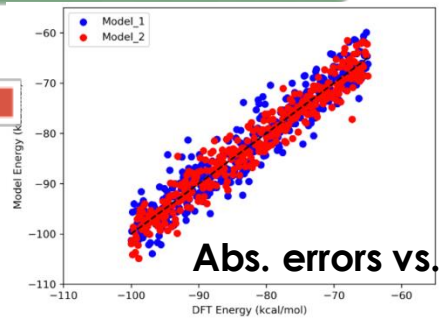
Experimental characterization

XRD
STM
STEM
RHEED
.....

Intrinsic variance in targeted phenomena amongst ensemble of FFs



Variation threshold exceeded



Abs. errors vs. DFT (or expt obs.)

ReaxFF – A Promising Model for Bi_2Se_3 Dynamics

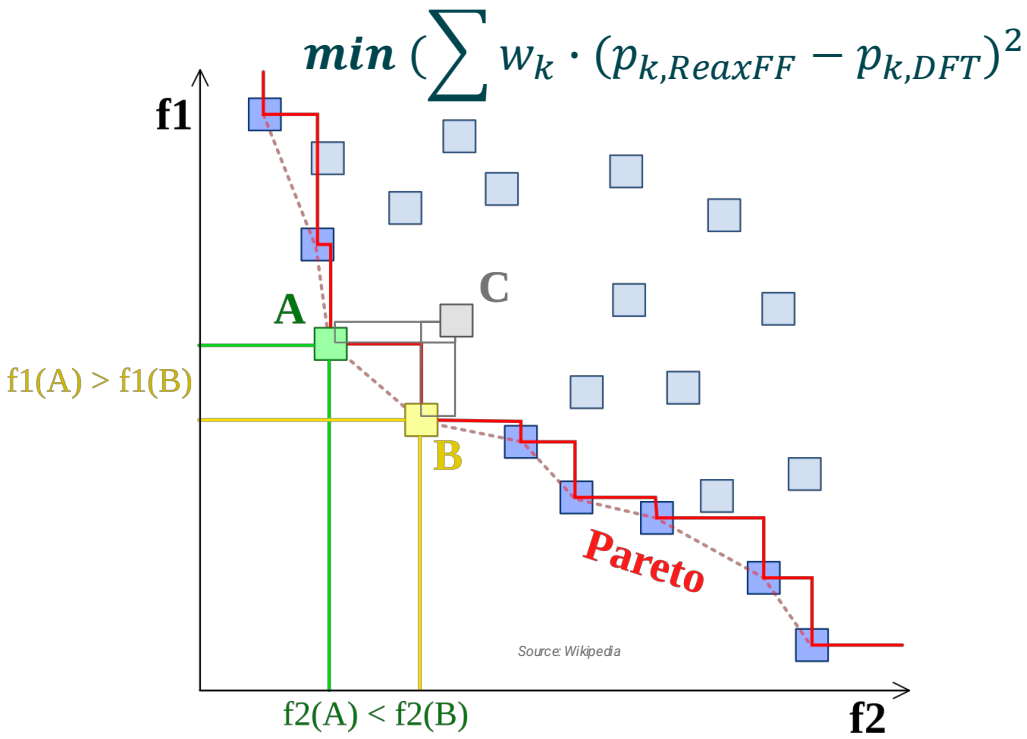
$$E_{system} = E_{bond} + E_{over} + E_{angle} + E_{torsion} + E_{vdWaals} + E_{Coulomb} + E_{Specific}$$

$$BO'_{ij} = exp\left[p_{bo,1} \cdot \left(\frac{r_{ij}}{r_o}\right)^{pbo,2}\right] + exp\left[p_{bo,3} \cdot \left(\frac{r_{ij}^{\pi}}{r_o}\right)^{pbo,4}\right] + exp\left[p_{bo,5} \cdot \left(\frac{r_{ij}^{\pi\pi}}{r_o}\right)^{pbo,6}\right]$$

$$\boldsymbol{min}\left(\sum w_k \cdot \left(p_{k,ReaxFF} - p_{k,DFT}\right)^2\right)$$

$p_k = training\ property\ k$

$w_k = weight\ of\ property\ k$

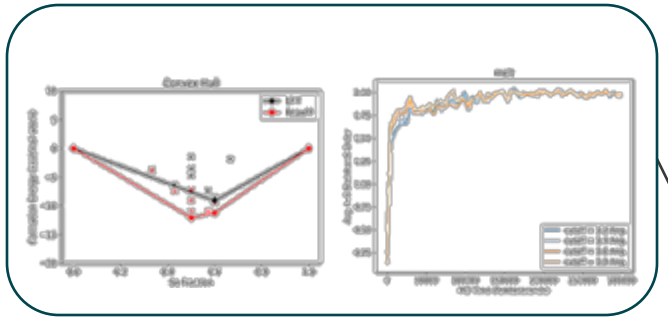


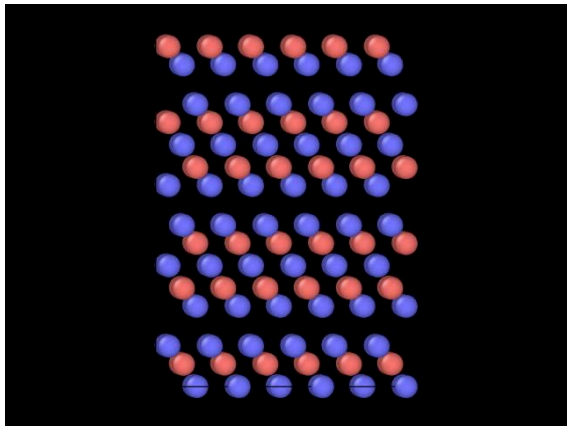
Under-constrained: Pareto front

Too complex: overflow or expensive function evaluation

Active learning based automated workflows lead to reliable FFs

Starting Force Field

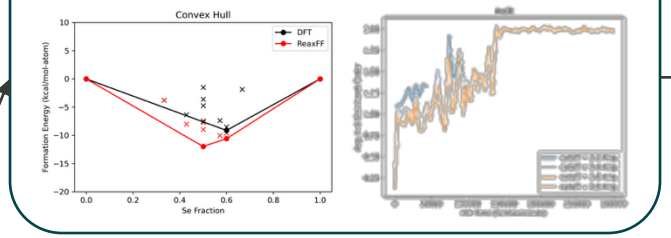




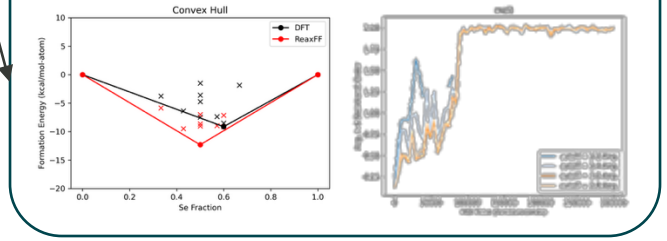
NVT Melt: 0 K to 2000 K, 6.67 K/ps

Iteration 2 (Data Sampling)

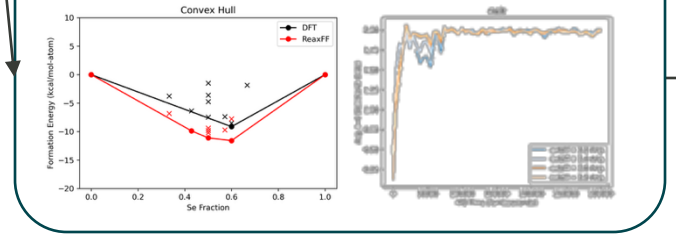
RMSE = 8391



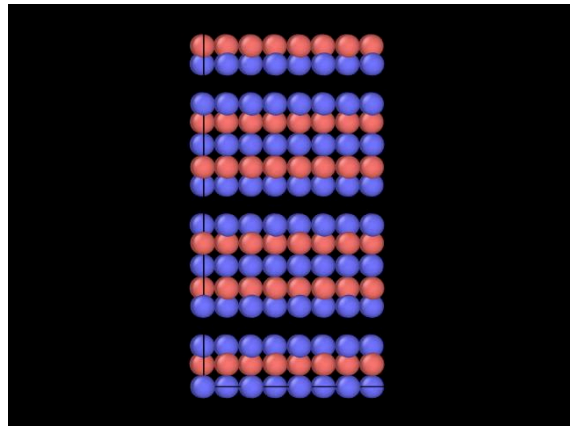
RMSE = 8173



RMSE = 8290

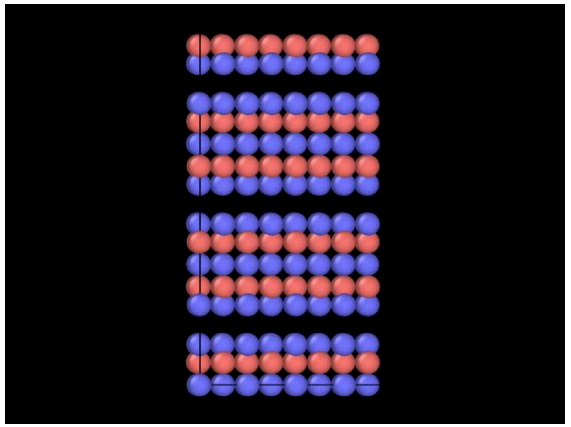


...



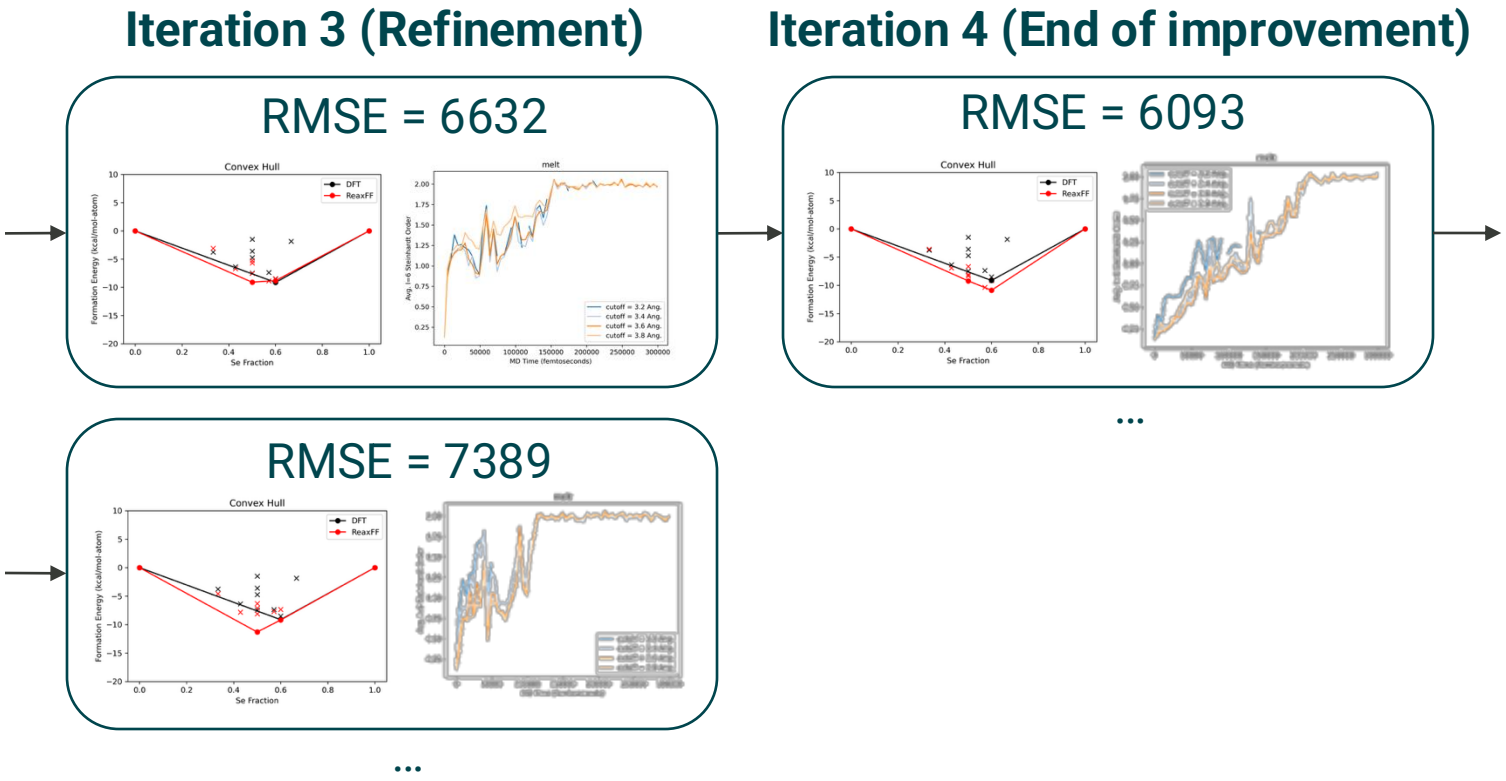
Selected for refitting

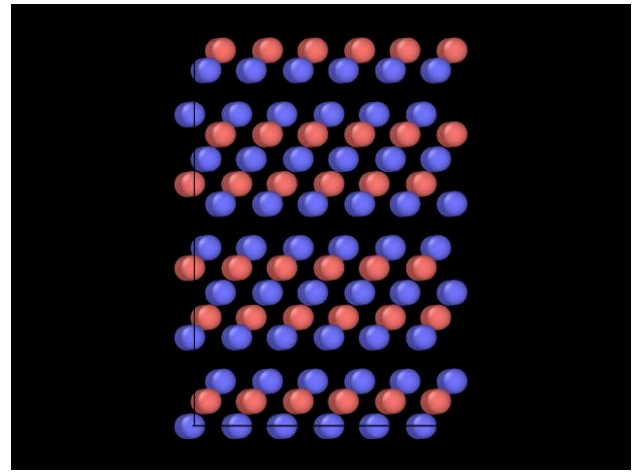
NVT Melt: 0 K to 2000 K, 6.67 K/ps



Rejected

Active learning based automated workflows lead to reliable FFs



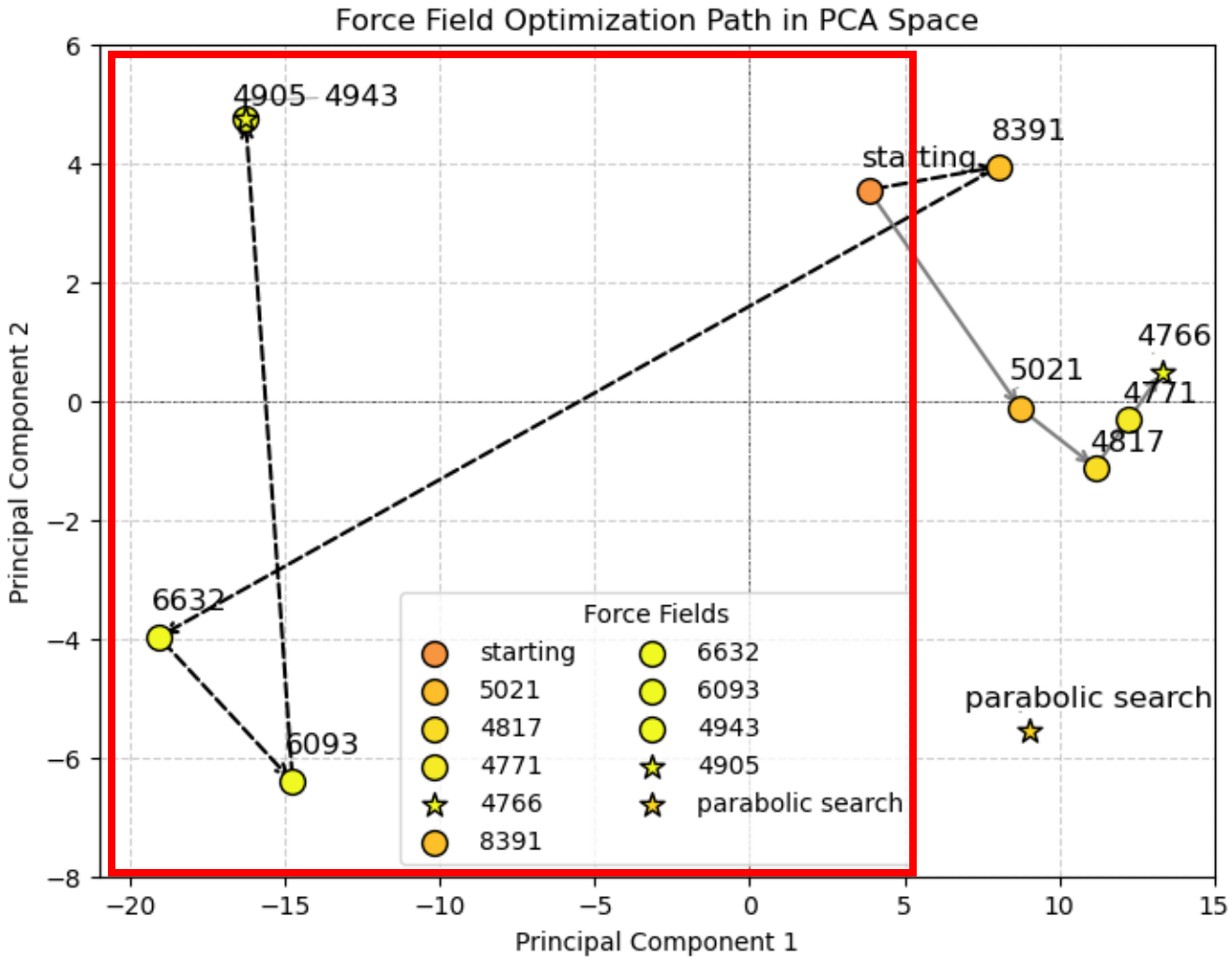


NVT Melt: 0 K to 2000 K, 6.67 K/ps

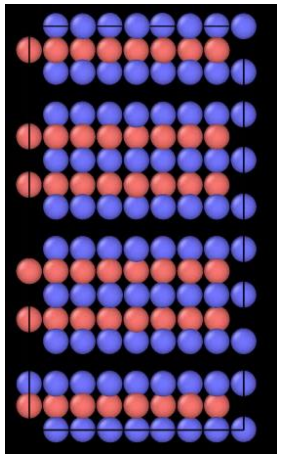
Objective function reduced and vdW-layering preserved up to ~750 K

4 serial iterations (100s of parallel fittings) performed in **2 days**

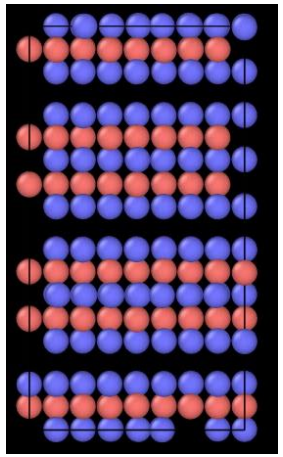
Active learning based automated workflows lead to reliable FFs



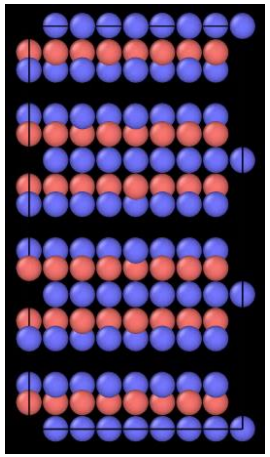
Example (RMSE = 4766)



MD Checks (RMSE = 4905)

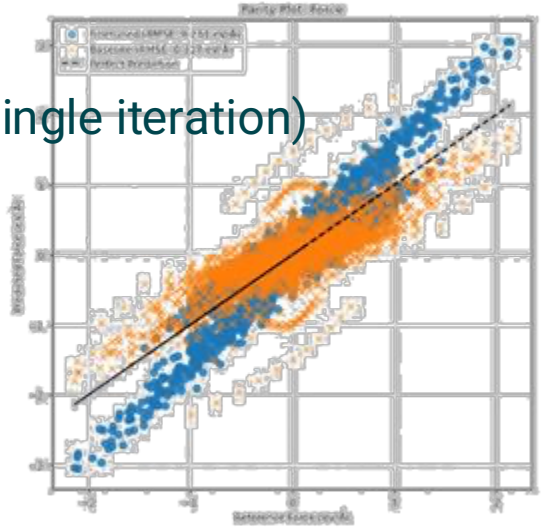


Parabolic Search



NPT Equilibration: 300 K

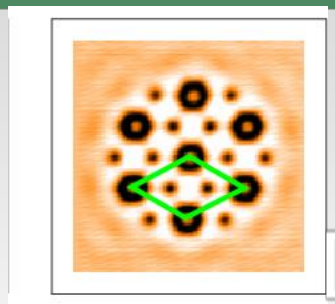
(single iteration)



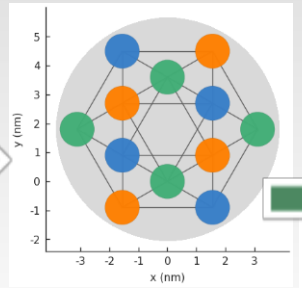
MACE fine-tuning shows similar levels of performance

Inverse learning methods – to bridge theory with expts. in new ways

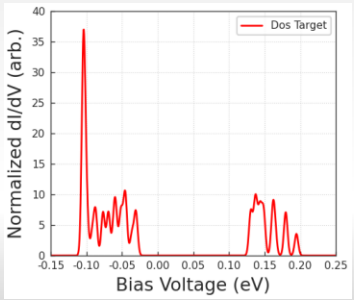
Spec2Ham



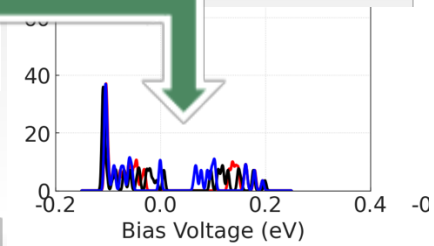
$H(t_a, t_b, v_a, v_b)$



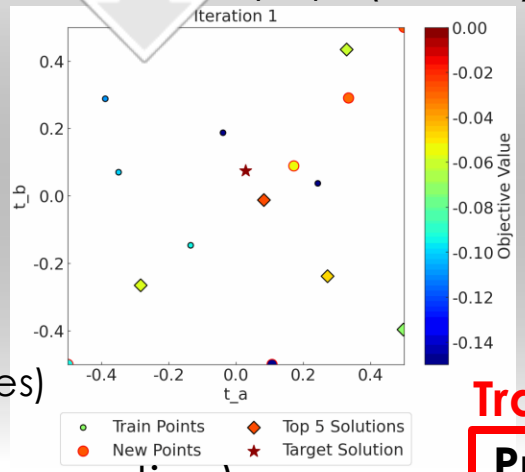
Solve at any level of accuracy (TB/RPA/DMFT)



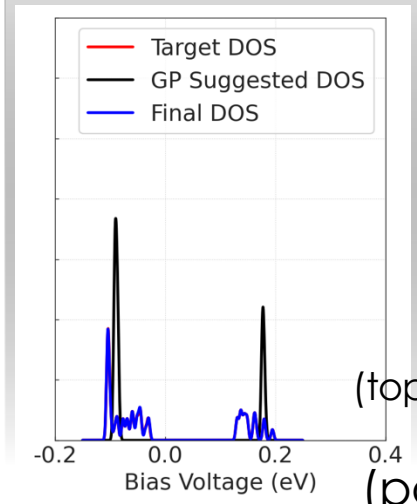
Loss ~ Wasserstein



BayOpt (50-iter.)



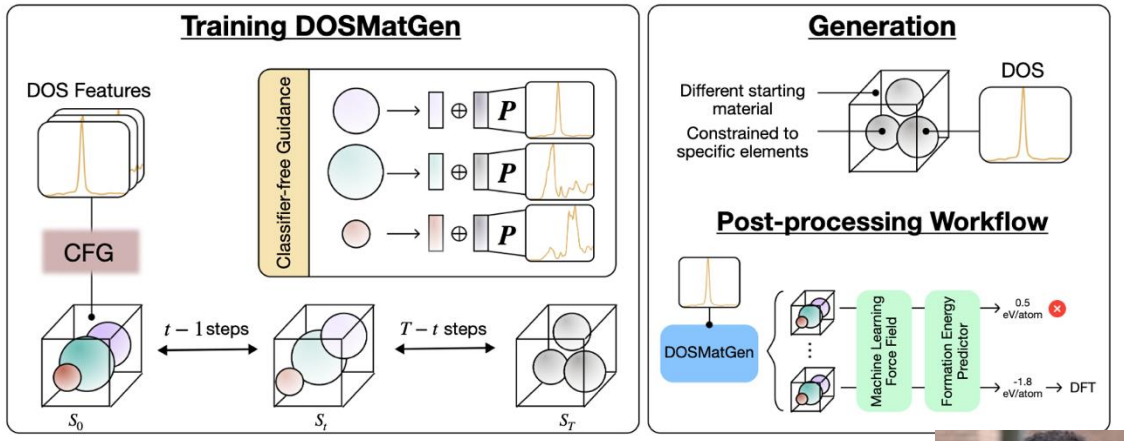
NumOpt (top 5 candidates)



(paper in preparation)

Spec2Struct (arXiv 2504.06249)

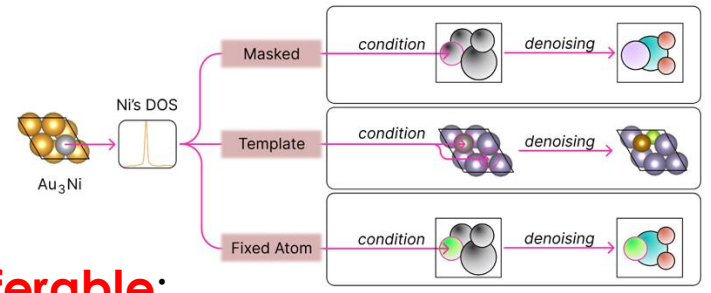
https://github.com/Q-CAD/Spec2Struct

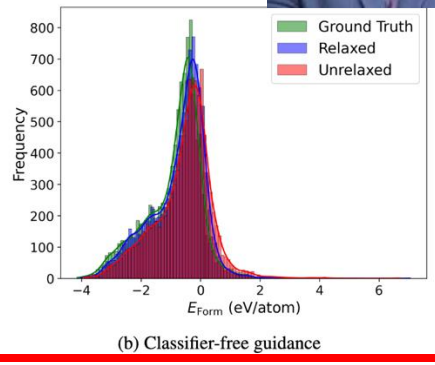


$$\nabla_{S_t} \log p_t(S_t|y) = \nabla_{S_t} \log p_t(S_t) + \nabla_{S_t} \log p_t(y|S_t),$$

Classifier-free guidance (

$$\nabla_{S_t} \log p_t(S_t, c), \quad c = \begin{cases} \varnothing & \text{with probability } p_{\text{uncond}} \\ y & \text{else} \end{cases}$$





Transferable:

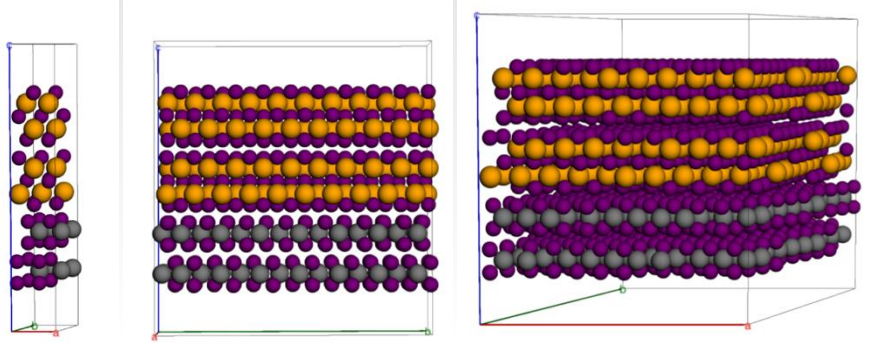
Pre-trained (50K PV) → Fine-tuned (~500 2D M-X + defects)

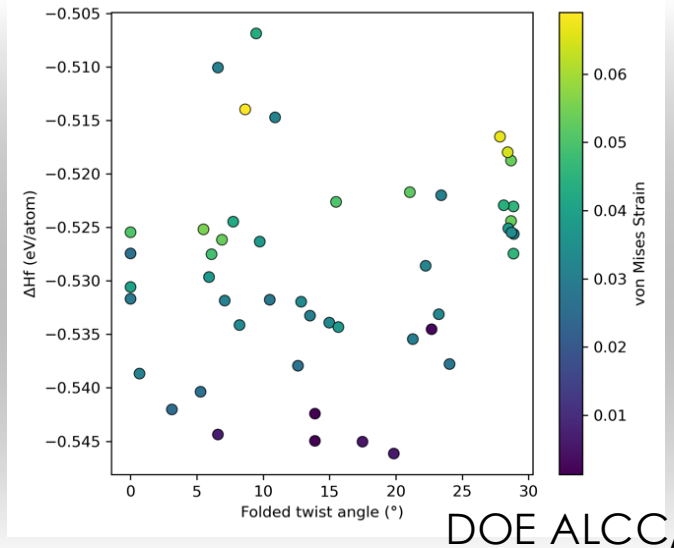
New RMG DFT capabilities (collaboration with Prof. Bernholc @ NCSU)

https://github.com/RMGDFT/rmgdft

pyRMG high-throughput DFT

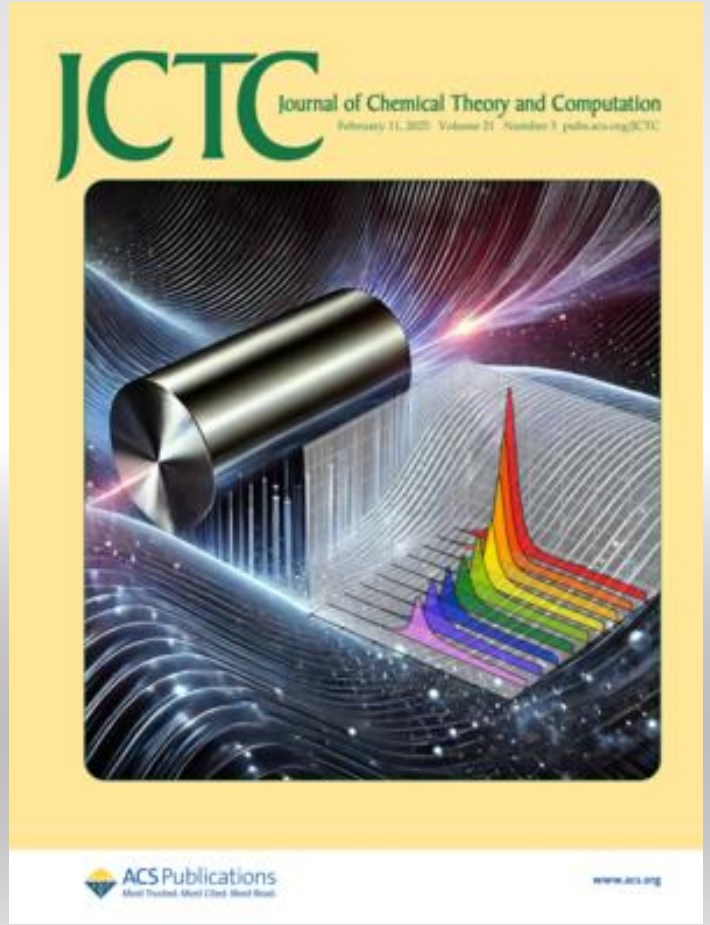
- ~few-tens-thousand elec. DFT / AIMD ; https://github.com/Q-CAD/pyRMG





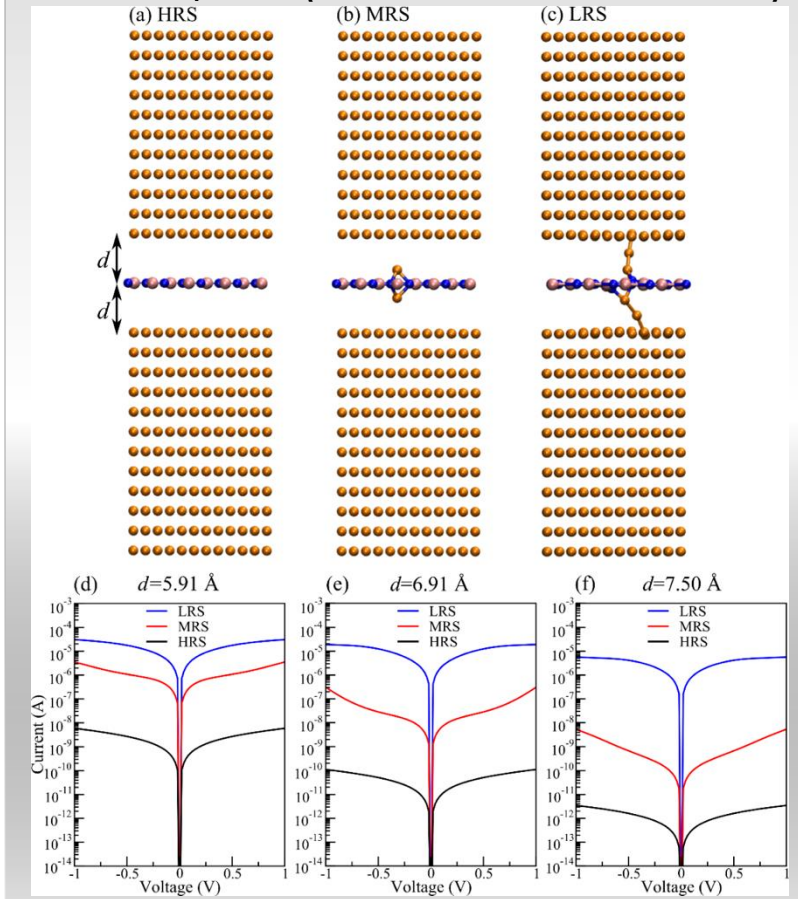
Large-scale RT TD-DFT

- ~up to 20-thousand elec. TDDFT;



Linear-scaling NEGF

- ~few-tens-thousand elec. Transport (ACS. Nano. coming soon)

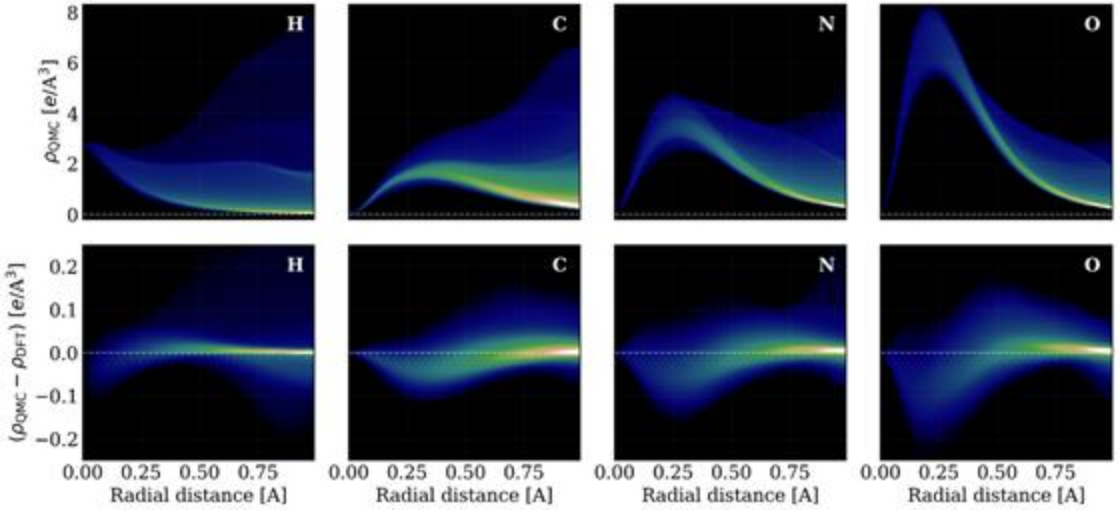


DOE ALCC/INCITE: "*Ab Initio Simulations of Out-of-Equilibrium Heterogeneous Quantum Materials*" & "*Integrated Exascale Computational Workflows for Accelerated Material Synthesis*"

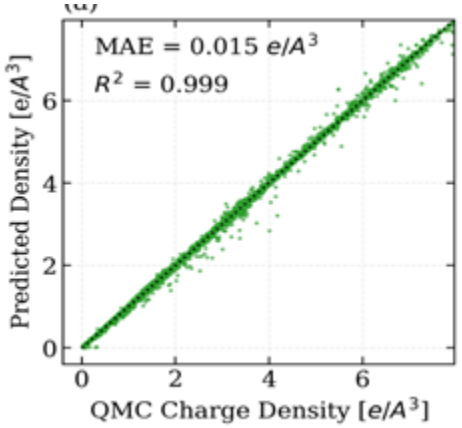
Benchmark QMC Datasets / Machine-learning QMC densitie(s)

QMCPACK https://qmcpack.org/

VMC → $\Psi(\mathbf{X}) = e^{J(\mathbf{X})} D(\mathbf{X})$ $E_T = \frac{1}{M}\sum_{m=1}^{M} \frac{\hat{H}\Psi_T(\mathbf{R}_m)}{\Psi_T(\mathbf{R}_m)} + \mathcal{O}\left(\frac{1}{\sqrt{M}}\right) \geq E_{exact}$

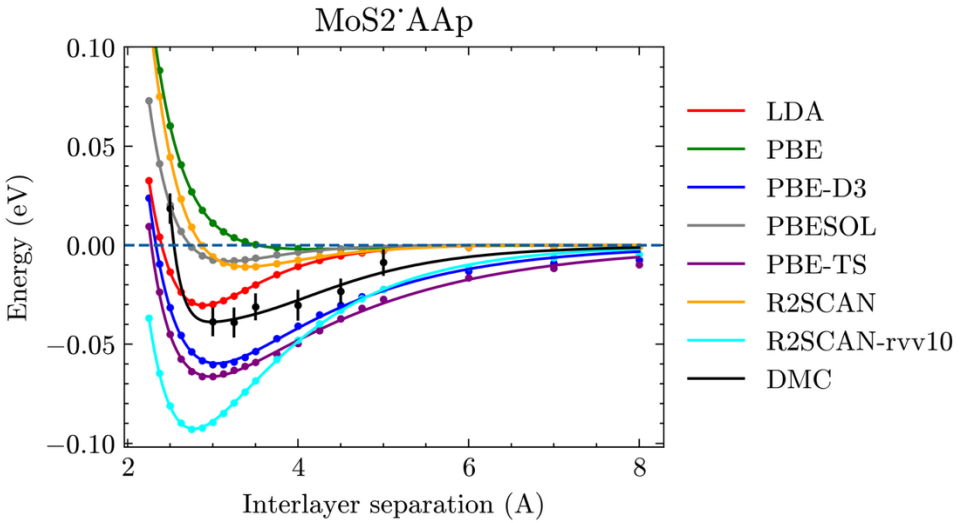


Gani @ ORNL
& Victor@GTech



QMC quality charge/spin/KE – densities using a surrogate GNN model (trained on our QM9/VMC dataset) (soon to be released)

Kayahan
@ORNL



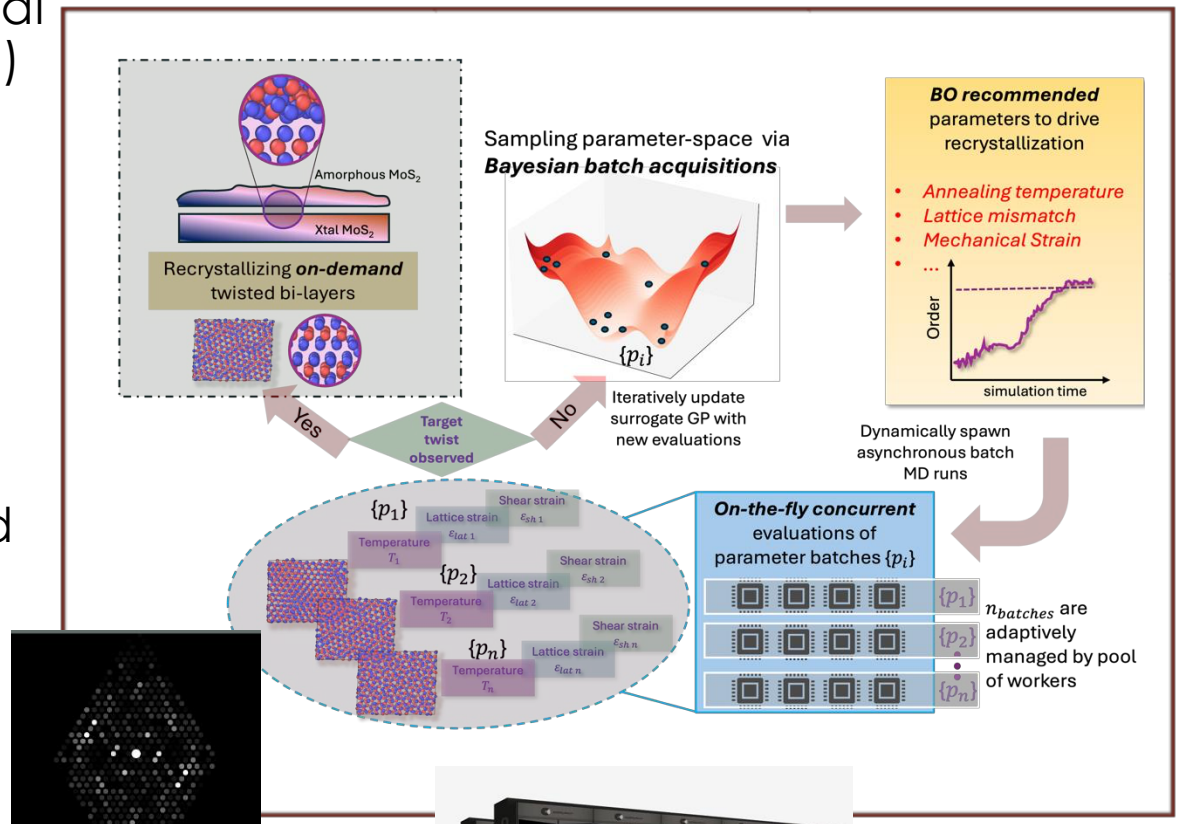
QMC database of monolayer, bilayer and bulk 2D materials (+ automated workflows/recipes)

(Center for Predictive Simulation of Functional Materials – DOE CMS @ ORNL)

Summary: Towards autonomous experimentation via integrating multi-fidelity theory, ML-models **and** active learning algorithms empowered by extreme-scale computing

- A digital platform enabled by adaptive and optimal task management based on large-scale (Exascale) resources can serve as testbeds for advanced/complex active learning algorithms
- Some use cases: accelerated FF fitting and digital-twins for autonomous (computational) synthesis
- ML for inverse problems: spec2struct & spec2Ham
- New ab initio development: large-scale DFT ground /excited-states / transport / QMC quality 2D datasets / ML QMC densities

Current work: on-the-fly real time inferences and seamless integration with AE platforms for MBE/PLD/STM/STEM.

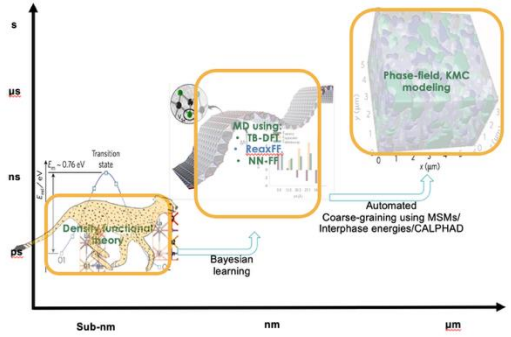


Email: ganeshp@ornl.gov

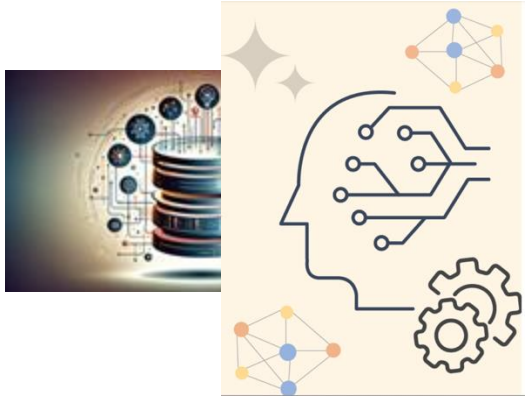
Few high-level discussion points for the workshop

- What do we mean by digital twins in materials science ?
- How do we assess the accuracy and provenance of the digital twins ?
- What are best practices to combine digital twins with physical experiments to accelerate discovery of new materials and its science ?
- How do we leverage AI/ML algorithms to extract pertinent information from digital twins ?
- What tools/infrastructure do we need to create/share digital twin datasets, AI-models and automation of physical experiments ?

Multi-scale theory/modeling/simulation



Database & AI - models



Autonomous experiments