A Data-driven Future for Quantum Chemistry

CMS 273: Miller/Bhattacharya
Final Presentation

Prof. Thomas Miller III

Prof. Kaushik Bhattacharya

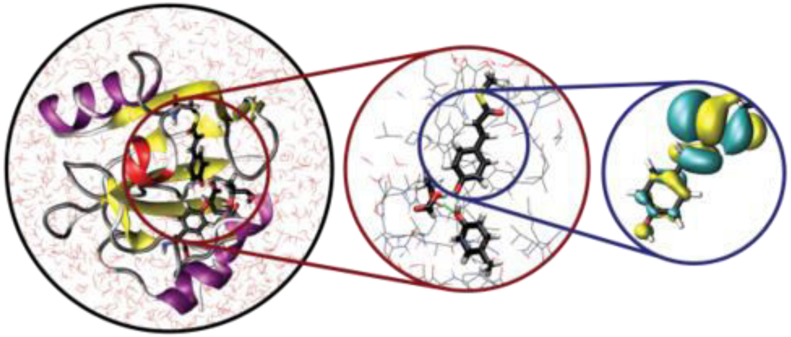
Ph.D. Matt Welborn

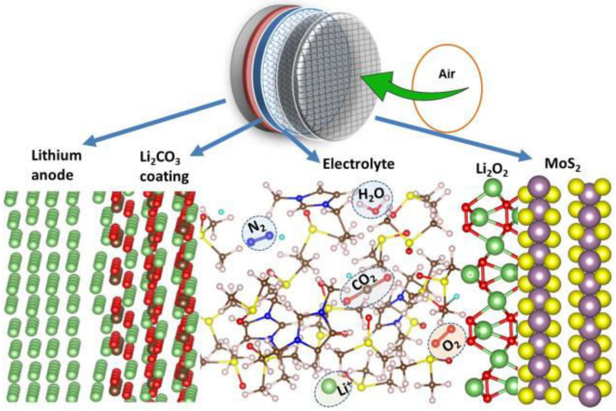
Grad students (Science/Engineering): Sherry Cheng, Ying Shi Teh

Grad students(CMS): Jialin Song, Nikola Kovachki, Dmitry Burov

Caltech

Computational Chemistry & Material Science

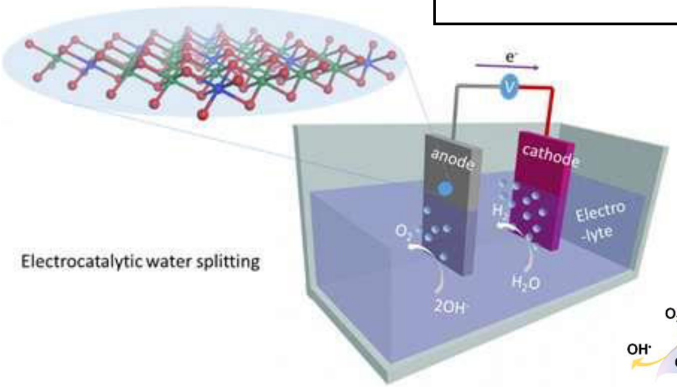




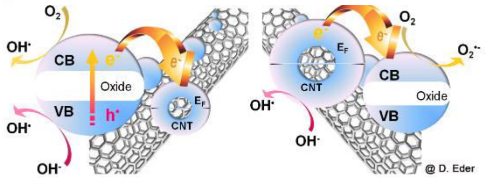
Source: Asadi *et al.*, *Nature* **555** (2018)

Schrödinger equation

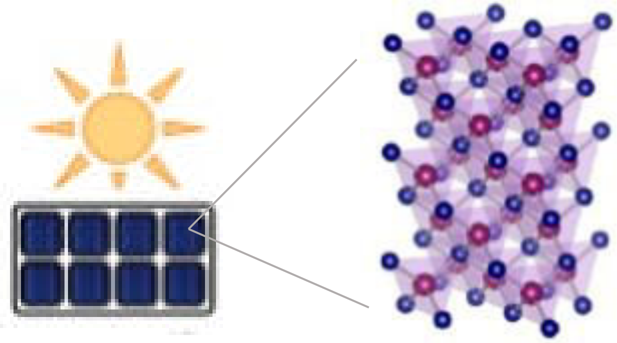
$$\hat{\mathcal{H}}\Psi(\mathbf{r}_1, \mathbf{r}_2, ..., \mathbf{r}_N) = E\Psi(\mathbf{r}_1, \mathbf{r}_2, ..., \mathbf{r}_N)$$



Source: KTH

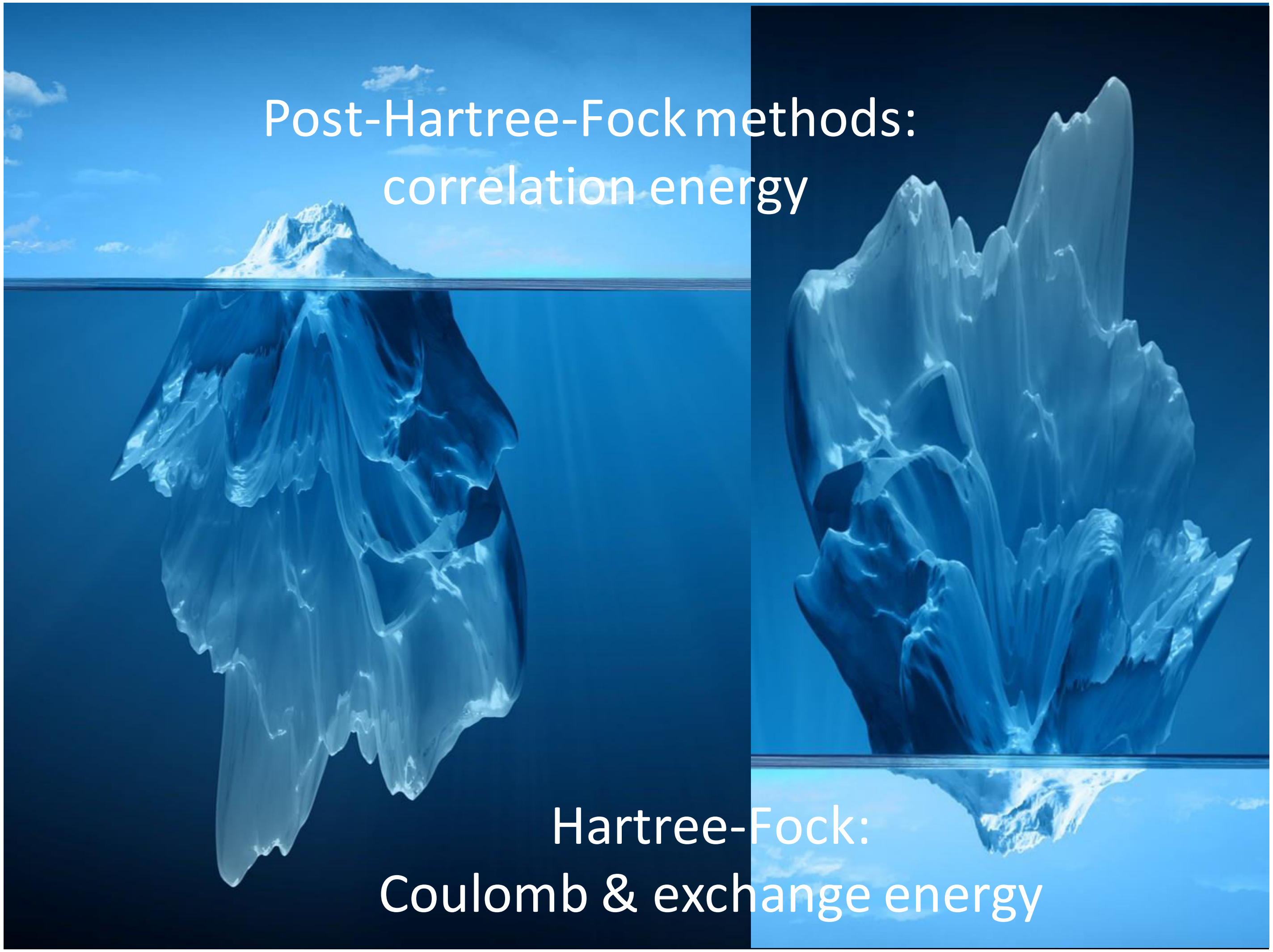


Source: IMDEA Materials Institute



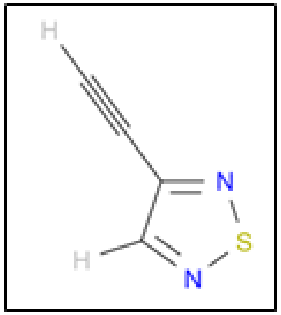
Energy Computation

Energy **Computation Cost**



Post-Hartree-Fock methods:
correlation energy

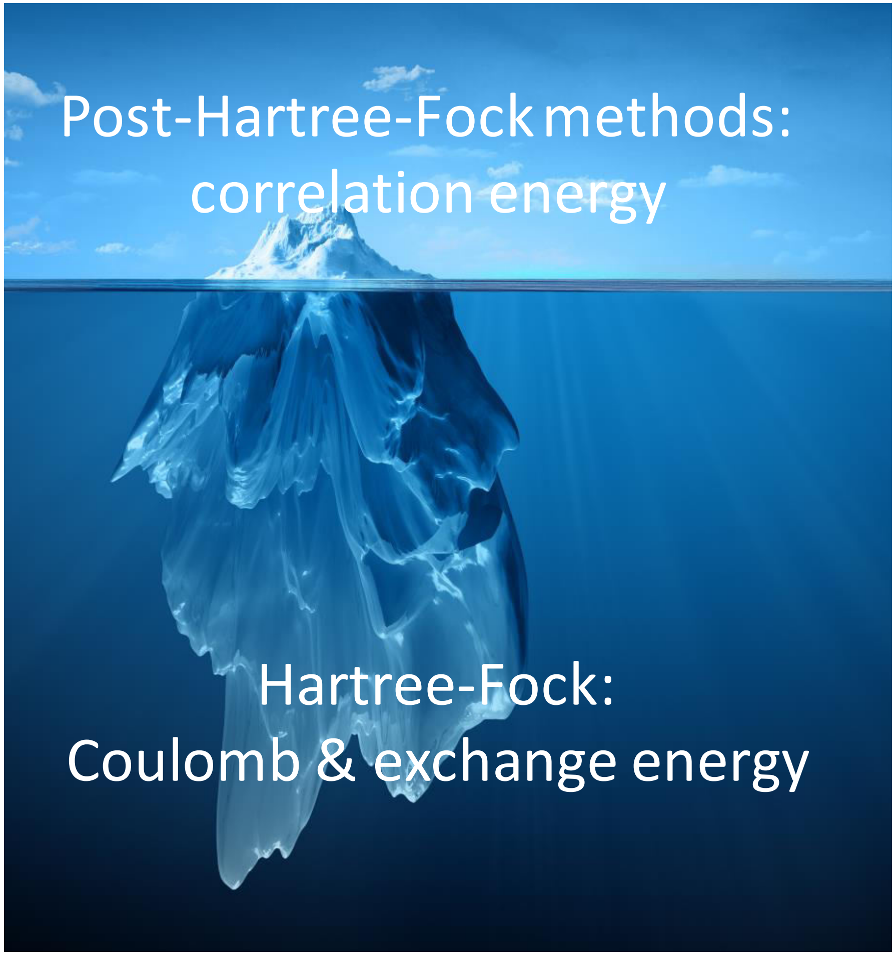
Hartree-Fock:
Coulomb & exchange energy



26 hours

2 minutes

Correlation Energy Computation



Chemical accuracy: 0.1 to 2 mH; not satisfied by HF.

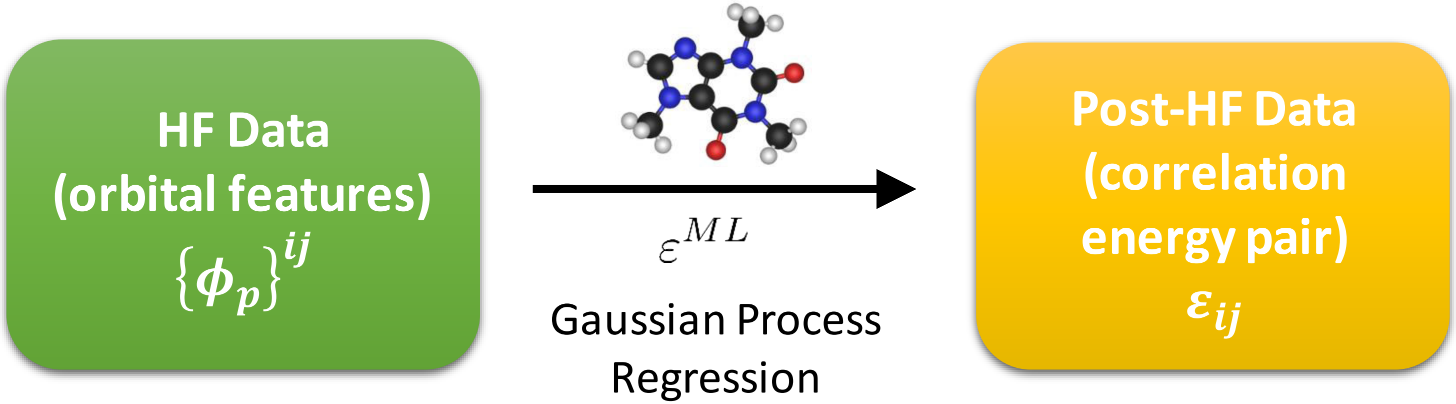
Hence, the need for expensive post-HF methods.

Post-HF methods estimate *correlation energy:*

$$E_c = \sum_{ij}^{occ} \varepsilon_{ij}$$

$$\varepsilon_{ij} = \varepsilon\left[\{\phi_p\}^{ij}\right]$$

Data-driven Approach



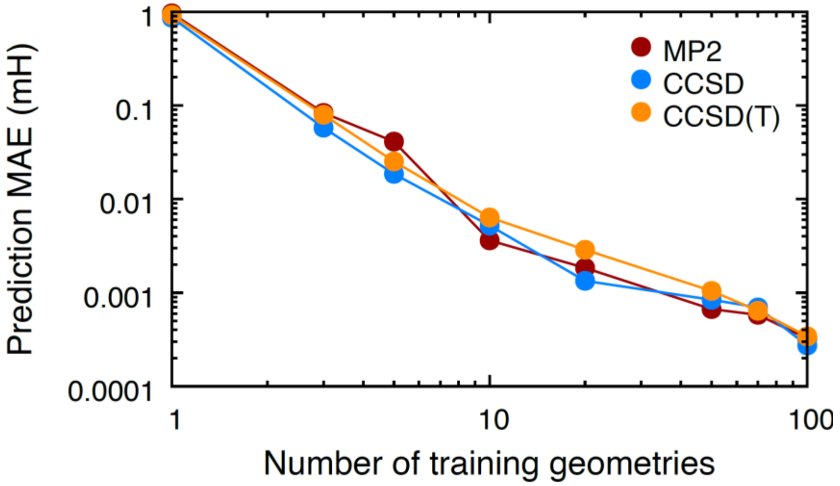
HF Data (orbital features) $\{\phi_p\}^{ij}$

ε^{ML}

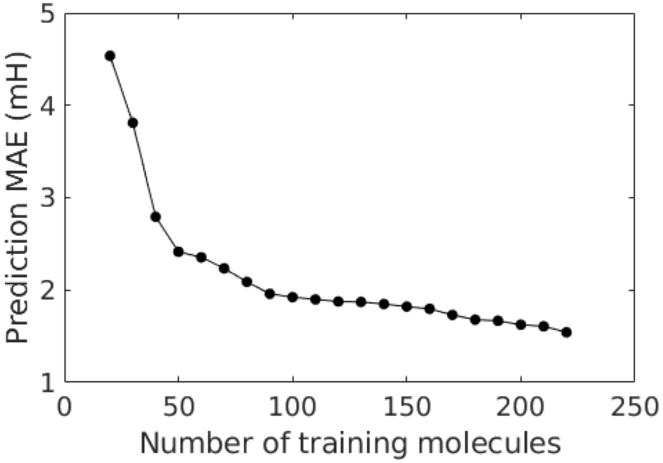
Gaussian Process Regression

Post-HF Data (correlation energy pair) ε_{ij}

Prediction on different H_2O geometries



Prediction on different molecules



Predict on 7000 organic molecules with at least 7 heavy atoms

Cheng, Welborn and Miller III, *ArXiv* (2019)

Data-driven Approach

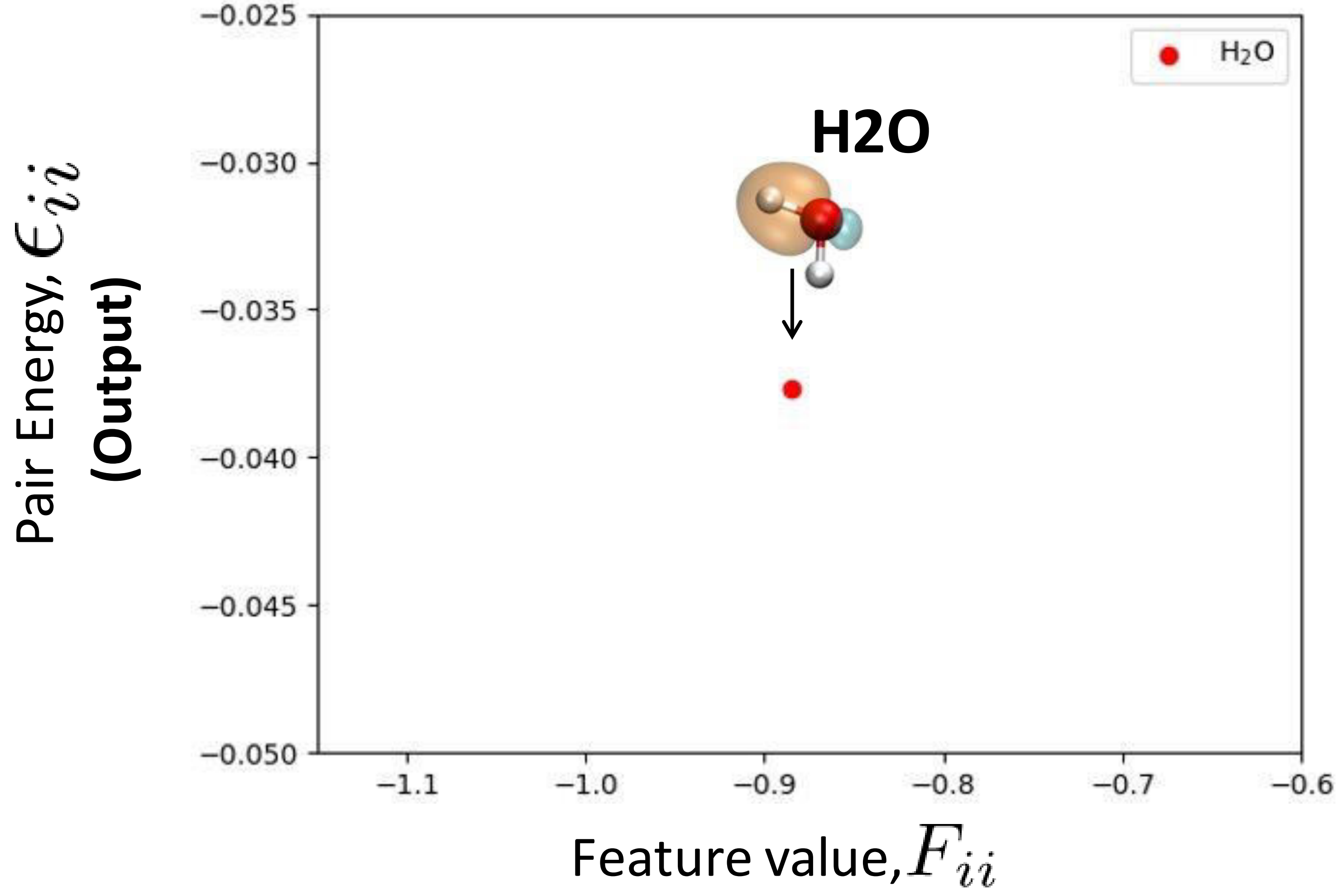
Our goal: scale to billions of molecules.

Approaches:

- Scalability & Transferability
  - Issue: GPR is constrained by memory and time
  - Approach: utilize clusters & their local linearity
- Leverage Multi-fidelity Data
  - Issue: different data volume based on fidelities
  - Approach: learn residual model between fidelities

Key Observation

Sigma-bond Orbital

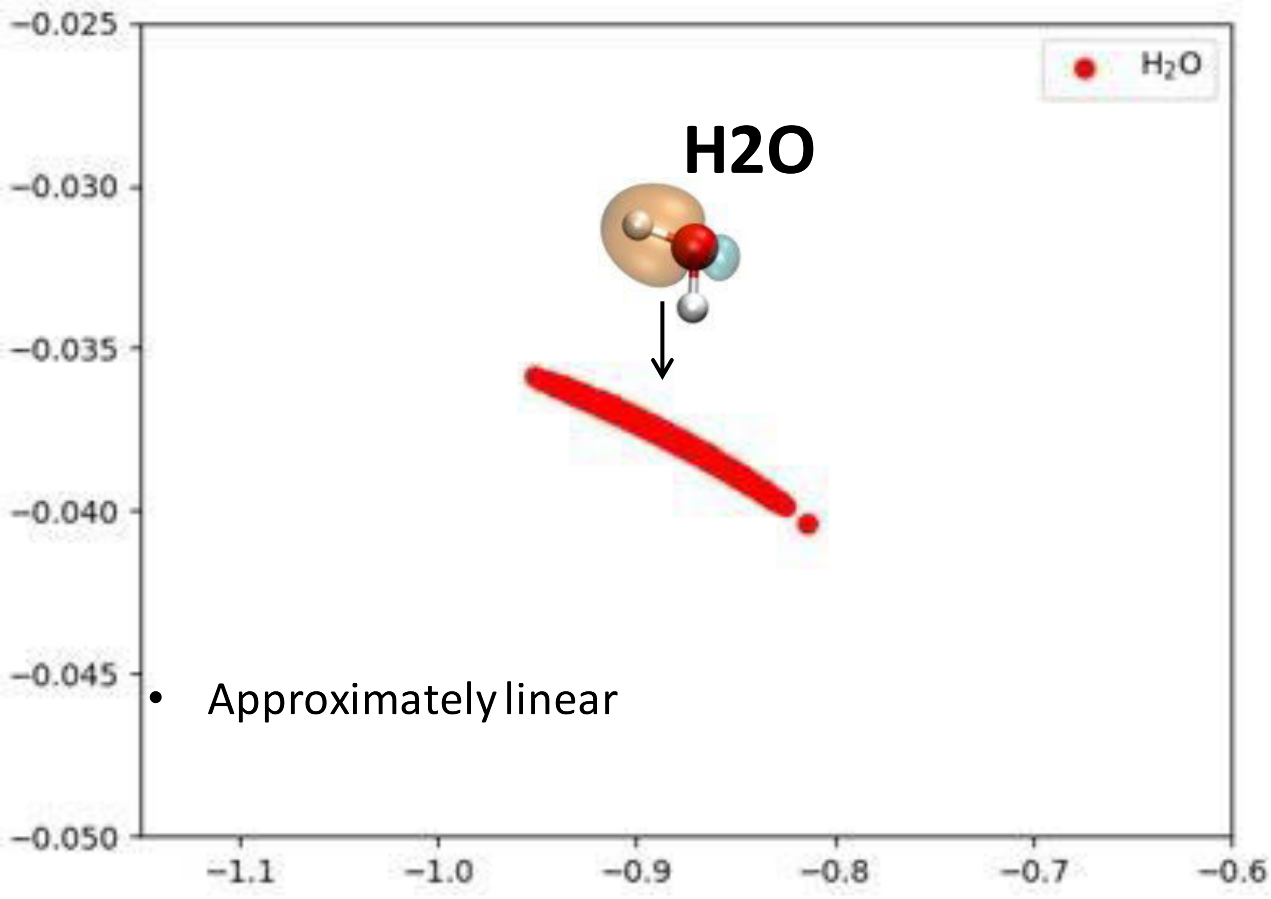


Pair Energy, ϵ_{ii}
(Output)

Feature value, F_{ii}
(Input)

Key Observation

Sigma-bond Orbital



H2O

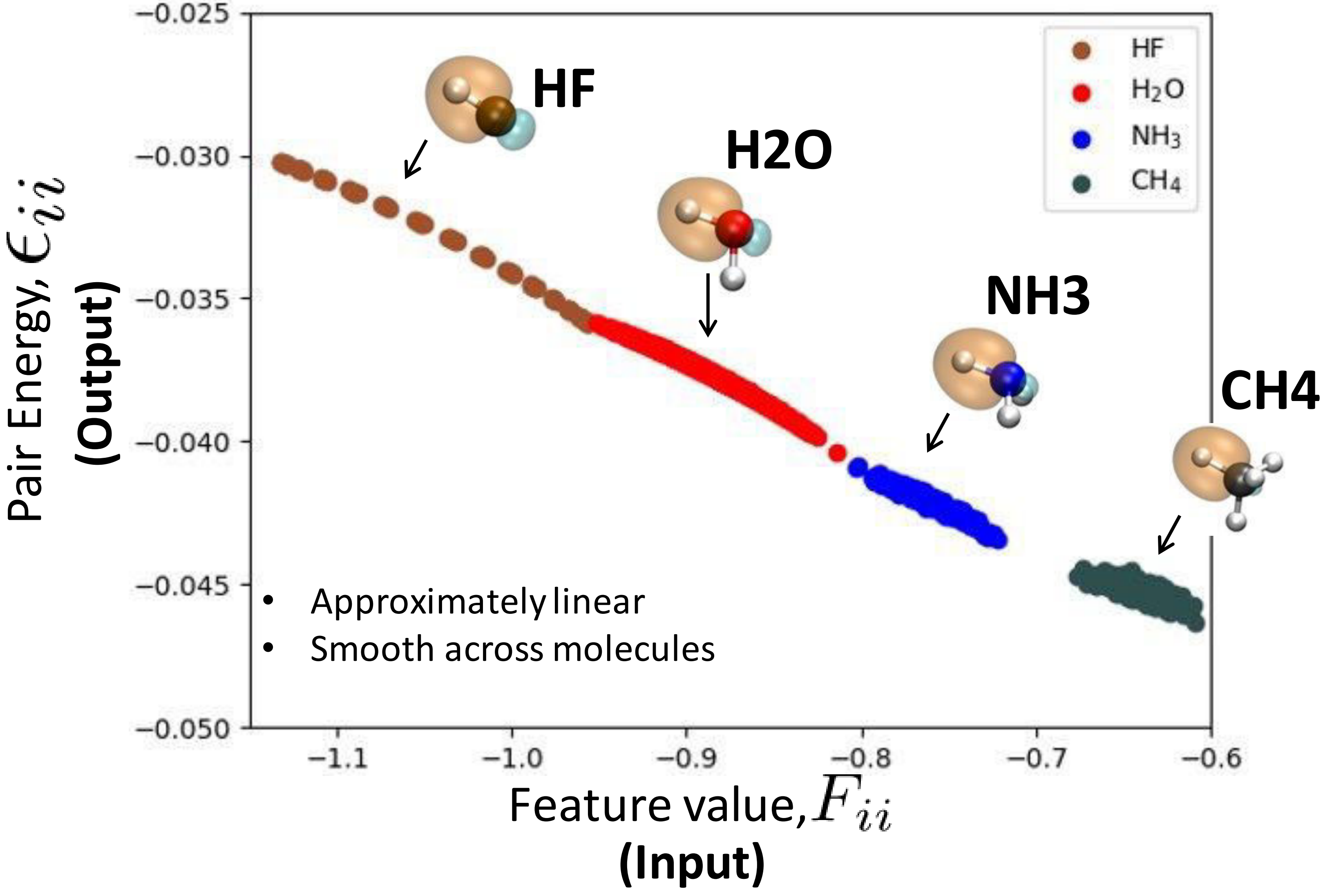
- Approximately linear

Pair Energy, ϵ_{ii} **(Output)**

Feature value, F_{ii} **(Input)**

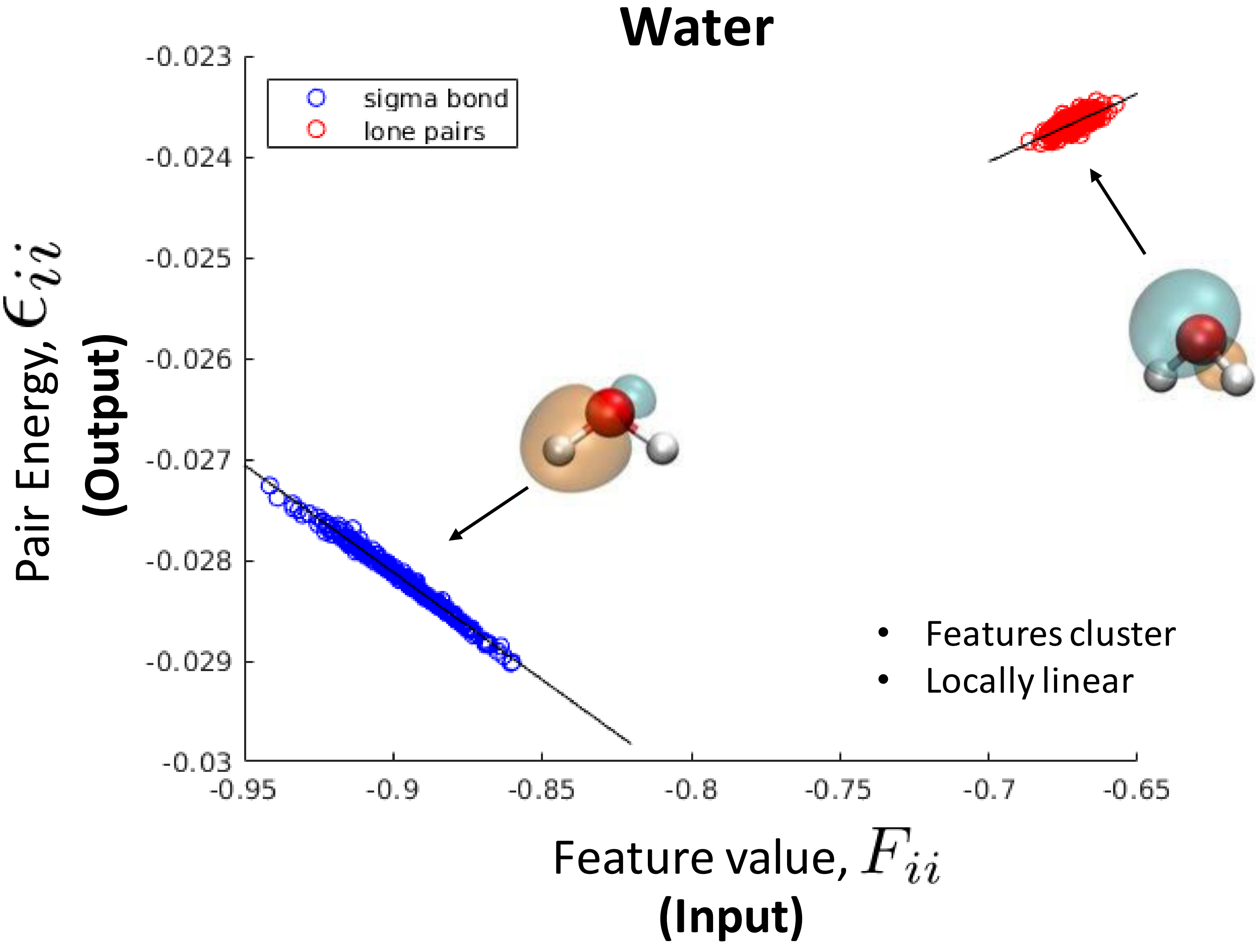
Key Observation

Sigma-bond Orbital



- Approximately linear
- Smooth across molecules

Key Observation



- Features cluster
- Locally linear

Clustering

- Pairs have different chemical properties
  - Sigma bond, lone pairs
  - Learning specialized models likely beneficial

- Reduce computational resource demand
  - Partition large datasets into smaller ones
  - Enable parallel training and large scale-up factor

- Learn connections among molecules
  - Inspect clusters to gain insights

Regression Clustering

Objective: $$\underset{S_1,\ldots,S_k}{\arg\min} \sum_{j=1}^{k} c(\{x_l, y_l\}_{l \in S_j})$$

Cost: $$c(\{x_l, y_l\}_{l \in S_j}) = \sum_{l \in S_j} |f_j(x_l) - y_l|^2$$

Regressor: $$f_j = \text{ OLS solution of } \{x_l, y_l\}_{l \in S_j}$$

Solution with greedy algorithm:

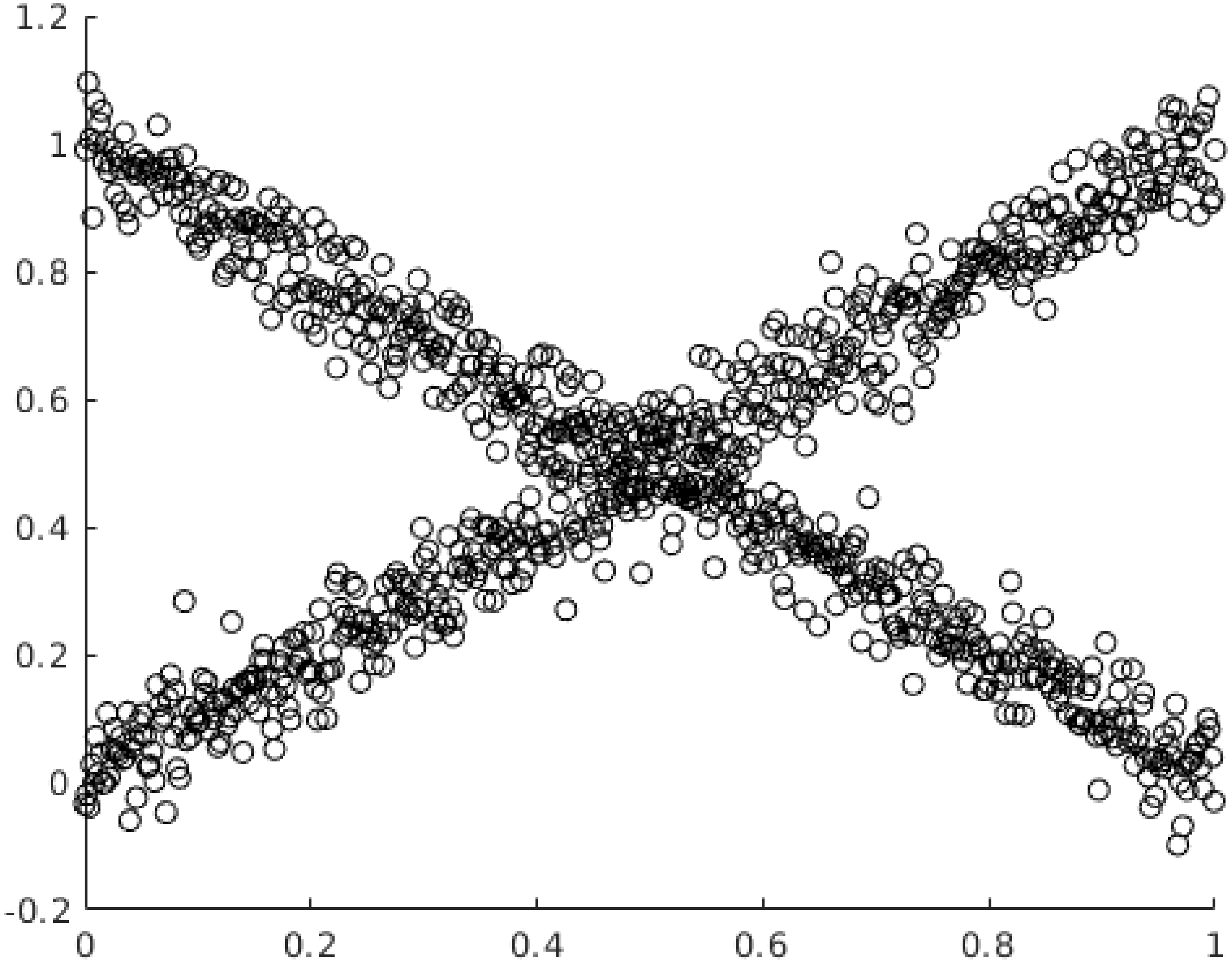
Iterate until converged:

$$S_j = \left\{ l : \underset{n \in \{1,\ldots,k\}}{\arg\min} |f_n(x_l) - y_l|^2 = j \right\}$$

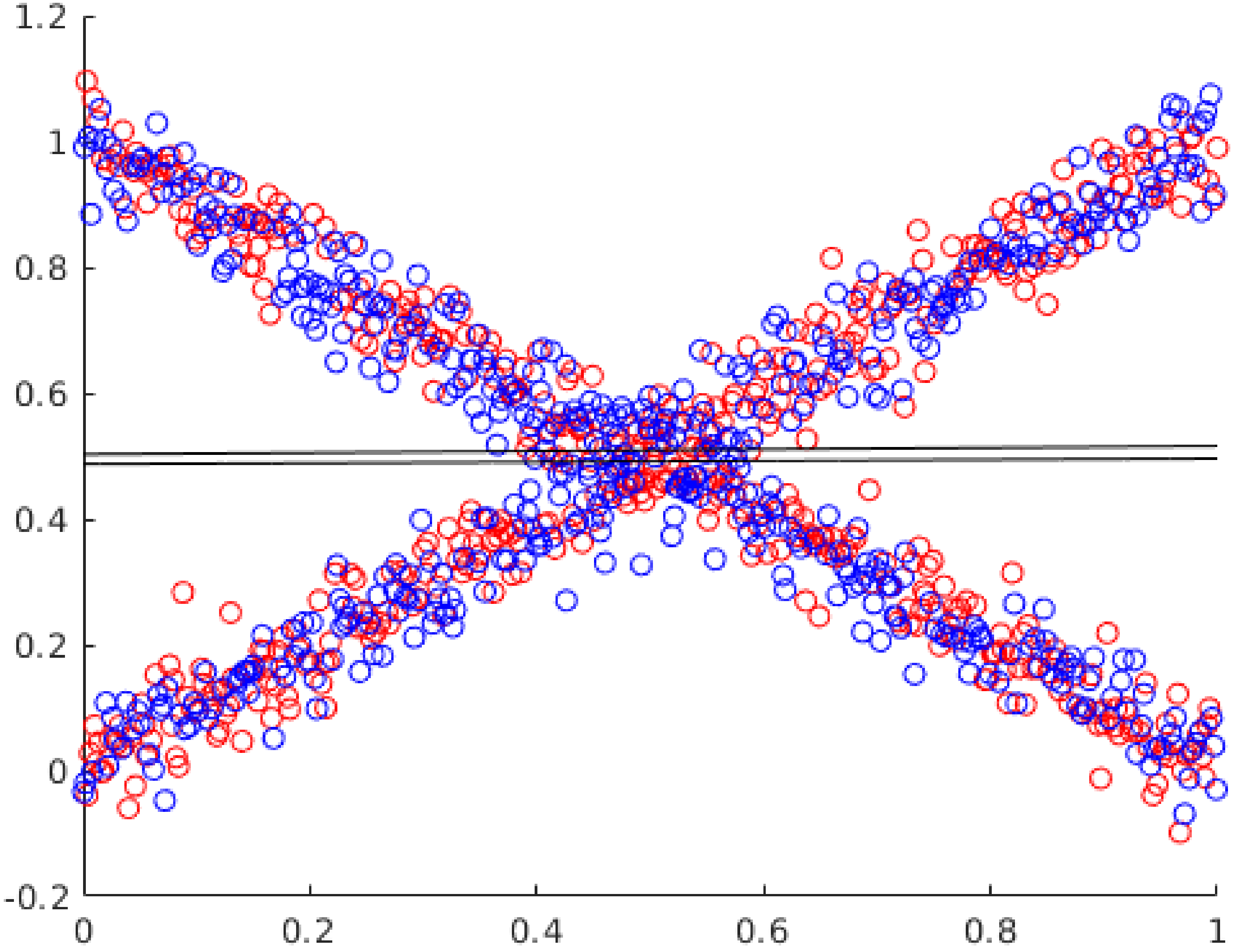
$$f_j = \text{ OLS solution of } \{x_l, y_l\}_{l \in S_j}$$

Spath, H. (1979)

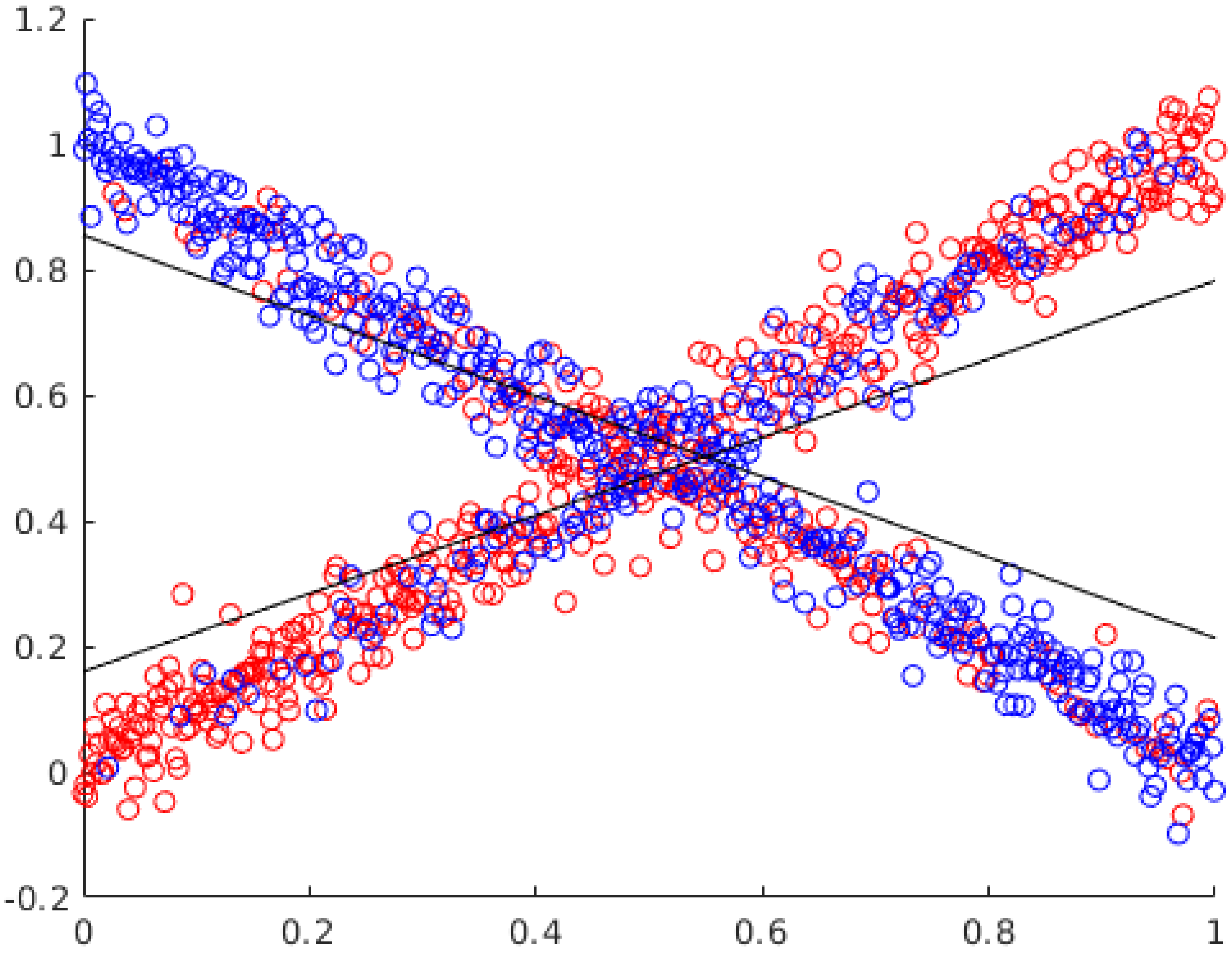
Regression Clustering



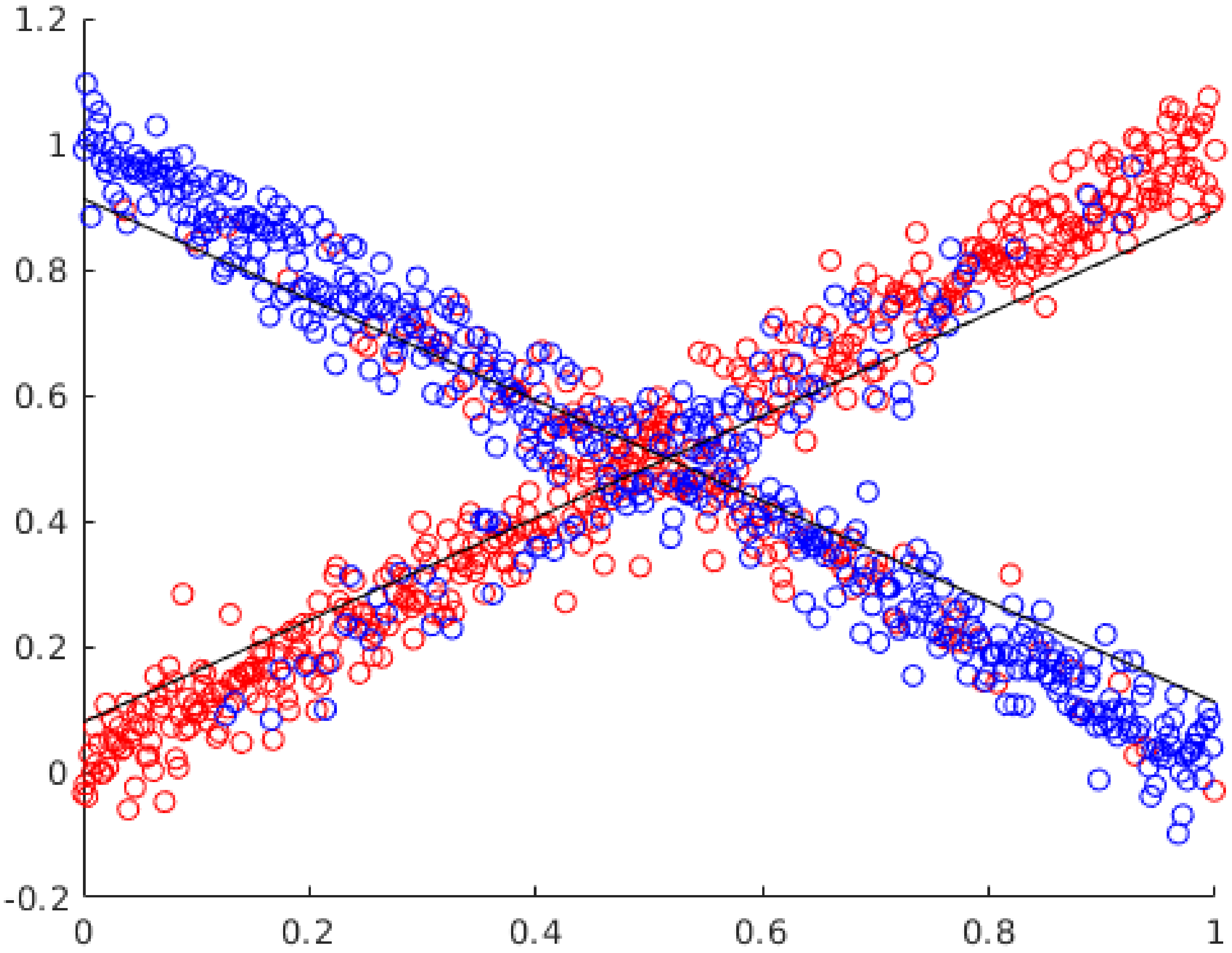
Regression Clustering



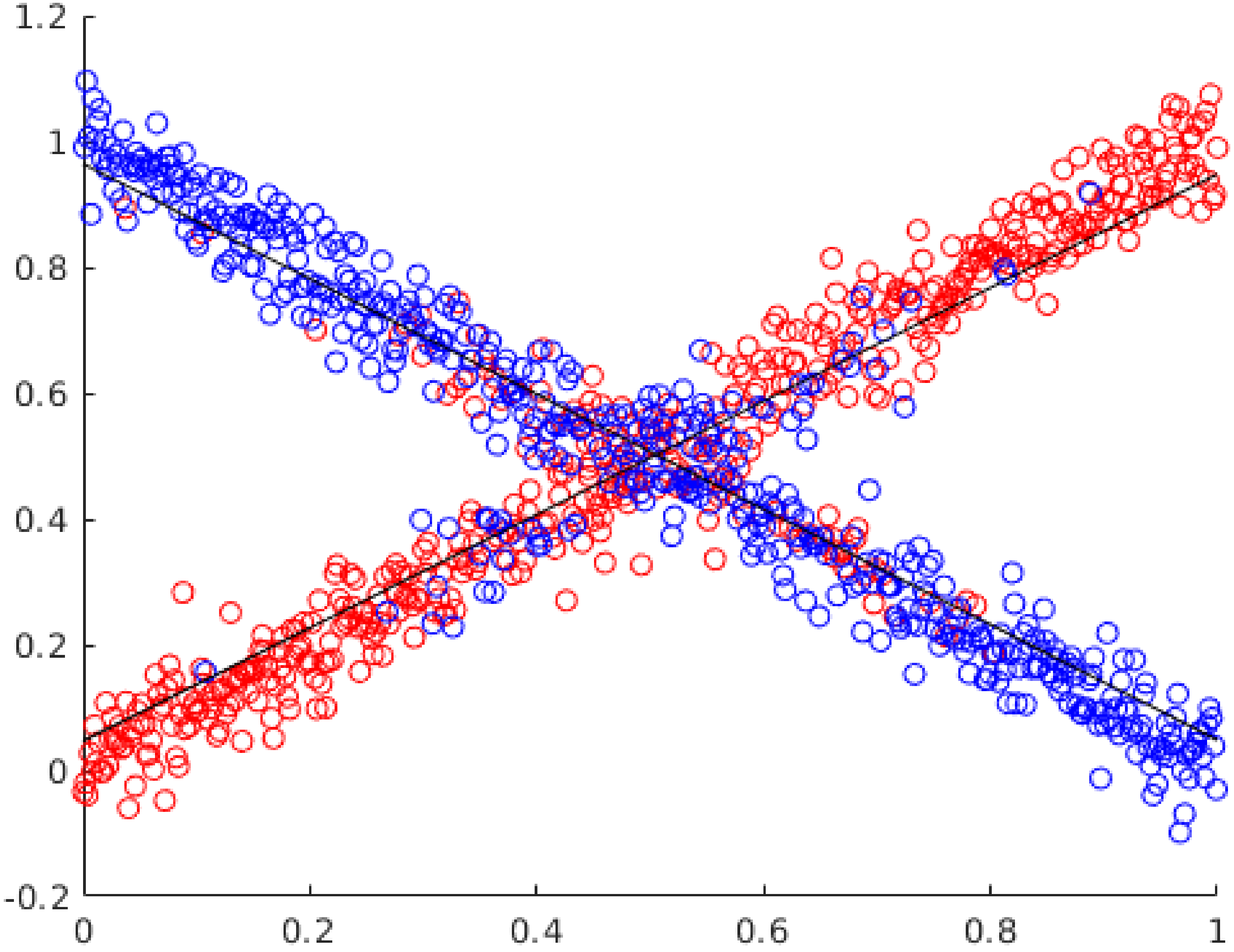
Regression Clustering



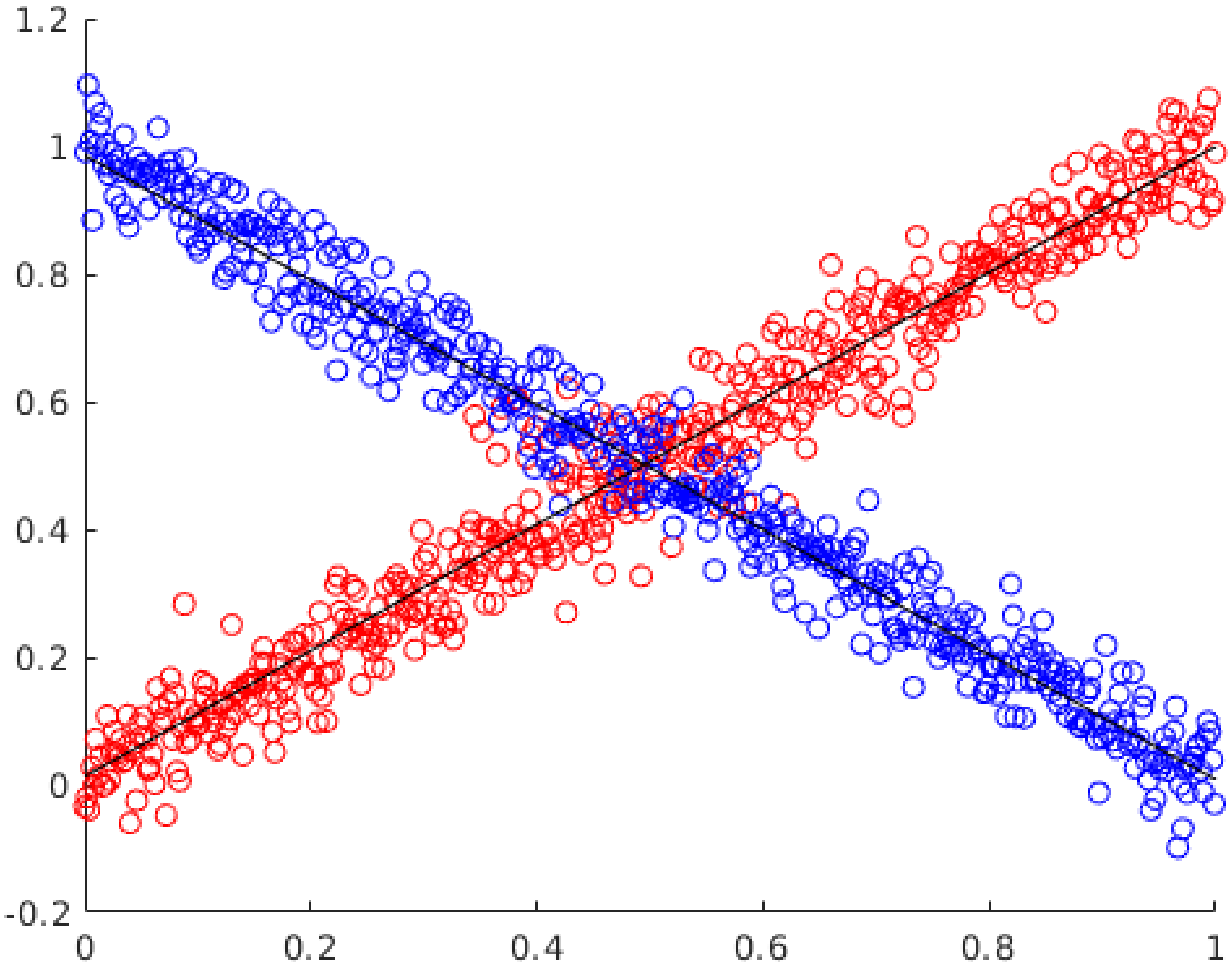
Regression Clustering



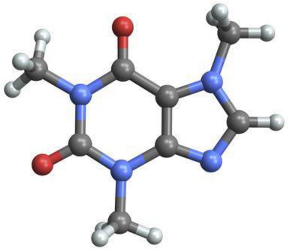
Regression Clustering



Regression Clustering

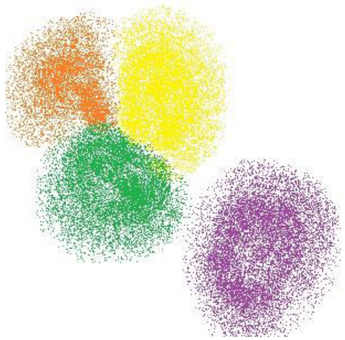


Training Process



$$\{\phi_k\}^{ij},\ \epsilon_{ij}$$

(Regression clustering)



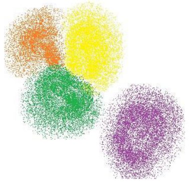
(Local regression)

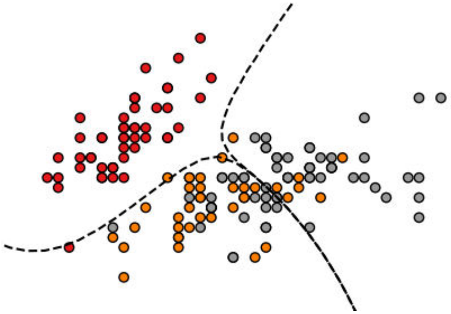
$$\epsilon(\{\phi_k\}^{ij})$$

Predicting

$$\{\phi_k\}^{ij}$$

(Classifier)

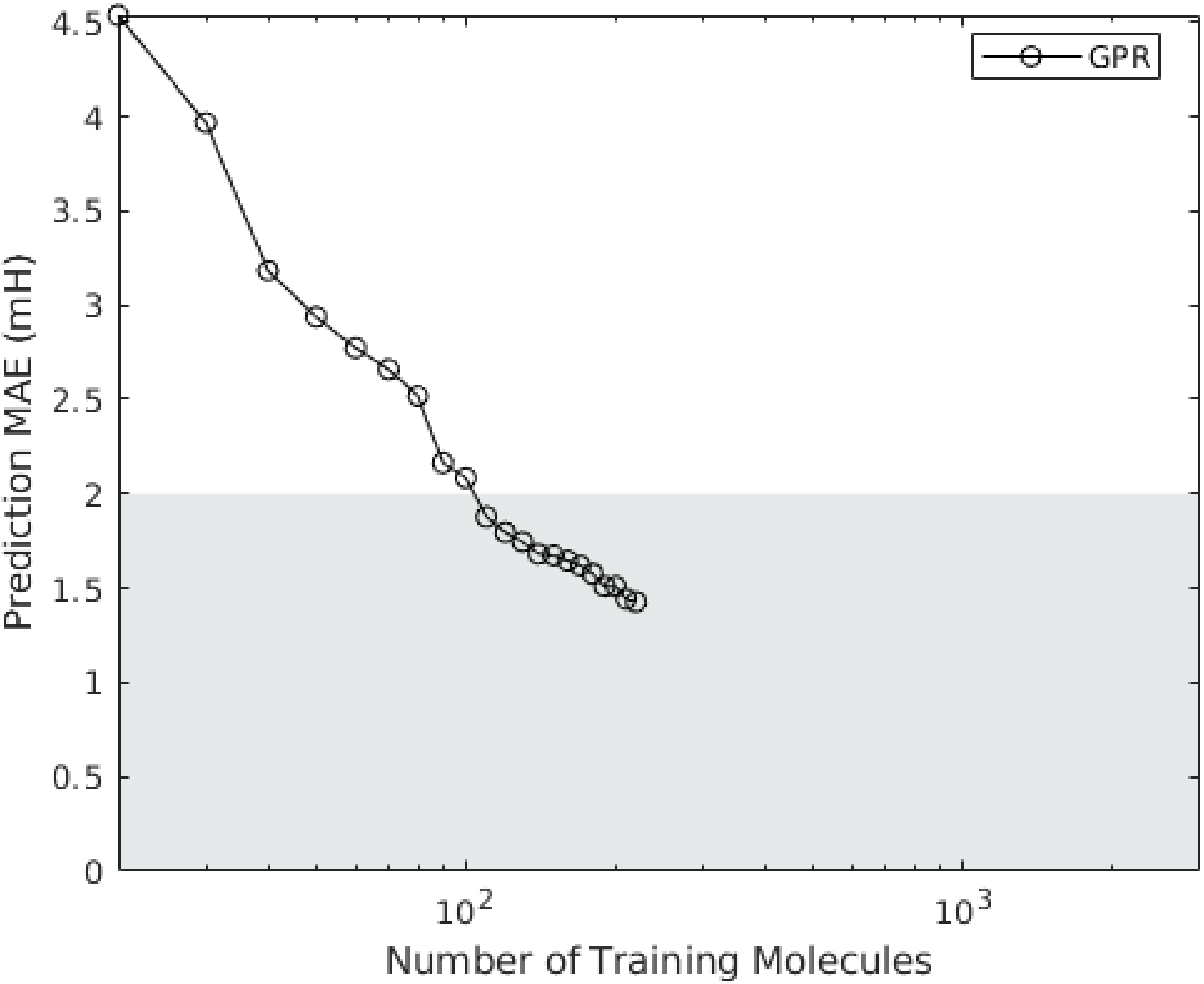




(Local regression models)

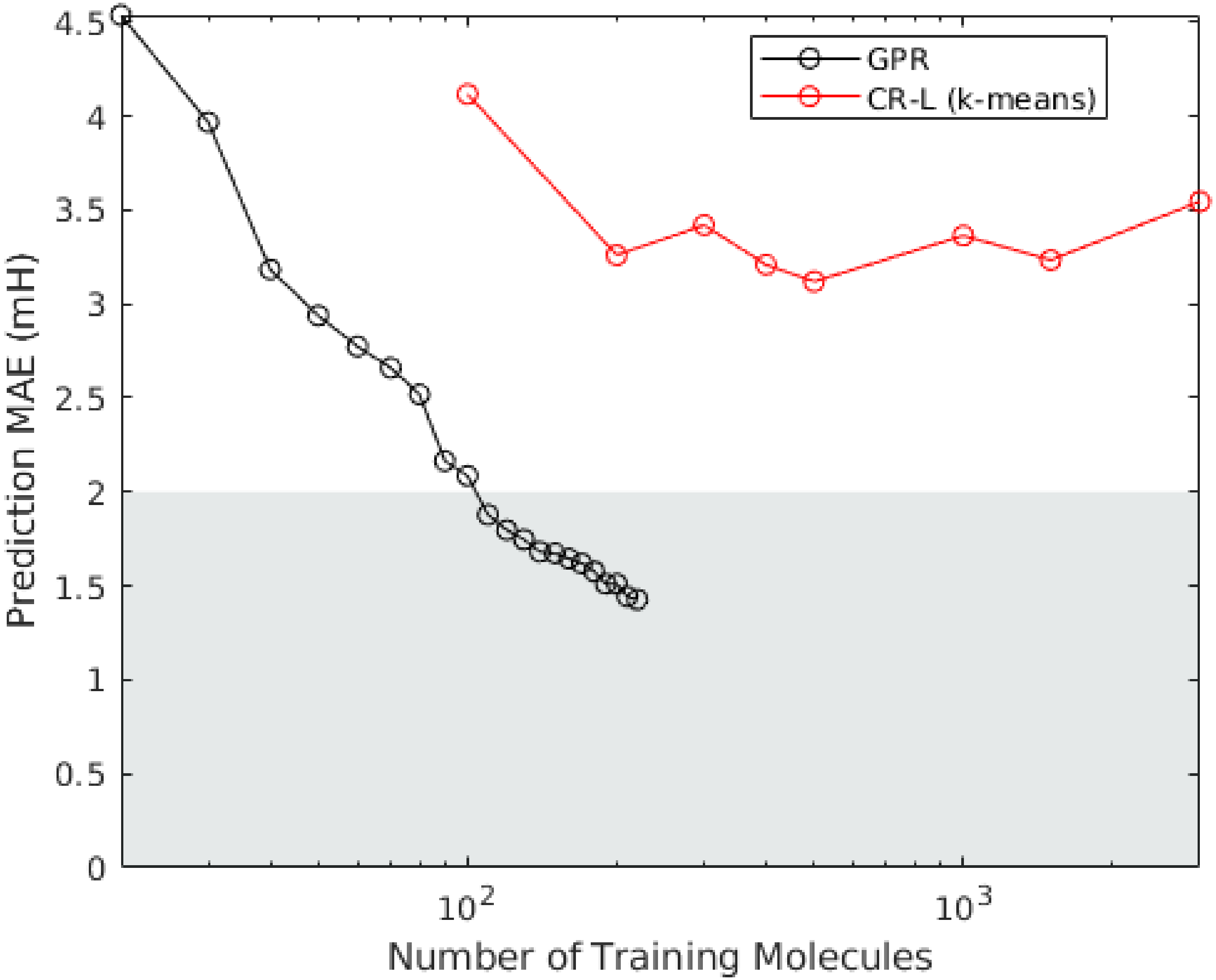
$$\epsilon\left(\{\phi_k\}^{ij}\right) \approx \epsilon_{ij}$$

QM7B – Training Molecules



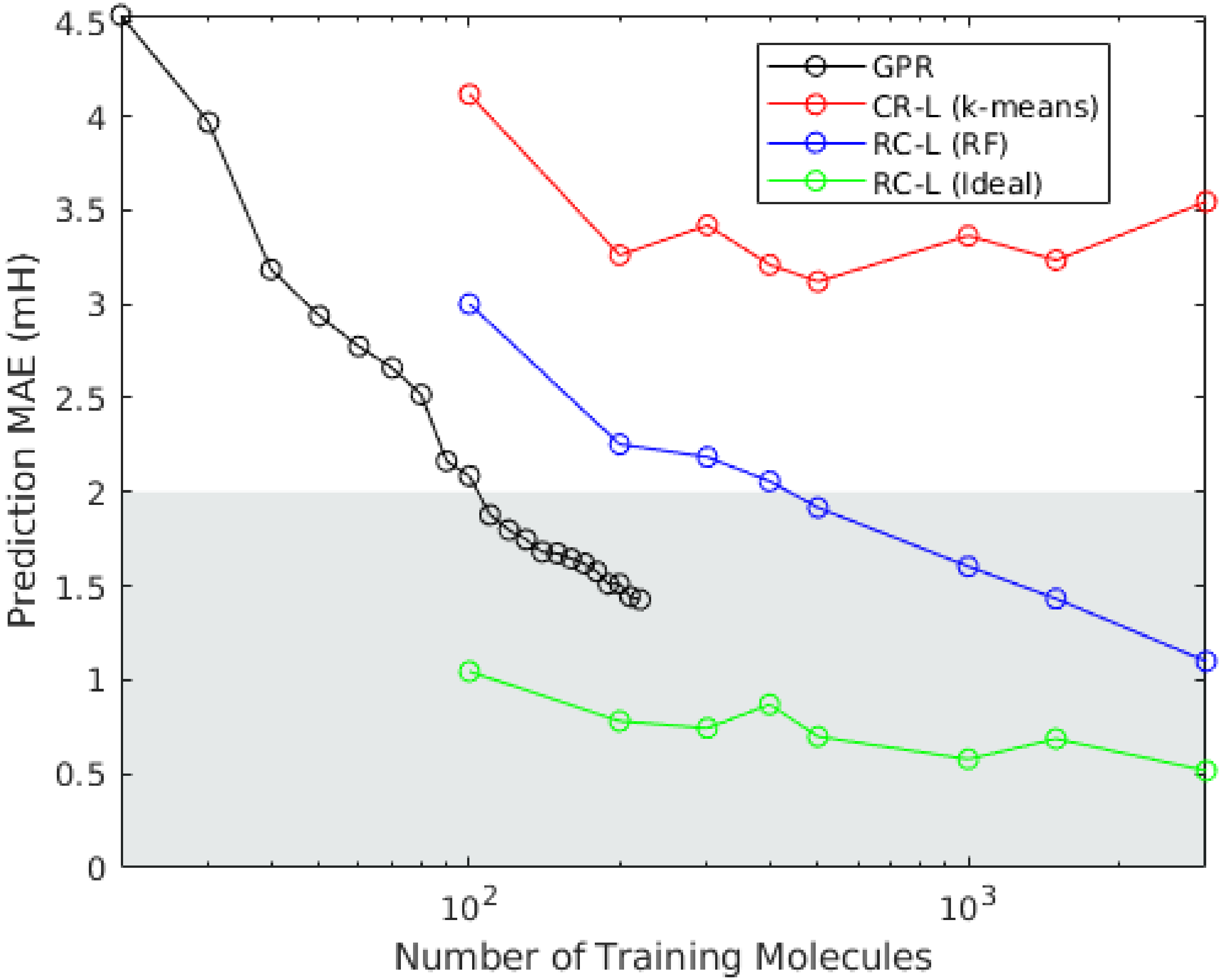
Cheng, Welborn and Miller III, *ArXiv* (2019)

QM7B – Training Molecules



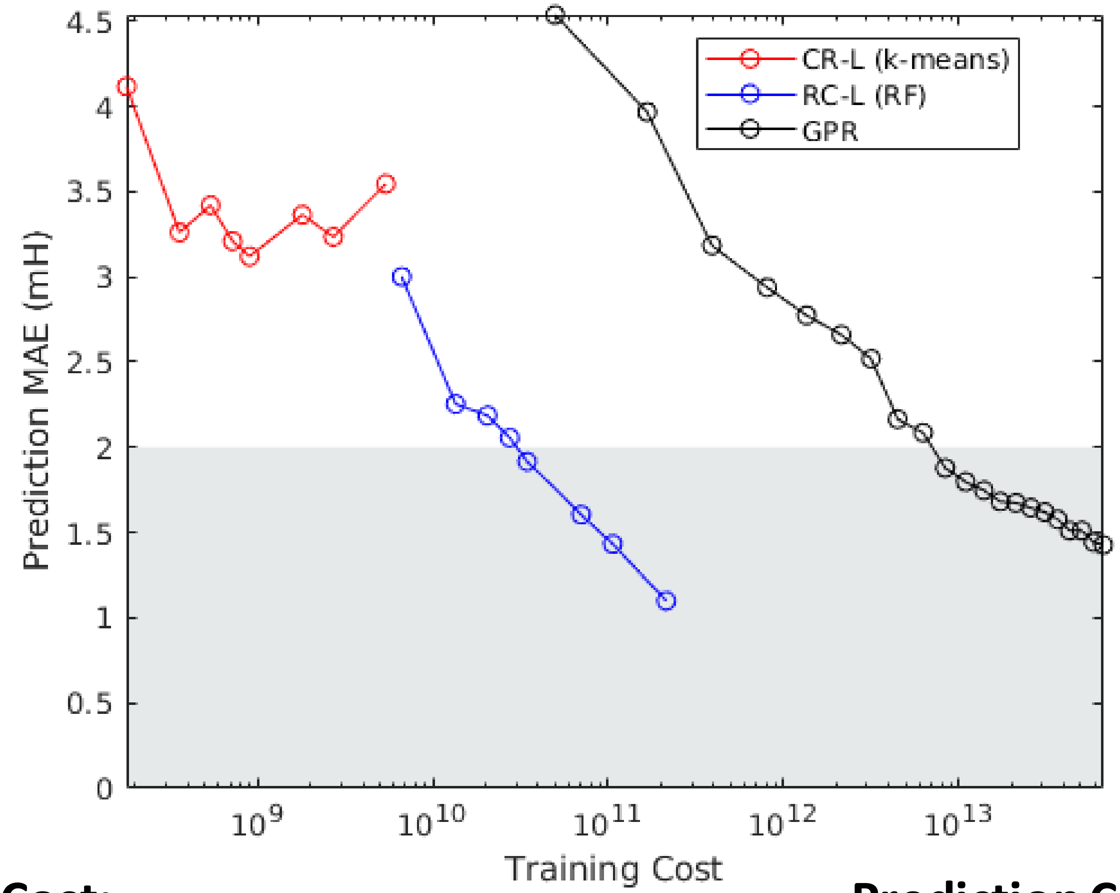
Cheng, Welborn and Miller III, *ArXiv* (2019)

QM7B – Training Molecules



Cheng, Welborn and Miller III, *ArXiv* (2019)

QM7B – Cost



Storage Cost:

RC-L (RF): $\mathcal{O}(1)$

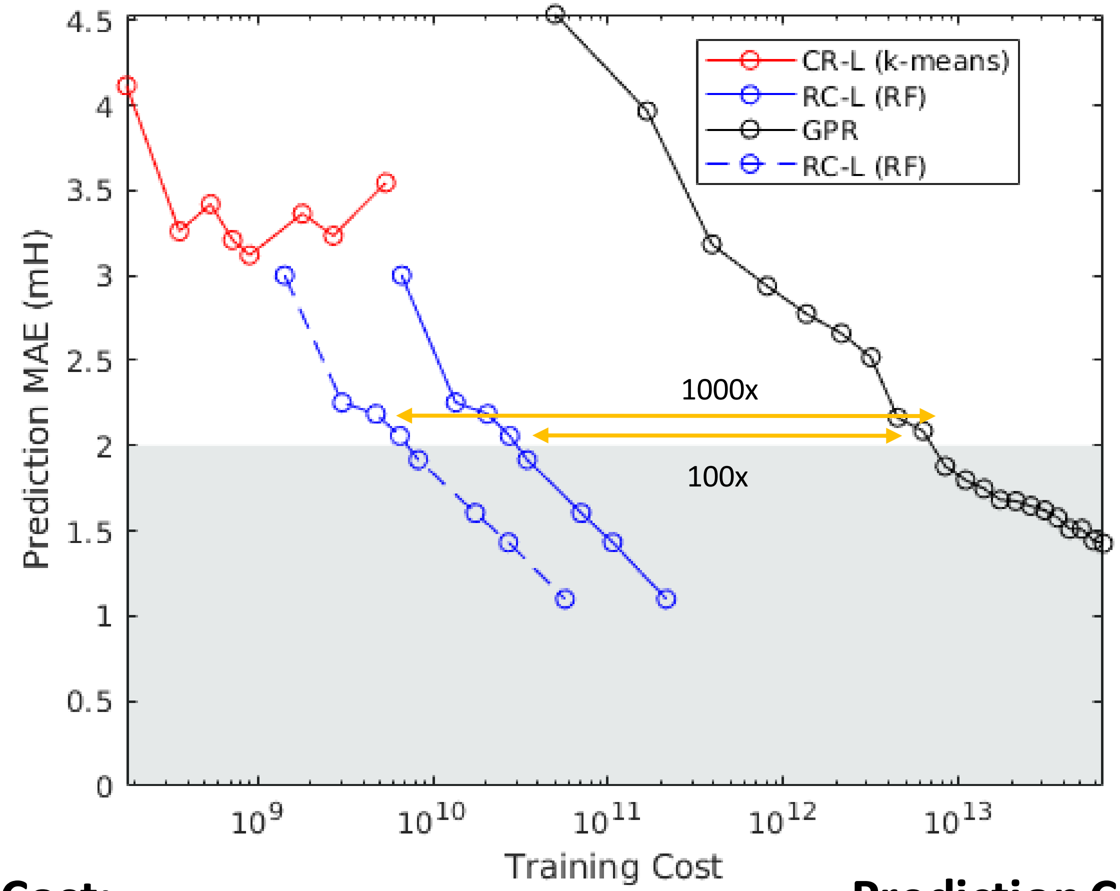
GPR: $\mathcal{O}(N^2)$

Prediction Cost:

RC-L (RF): $\mathcal{O}(1)$

GPR: $\mathcal{O}(N)$

QM7B – Cost



Storage Cost:

RC-L (RF): $\mathcal{O}(1)$

GPR: $\mathcal{O}(N^2)$

Prediction Cost:

RC-L (RF): $\mathcal{O}(1)$

GPR: $\mathcal{O}(N)$

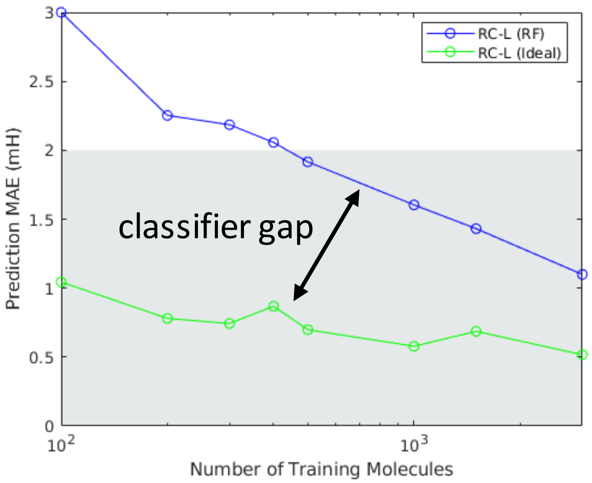
Method Overview

- Advantages:
  - Cheap to train/store/predict
  - Parallelizable
  - Can utilize big data
  - Chemically interpretable
  - Well-understood UQ
- Disadvantages:
  - More data to be chemically accurate (w/ linear regressors)
  - Dependent on the quality of the classifier
  - Not smooth at cluster boundaries (w/ current implementation)
  - Sensitive to initialization (local minima)

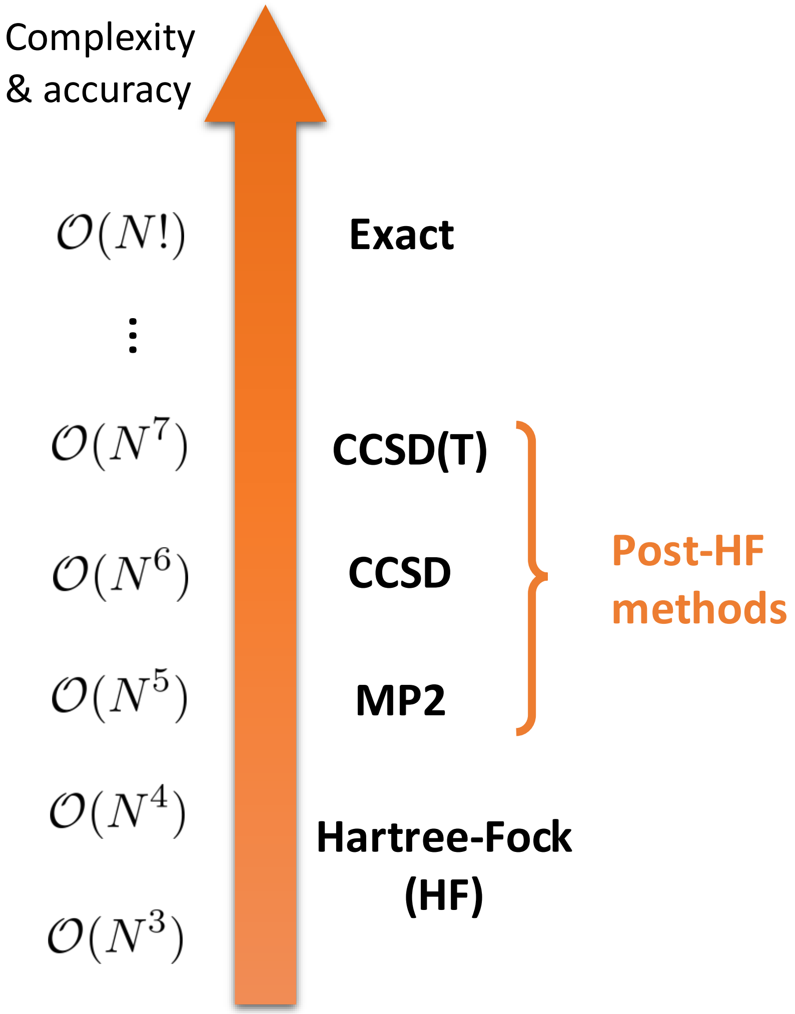
Ongoing Next Steps

- Better regressors
  - Capture non-linearity
  - Smooth cluster transitions

- Better classifier
  - Deep networks
  - Bayesian consensus
  - Cluster combinations



Multi-fidelity Hierarchy



Complexity & accuracy

$\mathcal{O}(N!)$ Exact

⋮

$\mathcal{O}(N^7)$ CCSD(T)

$\mathcal{O}(N^6)$ CCSD

$\mathcal{O}(N^5)$ MP2

Post-HF methods

$\mathcal{O}(N^4)$

$\mathcal{O}(N^3)$ Hartree-Fock (HF)

	QM7b	GDB-13
HF	1 min	4 min
MP2	1 hour	20 hours
CCSD	5 hours	9 days
CCSD(T)	1 day	3 years

Impossible to scale!

Leverage Multi-fidelity Data

- Data volume decreases as complexity increases.

- Can we bootstrap a prediction model for high-fidelity data (e.g., CCSD(T)) from low-fidelity data (e.g. MP2)?

- "Generating" more high-fidelity data to train a more accurate high-fidelity model.

Mathematical Formulation

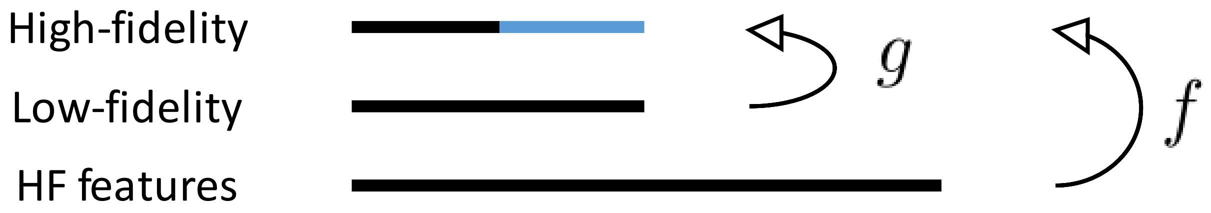
Learn direct mapping from HF features to high-fidelity data:

$$\min_{f\in\mathcal{F}} \sum_{i=1}^{n} (f(x_i) - \epsilon_i^{high})^2$$

Learn a residual model between low and high-fidelity data:

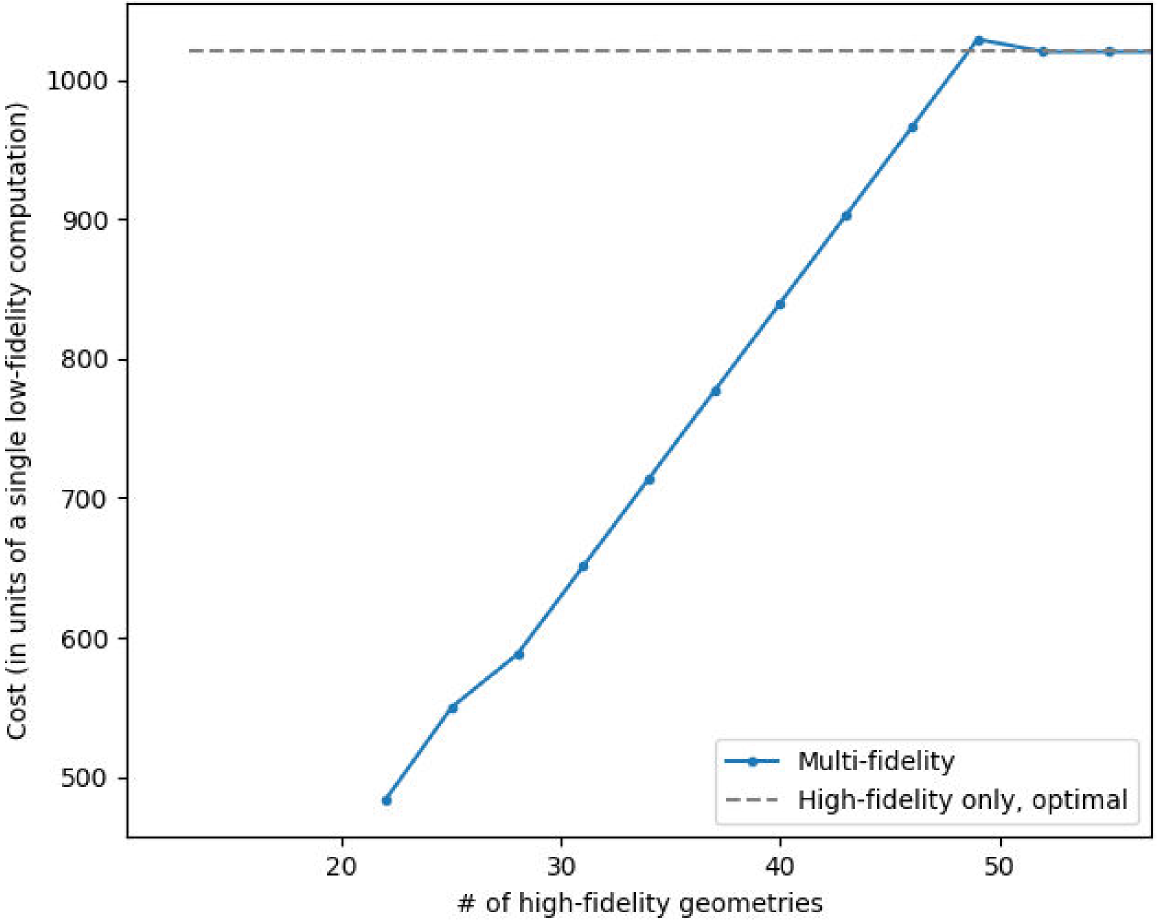
$$\delta = \epsilon^{high} - \epsilon^{low}$$
$$\min_{g\in\mathcal{F}} \sum_{i=1}^{n} (g(x_i) - \delta_i)^2$$

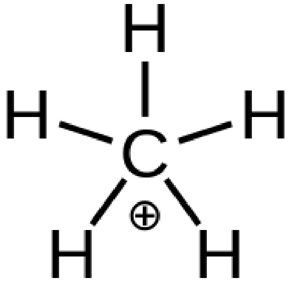
"Generate" High-fidelity Data



Q: can we train a more accurate high-fidelity model with the generated data?

Results: Chemical Accuracy



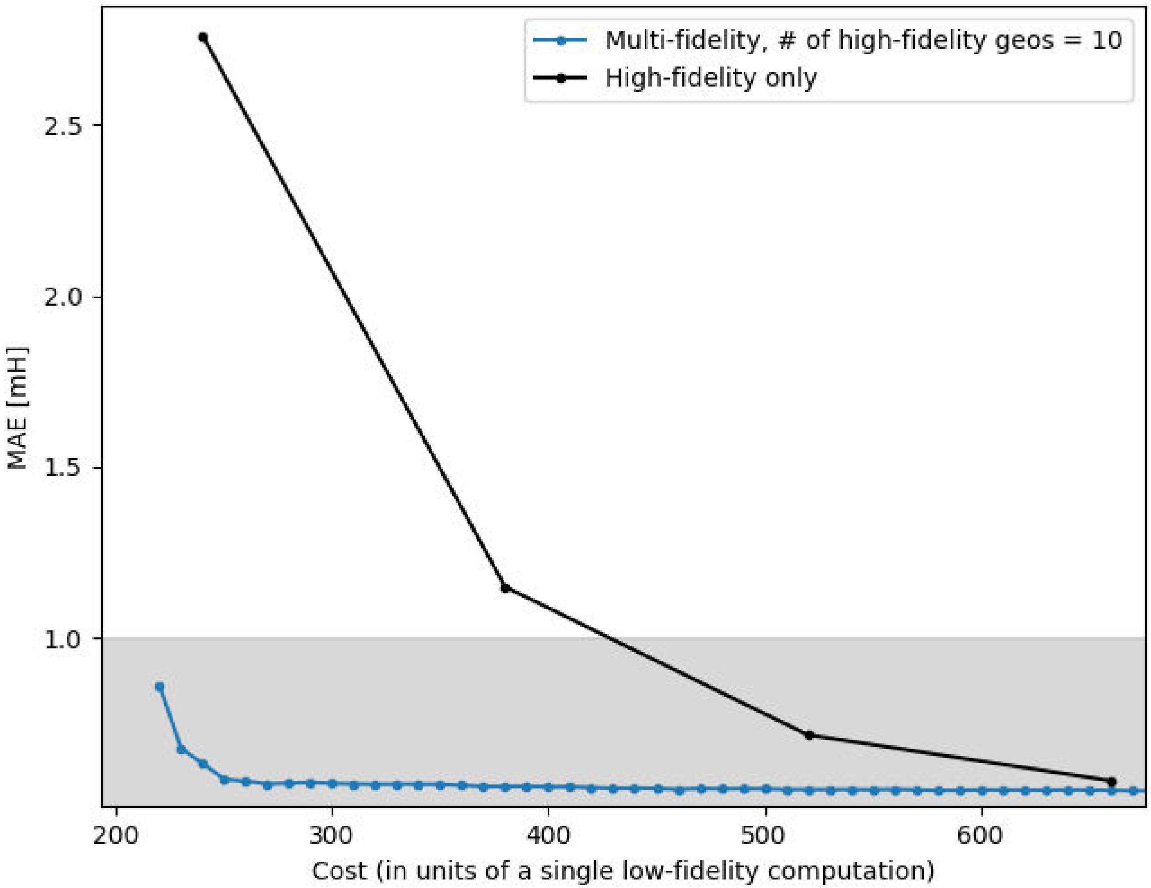


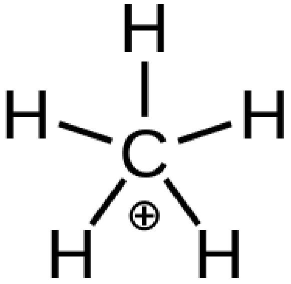
cost(high-fidelity) = 20 * cost(low-fidelity)

Testing on 4200 new geometries.

For a fixed number of high-fidelity geometries, determine the computational cost upon increasing the number of low-fidelity calculations to achieve a chemical accuracy of 0.5 mH.

Results: Varying Costs





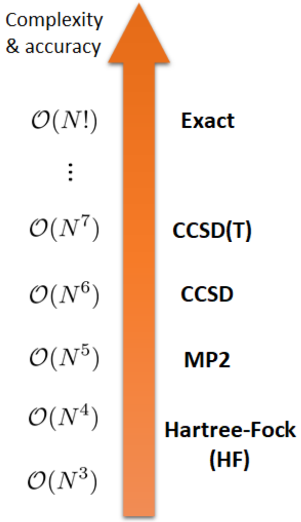
cost(high-fidelity) = 20 * cost(low-fidelity)

Testing on 4200 new geometries.

Gradually increase low-fidelity data and the multi-fidelity model achieves lower error at lower cost.

Future Directions

- Beyond two fidelities:
  - Go up towards a good approximation for exact computation.
  - What is the optimal way to define residuals?
    - CCSD(T) = MP2 + (CCSD(T) - MP2)
    - CCSD(T) = CCSD + (CCST(T) - CCSD)
    - CCSD(T) = MP2 + (CCSD – MP2) + (CCSD(T) - CCSD)



Complexity & accuracy

$\mathcal{O}(N!)$ Exact

$\vdots$

$\mathcal{O}(N^7)$ CCSD(T)

$\mathcal{O}(N^6)$ CCSD

$\mathcal{O}(N^5)$ MP2

$\mathcal{O}(N^4)$ Hartree-Fock (HF)

$\mathcal{O}(N^3)$

- Basis set hierarchy:
  - Varying the granularity of discretization to reduce costs of generating molecular orbital features.

	QM7b	GDB-13
HF	1 min	4 min

- Widely applicable:
  - Any application that exhibits hierarchy of different quality data can adopt our methodology.

Conclusion

Our goals:

- Scalability & Transferability ✓
  - Scale to 10X data
  - Transfer better to new molecules, 30% error reduction
  - Reduce training computation cost by a factor of 1000
- Leverage Multi-fidelity Data ✓
  - Chemically accurate high-fidelity model at 50% cost