Machine Learning in Molecular Sciences

My Background

2005 - 2009
B.S.
Chemical Engineering

2010 - 2015
Ph.D.
Chemical Engineering

2015 - 2018
Postdoctoral Training
Molecular Eng.

2019 - present
Assistant Professor
Chemical & Biomol. Eng.

Undergraduate Research at ND

Fulbright Scholarship, Spain, 2009

Born and raised in Puerto Rico

Research: Computational Molecular Engineering

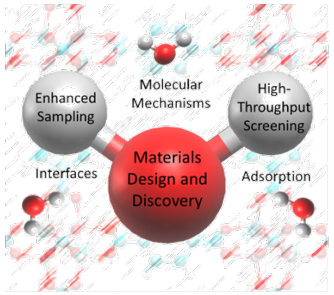
Research Methods

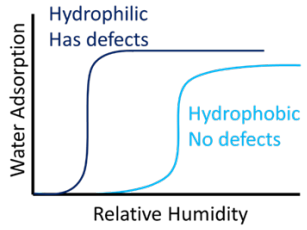
- Molecular Modeling
- Statistical Mechanics
- Materials Science
- Machine Learning and Data Science

Application Areas

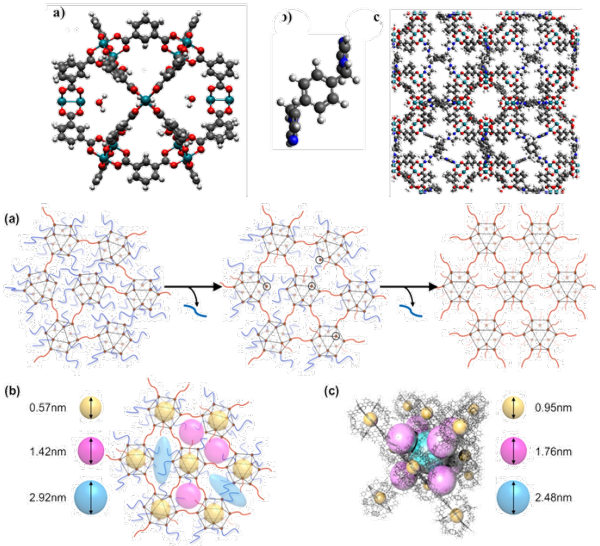
- Energy Storage
- Gas adsorption and Separations
- Molecular Design of Porous Materials
- Water farming, Water security
- Physical Properties of Novel Materials
- Materials Design and Discovery
- Self-Assembly of Porous Materials
- Interfacial Engineering

Water-Energy Nexus

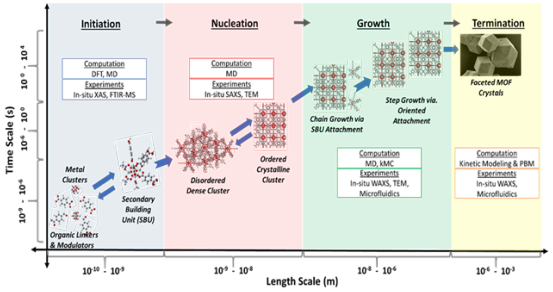


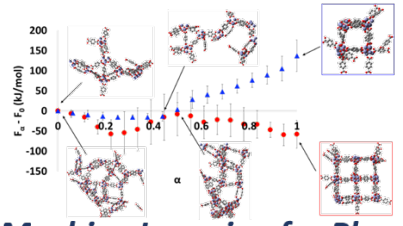


Soft Porous Coordination Polymers

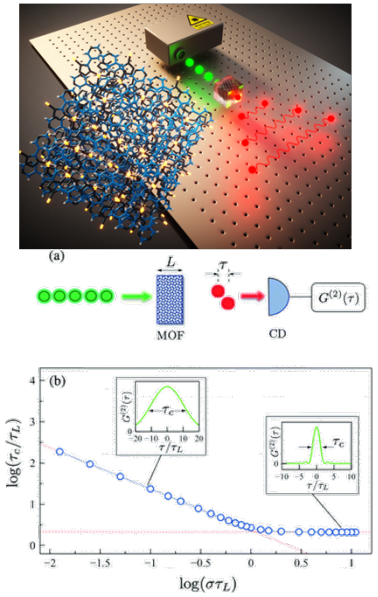


Self-Assembly of Porous Materials

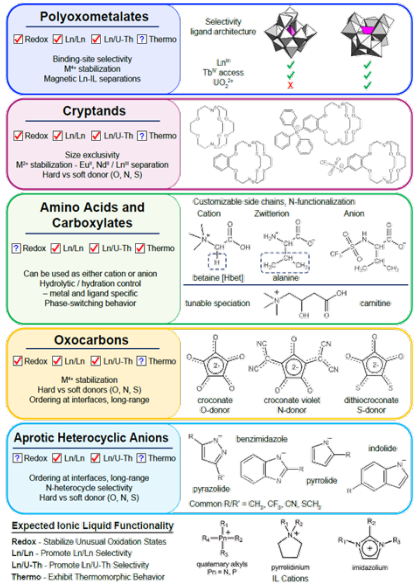




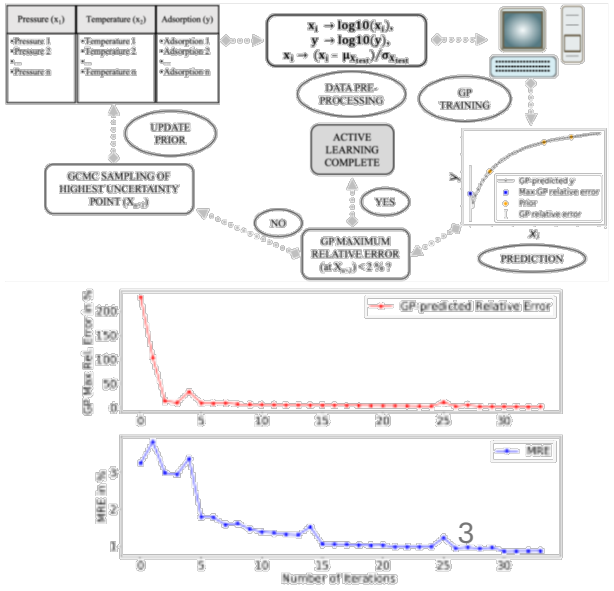
Quantum Technologies



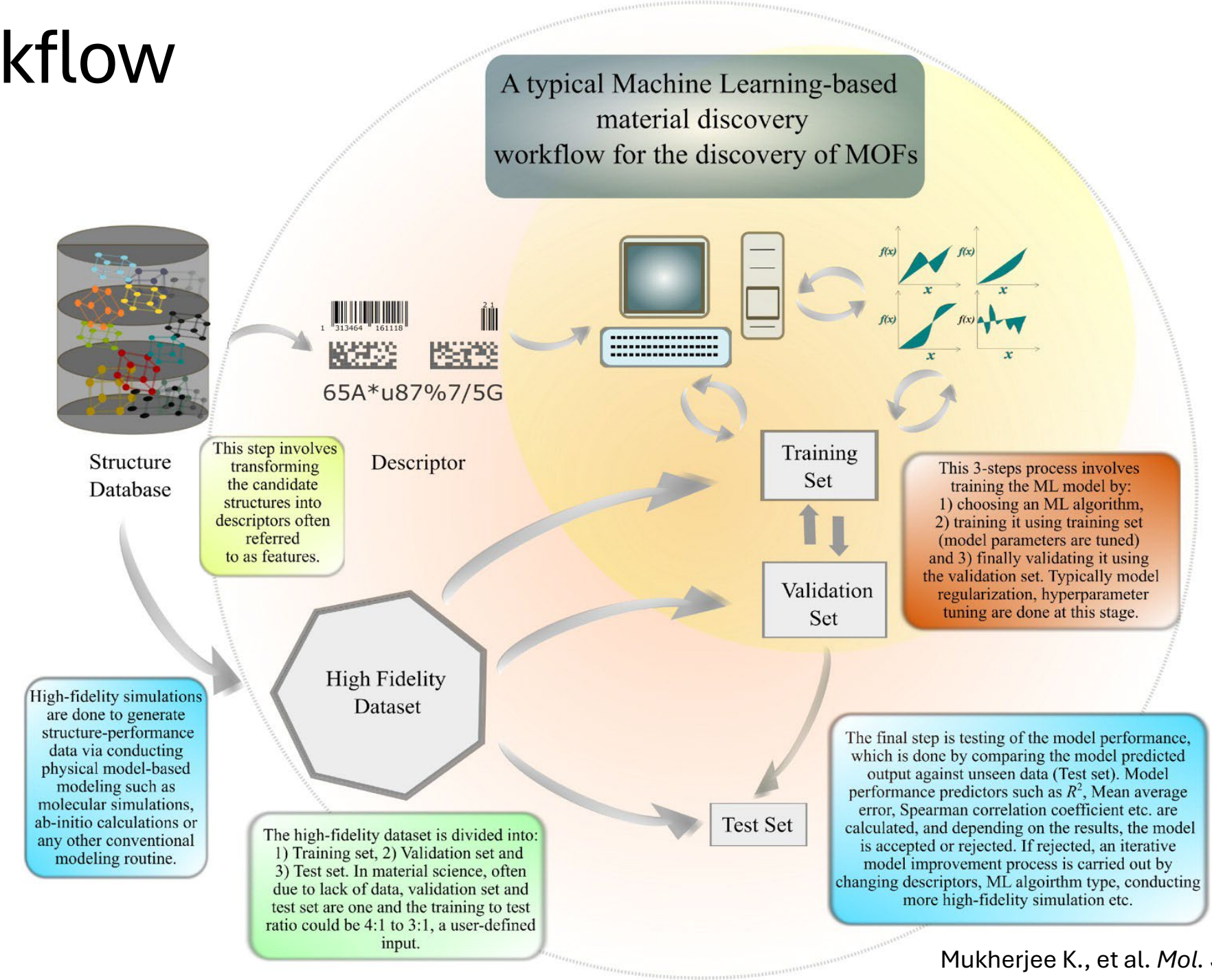
ILs for REE separations



Machine Learning for Phase Equilibria Predictions

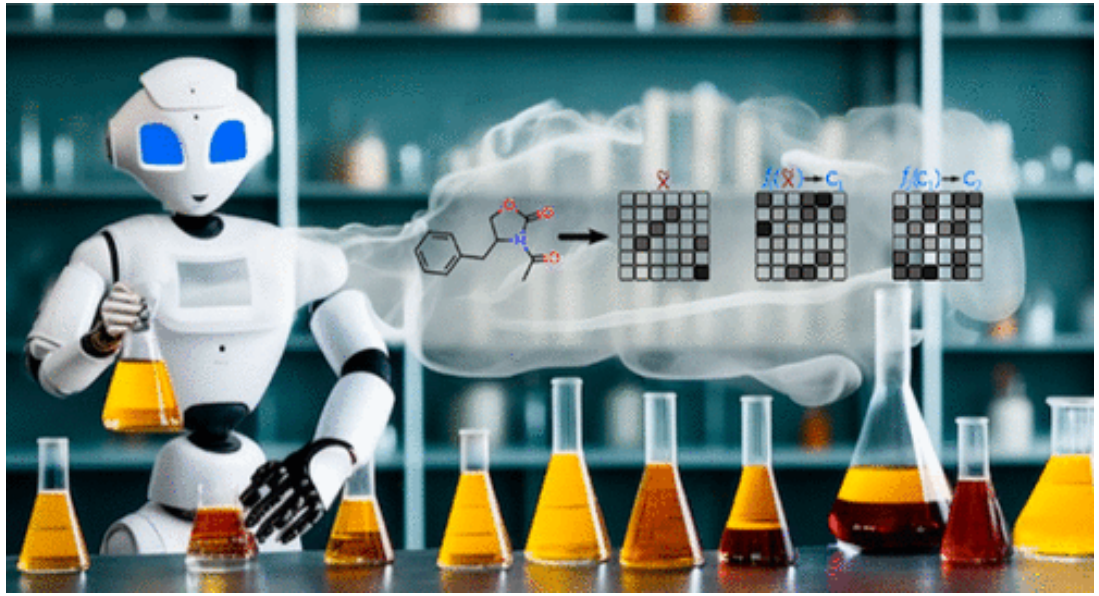


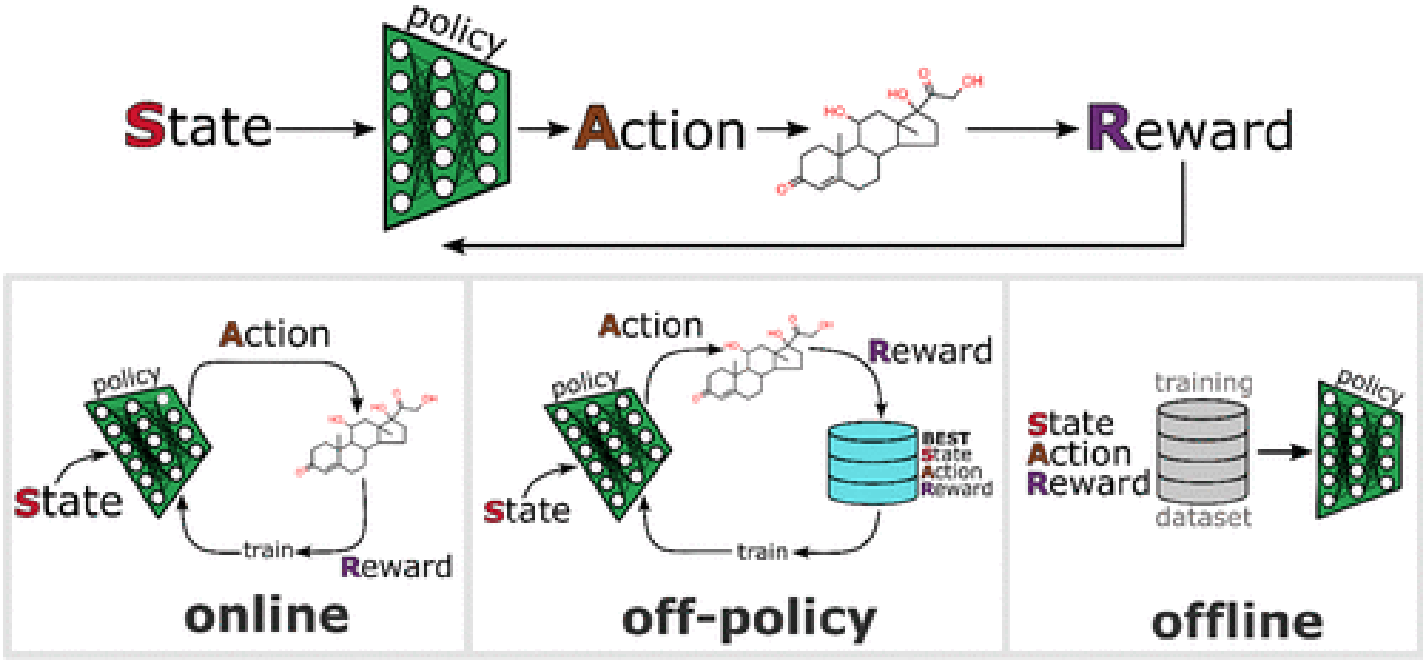
ML workflow



Mukherjee K., et al. *Mol. Sim.*, 2021

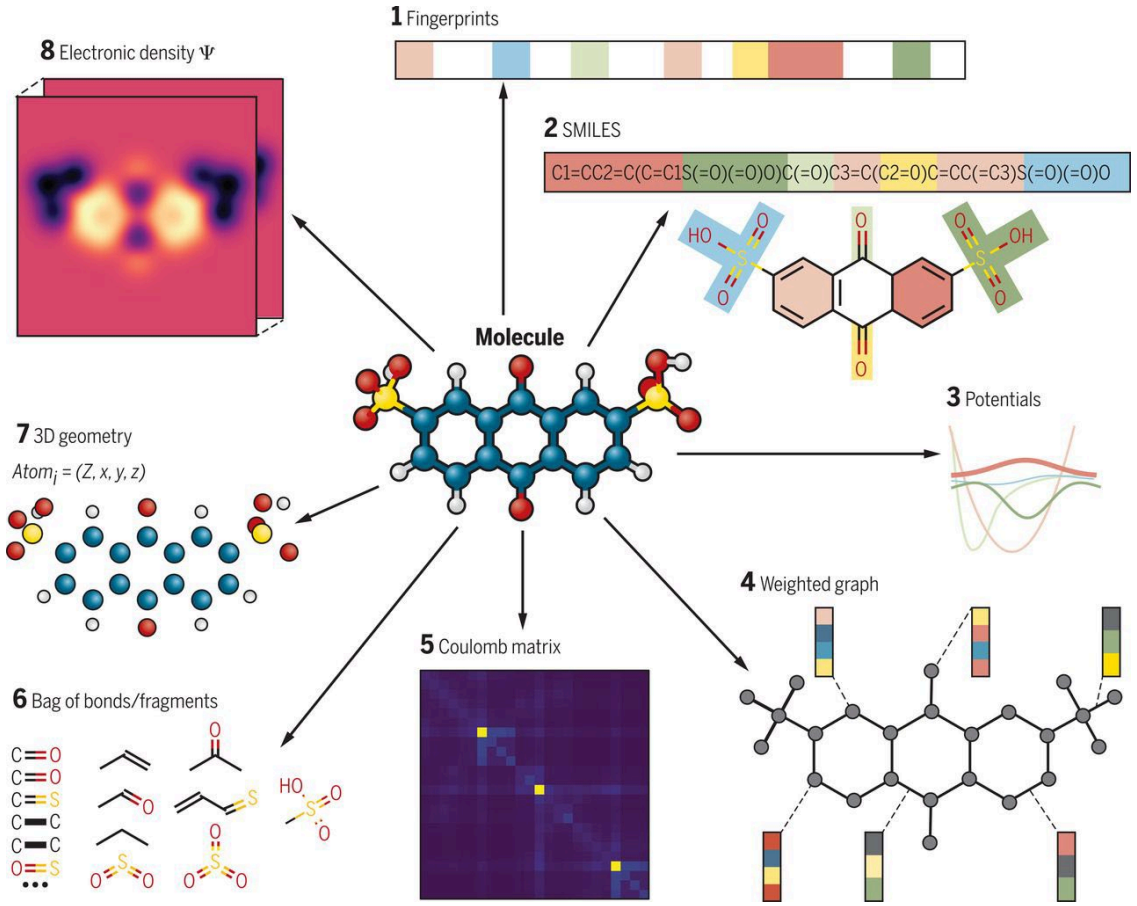
Generation of Chemistry

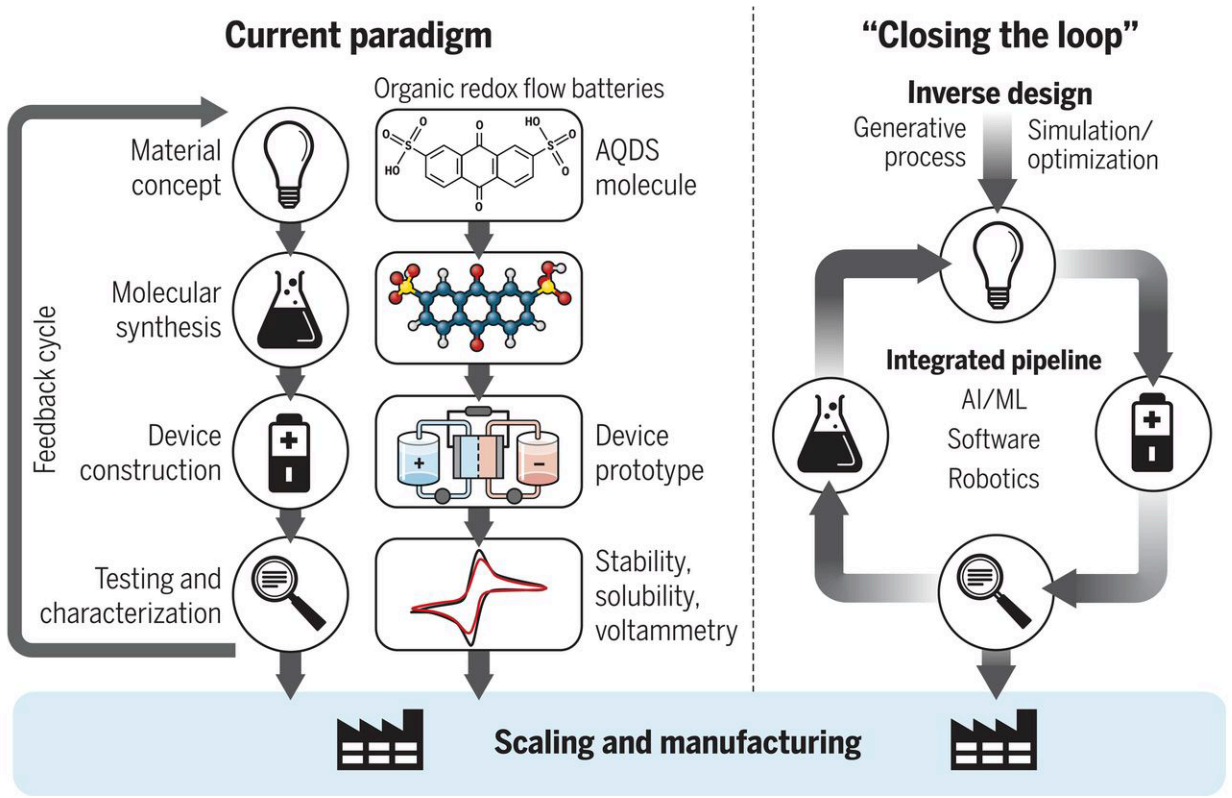




Anstine, D. M.; Isayev, O., Generative Models as an Emerging Paradigm in the Chemical Sciences. *Journal of the American Chemical Society* **2023,** *145* (16), 8736-8750.

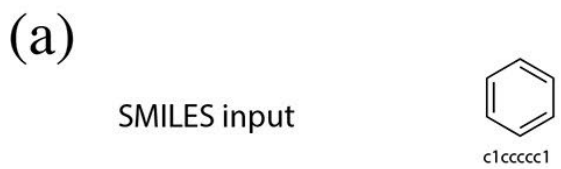
Molecular Descriptors and Inverse Design





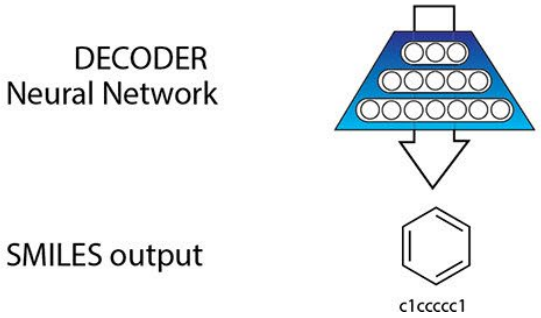
Sanchez-Lengeling and Aspuru-Guzik, Inverse molecular design using machine learning: Generative models for matter engineering, **2018**, 10.1126/science.aat2663

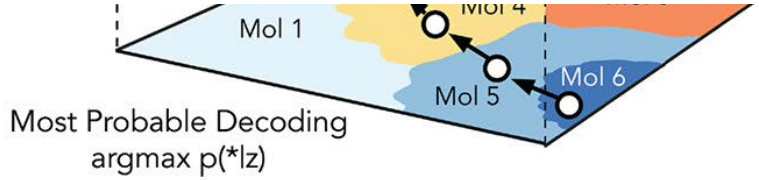
Digital Chemical Spaces



(b)

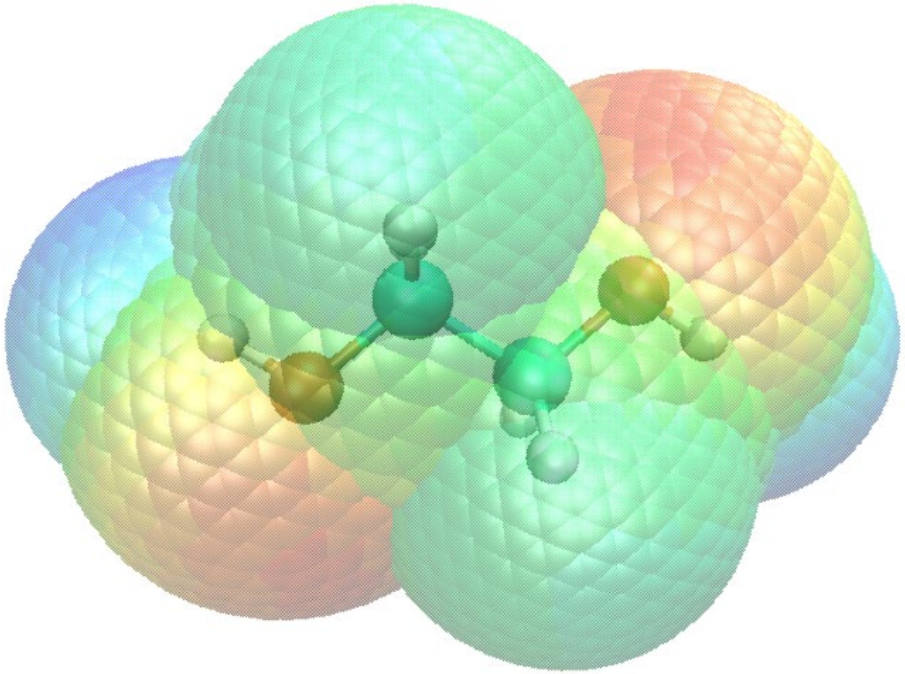
Two autoencoder systems were trained: one with 108 000 molecules from the QM9 data set of molecules with fewer than 9 heavy atoms (31) and another with 250 000 drug-like commercially available molecules extracted at random from the ZINC database. (32) We performed random optimization over hyperparameters specifying the deep autoencoder architecture and training, such as the choice between a recurrent or convolutional encoder, the number of hidden layers, layer sizes, regularization, and learning rates. The latent space representations for the QM9 and ZINC data sets had 156 dimensions and 196 dimensions, respectively.

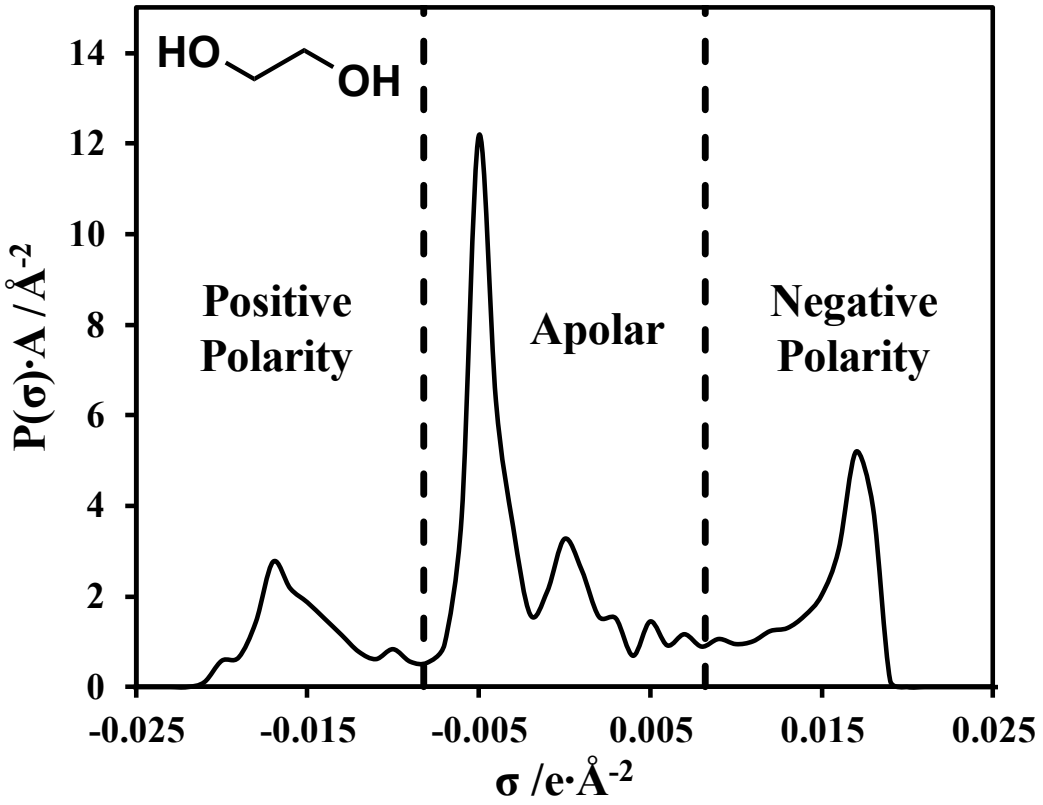




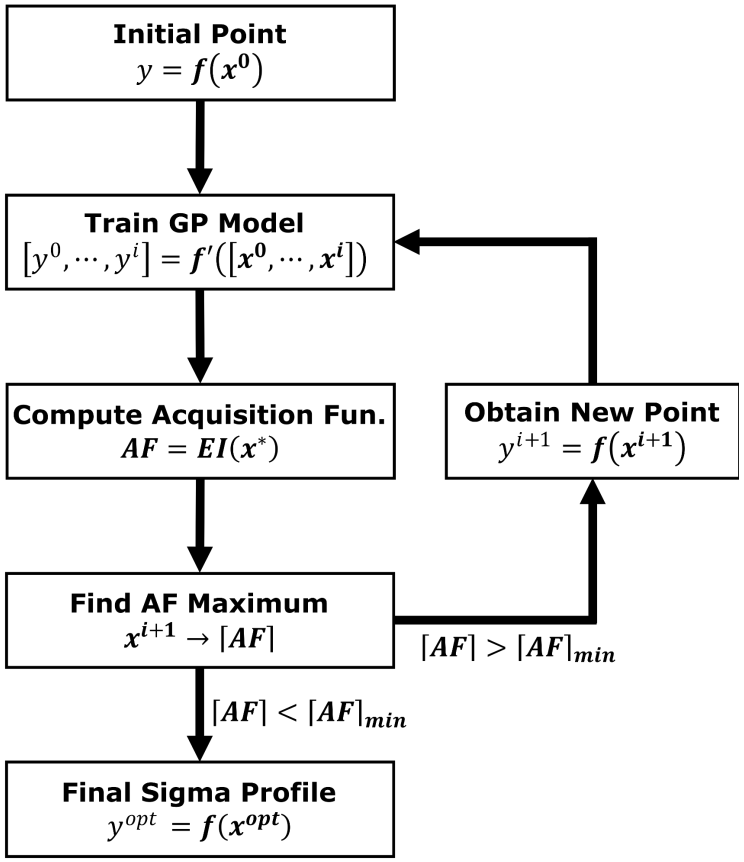
Gómez-Bombarelli, R., Wei, J. N., Duvenaud, D., Hernández-Lobato, J. M., Sánchez-Lengeling, B., Sheberla, D., ... & Aspuru-Guzik, A. (2018). Automatic chemical design using a data-driven continuous representation of molecules. ACS central science, 4(2), 268-276.

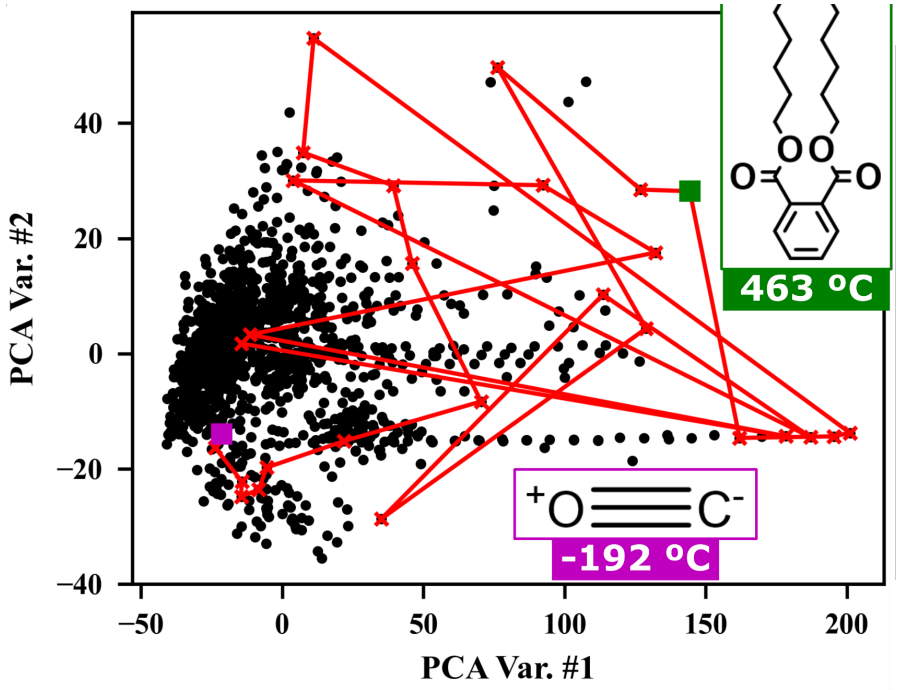
New Chemical Descriptors

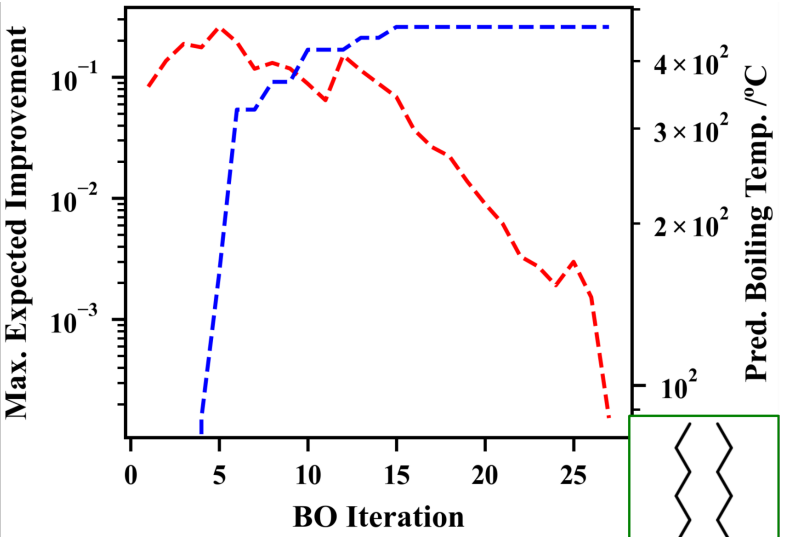




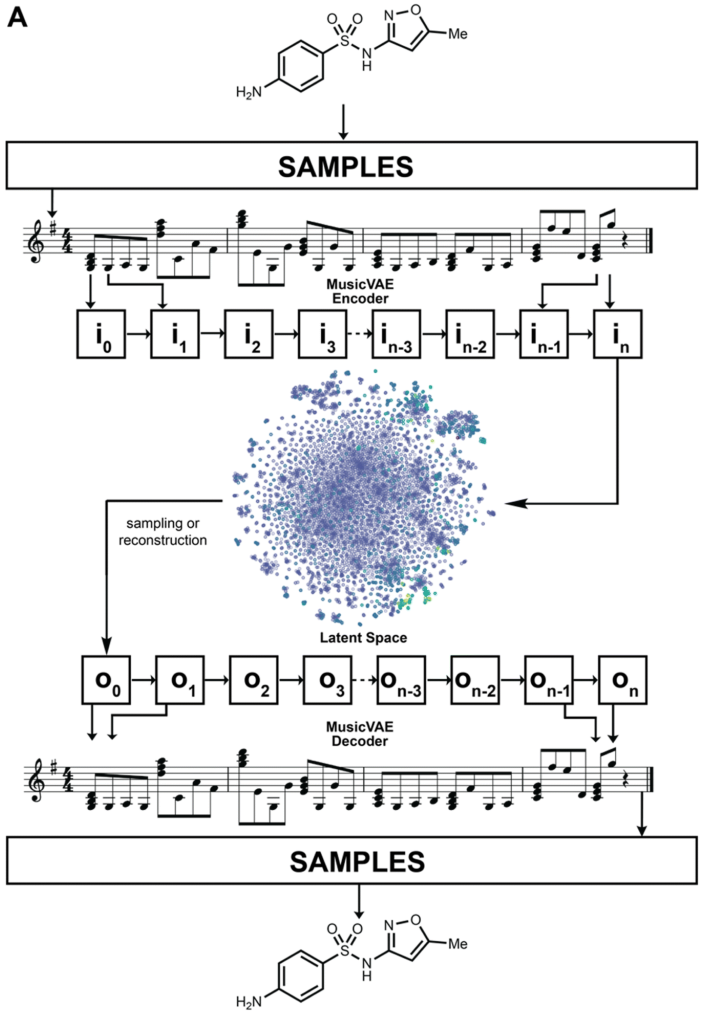
Navigating Chemical Digital Spaces

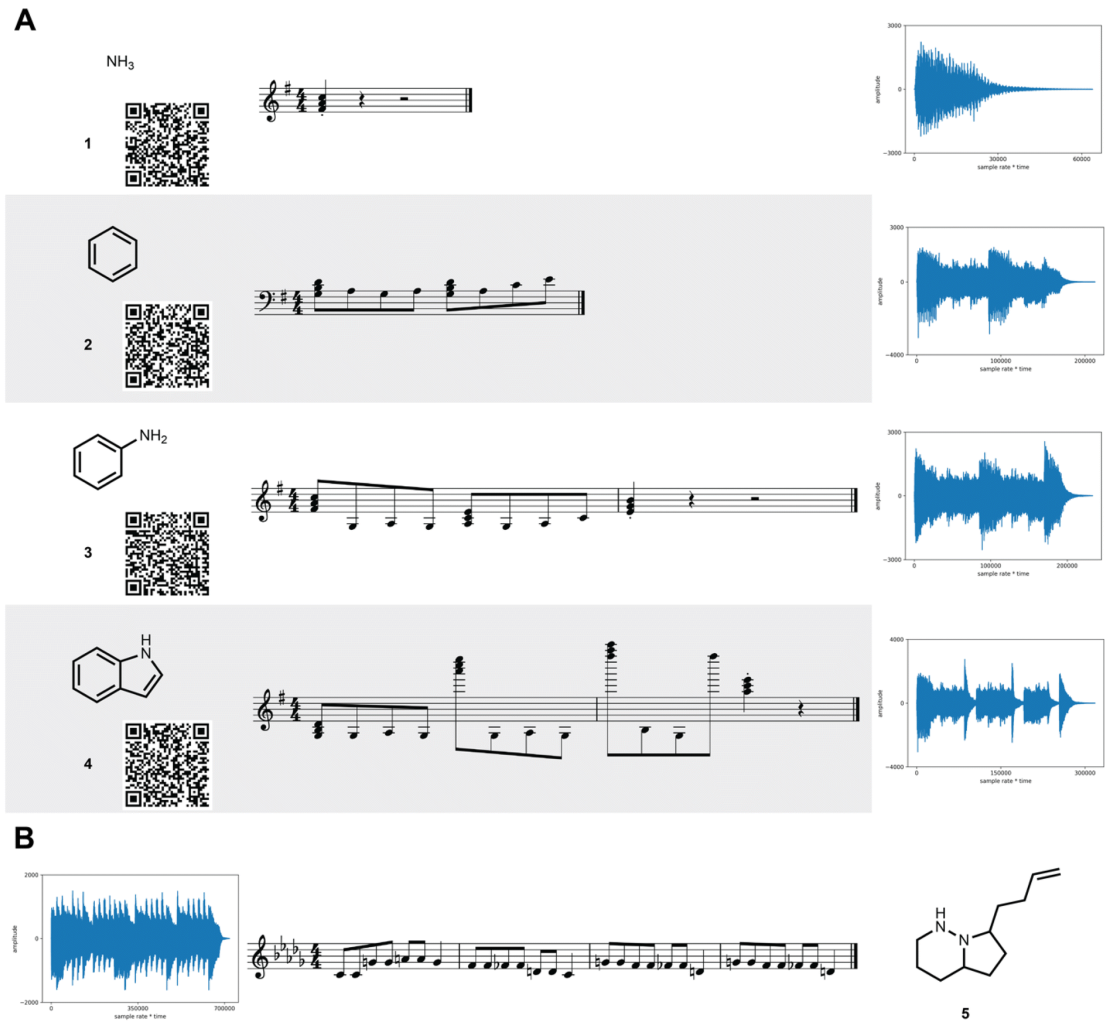






Molecular Sonification





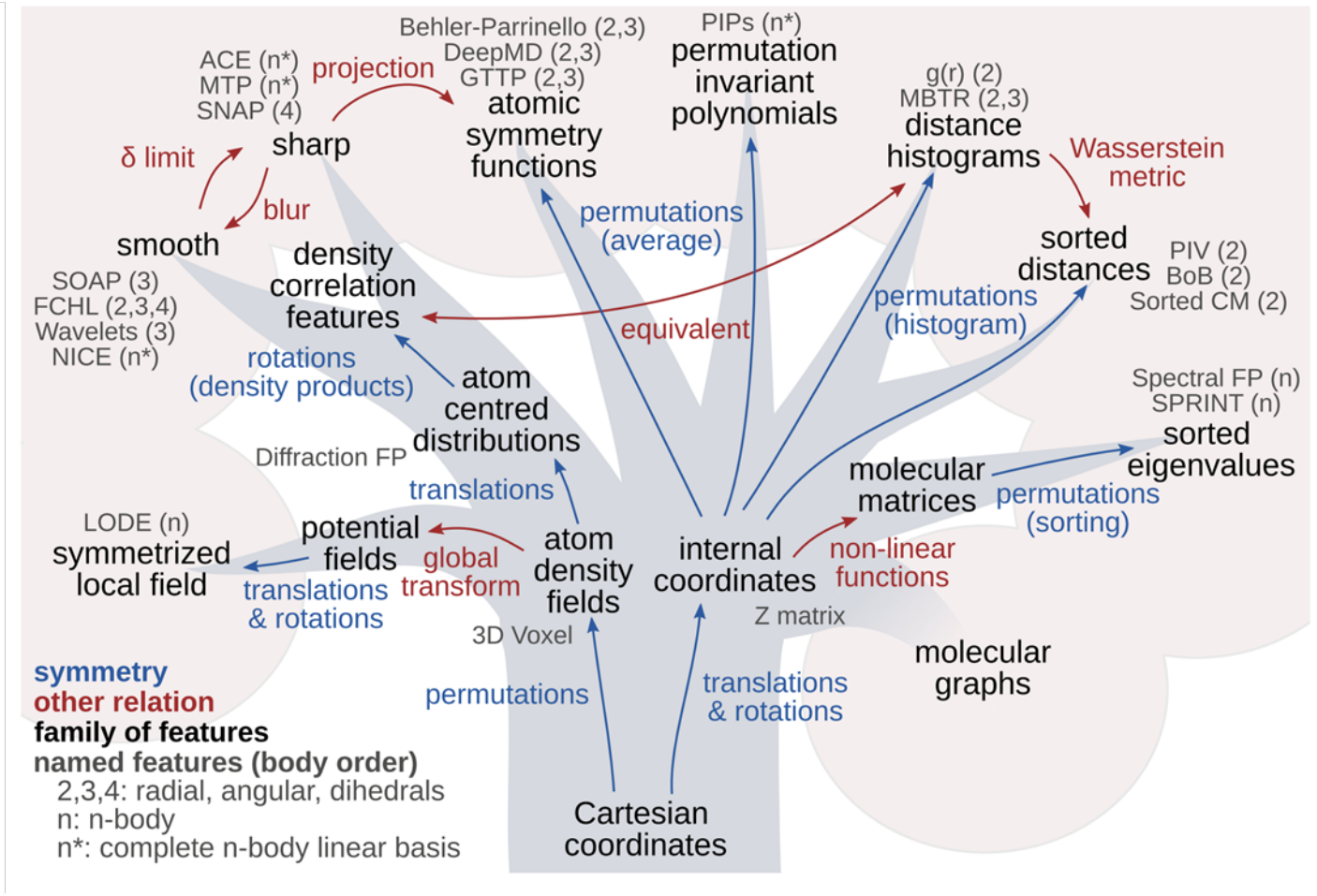
Mahjour, B.; Bench, J.; Zhang, R.; Frazier, J.; Cernak, T., Molecular sonification for molecule to music information transfer. *Digital Discovery* **2023,** *2* (2), 520-530.

Featurization

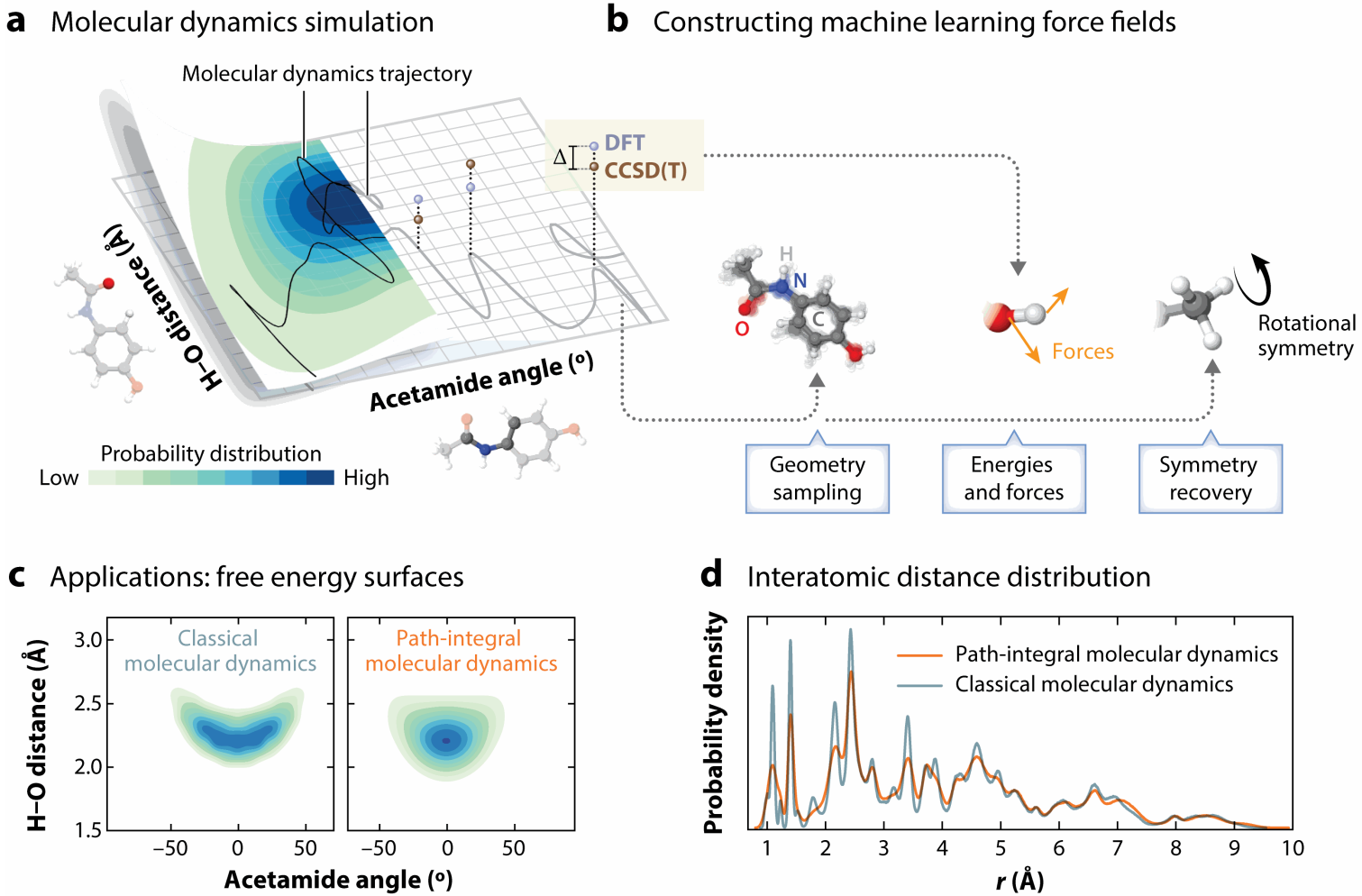
Physics-Inspired Structural Representations for Molecules and Materials

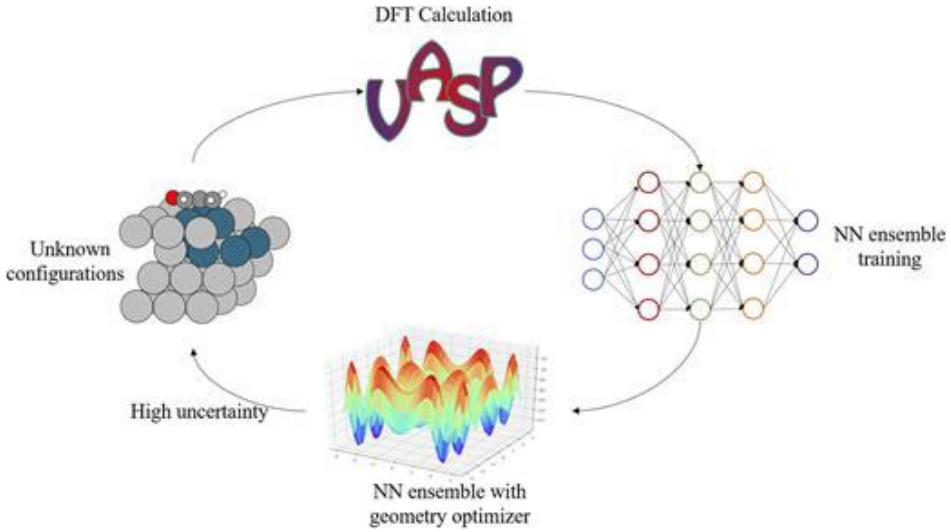
Felix Musil, Andrea Grisafi, Albert P. Bartók, Christoph Ortner, Gábor Csányi, and Michele Ceriotti*

Cite This: *Chem. Rev.* 2021, 121, 9759–9815 Read Online



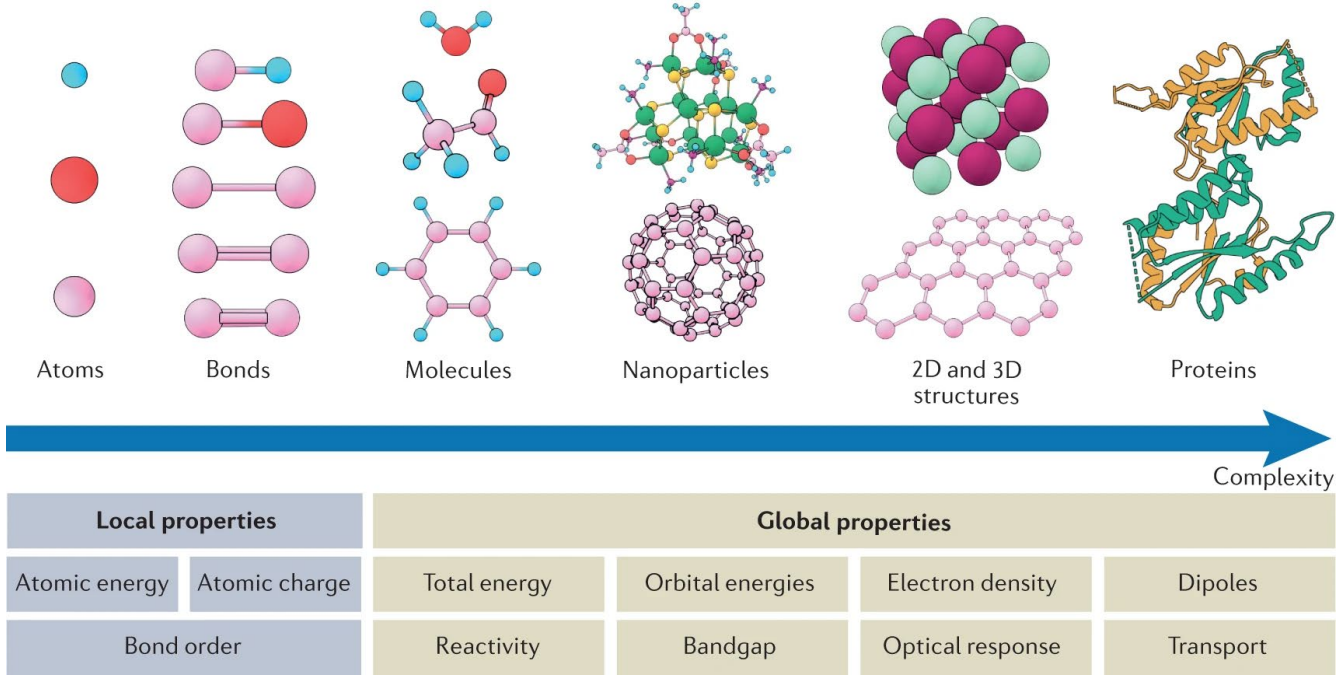
Molecular Simulations

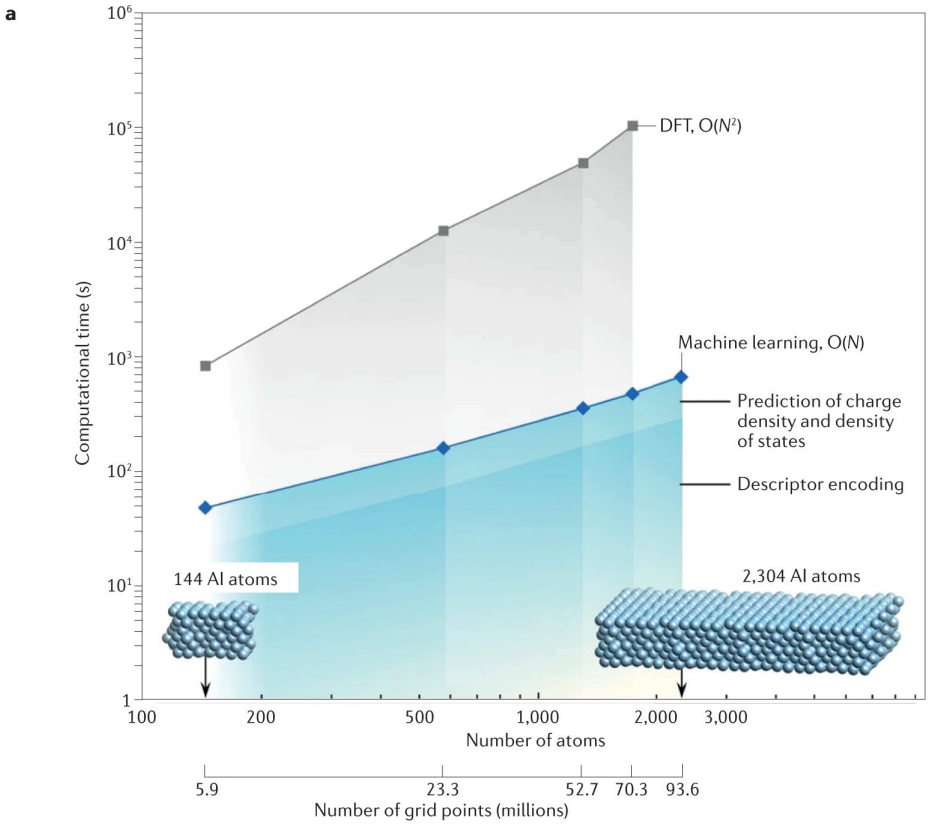




Noé F, et al. 2020.
Annu. Rev. Phys. Chetm. 71:361–90

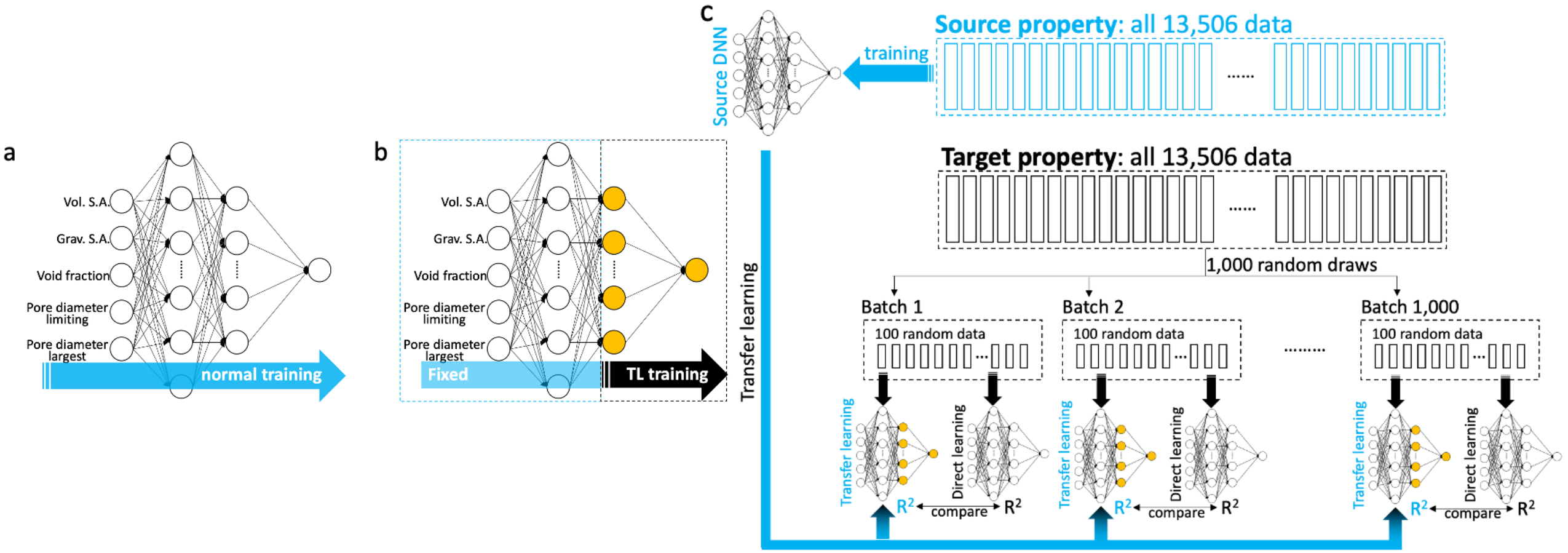
Molecular Simulations





Fedik, N.; Zubatyuk, R.; Kulichenko, M.; Lubbers, N.; Smith, J. S.; Nebgen, B.; Messerly, R.; Li, Y. W.; Boldyrev, A. I.; Barros, K.; Isayev, O.; Tretiak, S., Extending machine learning beyond interatomic potentials for predicting molecular properties. *Nature Reviews Chemistry* **2022,** *6* (9), 653-672.

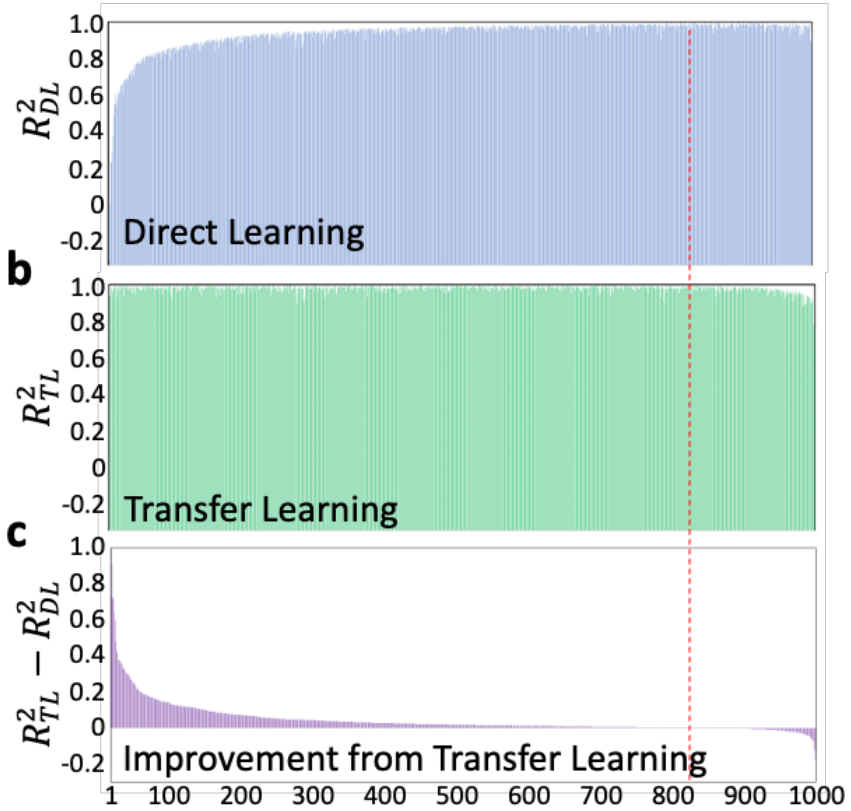
Transfer Learning

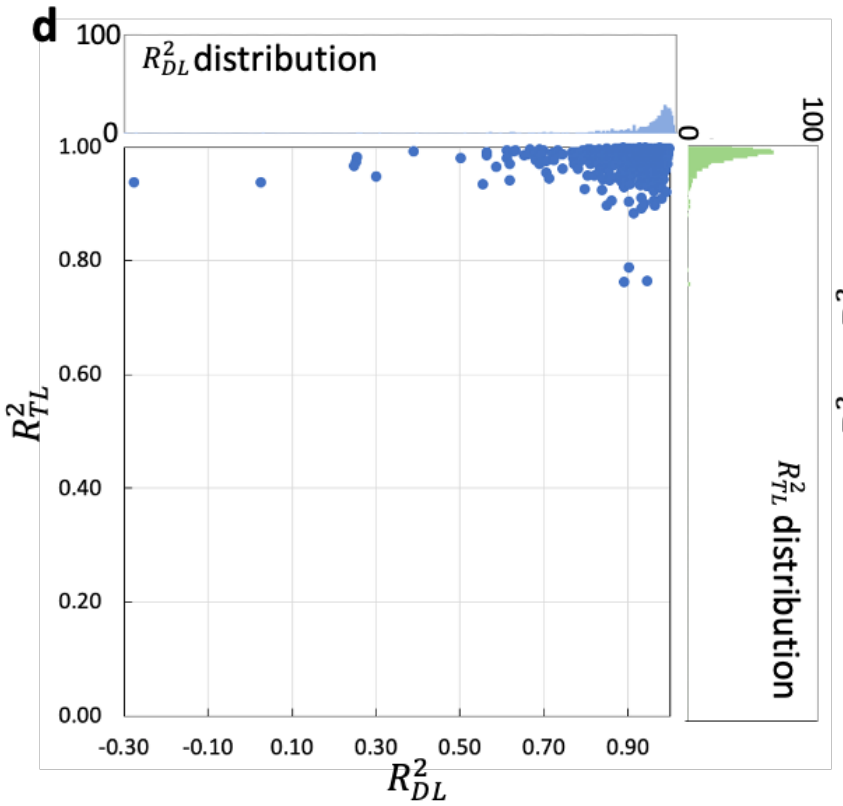


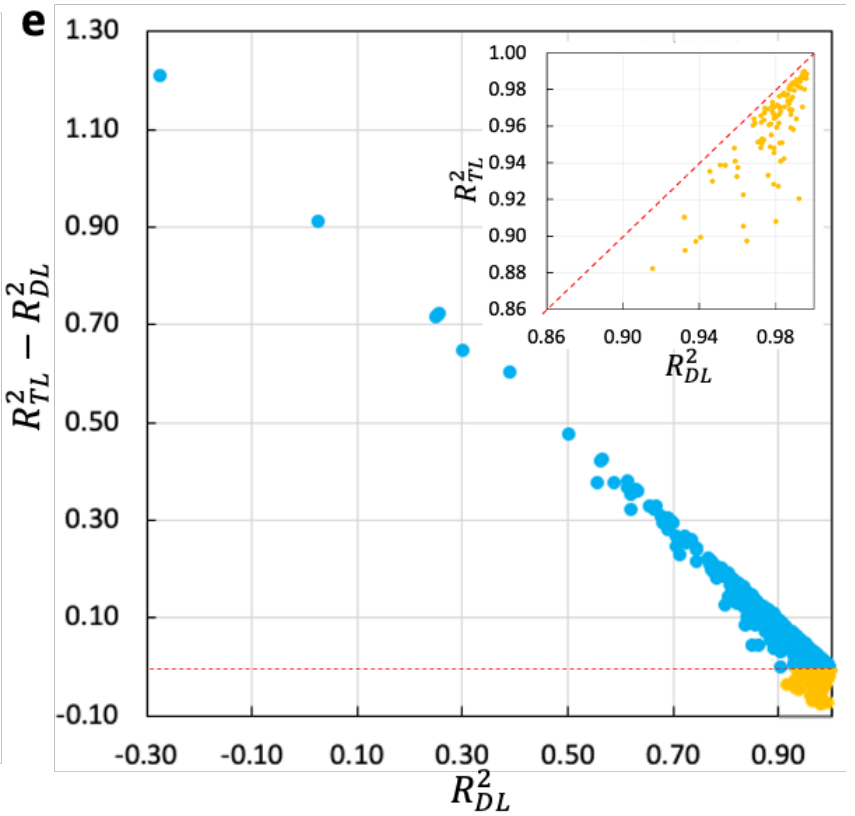
Ma, R. et al. *ACS Appl. Mat. Int.* . 2020

TL from hydrogen to methane

H2 (100 bar, 243 K) to CH4 (100 bar, 298 K)

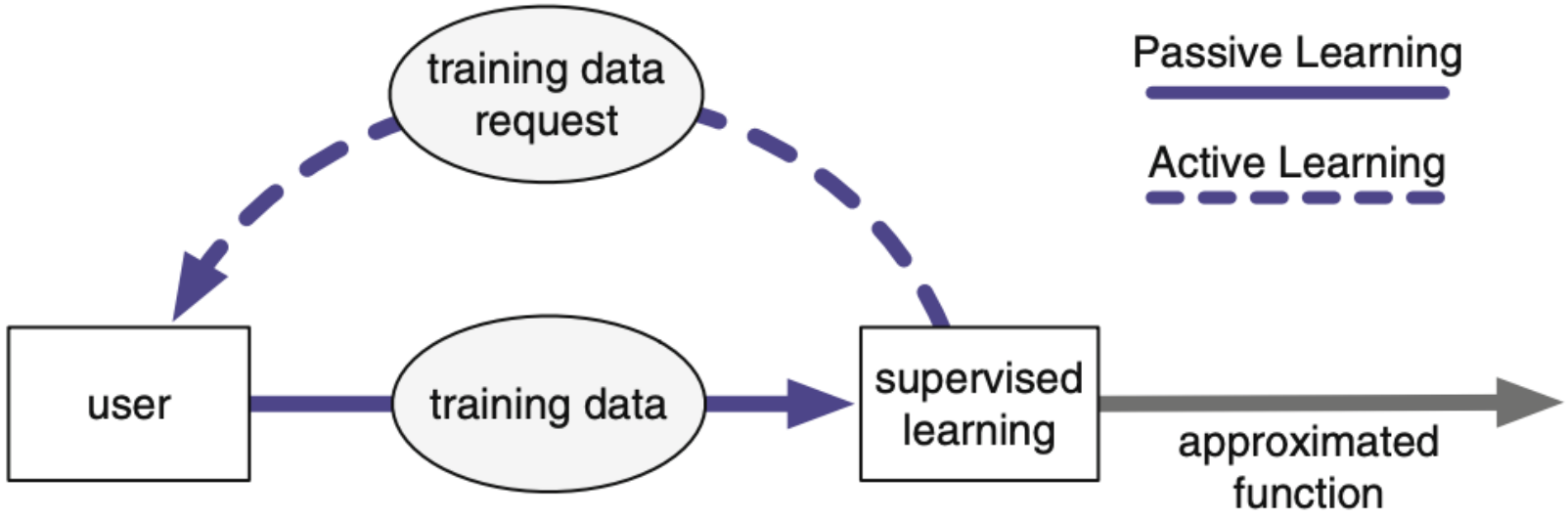






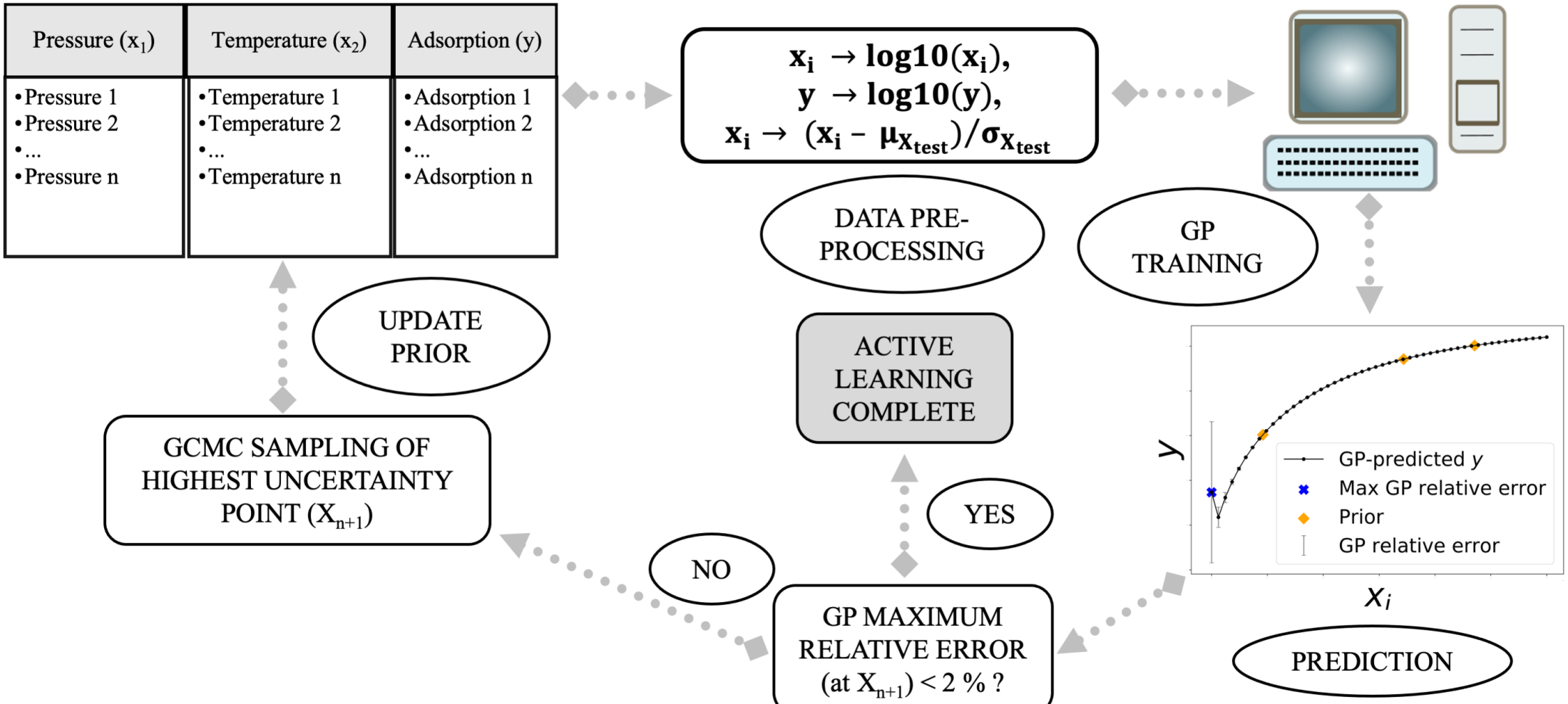
Ma, R. et al. *ACS Appl. Mat. Int.* . 2020

Active Learning



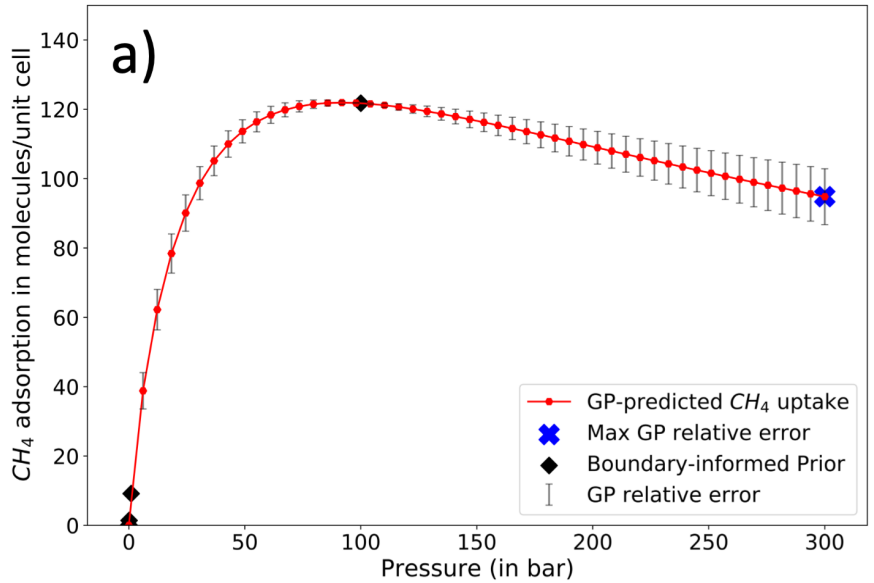
$$f \sim \mathcal{N}\left(\begin{bmatrix}\mu(x_1)\\ \mu(x_2)\\ \vdots\\ \mu(x_n)\end{bmatrix}, \begin{bmatrix}K(x_1,x_1) & \cdots & K(x_1,x_n)\\ K(x_2,x_1) & \ddots & \vdots\\ \vdots & \cdots & \vdots\\ K(x_n,x_1) & \cdots & K(x_n,x_n)\end{bmatrix}\right)$$

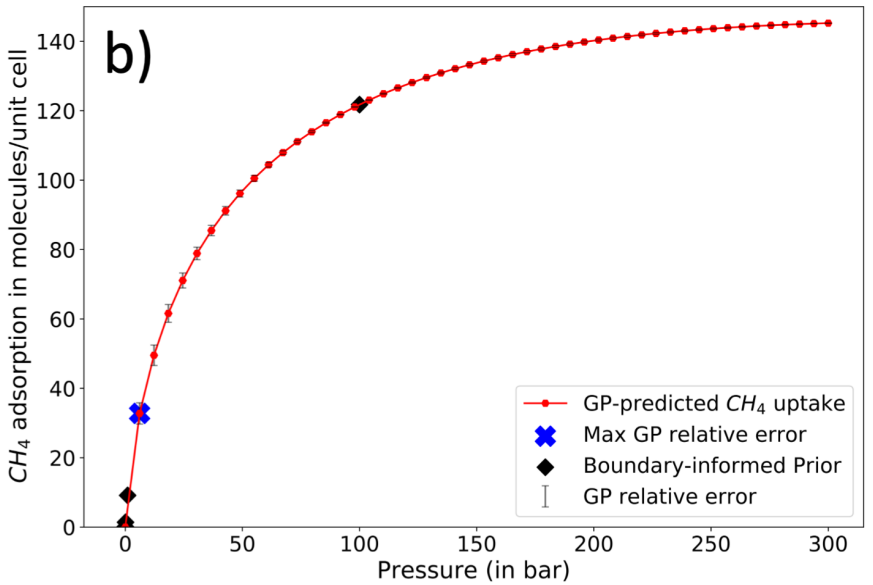
Sequential Design of Adsorption Simulations

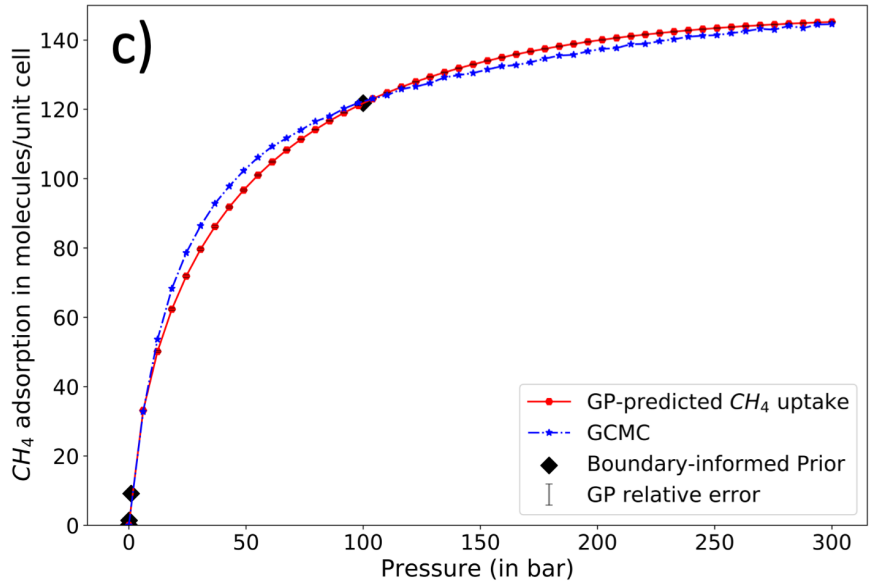


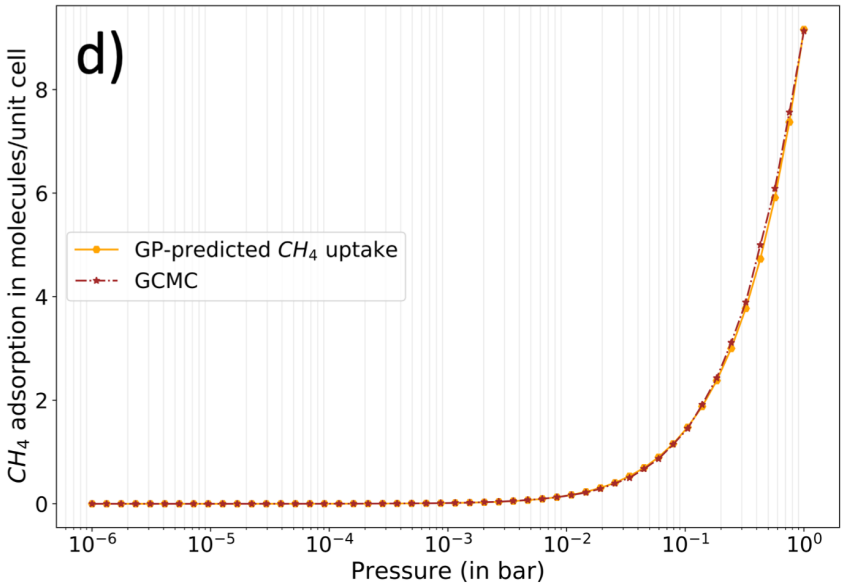
Mukherjee, K. et al. *Mol. Sys. Des. Eng.* 2022

Methane Isotherm



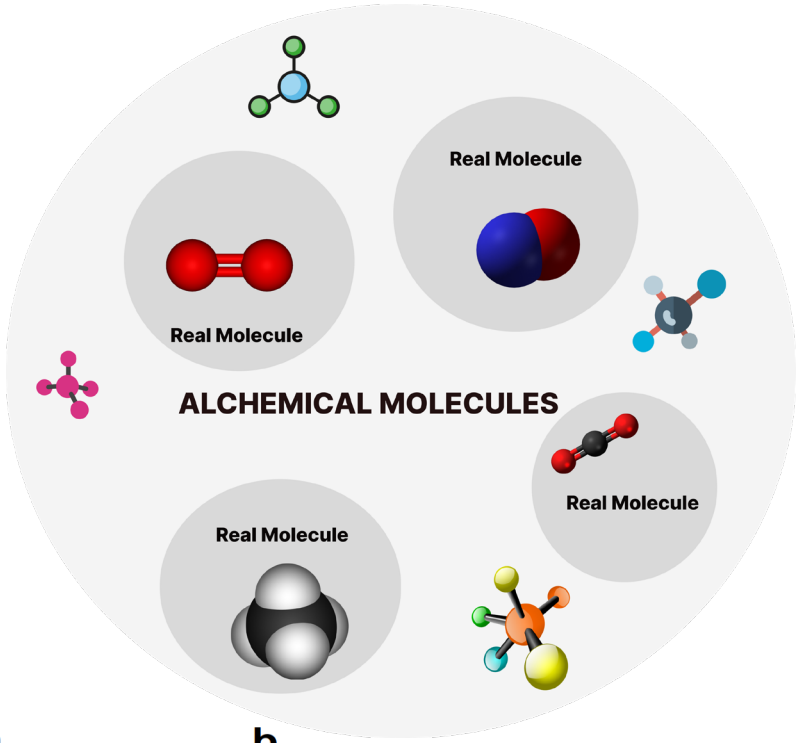


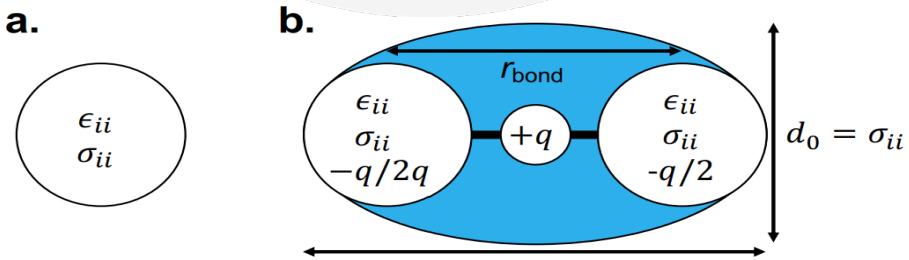




Mukherjee, K. et al. *Mol. Sys. Des. Eng.* 2022

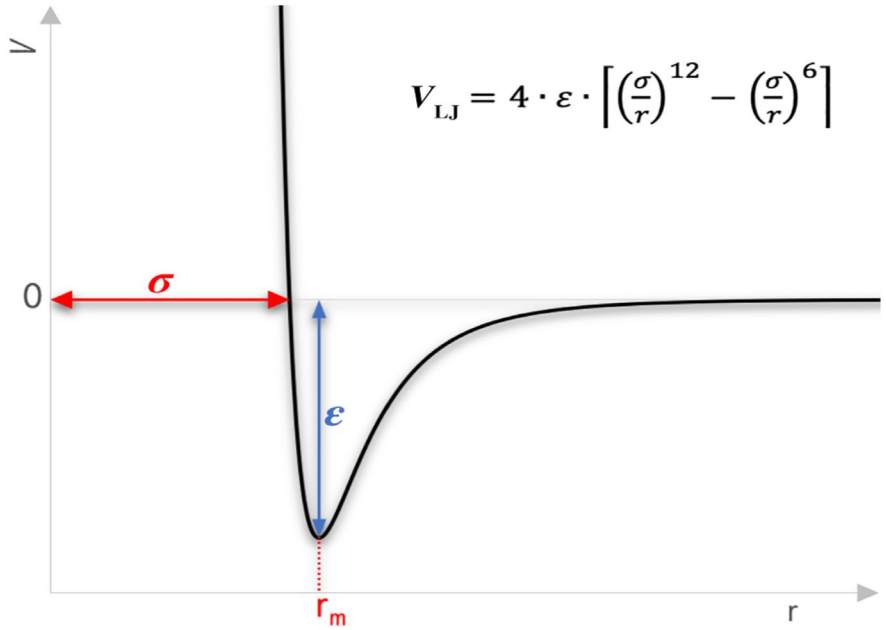
Universal Adsorption through Alchemical Species

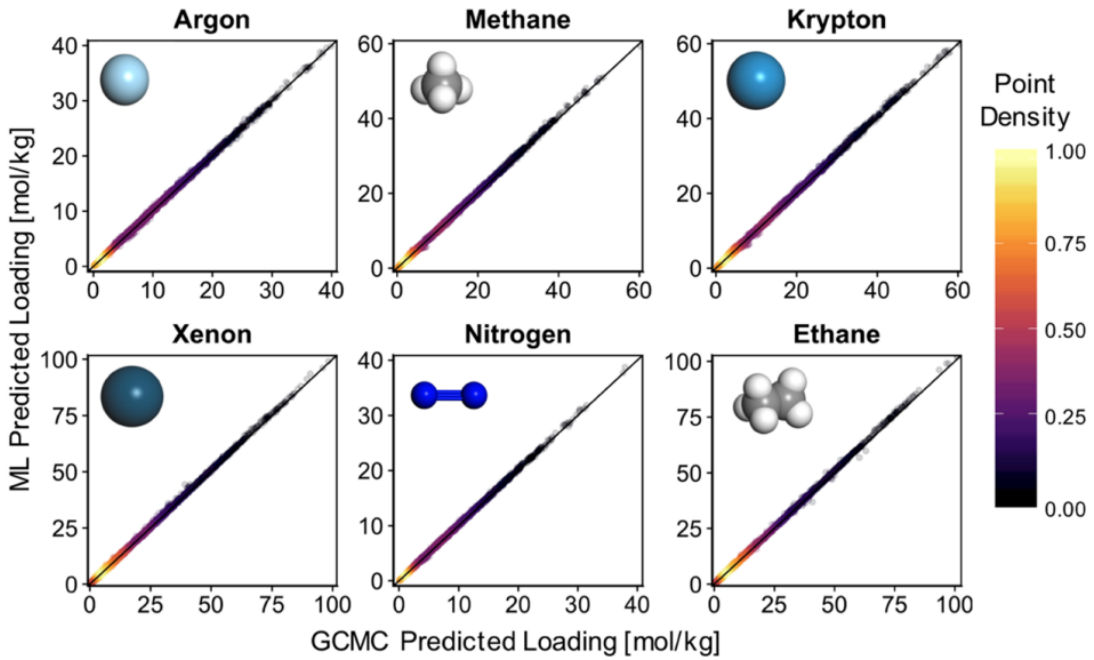




$\epsilon_{effective} = \epsilon_{ii}$
$\sigma_{effective} = \sigma_{ii}$
bond length = 0
charge = 0

$\epsilon_{effective} = \epsilon_{ii} + \epsilon_{ii}$
$\sigma_{effective} = (d_0 + d_1)/2$
bond length = r_{bond}
charge = q





Anderson.et al. J. Chem. Theory Comput. 16, 1271–1283 (2020).

AL in alchemical space

$$N \sim f(\mathrm{F}, \varepsilon, \sigma, l, q)$$

Starting Prior

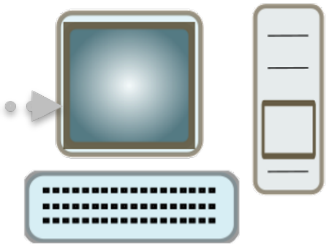
PC1 (X_1)	PC2 (X_2)	Fugacity (X_3)	Epsilon (X_4)	Sigma (X_5)	Bond length (X_6)	Charge (X_7)	Adsorption (y)
PC1 1	PC2 1	Fugacity 1	Epsilon 1	Sigma 1	Bond length 1	Charge 1	Adsorption 1
PC1 2	PC2 2	Fugacity 2	Epsilon 2	Sigma 2	Bond length 2	Charge 2	Adsorption 2
…	…	…	…	…	…	…	…
PC1 n	PC2 n	Fugacity n	Epsilon n	Sigma n	Bond length n	Charge n	Adsorption n

6000 total iterations

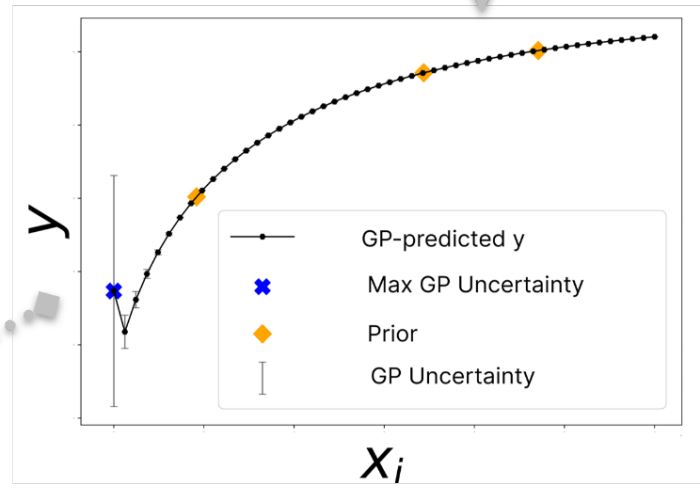
ACTIVE LEARNING COMPLETE

$X_i \rightarrow \mathbf{log10}(X_i)$,
$y \rightarrow \mathbf{log10}(y)$,
$X_i \rightarrow (X_i - \mu_{X_{test}})/\sigma_{X_{test}}$

For each bag_k



GP TRAINING



Compile the GP Mean uncertainty (at X_{n+1}) for each bag_k

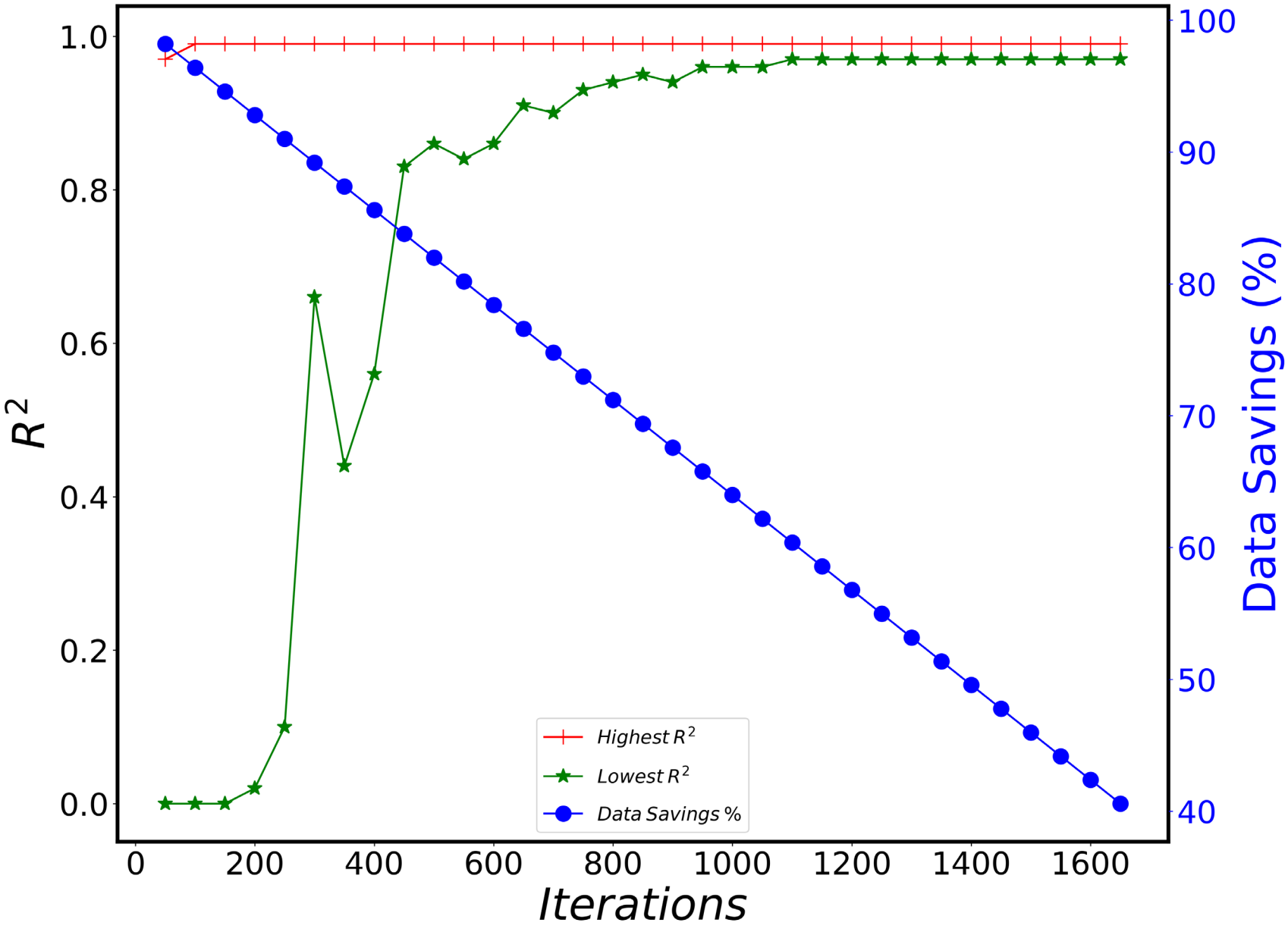
Determine the max. GP mean uncertainty across all bags and get the corresponding X_{n+1}

Lookup the corresponding y for the X_{n+1} with the max GP MAE

Update initial prior

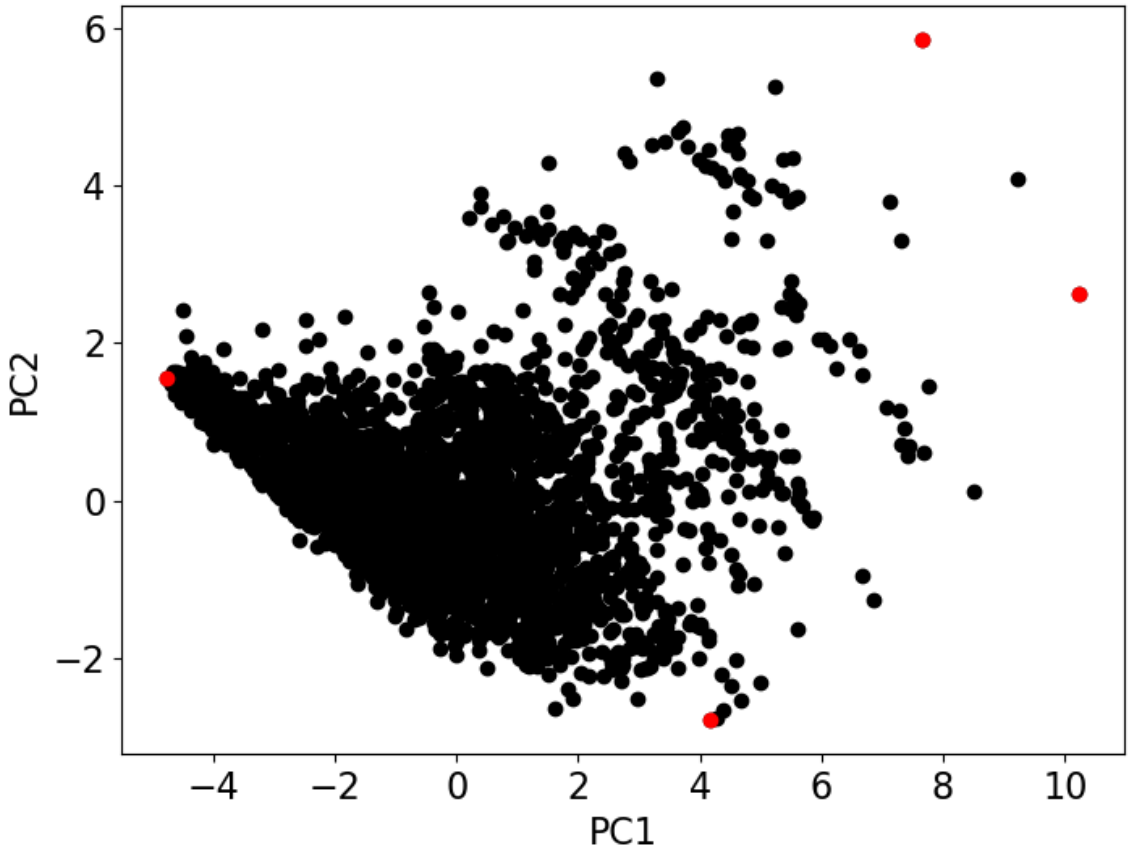
Significant Data Savings!

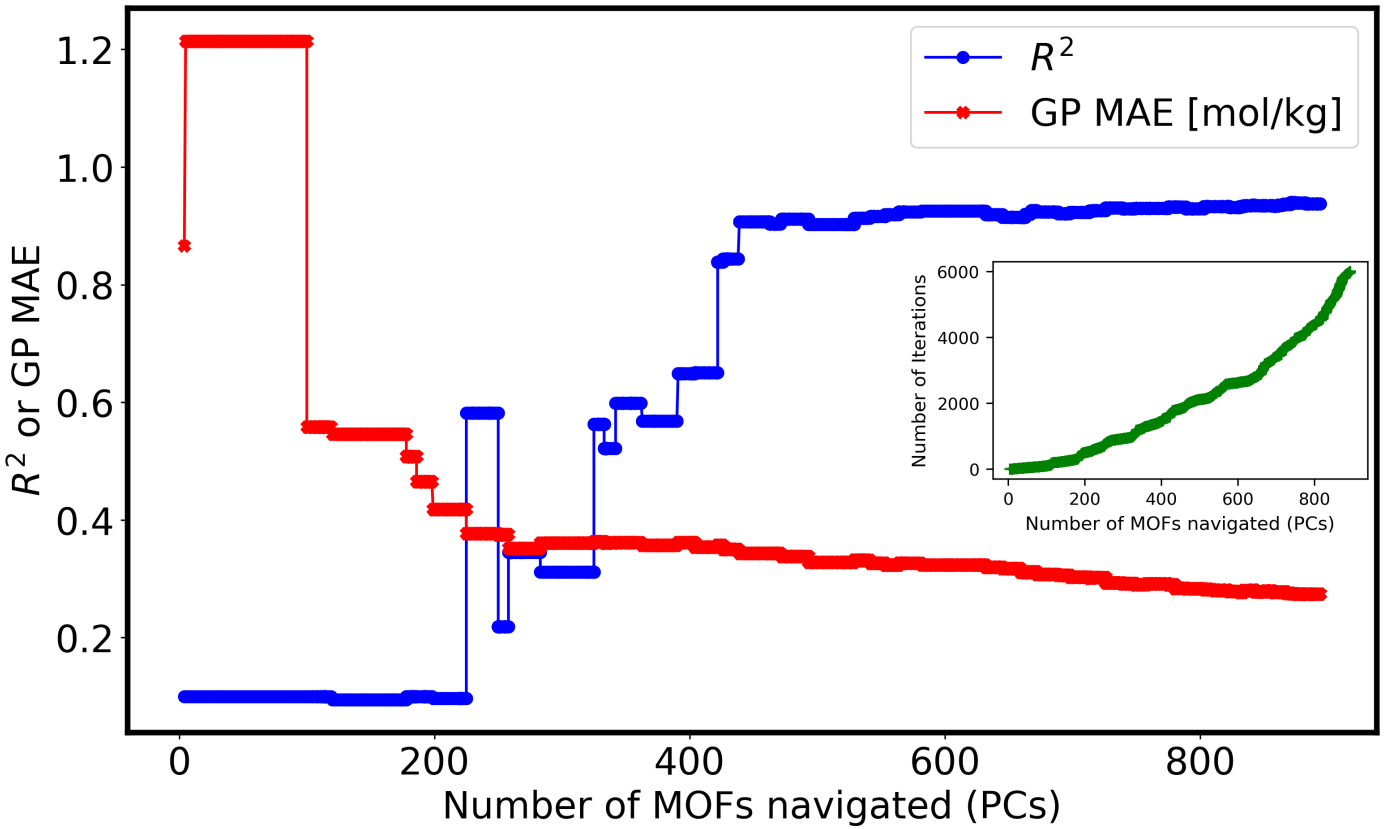
$$N \sim f(\mathrm{F}, \varepsilon, \sigma, l, q)$$



Navigating Alchemical and Material Spaces

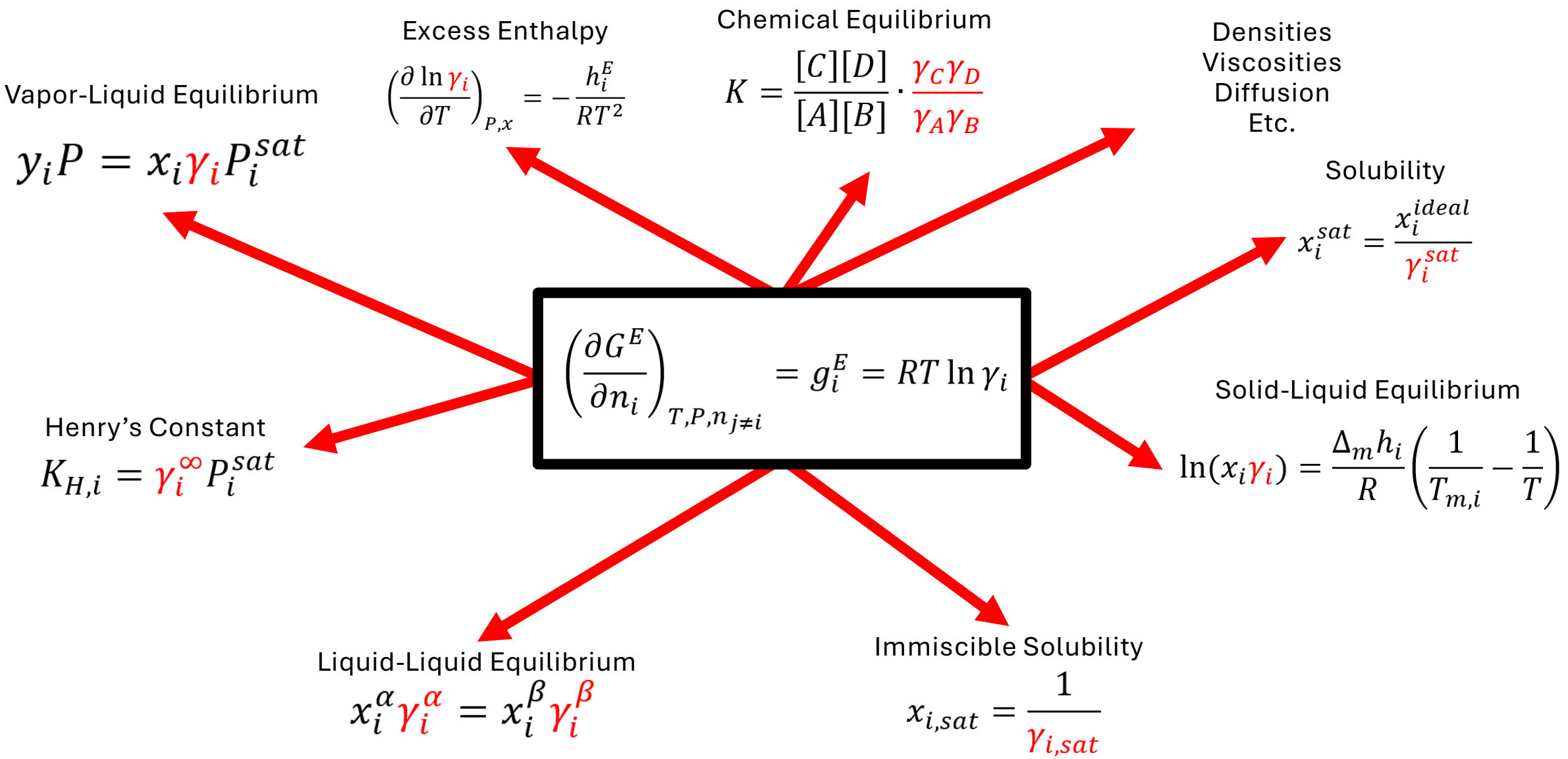
$$N \sim f(PC1, PC2, \mathrm{F}, \varepsilon, \sigma, l, q)$$



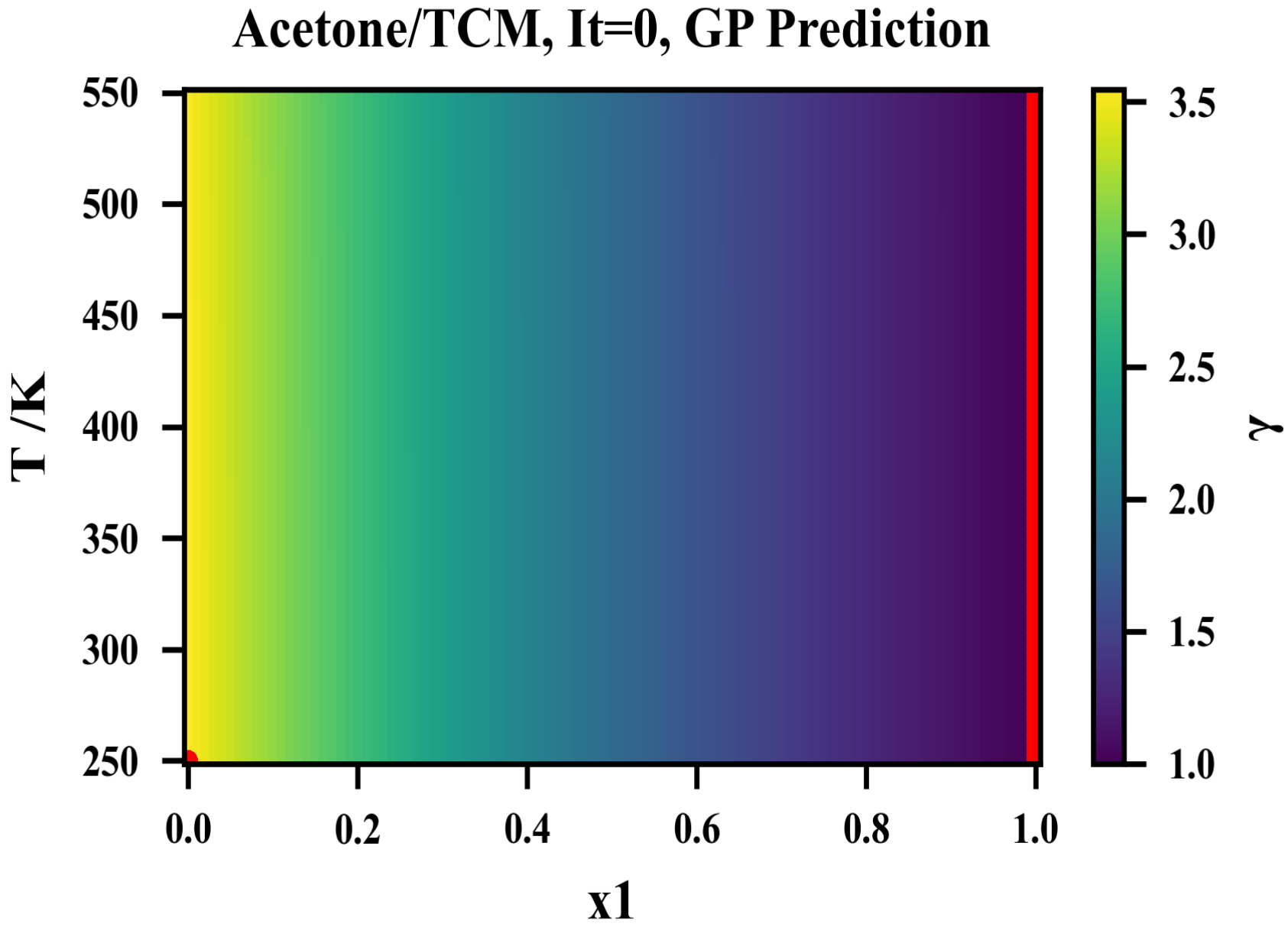


6,000 vs. 5,000,000 data

Active Learning on Activity Coefficients

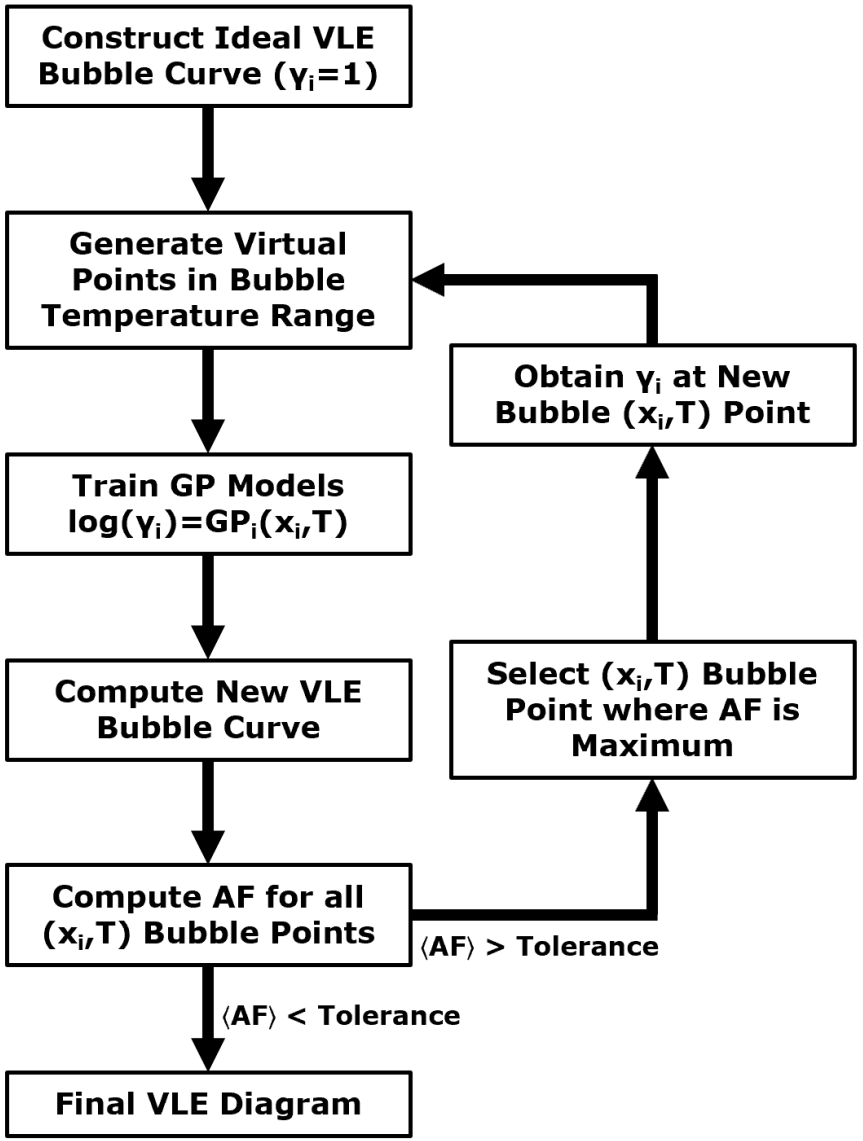


NRTL as ground truth; GPs for activity coefficients



Abranches, D. et al. *AIChE J.* 2023

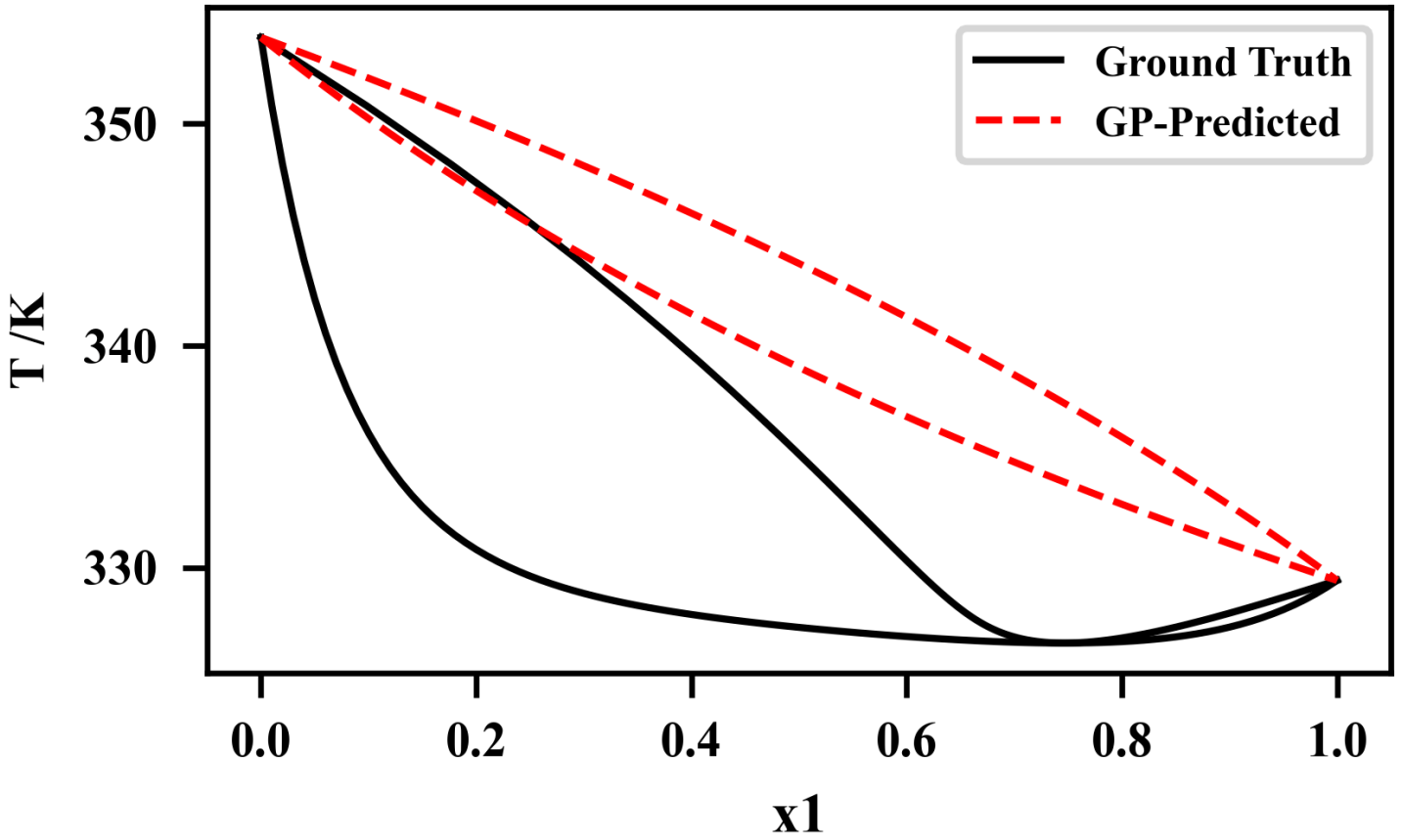
Thermodynamics-Informed AL



$$y_i P = x_i \gamma_i P_i^{sat}$$

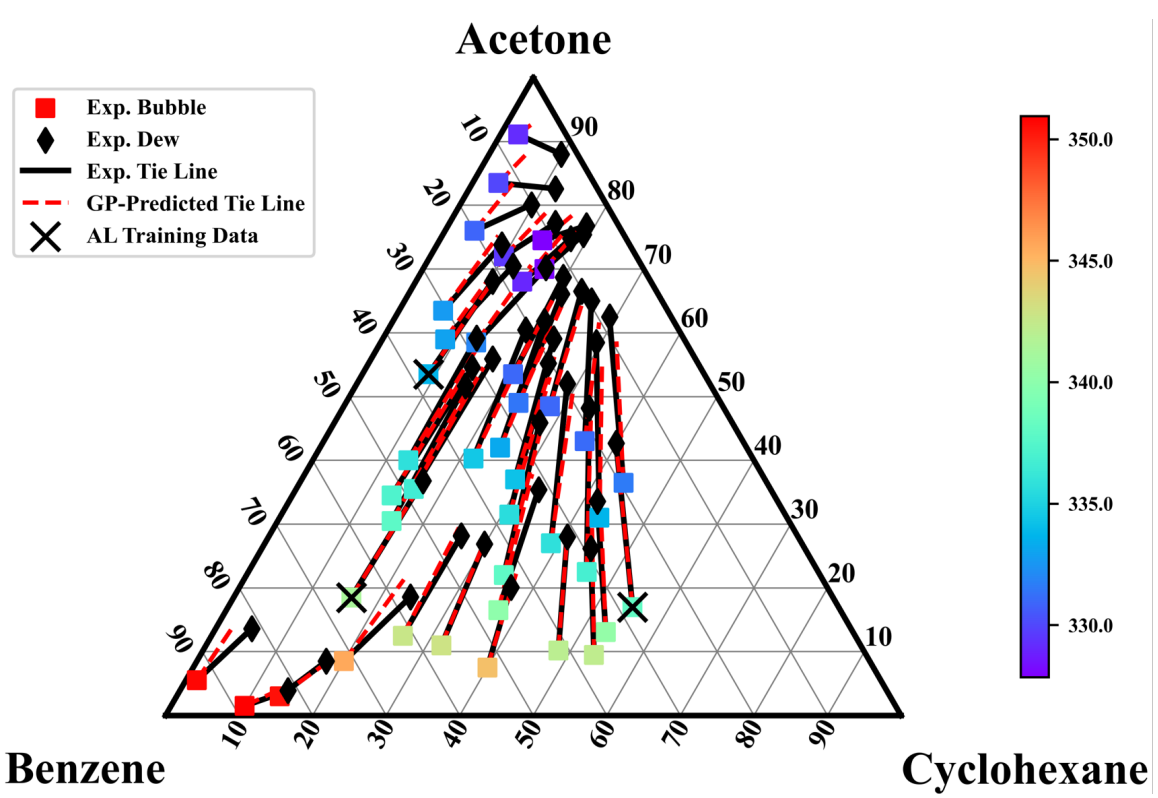
$$s_{y_i} = \left| \frac{\partial}{\partial \gamma_i} \left(\frac{x_i \gamma_i P_i^{sat}}{P} \right) \right| s_{\gamma_i} = \frac{x_i P_i^{sat}}{P} s_{\gamma_i}$$

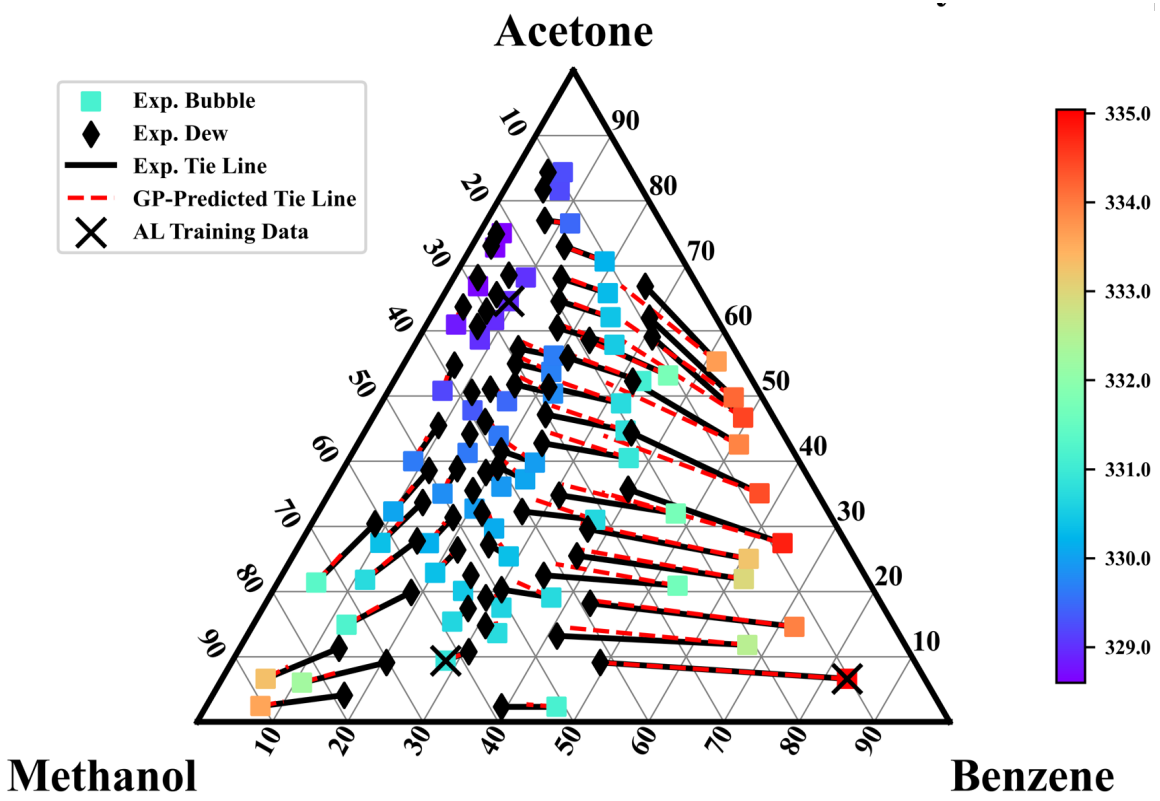
$$AF(x_i, T) = s_y^{\Sigma} = \sum_{i=1}^{N} s_{y_i}(x_i, T)$$



Abranches, D. et al. *AIChE J.* 2023

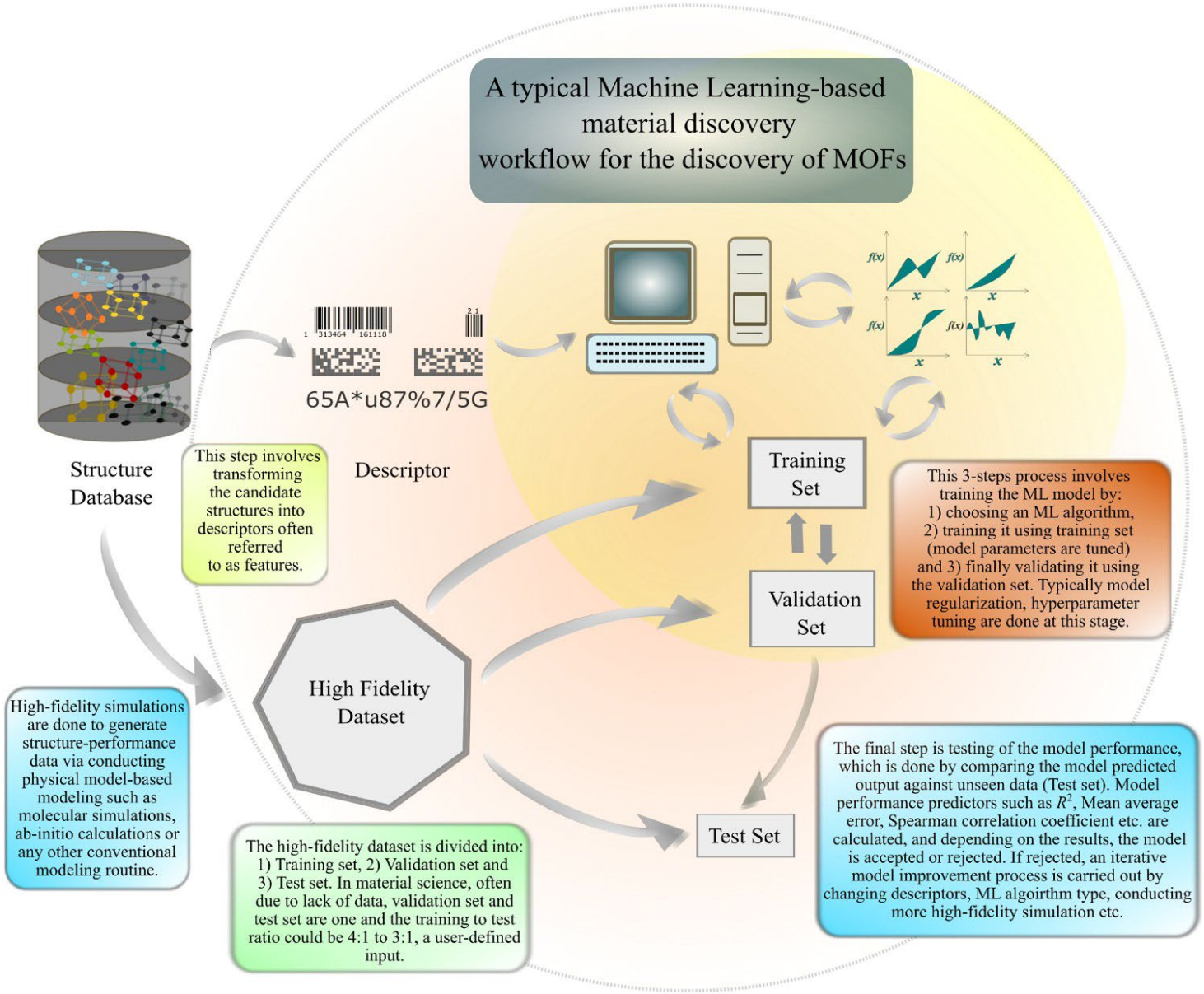
Multicomponent VLE with AL

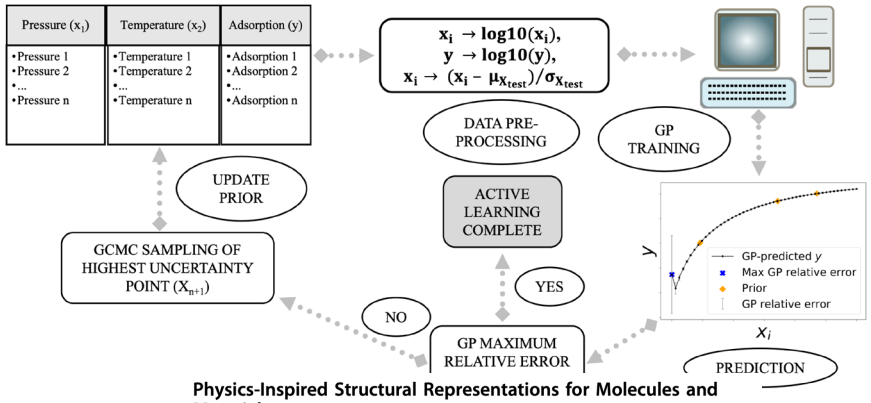




Abranches, D. et al. *AIChE J.* 2023

Summary





Physics-Inspired Structural Representations for Molecules and Materials

Felix Musil, Andrea Grisafi, Albert P. Bartók, Christoph Ortner, Gábor Csányi, and Michele Ceriotti*