JOINT WINTER SCHOOL
AI FOR SCIENCES
26 - 30 January 2026
NTU Main Campus | Singapore

AI Applications in Materials Chemistry

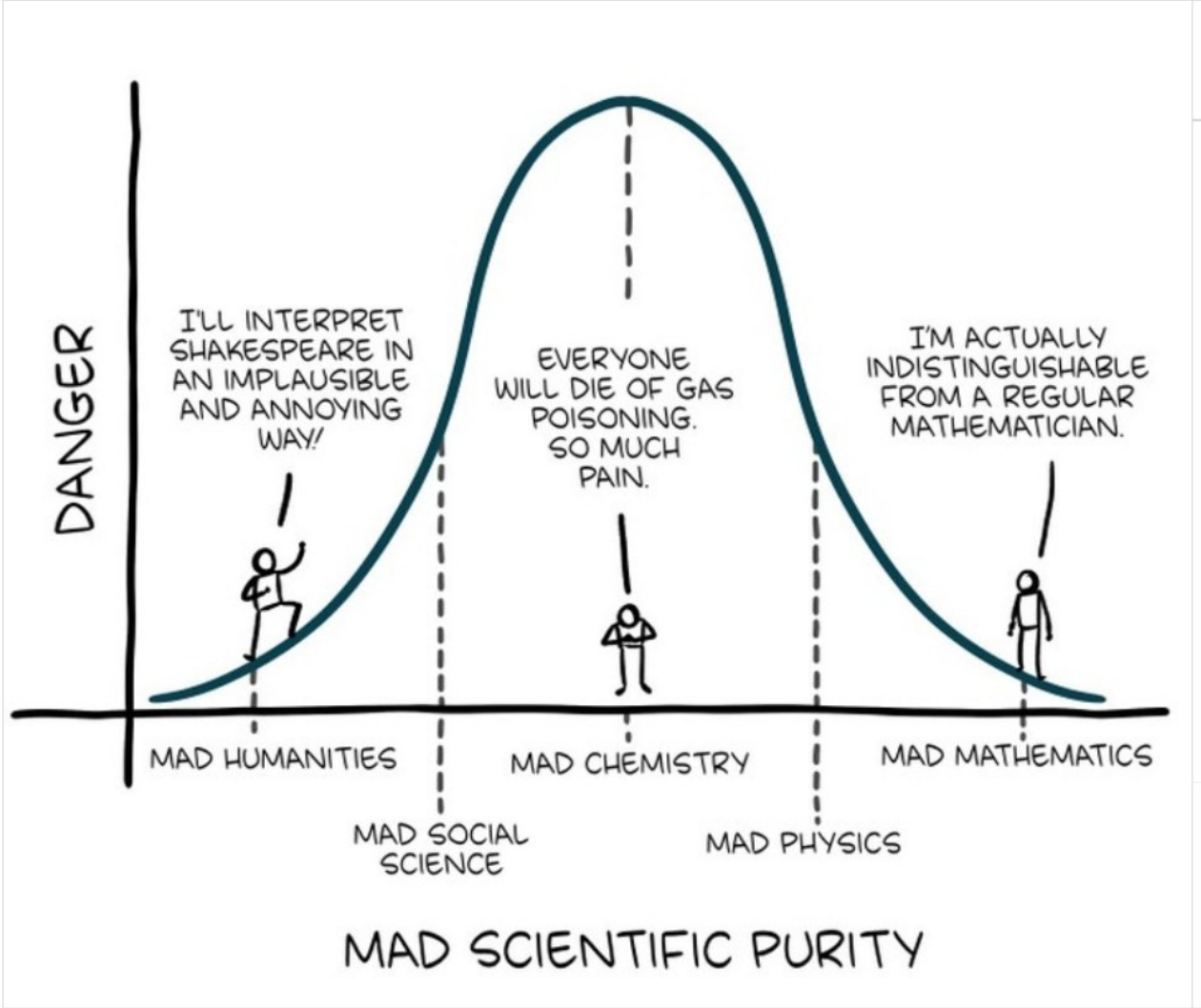
Rocio Semino

@RS_Chemist

rocio.semino@sorbonne-universite.fr

website: www.rociosemino.com

Chemistry versus other sciences



smbccomics: https://www.instagram.com/p/Cum2JYmugtV/

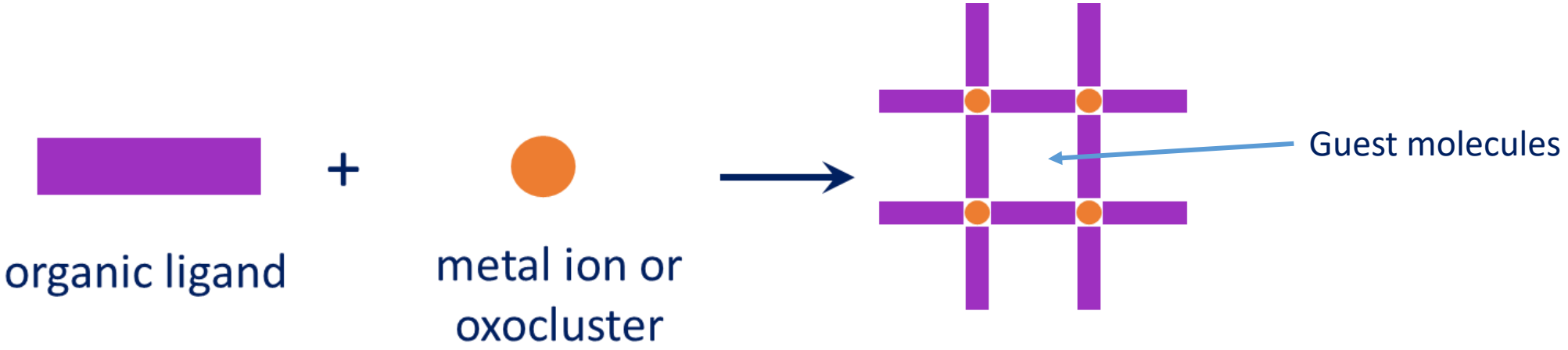
Overview of the course

- Introduction
- Machine Learning Potentials
- Property Prediction and High-Throughput Screening
- Generative and Inverse Design
- Material Phase Classification
- Synthesis Prediction and Optimization
- Self-Driving Laboratories

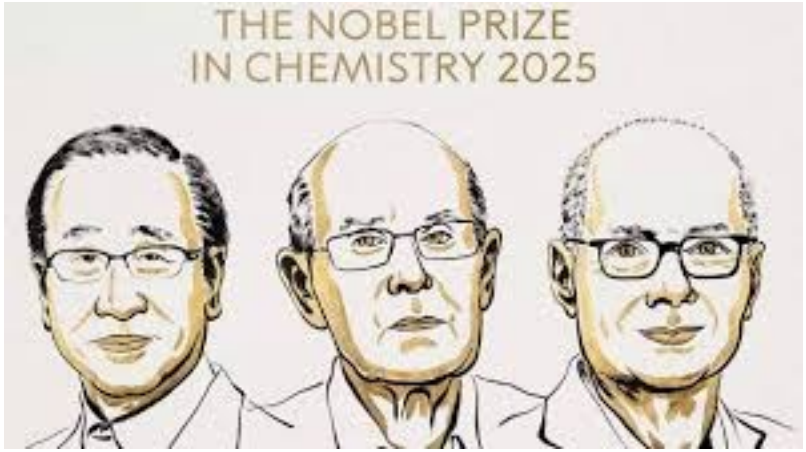
Overview of the course

- Introduction
- Machine Learning Potentials
- Property Prediction and High-Throughput Screening
- Generative and Inverse Design
- Material Phase Classification
- Synthesis Prediction and Optimization
- Self-Driving Laboratories

Introduction to metal-organic frameworks (MOFs)

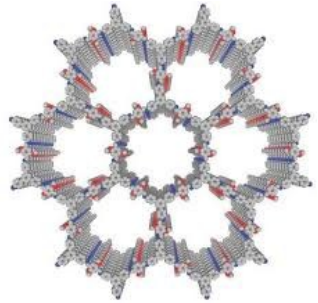


Applications: separation, storage, catalysis, drug delivery, proton conduction…

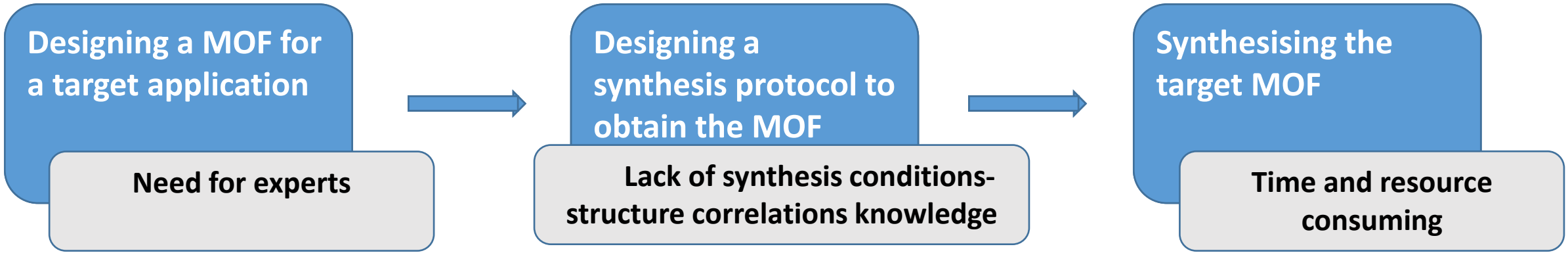


Other porous materials:

- Zeolites (SiO_4^{2-})
- Covalent-Organic Frameworks



Synthesis



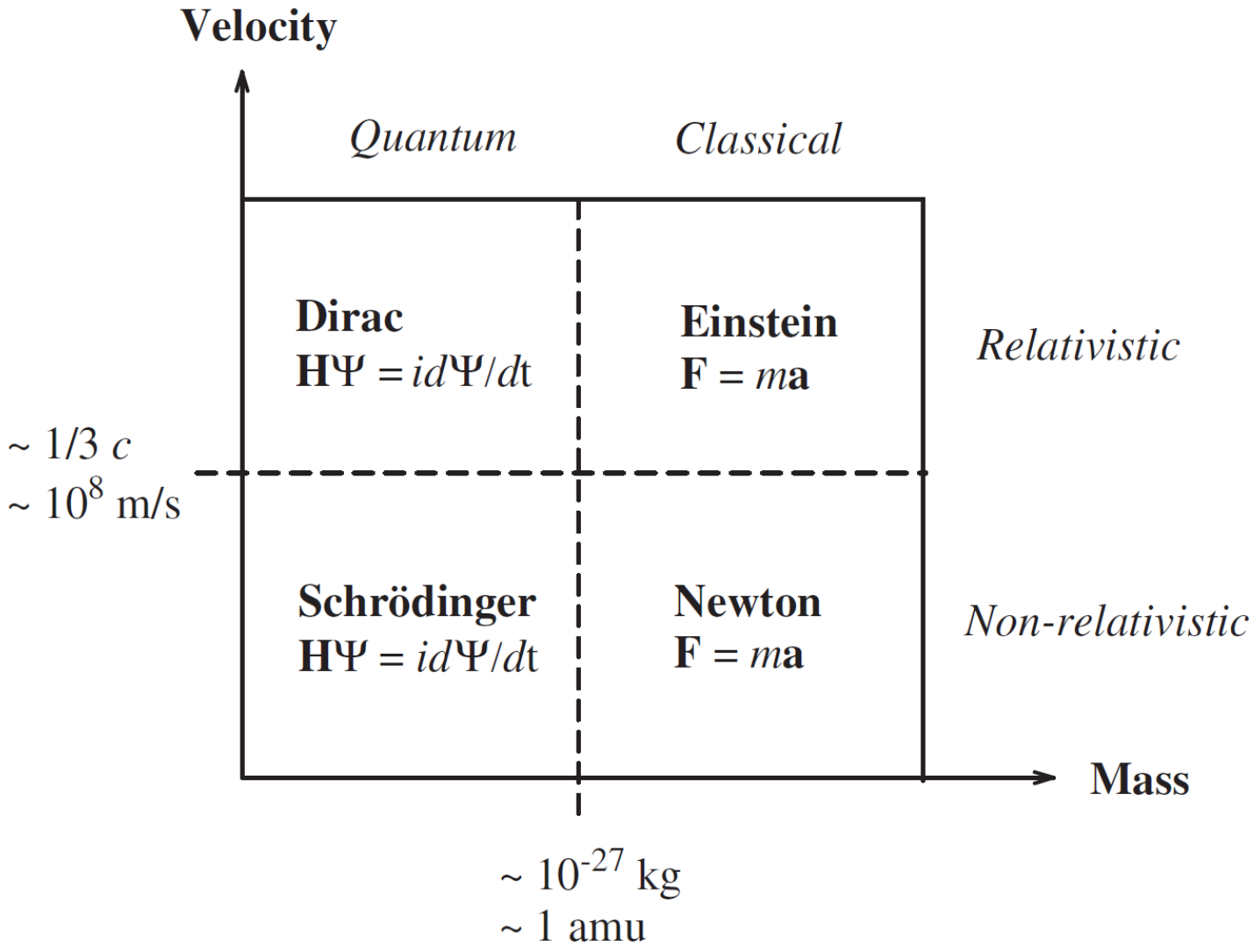
Opportunities to use artificial intelligence and machine learning methods

Computational modelling

1) **Type and number of particles**: nuclei & electrons? atoms? Molecular fragments? molecules? macromolecules?
2) **Initial condition**: determines the part of the phase space* that we will explore
3) **Interaction** between particles (forces)
4) **Dynamics**: to study the time evolution of the system

*set of all possible states of the system in the coordinates / momenta space

Computational modelling



- $m_{electron} = 9.1 \times 10^{-31}$ kg
- $m_{proton} = 1.67 \times 10^{-27}$ kg
- c = speed of light = 3×10^{8} m s^{-1}

Fig. from: "Introduction to Computational Chemistry", 2017, Frank Jensen, John Wiley & Sons, Chichester, UK.

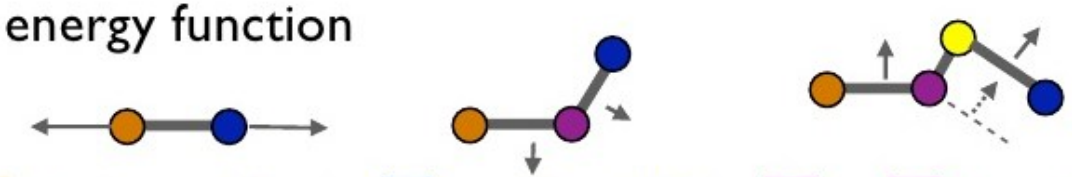
Computational modelling

	AB INITIO METHODS	**FORCE FIELD-BASED METHODS**
TYPE & NUMBER OF PARTICLES	Electrons, nuclei < 1K nuclei	Atoms, molec. fragments, molecules < 100K units
INITIAL CONDITIONS	Molecules, Fragments (cluster), periodic systems (small unit cells)	Periodic systems (large unit cells possible), solutions
INTERACTION	Schrödinger or Kohn Sham (DFT) equations	Force field
DYNAMICS	Time-dependent Schrödinger or DFT equations	Newton or Langevin equations

Force field-based molecular dynamics

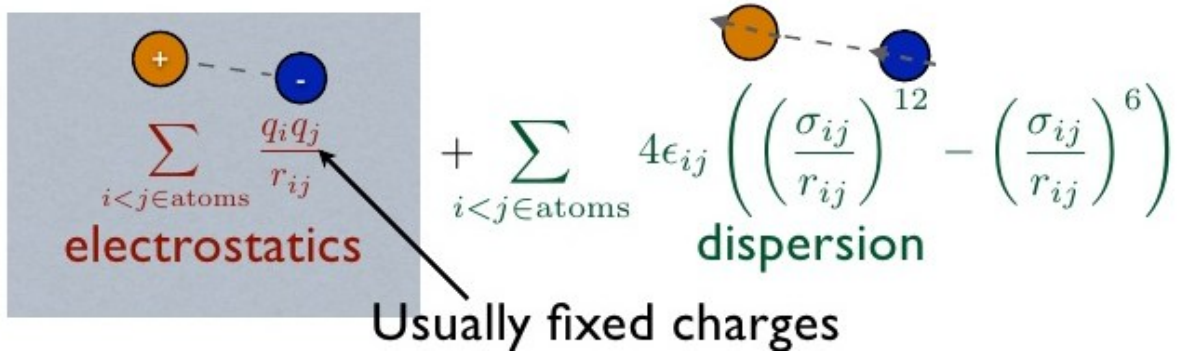
1) Define a force field → 2) Choose initial r , v → 3) Compute f → 4) Update r and v → 5) Apply T, P… → 6) Repeat 3-5 → 7)Compute properties

Typical energy function



$$E = \sum_{b \in \text{bonds}} k_b (r_b - r_b^0)^2 + \sum_{a \in \text{angles}} \kappa_a (\theta_a - \theta_a^0)^2 + \sum_{d \in \text{dihedrals}} \sum_n l_{dn} \cos(n\pi)$$

bond stretch | angle torsion | dihedrals



$$+ \sum_{i<j \in \text{atoms}} \frac{q_i q_j}{r_{ij}} + \sum_{i<j \in \text{atoms}} 4\epsilon_{ij} \left(\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right)$$

electrostatics | dispersion

Usually fixed charges

Frenkel, D.; Smit, B. Understanding Molecular Simulations. Academic Press: San Diego, California, USA (2002)

Force field-based molecular dynamics

$$-\frac{\partial \boldsymbol{V}}{\partial \boldsymbol{r}} = m\,\frac{\partial^2 \boldsymbol{r}}{\partial t^2}$$

$$\boldsymbol{F} = -\nabla \boldsymbol{V}$$

$$\boldsymbol{F_i} = \sum_{i \neq j}^{N} \boldsymbol{F_{ji}}$$

Frenkel, D.; Smit, B. Understanding Molecular Simulations. Academic Press: San Diego, California, USA (2002)

Force field-based molecular dynamics

The output of 6) is a **trajectory**, from which we can obtain the value of several **observables**, by averaging over particles and configurations

$$A_{obs} = \lim_{t_{obs}\to\infty} \frac{1}{t_{obs}} \int_0^{t_{obs}} A\big(\boldsymbol{\Gamma}(t)\big)dt = \frac{1}{\tau_{obs}} \sum_{\tau=1}^{\tau_{obs}} A\big(\boldsymbol{\Gamma}(t)\big) = \frac{1}{N_{config}} \sum_{\alpha=1}^{N_{config}} \rho_\alpha A_\alpha$$

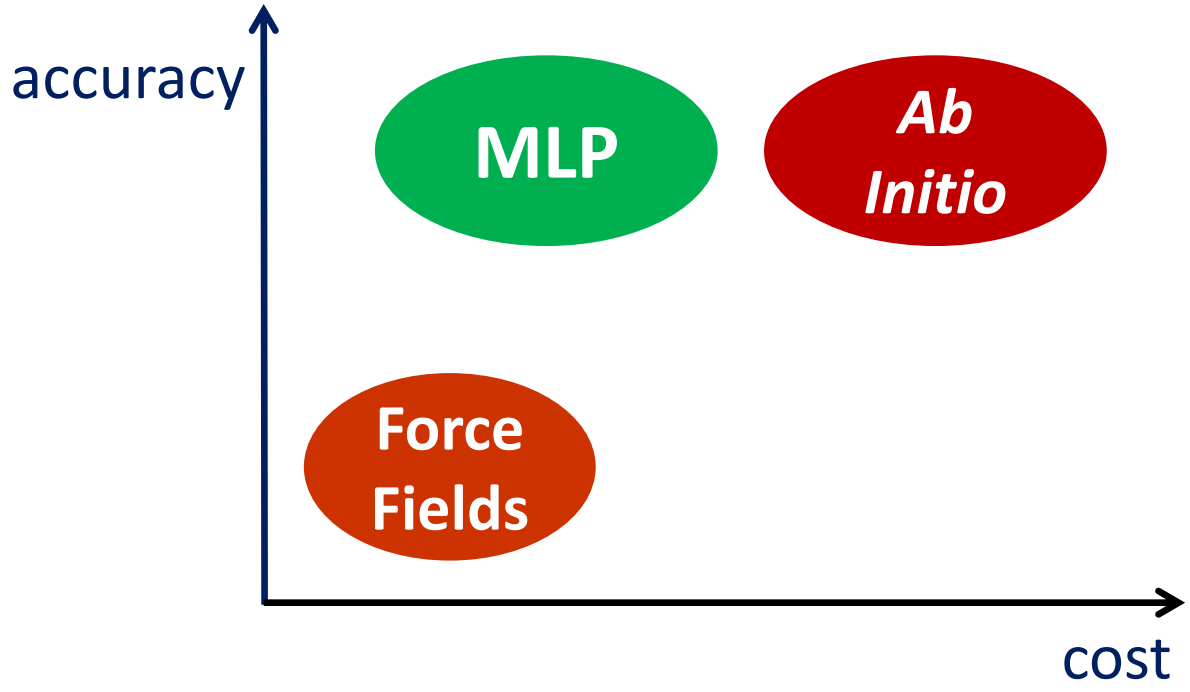
Frenkel, D.; Smit, B. Understanding Molecular Simulations. Academic Press: San Diego, California, USA (2002)

Overview of the course

- Introduction
- Machine Learning Potentials
- Property Prediction and High-Throughput Screening
- Generative and Inverse Design
- Material Phase Classification
- Synthesis Prediction and Optimization
- Self-Driving Laboratories

Machine Learning Potentials

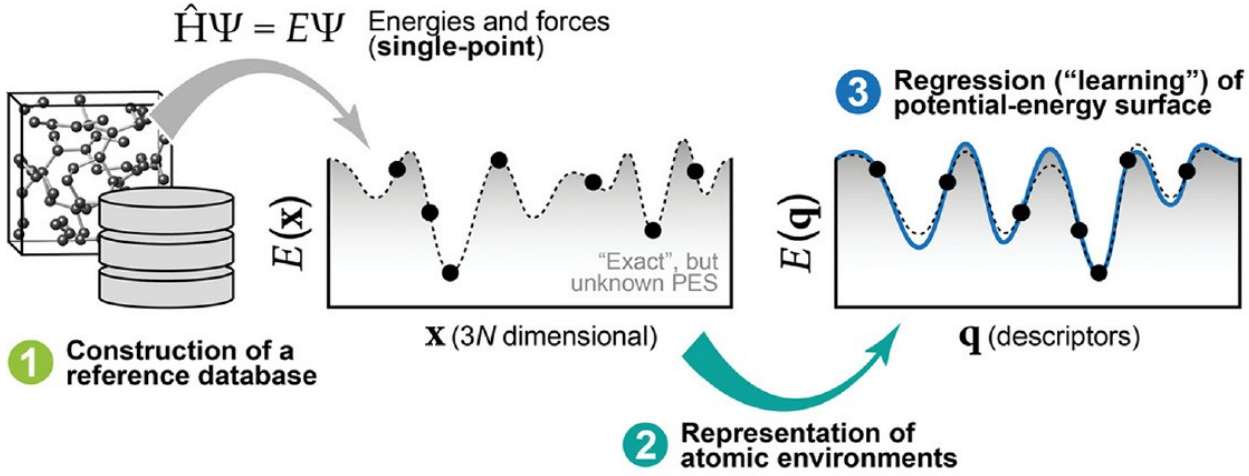
MLPs bridge the gap between high-accuracy but slow quantum mechanical (DFT) methods and fast but less accurate classical force fields:

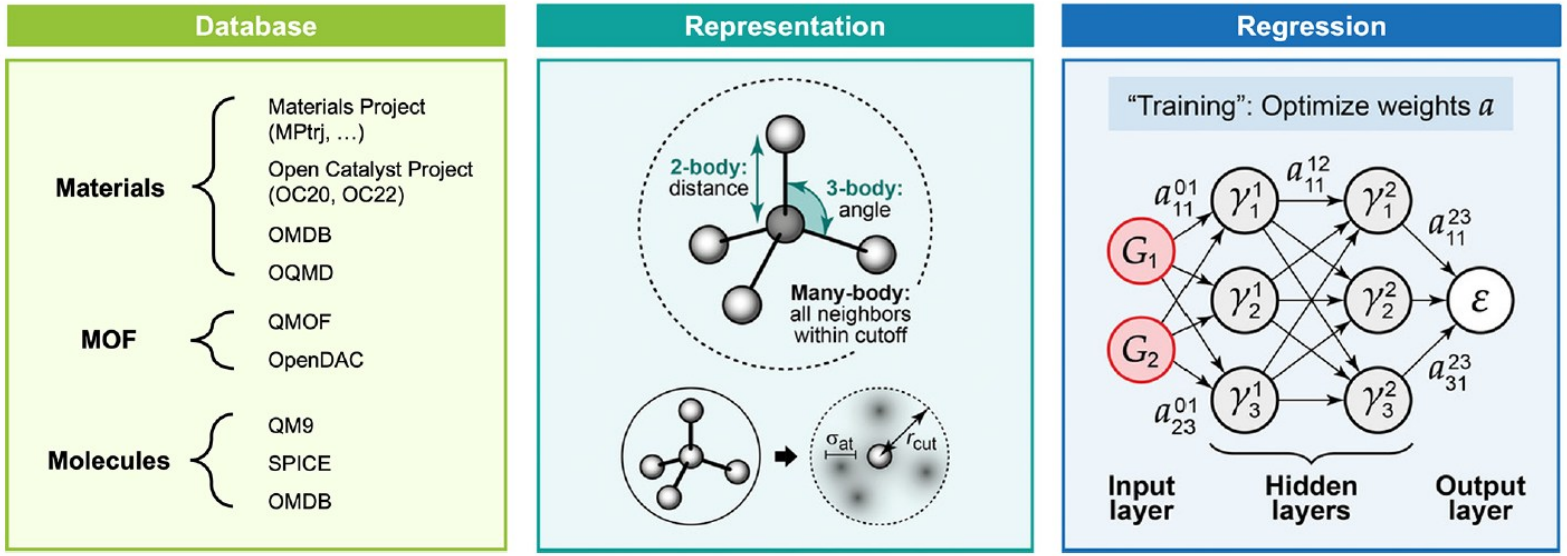


May allow the breaking and formation of chemical bonds

J. Behler, Int. J. Quantum Chem., 115, 1032-1050 (2015)

Machine Learning Potentials

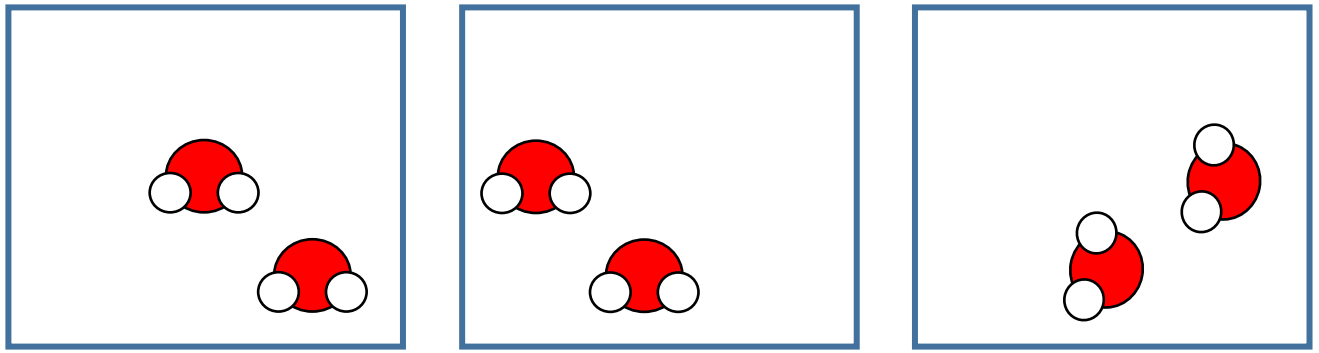




V. L. Deringer, M.A. Caro, G. Csanyi, Adv. Mater., 31, 1902765 (2019)

Methodology: descriptors

Cartesian coordinates are a bad descriptor:



Same energy,
different xyz…

We need:

- Translational invariance
- Rotational invariance
- Permutation invariance
- Any applicable point group symmetry invariance

J. Behler, J. Chem. Phys., 145, 170901 (2016)

Methodology: descriptors

Behler-Parrinello Symmetry functions

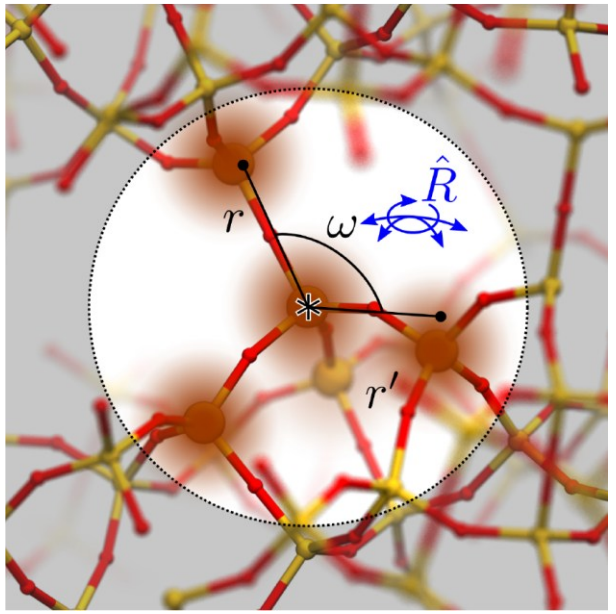
$$G_i^{atom,rad} = \sum_{j=1}^{N_{atom}} \left(e^{-\eta(R_{ij}-R_s)^2} \times f_c(R_{ij}) \right)$$

$$G_i^{atom,ang} = 2^{1-\zeta} \sum_{j,k\neq i}^{all} \left((1+\lambda\cos\theta_{ijk})^{\zeta} \times e^{-\eta(R_{ij}^2+R_{ik}^2+R_{jk}^2)^2} \times f_c(R_{ij}) \times f_c(R_{ik}) \times f_c(R_{jk}) \right)$$

Smooth overlap of atomic positions (SOAP):

$$\psi^{\alpha}_{\mathcal{X}_j}(\mathbf{r}) = \sum_{i\in\mathcal{X}_j^{\alpha}} g(\mathbf{r}-\mathbf{r}_{ij}) f_c(r_{ij})$$

$$\langle \alpha r \alpha' r' \omega | \mathcal{X}_j \rangle = \int \mathrm{d}\hat{R}\, r r' \psi^{\alpha}_{\mathcal{X}_j}(r\hat{R}\hat{\mathbf{e}}_z) \times \psi^{\alpha'}_{\mathcal{X}_j}(r'\hat{R}(\omega\hat{\mathbf{e}}_z + \sqrt{1-\omega^2}\hat{\mathbf{e}}_x))$$

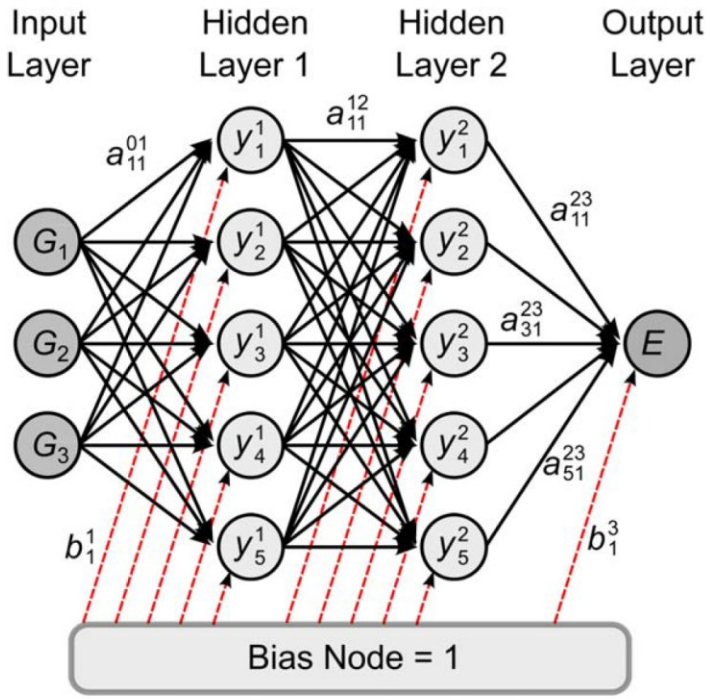


A. P. Bartók, R. Kondor, G. Csányi Phys. Rev. B 87, 184115 (2013)

J. Behler, J. Chem. Phys., 145, 170901 (2016)

Methodology: Neural Networks

Training ML models on high-quality configurations and forces derived from DFT calculations to learn the **Potential Energy Surface**



$$y_i^j = f_i^j\left(b_i^j + \sum_{k=1}^{N_{j-1}} a_{k,i}^{j-1,j}.y_k^{j-1}\right)$$

- Neurons = y
- Weight parameters = a
- Bias weights = b
- Activation or basis function = f

But a large database of benchmark configurations is needed…

J. Behler, Int. J. Quantum Chem., 115, 1032-1050 (2015)

Methodology: further improvements

- Efficiently sampling the PES through metadynamics to reduce the number of expensive DFT calculations needed for training (active learning approach)
- **MACE** explicitly preserves rotation, translation, and permutation invariances. While earlier MLP generations often relied on local descriptors invariant to these transformations (such as symmetry functions), these newer models utilize **message passing** to capture directional interactions and complex atomic environments more effectively.
  - Smaller databases needed to train them.
  - Universality.

S. Vandenhaute, M. Cools-Ceuppens, S. DeKeyser, T. Verstraelen and V. Van Speybroeck, npj Comput. Mater., 9, 1–8 (2023)

A. M. Elena, P. D. Kamath, T. Jaffrelot Inizan, A. S. Rosen, F. Zanca and K. A. Persson, npj Comput. Mater. 11, 125 (2025)

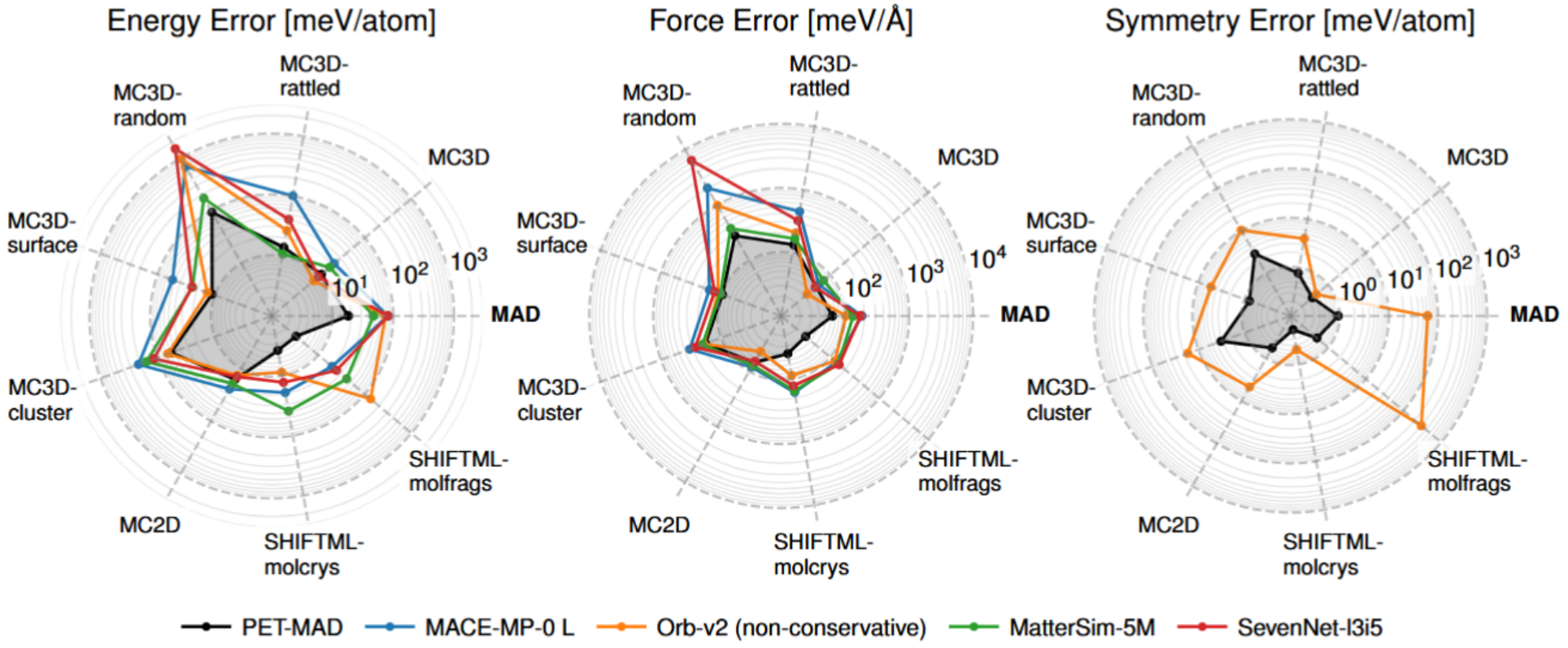
Matbench Discovery benchmark

Model	Org	r_{cut}	Links	Date Added	Targets	Params	Training Set	RMSD ↓	κ_{SRME} ↓	R^2 ↑	MAE ↓	Prec ↑	DAF ↑	F1 ↑	Acc ↑	CPS
PET-OAM-XL	EPFL	n/a		2026-01-10	EFS_G	730M	6.6M (113M) OMat24+sAlex+MPtrj	0.060	0.119	0.864	0.019	0.929	6.075	0.924	0.977	0.898
eSEN-30M-OAM		6 Å		2025-03-17	EFS_G	30.2M	6.6M (113M) OMat24+MPtrj+sAlex	0.061	0.170	0.866	0.018	0.928	6.069	0.925	0.977	0.888
EquFlash	SAMSUNG ELECTRONICS	6 Å		2025-06-23	EFS_G	28.7M	6.6M (113M) OMat24+MPtrj+sAlex	0.060	0.158	0.871	0.019	0.915	5.983	0.919	0.975	0.888
Nequip-OAM-XL		6 Å		2025-11-30	EFS_G	32.1M	6.6M (113M) OMat24+sAlex+MPtrj	0.063	0.125	0.872	0.020	0.897	5.869	0.906	0.971	0.886
MatRIS-10M-OAM		6 Å		2025-10-29	EFS_GM	10.4M	6.6M (113M) OMat24+sAlex+MPtrj	0.060	0.218	0.871	0.019	0.923	6.039	0.921	0.976	0.877
SevenNet-Omni-i12		6 Å		2026-01-12	EFS_G	54.9M	243M COSMOSDataset	0.062	0.192	0.868	0.021	0.910	5.954	0.906	0.971	0.873
Nequip-OAM-L		6 Å		2025-09-08	EFS_G	9.6M	6.6M (113M) OMat24+sAlex+MPtrj	0.065	0.166	0.865	0.022	0.890	5.823	0.893	0.967	0.870
TACE-v1-OAM-M		6 Å		2026-01-06	EFS_G	18.8M	6.6M (113M) OMat24+sAlex+MPtrj	0.065	0.173	0.865	0.022	0.879	5.749	0.889	0.965	0.867
GRACE-2L-OAM-L	ICAMS	6 Å		2025-09-09	EFS_G	26.4M	6.6M (113M) OMat24+sAlex+MPtrj	0.064	0.169	0.862	0.022	0.893	5.840	0.883	0.964	0.865
ORB v3		6 Å		2025-04-05	EFS_G	25.5M	6.47M (133M) MPtrj+Alex+OMat24	0.075	0.210	0.821	0.024	0.904	5.912	0.905	0.971	0.860
Allegro-OAM-L		7 Å		2025-09-08	EFS_G	9.7M	6.6M (113M) OMat24+sAlex+MPtrj	0.065	0.319	0.868	0.022	0.867	5.674	0.895	0.966	0.840
GRACE-2L-OAM	ICAMS	6 Å		2025-02-06	EFS_G	12.6M	6.6M (113M) OMat24+sAlex+MPtrj	0.067	0.294	0.862	0.023	0.883	5.774	0.880	0.963	0.837
DPA-3.1-3M-FT	AISI	6 Å		2025-06-05	EFS_G	3.27M	163M OpenLAM	0.069	0.469	0.869	0.023	0.866	5.667	0.884	0.963	0.802
eSEN-30M-MP		6 Å		2025-03-17	EFS_G	30.1M	146k (1.58M) MPtrj	0.075	0.340	0.822	0.033	0.804	5.260	0.831	0.946	0.797
MACE-MPA-0		6 Å		2024-12-09	EFS_G	9.06M	3.37M (12M) MPtrj+sAlex	0.073	0.412	0.842	0.028	0.853	5.582	0.852	0.954	0.795
MatRIS-10M-MP		6 Å		2025-10-29	EFS_GM	10.4M	146k (1.58M) MPtrj	0.072	0.489	0.824	0.031	0.829	5.422	0.847	0.951	0.778
AlphaNet-v1-OMA		5 Å		2025-05-12	EFS_G	4.65M	6.6M (113M) OMat24+sAlex+MPtrj	0.079	0.643	0.831	0.024	0.879	5.747	0.901	0.968	0.769
MatterSim v1 5M		5 Å		2024-12-16	EFS_G	4.55M	17M MatterSim	0.073	0.575	0.863	0.024	0.895	5.852	0.862	0.959	0.767

https://matbench-discovery.materialsproject.org/

State-of-the-art: PET-MAD

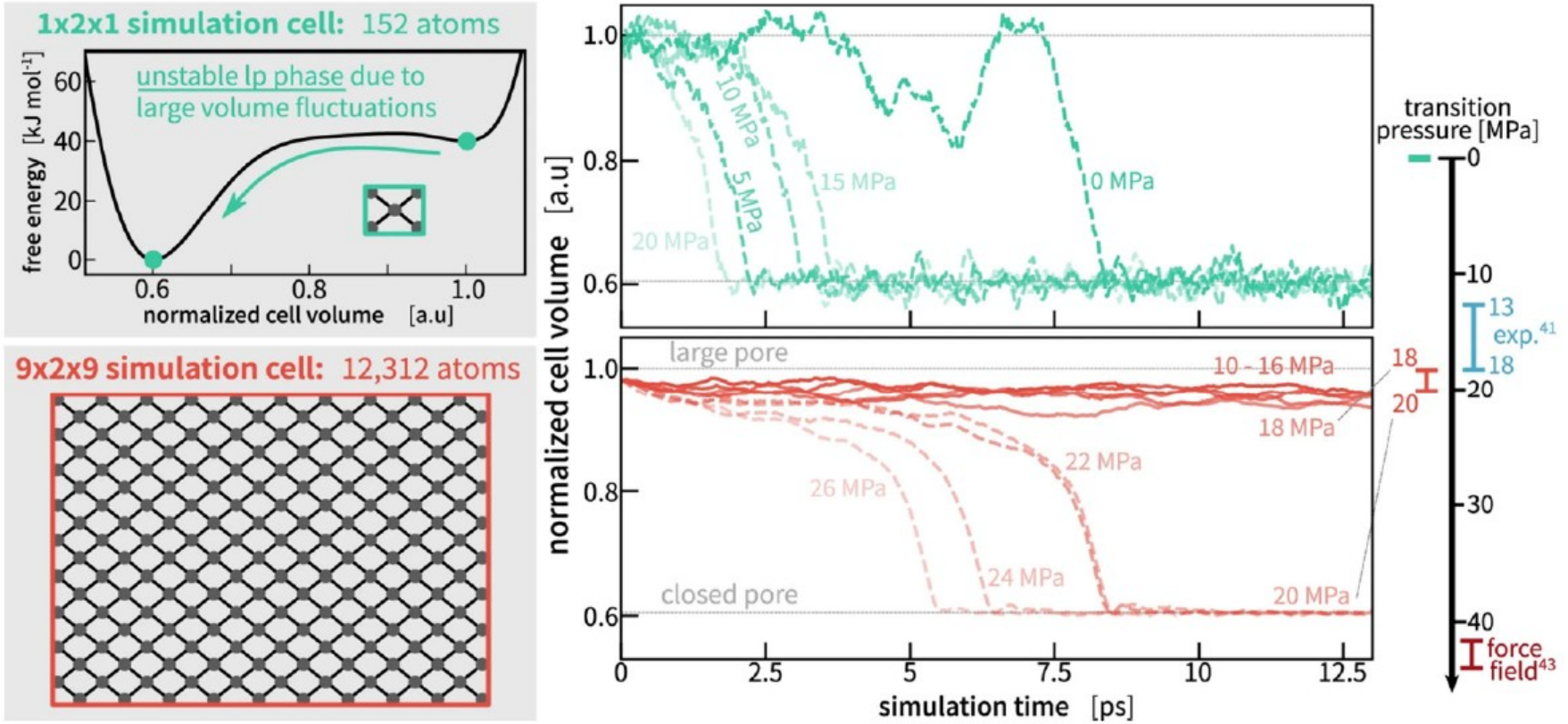
- Transformer-based graph neural network
- Fast fine-tuning possible through a low-rank adaptation technique



A. Mazitov, F. Bigi, M. Kellner, P. Pegolo, D. Tisi, G. Fraux, S. Pozdnyakov, P. Loche, M. Ceriotti, Nat Commun 16, 10653 (2025)

Example: phase transitions

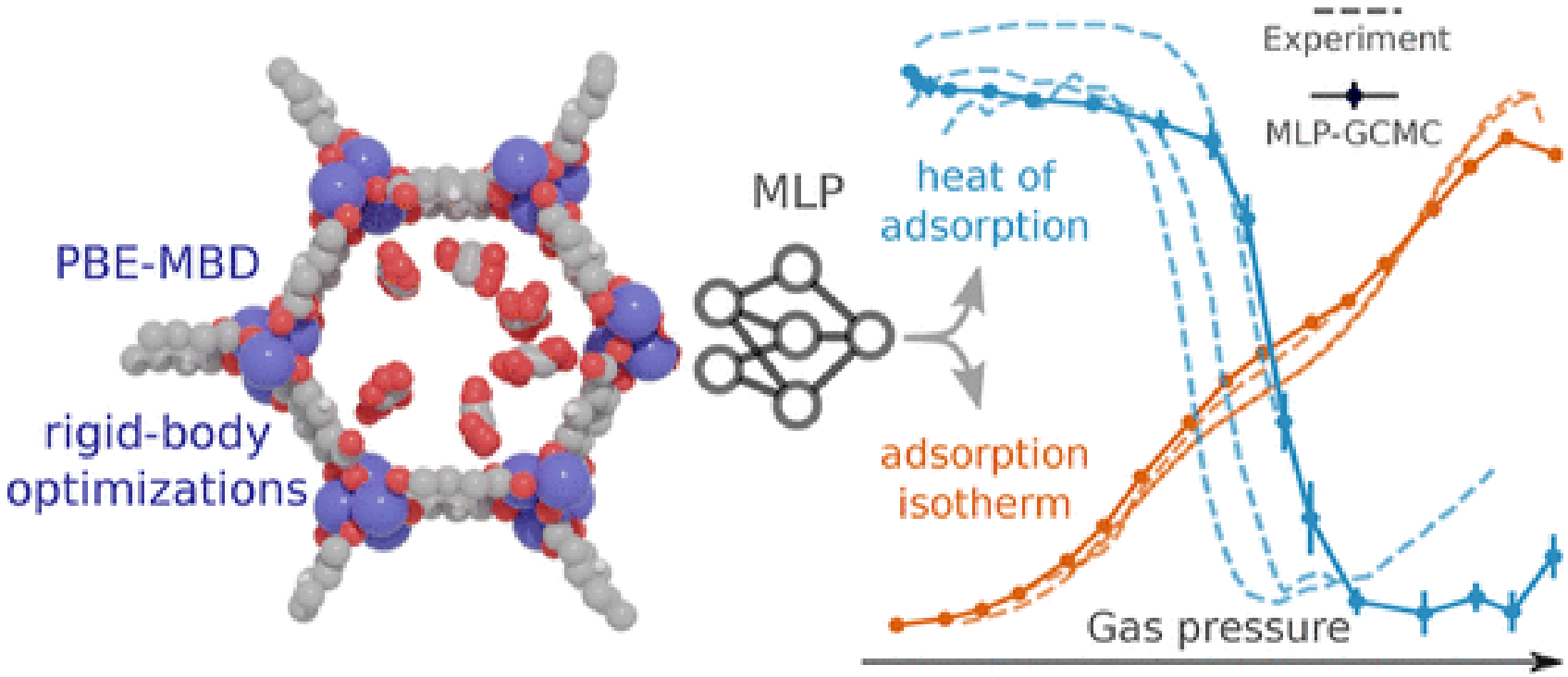
Accurately reproducing experimental transition pressures in flexible materials like MIL-53(Al) using large supercells:



S. Vandenhaute, M. Cools-Ceuppens, S. DeKeyser, T. Verstraelen and V. Van Speybroeck, npj Comput. Mater., 9, 1–8 (2023)

Example: modelling open metal sites

Capturing guest-induced framework dynamics to accurately predict CO_2 adsorption isotherms – difficult with standard force fields



Limitation: lack of transferability

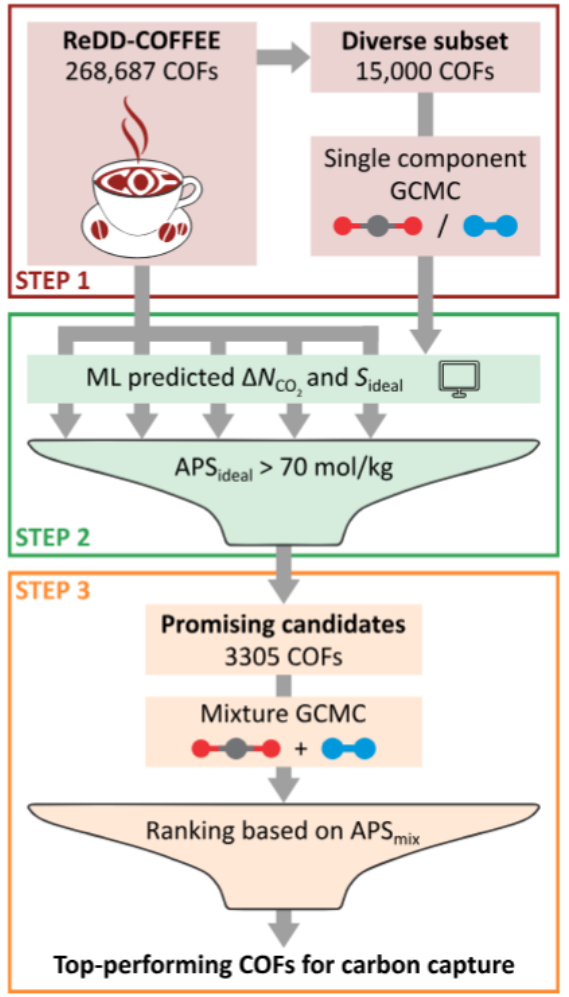
R. Goeminne, L. Vanduyfhuys, V. Van Speybroeck, T. J. Verstraelen, Chem. Theory Comput., 19, 6313−6325 (2023)

Overview of the course

- Introduction
- Machine Learning Potentials
- Property Prediction and High-Throughput Screening
- Generative and Inverse Design
- Material Phase Classification
- Synthesis Prediction and Optimization
- Self-Driving Laboratories

Property prediction and high-throughput screening

- High-throughput screening: Using (computational) tools to identify top-performing materials to direct lab resources toward the most promising candidates.
- Many prototypes, candidate applications & operating conditions → faster methods are needed
- Impact in carbon capture, energy storage and other societal applications

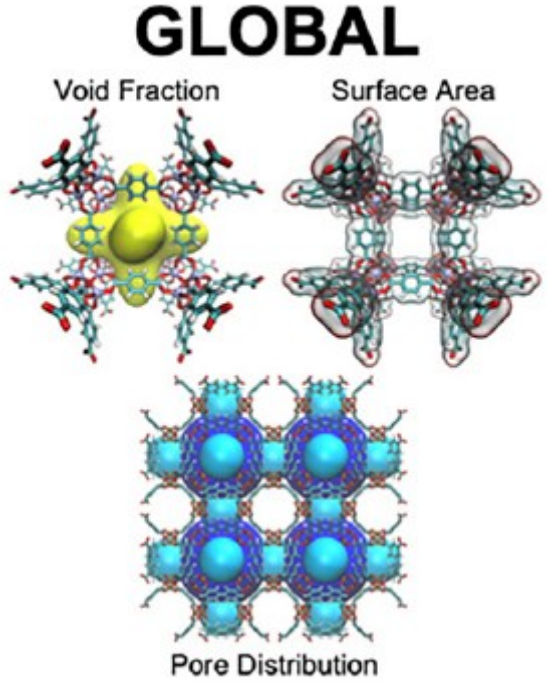


D. A. Gomez-Gualdron, T. Gercina de Vilas, K. Ardila, F. Fajardo-Rojas, A. J. Pak; Mater. Horiz., Advance Article, 10.1039/d5mh01467k (2026)

J. S. De Vos, et al, Chem. Mater. 36, 4315−4330 (2024)

Methodology: descriptor engineering

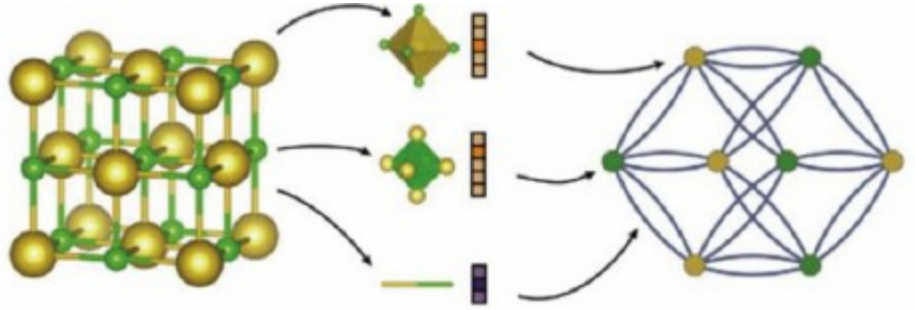
- Global features: Representing materials via global statistics (geometric descriptors like pore diameter and surface area) and processing conditions.
- Support vector regressors, ANNs and decision trees perform better than linear regression in predicting adsorption properties → non linear relationship between adsorption properties and global features

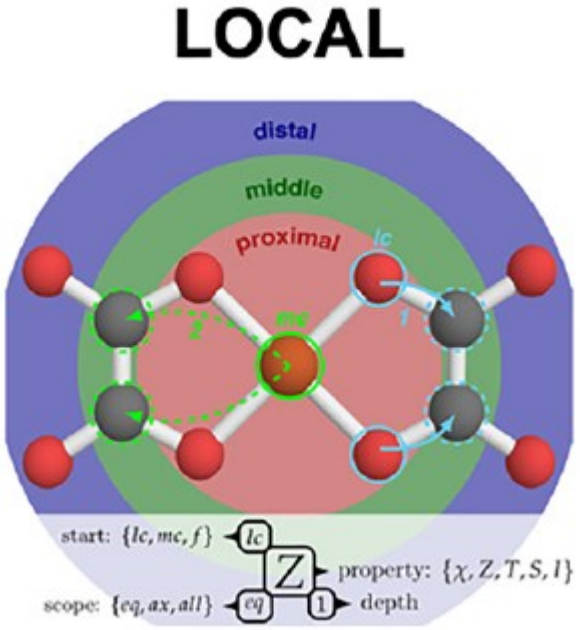


J. Cui et al, Nat. Commun., 14, 7043 (2023)

Methodology: descriptor engineering

- Local features: at the atomic level, such as graphs representations in which bonds are edges and atoms are nodes, atomic property weighted radial distribution functions (AP-RDF) and revised autocorrelation functions (RAC).



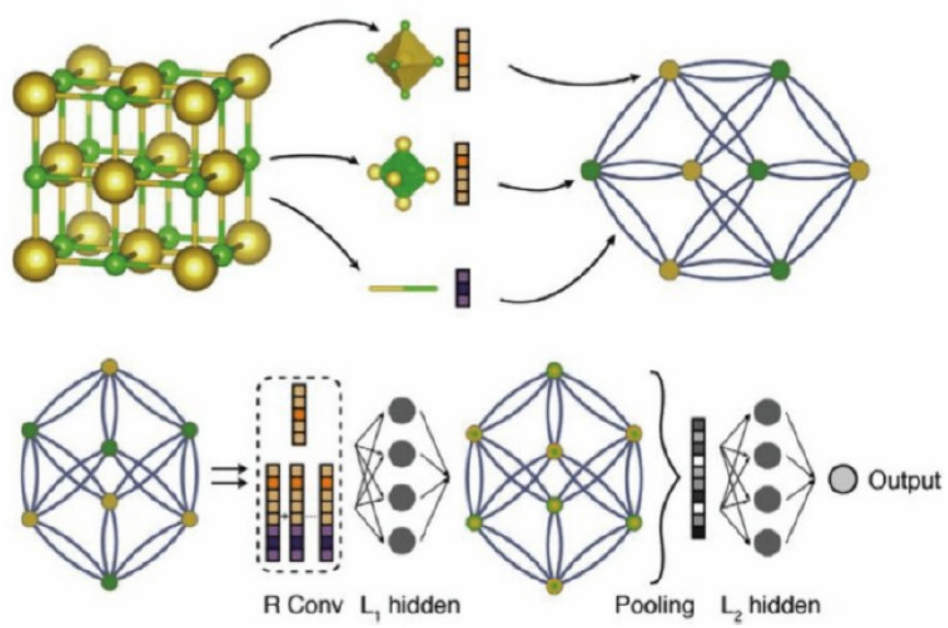


T. Xie and J. C. Grossman, Phys. Rev. Lett. 120, 145301 (2018)

D. A. Gomez-Gualdron, T. Gercina de Vilas, K. Ardila, F. Fajardo-Rojas, A. J. Pak; Mater. Horiz., Advance Article, 10.1039/d5mh01467k (2026)

Methodology: graph neural networks

Graph Neural Networks (GNNs): using graph representations as features, to learn properties. Crystal Graph Convolutional Neural Network (CGCNN) to treat periodicity.



T. Xie and J. C. Grossman, Phys. Rev. Lett. 120, 145301 (2018)

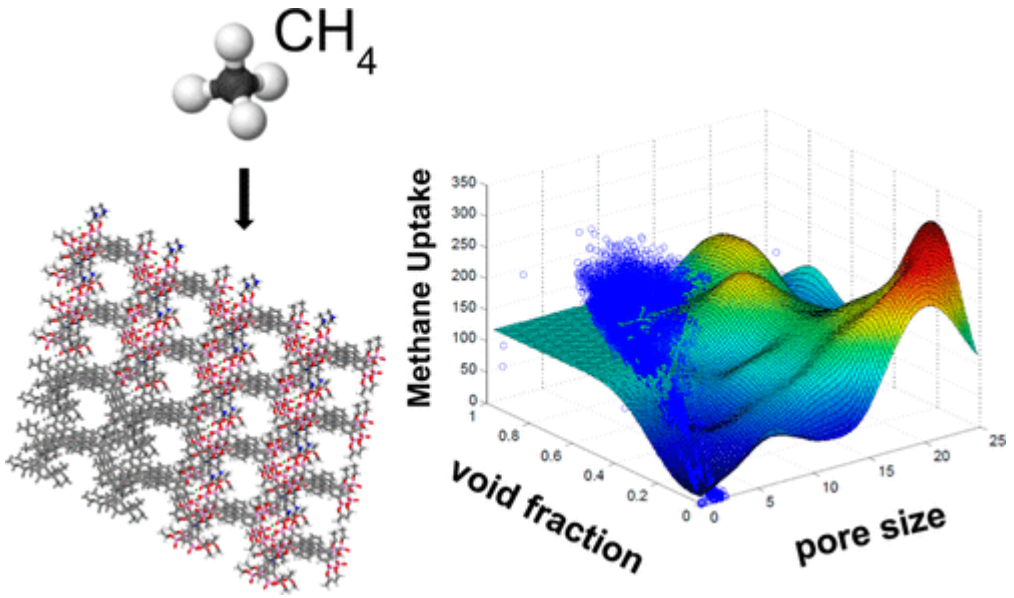
Methodology: transformers

- **Transformer Architectures:** primarily investigated in natural language processing, transformers are a neural network architecture that includes (trainable) matrix operations that embody the concept of attention.
- Typically, a MLP is trained to aggregate this attention-weighted ‘‘transformation’’ into the final prediction.
- Their large number of hyperparameters offers significant potential and versatility to capture complex relationships within structural sequence data.

D. A. Gomez-Gualdron, T. Gercina de Vilas, K. Ardila, F. Fajardo-Rojas, A. J. Pak; Mater. Horiz., Advance Article, 10.1039/d5mh01467k (2026)

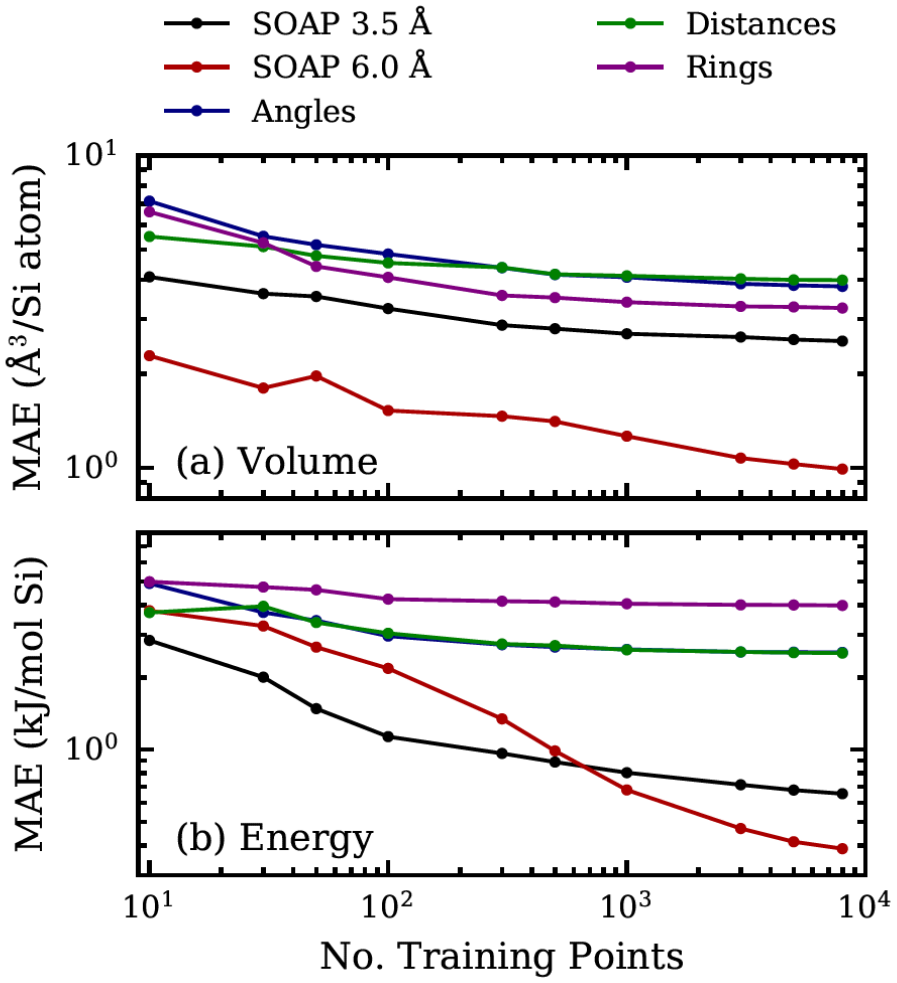
Example: predicting CH_4 uptake

The first large scale structure-property relationship (QSPR) model for MOFs, trained on void fraction and pore size to screen MOFs for CH_4 uptake via multilinear regression models, decision trees, and nonlinear support vector machines.



M. Fernandez, T. K. Woo, C. E. Wilmer, R. Q. Snurr, J. Phys. Chem. C, 117, 15, 7681–7689 (2013)

Example: energy and molar volume prediction



- Sparse Kernel Ridge Regression with Gaussian kernels
- Basis of 2000 environments chosen by FPS

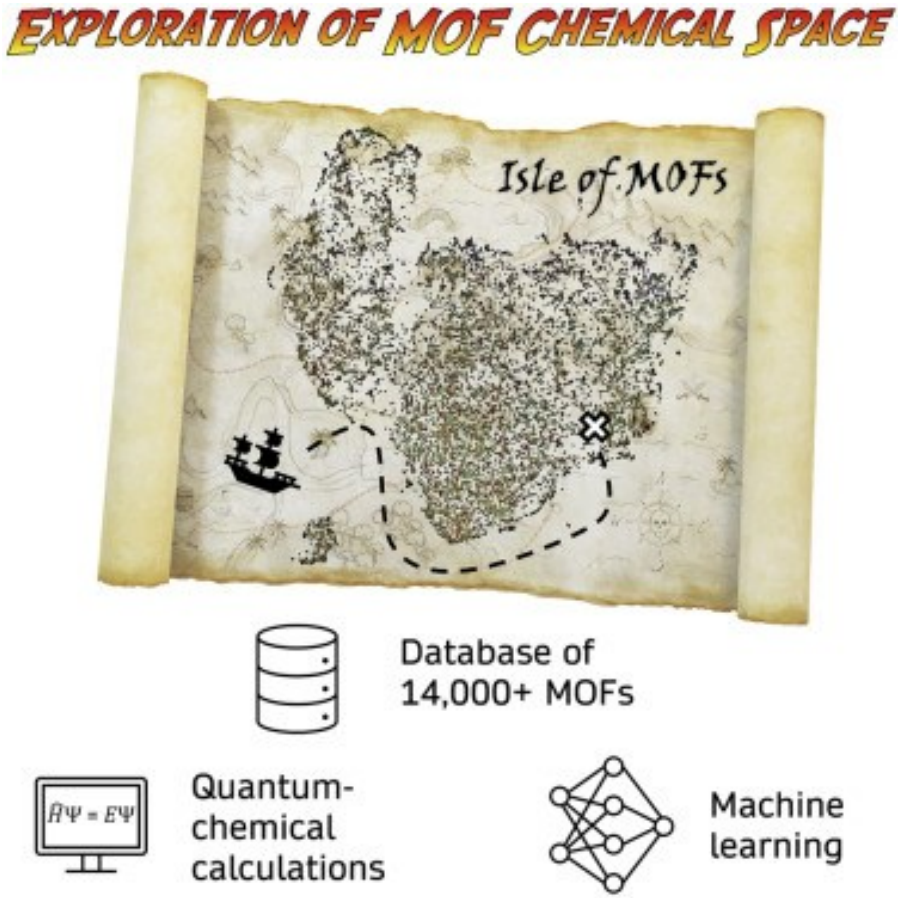
$$y(\mathcal{A}) = \sum_{\mathcal{X}\in\mathcal{A}} y(\mathcal{X}) = \sum_{\mathcal{X}\in\mathcal{A}} \sum_{\mathcal{X}_j\in M} x_j k(\mathcal{X}, \mathcal{X}_j)$$

SOAP vs. *ad hoc* descriptors:
SOAP is the best descriptor to represent structural diversity of zeolites

B. A. Helfrecht, R. Semino, G. Pireddu, S. M. Auerbach, M. Ceriotti, JCP, 151, 154112 (2019)

Example: band gap estimation

- **Bandgap Discovery:** Using CGCNN to screen databases for conductive or semiconducting MOFs



A. S. Rosen, S. M. Lyer, D. Ray, Z. Yao, A. Aspuru-Guzik, L. Gagliardi, J. M. Notestein, R. Q. Snurr, Matter, 4, 1578 – 1597 (2021)

Overview of the course

- Introduction
- Machine Learning Potentials
- Property Prediction and High-Throughput Screening
- Generative and Inverse Design
- Material Phase Classification
- Synthesis Prediction and Optimization
- Self-Driving Laboratories

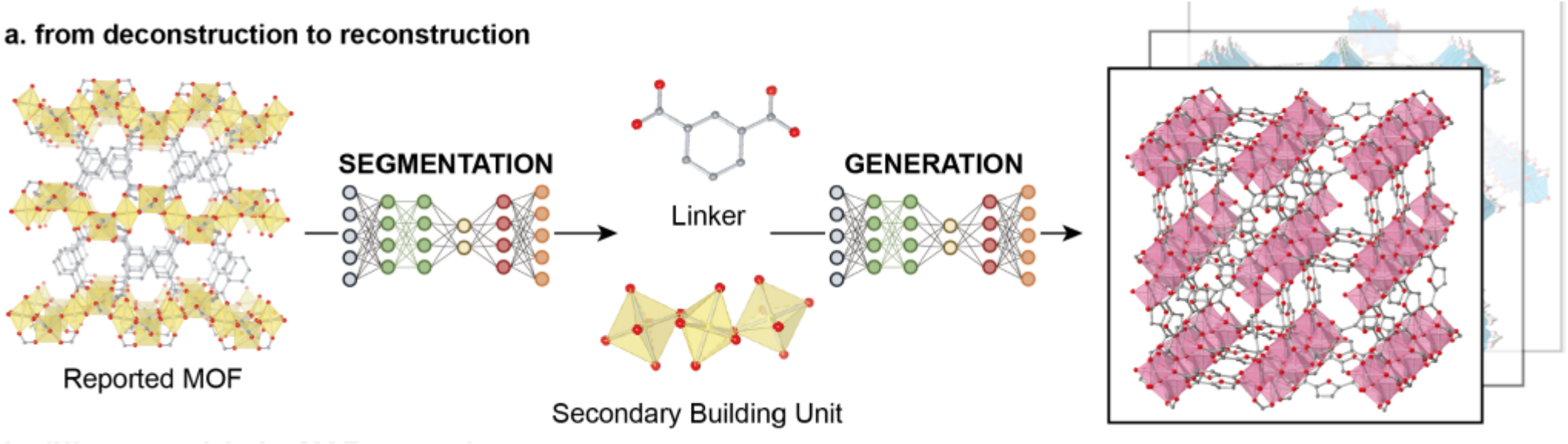
Generative and Inverse Design

- While traditional prediction moves from structure to property, **inverse design** establishes a target property value and uses a model to yield a structural design that meets it.
- Intuitively, we could take the same kind of vector of features used to predict a property and optimize it to maximize performance.

D. A. Gomez-Gualdron, T. Gercina de Vilas, K. Ardila, F. Fajardo-Rojas, A. J. Pak; Mater. Horiz., Advance Article, 10.1039/d5mh01467k (2026)

Variational autoencoders

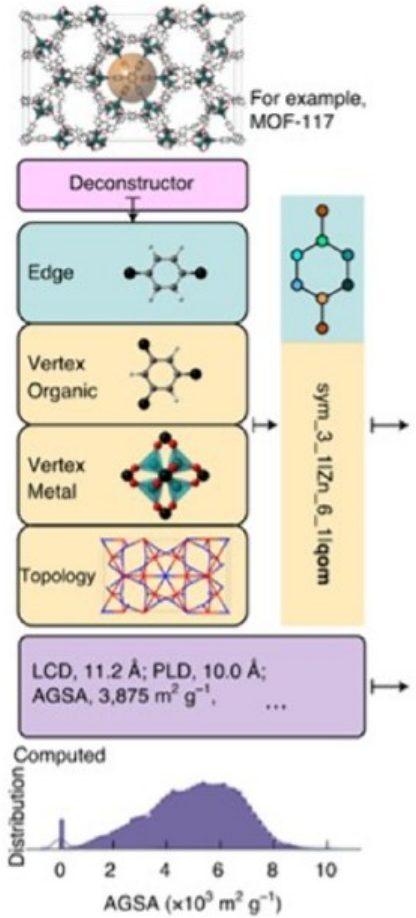
- **VAE :** an encoder to learn a continuous **latent space** representation and a decoder to reconstruct realistic structures from optimized latent vectors

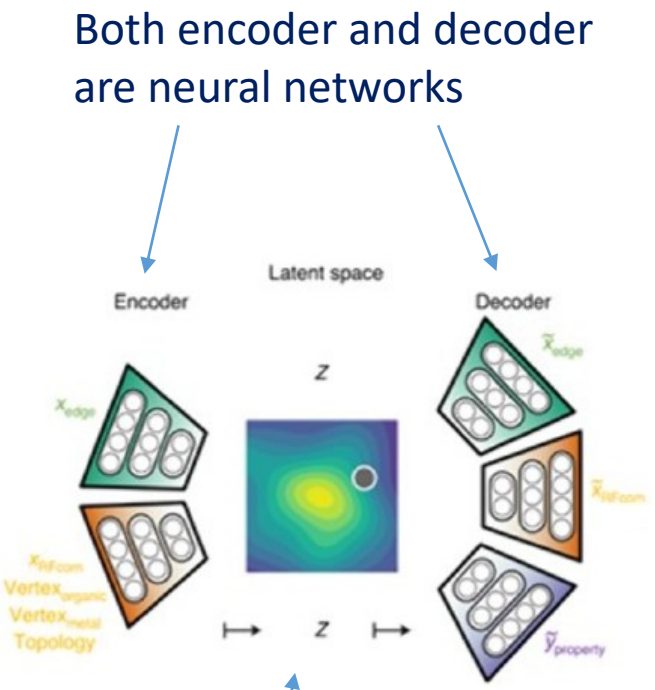


C. Duan et al, arXiv: 2508.13197v1 (2025)

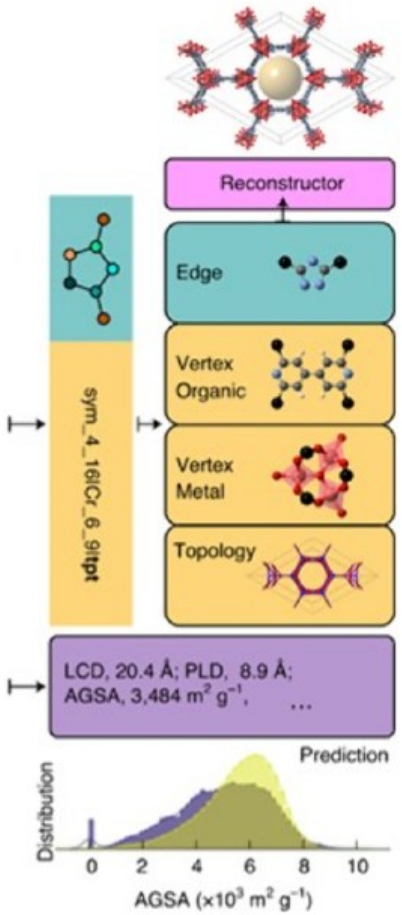
Example: SmVAE

- **Variational Autoencoder** + an additional model that predicts properties from latent vectors
- SmVAE: used to predict CO_2 uptake





Latent space: MOFs are represented by topology and building blocks.

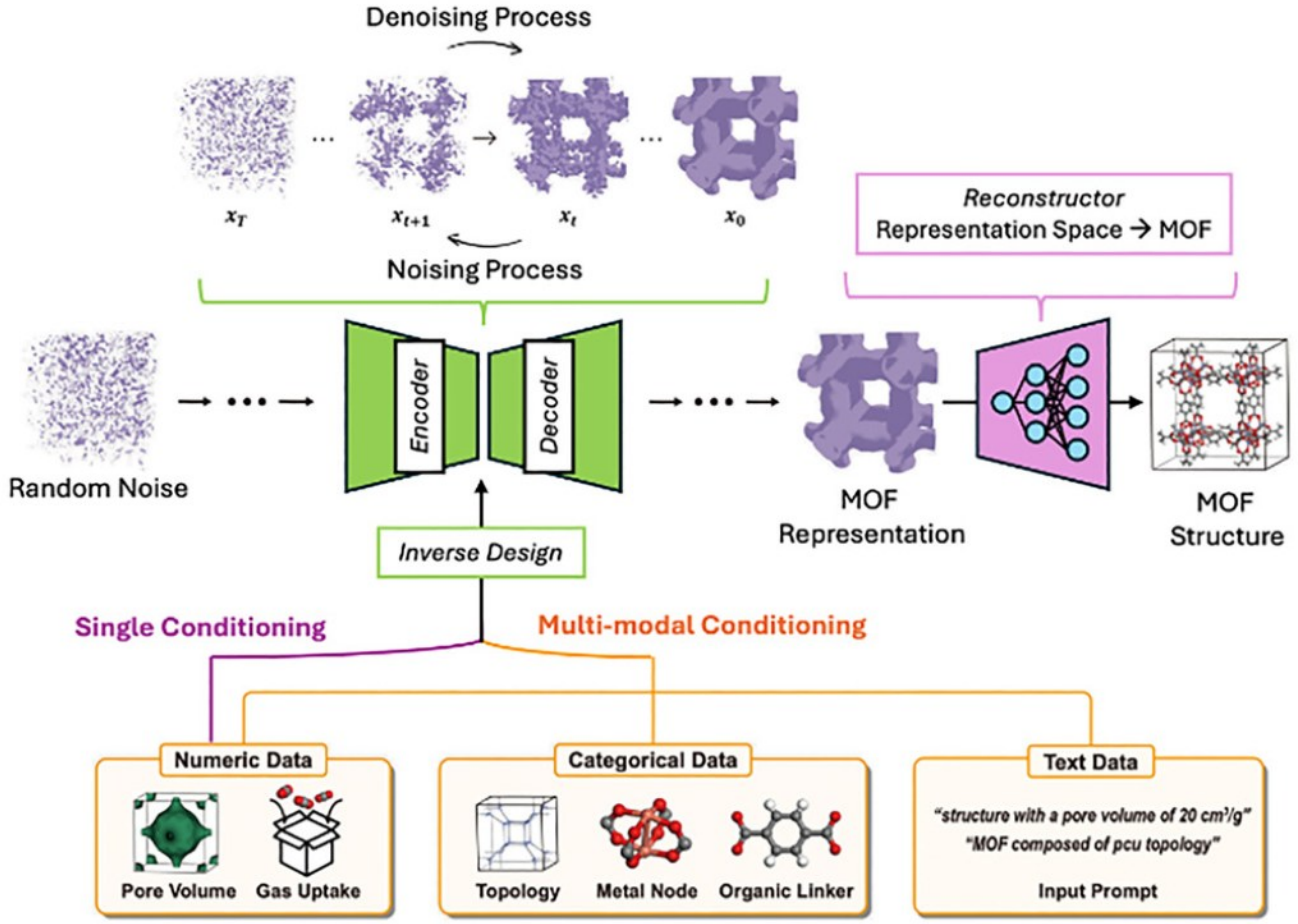


Z. Yao *et al*, Nature Machine Intelligence volume 3, pages 76–86 (2021)

Diffusion models

Denoising Diffusion Probabilistic Models: Learning to reverse a denoising process to recover a MOF

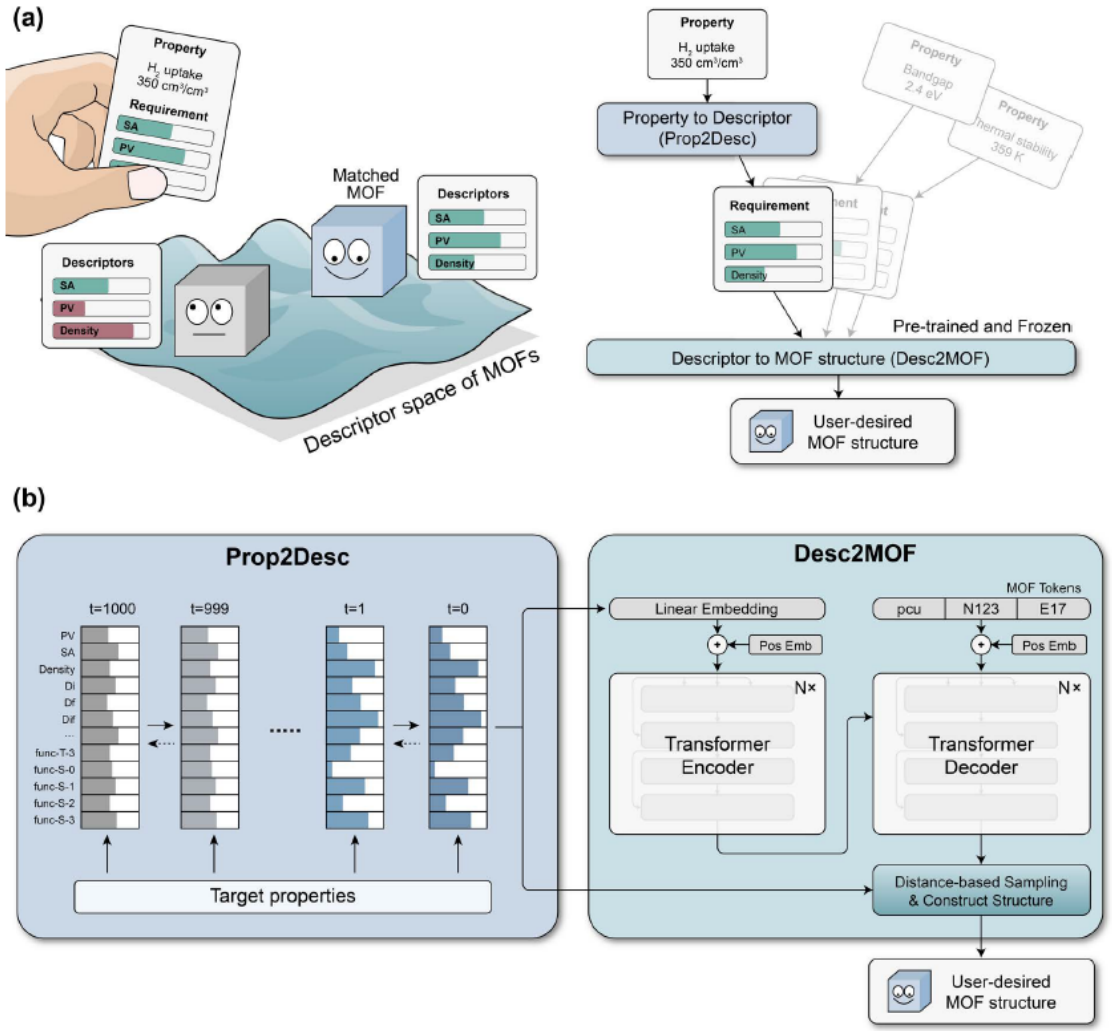
- To train it, we randomly add (Gaussian) noise to MOFs and then train the denoiser (a neural network)
- Once trained, the learning process can be **conditioned** on a property of interest



J. Park, Y. Lee and J. Kim, Nat. Commun., 16, 34 (2025)

Example: EGMOF

- A hybrid model that generates chemically informed descriptors via diffusion and translates them into MOF structures via a transformer for high-performance hydrogen storage
- Prop2Desc is a diffusion model that generates a descriptor intermediate to structure and property
- Desc2MOF a transformer that predicts MOF structure from the descriptor.



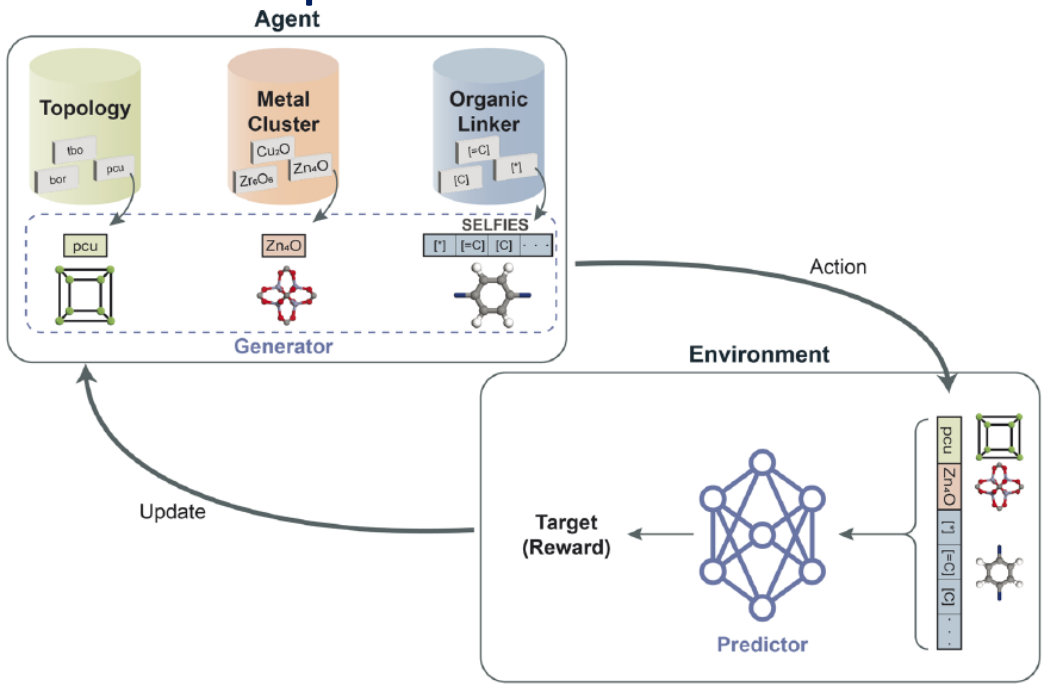
S. Han, Y. Kang, T. Bae, V. Bernales, A. Aspuru-Guzik, J. Kim, arXiv: 2511.03122v1

Reinforcement learning & Example

A reinforcement learning framework used to generate MOFs with unprecedented CO_2/H_2O selectivity for capturing CO_2 from humid air

An agent learns a "policy" to generate structural strings (SMILES/SELFIES) that maximize a reward function based on desired performance metrics

The agent (generator) generates a MOF structure as an action. The environment (predictor) evaluates the action by predicting the property of the new MOF structure and returns a reward as an update to the agent. The agent then generates the next round of MOFs based on the received reward.



H. Park, S. Majumdar, X. Zhang, J. Kim, B. Smit, Digital Discovery, 3, 728–741 (2024)

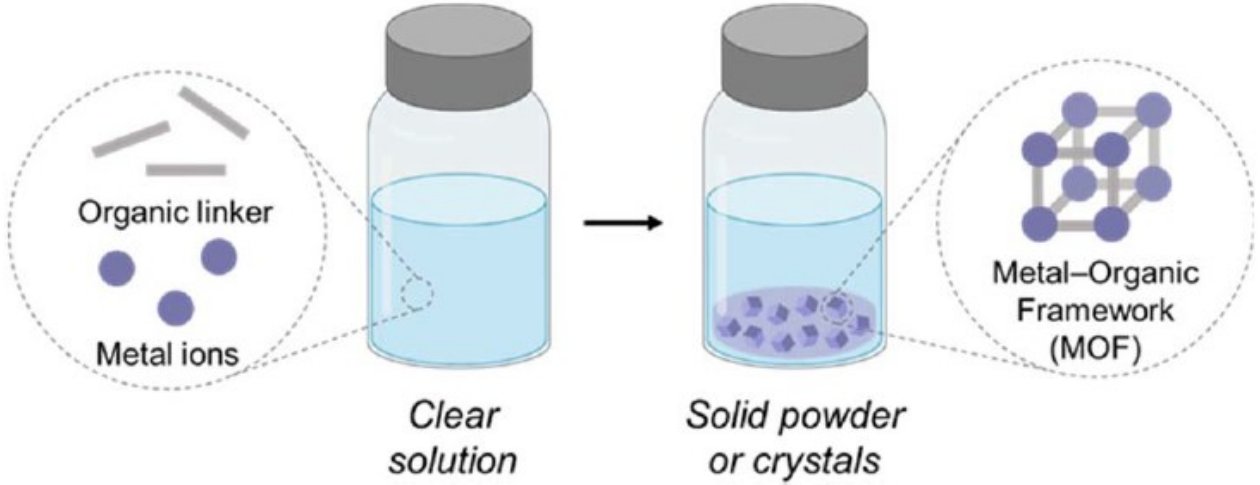
Overview of the course

- Introduction
- Machine Learning Potentials
- Property Prediction and High-Throughput Screening
- Generative and Inverse Design
- Material Phase Classification
- Synthesis Prediction and Optimization
- Self-Driving Laboratories

Introduction to Material Phase Classification

Classification identifies the physical or chemical state of a material (e.g., solid, liquid, residue, different polymorphs)

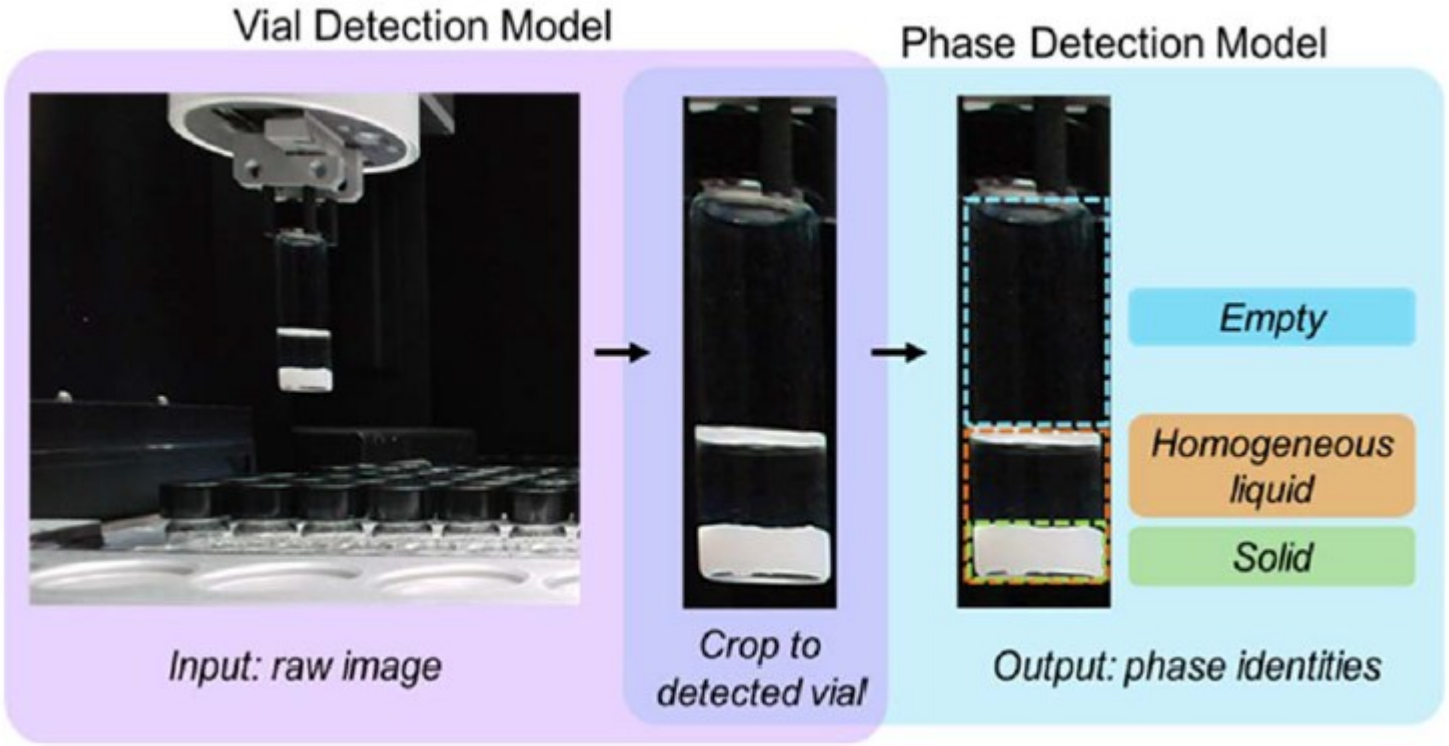
- To automatically monitor reaction progress and identify successful synthesis outcomes (high throughput synthesis)
- To analyze simulation trajectories and identify complex structural patterns



M. A. Gaidimas et al, Digital Discovery (2025), doi:10.1039/D5DD00384A

Computer vision

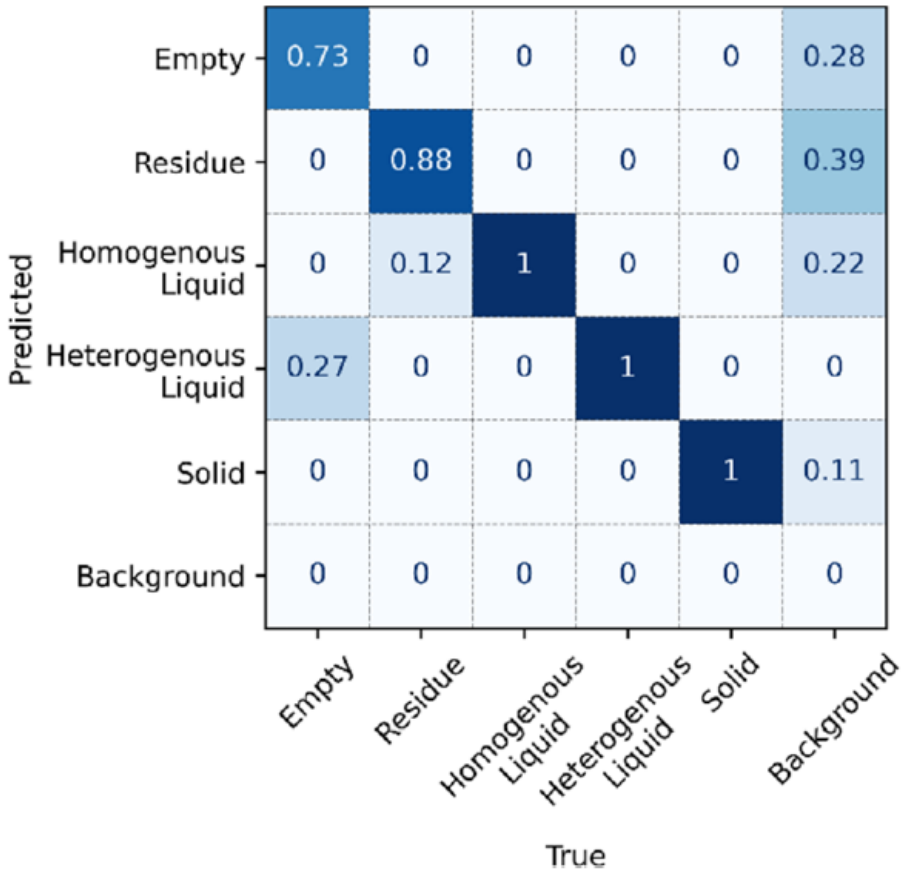
Computer Vision: object detection models like “you only look once” (**YOLO)** to rapidly classify material phases in sample vials from digital images. There are also hierarchical versions in which the vessel and phase are detected separately:

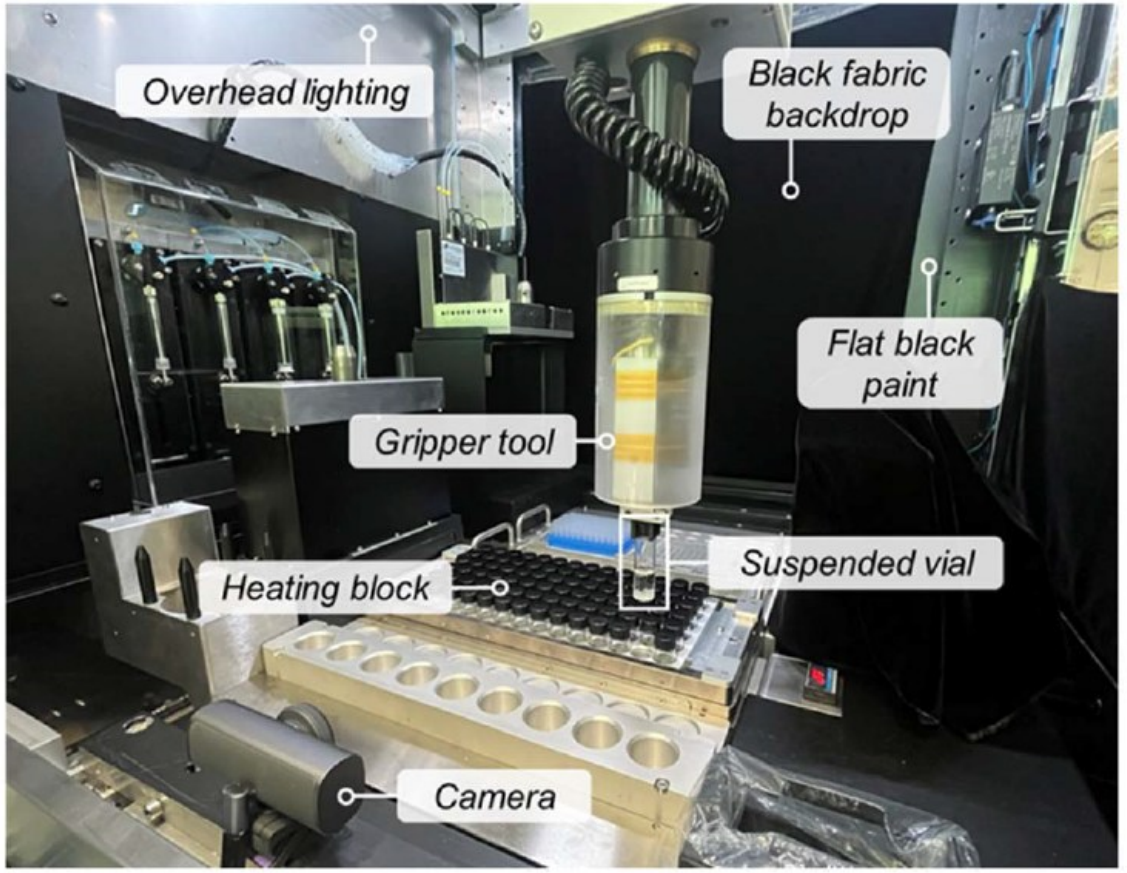


M. A. Gaidimas et al, Digital Discovery (2025), doi:10.1039/D5DD00384A

Example: monitoring MOF crystallization in real-time

A computer vision pipeline that classifies vial contents into five categories coupled with an automated synthesis system.

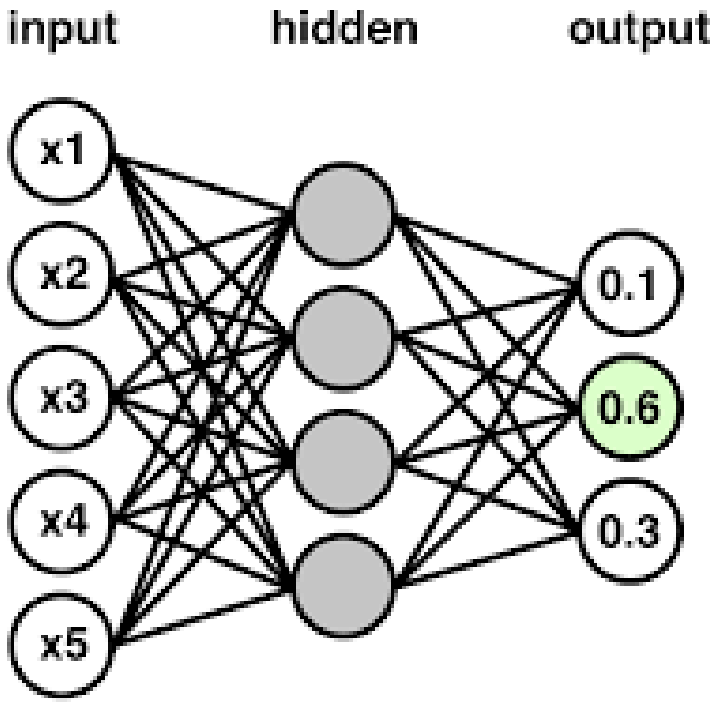




M. A. Gaidimas et al, Digital Discovery (2025), doi:10.1039/D5DD00384A

Neural network classifiers

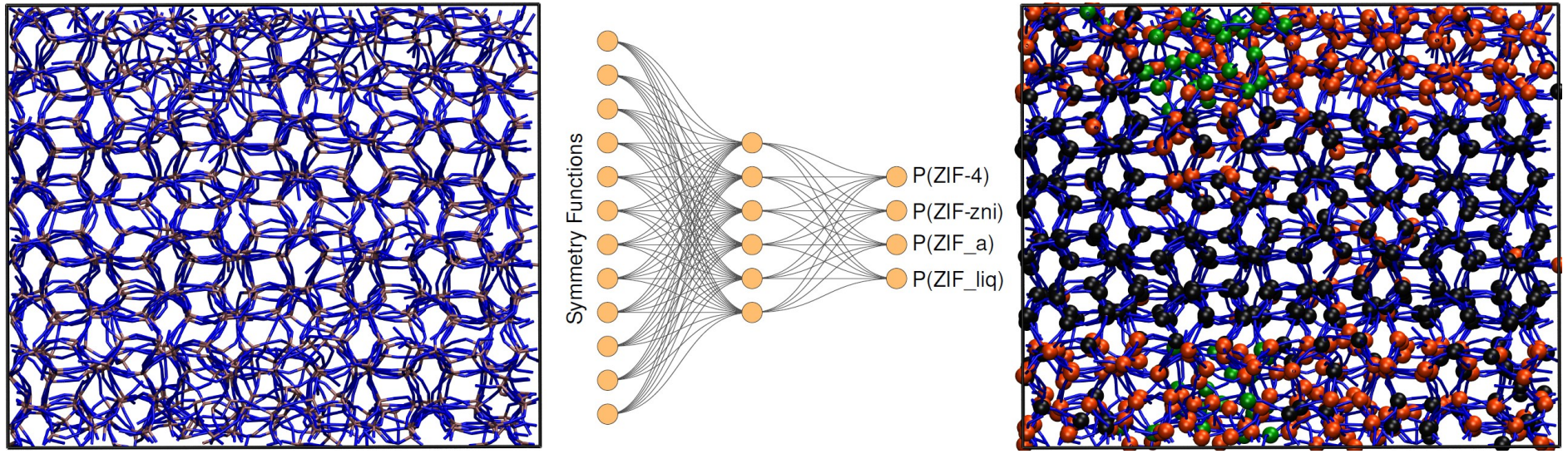
A neural network that outputs probability of belonging to each of a series of pre-defined classes:



Example: monitoring MOF phase transitions

$$G_i^{atom,rad} = \sum_{j=1}^{N_{atom}} \left(e^{-\eta(R_{ij}-R_s)^2} \times f_c(R_{ij}) \right)$$

$$G_i^{atom,ang} = 2^{1-\zeta} \sum_{j,k\neq i}^{all} \left((1+\lambda\cos\theta_{ijk})^{\zeta} \times e^{-\eta(R_{ij}^2+R_{ik}^2+R_{jk}^2)^2} \times f_c(R_{ij}) \times f_c(R_{ik}) \times f_c(R_{jk}) \right)$$



E. Méndez & R. Semino; J. Mater. Chem. A, **12**, 4572 – 4582 (2024)

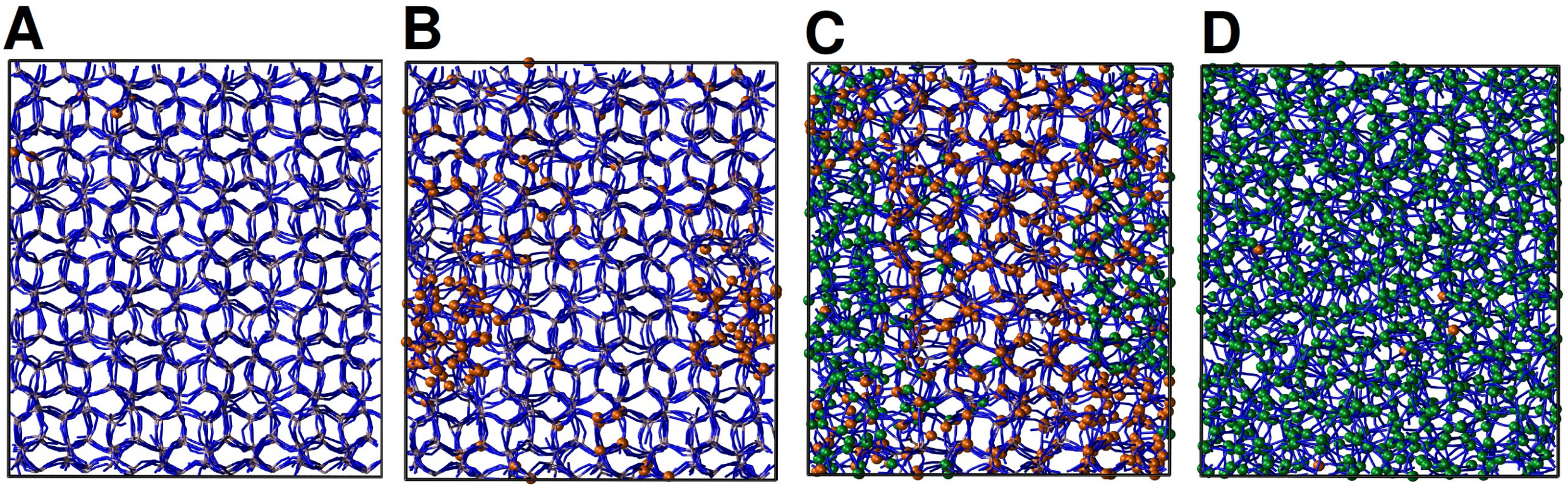
Example: monitoring MOF phase transitions

	ZIF-4	ZIF-zni	ZIF_a	ZIF_liq
ZIF-4	**0.951**	0.000	0.004	0.045
ZIF-zni	0.000	**0.946**	0.044	0.010
ZIF_a	0.000	0.019	**0.895**	0.086
ZIF_liq	0.012	0.002	0.098	**0.888**

90% accuracy with only structure-based information!

E. Méndez & R. Semino; J. Mater. Chem. A, **12**, 4572 – 4582 (2024)

Example: monitoring MOF phase transitions



E. Méndez & R. Semino; J. Mater. Chem. A, **12**, 4572 – 4582 (2024)

Overview of the course

- Introduction
- Machine Learning Potentials
- Property Prediction and High-Throughput Screening
- Generative and Inverse Design
- Material Phase Classification
- Synthesis Prediction and Optimization
- Self-Driving Laboratories

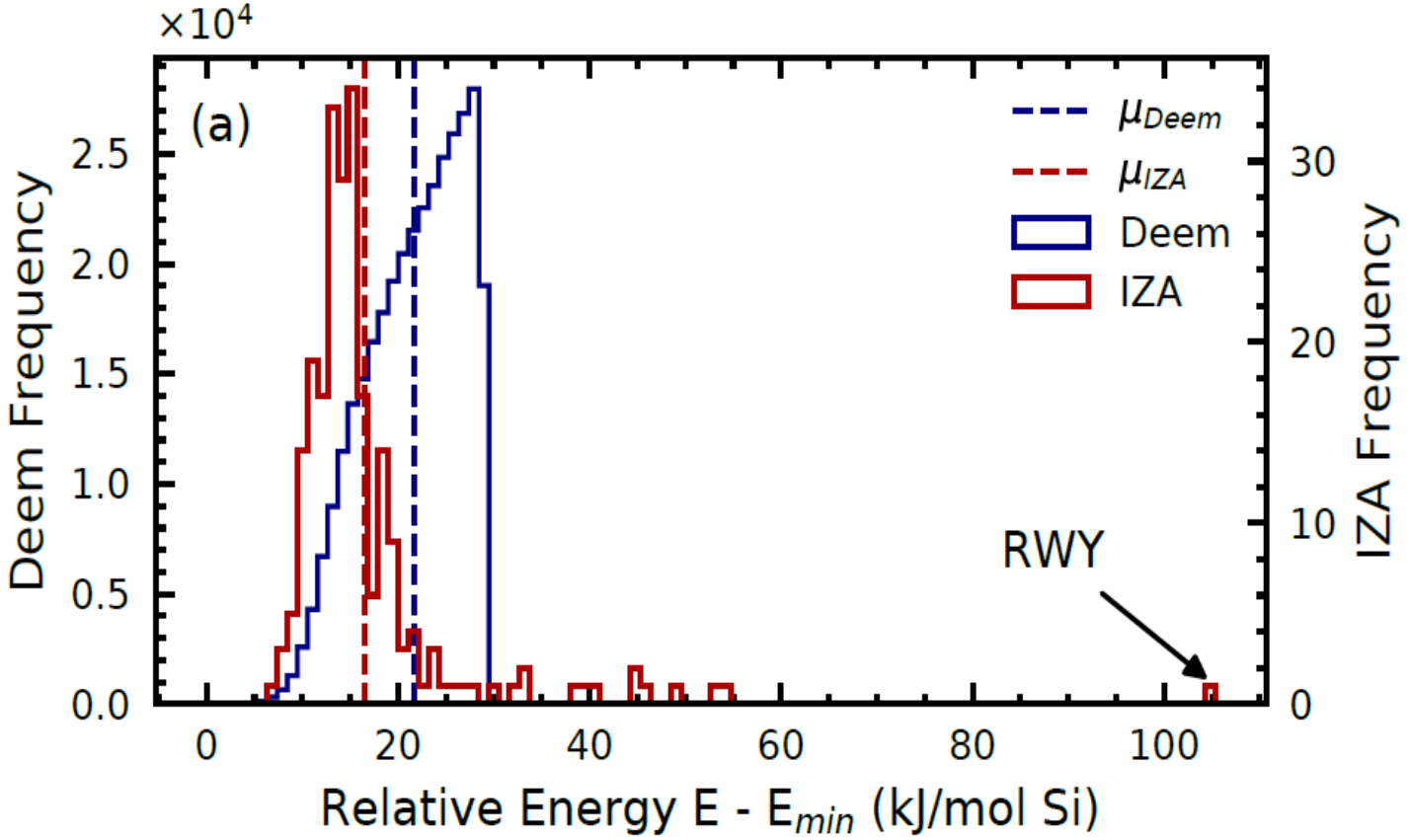
Introduction to Synthesis Prediction and Optimization

"Synthesizability" gap:

The zeolite conundrum:

- Real zeolites (IZA) **--> 247**
- Deem hypothetical zeolites database **--> 2.6 M**

None of them yet synthesized

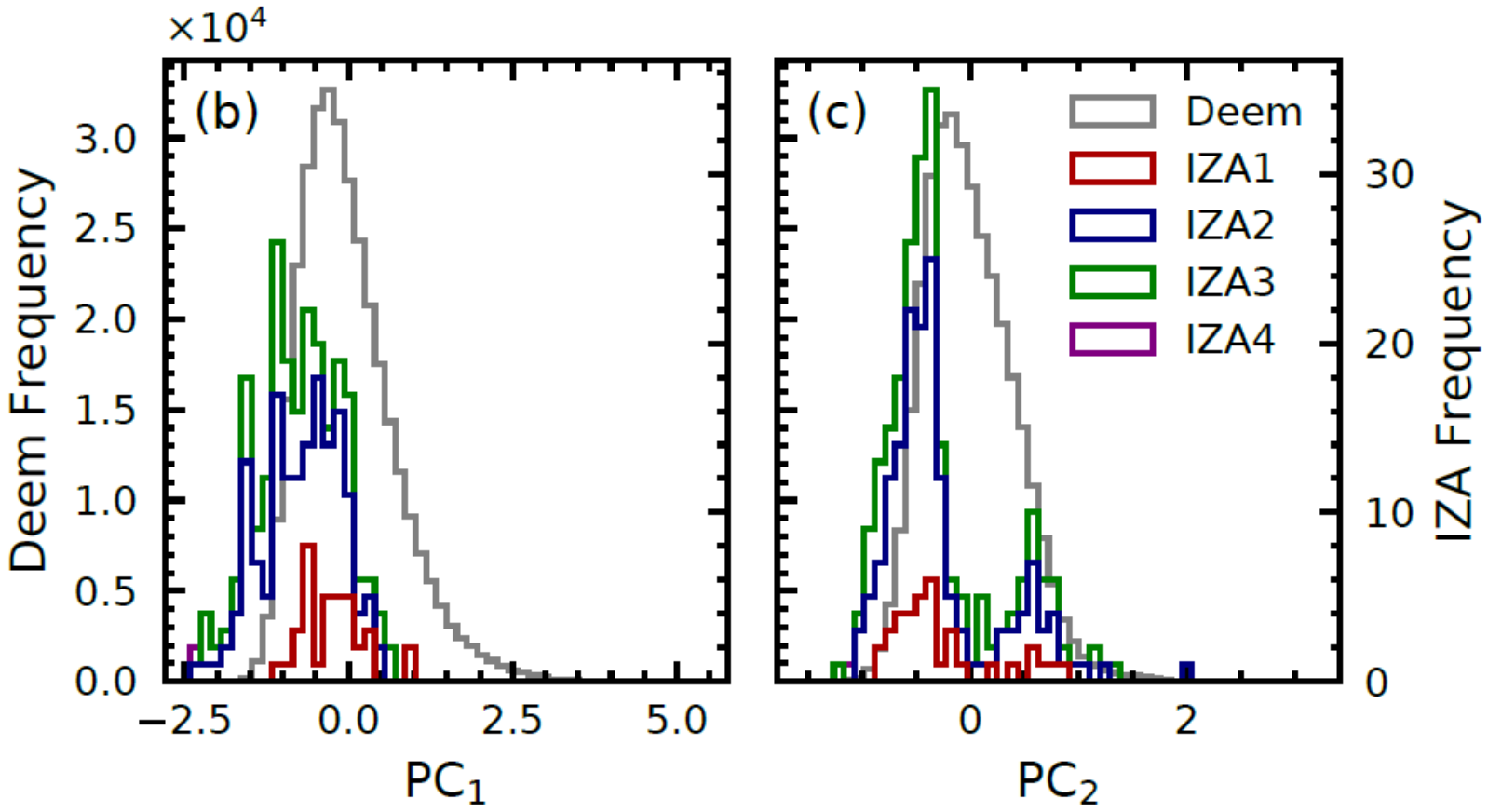


Predicting the likelihood of successful crystallization and optimizing synthesis conditions

R. Pophale *et al*, PhysChemChemPhys, 13, 12407 (2011)

B. A. Helfrecht *et al*; Digit. Disc., 1, 779-789 (2022)

Introduction to Synthesis Prediction and Optimization



IZA1: Si + O

IZA2: Si + O + Other T atoms

IZA3: O + Other T atoms

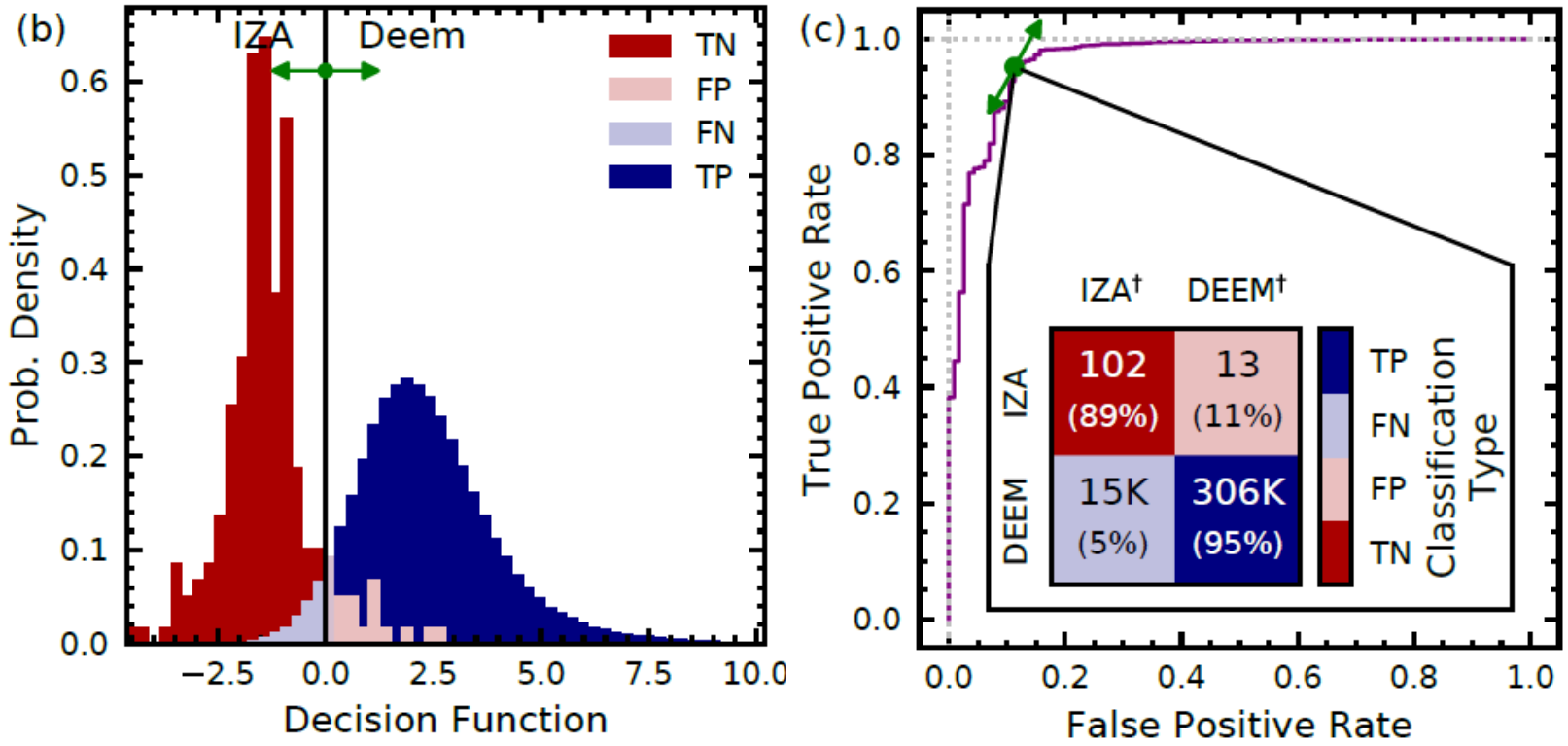
IZA4: RWY

Nor energy nor unsupervised clustering allow to distinguish known zeolites from hypothetical ones...

B. A. Helfrecht *et al*; Digit. Disc., 1, 779-789 (2022)

Example: synthesizability of hypothetical zeolites

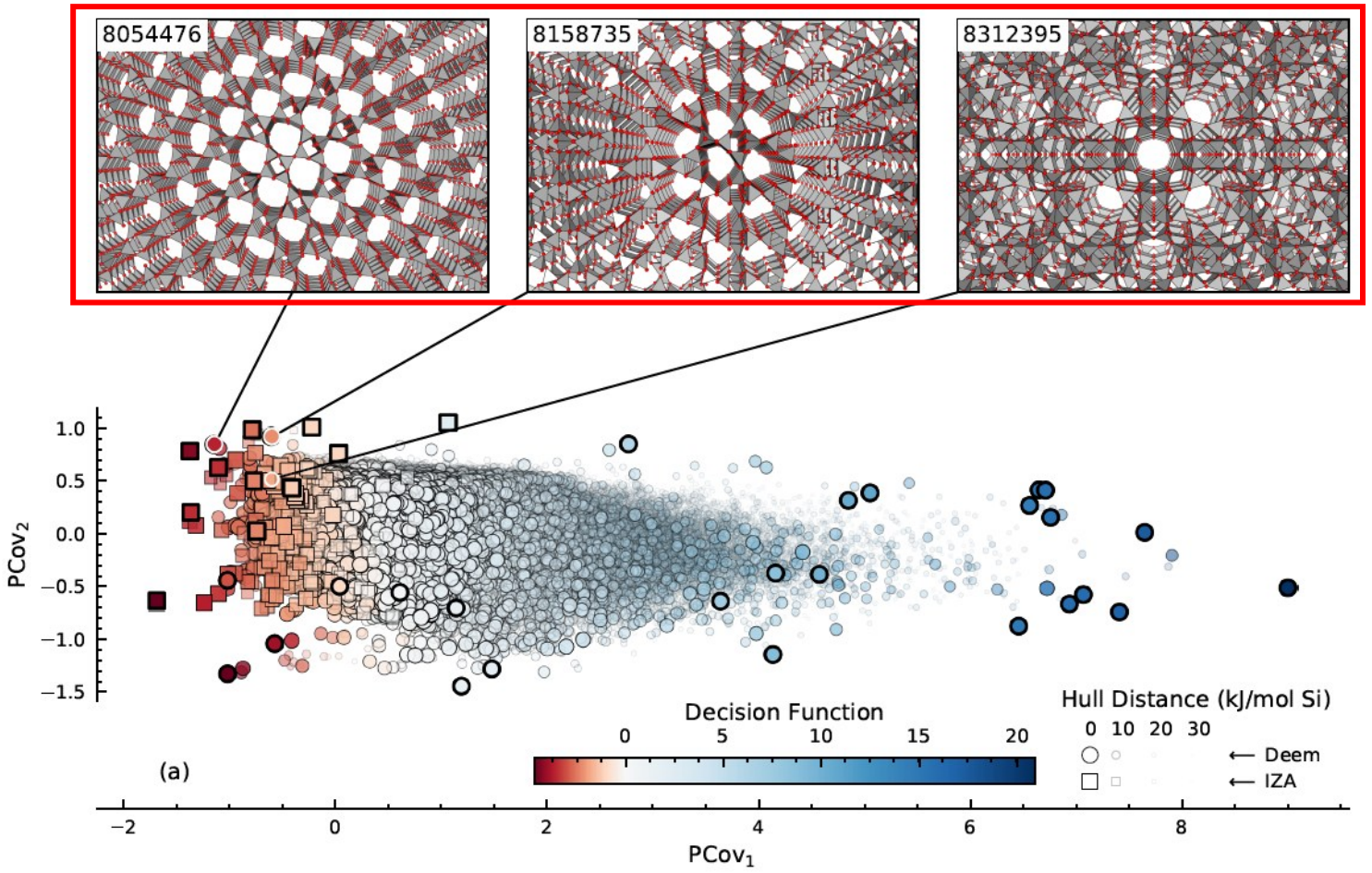
Supervised learning: Support Vector Machine with a class-balanced classification scheme



B. A. Helfrecht *et al*; Digit. Disc., 1, 779-789 (2022)

Example: synthesizability of hypothetical zeolites

SVM combined with lattice energy to yield a synthesizability map:



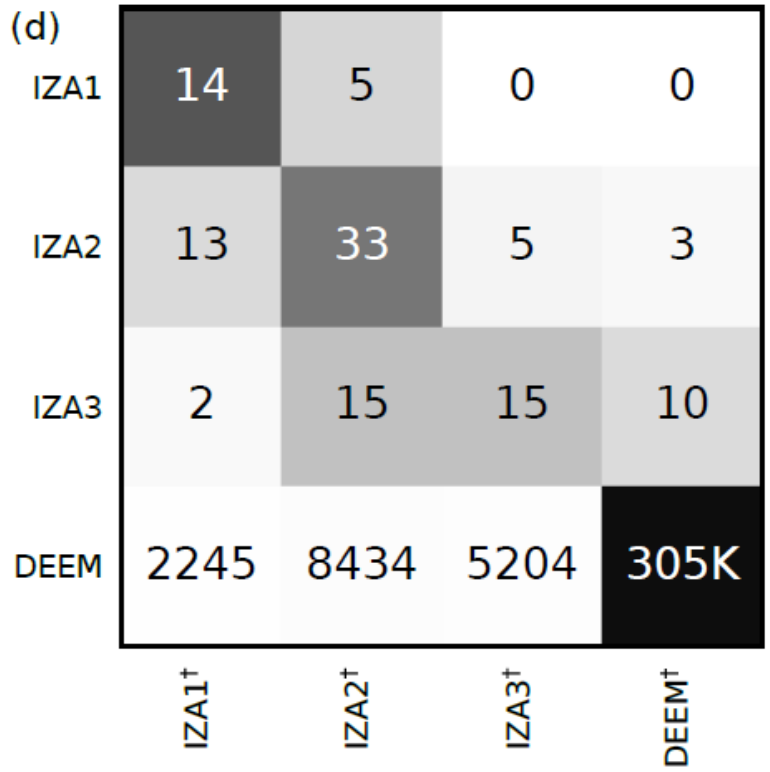
Three good candidates in the 55–60, 60–65, and 65–70 A^3/Si volume ranges respectively

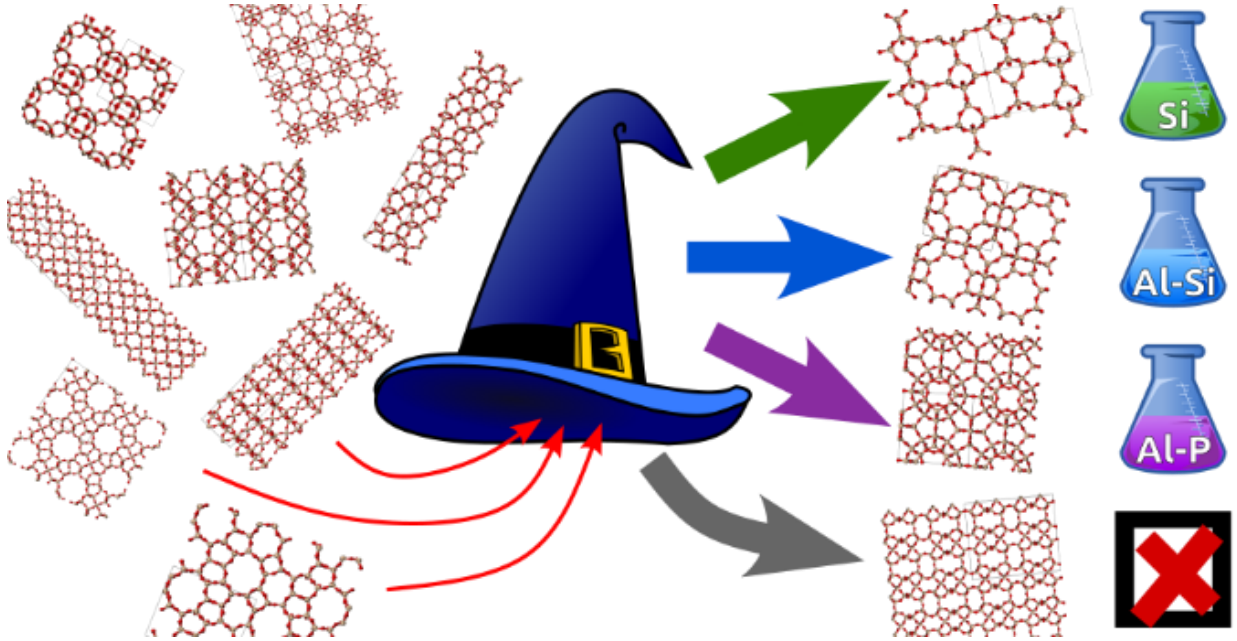
- Circles = DEEM
- Red = classified as IZA
- Large = more stable

B. A. Helfrecht *et al*; Digit. Disc., 1, 779-789 (2022)

Example: synthesizability of hypothetical zeolites

SVM combined with lattice energy to yield a synthesizability map:





Guidance for synthesis conditions exploration

B. A. Helfrecht *et al*; Digit. Disc., 1, 779-789 (2022)

Example 2: Synthesizability classification

A binary ML model developed through GPT-4 that predicts whether a given set of precursors and synthesis conditions will form a crystalline MOF with >90% accuracy:

ChemPrompt Engineering

Prompt:
Answer the question as truthfully as possible using the provided context.

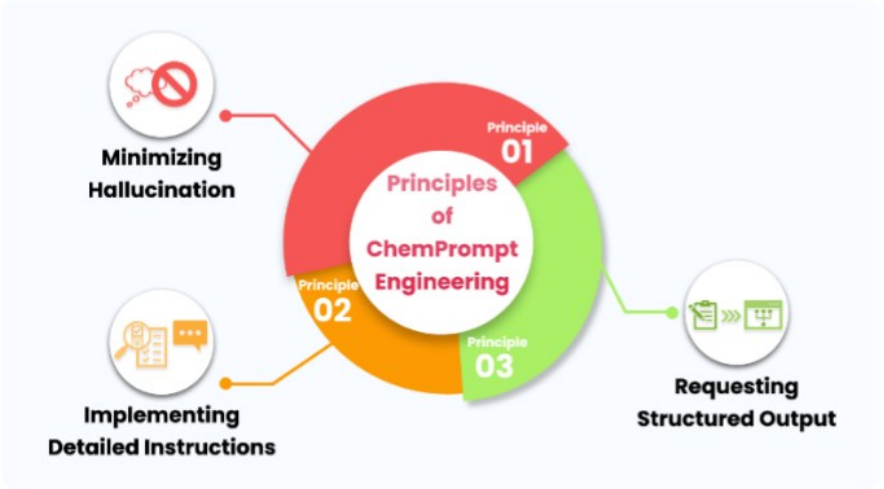
Please summarize the following details in a table: compound name or chemical formula (if the name is not provided), metal source, organic linker(s), solvent(s), reaction temperature, and reaction time. If any information is not provided or you are unsure, use "N/A". Please ignore information related to organic linker synthesis, MOF postsynthetic modification or metalation.

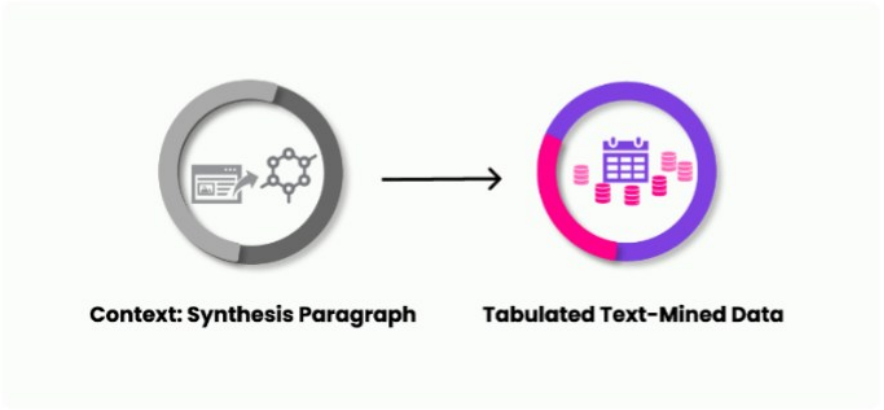
The table should have 6 columns, all in lowercase:| compound name | metal source | linker| solvent | reaction temperature | reaction time |

Input:
In a 100 mL media bottle were dissolved 1,3,5-benzenetricarboxylic acid (210 mg) and ZrOCl2·8H2O (970 mg) in a solution containing DMF (30 mL) and formic acid (30 mL). The bottle was sealed and heated in a 100 °C isothermal oven for a day. White powder of MOF-808 was collected by centrifugation.

Output:

compound name	metal source	linker	solvent	reaction temperature	reaction time
MOF-808	ZrOCl2·8H2O	1,3,5-benzenetricarboxylic acid	DMF	100 °C	24 h





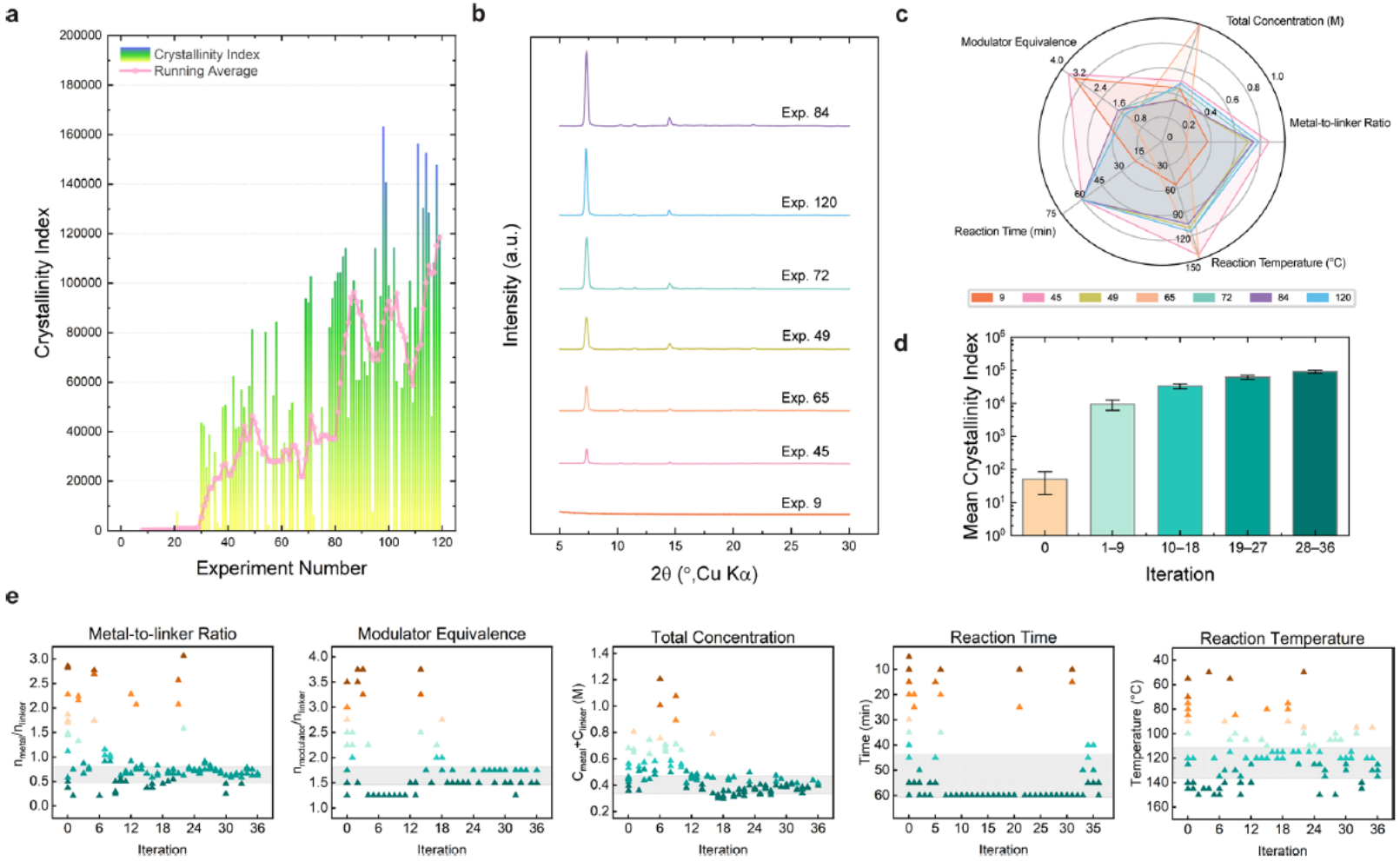
Zheng, Z., Zhang, O., Borgs, C., Chayes, J. T. & Yaghi, O. M. *J. Am. Chem. Soc.* 145, 18048–18062 (2023)

Example 3: Agentic AI

Multi-agent systems where different LLMs handle specialized tasks

Z. Zheng, O. Zhang, H. L. Nguyen, N. Rampal, A. H. Alawadhi, Z. Rong, T. Head-Gordon, C. Borgs, J. T. Chayes, O. M. Yaghi, *ACS Cent. Sci.*, 9, 2161−2170 (2023)

Methodology: Agentic AI



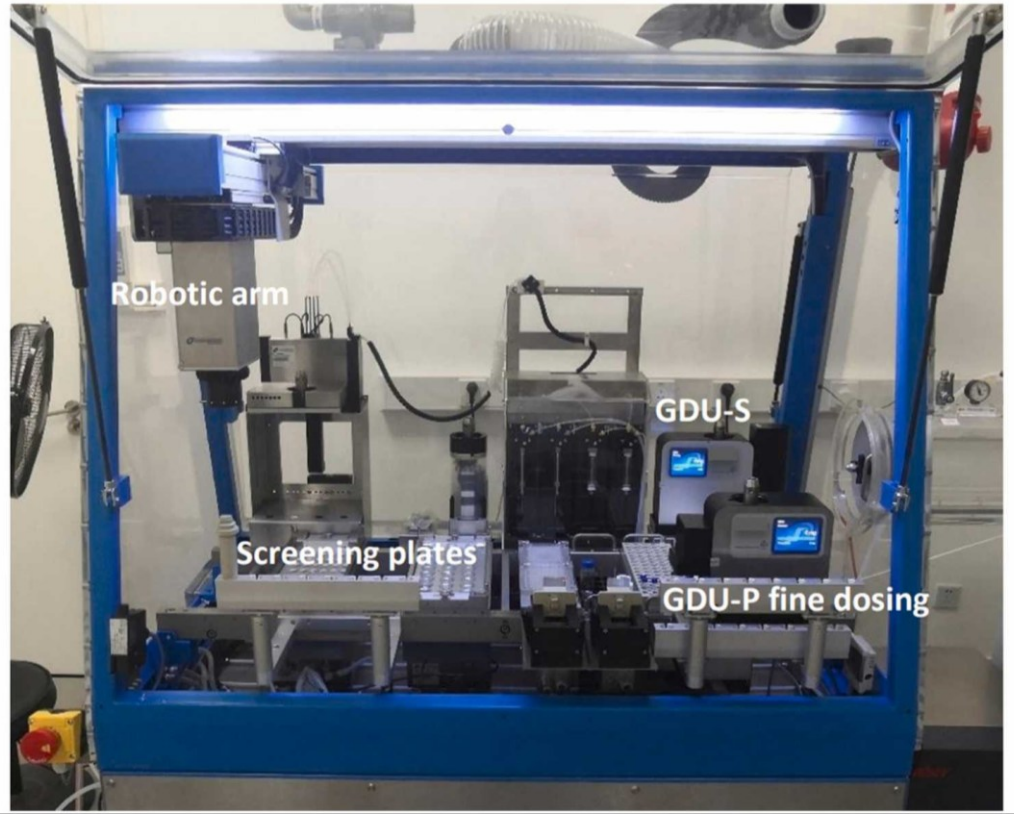
Z. Zheng, O. Zhang, H. L. Nguyen, N. Rampal, A. H. Alawadhi, Z. Rong, T. Head-Gordon, C. Borgs, J. T. Chayes, O. M. Yaghi, *ACS Cent. Sci.*, 9, 2161−2170 (2023)

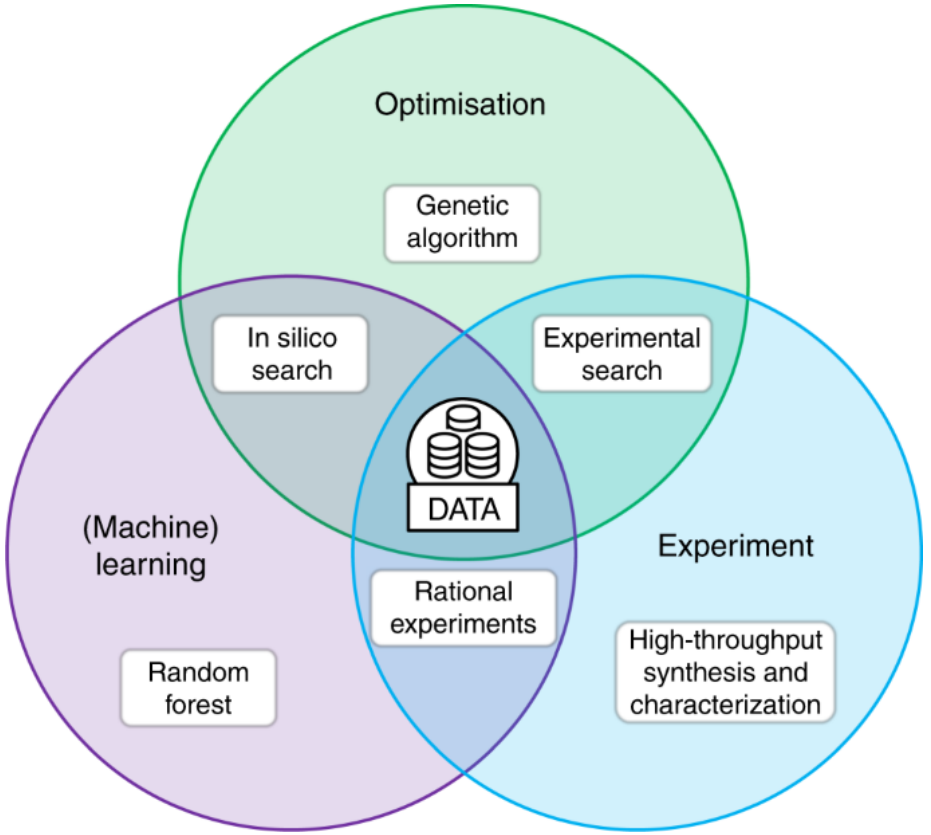
Overview of the course

- Introduction
- Machine Learning Potentials
- Property Prediction and High-Throughput Screening
- Generative and Inverse Design
- Material Phase Classification
- Synthesis Prediction and Optimization
- Self-Driving Laboratories

Self-Driving Laboratories

Self-Driving Laboratories integrate AI-driven decision-making with automated hardware to conduct scientific discovery with minimal human intervention.





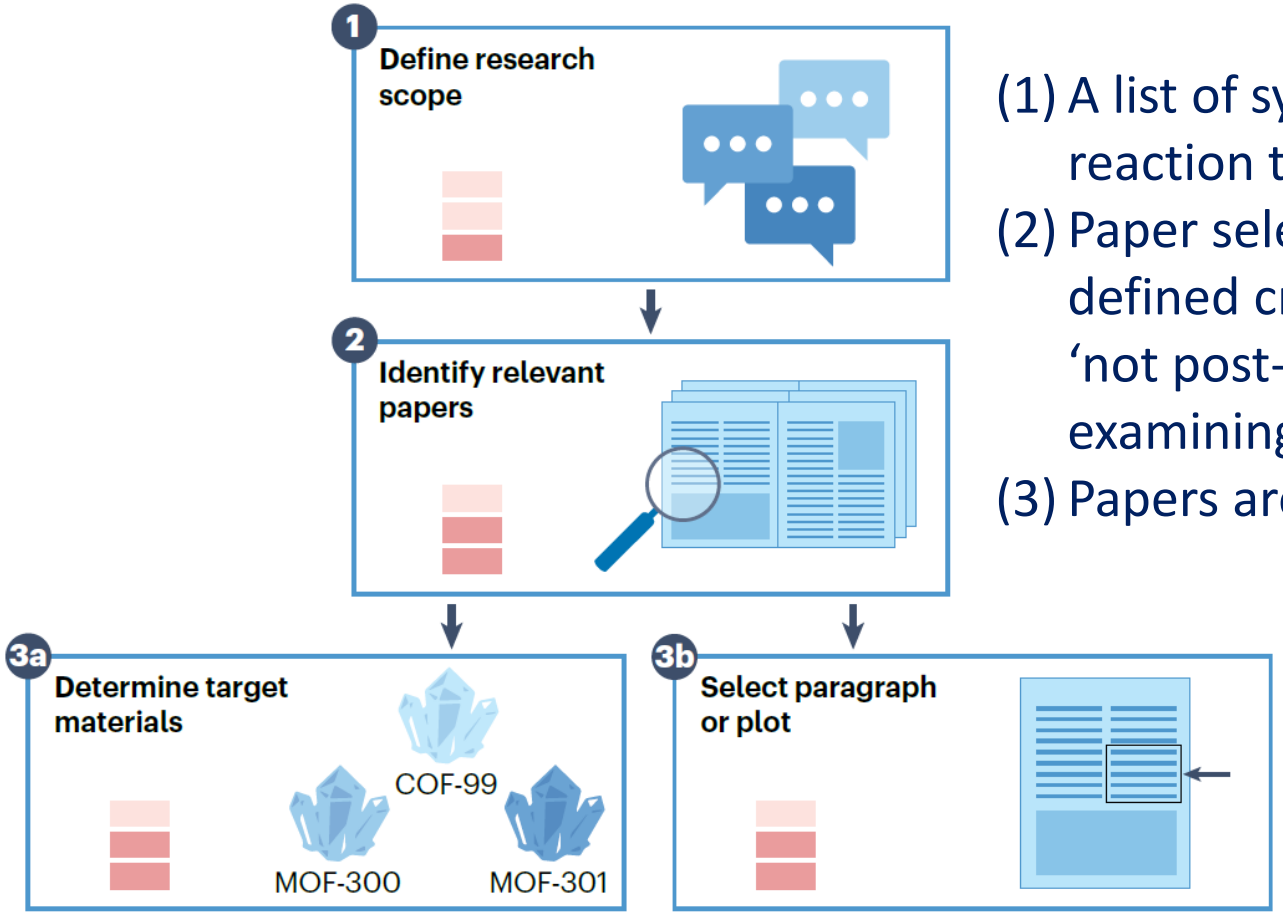
J.-M. Lu, J.-Z. Pan, Y.-M. Mo, Q. Fang, Artificial Intelligence Chemistry, 2, 100057 (2024)
Y. Zhao, Y. Zhao, J. Wang, Z. Wang, Ind. Eng. Chem. Res., 64, 4637−4668 (2025)

Self-Driving Laboratories

G. Tom *et al*, Chem. Rev., 124, 9633−9732 (2024)

Natural Language Processing

LLMs and rule-based extractors to mine data (synthesis conditions, property values) from thousands of scientific papers

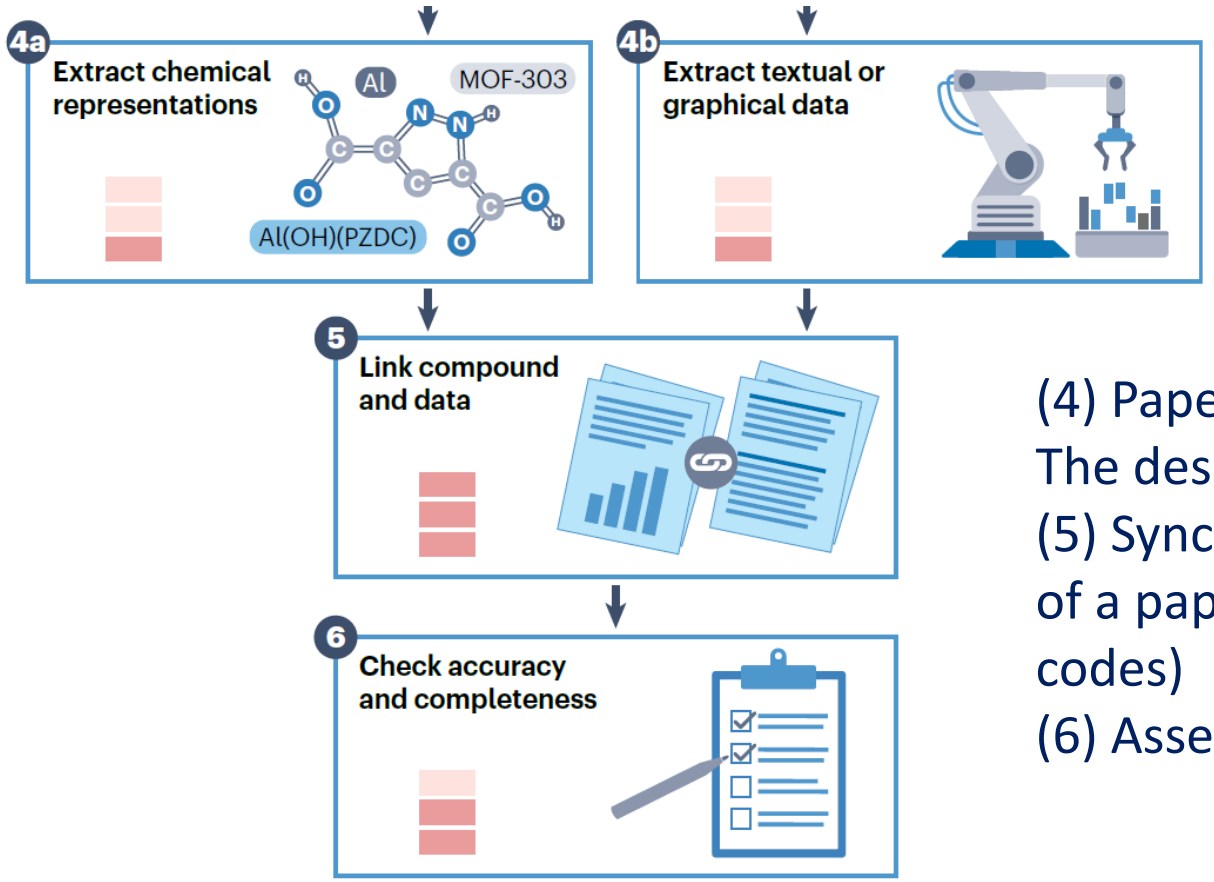


(1) A list of synthesis conditions including metal, linker, solvent, reaction time and temperature is created.
(2) Paper selection using an LLM trained to recognize user-defined criteria (for example, 'MOF synthesis', 'experimental', 'not post-synthetic modification', 'not a review paper') by examining titles and abstracts in a library of papers.
(3) Papers are fed into the model to extract specific parameters.

Zheng, Z., Rampal, N., Inizan, T.J., Borgs, C., Chayes, J. T, Yaghi, O. M., *Nat Rev Mater,* 10, 369–381 (2025)

Natural Language Processing

LLMs and rule-based extractors to mine data (synthesis conditions, property values) from thousands of scientific papers



(4) Papers are fed into the model to extract specific parameters The desired output format is specified in the prompt.
(5) Synchronization of data that is located in different sections of a paper (abbreviations, general procedures, reference codes)
(6) Assessment of LLM performance.

Zheng, Z., Rampal, N., Inizan, T.J., Borgs, C., Chayes, J. T, Yaghi, O. M., *Nat Rev Mater,* 10, 369–381 (2025)