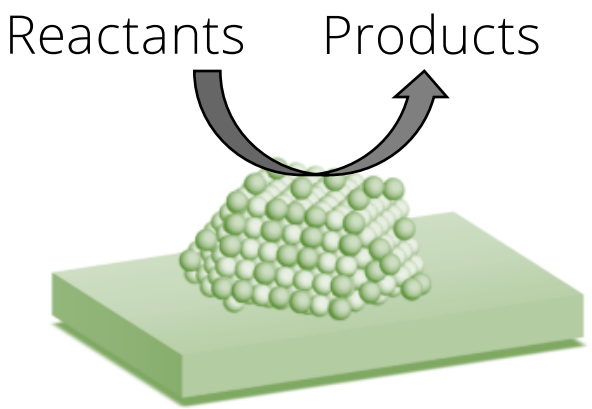
Big-Data Analytics and Machine Learning for Materials Science

Lucas Foppa

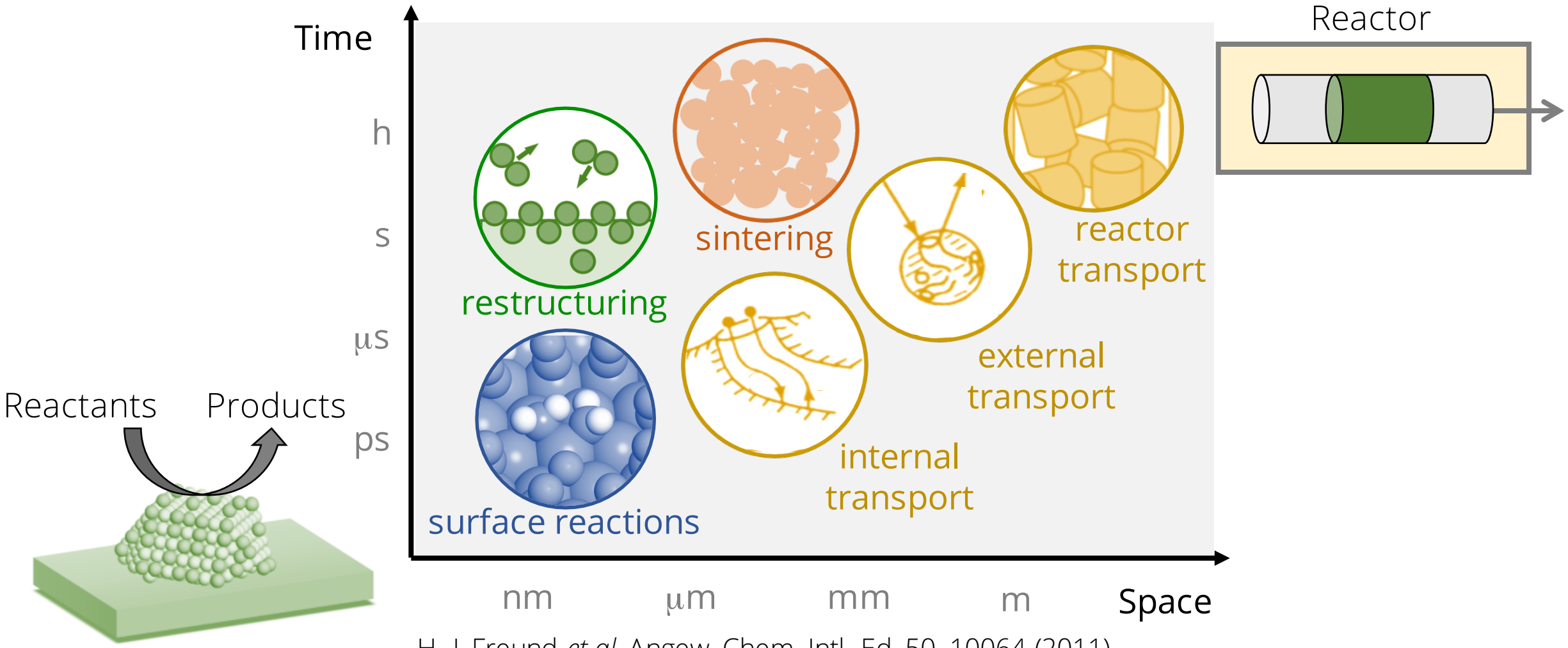
The NOMAD Laboratory at the Fritz Haber Institute of the Max Planck Society

Heterogeneous Catalysis: A Materials' Function Governed by an Intricate Interplay of Many Processes



H.-J. Freund *et al*. Angew. Chem. Intl. Ed. 50, 10064 (2011)
R. Schlögl Angew. Chem. Intl. Ed. 54, 3465 (2015)

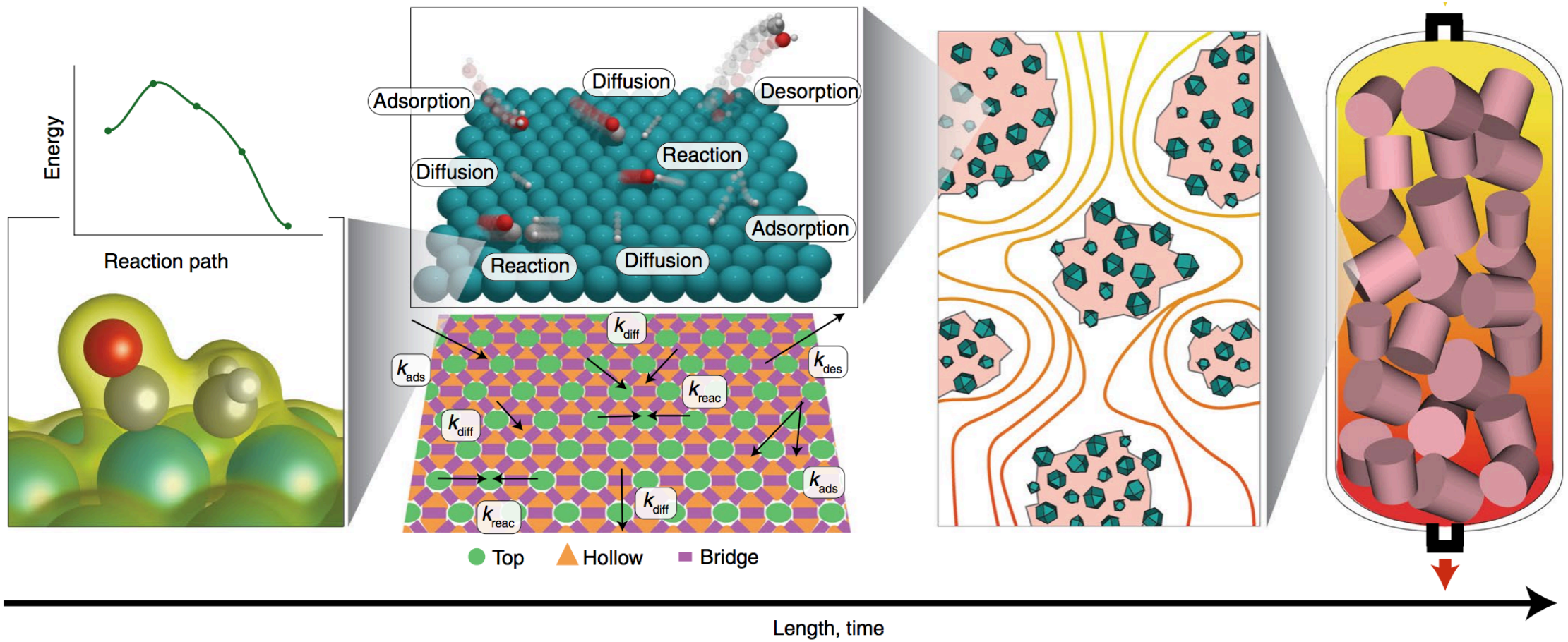
Heterogeneous Catalysis: A Materials' Function Governed by an Intricate Interplay of Many Processes



H.-J. Freund *et al.* Angew. Chem. Intl. Ed. 50, 10064 (2011)
R. Schlögl Angew. Chem. Intl. Ed. 54, 3465 (2015)

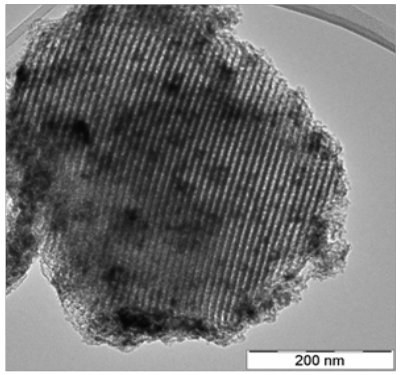
Heterogeneous Catalysis: A Materials' Function Governed by an Intricate Interplay of Many Processes

Explicit Atomistic Models of the Full Catalytic Progression Unfeasible



A. Bruix et al. Nat. Catal. 2, 659 (2019)

Identifying Correlations Describing Materials in Consistent Data with Artificial Intelligence

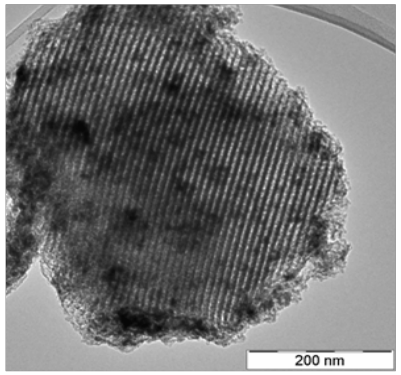


Systematic Experiments

- Synthesis
- Characterization
- Performance Testing

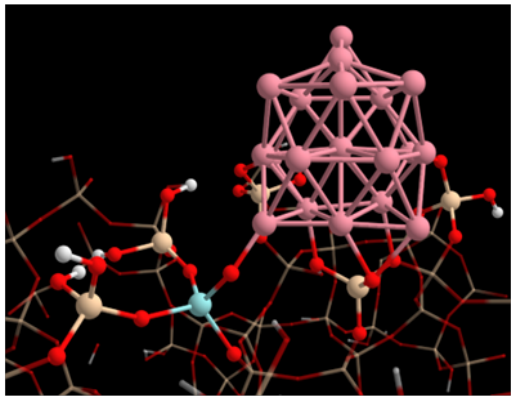
A. Trunschke *et al.* Top. Catal. 63, 1683 (2020), L. Foppa *et al.* MRS Bull. 46, 1016 (2021), R. Miyazaki *et al.* J. Am. Chem. Soc. 146, 5433 (2024)

Identifying Correlations Describing Materials in Consistent Data with Artificial Intelligence



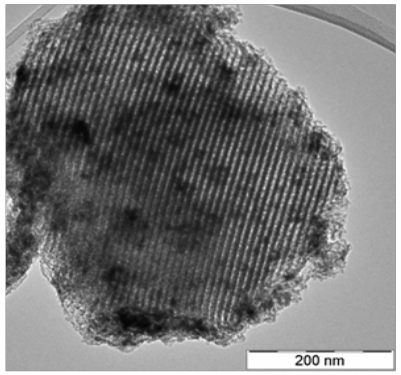
Systematic Experiments
- Synthesis
- Characterization
- Performance Testing

Atomistic Simulations
- Materials Properties
- Environment Effect



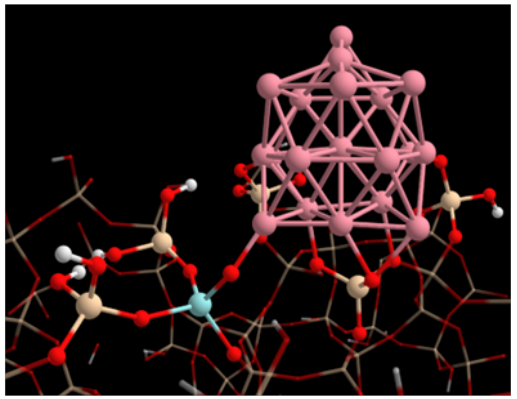
A. Trunschke *et al.* Top. Catal. 63, 1683 (2020), L. Foppa *et al.* MRS Bull. 46, 1016 (2021), R. Miyazaki *et al.* J. Am. Chem. Soc. 146, 5433 (2024)

Identifying Correlations Describing Materials in Consistent Data with Artificial Intelligence



Systematic Experiments
- Synthesis
- Characterization
- Performance Testing

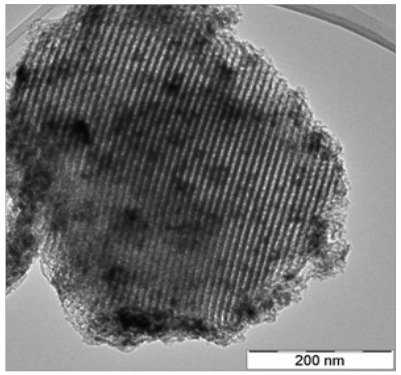
Atomistic Simulations
- Materials Properties
- Environment Effect



Consistent Dataset
- Target
  - ➢ Measured Performance
- Candidate Descriptive Parameters (Primary Features)
  - ➢ Measured or Calculated
  - ➢ Reflecting Possible Underlying Processes

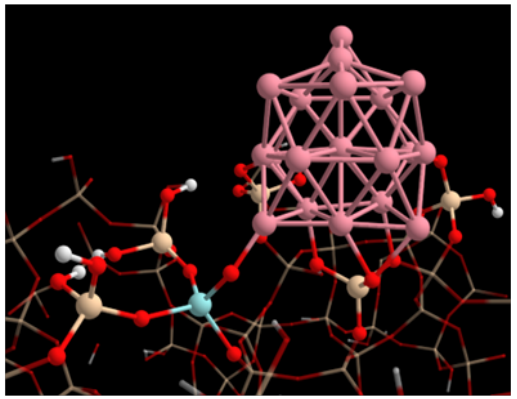
A. Trunschke *et al.* Top. Catal. 63, 1683 (2020), L. Foppa *et al.* MRS Bull. 46, 1016 (2021), R. Miyazaki *et al.* J. Am. Chem. Soc. 146, 5433 (2024)

Identifying Correlations Describing Materials in Consistent Data with Artificial Intelligence



Systematic Experiments
- Synthesis
- Characterization
- Performance Testing

Atomistic Simulations
- Materials Properties
- Environment Effect

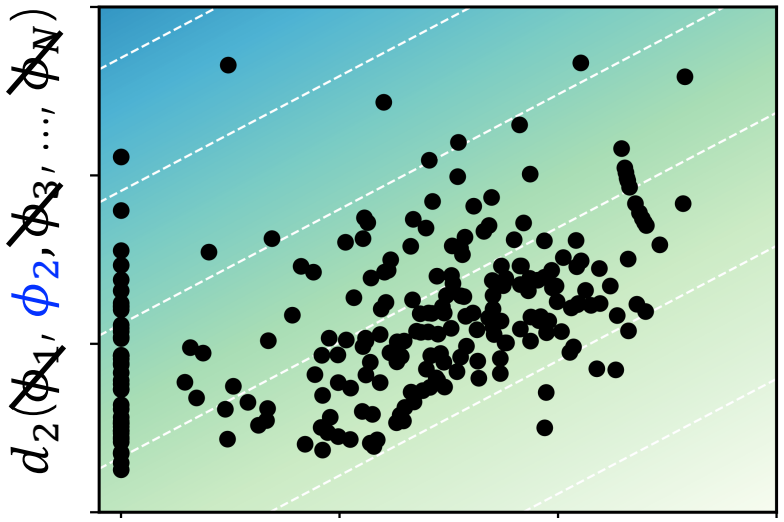


Consistent Dataset
- Target
  - ➢ Measured Performance
- Candidate Descriptive Parameters (Primary Features)
  - ➢ Measured or Calculated
  - ➢ Reflecting Possible Underlying Processes

AI

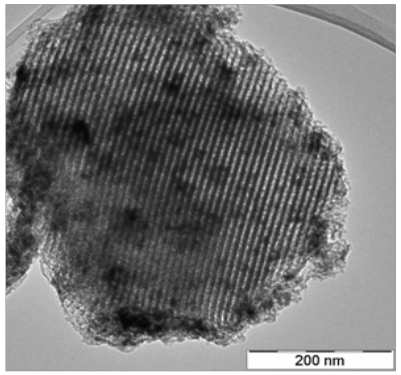
"Materials Genes"

Performance f (Key Parameters)



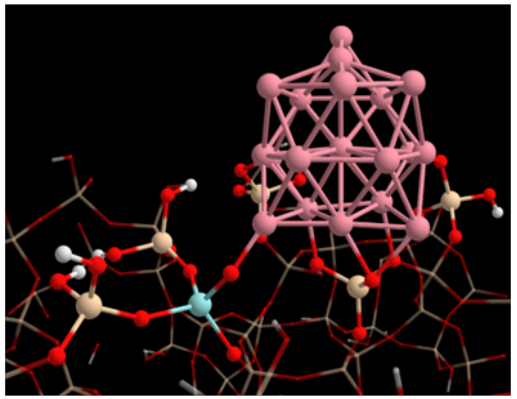
A. Trunschke *et al.* Top. Catal. 63, 1683 (2020), L. Foppa *et al.* MRS Bull. 46, 1016 (2021), R. Miyazaki *et al.* J. Am. Chem. Soc. 146, 5433 (2024)

Identifying Correlations Describing Materials in Consistent Data with Artificial Intelligence



Systematic Experiments
- Synthesis
- Characterization
- Performance Testing

Atomistic Simulations
- Materials Properties
- Environment Effect



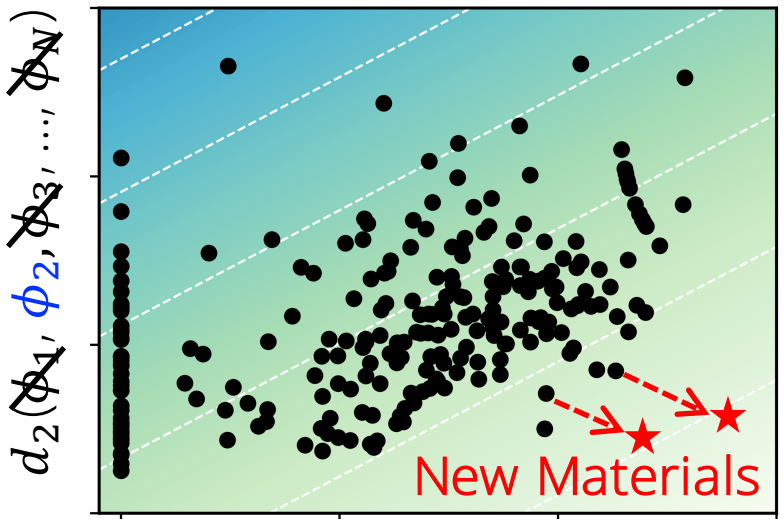
Insights
Predictions

Consistent Dataset
- Target
  - ➢ Measured Performance
- Candidate Descriptive Parameters (Primary Features)
  - ➢ Measured or Calculated
  - ➢ Reflecting Possible Underlying Processes

AI

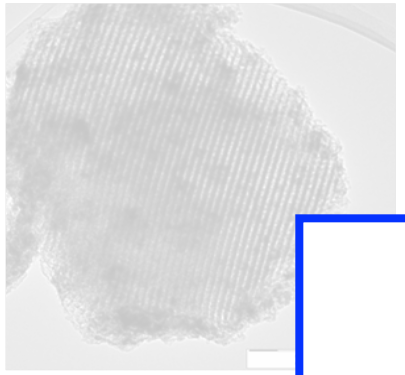
"Materials Genes"

Performance
f (Key Parameters)



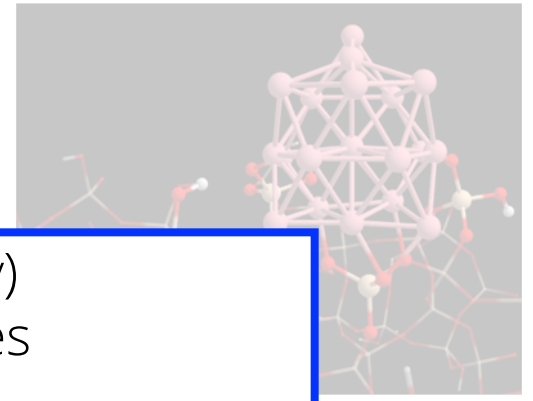
A. Trunschke *et al.* Top. Catal. 63, 1683 (2020), L. Foppa *et al.* MRS Bull. 46, 1016 (2021), R. Miyazaki *et al.* J. Am. Chem. Soc. 146, 5433 (2024)

Identifying Correlations Describing Materials in Consistent Data with Artificial Intelligence



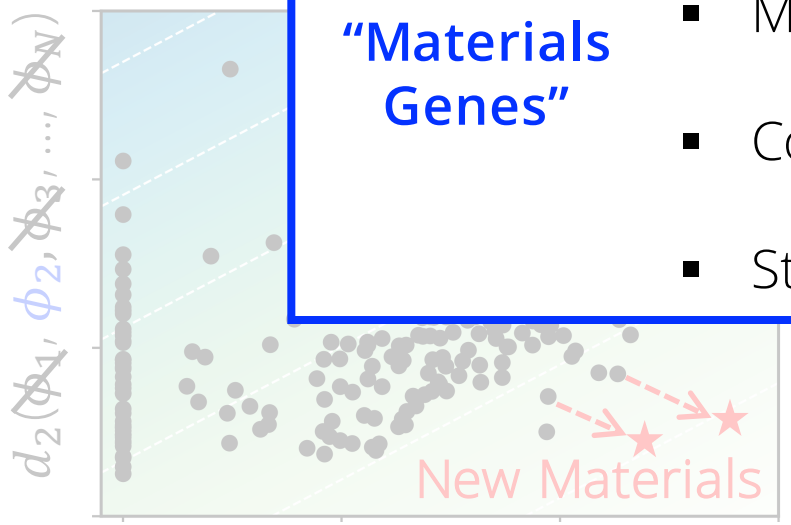
Systematic Experiments
- Synthesis
- Characterization

Atomistic Simulations
- Materials Properties
- Environment Effect



"Materials Genes"

- Capture Complex Patterns But Do Not (Necessarily) Provide Full Understanding of Underlying Processes
- Multiple Parameters
- Correlations (= Rules) Different from Laws
- Statistical Description (Uncertainty, Domain of Applicability)



"Materials Genes"

AI

Primary
- Measured or Calculated
- Reflecting Possible Underlying Processes

A. Trunschke *et al.* Top. Catal. 63, 1683 (2020), L. Foppa *et al.* MRS Bull. 46, 1016 (2021), R. Miyazaki *et al.* J. Am. Chem. Soc. 146, 5433 (2024)

Requirements for AI Methods

- Identification of Key Parameters to Describe Target, From Many Offered Parameters
- Efficient with "Small" Data (e.g., 10^2)

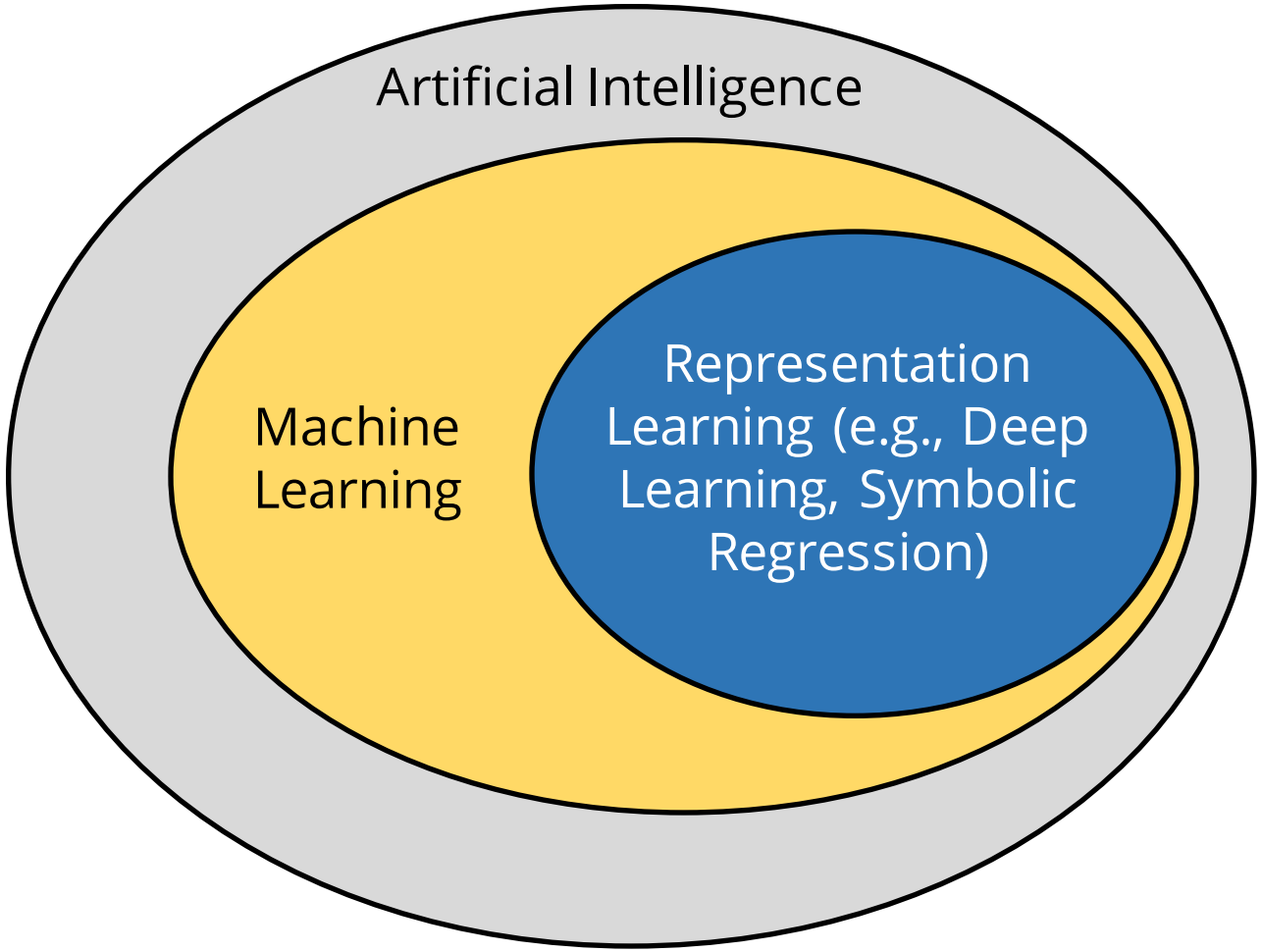


Figure from L. Ghiringhelli

Requirements for AI Methods

- Identification of Key Parameters to Describe Target, From Many Offered Parameters
- Efficient with "Small" Data (e.g., 10^2)
- Methods of Choice:
  - SISSO
  - Subgroup Discovery

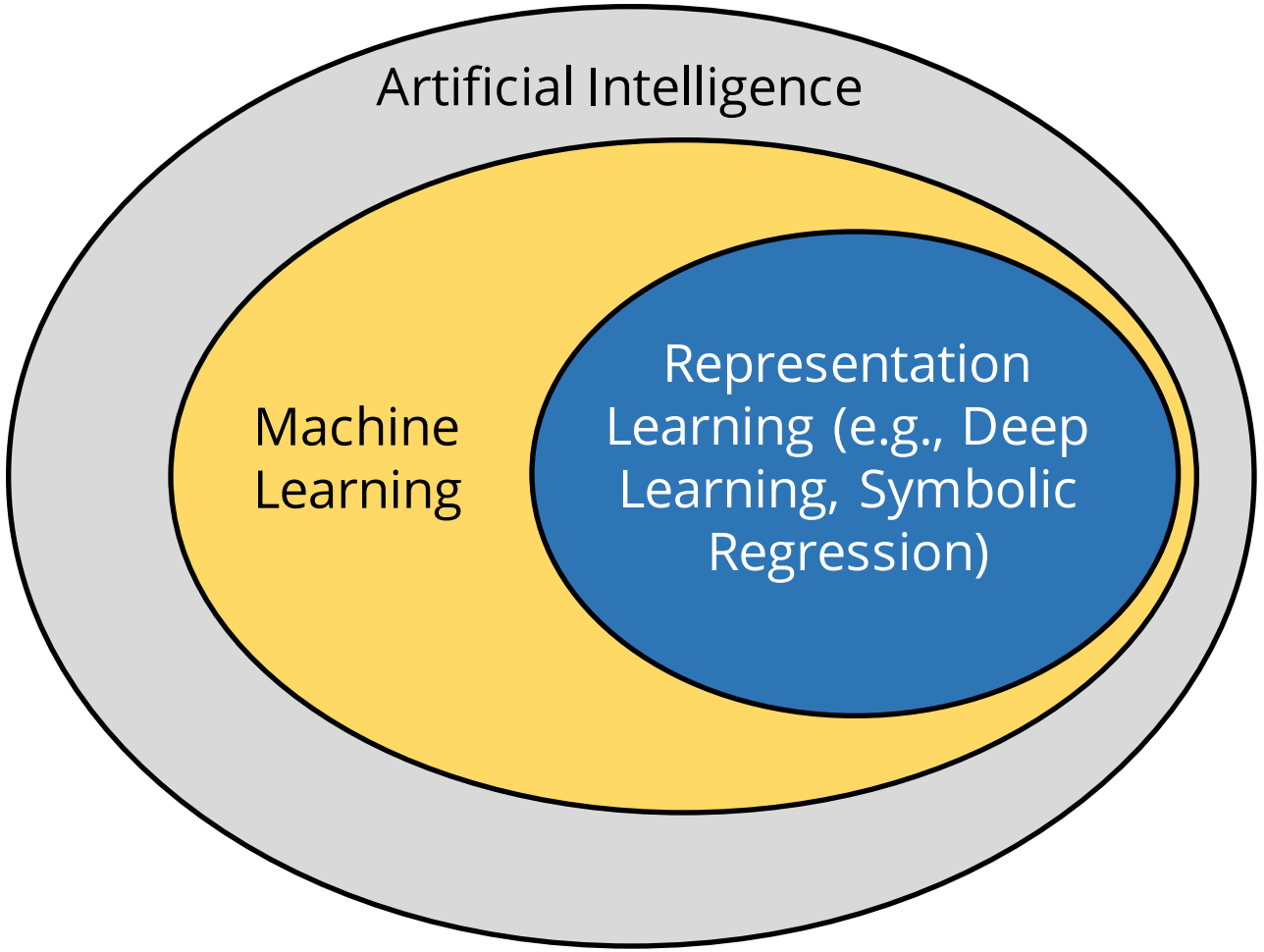


Figure from L. Ghiringhelli

Sure Independence Screening and Sparsifying Operator (SISSO)

1. Choice of Input Parameters (Primary Features, φ_i)

R. Ouyang *et al.* Phys. Rev. Mater. 2, 083802 (2018)
T. A. R. Purcell *et al.* J. Chem. Phys. 159, 114110 (2023)

Sure Independence Screening and Sparsifying Operator (SISSO)

1. Choice of Input Parameters (Primary Features, φ_i)

2. Feature Construction

Primary Feature Space, Φ_0 → Candidate Feature Space, Φ_N

Unary and Binary Mathematical Operators Iteratively Applied

e.g.

$$\varphi_1, \varphi_2 \xrightarrow{q=1} (\varphi_1)^2,\ |\varphi_1 - \varphi_2| \xrightarrow{q=2} \frac{|\varphi_1 - \varphi_2|}{(\varphi_1)^2}$$

R. Ouyang *et al.* Phys. Rev. Mater. 2, 083802 (2018)
T. A. R. Purcell *et al.* J. Chem. Phys. 159, 114110 (2023)

Sure Independence Screening and Sparsifying Operator (SISSO)

1. Choice of Input Parameters (Primary Features, φ_i)

2. Feature Construction

Primary Feature Space, $\mathbf{\Phi_0}$ → Candidate Feature Space, $\mathbf{\Phi_N}$

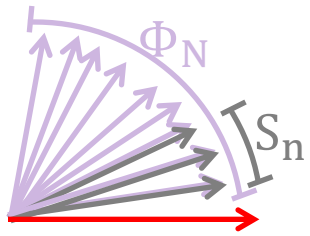
Unary and Binary Mathematical Operators Iteratively Applied

e.g.

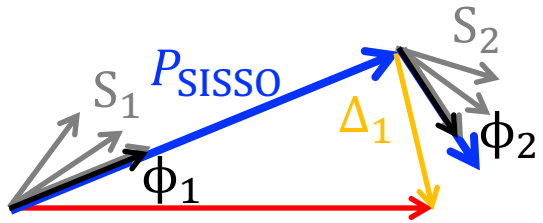
$$\varphi_1,\varphi_2 \xrightarrow{q=1} (\varphi_1)^2,\ |\varphi_1-\varphi_2| \xrightarrow{q=2} \frac{|\varphi_1-\varphi_2|}{(\varphi_1)^2}$$

3. Compressed Sensing

Selection SIS → Candidate Feature Subspaces, $\mathbf{S_n}$



P or $\Delta_1, \Delta_2, \Delta_3, \ldots$



P: property
Δ_n: residuals

R. Ouyang *et al.* Phys. Rev. Mater. 2, 083802 (2018)
T. A. R. Purcell *et al.* J. Chem. Phys. 159, 114110 (2023)

Sure Independence Screening and Sparsifying Operator (SISSO)

1. Choice of Input Parameters (Primary Features, φ_i)

2. Feature Construction

Primary Feature Space, $\mathbf{\Phi_0}$ → Candidate Feature Space, $\mathbf{\Phi_N}$

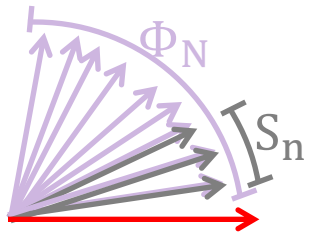
Unary and Binary Mathematical Operators Iteratively Applied

e.g.

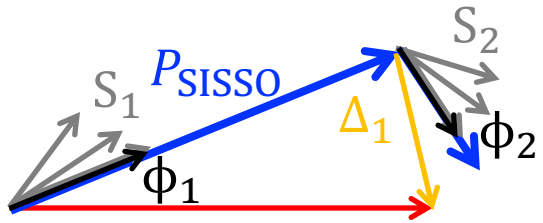
$$\varphi_1, \varphi_2 \xrightarrow{q=1} (\varphi_1)^2,\ |\varphi_1 - \varphi_2| \xrightarrow{q=2} \frac{|\varphi_1 - \varphi_2|}{(\varphi_1)^2}$$

3. Compressed Sensing

Selection: SIS → Candidate Feature Subspaces, $\mathbf{S_n}$ → Regularization: ℓ_0, SO



P or $\Delta_1, \Delta_2, \Delta_3, \ldots$



P: property
Δ_n: residuals

Descriptor and Model

Fitted Coefficients

$$P^{(\mathrm{SISSO})} = c_0 + \sum_{i=1}^{D} c_i d_i$$

Analytical Expressions

R. Ouyang *et al.* Phys. Rev. Mater. 2, 083802 (2018)
T. A. R. Purcell *et al.* J. Chem. Phys. 159, 114110 (2023)

Sure Independence Screening and Sparsifying Operator (SISSO)

1. Choice of Input Parameters (Primary Features, φ_i)

2. Feature Construction

Primary Feature Space, $\mathbf{\Phi_0}$ → Candidate Feature Space, $\mathbf{\Phi_N}$

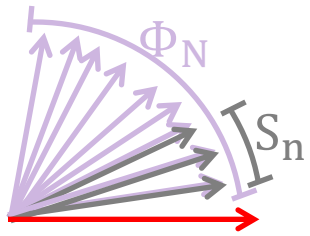
Unary and Binary Mathematical Operators Iteratively Applied

e.g.

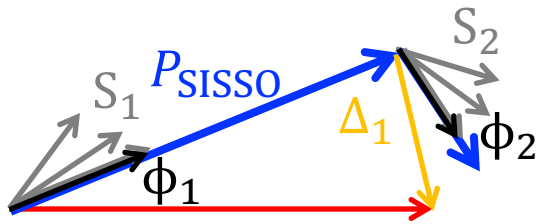
$$\varphi_1, \varphi_2 \xrightarrow{q=1} (\varphi_1)^2,\ |\varphi_1 - \varphi_2| \xrightarrow{q=2} |\varphi_1 - \varphi_2|,\ \frac{|\varphi_1 - \varphi_2|}{(\varphi_1)^2}$$

3. Compressed Sensing

Selection: SIS → Candidate Feature Subspaces, $\mathbf{S_n}$ → Regularization: ℓ_0 SO



P or $\Delta_1, \Delta_2, \Delta_3, \ldots$



P: property
Δ_n: residuals

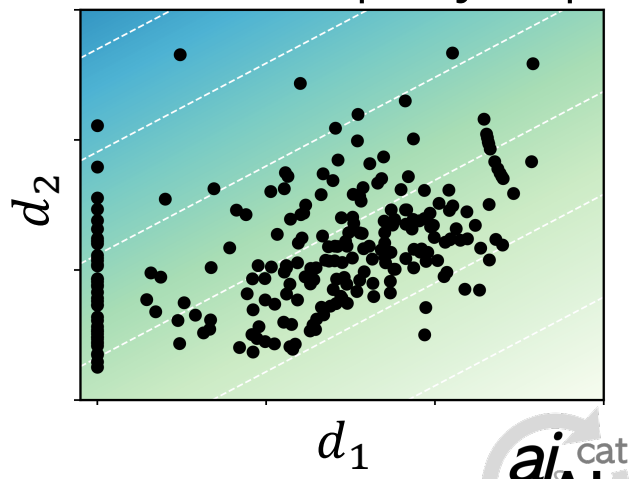
Descriptor and Model

Fitted Coefficients

$$P^{(\mathrm{SISSO})} = c_0 + \sum_{i=1}^{D} c_i d_i$$

Analytical Expressions

"Materials-Property Maps"



R. Ouyang *et al.* Phys. Rev. Mater. 2, 083802 (2018)
T. A. R. Purcell *et al.* J. Chem. Phys. 159, 114110 (2023)

Sure Independence Screening and Sparsifying Operator (SISSO)

1. Choice of Input Parameters (Primary Features, φ_i)

Low D (e.g., 2 or 3)

2. Feature Construction

Primary Feature Space, $\mathbf{\Phi_0}$ → Candidate Feature Space, $\mathbf{\Phi_N}$

Unary and Binary Mathematical Operators Iteratively Applied

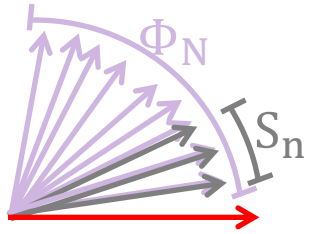
e.g.

$$\varphi_1, \varphi_2 \xrightarrow{q=1} (\varphi_1)^2,\ |\varphi_1 - \varphi_2| \xrightarrow{q=2} \frac{|\varphi_1 - \varphi_2|}{(\varphi_1)^2}$$

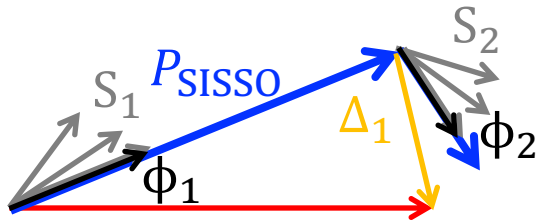
Immense Number (e.g., billions)

3. Compressed Sensing

Selection: SIS → Candidate Feature Subspaces, $\mathbf{S_n}$ → Regularization: ℓ_0 SO



P or $\Delta_1, \Delta_2, \Delta_3, \ldots$



P: property
Δ_n: residuals

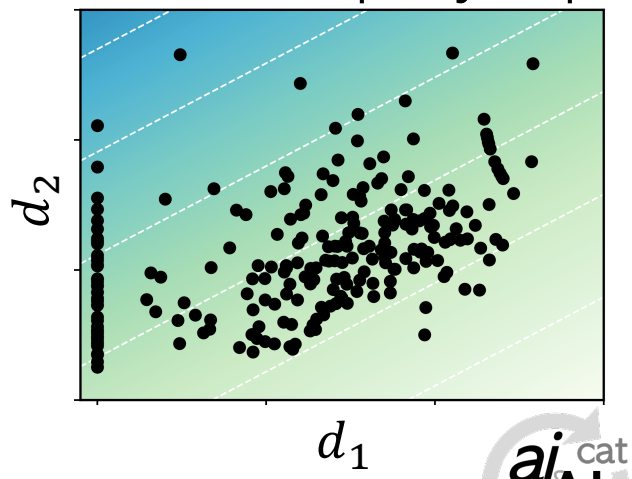
Descriptor and Model

Fitted Coefficients

$$P^{(SISSO)} = c_0 + \sum_{i=1}^{D} c_i d_i$$

Analytical Expressions

"Materials-Property Maps"



R. Ouyang *et al.* Phys. Rev. Mater. 2, 083802 (2018)
T. A. R. Purcell *et al.* J. Chem. Phys. 159, 114110 (2023)

Sure Independence Screening and Sparsifying Operator (SISSO)

1. Choice of Input Parameters (Primary Features, φ_i)

Low D (e.g., 2 or 3)

2. Feature Construction

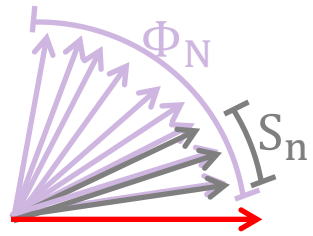
Primary Feature Space, Φ_0 → Candidate Feature Space, Φ_N

Immense Number (e.g., billions)

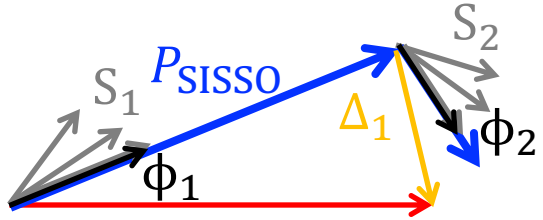
3. Compressed Sensing

Selection SIS → Candidate Feature Subspaces, $\mathbf{S_n}$ → ℓ_0 Regularization SO

$$\mathcal{L}oss^{\ell_0} = RMSE + \lambda \|\boldsymbol{c}\|_0$$



P or $\Delta_1, \Delta_2, \Delta_3, \ldots$



P: property
Δ_n: residuals

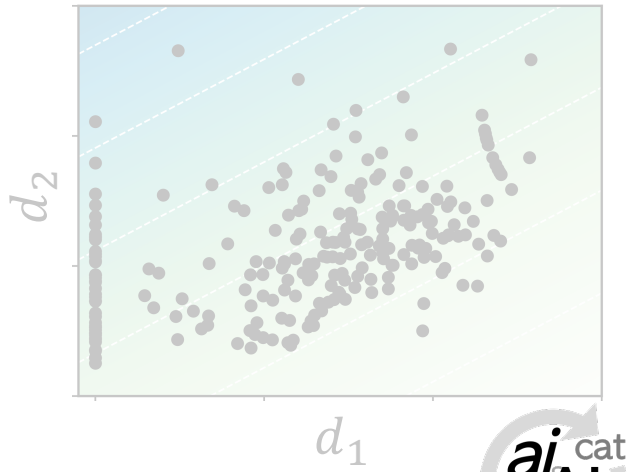
Descriptor and Model

Fitted Coefficients

$$P^{(SISSO)} = c_0 + \sum_{i=1}^{D} c_i d_i$$

lytical Expressions

"Materials-Property Maps"



SISSO Handles Larger (Correlated) Feature Spaces than the **LASSO Method** (ℓ_1 Regularization)

$$\mathcal{L}oss^{\ell_1} = RMSE + \lambda \|\boldsymbol{c}\|_1$$

↑ Regularization Strength

R. Ouyang *et al.* Phys. Rev. Mater. 2, 083802 (2018)
T. A. R. Purcell *et al.* J. Chem. Phys. 159, 114110 (2023)

Sure Independence Screening and Sparsifying Operator (SISSO)

1. Choice of Input Parameters (Primary Features, φ_i)

2. Feature Construction

Primary Feature Space, $\mathbf{\Phi_0}$ → Candidate Feature Space, $\mathbf{\Phi_N}$

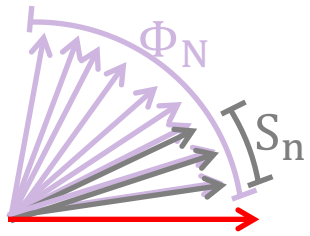
Unary and Binary Mathematical Operators Iteratively Applied

e.g.

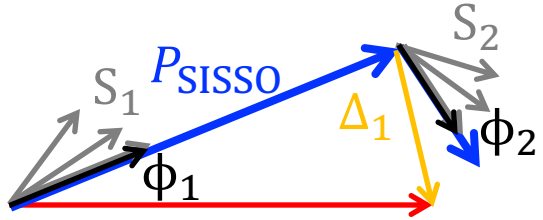
$\varphi_1, \varphi_2 \xrightarrow{q=1} (\varphi_1)^2,\ |\varphi_1 - \varphi_2| \xrightarrow{q=2} \frac{|\varphi_1 - \varphi_2|}{(\varphi_1)^2}$

3. Compressed Sensing

Selection: SIS → Candidate Feature Subspaces, $\mathbf{S_n}$ → Regularization: ℓ_0, SO



P or $\Delta_1, \Delta_2, \Delta_3, \ldots$



P: property
Δ_n: residuals

Descriptor and Model

Fitted Coefficients

$$P^{(SISSO)} = c_0 + \sum_{i=1}^{D} c_i d_i$$

Analytical Expressions

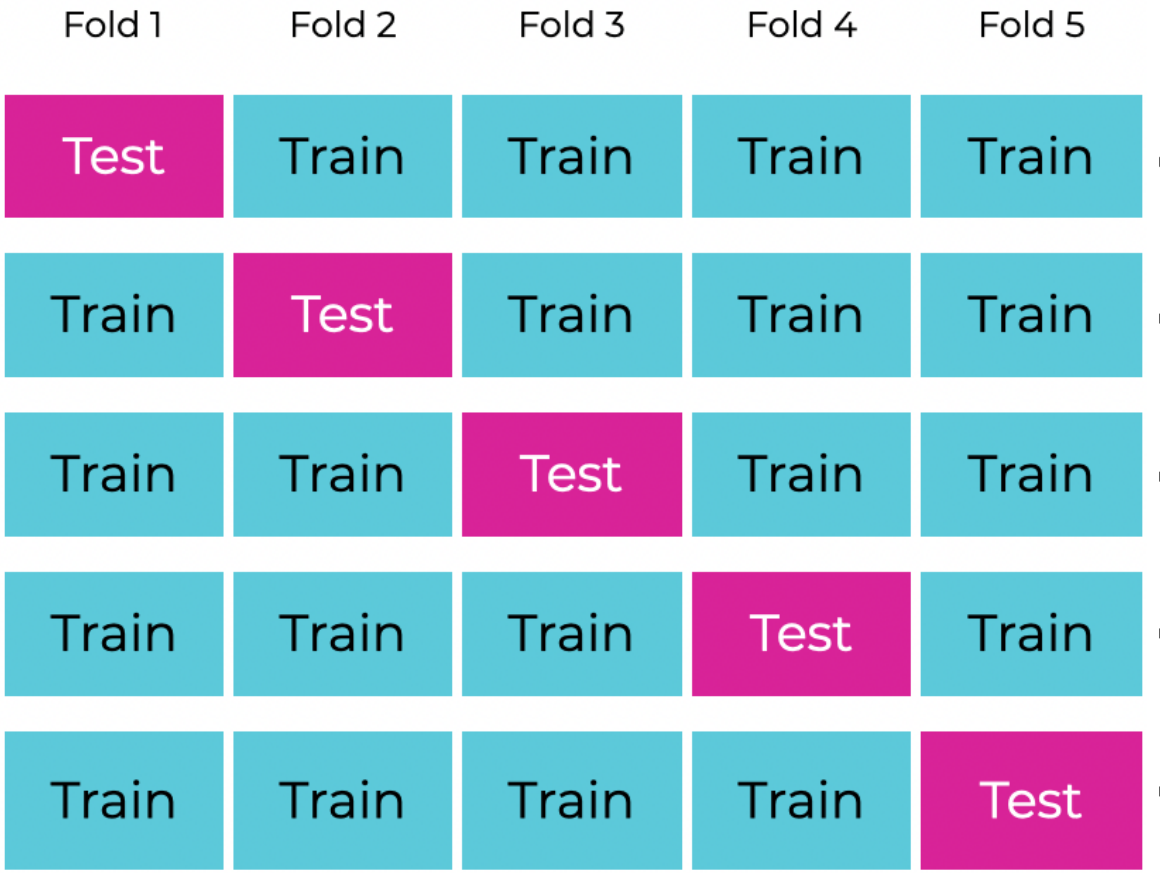
SISSO Hyperparameters:
D (Dimensionality) and q (Rung)

R. Ouyang *et al.* Phys. Rev. Mater. 2, 083802 (2018)
T. A. R. Purcell *et al.* J. Chem. Phys. 159, 114110 (2023)

Cross Validation Schemes

- Goals
  - Determine Hyperparameters
  - Estimate Model Performance on Unseen Data
- How?
  - Splitting Dataset into Multiple Folds
  - Training On Some Folds
  - Testing Model Predictions On Other Folds
- Different Schemes Appropriate for Different Problems, e.g., *K*-fold, Nested, and Leave-One-Out

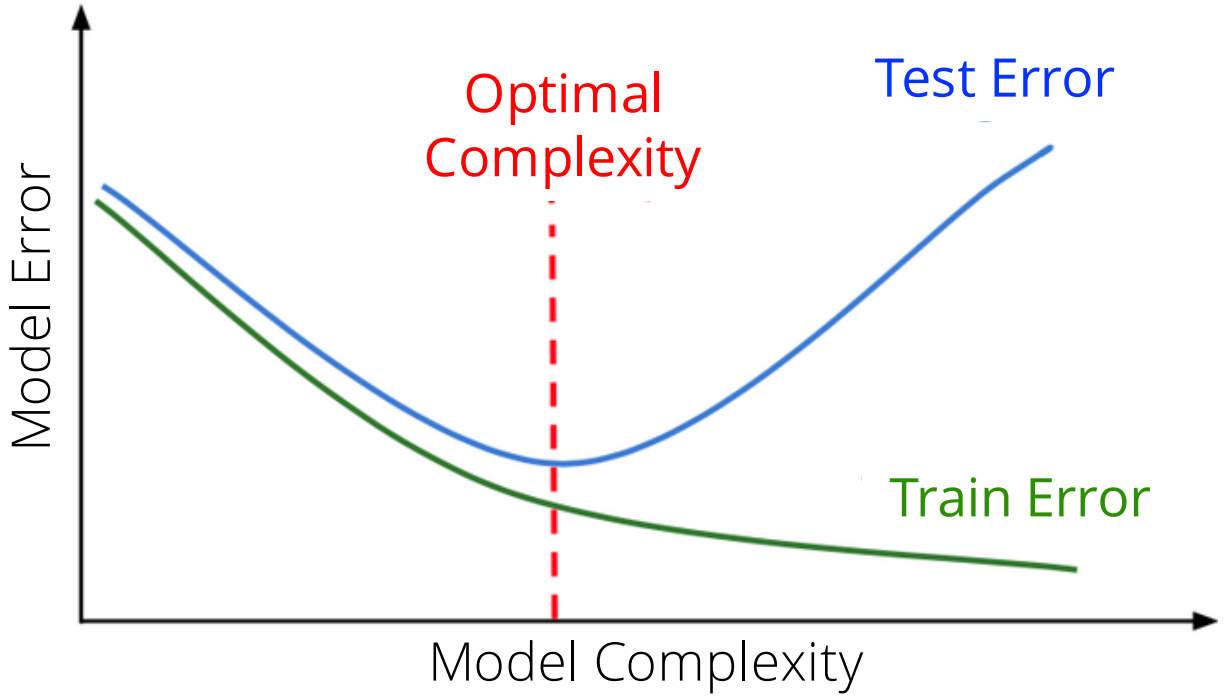
E.g., k-Fold Cross Validation



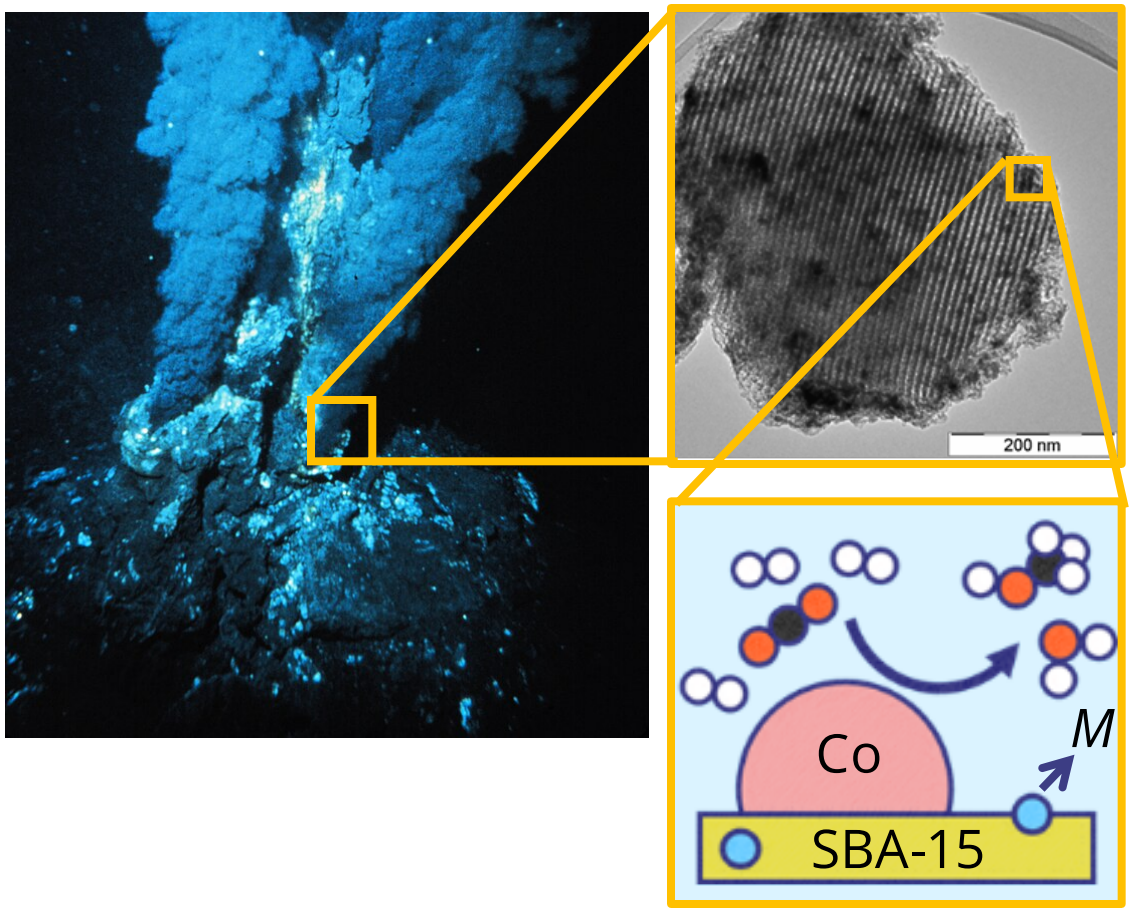
Cross Validation Schemes

- Goals
  - Determine Hyperparameters
  - Estimate Model Performance on Unseen Data
- How?
  - Splitting Dataset into Multiple Folds
  - Training On Some Folds
  - Testing Model Predictions On Other Folds
- Different Schemes Appropriate for Different Problems, e.g., *K*-fold, Nested, and Leave-One-Out

E.g., for Hyperparameters Controlling SISSO Model Complexity D (Dimensionality) and q (Rung)

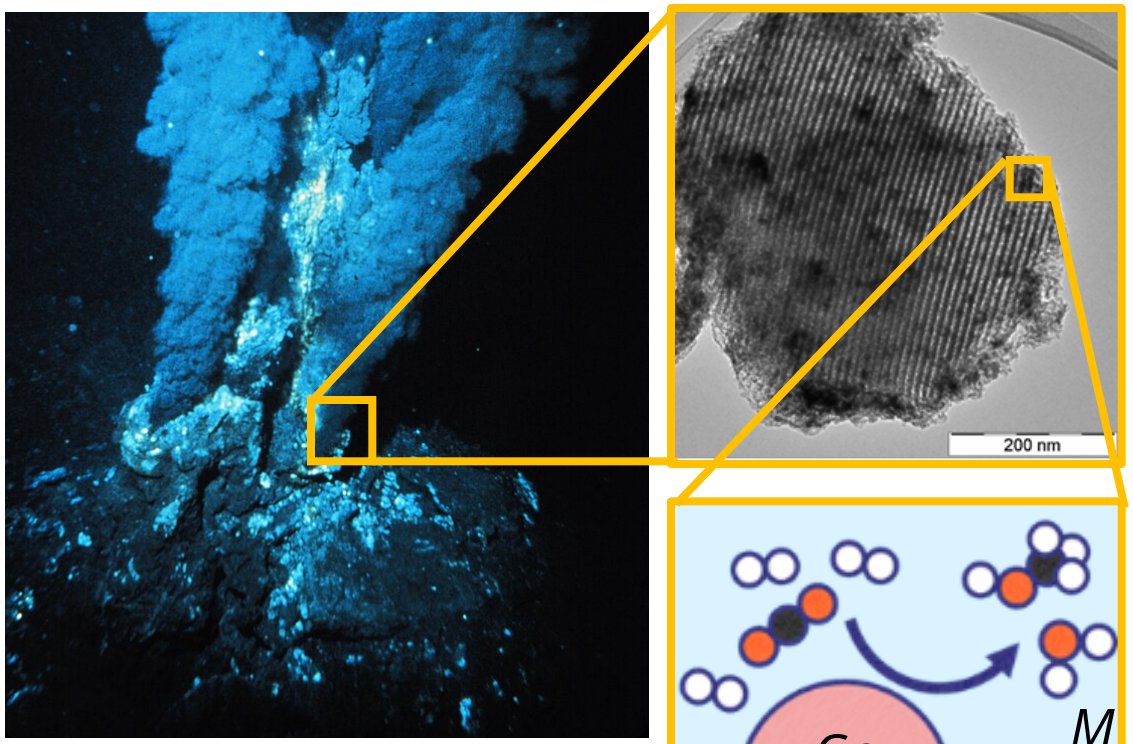


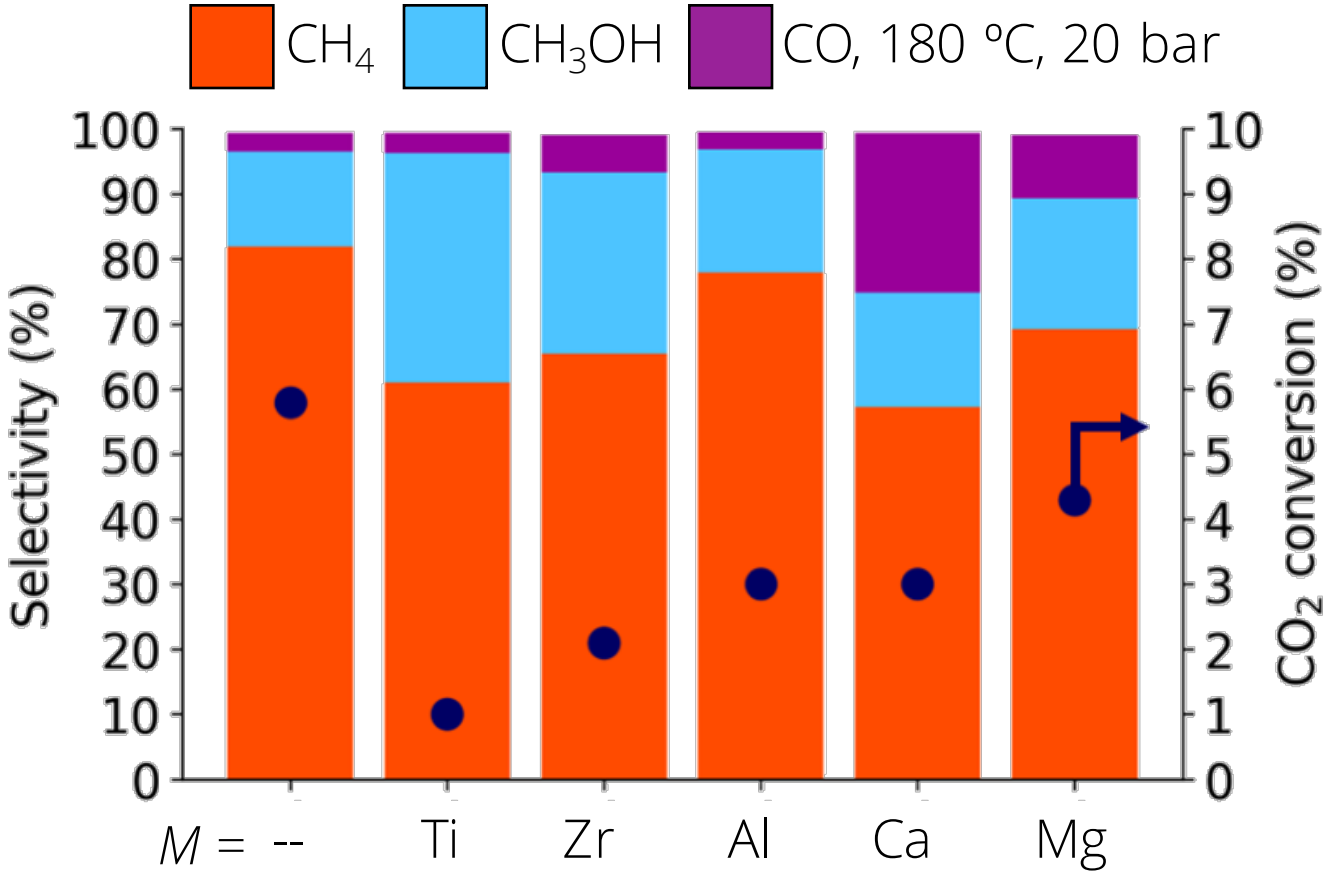
Application: Describing the Performance of Silica-Supported Cobalt Catalysts (Co/*M*-SiO_2) in CO_2 Hydrogenation



K. Belthle *et al.* J. Am. Chem. Soc. 144, 21232 (2022)

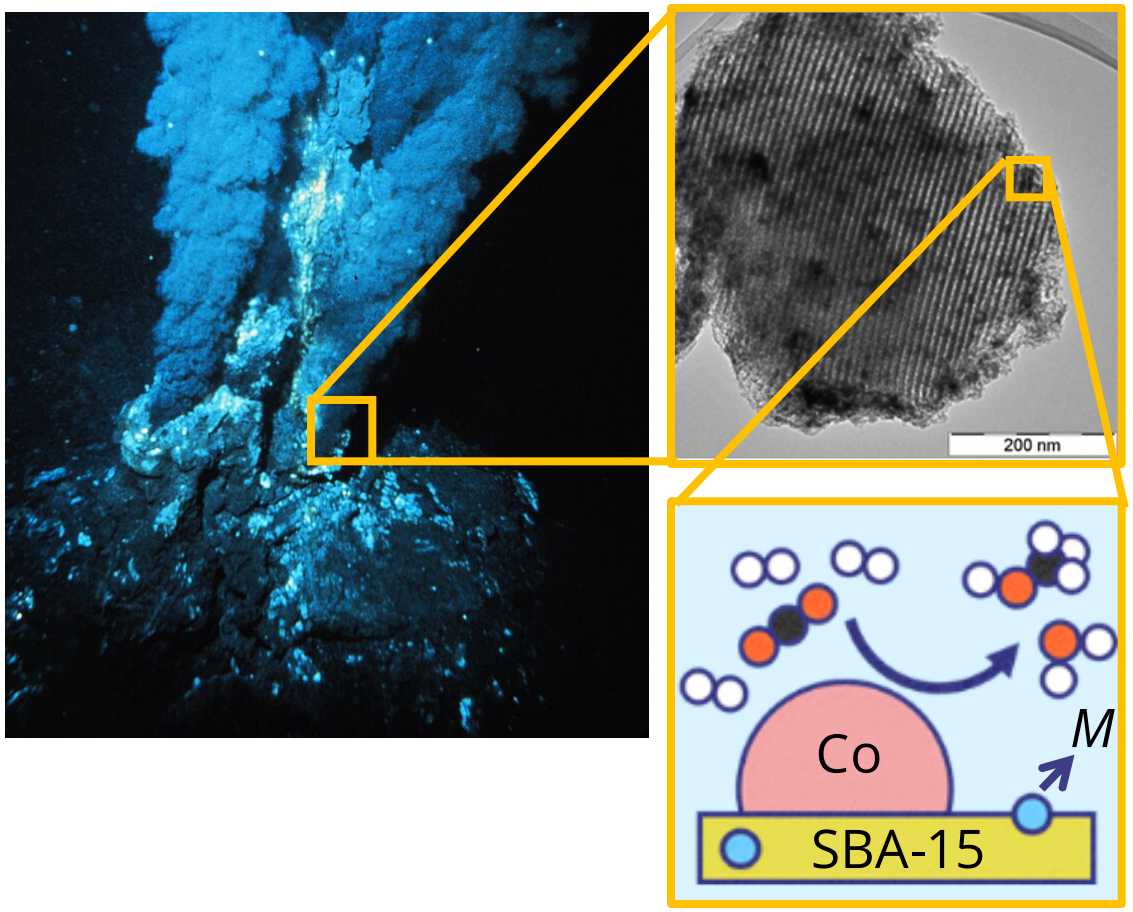
Application: Describing the Performance of Silica-Supported Cobalt Catalysts (Co/*M*-SiO_2) in CO_2 Hydrogenation

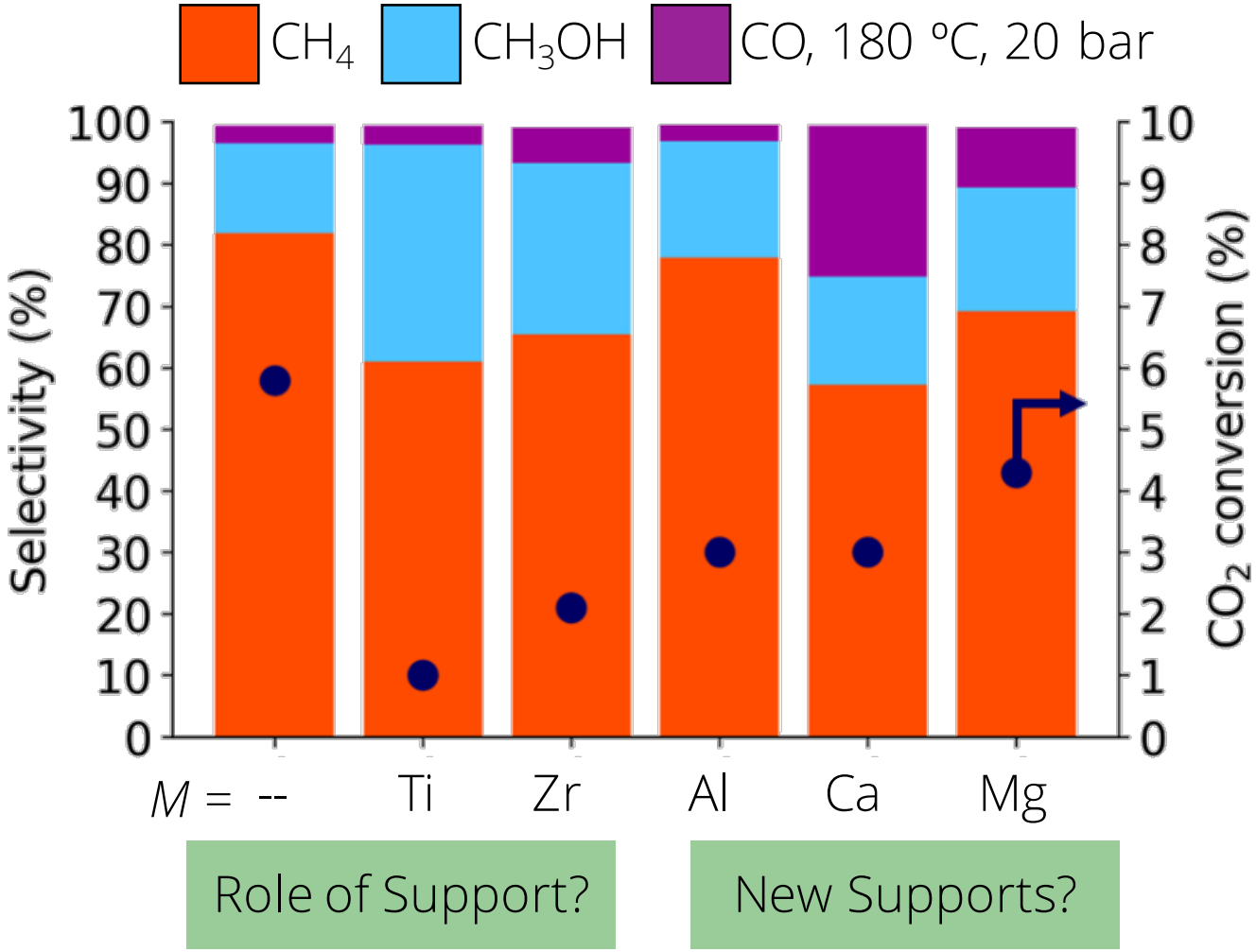




K. Belthle *et al.* J. Am. Chem. Soc. 144, 21232 (2022)

Application: Describing the Performance of Silica-Supported Cobalt Catalysts (Co/*M*-SiO_2) in CO_2 Hydrogenation





K. Belthle *et al.* J. Am. Chem. Soc. 144, 21232 (2022)

Application: Describing the Performance of Silica-Supported Cobalt Catalysts (Co/*M*-SiO_2) in CO_2 Hydrogenation

Creating an Experimental-Theoretical Dataset

Measured Performance

Methanol Selectivity

Measured at
4 Reaction Conditions

(6*4=24 data points)

R. Miyazaki *et al.* JACS 146, 5433 (2024)

Application: Describing the Performance of Silica-Supported Cobalt Catalysts (Co/*M*-SiO_2) in CO_2 Hydrogenation

Creating an Experimental-Theoretical Dataset

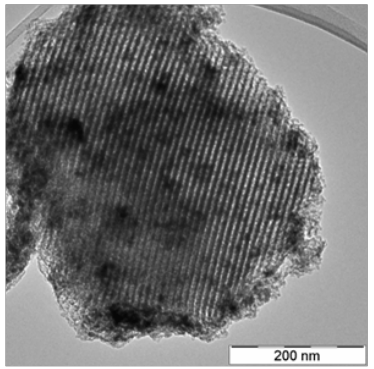
Measured Performance

Methanol Selectivity

Measured at
4 Reaction Conditions

(6*4=24 data points)

Experimental



- H_2-TPR
- SEM-EDX
- N_2 adsorption
- CO_2/NH_3-TPD, …

33 Candidate Descriptive Parameters

R. Miyazaki *et al.* JACS 146, 5433 (2024)

Application: Describing the Performance of Silica-Supported Cobalt Catalysts (Co/*M*-SiO_2) in CO_2 Hydrogenation

Creating an Experimental-Theoretical Dataset

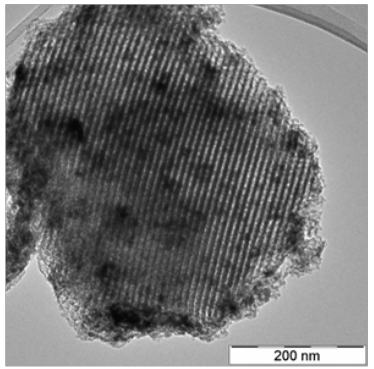
Measured Performance

Methanol Selectivity

Measured at
4 Reaction Conditions

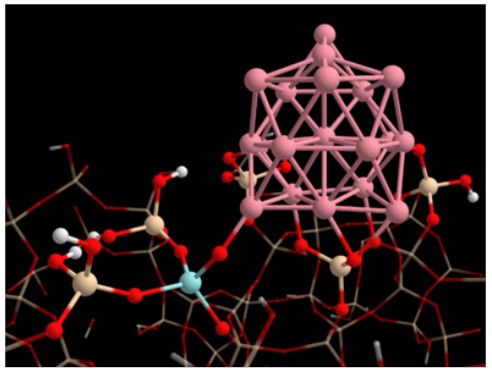
(6*4=24 data points)

Experimental



- H_2-TPR
- SEM-EDX
- N_2 adsorption
- CO_2/NH_3-TPD, …

Theoretical



33 Candidate Descriptive Parameters

R. Miyazaki *et al.* JACS 146, 5433 (2024)

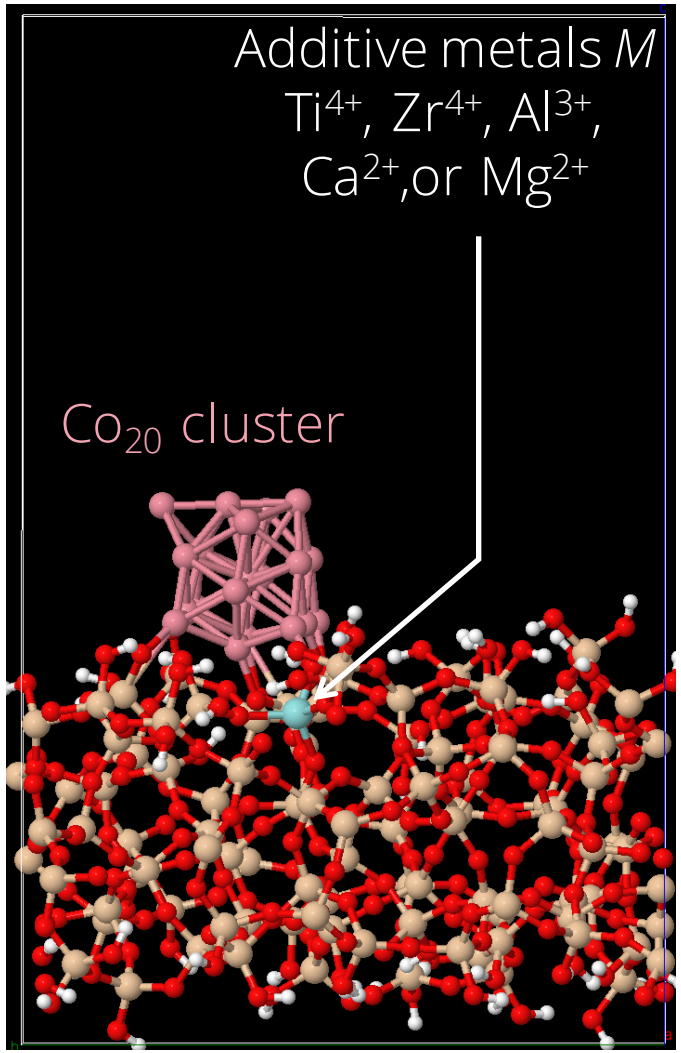
Application: Describing the Performance of Silica-Supported Cobalt Catalysts (Co/*M*-SiO_2) in CO_2 Hydrogenation

Theoretical Candidate Descriptive Parameters

- DFT-RPBE simulations of Co/*M*-SiO_2 interface

FHI-aims
The ab initio materials simulation package

$E_{ads}^{CO_2}$	CO_2 adsorption energy	eV
Δ_{C-O}	C–O bond elongation of adsorbed CO_2	Å
$\Delta_{\angle O-C-O}$	O–C–O angle bending of adsorbed CO_2	degree
q_{CO_2}	Hirshfeld charge of adsorbed CO_2	e
q_M	Hirshfeld charge of the additive metal	e
E_{CH_3O}	CH_3O Formation energy	eV
E_{COOH}	COOH Formation energy	eV
E_{CO+O}	CO + O formation energy	eV
E_{HCOO}	HCOO formation energy	eV
E_{ads}^{O}	O adsorption energy	eV



R. Miyazaki *et al.* JACS 146, 5433 (2024)

Application: Describing the Performance of Silica-Supported Cobalt Catalysts (Co/*M*-SiO_2) in CO_2 Hydrogenation

Creating an Experimental-Theoretical Dataset

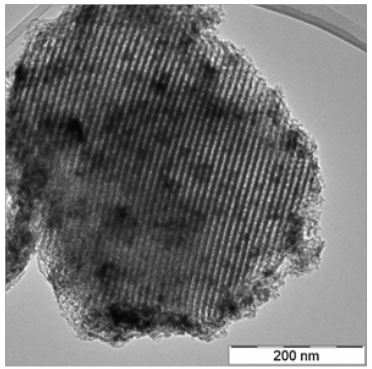
Measured Performance

Methanol Selectivity

Measured at
4 Reaction Conditions

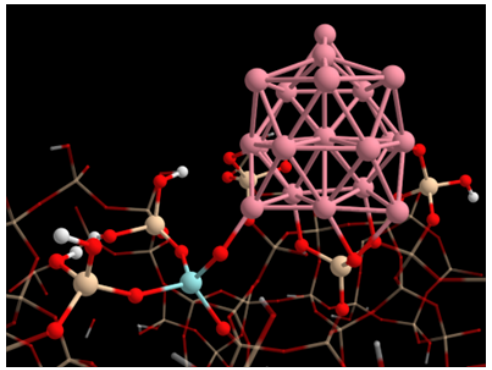
(6*4=24 data points)

Experimental



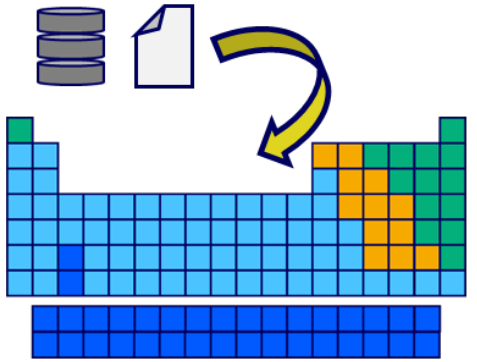
- H_2-TPR
- SEM-EDX
- N_2 adsorption
- CO_2/NH_3-TPD, …

Theoretical



- DFT-RPBE Simulations

Elemental



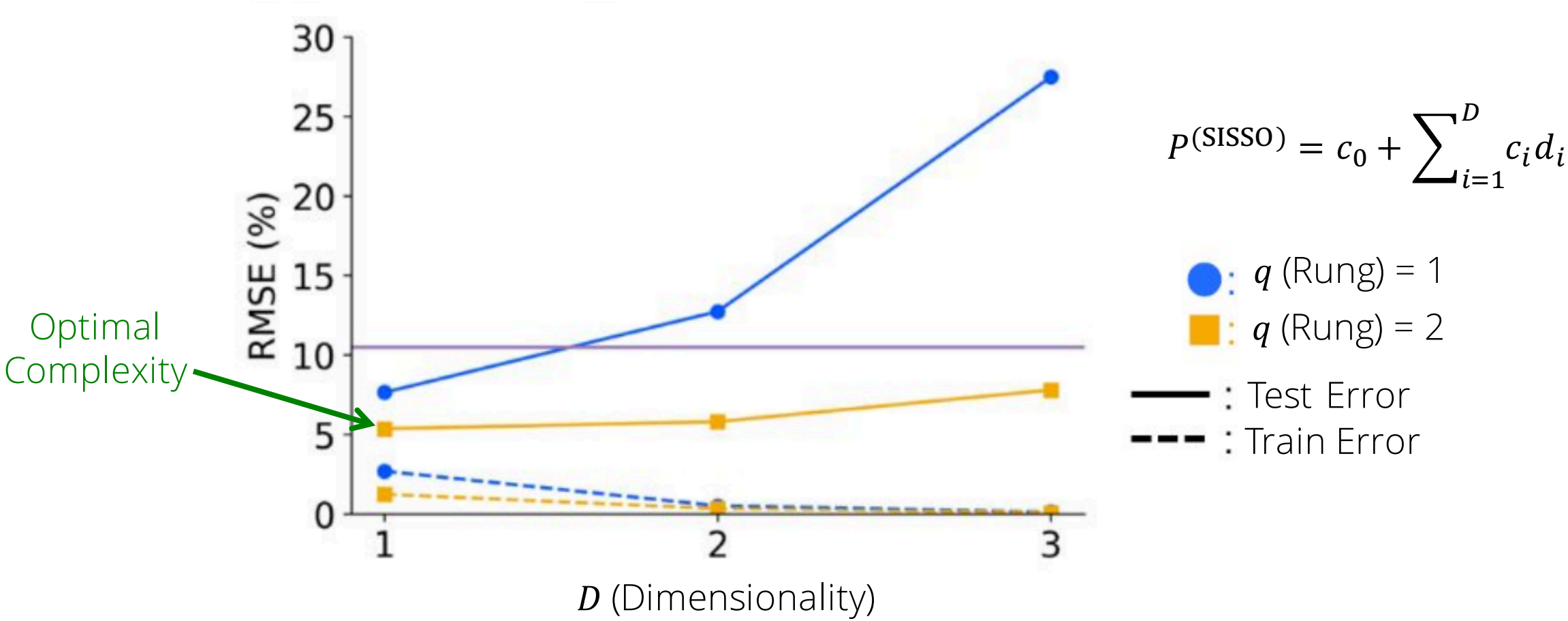
- Experiment
- Calculation

33 Candidate Descriptive Parameters

R. Miyazaki *et al.* JACS 146, 5433 (2024)

Application: Describing the Performance of Silica-Supported Cobalt Catalysts (Co/*M*-SiO_2) in CO_2 Hydrogenation

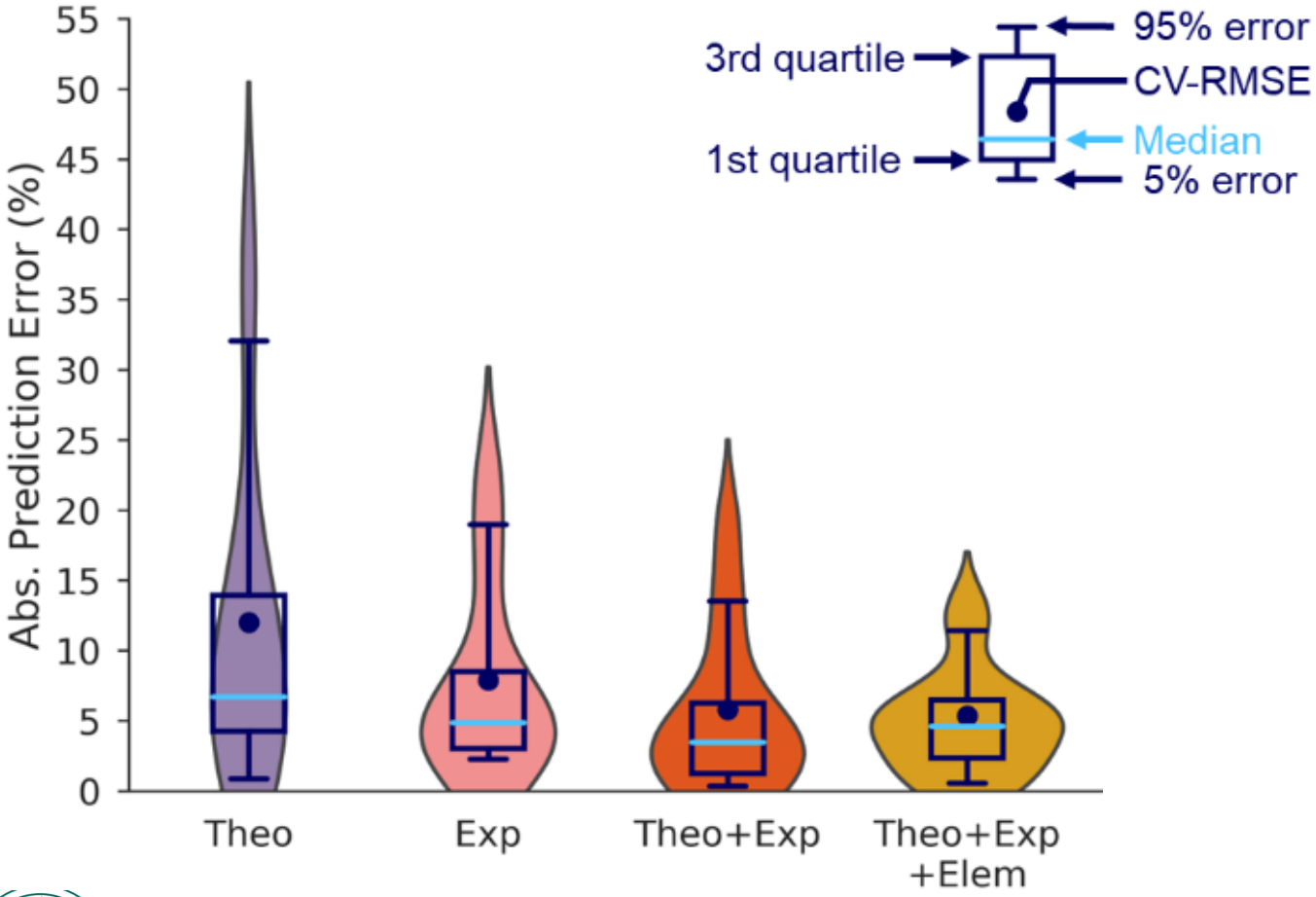
Determination of SISSO Model Hyperparameters (Leave-One-Material-Out Cross Validation)



R. Miyazaki *et al.* JACS 146, 5433 (2024)

Application: Describing the Performance of Silica-Supported Cobalt Catalysts (Co/*M*-SiO_2) in CO_2 Hydrogenation

Error Distributions Associated to SISSO Models for Methanol Selectivity



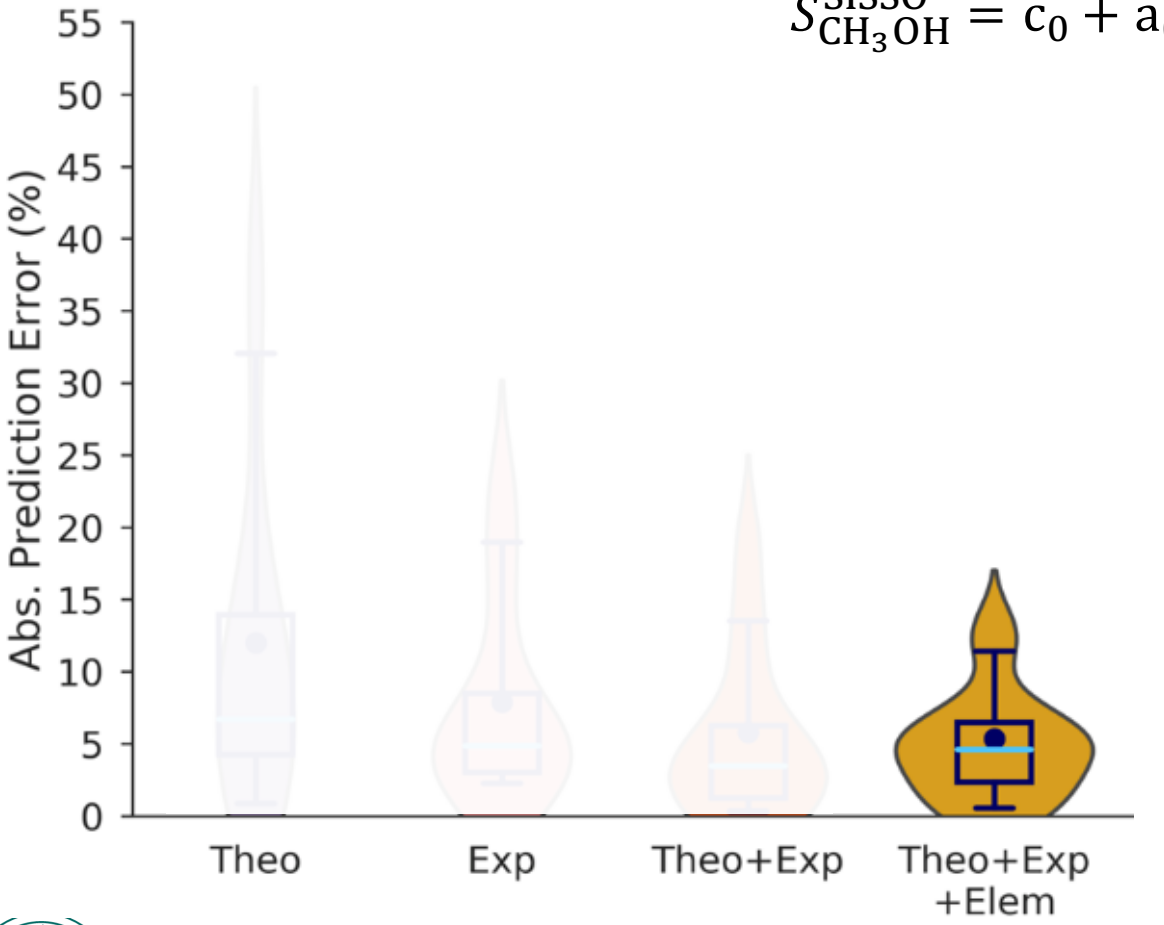
- Improved Description Combining Theoretical (Theo), Experimental (Exp), and Elemental (Elem) Primary Features

R. Miyazaki *et al.* JACS 146, 5433 (2024)

Application: Describing the Performance of Silica-Supported Cobalt Catalysts (Co/*M*-SiO_2) in CO_2 Hydrogenation

CO_2 Conversion (Exp)

$$S^{\text{SISSO}}_{\text{CH}_3\text{OH}} = c_0 + a_0 \left\{ \frac{M_{\text{cov}} * X^{RGSV}_{CO2}}{E_{CH3O} * T_{H2}} \right\}$$



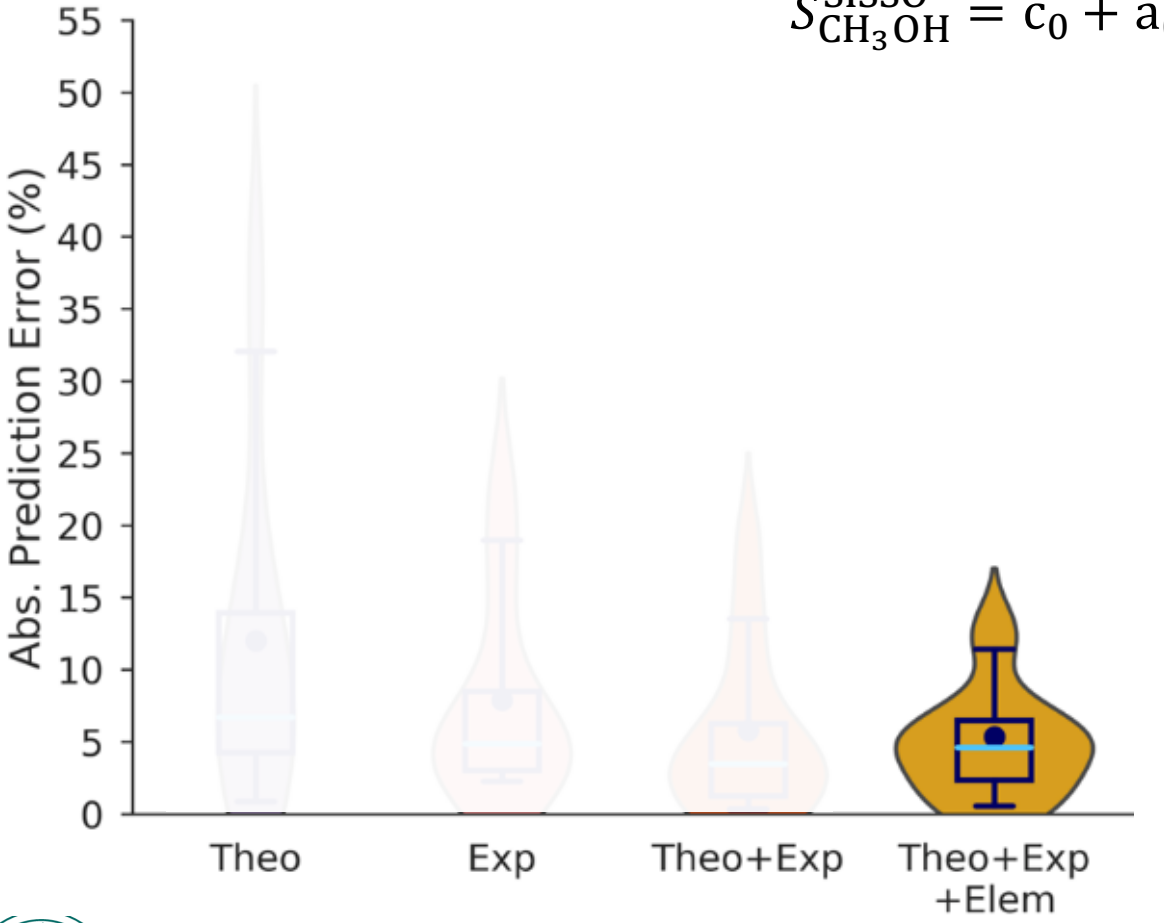
R. Miyazaki *et al.* JACS 146, 5433 (2024)

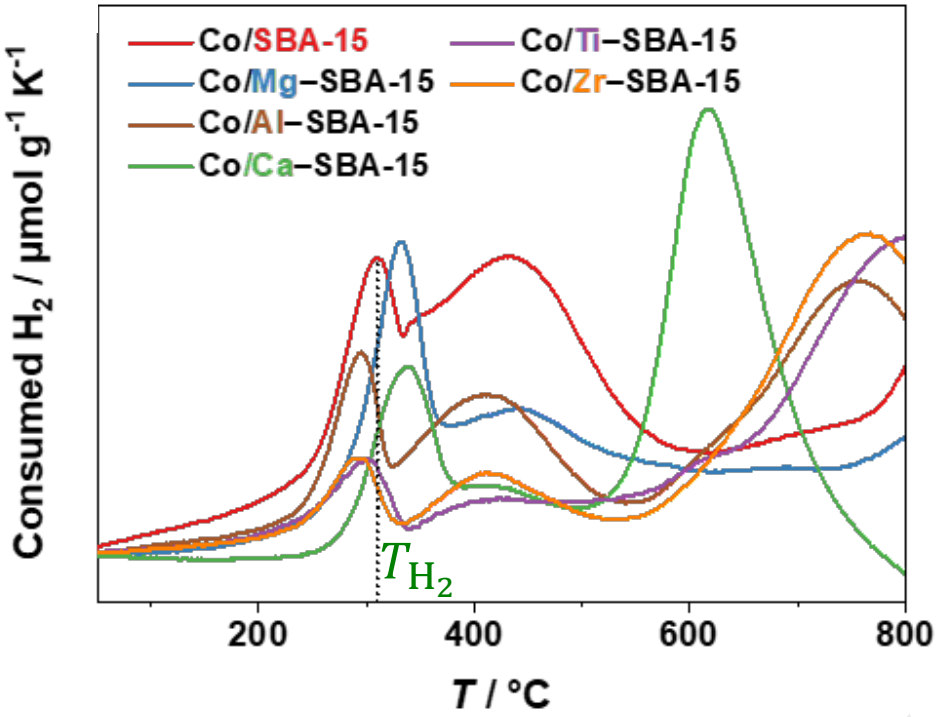
Application: Describing the Performance of Silica-Supported Cobalt Catalysts (Co/*M*-SiO_2) in CO_2 Hydrogenation

$$S_{CH_3OH}^{SISSO} = c_0 + a_0 \left\{ \frac{M_{cov} * X_{CO2}^{RGSV}}{E_{CH3O} * T_{H2}} \right\}$$

CO_2 Conversion (Exp)

Temperature of 1st H_2-TPR Signal (Exp)





R. Miyazaki *et al.* JACS 146, 5433 (2024)

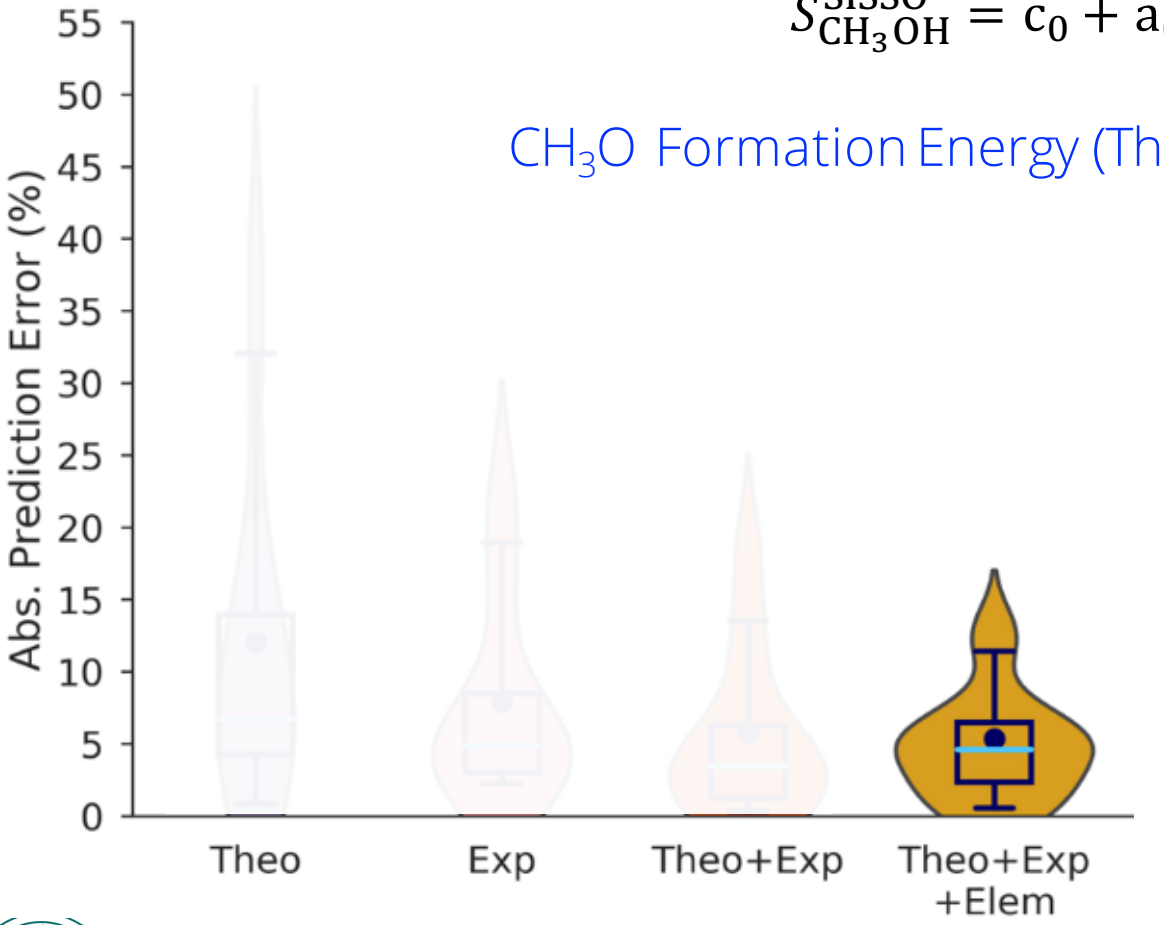
Application: Describing the Performance of Silica-Supported Cobalt Catalysts (Co/*M*-SiO_2) in CO_2 Hydrogenation

$$S^{\text{SISSO}}_{\text{CH}_3\text{OH}} = c_0 + a_0 \left\{ \frac{M_{\text{cov}} * X^{RGSV}_{CO2}}{E_{CH3O} * T_{H2}} \right\}$$

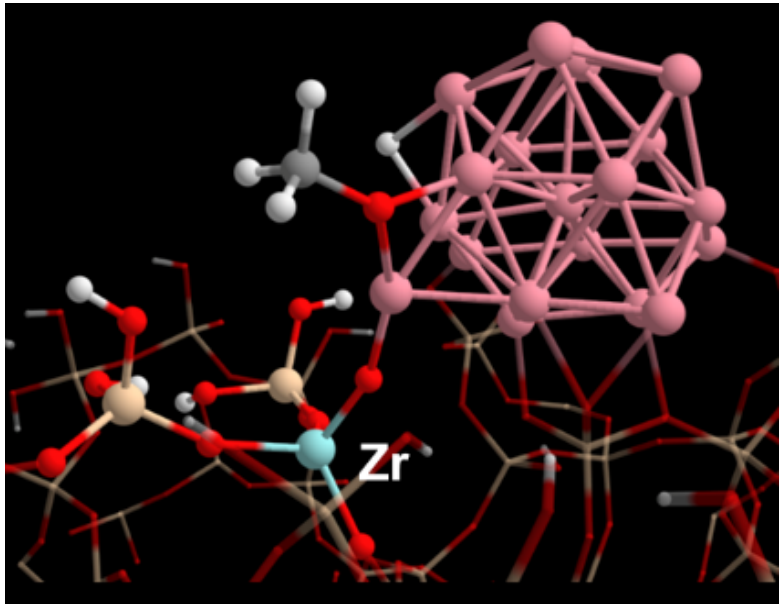
CO_2 Conversion (Exp)

CH_3O Formation Energy (Theo)

Temperature of 1st H_2-TPR Signal (Exp)



M	E_{CH_3O}(eV)
-	-1.49
Ti	-1.50
Zr	**-1.39**
Al	-0.97
Ca	-1.13
Mg	-1.55



R. Miyazaki *et al.* JACS 146, 5433 (2024)

Application: Describing the Performance of Silica-Supported Cobalt Catalysts (Co/*M*-SiO_2) in CO_2 Hydrogenation

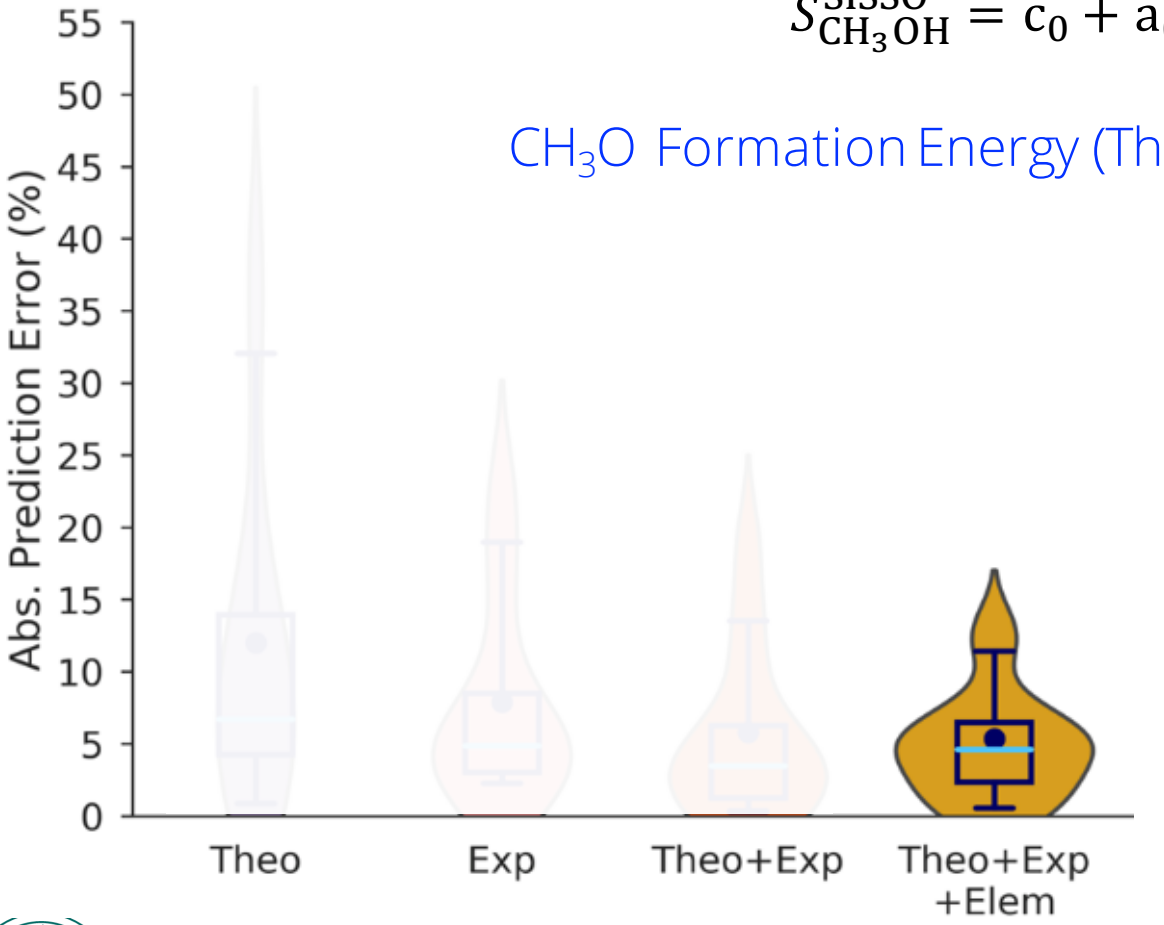
M Covalent Atomic Radius (Elem)

CO_2 Conversion (Exp)

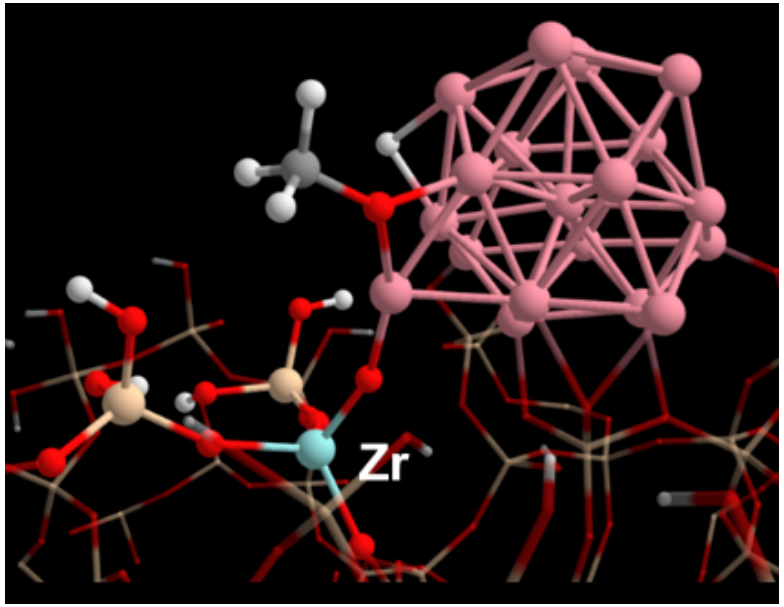
$$S^{\text{SISSO}}_{\text{CH}_3\text{OH}} = c_0 + a_0 \left\{ \frac{M_{\text{cov}} * X^{RGSV}_{CO2}}{E_{CH3O} * T_{H2}} \right\}$$

CH_3O Formation Energy (Theo)

Temperature of 1st H_2-TPR Signal (Exp)



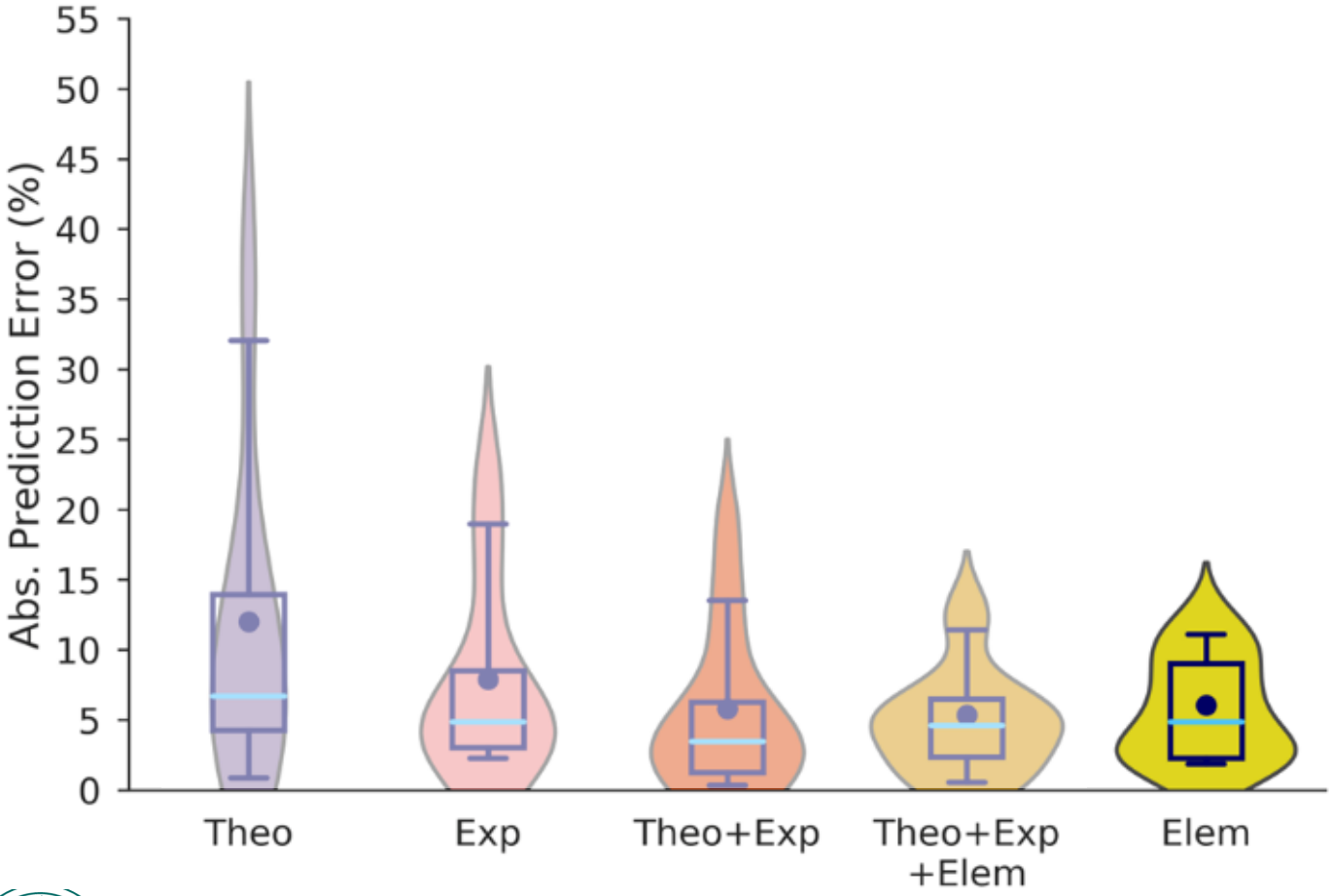
M	E_{CH_3O}(eV)
-	-1.49
Ti	-1.50
Zr	**-1.39**
Al	-0.97
Ca	-1.13
Mg	-1.55



R. Miyazaki *et al.* JACS 146, 5433 (2024)

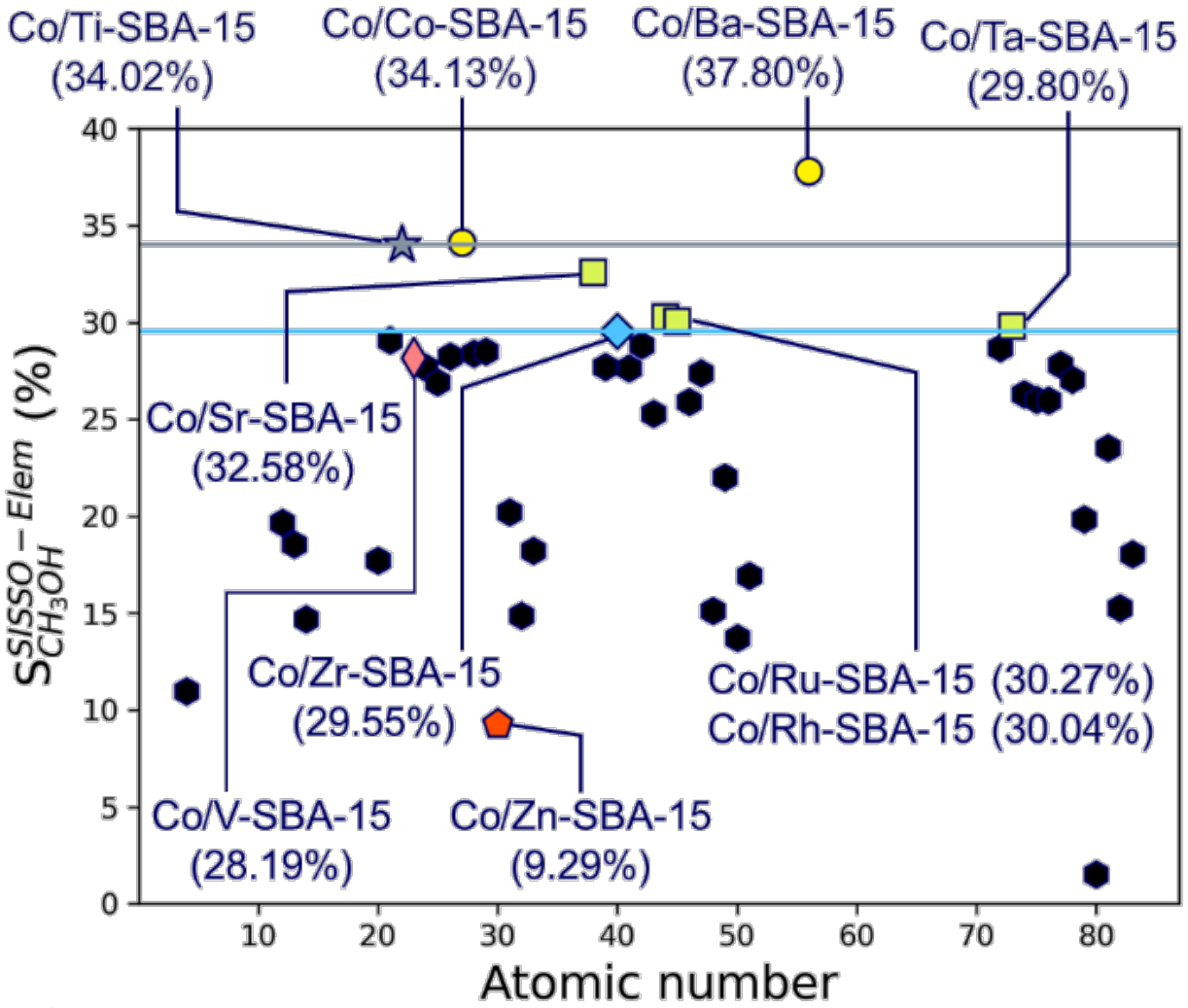
Application: Describing the Performance of Silica-Supported Cobalt Catalysts (Co/*M*-SiO_2) in CO_2 Hydrogenation

SISSO Model Based Only on Elemental Parameters Suggests New Supports



R. Miyazaki *et al.* JACS 146, 5433 (2024)

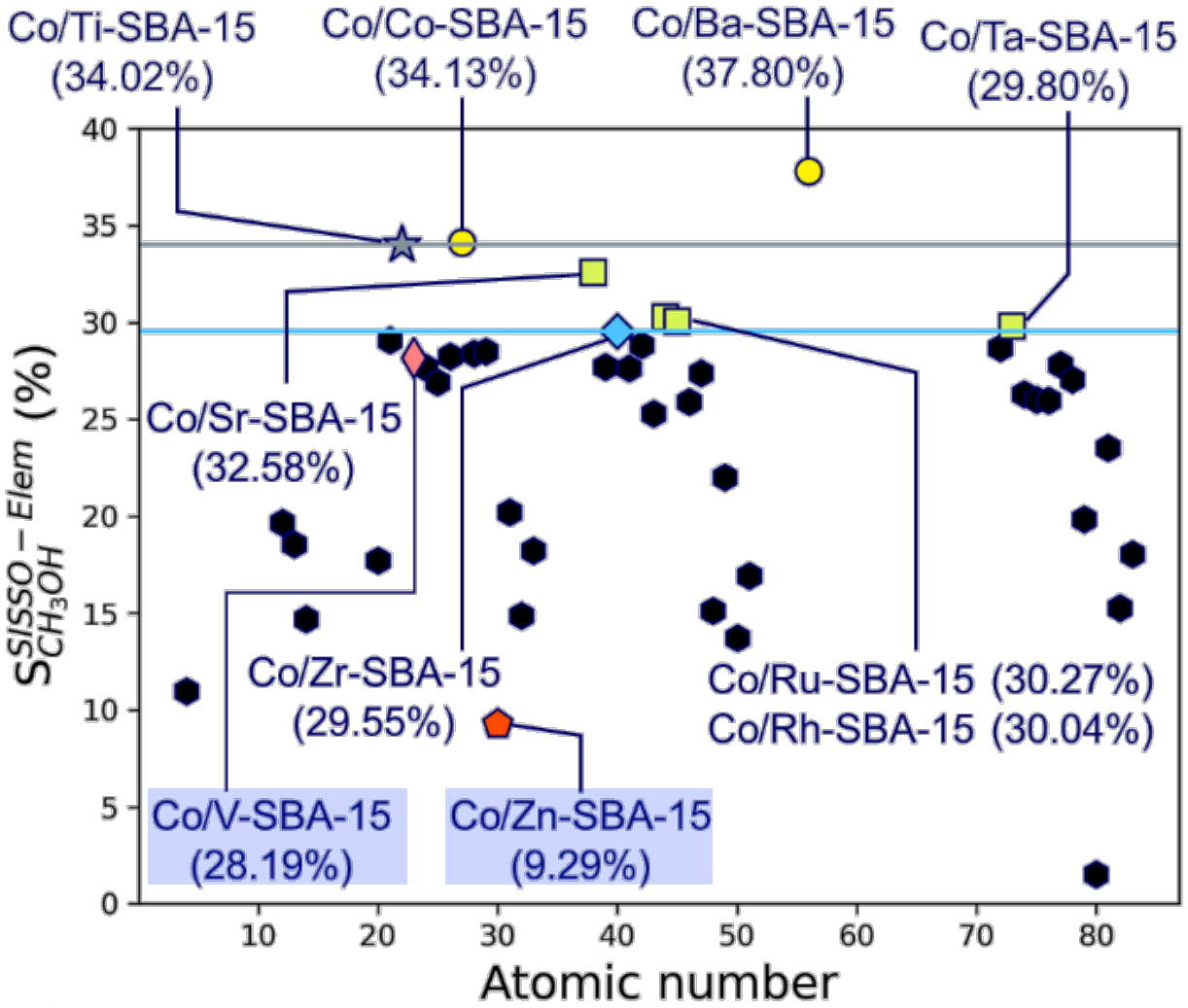
Application: Describing the Performance of Silica-Supported Cobalt Catalysts (Co/*M*-SiO_2) in CO_2 Hydrogenation



SISSO Model Based Only on Elemental Parameters Suggests New Supports

R. Miyazaki *et al.* JACS 146, 5433 (2024)

Application: Describing the Performance of Silica-Supported Cobalt Catalysts (Co/*M*-SiO_2) in CO_2 Hydrogenation



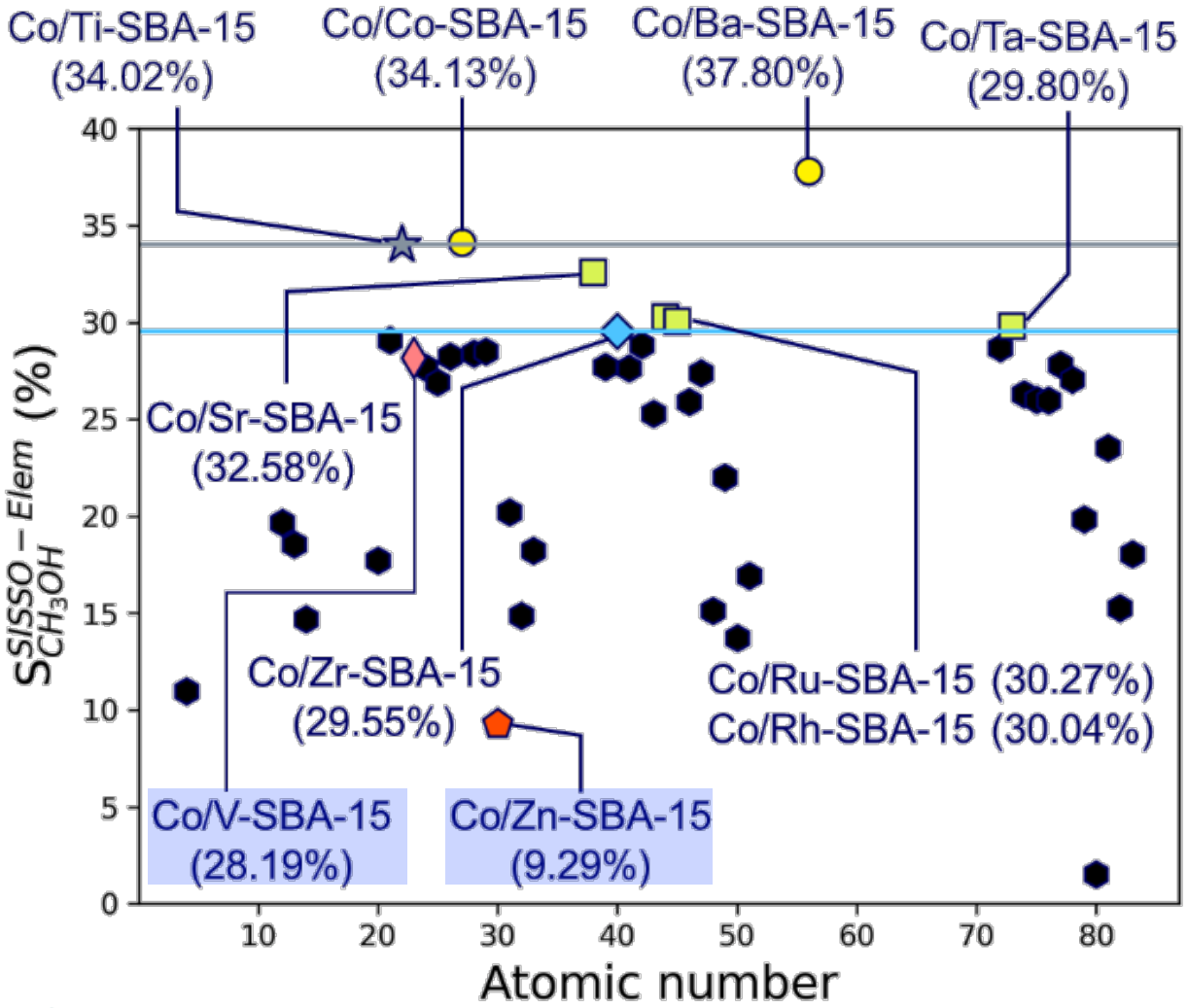
SISSO Model Based Only on Elemental Parameters Suggests New Supports

material	$S^{\mathrm{SISSO}}_{\mathrm{CH_3OH}}$ (predicted)	$\sigma^{\mathrm{SISSO}}_{S_{\mathrm{CH_3OH}}}$
Co/**V**-SiO_2	28.19	3.9
Co/**Zn**–SiO_2	9.29	20.6

↑ Standard Deviation of Predictions of SISSO Models

R. Miyazaki *et al.* JACS 146, 5433 (2024)

Application: Describing the Performance of Silica-Supported Cobalt Catalysts (Co/*M*-SiO_2) in CO_2 Hydrogenation



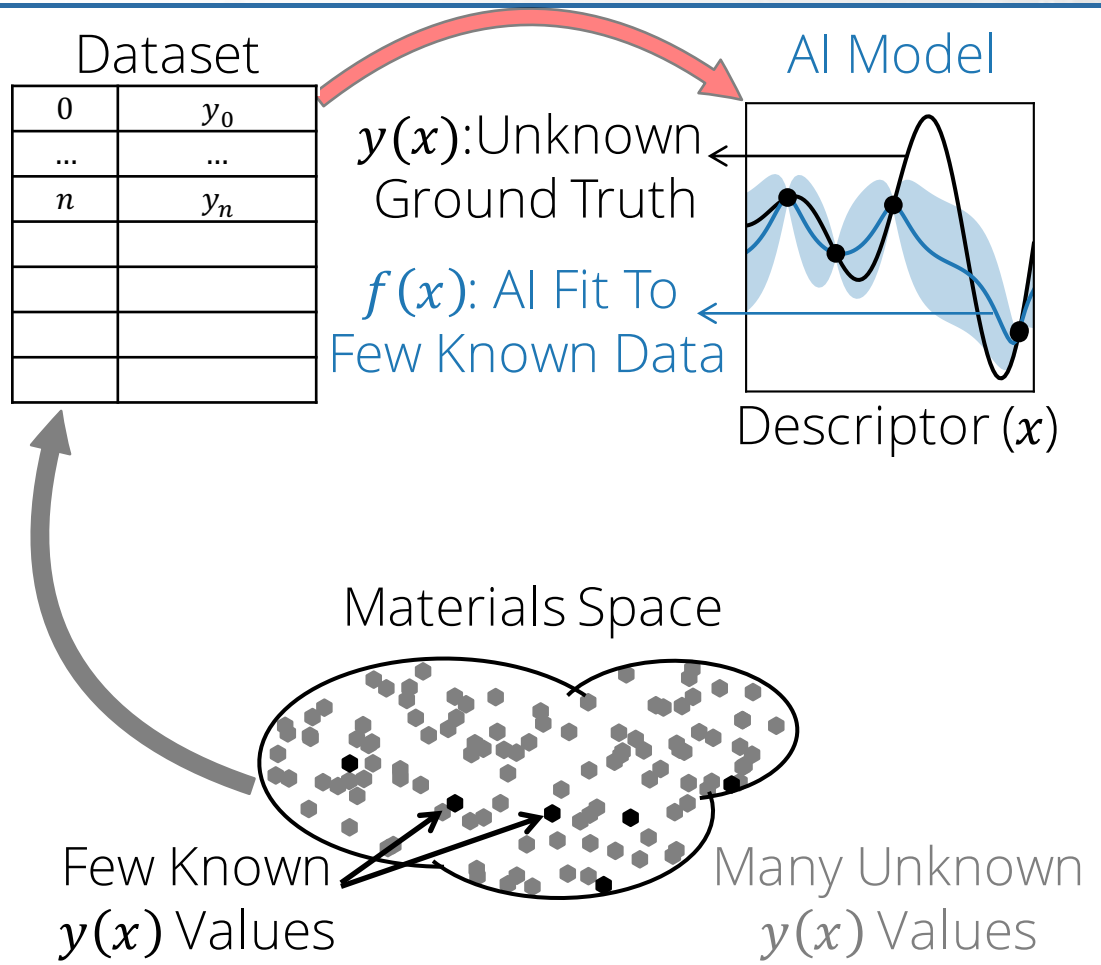
SISSO Model Based Only on Elemental Parameters Suggests New Supports

material	$S^{SISSO}_{CH_3OH}$ (predicted)	$\sigma^{SISSO}_{S_{CH_3OH}}$	$S^{exp}_{CH_3OH}$ (measured)
Co/**V**-SiO_2	28.19	3.9	32.5
Co/**Zn**–SiO_2	9.29	20.6	16.9

↑ Standard Deviation of Predictions of SISSO Models

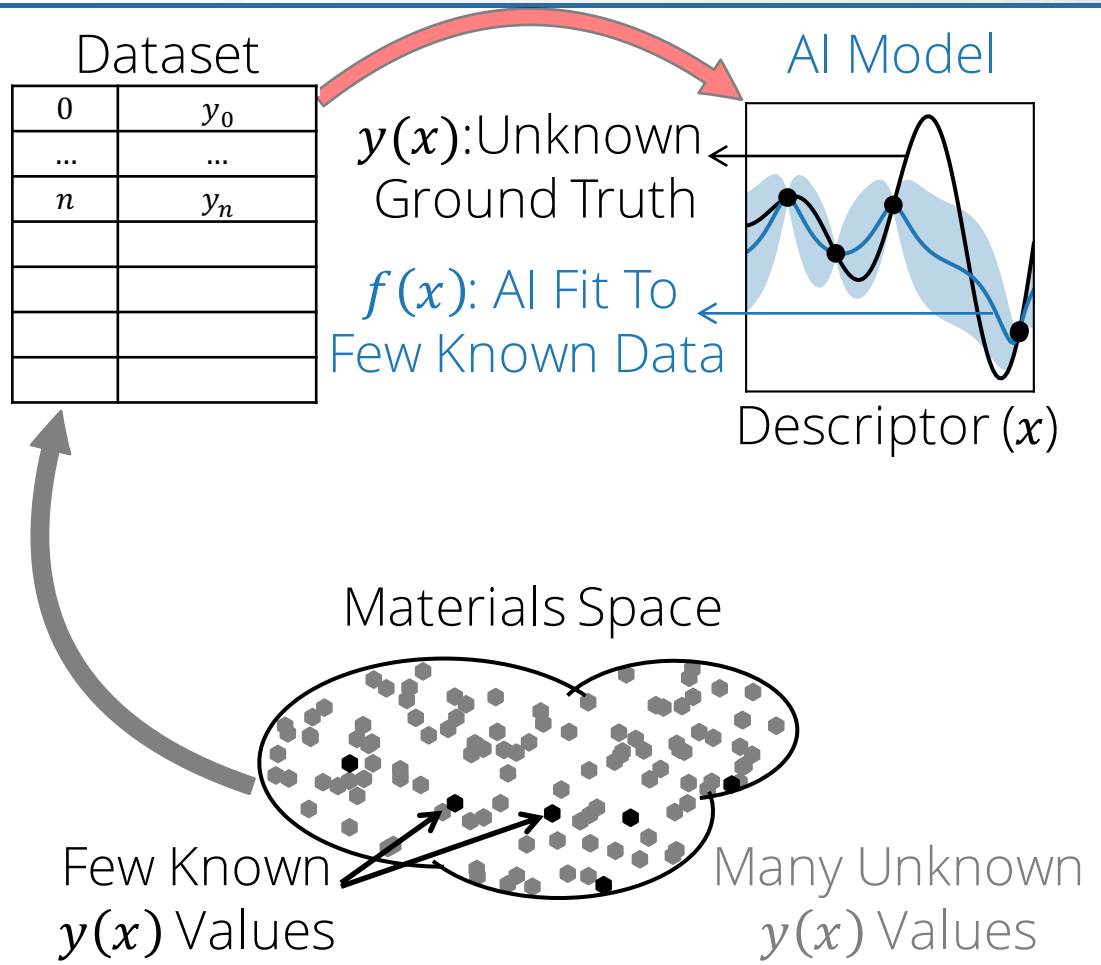
R. Miyazaki *et al.* JACS 146, 5433 (2024)

AI Models of Materials Properties



AI Models of Materials Properties

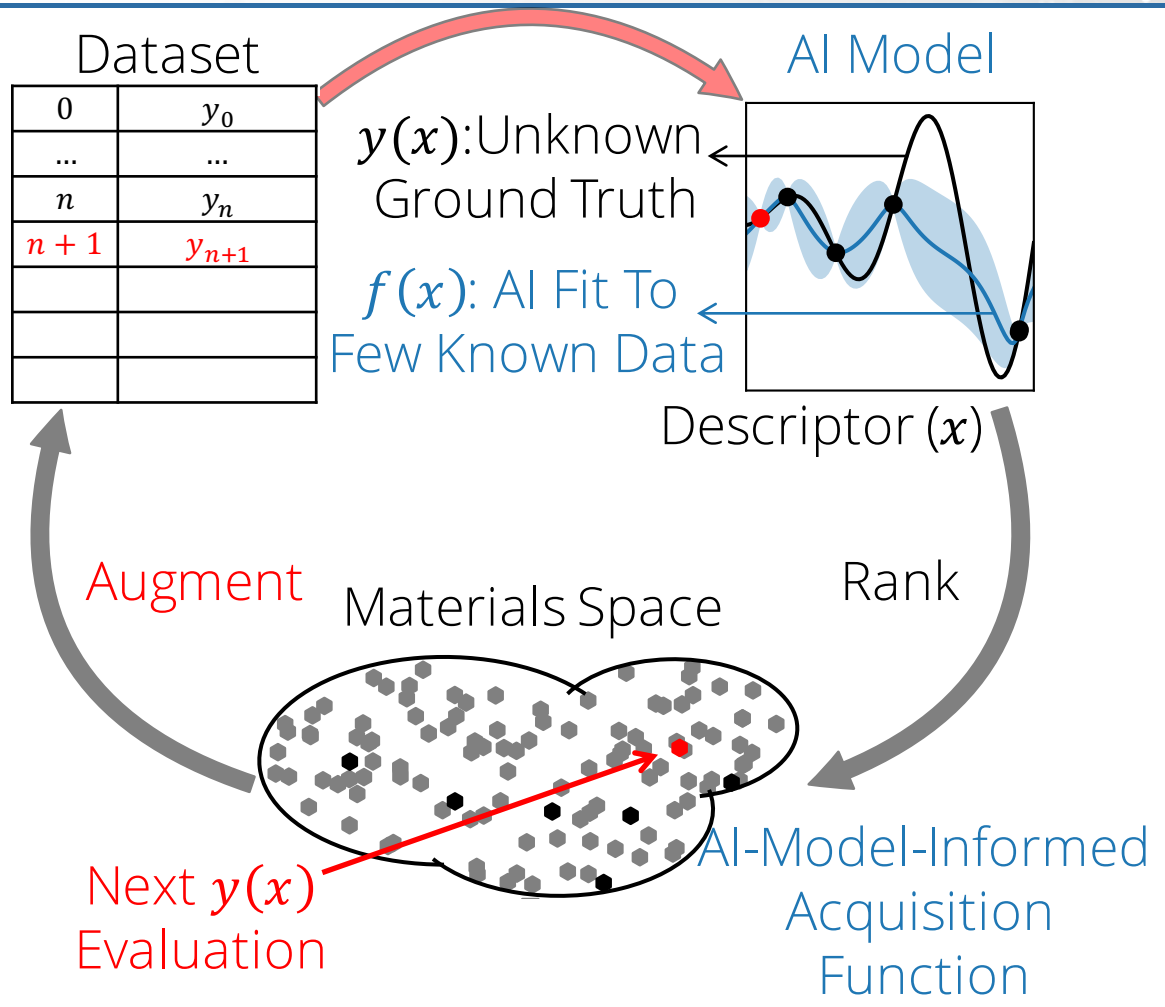
- Model Performance Assessed in Average
- Good Average Performance Might Not Translate to Good Description of Promising Materials



S. Kauwe *et al.*, Comput. Mat. Sci. 174, 109498 (2020)
B. Meredig *et al.*, Mol. Syst. Des. Eng. 3, 819 (2018)

From Prediction to Action: AI-Guided Workflows for the Discovery of Materials with Optimal Properties

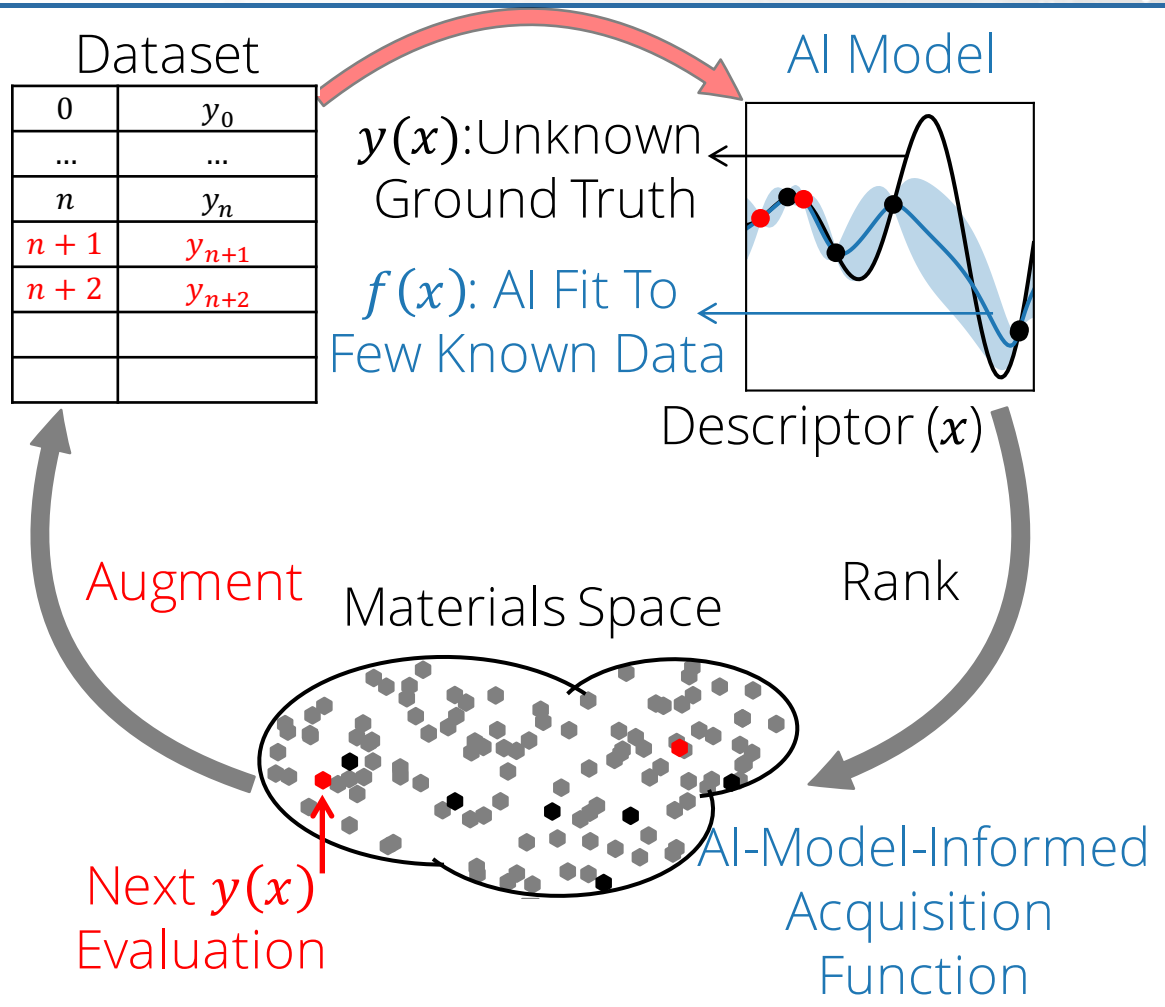
- Iterative Modelling and Data Acquisition Approach to Identify Materials with Desired Property Values (e.g., High)



B. Shahriari *et al.*, Proc. IEEE, 104, 1 148 (2016)

From Prediction to Action: AI-Guided Workflows for the Discovery of Materials with Optimal Properties

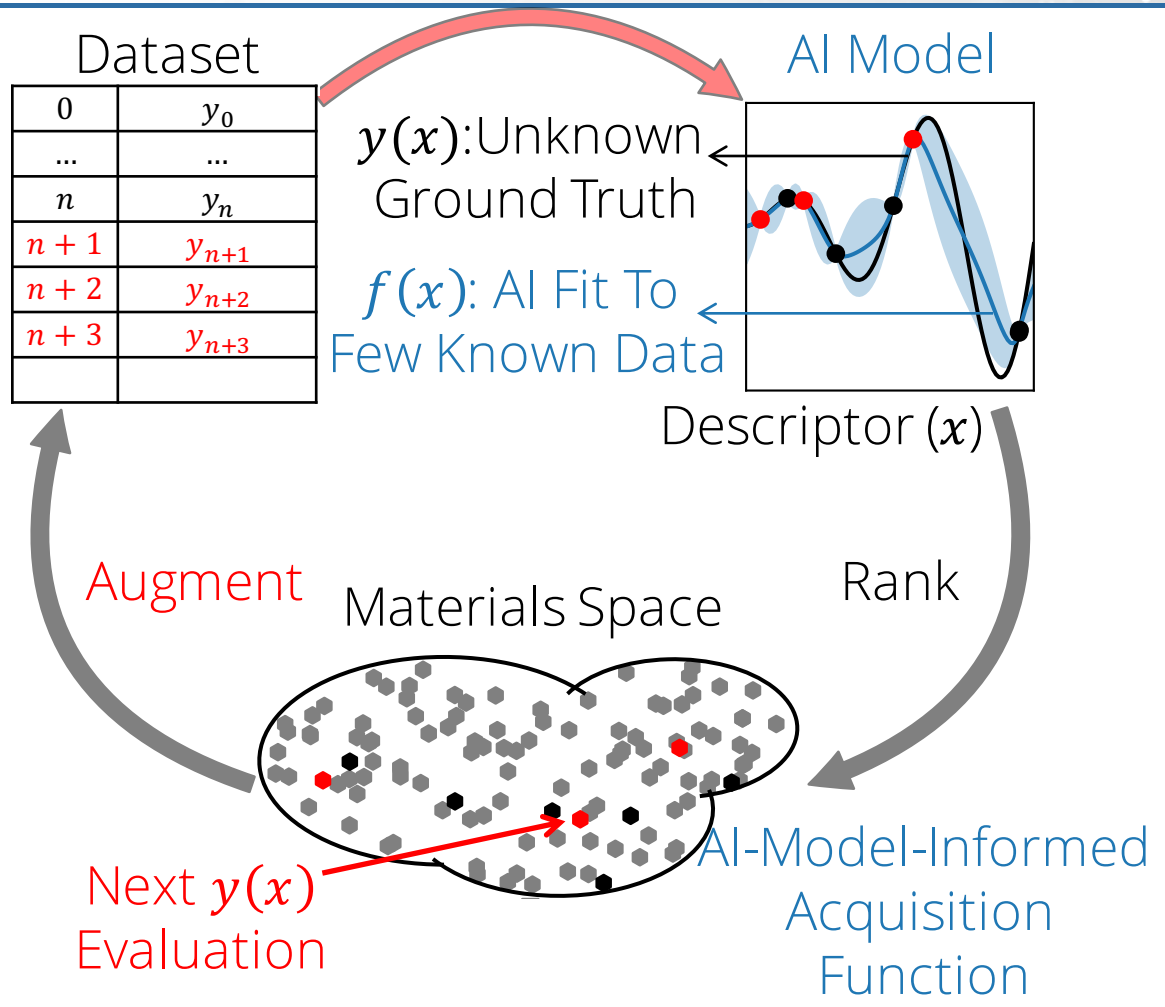
- Iterative Modelling and Data Acquisition Approach to Identify Materials with Desired Property Values (e.g., High)



B. Shahriari *et al.*, Proc. IEEE, 104, 1 148 (2016)

From Prediction to Action: AI-Guided Workflows for the Discovery of Materials with Optimal Properties

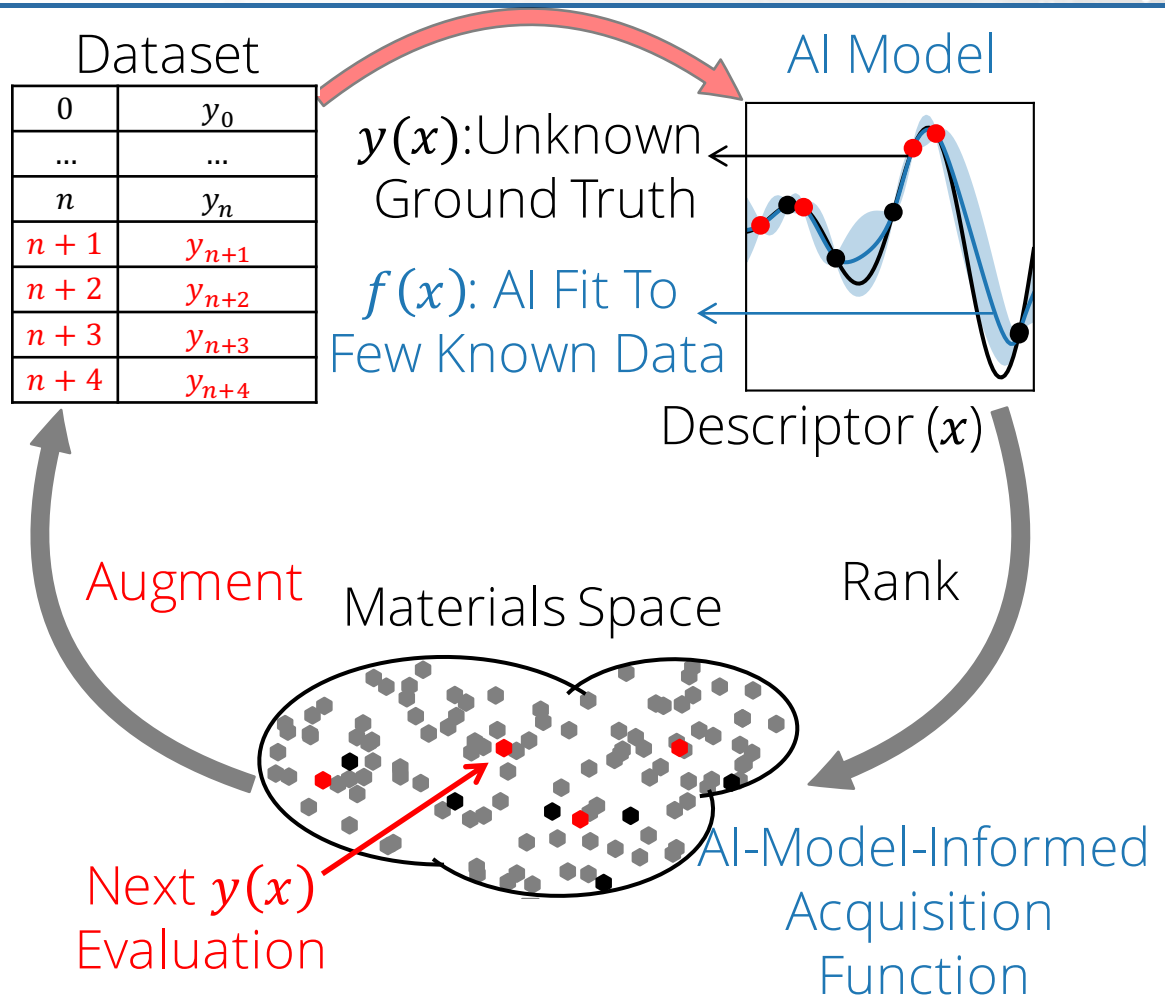
- Iterative Modelling and Data Acquisition Approach to Identify Materials with Desired Property Values (e.g., High)



B. Shahriari *et al.*, Proc. IEEE, 104, 1 148 (2016)

From Prediction to Action: AI-Guided Workflows for the Discovery of Materials with Optimal Properties

- Iterative Modelling and Data Acquisition Approach to Identify Materials with Desired Property Values (e.g., High)

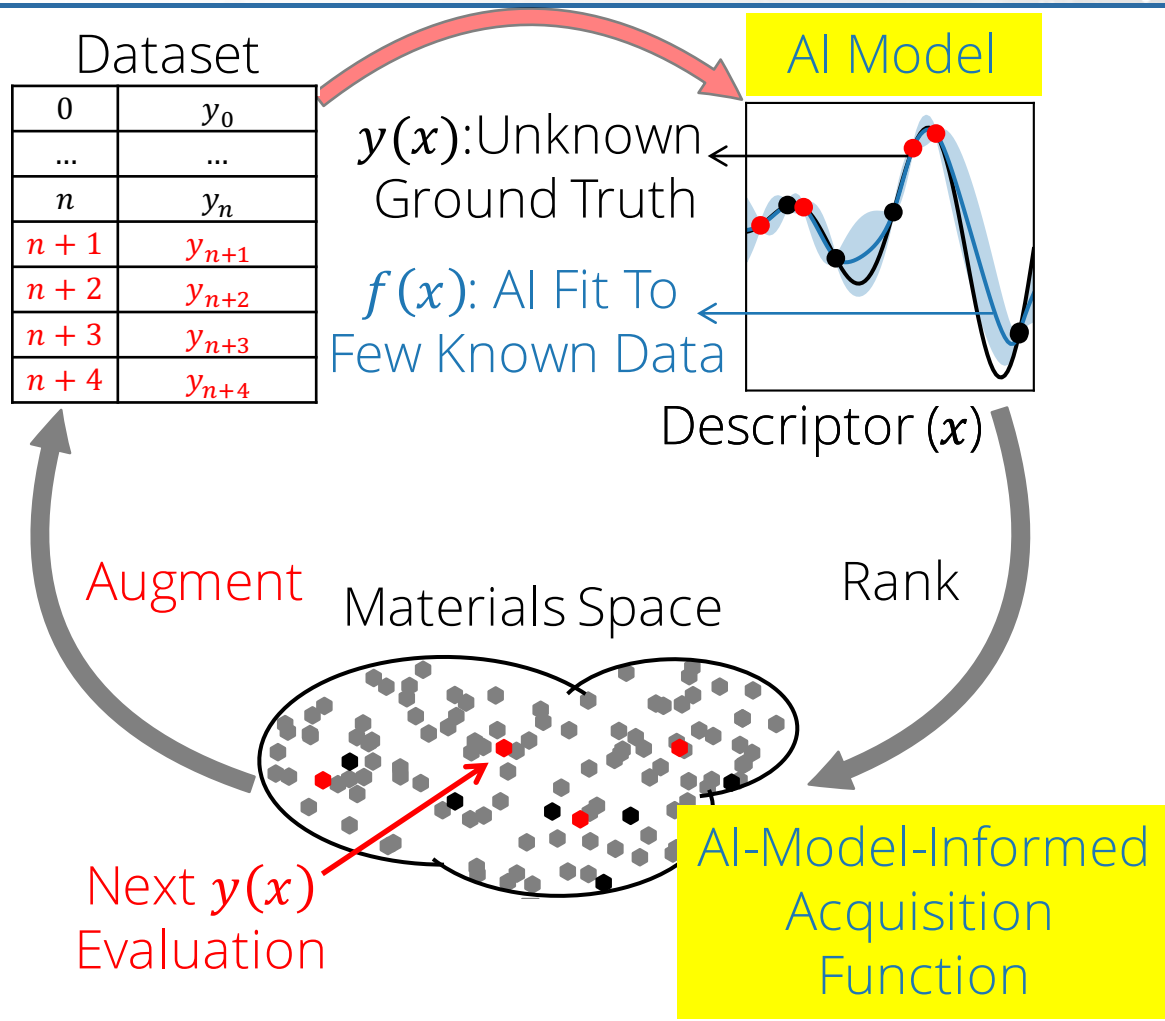


B. Shahriari *et al.*, Proc. IEEE, 104, 1 148 (2016)

From Prediction to Action: AI-Guided Workflows for the Discovery of Materials with Optimal Properties

- Iterative Modelling and Data Acquisition Approach to Identify Materials with Desired Property Values (e.g., High)

1. Impact of Choices of AI Model and Acquisition Function?



0	y_0
...	...
n	y_n
$n+1$	y_{n+1}
$n+2$	y_{n+2}
$n+3$	y_{n+3}
$n+4$	y_{n+4}

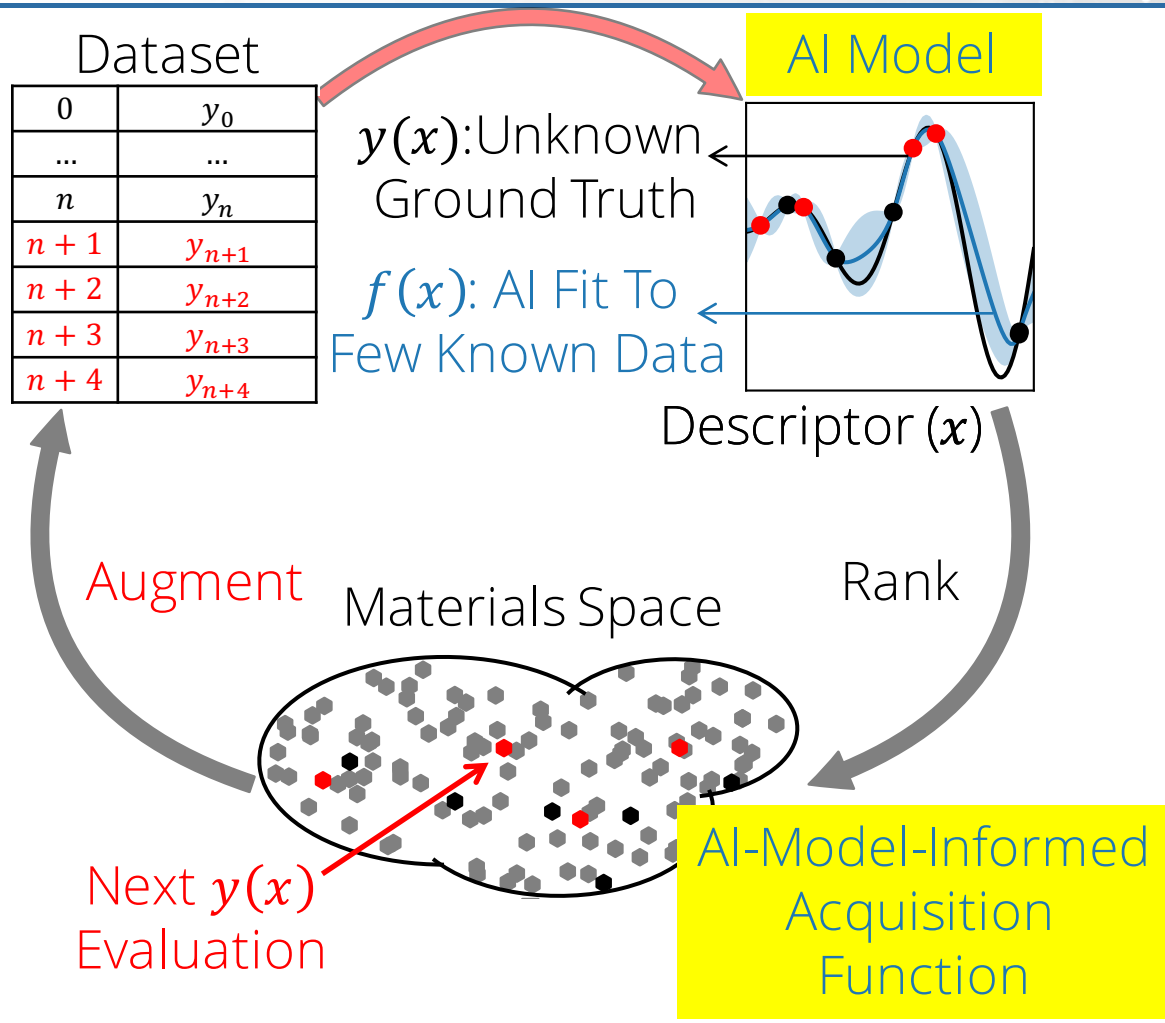
B. Shahriari *et al.*, Proc. IEEE, 104, 1 148 (2016)

From Prediction to Action: AI-Guided Workflows for the Discovery of Materials with Optimal Properties

- Iterative Modelling and Data Acquisition Approach to Identify Materials with Desired Property Values (e.g., High)

1. Impact of Choices of AI Model and Acquisition Function?

2. SISSO-Guided Workflow?



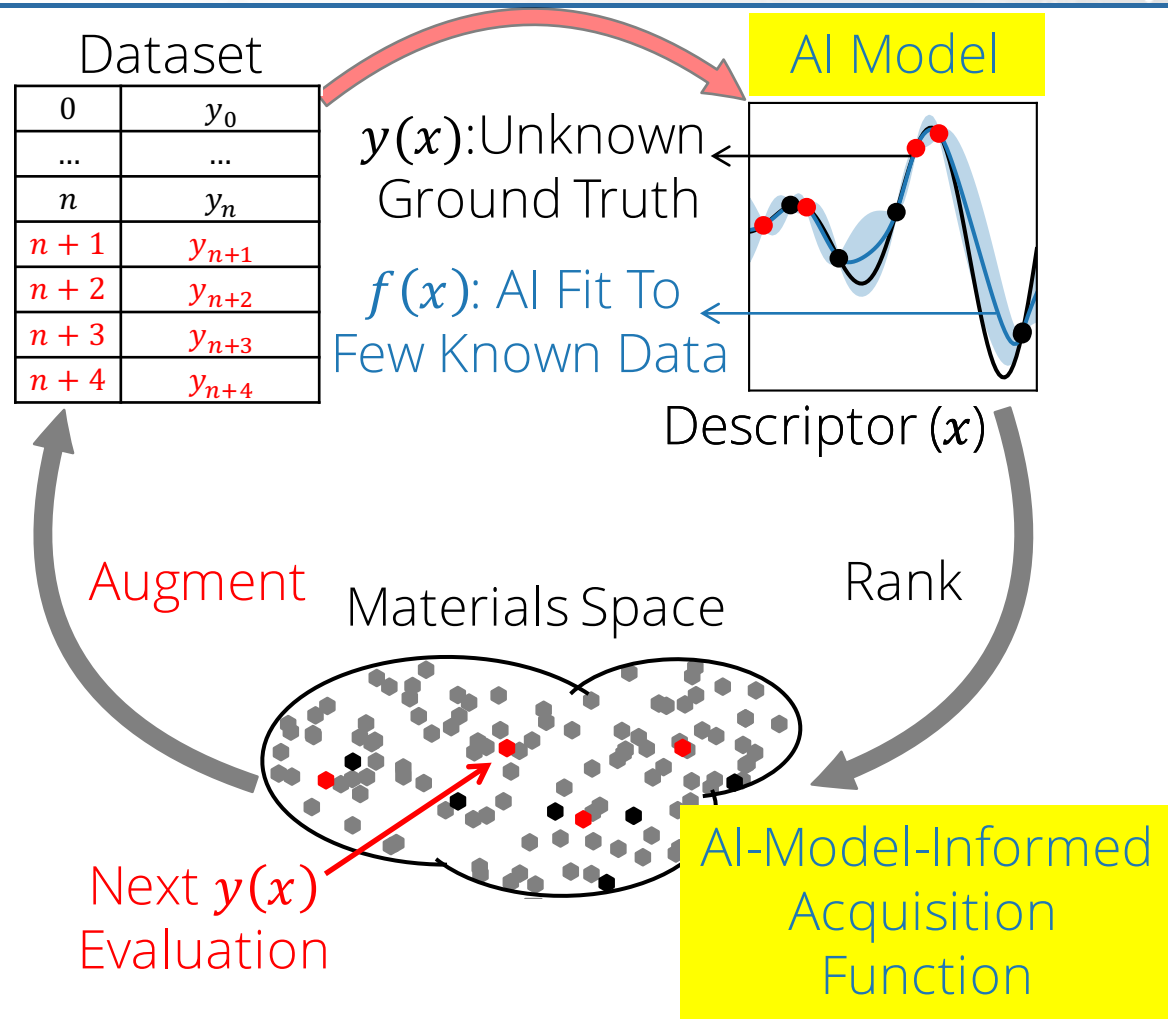
B. Shahriari *et al.*, Proc. IEEE, 104, 1 148 (2016)

From Prediction to Action: AI-Guided Workflows for the Discovery of Materials with Optimal Properties

- Iterative Modelling and Data Acquisition Approach to Identify Materials with Desired Property Values (e.g., High)

1. Impact of Choices of AI Model and Acquisition Function?

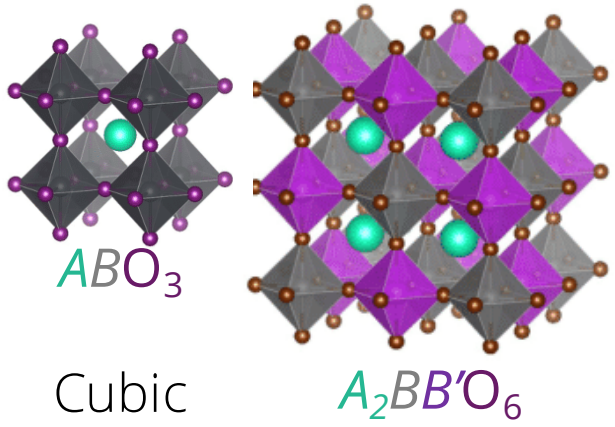
2. SISSO-Guided Workflow?

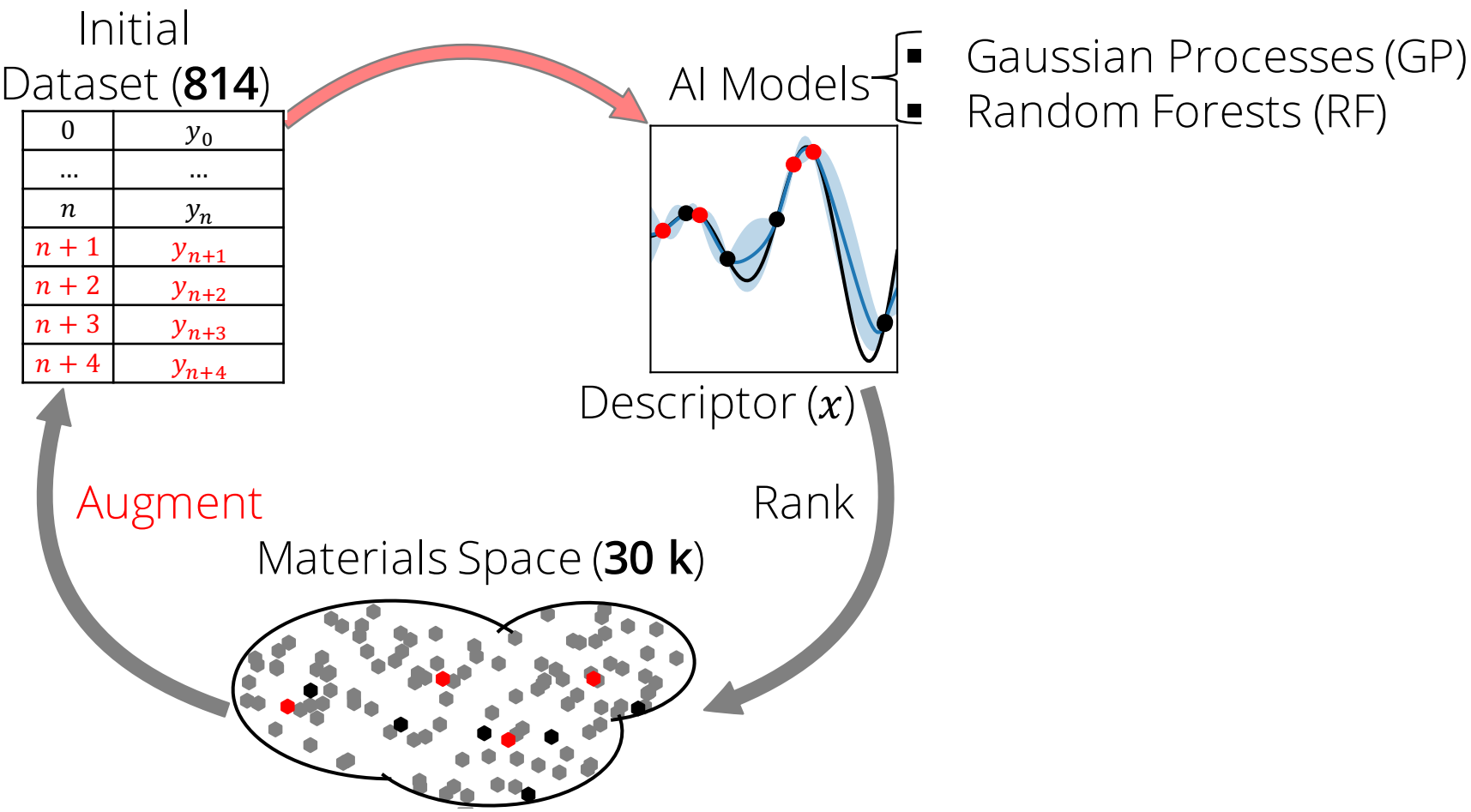


B. Shahriari *et al.*, Proc. IEEE, 104, 1 148 (2016)

Workflow 1: Identifying Perovskites with High Bulk Modulus

- Target Property (y):
  Bulk Modulus (B_0)



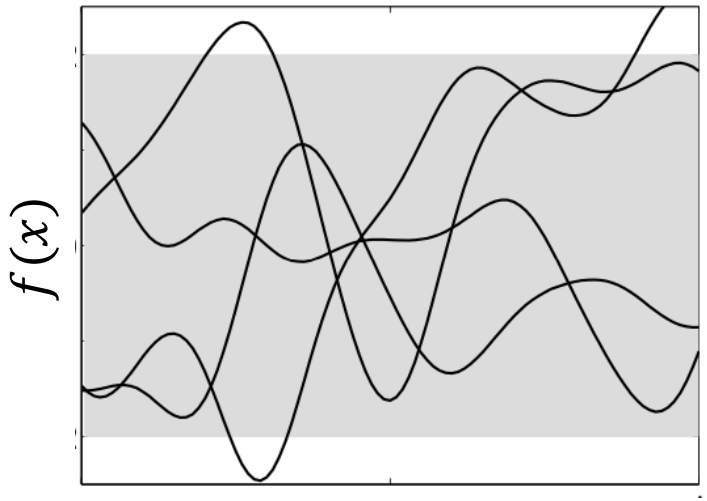


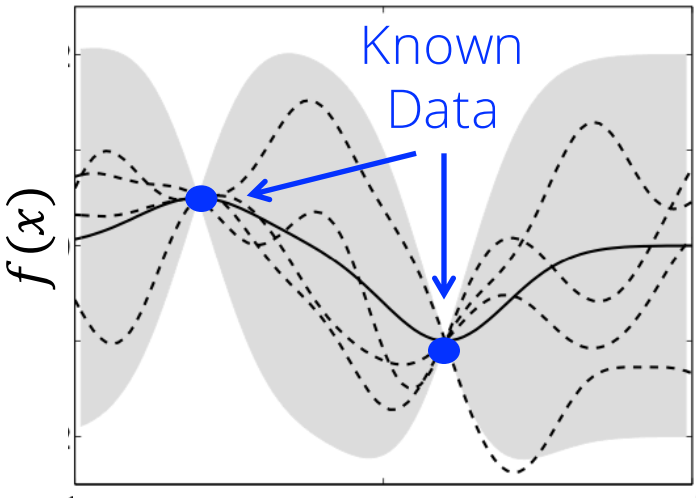
FHI-aims DFT-PBEsol
The ab initio materials simulation package

M. Boley *et al*. Modelling Simul. Mater. Sci. Eng. 32, 063301 (2024)

Gaussian Processes (GP)

- Probabilistic Model Providing Mean Predictions and Confidence Intervals (Uncertainty)
  - Learns a Distribution Over All Possible Functions that Fit the Data





Prior Distribution

- Distribution of Functions Generated Using Assumptions on Continuity and Smoothness (Kernel)
- All Functions Are Possible Before Seeing Data

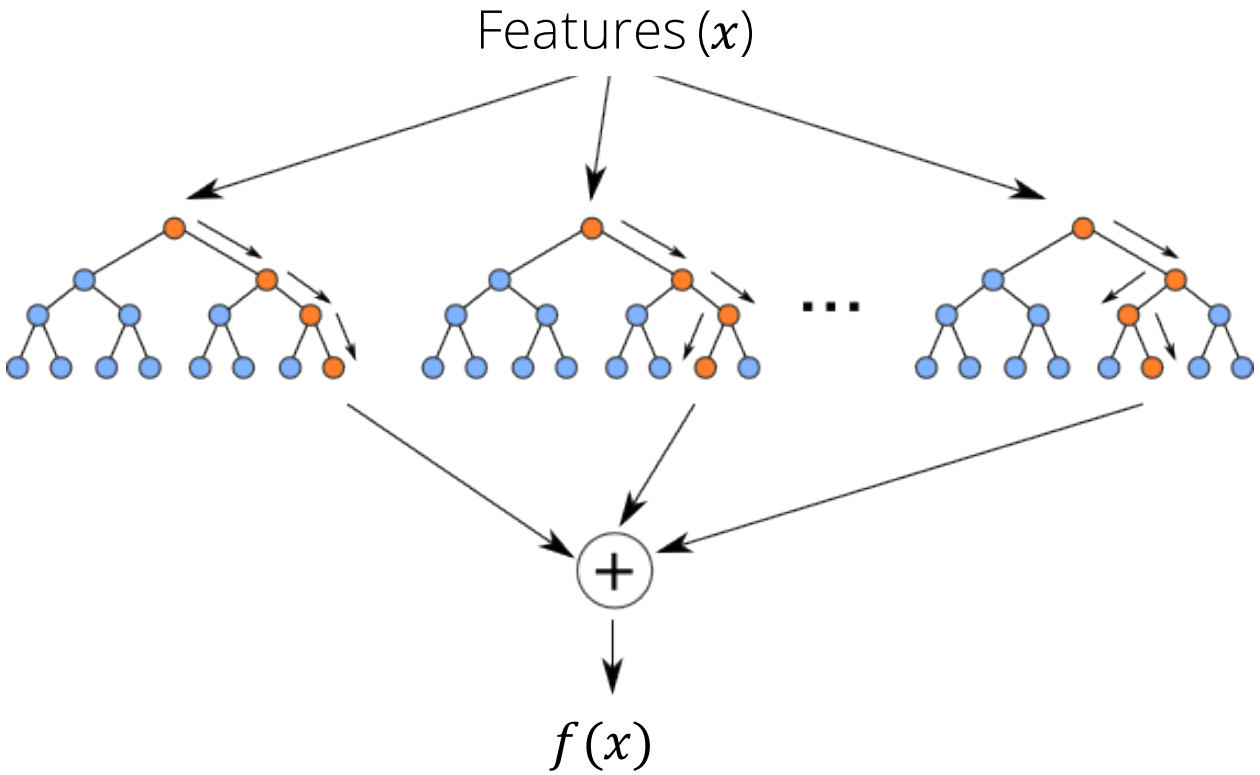
Posterior Distribution

- After Seeing Data, Some Functions Become More Important than Others

Random Forests (RFs)

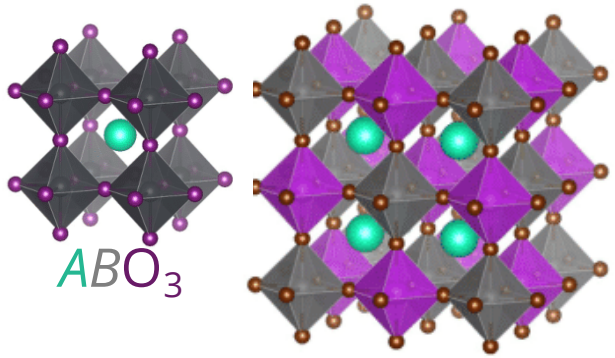
- Combines the Predictions of Many (Ensemble) of Decision Trees, Each Trained on a Subset of Data and Features

- Variability of Predictions Across Trees Estimates Prediction Uncertainty



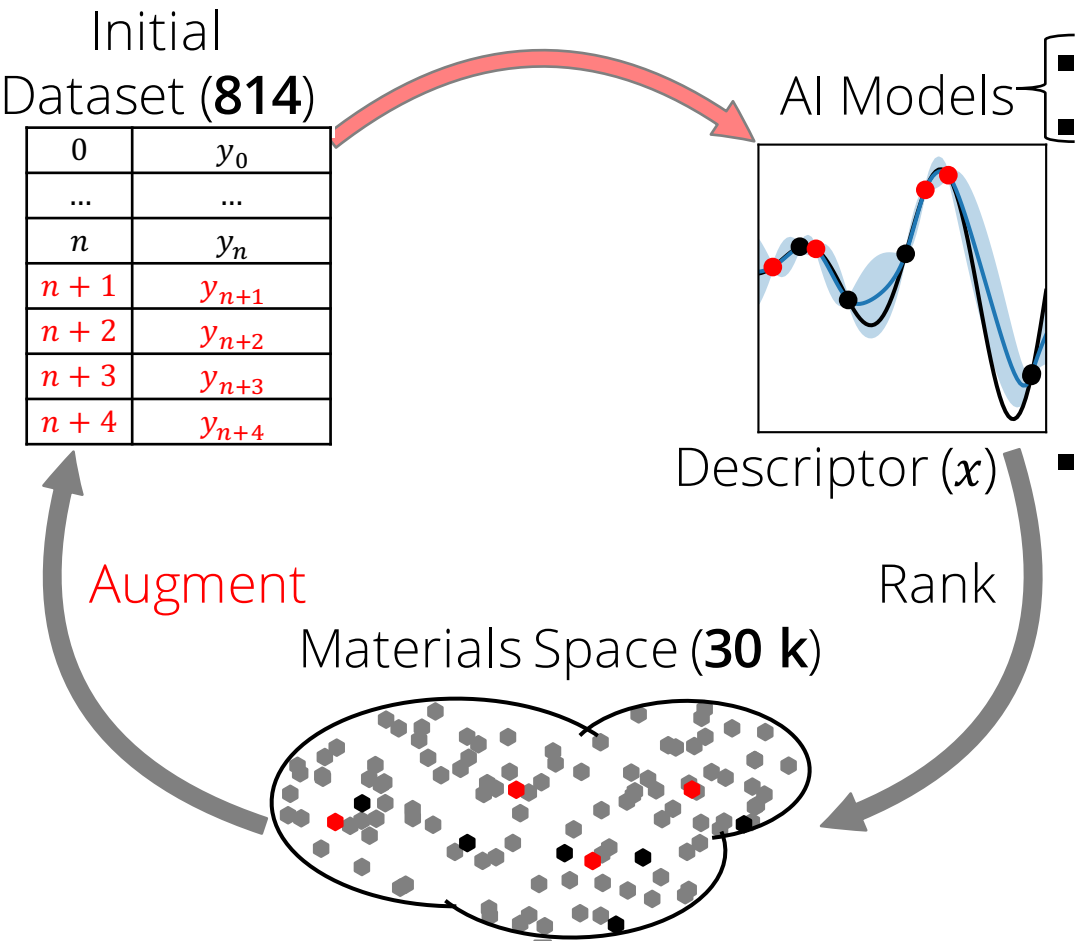
Workflow 1: Identifying Perovskites with High Bulk Modulus

- Target Property (y):
  Bulk Modulus (B_0)



ABO_3 Cubic

$A_2BB'O_6$



- Gaussian Processes (GP)
- Random Forests (RF)

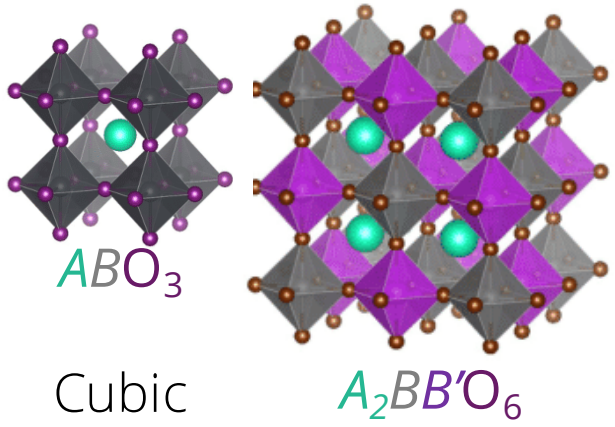
- Elemental Properties of ***A*** and ***B*** (e.g. Radii, Electronegativity, Ionization Potential)

FHI-aims DFT-PBEsol
The ab initio materials simulation package

M. Boley *et al.* Modelling Simul. Mater. Sci. Eng. 32, 063301 (2024)

Workflow 1: Identifying Perovskites with High Bulk Modulus

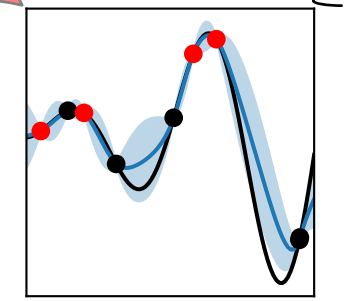
- Target Property (y):
  Bulk Modulus (B_0)



Initial Dataset (**814**)

0	y_0
...	...
n	y_n
$n+1$	y_{n+1}
$n+2$	y_{n+2}
$n+3$	y_{n+3}
$n+4$	y_{n+4}

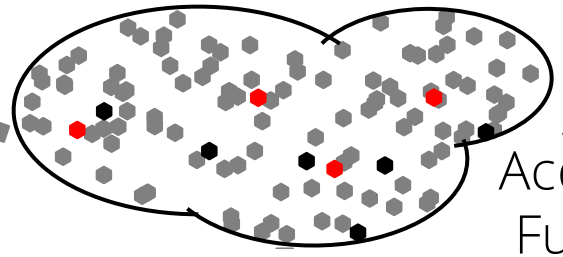
AI Models
- Gaussian Processes (GP)
- Random Forests (RF)



Descriptor (x)
- Elemental Properties of ***A*** and ***B*** (e.g. Radii, Electronegativity, Ionization Potential)

Rank

Materials Space (**30 k**)



Augment

Acquisition Functions
- Pure Exploitation (XT)
- Expected Improvement (EI)

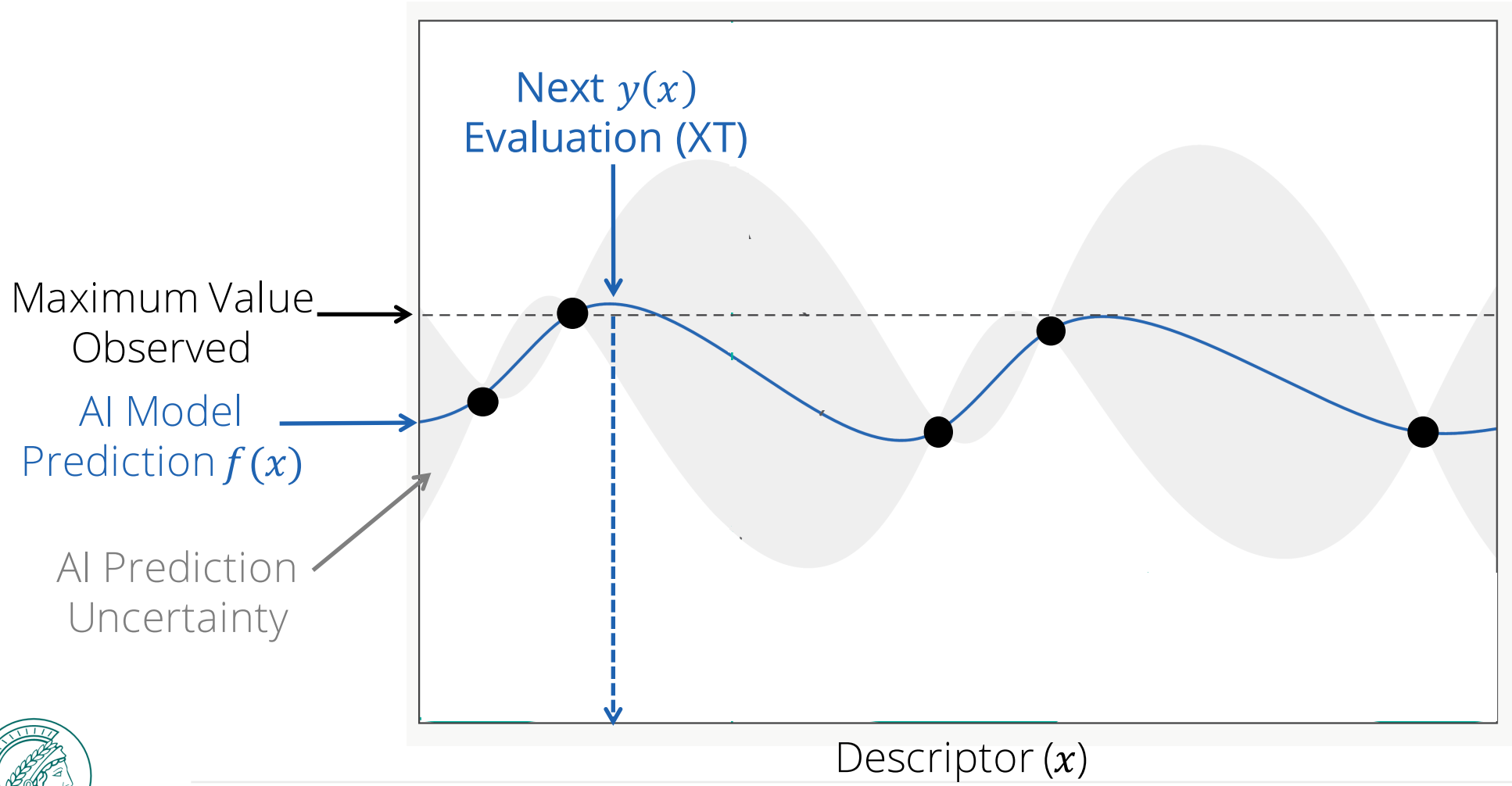
FHI-aims The ab initio materials simulation package DFT-PBEsol

M. Boley *et al.* Modelling Simul. Mater. Sci. Eng. 32, 063301 (2024)

Workflow 1: Identifying Perovskites with High Bulk Modulus

- **Reward:** Maximum B_0 Observed So Far

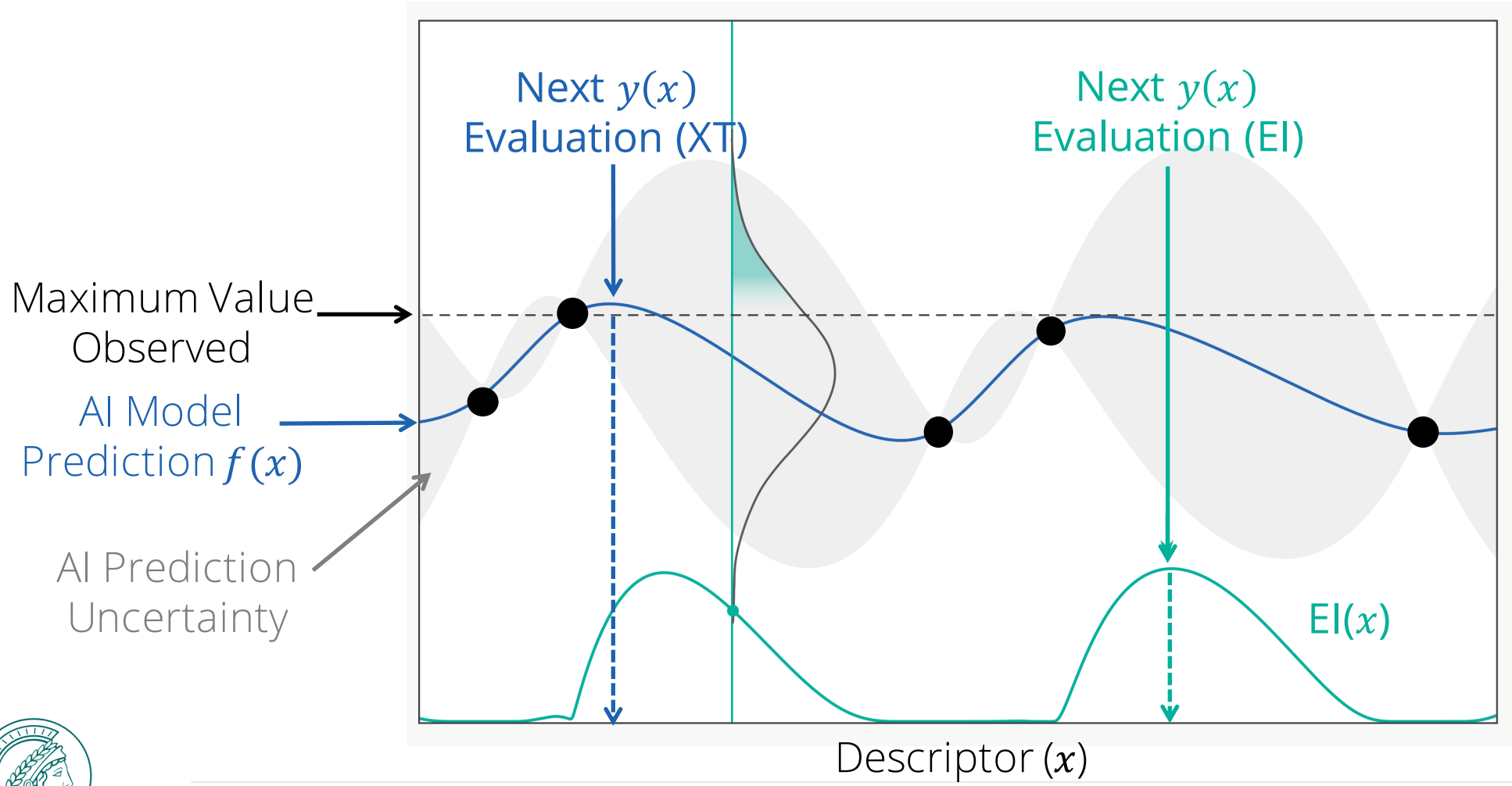
- Pure Exploitation (XT)



- XT Only Uses Predictions

Workflow 1: Identifying Perovskites with High Bulk Modulus

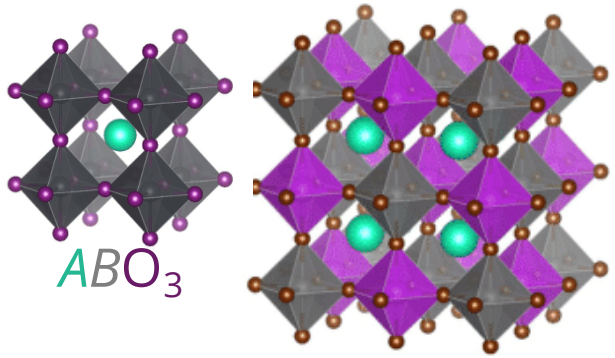
- **Reward:** Maximum B_0 Observed So Far
  - Pure Exploitation (XT)
  - Expected Improvement (EI)



- XT Only Uses Predictions
- EI Uses Predictions and Uncertainties
- EI Balances Model Exploitation and Exploration of Materials Space

Workflow 1: Identifying Perovskites with High Bulk Modulus

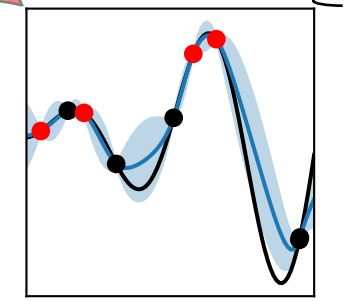
- Target Property (y):
  Bulk Modulus (B_0)



Initial Dataset (**814**)

0	y_0
...	...
n	y_n
$n+1$	y_{n+1}
$n+2$	y_{n+2}
$n+3$	y_{n+3}
$n+4$	y_{n+4}

AI Models
- Gaussian Processes (GP)
- Random Forests (RF)



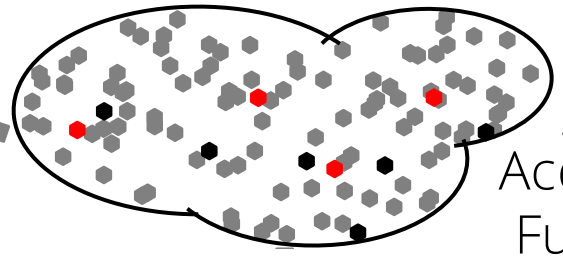
Descriptor (x)

Four Methods
- GP-XT
- GP-EI
- RF-XT
- RF-EI

Augment

Rank

Materials Space (**30 k**)

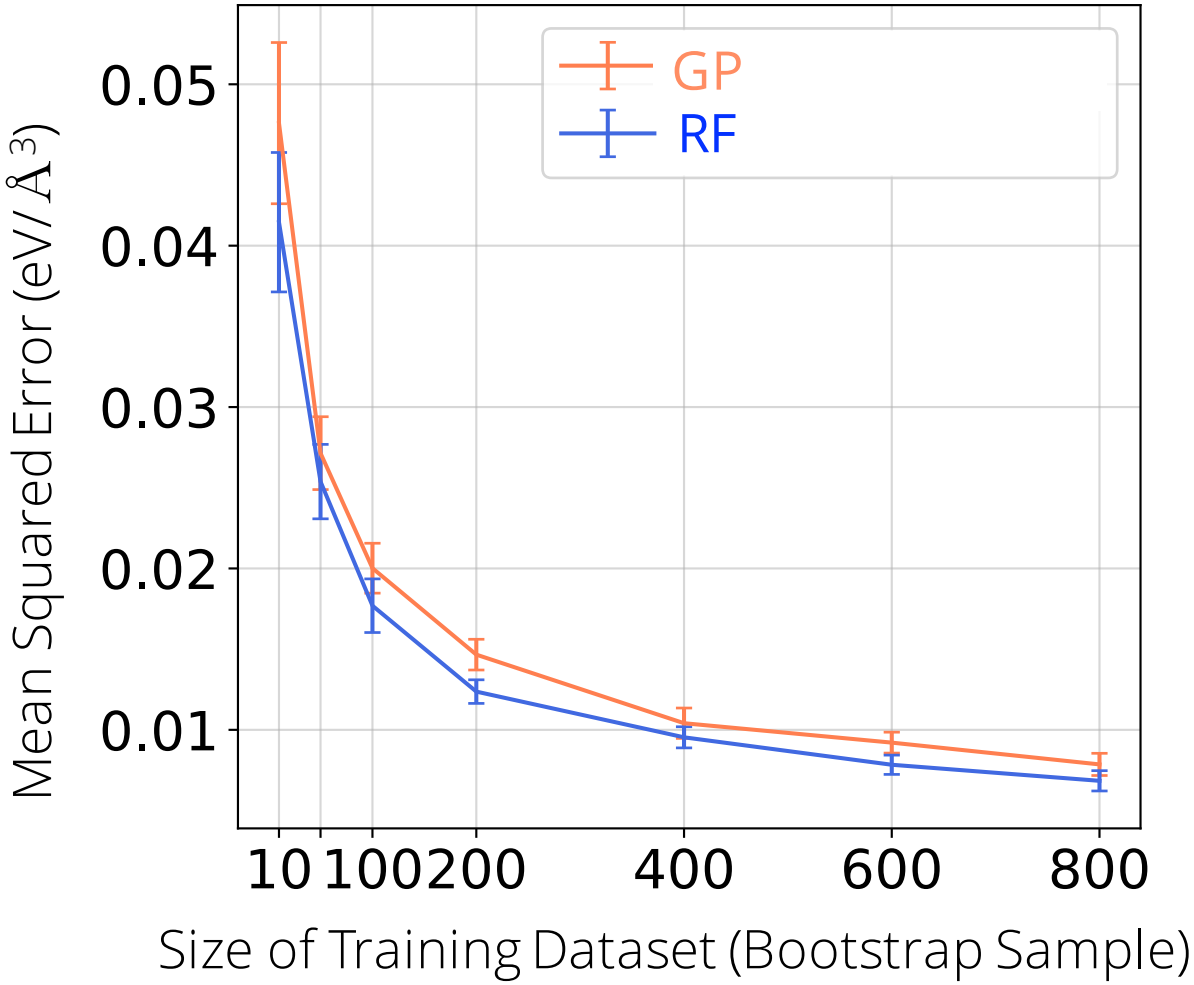


Acquisition Functions
- Pure Exploitation (XT)
- Expected Improvement (EI)

FHI-aims DFT-PBEsol
The ab initio materials simulation package

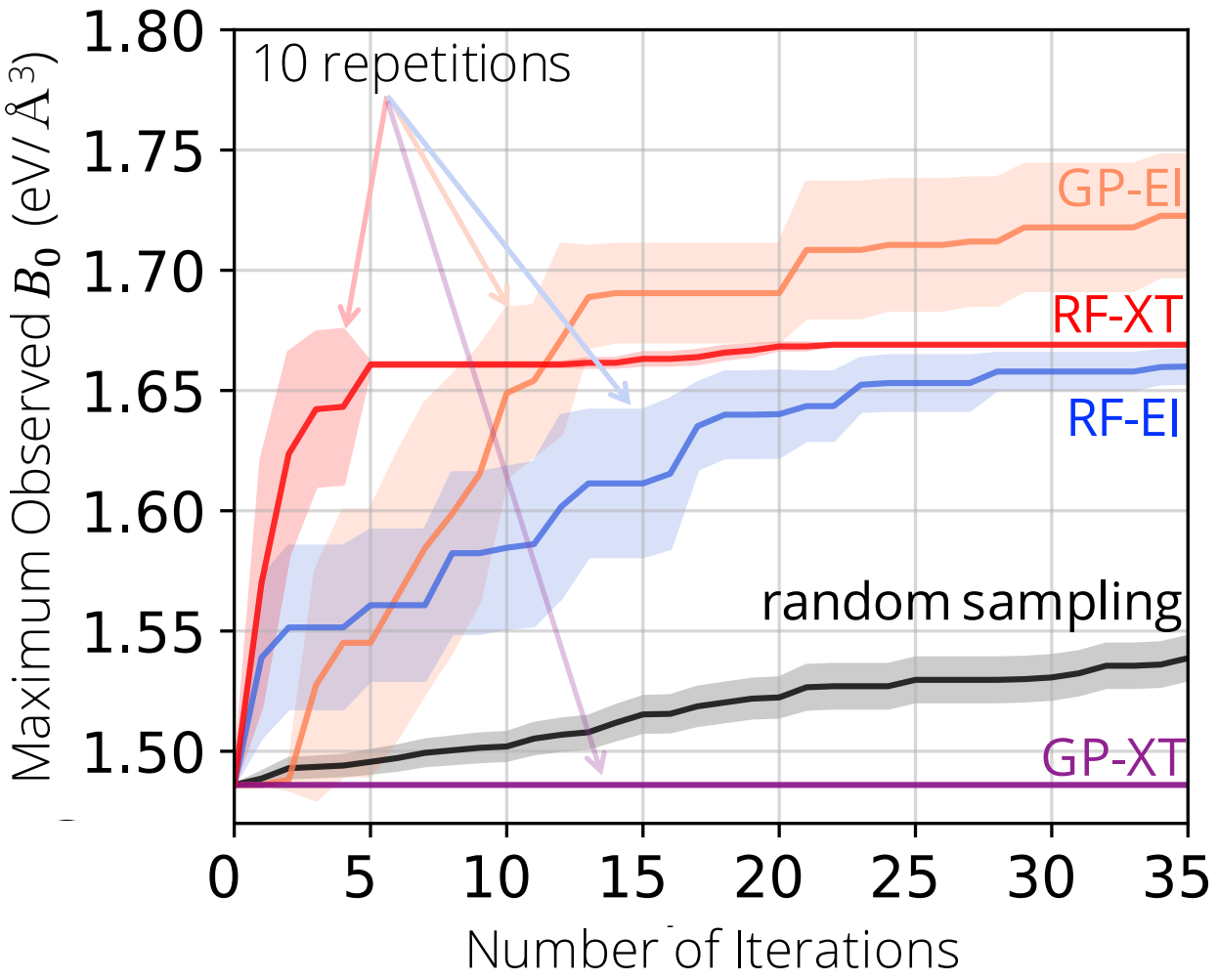
M. Boley *et al.* Modelling Simul. Mater. Sci. Eng. 32, 063301 (2024)

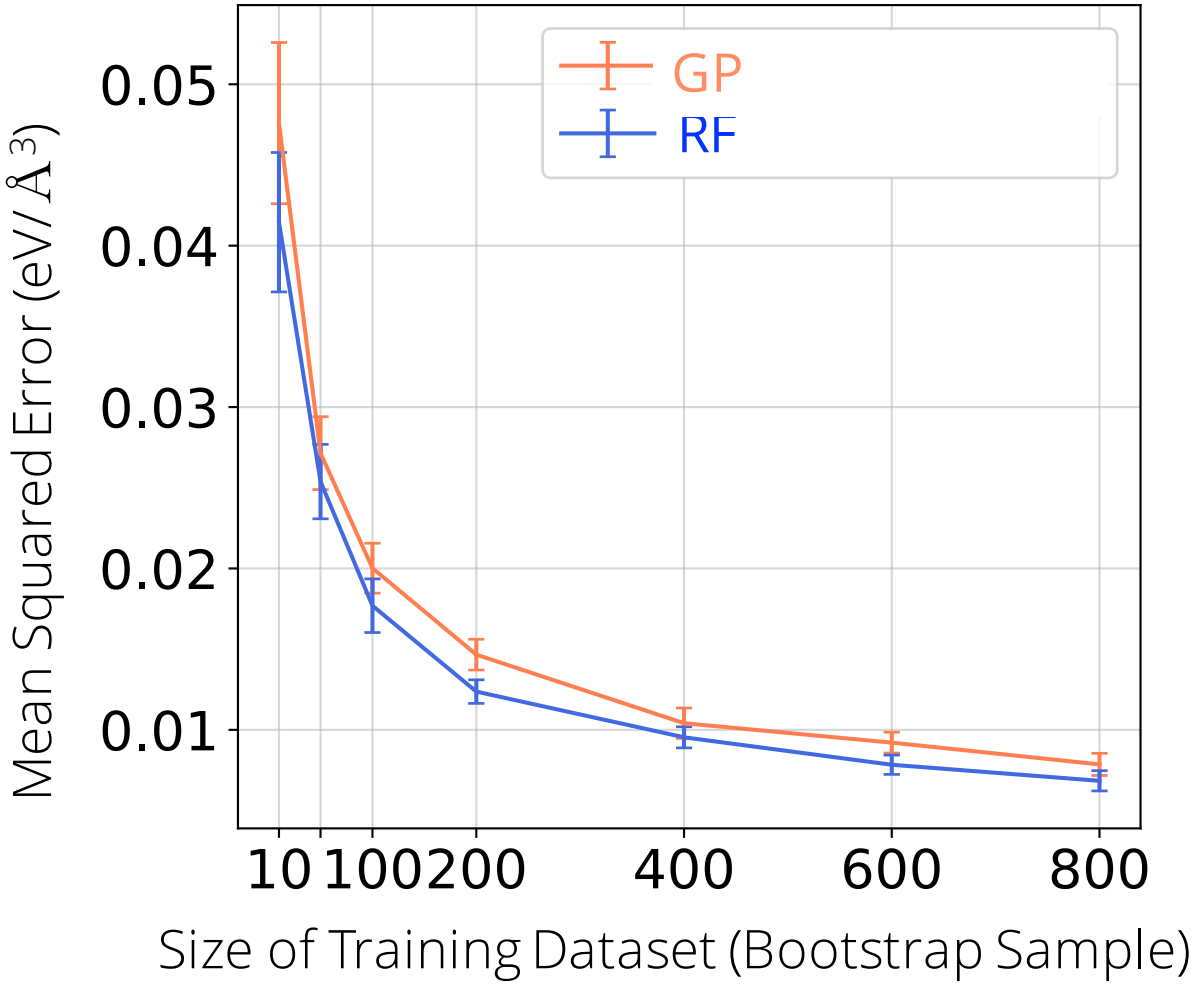
Workflow 1: Identifying Perovskites with High Bulk Modulus



M. Boley *et al.* Modelling Simul. Mater. Sci. Eng. 32, 063301 (2024)

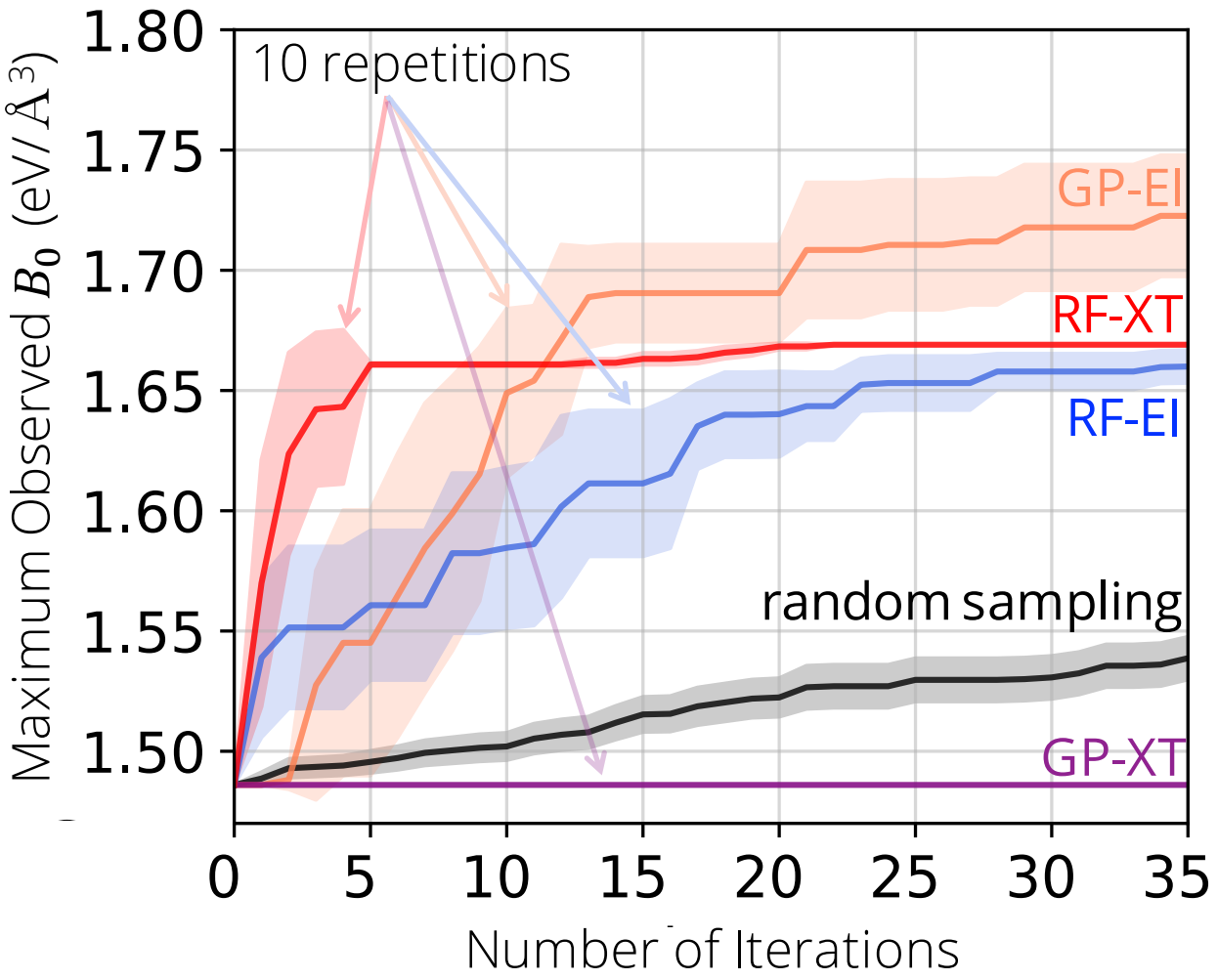
Workflow 1: Identifying Perovskites with High Bulk Modulus





M. Boley *et al.* Modelling Simul. Mater. Sci. Eng. 32, 063301 (2024)

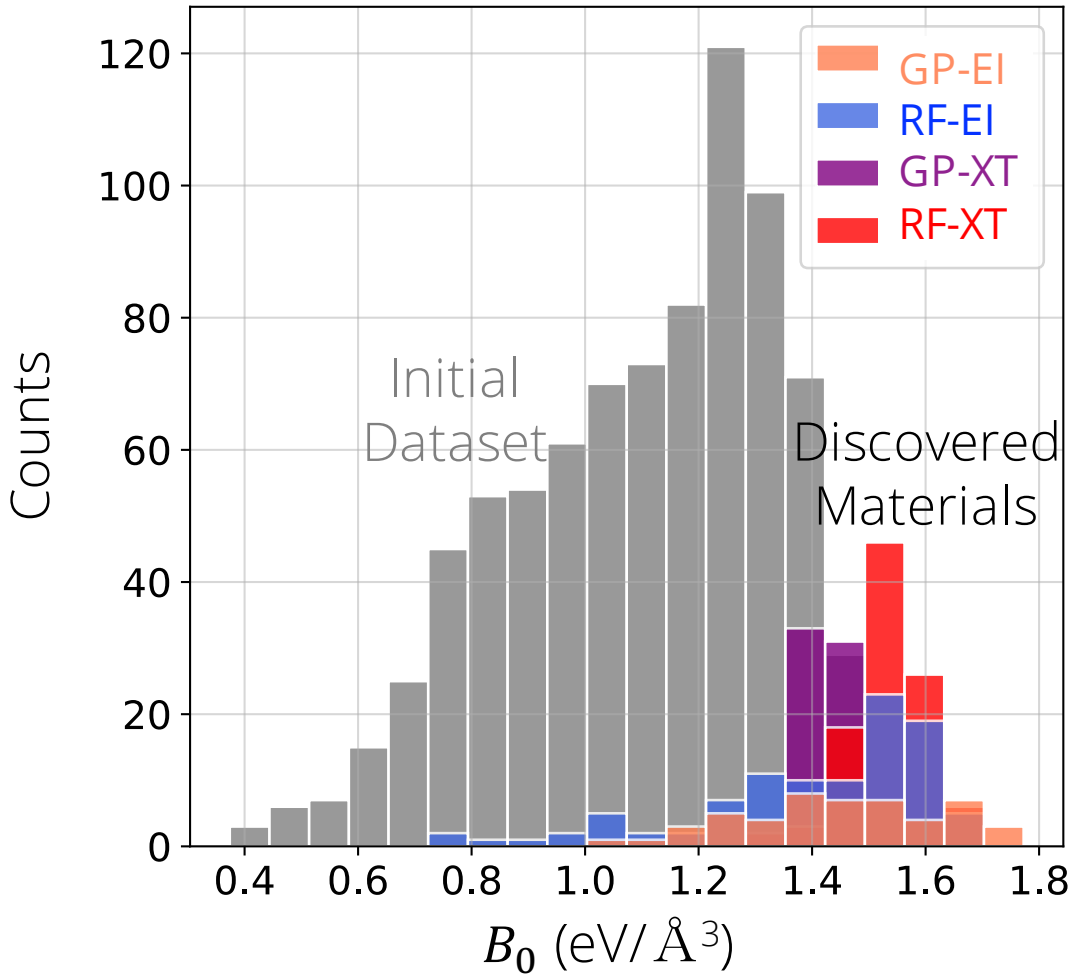
Workflow 1: Identifying Perovskites with High Bulk Modulus

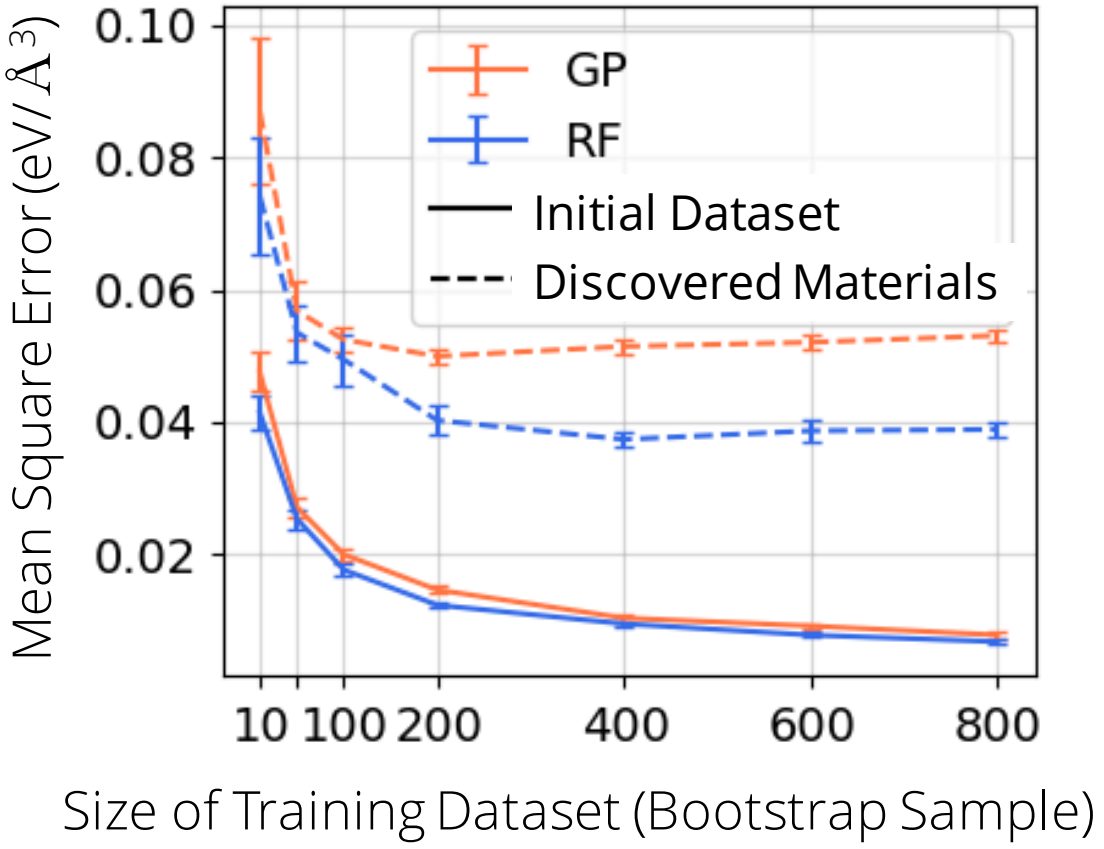


- Efficiency of AI-Guided Workflow Depends On the Combination of AI Model and Acquisition Function
- Exploration (e.g., Guided By Uncertainty Estimates) Might Be Crucial
- Importance of Reliable Uncertainty Quantification (Overconfidence Issue)

M. Boley *et al.* Modelling Simul. Mater. Sci. Eng. 32, 063301 (2024)

Workflow 1: Identifying Perovskites with High Bulk Modulus





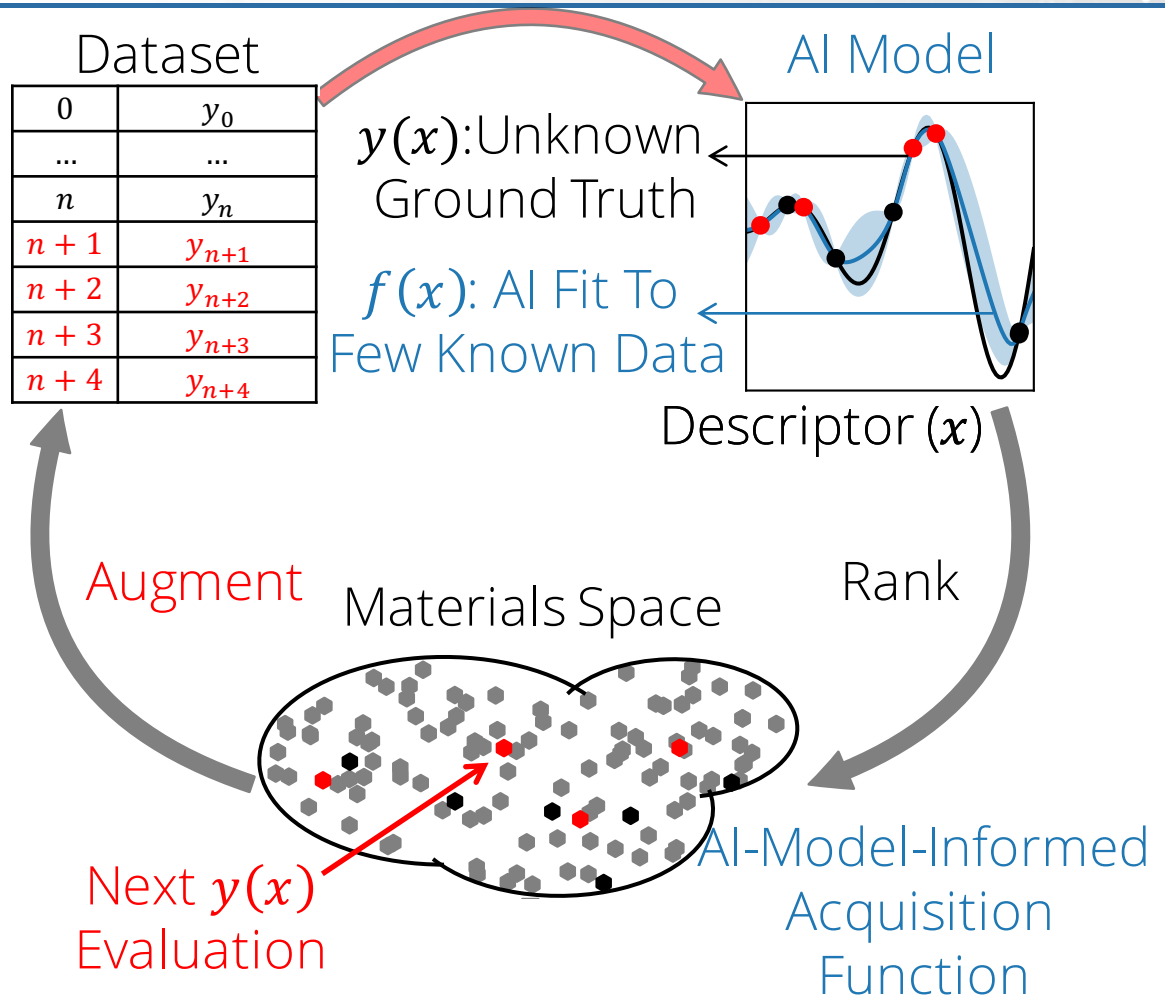
M. Boley *et al.* Modelling Simul. Mater. Sci. Eng. 32, 063301 (2024)

From Prediction to Action: AI-Guided Workflows for the Discovery of Materials with Optimal Properties

- Iterative Modelling and Data Acquisition Approach to Identify Materials with Desired Property Values (e.g., High)

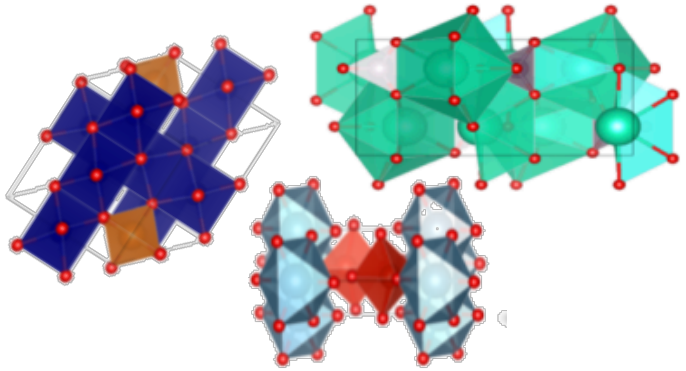
1. Impact of Choices of AI Model and Acquisition Function?

2. SISSO-Guided Workflow?



Workflow 2: Identifying Acid-Stable Oxides for Water Splitting

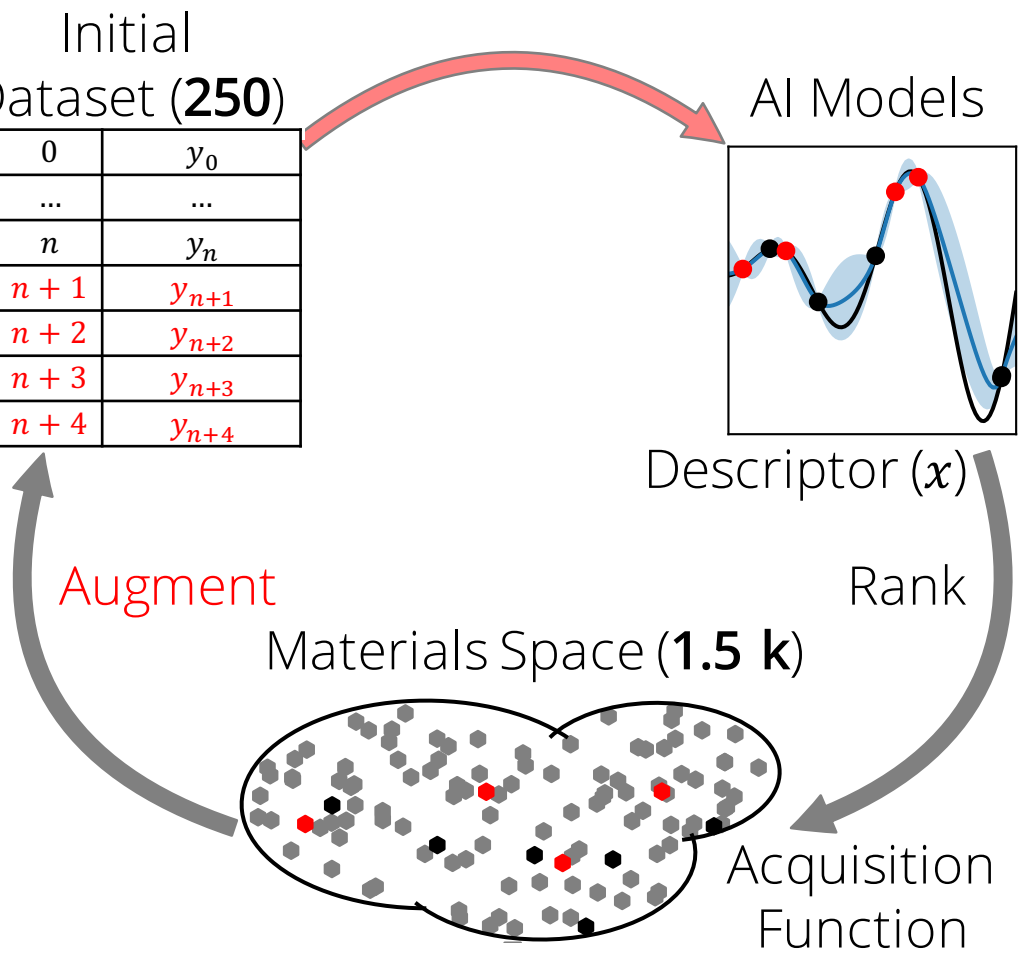
- **Target Property** (y): Decomposition Energy at OER Conditions pH=0, U=1.23 V ($\Delta G_{\mathbf{Pbx}}^{\mathbf{OER}}$)



Binary and Ternary Oxides

FHI-aims DFT-HSE06

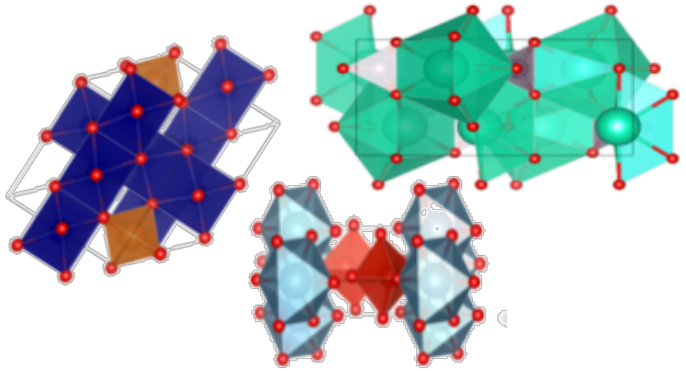
The ab initio materials simulation package



A. S. Nair *et al.* npj Comput. Mat. 11, 150 (2025)

Workflow 2: Identifying Acid-Stable Oxides for Water Splitting

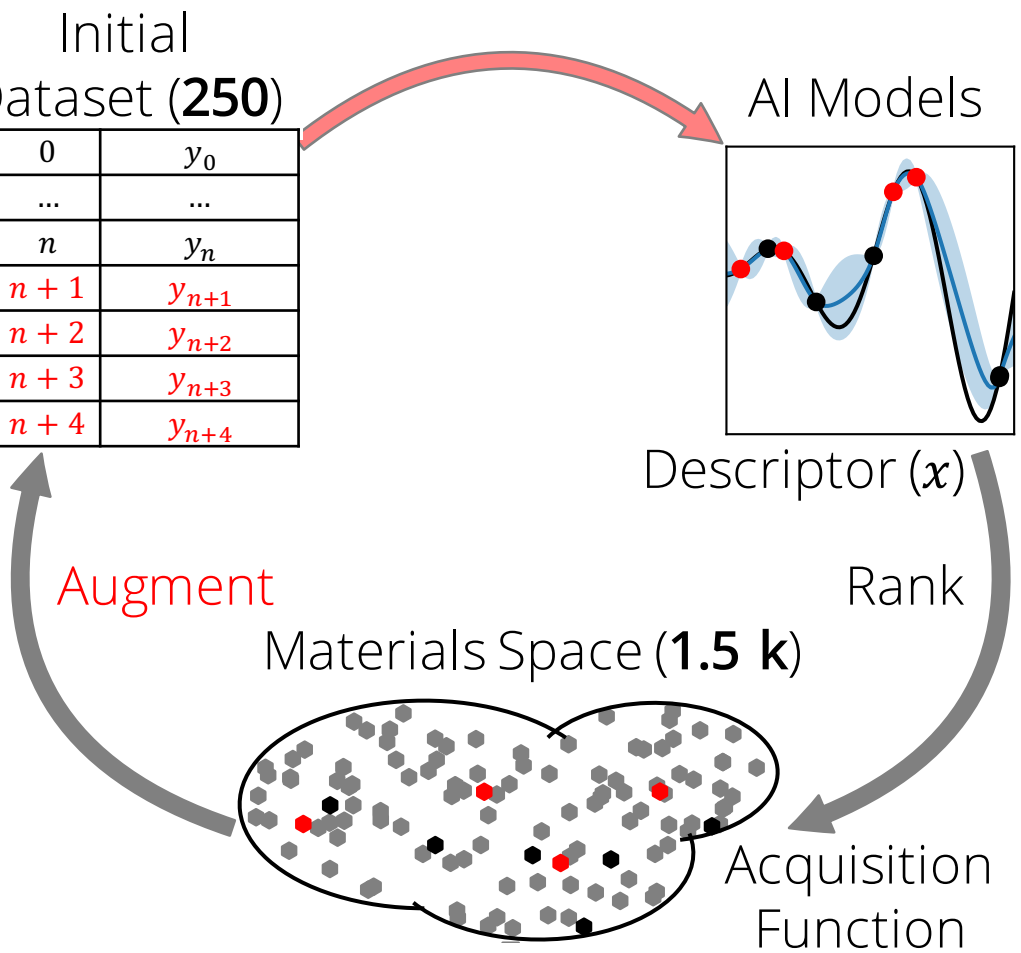
- **Target Property** (y): Decomposition Energy at OER Conditions pH=0, U=1.23 V ($\Delta G_{\mathbf{Pbx}}^{\mathbf{OER}}$)



Binary and Ternary Oxides

FHI-aims DFT-HSE06

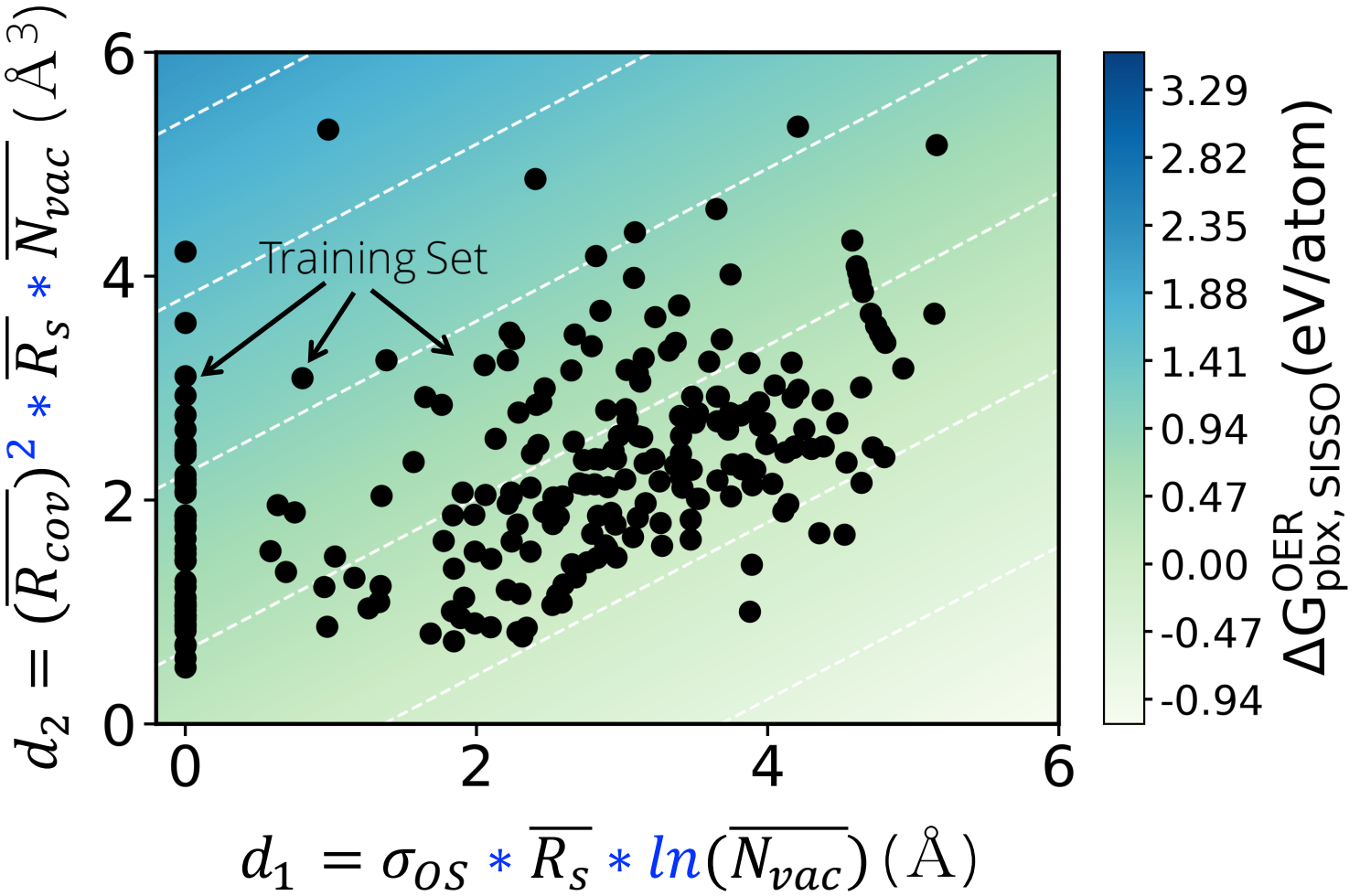
The ab initio materials simulation package



- SISSO Approach
- Primary Features: 14 Physical Descriptive Parameters Depending on the Composition (e.g. Atomic Radii, Oxidation States)

A. S. Nair *et al.* npj Comput. Mat. 11, 150 (2025)

Workflow 2: Identifying Acid-Stable Oxides for Water Splitting



Key Parameters in the Descriptor

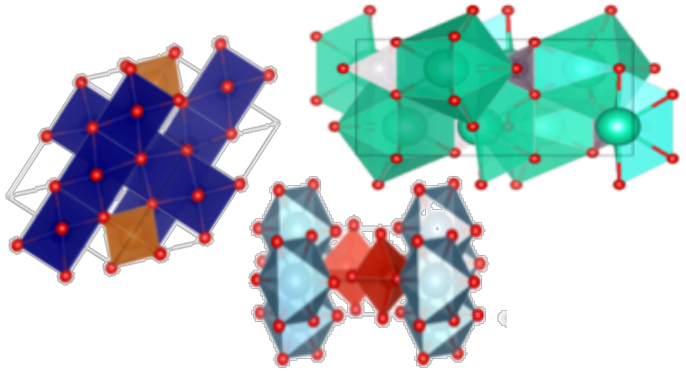
- σ_{OS}: Std. Dev. of Oxidation States
- $\overline{R_s}$: *s* Orbital Radius
- $\overline{N_{vac}}$: Number of Vacant Orbitals
- $\overline{R_{cov}}$: Covalent Radius

Average Across Cations

A. S. Nair *et al.* npj Comput. Mat. 11, 150 (2025)

Workflow 2: Identifying Acid-Stable Oxides for Water Splitting

- **Target Property** (y): Decomposition Energy at OER Conditions pH=0, U=1.23 V ($\Delta G_{\mathrm{Pbx}}^{\mathrm{OER}}$)



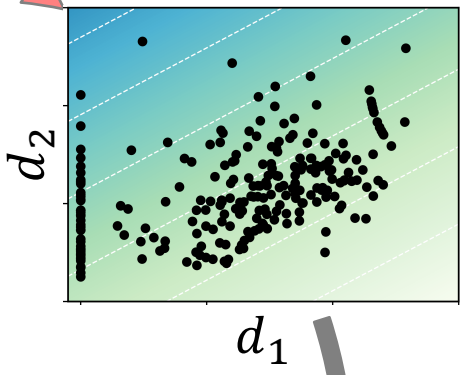
Binary and Ternary Oxides

FHI-aims DFT-HSE06
The ab initio materials simulation package

Initial Dataset (**250**)

0	y_0
...	...
n	y_n
$n+1$	y_{n+1}
$n+2$	y_{n+2}
$n+3$	y_{n+3}
$n+4$	y_{n+4}

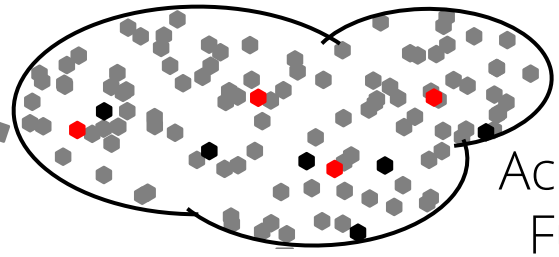
AI Models



Augment

Rank

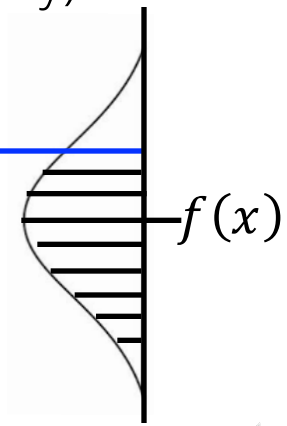
Materials Space (**1.5 k**)



Acquisition Function

- SISSO Approach
  - **Reward:** No. of Stable Materials
  - Probability of Feasibility POF (Prediction and Uncertainty)

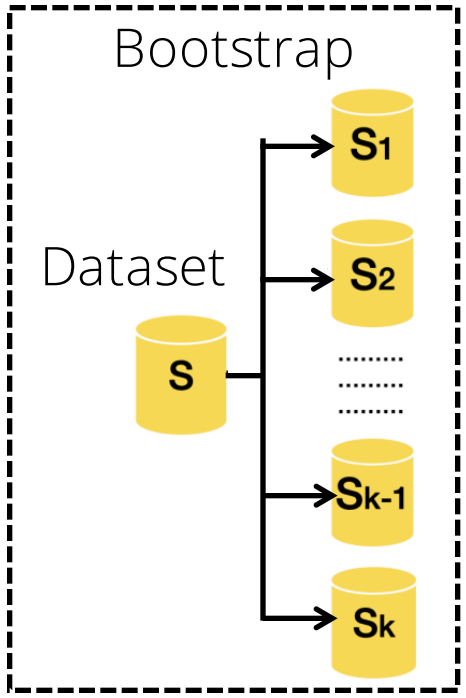
$\Delta G_{\mathrm{Pbx}}^{\mathrm{OER}} < 0.10$ eV/atom



A. S. Nair *et al.* npj Comput. Mat. 11, 150 (2025)

Workflow 2: Identifying Acid-Stable Oxides for Water Splitting

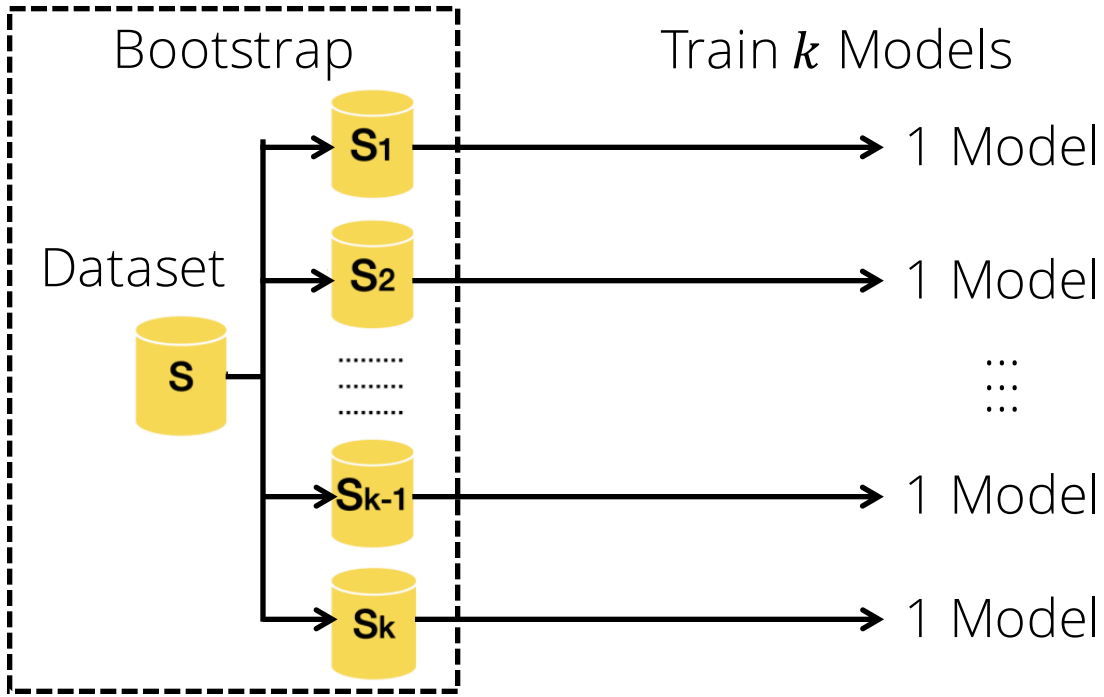
Creating Ensembles of SISSO Models: 1. (Standard) Bagging



A. S. Nair *et al.* npj Comput. Mat. 11, 150 (2025)

Workflow 2: Identifying Acid-Stable Oxides for Water Splitting

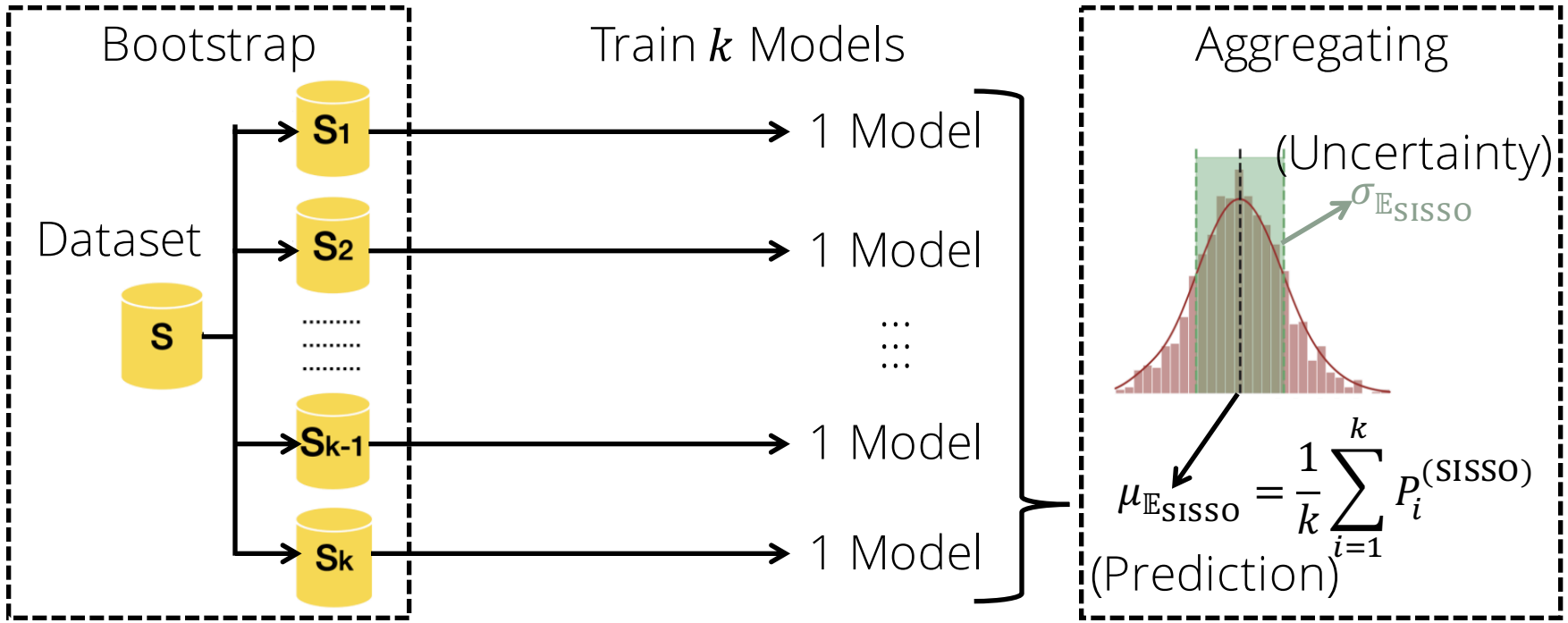
Creating Ensembles of SISSO Models: 1. (Standard) Bagging



A. S. Nair *et al.* npj Comput. Mat. 11, 150 (2025)

Workflow 2: Identifying Acid-Stable Oxides for Water Splitting

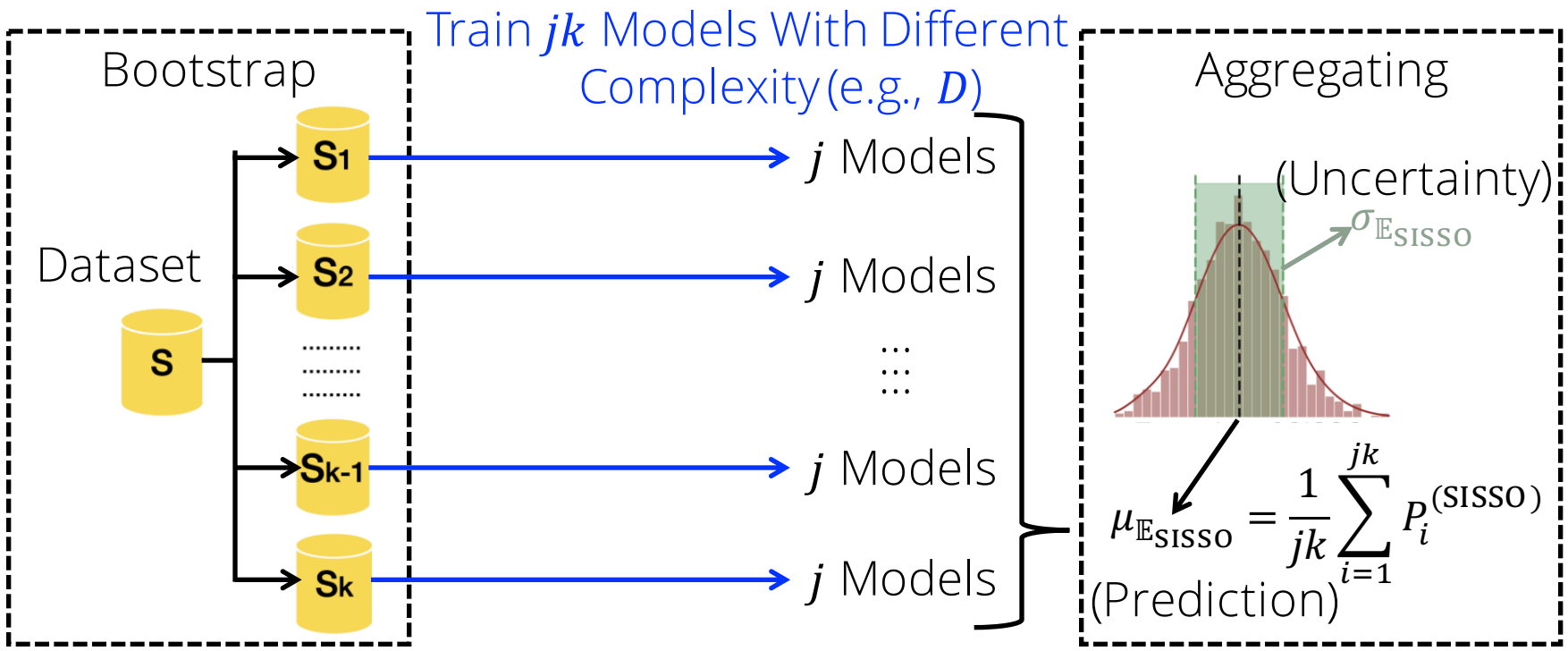
Creating Ensembles of SISSO Models: 1. (Standard) Bagging



A. S. Nair *et al.* npj Comput. Mat. 11, 150 (2025)

Workflow 2: Identifying Acid-Stable Oxides for Water Splitting

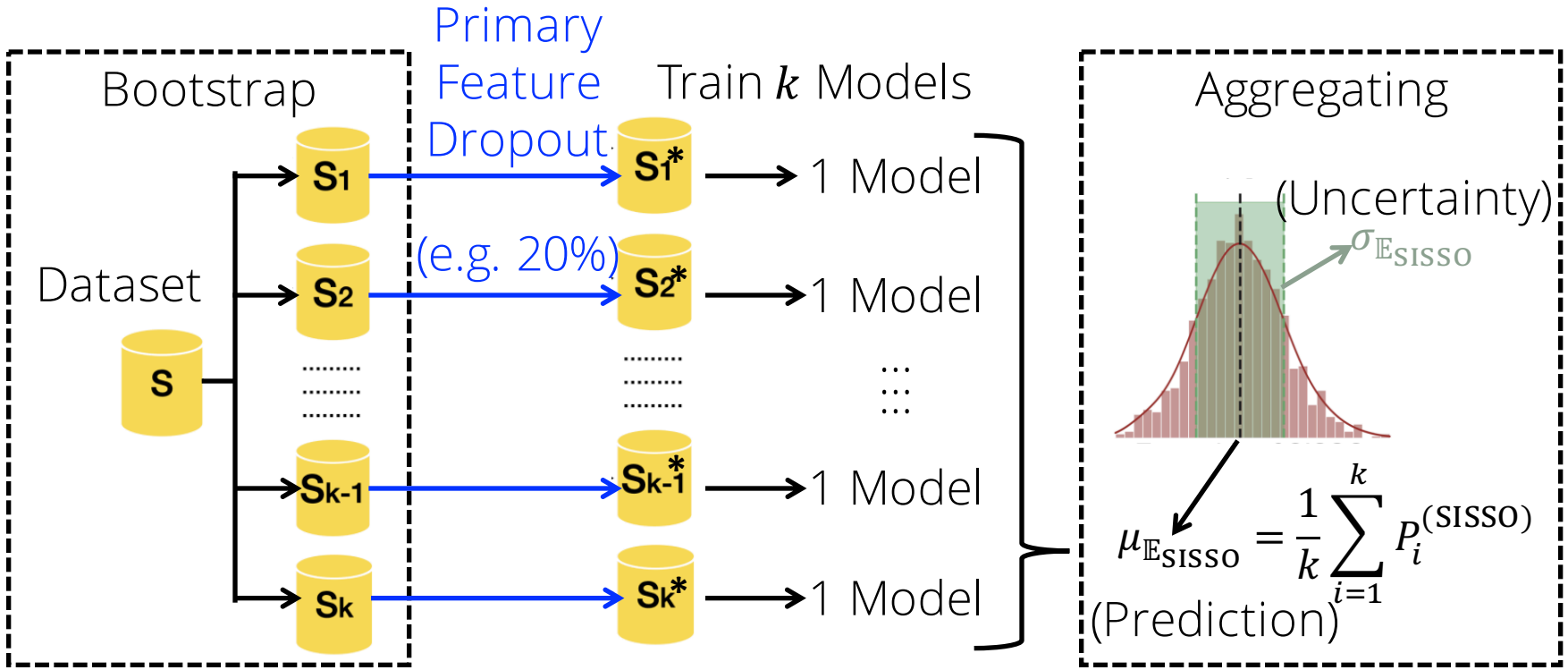
Creating Ensembles of SISSO Models: 2. Bagging with Multiple Model Complexities



A. S. Nair *et al.* npj Comput. Mat. 11, 150 (2025)

Workflow 2: Identifying Acid-Stable Oxides for Water Splitting

Creating Ensembles of SISSO Models: 3. Bagging with Monte-Carlo Dropout of Primary Features



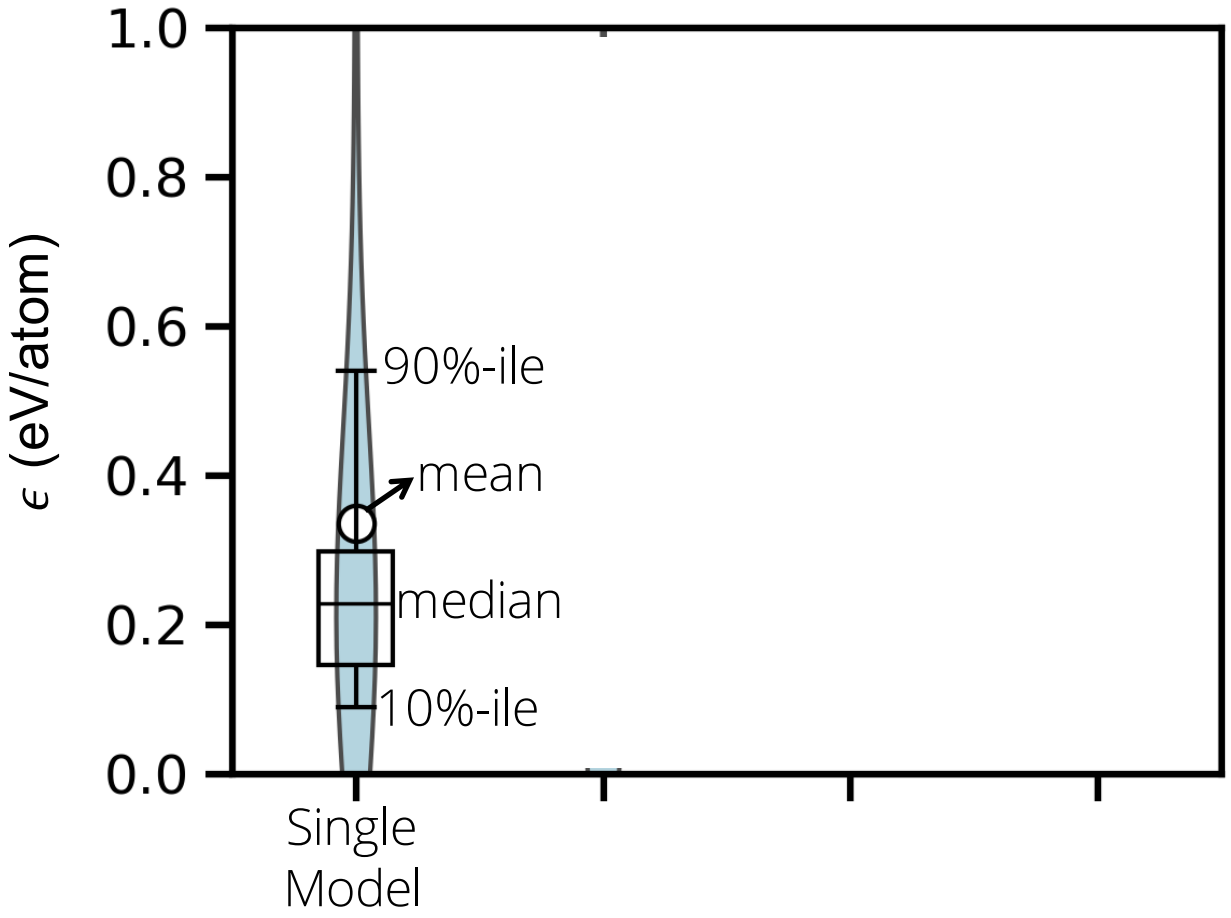
A. S. Nair *et al.* npj Comput. Mat. 11, 150 (2025)

Workflow 2: Identifying Acid-Stable Oxides for Water Splitting

Assessing the Quality of Mean Predictions of SISSO-Model Ensembles

Absolute Prediction Error
(20% Left-Out Materials)

$$\epsilon = |\Delta G^{\mathrm{OER}}_{\mathrm{Pbx}} - \Delta G^{\mathrm{OER}}_{\mathrm{Pbx},\mathbb{E}_{\mathrm{SISSO}}}|$$



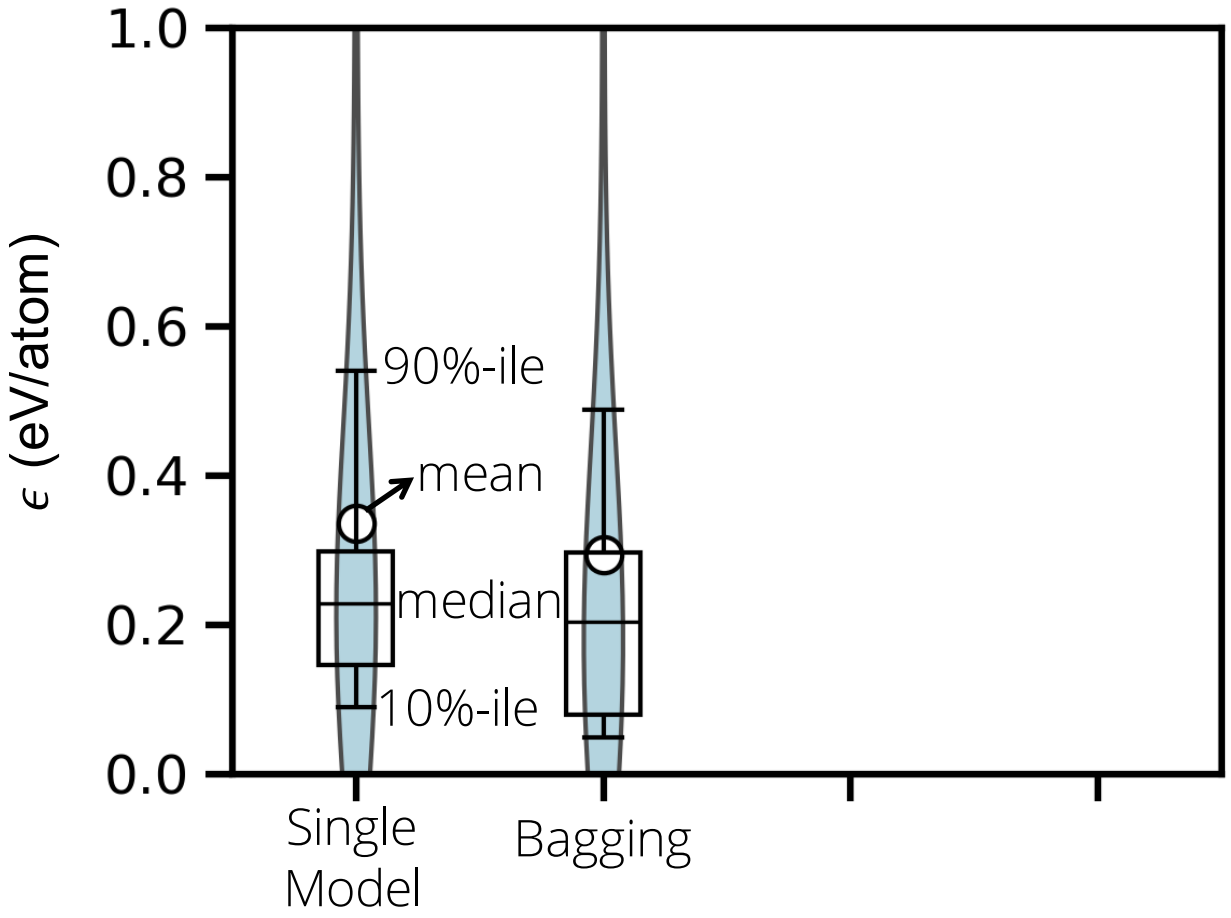
A. S. Nair *et al.* npj Comput. Mat. 11, 150 (2025)

Workflow 2: Identifying Acid-Stable Oxides for Water Splitting

Assessing the Quality of Mean Predictions of SISSO-Model Ensembles

Absolute Prediction Error
(20% Left-Out Materials)

$$\epsilon = |\Delta G^{\mathrm{OER}}_{\mathrm{Pbx}} - \Delta G^{\mathrm{OER}}_{\mathrm{Pbx},\mathbb{E}_{\mathrm{SISSO}}}|$$



A. S. Nair *et al.* npj Comput. Mat. 11, 150 (2025)

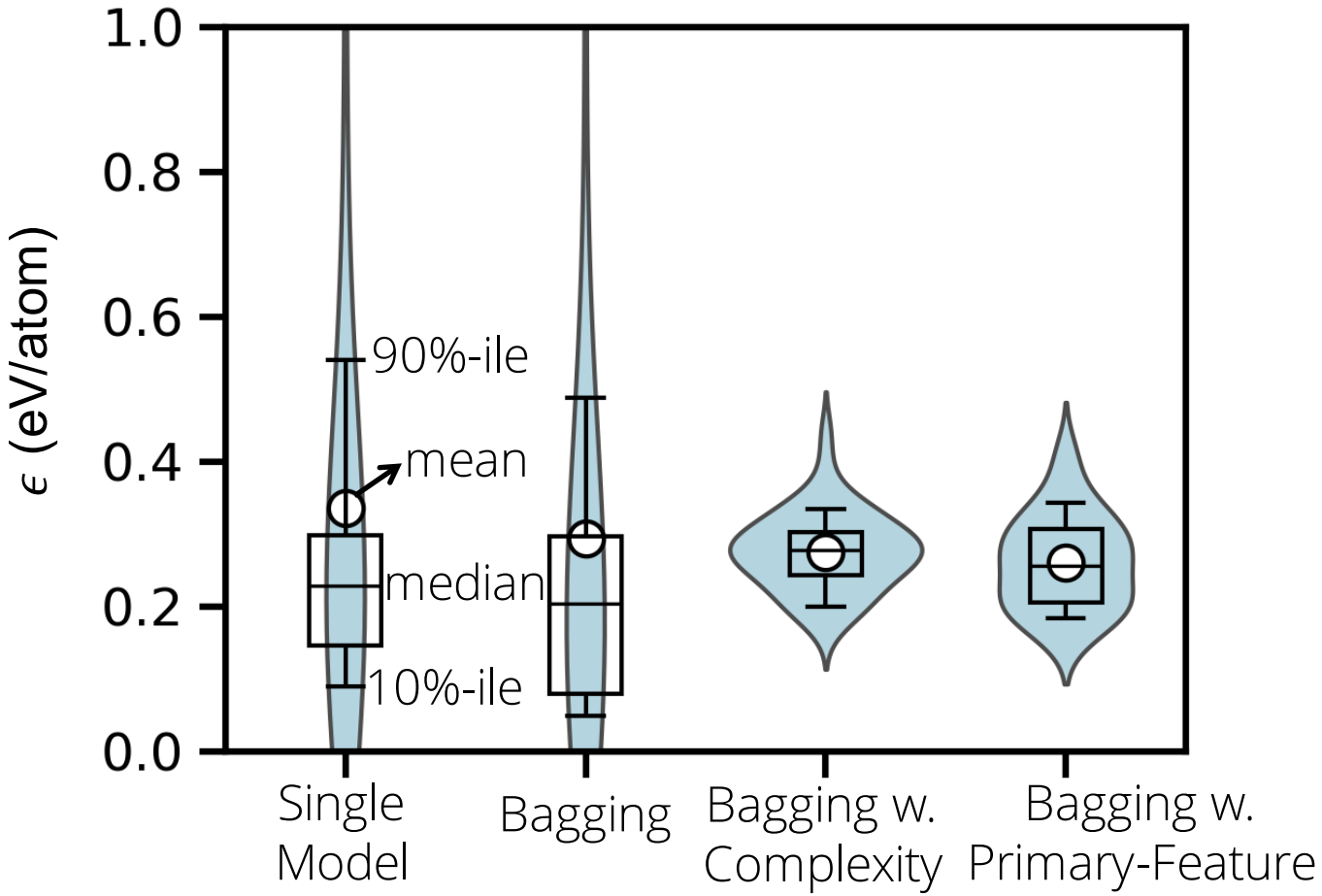
Workflow 2: Identifying Acid-Stable Oxides for Water Splitting

Assessing the Quality of Mean Predictions of SISSO-Model Ensembles

Absolute Prediction Error (20% Left-Out Materials)

$$\epsilon = |\Delta G^{\mathrm{OER}}_{\mathrm{Pbx}} - \Delta G^{\mathrm{OER}}_{\mathrm{Pbx},\mathbb{E}_{\mathrm{SISSO}}}|$$

- Proposed Ensemble Approaches Mitigate High Errors



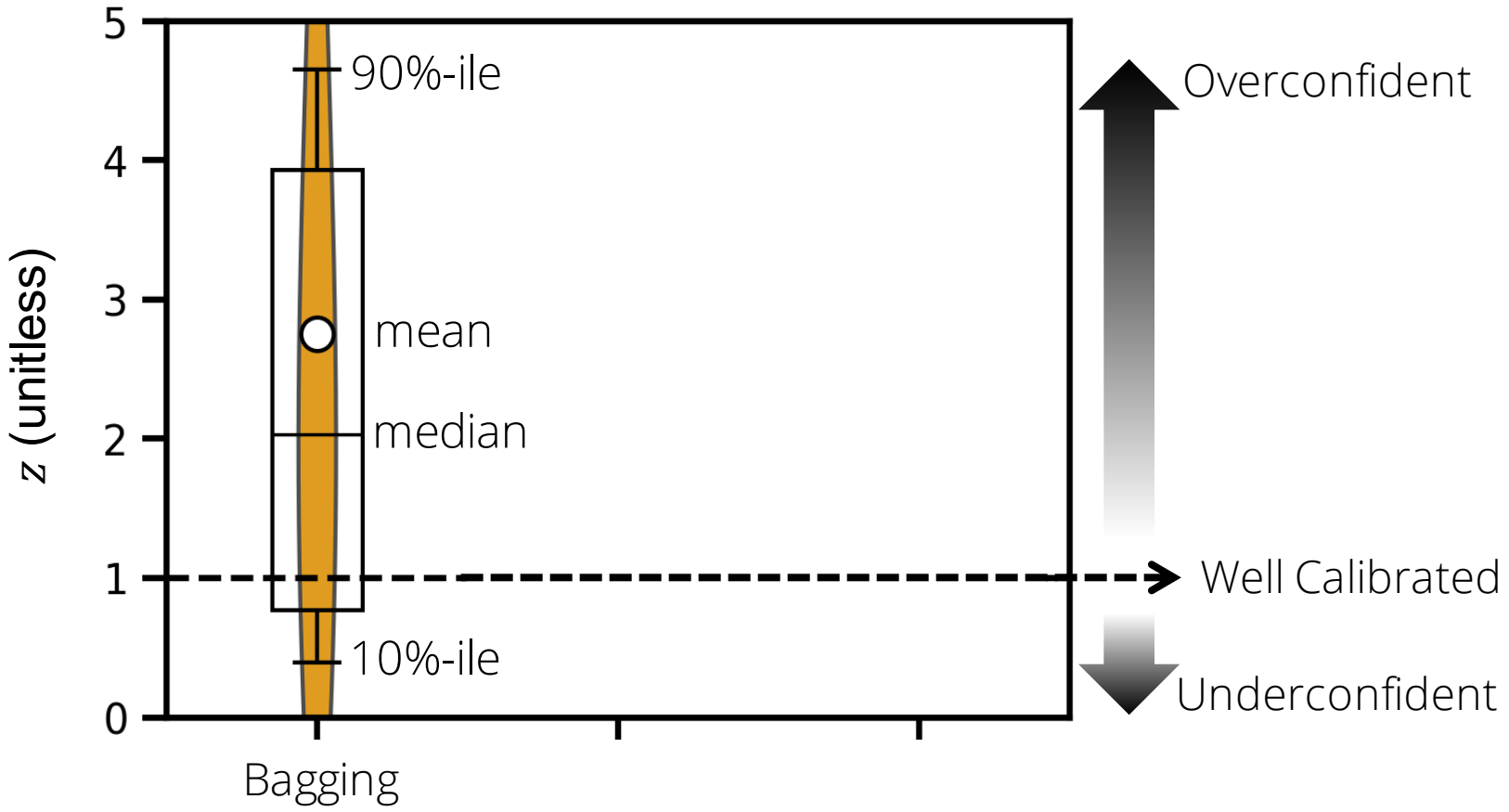
A. S. Nair *et al.* npj Comput. Mat. 11, 150 (2025)

Workflow 2: Identifying Acid-Stable Oxides for Water Splitting

Assessing the Quality of Uncertainty Estimates of SISSO-Model Ensembles

Miscalibration Score
(20% Left-Out Materials)

$$z = \frac{|\Delta G^{\mathrm{OER}}_{\mathrm{Pbx}} - \Delta G^{\mathrm{OER}}_{\mathrm{Pbx},\mathbb{E}_{\mathrm{SISSO}}}|}{\sigma_{\mathbb{E}_{\mathrm{SISSO}}}}$$



A. S. Nair *et al.* npj Comput. Mat. 11, 150 (2025)

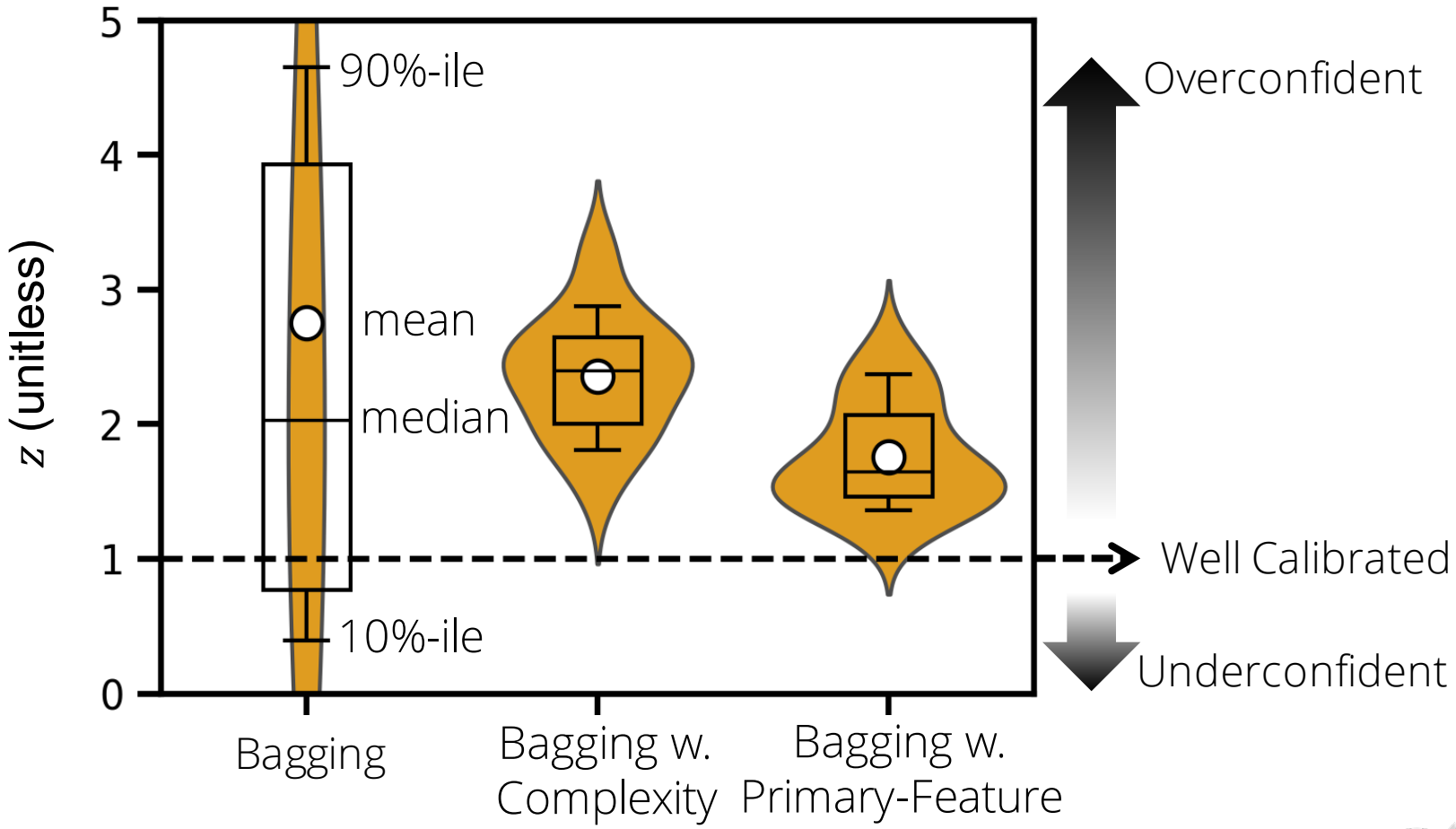
Workflow 2: Identifying Acid-Stable Oxides for Water Splitting

Assessing the Quality of Uncertainty Estimates of SISSO-Model Ensembles

Miscalibration Score
(20% Left-Out Materials)

$$z = \frac{|\Delta G^{\mathrm{OER}}_{\mathrm{Pbx}} - \Delta G^{\mathrm{OER}}_{\mathrm{Pbx},\mathbb{E}_{\mathrm{SISSO}}}|}{\sigma_{\mathbb{E}_{\mathrm{SISSO}}}}$$

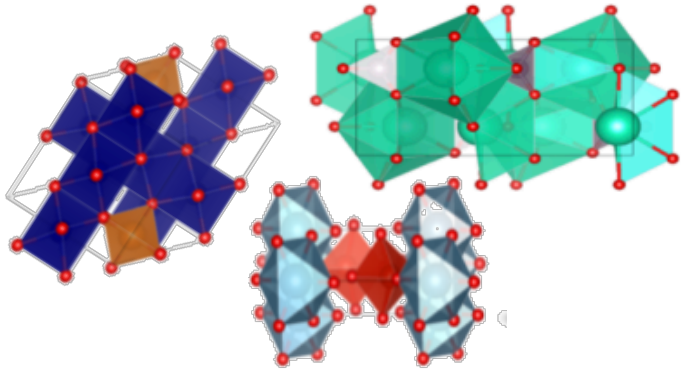
- Use of Primary-Feature Dropout Mitigates Overconfidence Issue



A. S. Nair *et al.* npj Comput. Mat. 11, 150 (2025)

Workflow 2: Identifying Acid-Stable Oxides for Water Splitting

- **Target Property** (y): Decomposition Energy at OER Conditions pH=0, U=1.23 V ($\Delta G_{\mathrm{Pbx}}^{\mathrm{OER}}$)



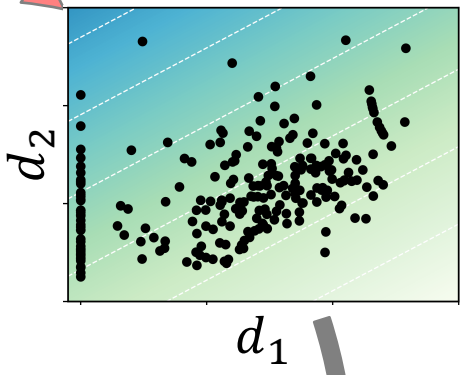
Binary and Ternary Oxides

FHI-aims The ab initio materials simulation package DFT-HSE06

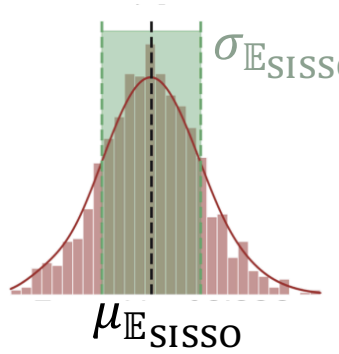
Initial Dataset (**250**)

0	y_0
...	...
n	y_n
$n+1$	y_{n+1}
$n+2$	y_{n+2}
$n+3$	y_{n+3}
$n+4$	y_{n+4}

AI Models

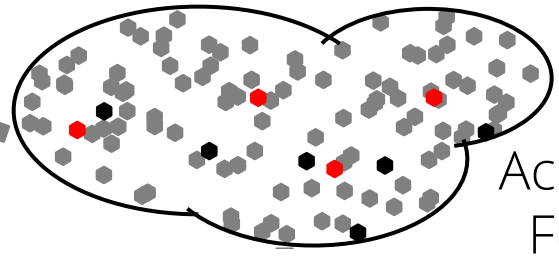


- SISSO Ensembles (Primary-Feature Dropout)



Rank

Materials Space (**1.5 k**)



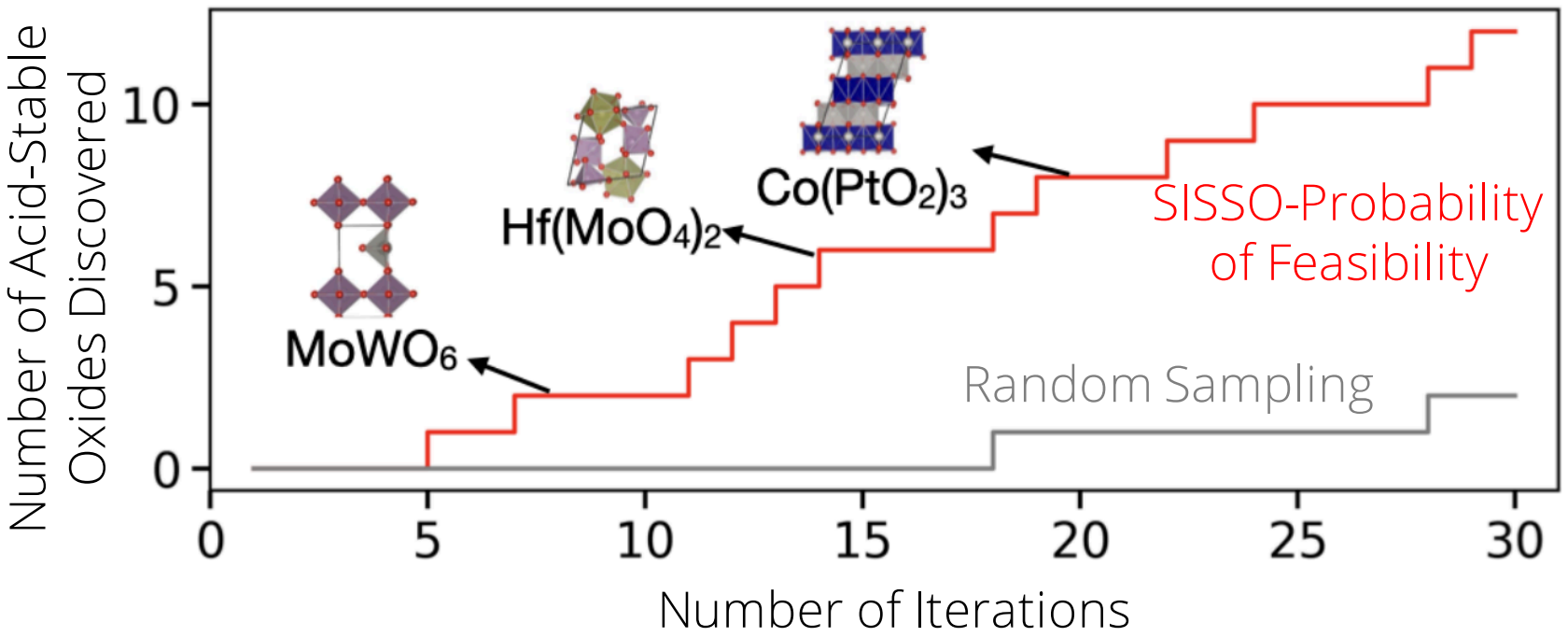
Augment

Acquisition Function

- Probability of Feasibility

A. S. Nair *et al.* npj Comput. Mat. 11, 150 (2025)

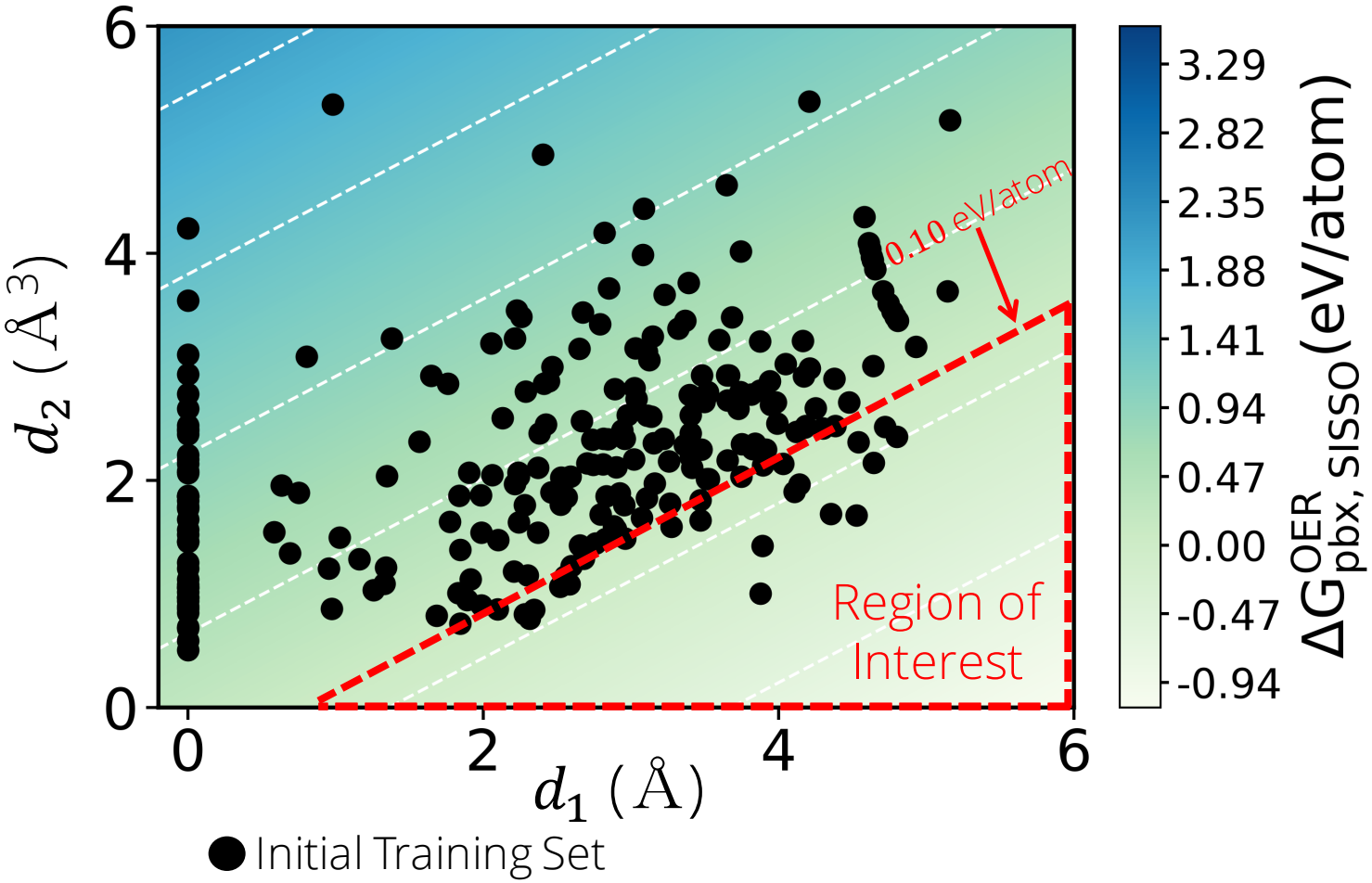
Workflow 2: Identifying Acid-Stable Oxides for Water Splitting



- Workflow Enabled the Discovery of 12 Stable Oxides, Out of 1.5 k Candidates in 30 Iterations

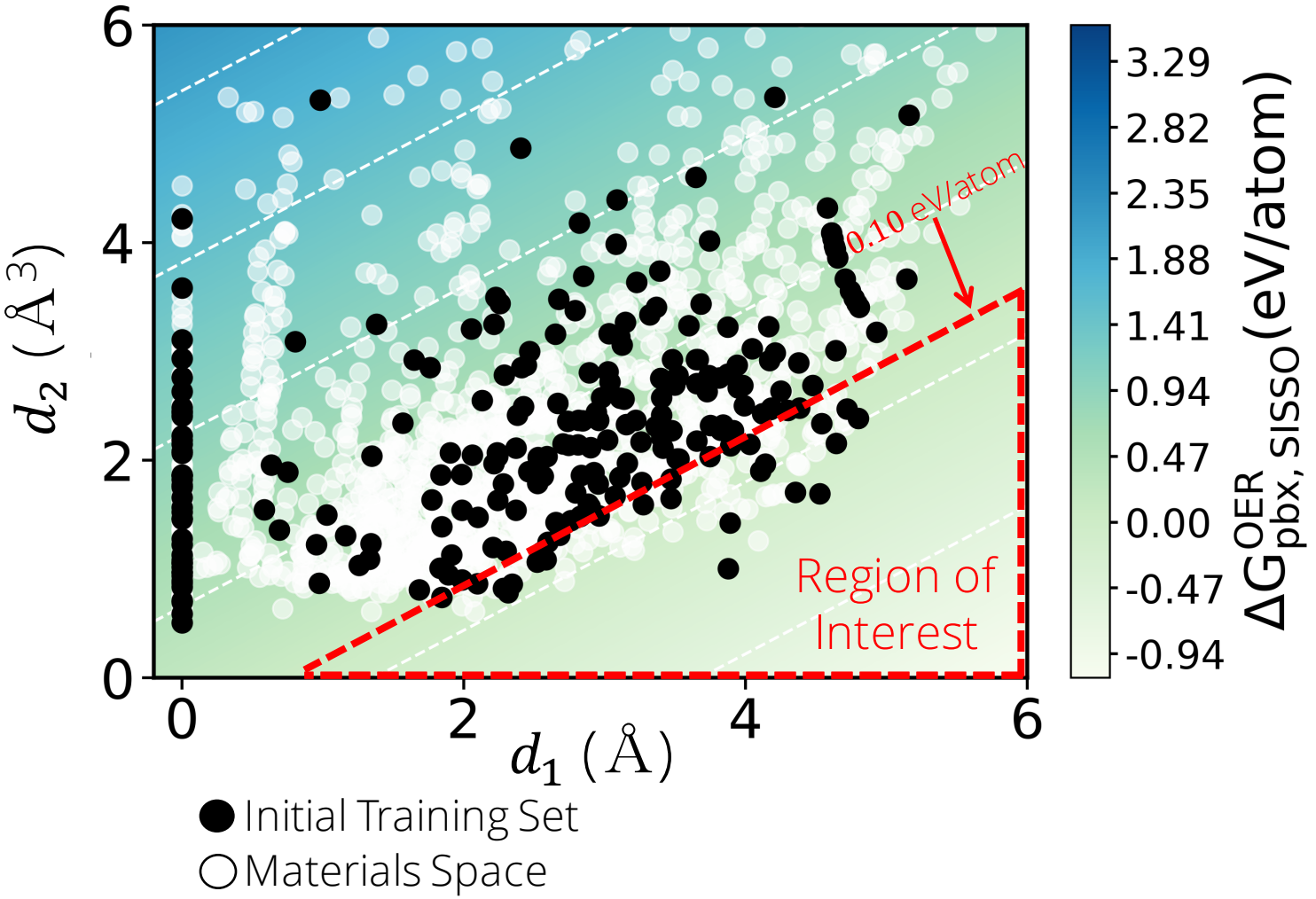
A. S. Nair *et al.* npj Comput. Mat. 11, 150 (2025)

Workflow 2: Identifying Acid-Stable Oxides for Water Splitting



A. S. Nair *et al.* npj Comput. Mat. 11, 150 (2025)

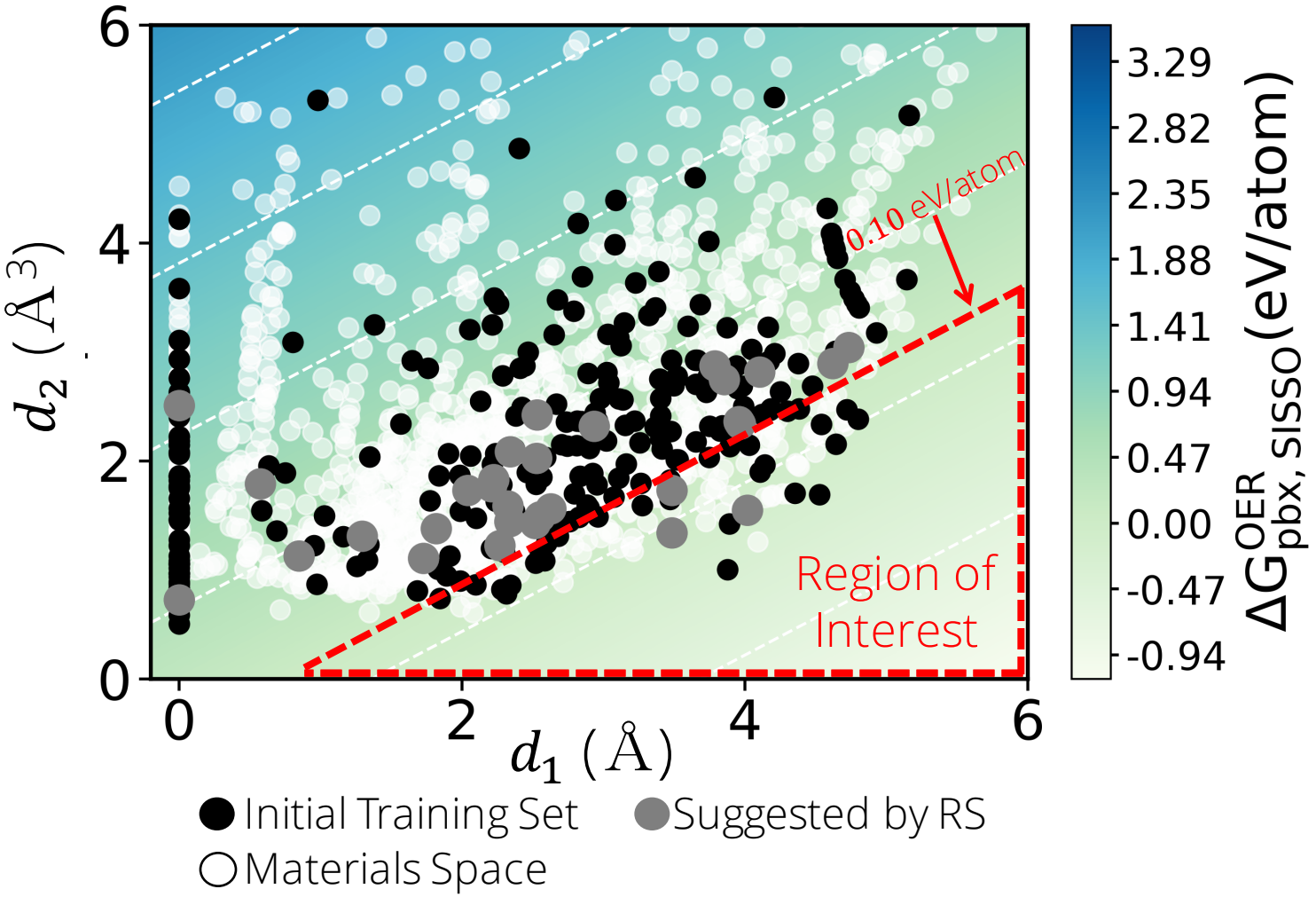
Workflow 2: Identifying Acid-Stable Oxides for Water Splitting



- Only Few Materials Stable

A. S. Nair *et al.* npj Comput. Mat. 11, 150 (2025)

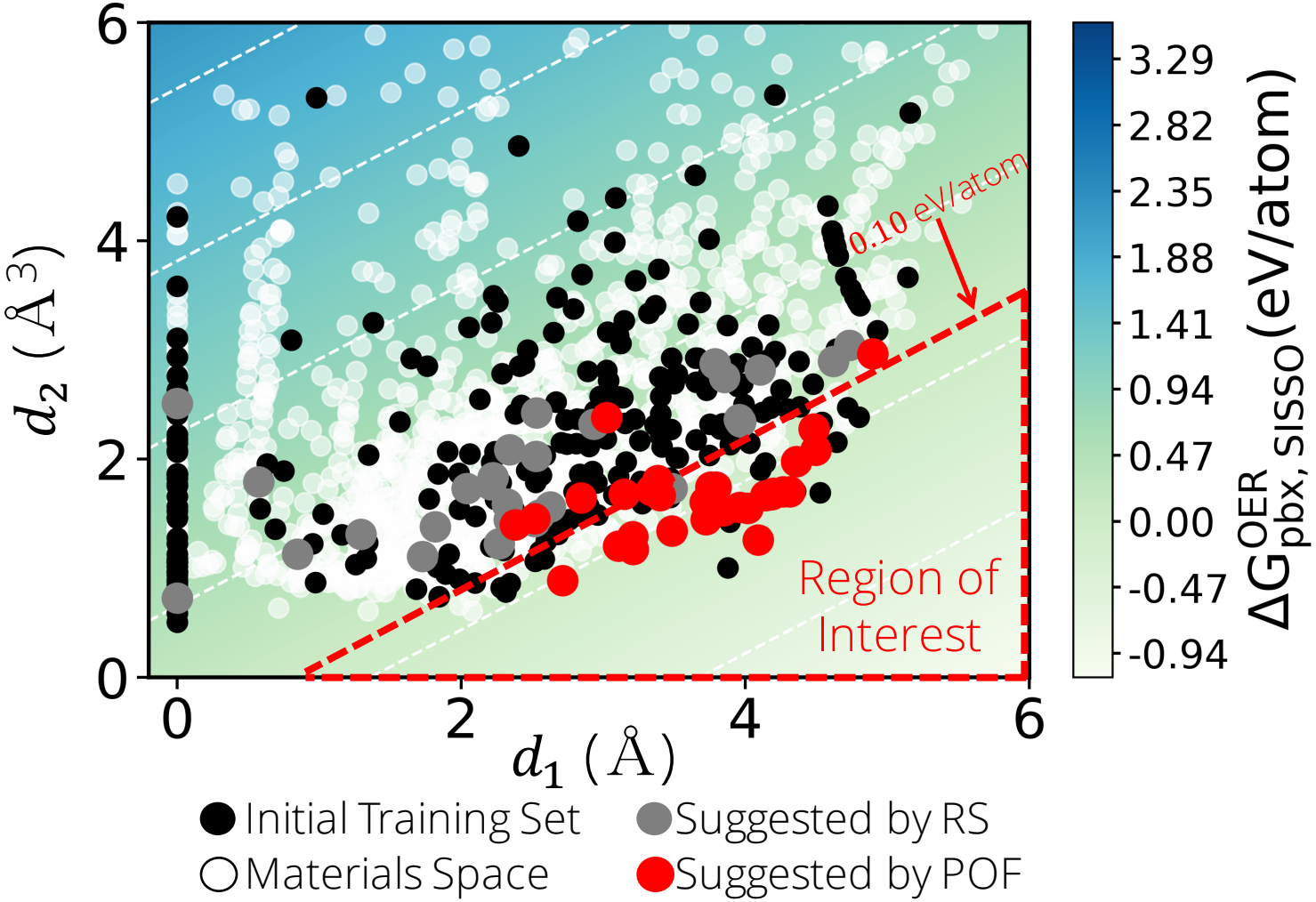
Workflow 2: Identifying Acid-Stable Oxides for Water Splitting



- Only Few Materials Stable

A. S. Nair *et al.* npj Comput. Mat. 11, 150 (2025)

Workflow 2: Identifying Acid-Stable Oxides for Water Splitting

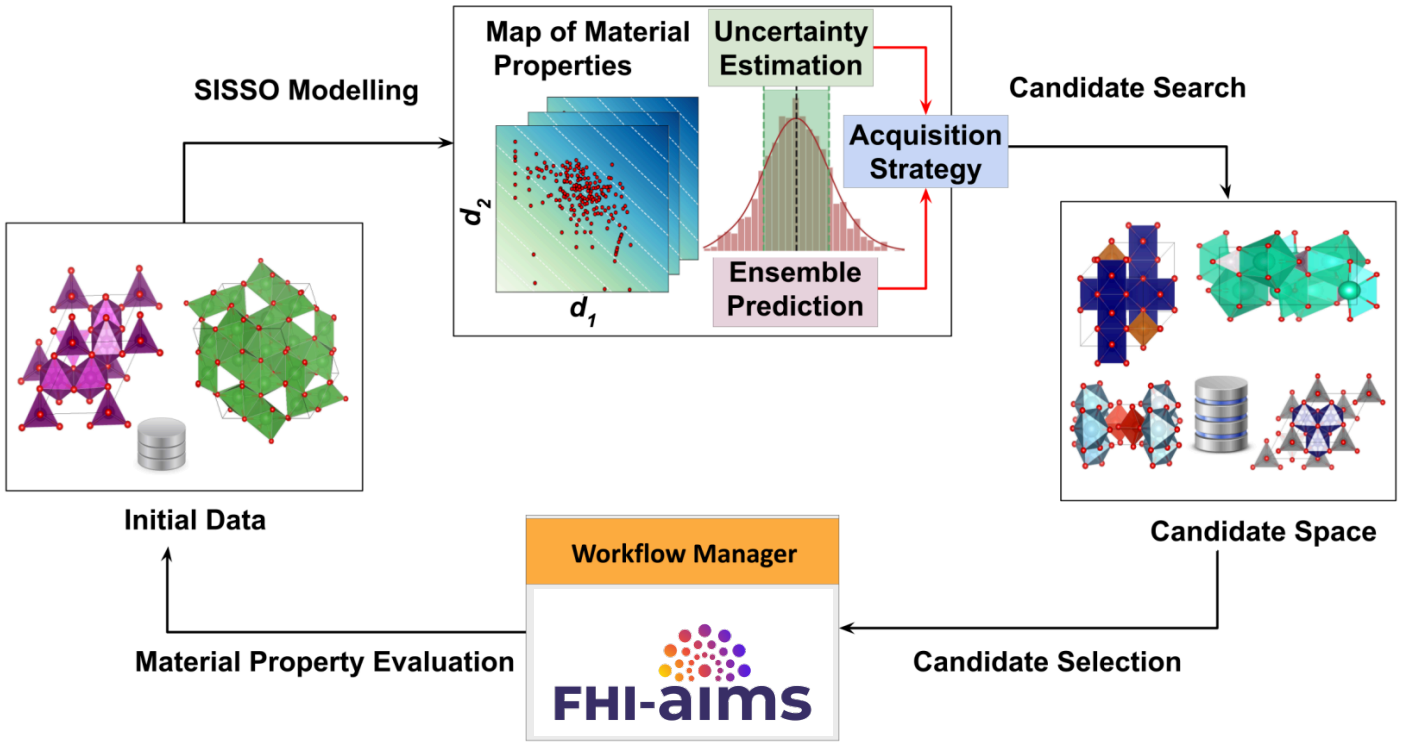


- Only Few Materials Stable
- Most Stable Materials Correctly Predicted by Initial SISSO Model
- Some Stable Oxides Identified by Workflow Incorrectly Predicted Unstable By Initial Model

A. S. Nair *et al.* npj Comput. Mat. 11, 150 (2025)

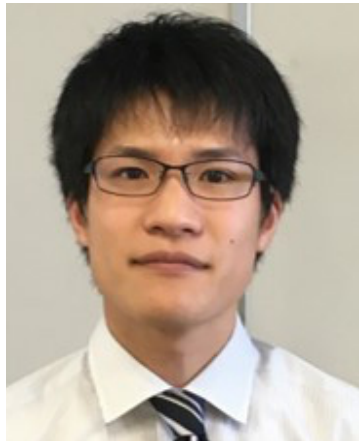
Conclusions

- Consistent Experimental/Theoretical Data + SISSO for Identifying "Materials Genes"
- AI-Guided Workflows Depends on Combination of AI Model and Acquisition Function
- Ensembles of SISSO Models for Quantifying Prediction Uncertainty
- SISSO-Guided Workflows for the Identification of Materials Based on High-Quality Calculations or Experiments



Acknowledgments

Matthias Scheffler



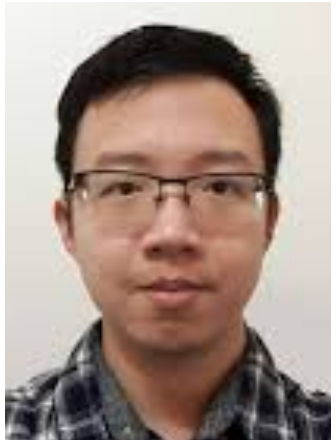
Ray Miyazaki

Akhil Nair

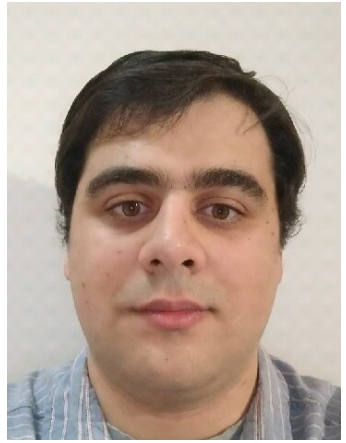
Kendra Belthle

Harun Tüysüz

Mario Boley



Felix Luong



Simon Teshuva

Daniel Schmidt

MAX-PLANCK-INSTITUT
FÜR KOHLENFORSCHUNG

MONASH University

Big-Data Analytics and Machine Learning for Materials Science

Lucas Foppa

The NOMAD Laboratory at the Fritz Haber Institute of the Max Planck Society