Accurate Materials Predictions with DFT & Machine Learning

Noa Marom

Materials Science & Engineering

Carnegie Mellon University

Machine Learning in Materials Simulations

Machine learning: A statistical model is built based on available "training" data to predict the results of future experiments

Applications in computational materials science:

- Machine learned inter-atomic potentials
- Machine learned DFT functionals
- Clustering
- Identifying correlations in data
- Feature selection
- Property prediction
- Optimization (e.g., Bayesian optimization)

Ingredients:

- Training data
- Representation
- Model type
- Model hyperparameters
- Validation

ML models can only interpolate, not extrapolate

It may be challenging to learn from "small data". Incorporating physical knowledge into models can help

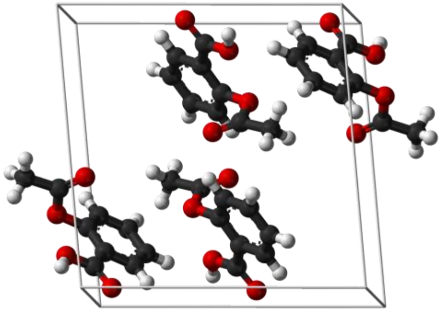
The application of ML models in materials simulations is usually not "black box" and some customization is required

A Machine Learned Model for Molecular Crystal Volume Estimation

Molecular Crystals

Used for *e.g.*, pharmaceuticals, organic electronics

Weak dispersion (van der Waals) interactions produce potential energy landscapes with many local minima close in energy



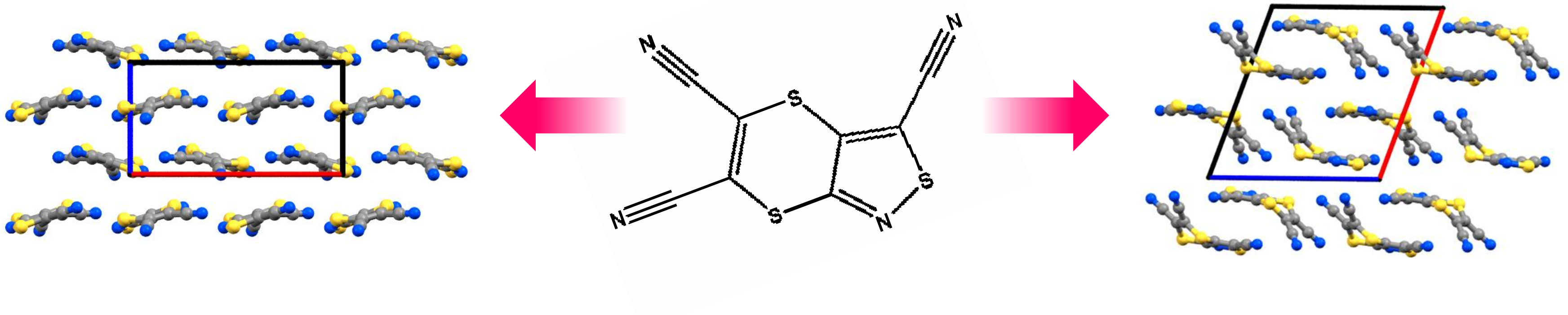
Aspirin crystal

Molecular crystals often exhibit polymorphism, the ability of the same molecule to crystallize in several structures

Polymorphs may have different physical/chemical properties!

The challenge: given a 2D stick diagram of a molecule, predict all of its possible polymorphs

Requires searching a high-dimensional space with a high accuracy



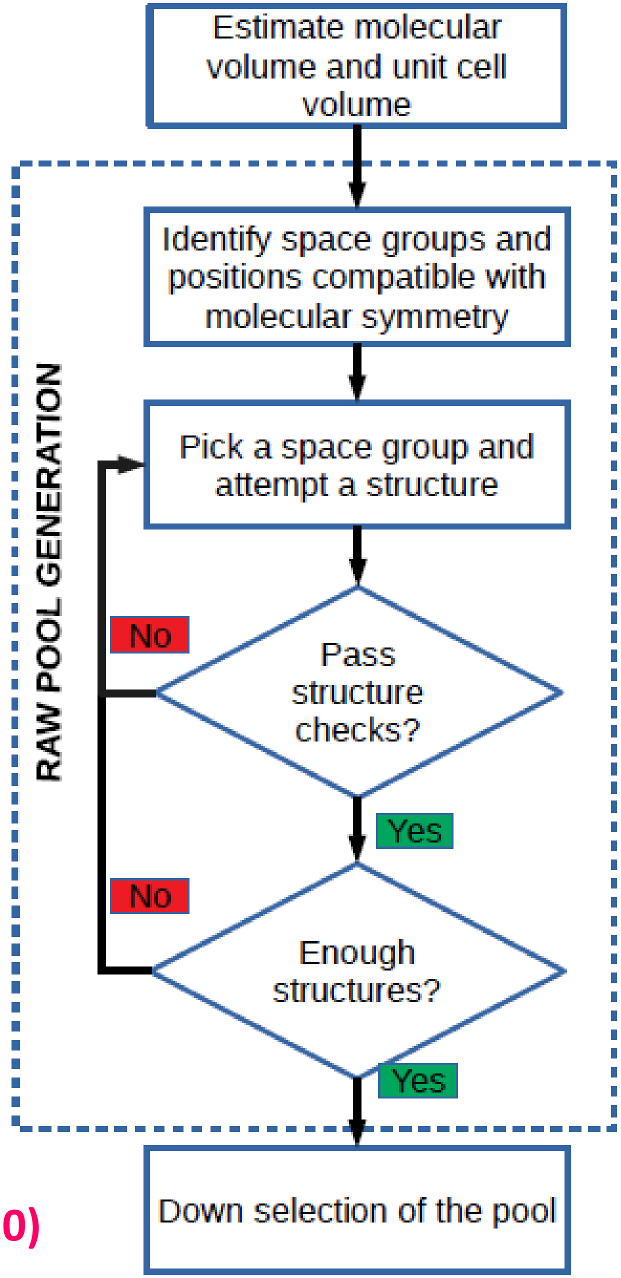
Molecular Solid Form Volume

The molecular solid form volume is the effective volume occupied by a molecule in a crystal:

$$V_M = \frac{V_{cell}}{Z}$$

Crystal structure prediction workflows often begin by estimating the solid form volume to define the search space

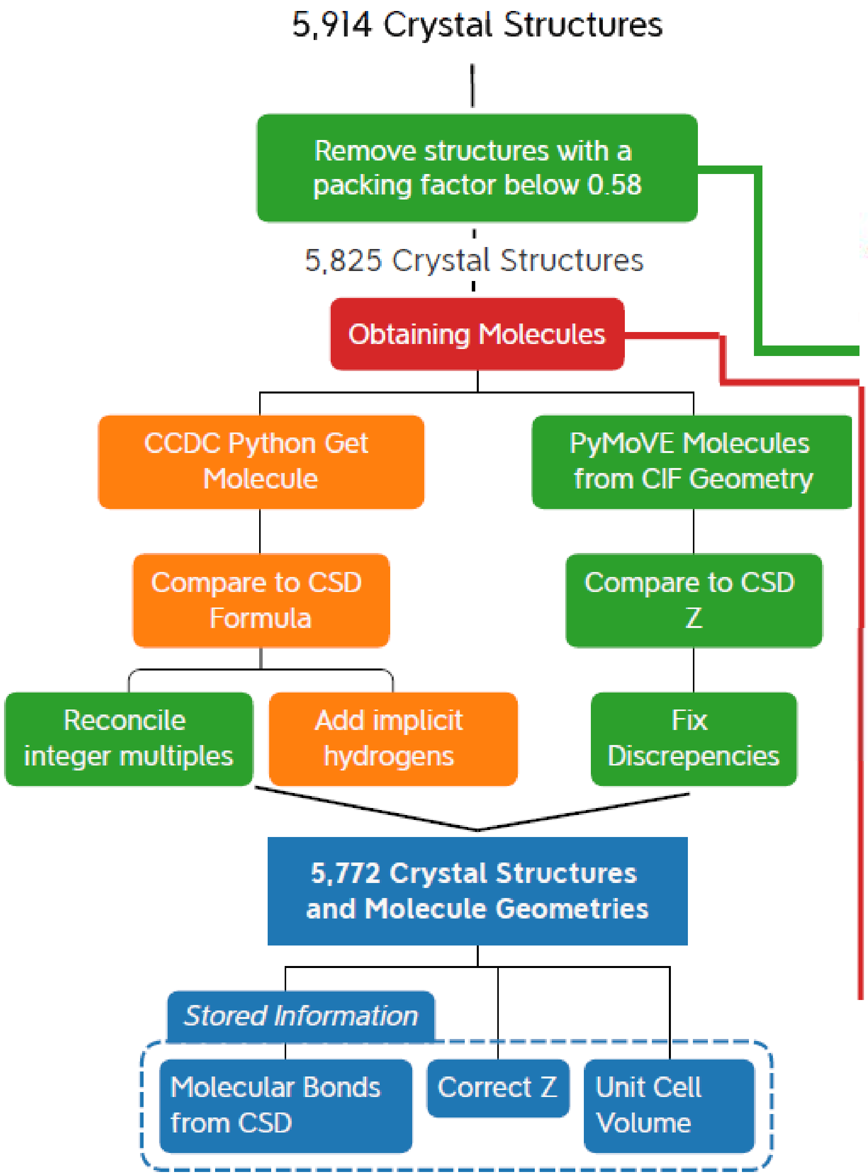
Workflow of the Genarris random structure generator for molecular crystals


Estimate molecular volume and unit cell volume
RAW POOL GENERATION
Identify space groups and positions compatible with molecular symmetry
Pick a space group and attempt a structure
No
Pass structure checks?
Yes
No
Enough structures?
Yes
Down selection of the pool

We developed a machine learned model to predict V_M, given the single molecule structure

R. Tom, T. Rose, I. Bier, H. O'Brien, A. Vazquez-Mayagoitia, and N. Marom, *Comput. Phys. Commun.* 250, 107170 (2020)

ML Model for Volume Estimation: Training Data



5,914 Crystal Structures

Remove structures with a packing factor below 0.58

5,825 Crystal Structures

Obtaining Molecules

CCDC Python Get Molecule

PyMoVE Molecules from CIF Geometry

Compare to CSD Formula

Compare to CSD Z

Reconcile integer multiples

Add implicit hydrogens

Fix Discrepencies

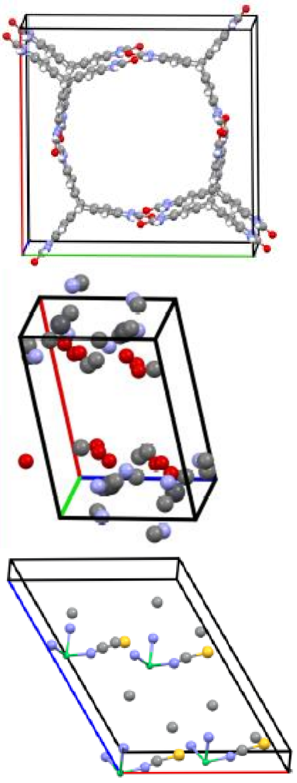
5,772 Crystal Structures and Molecule Geometries

Stored Information

Molecular Bonds from CSD

Correct Z

Unit Cell Volume



The performance of ML models depends on the quality of the training data

The data should be consistent and reliable

A set of polymorphic crystal structures characterized in ambient temperature and pressure conditions was extracted from CSD 2019

Porous structures were removed

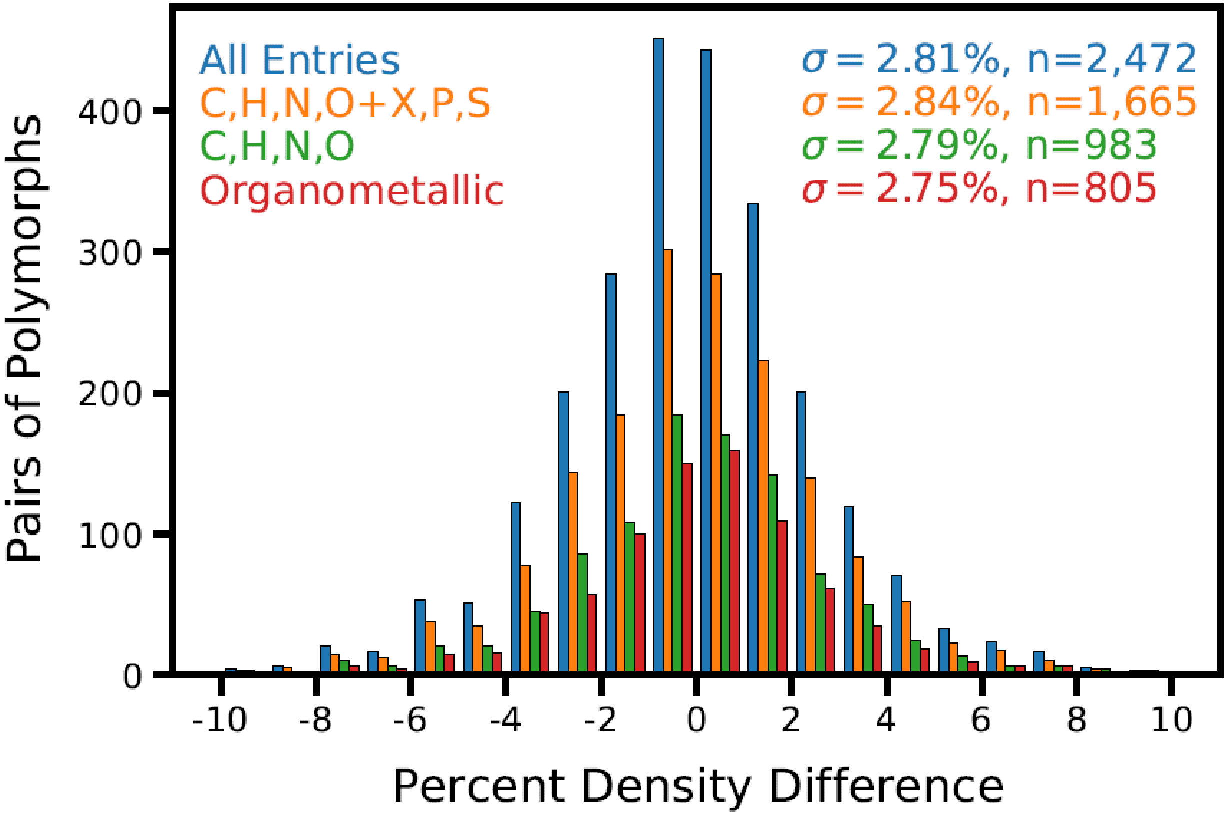
Problems, such as discrepancies in Z values and chemical formula were corrected

I. Bier and N. Marom, *J. Phys. Chem. A* 124, 10330 (2020)

ML Model for Volume Estimation: Training Data

The final training set contained 2,472 unique pairs of polymorphs

The standard deviation of the percent density difference between polymorphs may be considered as a lower bound for the error of a ML model

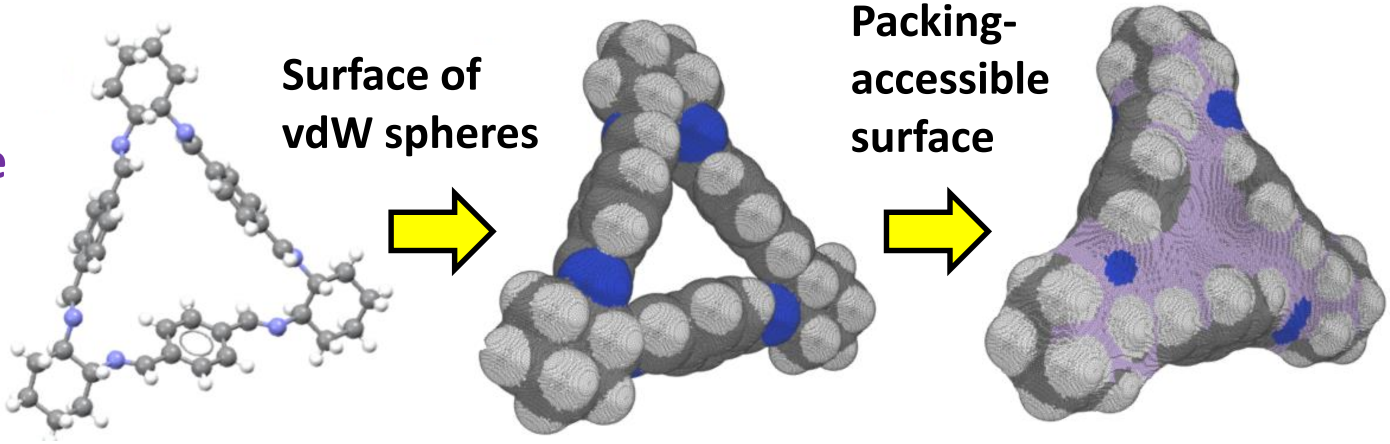


I. Bier and N. Marom, *J. Phys. Chem. A* 124, 10330 (2020)

ML Model for Volume Estimation: Model Features

The ML model is based on a combination of geometric and chemical descriptors that capture the salient features of molecular crystals

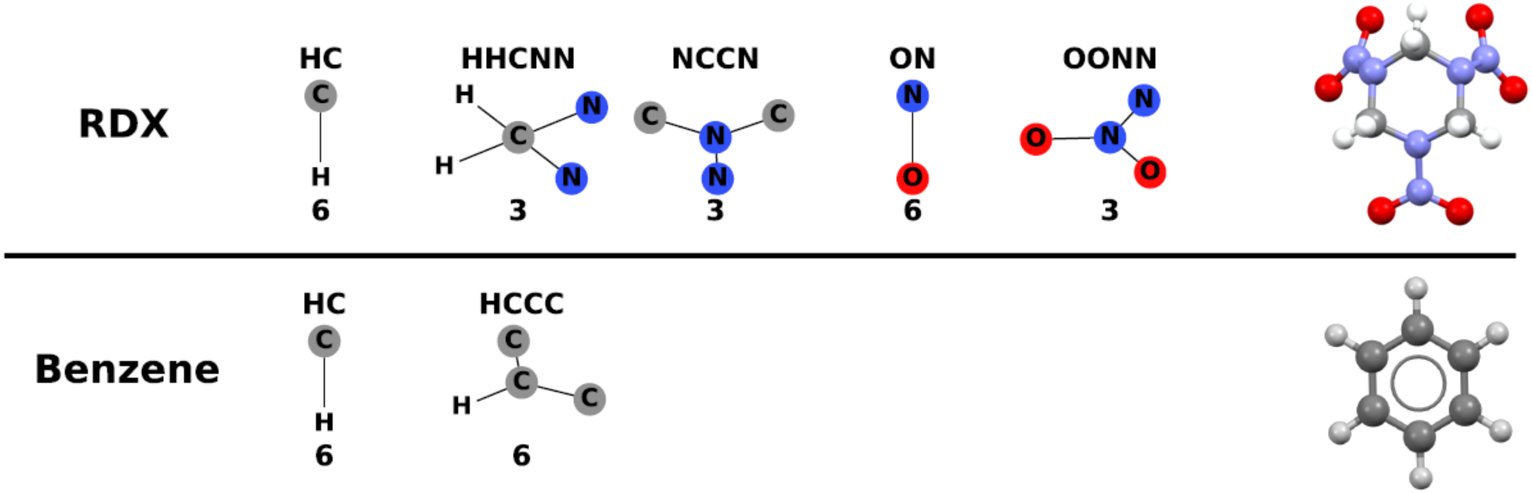
Geometric descriptor: volume enclosed by the packing accessible surface



Surface of vdW spheres

Packing-accessible surface

Chemical descriptor: molecular topological fragments



I. Bier and N. Marom, *J. Phys. Chem. A* 124, 10330 (2020)

ML Model for Volume Estimation: Model Training

The predicted solid form volume is given by:

$$V_M = \beta_0 V_0 + \sum_{i=1}^{n} \beta_i f_i$$

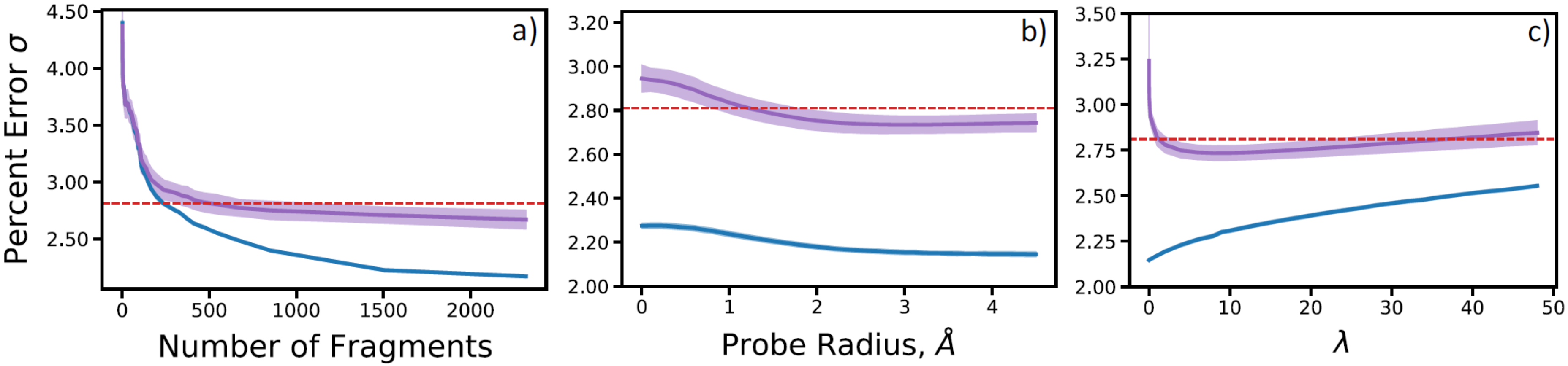
The coefficients are found by minimizing the ridge regression loss function:

$$L(\beta) = \sum_{j=1}^{N} \left(V_{CSD,j} - V_{M,j}\right)^2 + \lambda \sum_{i=0}^{n} {\beta_i}^2$$

The ML model has three hyper-parameters:

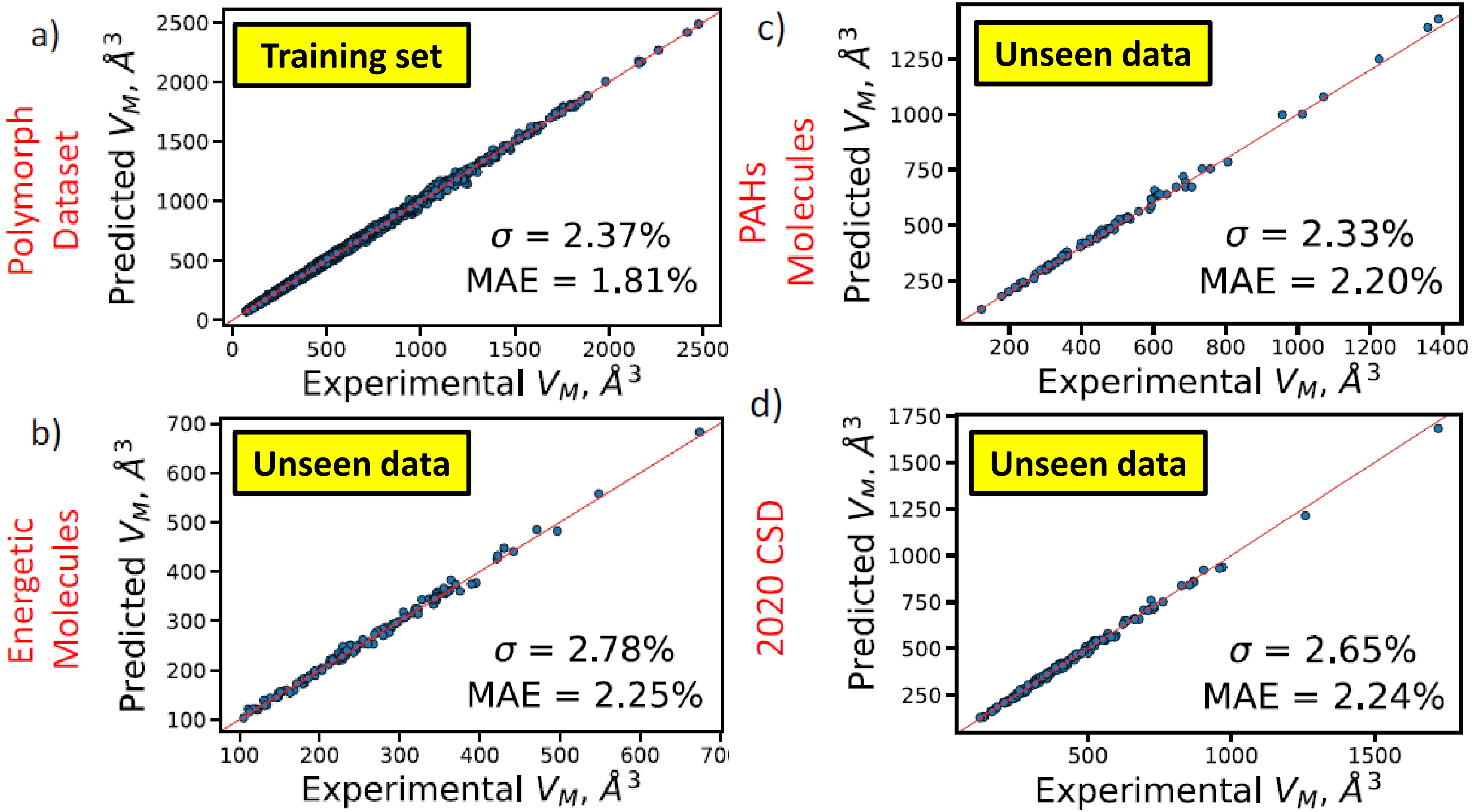
- Number of molecular topological fragments
- Probe radius for packing-accessible surface construction, α
- Ridge regression regularization parameter, λ

- The parameters were optimized by a 3D grid search over 54,810 combinations
- 10-fold cross validation was performed for each set of parameters
- Optimal values found: 2,231 fragments; α = 3 Å; λ = 10



I. Bier and N. Marom, *J. Phys. Chem. A* 124, 10330 (2020)

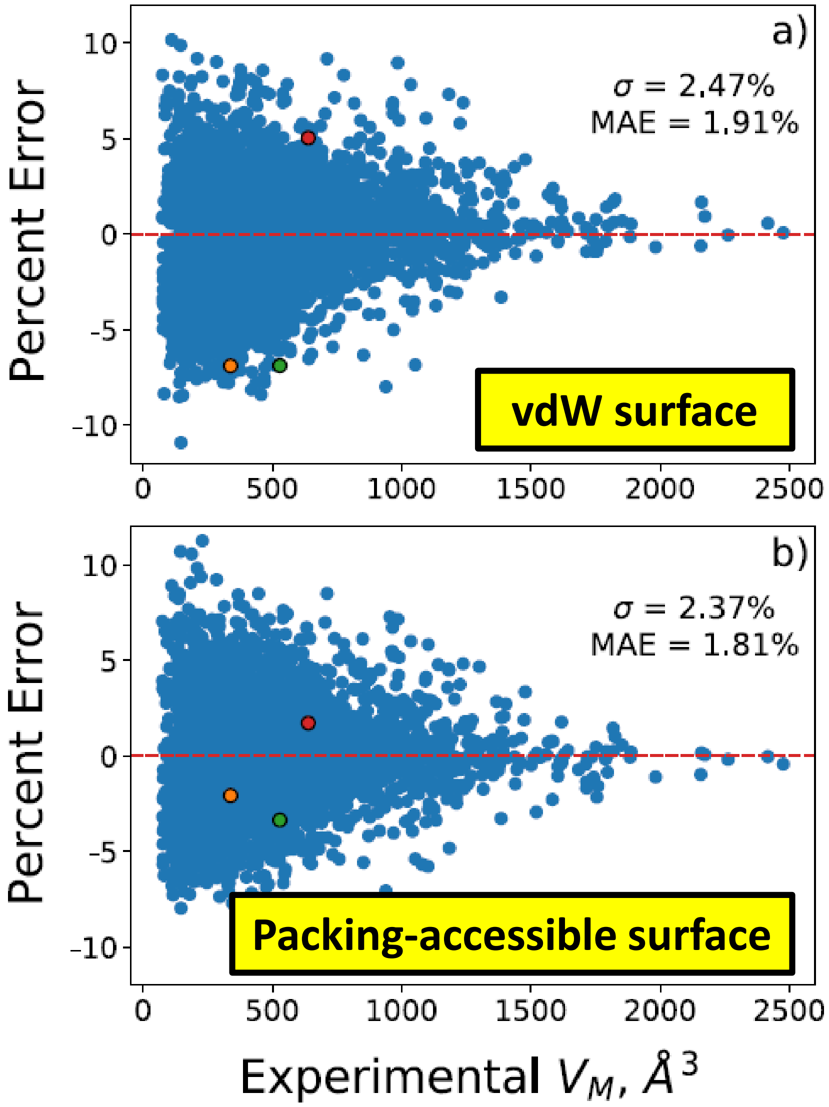
ML Model for Volume Estimation: Results

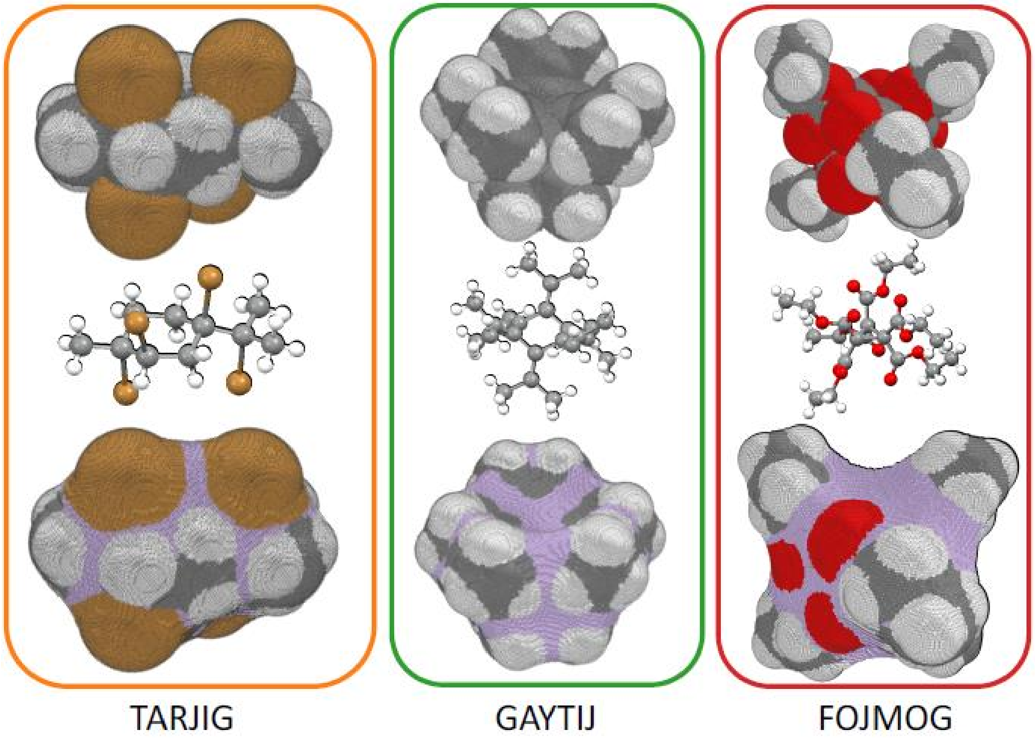


The model performs well for the training set and *three sets of unseen data* with errors below the presumed lower bound

I. Bier and N. Marom, *J. Phys. Chem. A* 124, 10330 (2020)

ML Model for Volume Estimation: Results

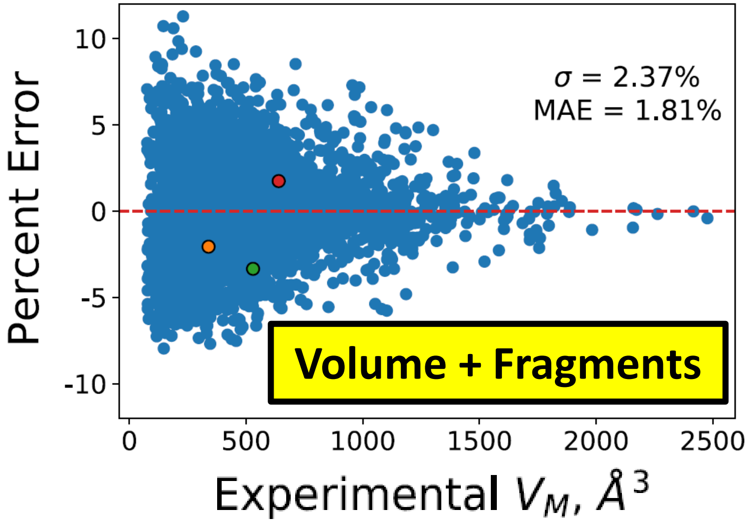




The volume enclosed by the packing-accessible surface captures the effect of sterically hindered regions and voids

I. Bier and N. Marom, *J. Phys. Chem. A* 124, 10330 (2020)

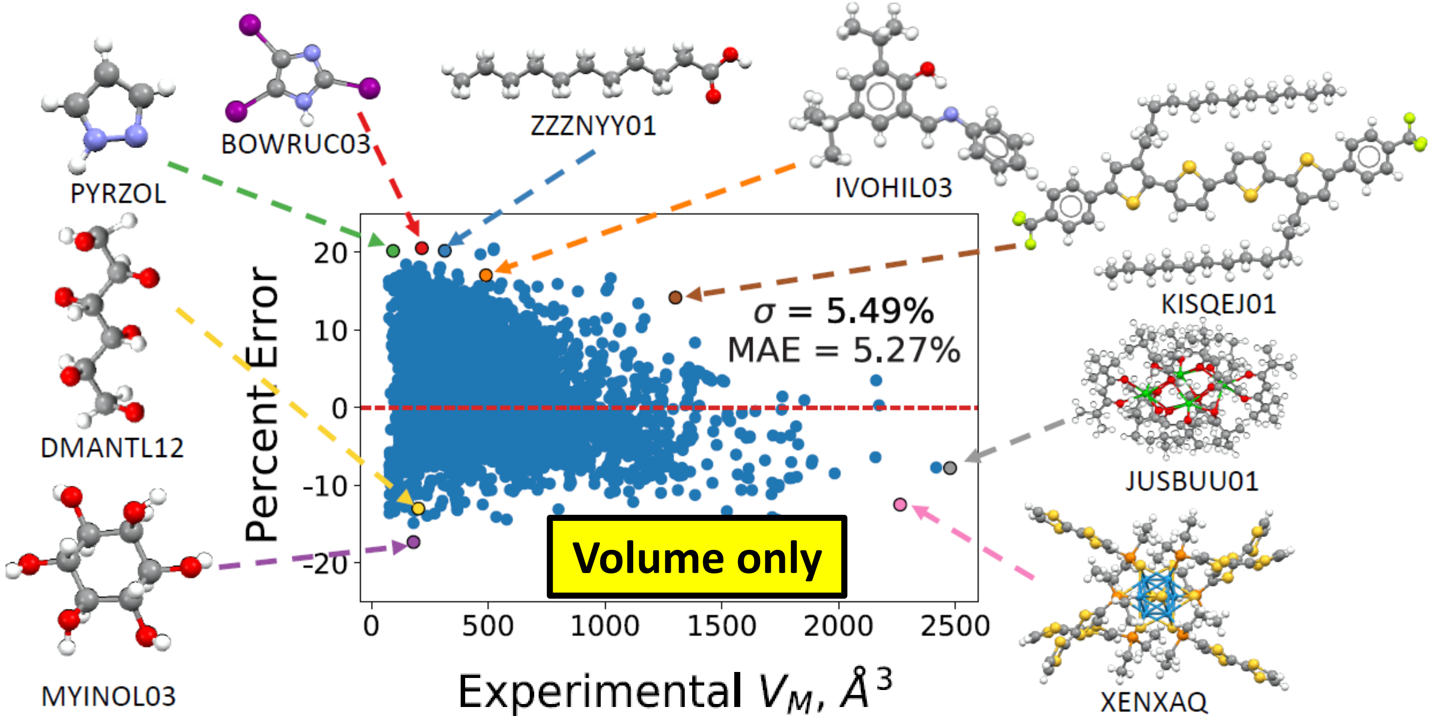
ML Model for Volume Estimation: Results



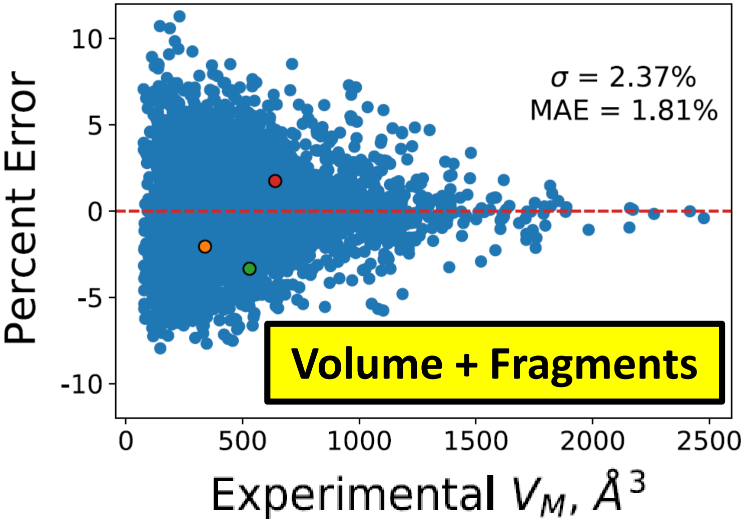
A model based only on the volume enclosed by the packing-accessible surface, without chemical information, has a broader error distribution

Outliers include materials with strong attractive interactions, such as H-bonds, repulsive groups, such as halogens, N lone-pairs, and alkyl side chains

I. Bier and N. Marom, *J. Phys. Chem. A* 124, 10330 (2020)

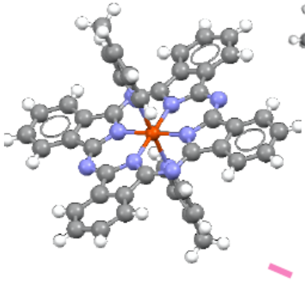


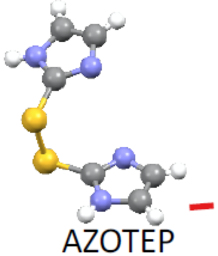
ML Model for Volume Estimation: Results

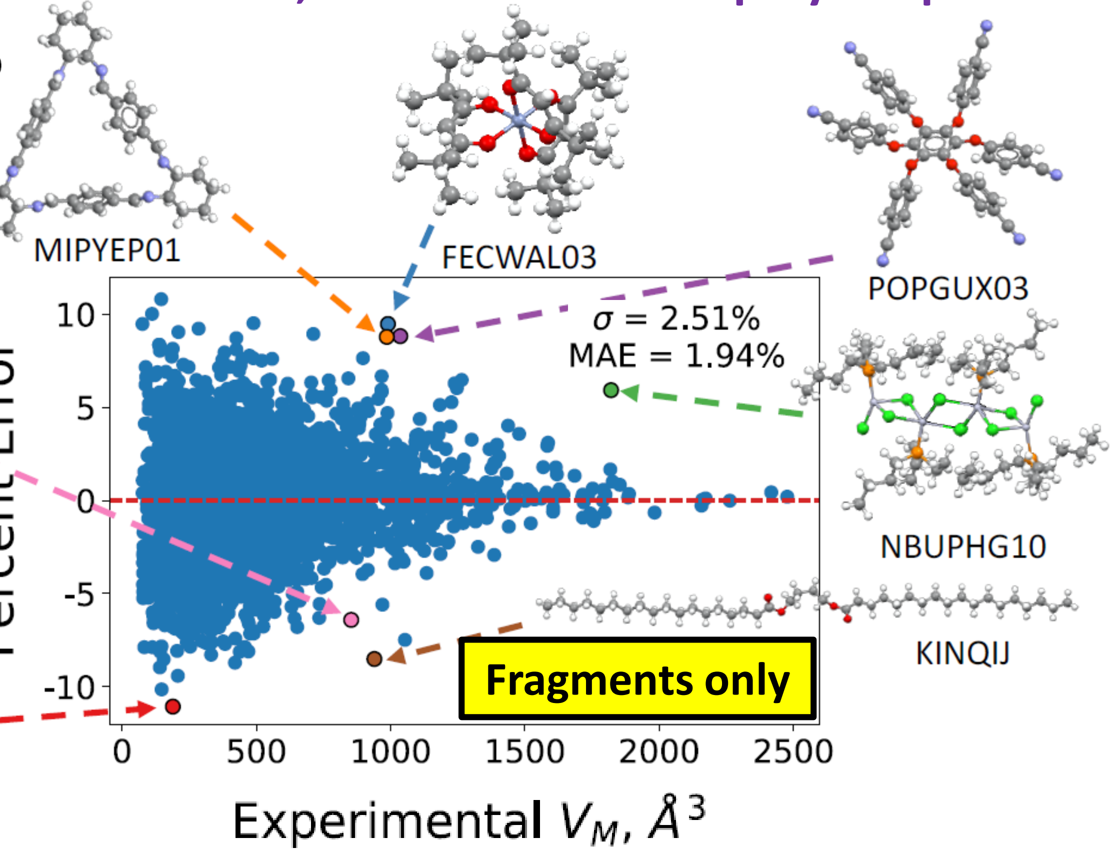


A model based only on topological fragments, without volume information has a broader error distribution

Outliers have sterically hindered regions, groups that do not participate in intermolecular interactions, or conformational polymorphs







Including both geometric and chemical information is essential to the performance of the ML model

I. Bier and N. Marom
J. Phys. Chem. A 124, 10330 (2020)

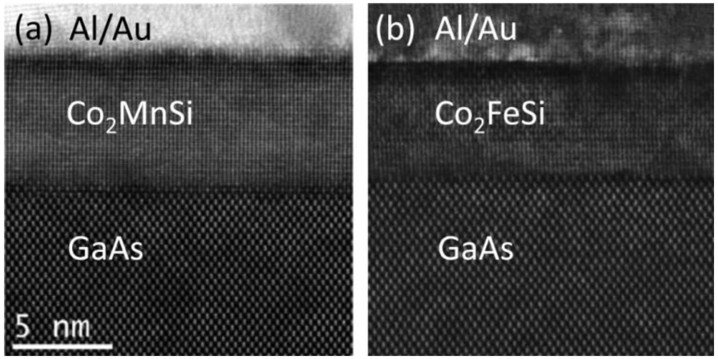
Machine Learning the Hubbard U Parameter in DFT+U

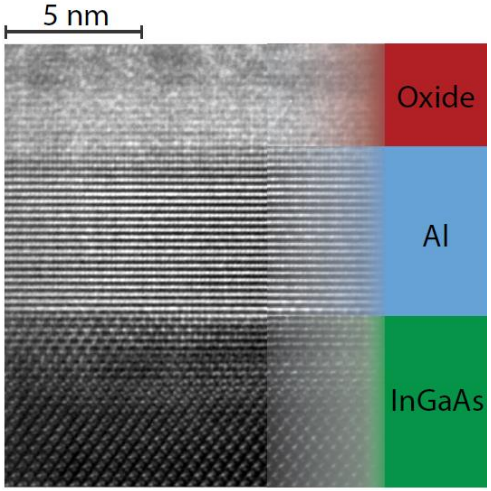
Hybrid Interfaces

A hybrid interface between two dissimilar materials may exhibit unique physical properties that do not exist in either bulk material

Spin injection at an interface between a ferromagnet and a semiconductor enables the implementation of a spin valve

T. A. Peterson *et al.*, *Phys. Rev. B* 94, 235309 (2016);





A superconductor/ semiconductor interface may enable the realization of networks of qubits based on Majorana zero modes

J. Shabani *et al.*, *Phys. Rev. B* 93, 155402 (2016);

Our goal is to develop computational tools for predicting the structure and properties of hybrid interfaces

Periodic Slab Models of Interfaces

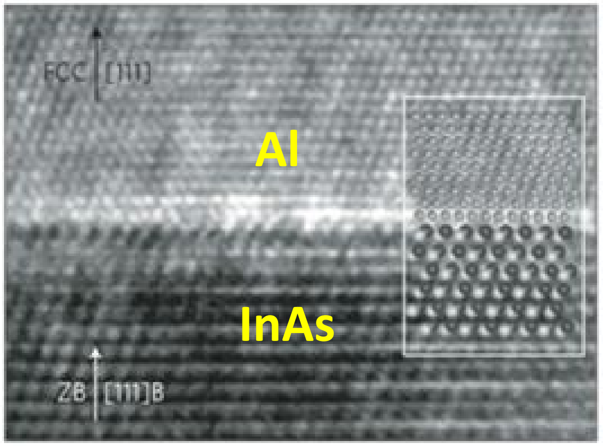
Many DFT codes are based on plane-wave basis sets and therefore impose 3D periodic boundary conditions

The interface must be commensurate in the x-y plane, which may require large supercells

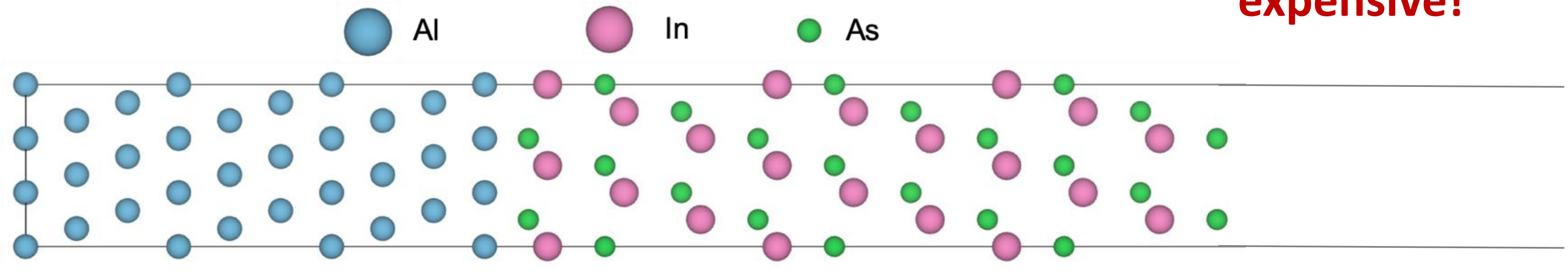
Often, a large number of layers of each material is needed to avoid quantum confinement effects

For a surface, vacuum space must be added along z to avoid spurious interactions between periodic replicas

Hydrogen passivation of dangling bonds at the surface may be required to eliminate spurious states



DFT simulations of interfaces are technically involved and computationally expensive!



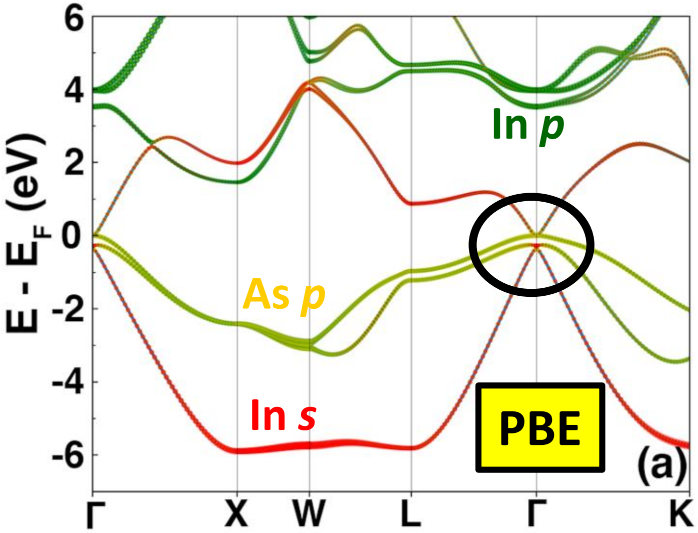
Band Structure of InAs

The Perdew-Burke-Ernzerhof (PBE) generalized gradient approximation:

J. P. Perdew, K. Burke, M. Ernzerhof, *Phys. Rev. Lett*. 77, 3865 (1996); 78, 1396 (1997)

- Includes a dependence on the density and its gradient (semi-local functional)
- Computationally efficient
- Suffers from the self-interaction error

PBE produces no band gap for InAs

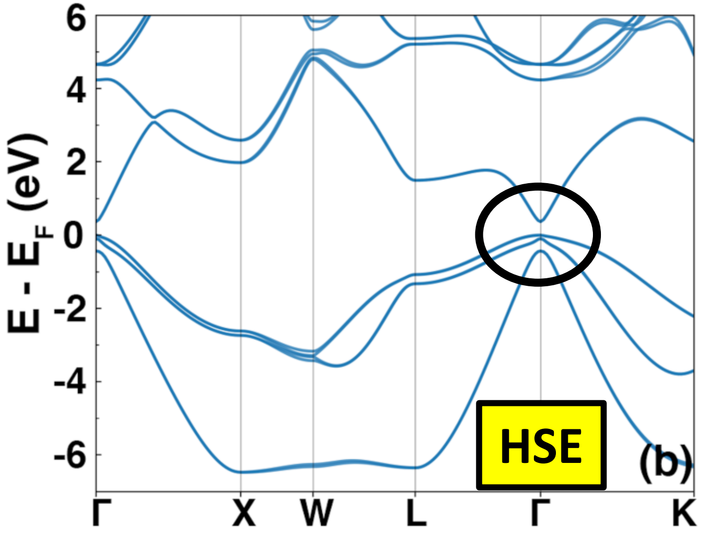


The Heyd-Scuzeria-Ernzerhof range-separated hybrid functional (HSE)

J. Heyd, G. E. Scuseria, M. Ernzerhof, *J. Chem. Phys*. 118, 8207 (2003); 124, 219906 (2006)

- A fraction of exact (Fock) exchange is mixed with the PBE exchange and correlation
- The Coulomb potential is split into short-range (SR) and long-range (LR) parts
- Has 25% exact exchange in the SR and reduces to PBE in the LR

HSE mitigates SIE and produces a gap for InAs but at a high computational cost



DFT+U(BO)

DFT+U

A Hubbard-like term, U_{eff} = U - J, is added to the DFT energy, where U is the on-site Coulomb repulsion interaction and J is the exchange interaction:

$$E_{tot} = E_{DFT} + \frac{U-J}{2}\sum_{\sigma} n_{m,\sigma} - n_{m,\sigma}^2$$

S. L. Dudarev, G. A. Botton, S. Y. Savrasov, C. J. Humphreys, A. P. Sutton, *Phys. Rev. B* 57, 1505 (1998)

Offers a balance of accuracy and efficiency

U_{eff} is a system dependent parameter

We machine learn U_{eff} by Bayesian optimization (BO)

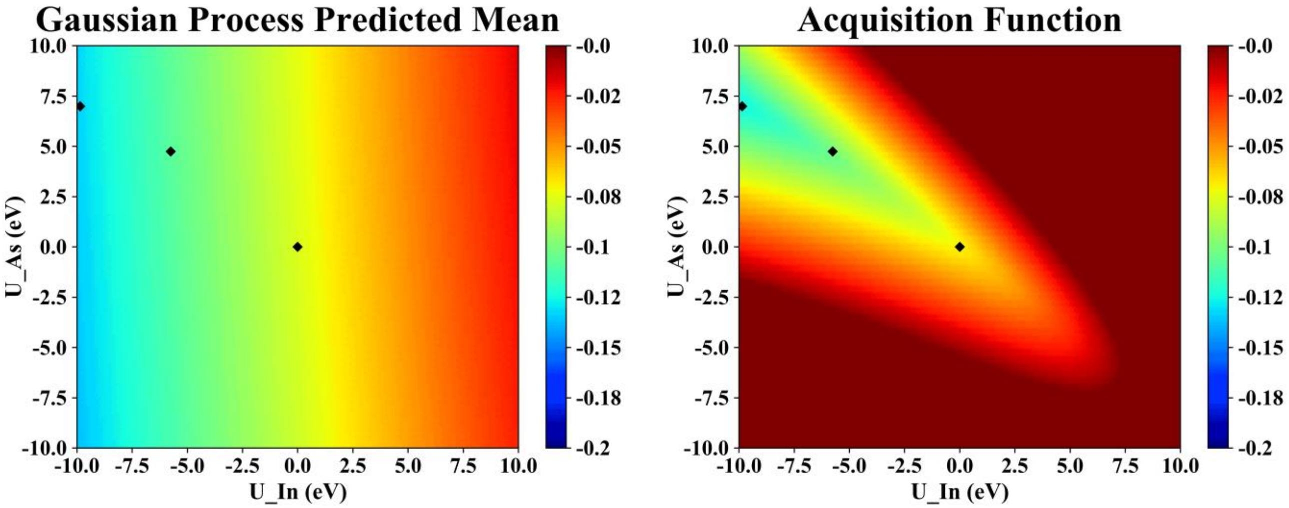
The objective function is formulated to reproduce the HSE band gap and band structure as closely as possible:

$$f(\vec{U}) = -\alpha_1\left(E_g^{HSE} - E_g^{PBE+U}\right)^2 - \alpha_2(\Delta Band)^2$$

$$\Delta Band = \sqrt{\frac{1}{N_E}\sum_{i=1}^{N_k}\sum_{j=1}^{N_b}\left(\varepsilon_{HSE}^{j}[k_i] - \varepsilon_{PBE+U}^{j}[k_i]\right)^2}$$

M. Yu, S. Yang, C. Wu, and N. Marom, npj Computational Materials 6, 180 (2020)

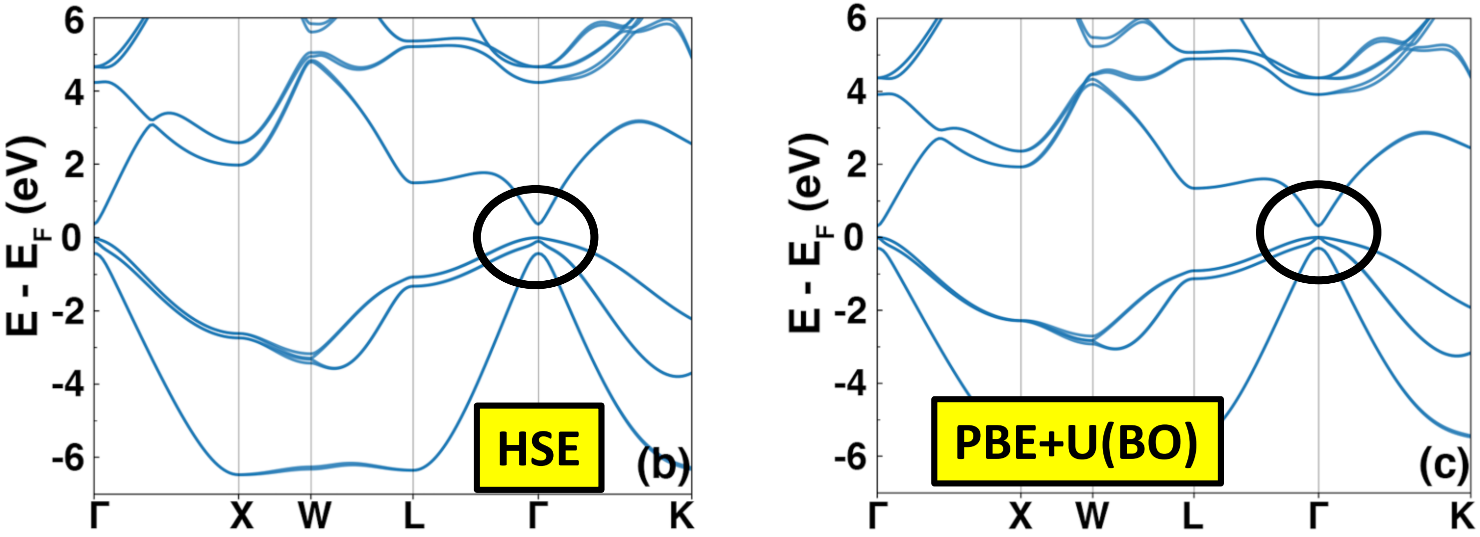
DFT+U(BO)



2D BO is performed to find the optimal U values for In-*p* and As-*p*

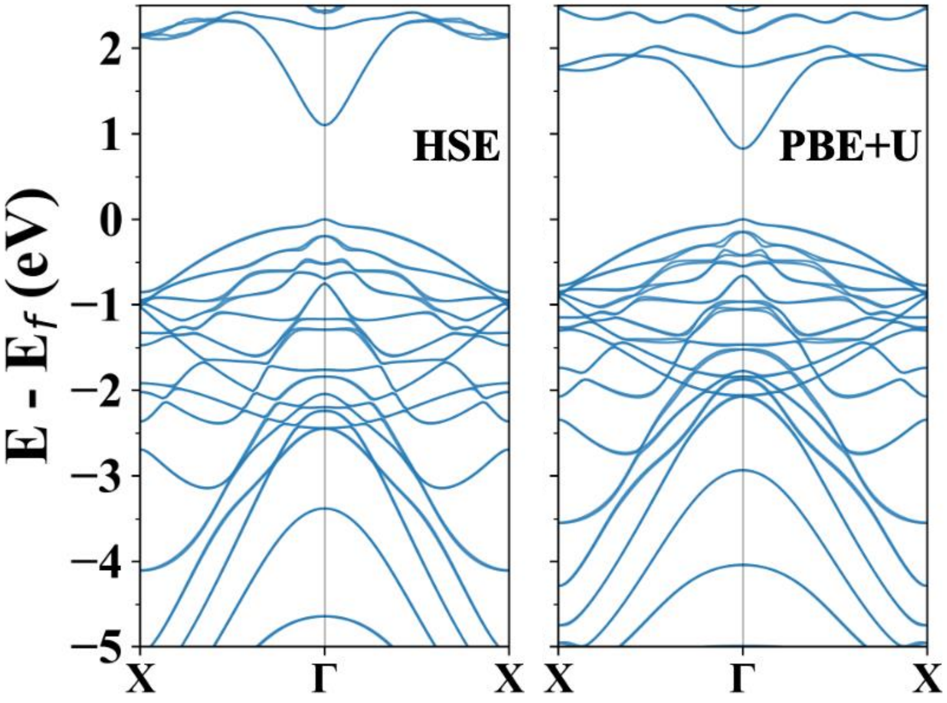
Negative values of U are allowed

PBE+U(BO) produces a comparable band structure to HSE at a fraction of the computational cost



M. Yu, S. Yang, C. Wu, and N. Marom, npj Computational Materials 6, 180 (2020)

Electronic Structure of InAs and InSb Surfaces

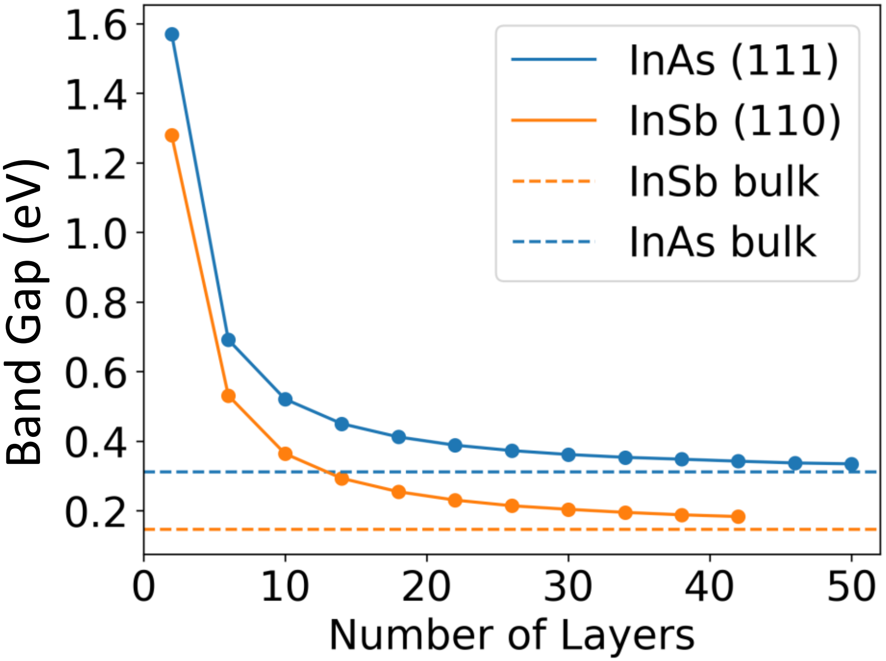


The parameters obtained for bulk InAs are transferrable to a surface slab with 11 layers (largest we could calculate with HSE)

M. Yu, S. Yang, C. Wu, and N. Marom, npj Computational Materials <u>6</u>, 180 (2020)

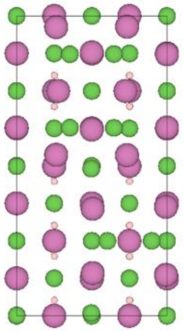
40-50 atomic layers are required to converge the electronic structure of InAs and InSb surfaces to the bulk limit

S. Yang *et al.*, arXiv 2012.14935 (2020)

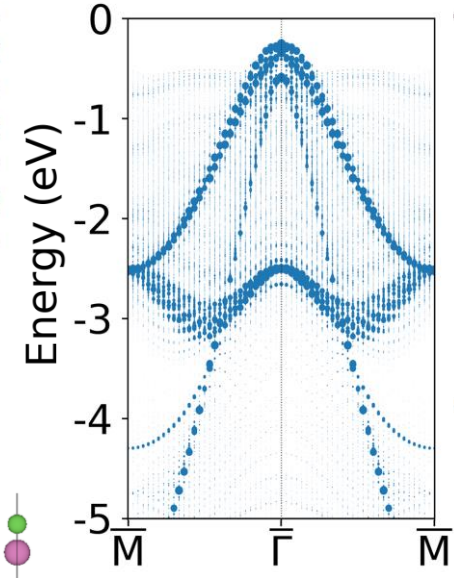


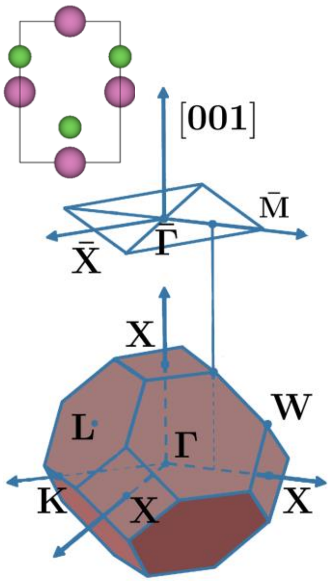
Bulk Band Unfolding

A slab with 20 layers is used to simulate the β2(2x4) reconstruction of InAs(001)



Z Unfolding

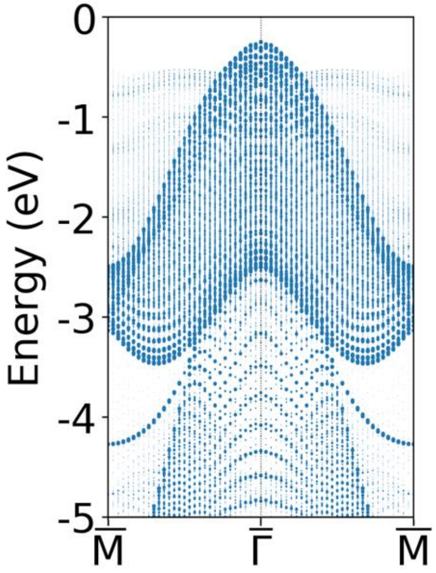


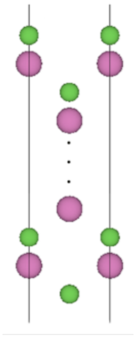


Bulk unfolding onto the primitive cell eliminates $k_z=|\Gamma X|$ bands

The band structure is in agreement with ARPES

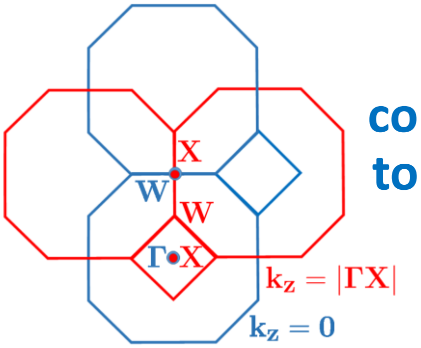
XY Unfolding





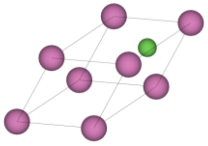
Unfolding in the xy plane onto a 1x1xZ slab produces a dense band structure

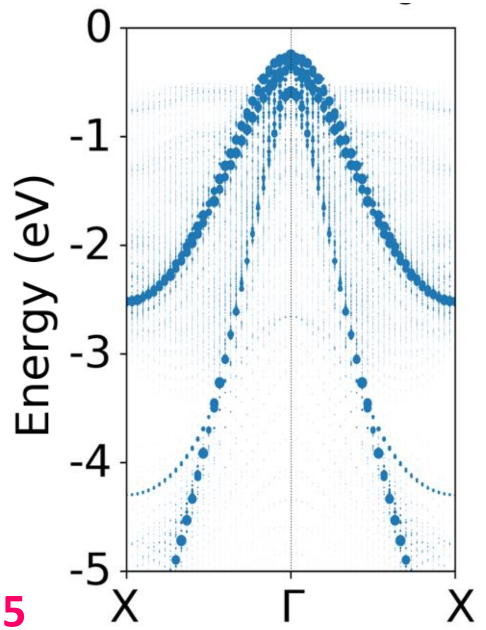
Z-unfolding onto a bulk unit cell oriented in (001) yields more bulk-like band structure



Bands corresponding to $k_z=|\Gamma X|$ are present

Bulk Unfolding





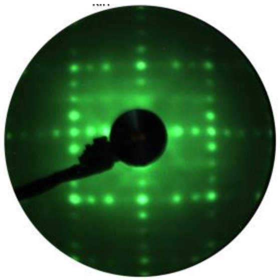
arXiv 2012.14935

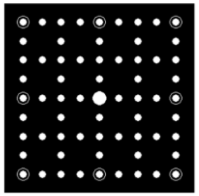
InAs(001) Surface Reconstructions

LEED shows superposition of 2x4 and 4x2 reconstructions

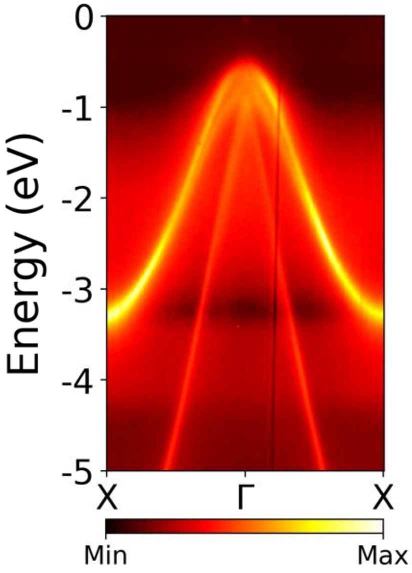
Different reconstructions exhibit different signatures of surface states but have similar band bending

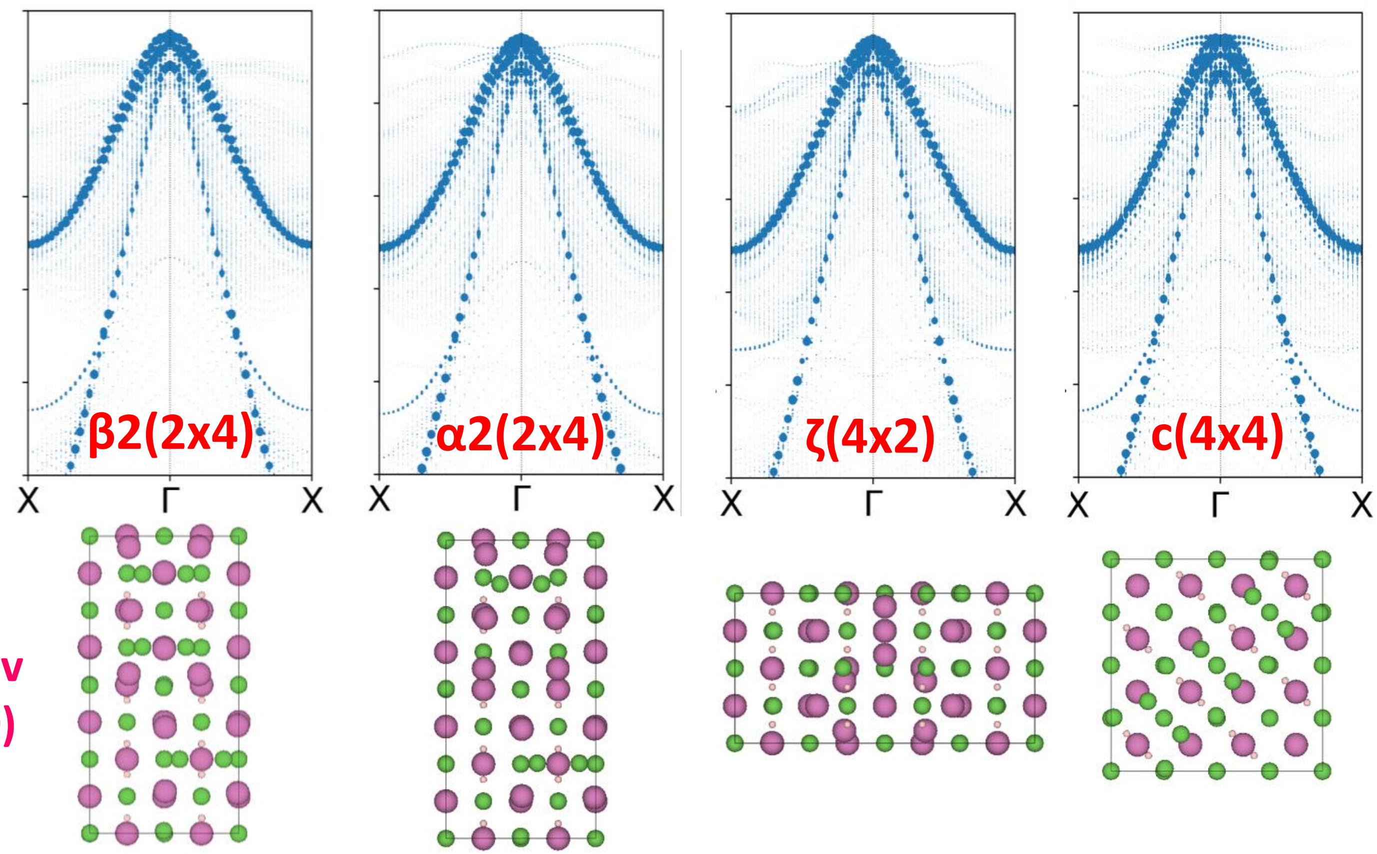
DFT supports the coexistence of 2x4 and 4x2 domains





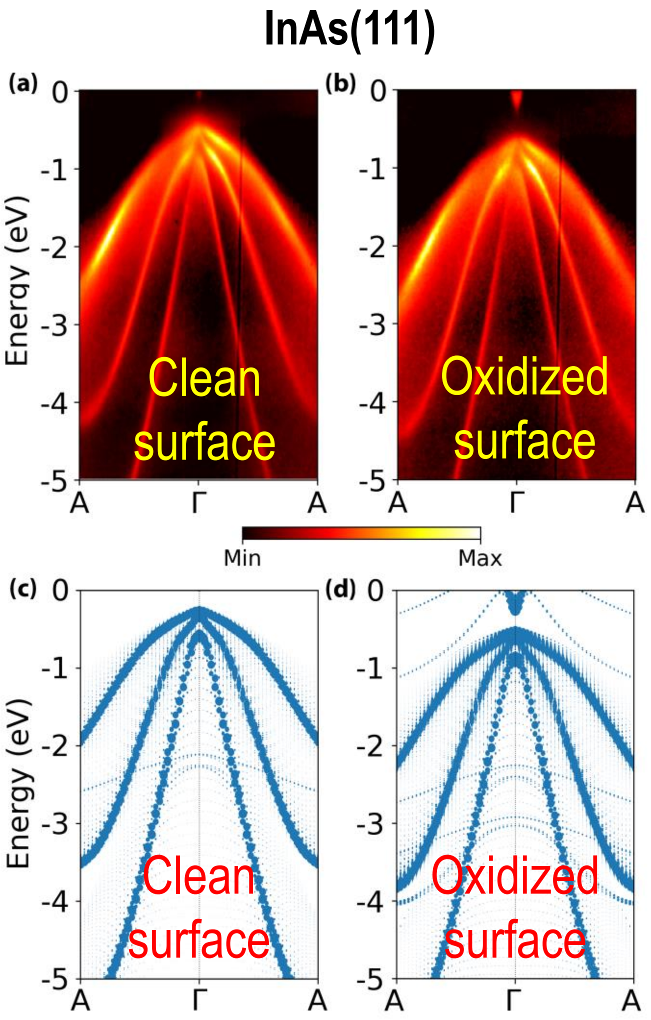
Surface sensitive ARPES would be needed to detect surface states





S. Yang *et al.*, arXiv 2012.14935 (2020)

Effect of Oxidation on InAs(111) vs InSb(110)

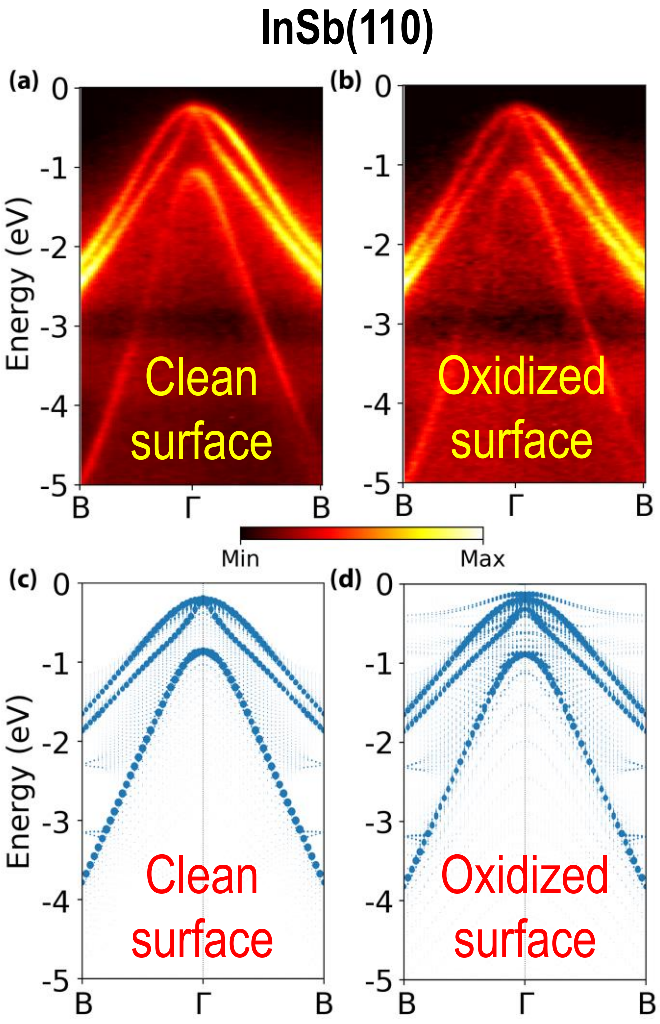


PBE+U(BO) is in agreement with ARPES experiments

For InAs(111) oxidation leads to band bending and the appearance of an electron pocket

For InSb(110) oxidation does not cause band bending and no electron pocket appears

This is due to stronger charge transfer from surface As to O than from Sb to O



S. Yang, N. Schröter, V. Strocov, S. Schuwalow, M. Rajpalke , K. Ohtani, P. Krogstrup, G. Winkler, J. Gukelberger, D. Gresch, G. Aeppli, R. Lutchyn, N. Marom, arXiv 2012.14935 (2020)

Acknowledgements

NSF

SCIENCE TO SHAPE THE FUTURE OF THE ARMY
ARMY RESEARCH OFFICE
1951

U.S. DEPARTMENT OF ENERGY

NERSC

Argonne NATIONAL LABORATORY

Download PyMoVE: https://github.com/manny405/PyMoVE

Download GAtor, Genarris, and Ogre: www.noamarom.com

GAtor Genarris Ogre