Science Society Clubhouse

Accurate and transferable molecular-orbital-based machine learning (MOB-ML) for molecular modelling

Sherry Lixue Cheng

Miller group
Division of Chemistry & Chemical Engineering
California Institute of Technology

Caltech

State-of-the-art of quantum simulations

$$H\Psi = E\Psi$$

Largest treatable system size

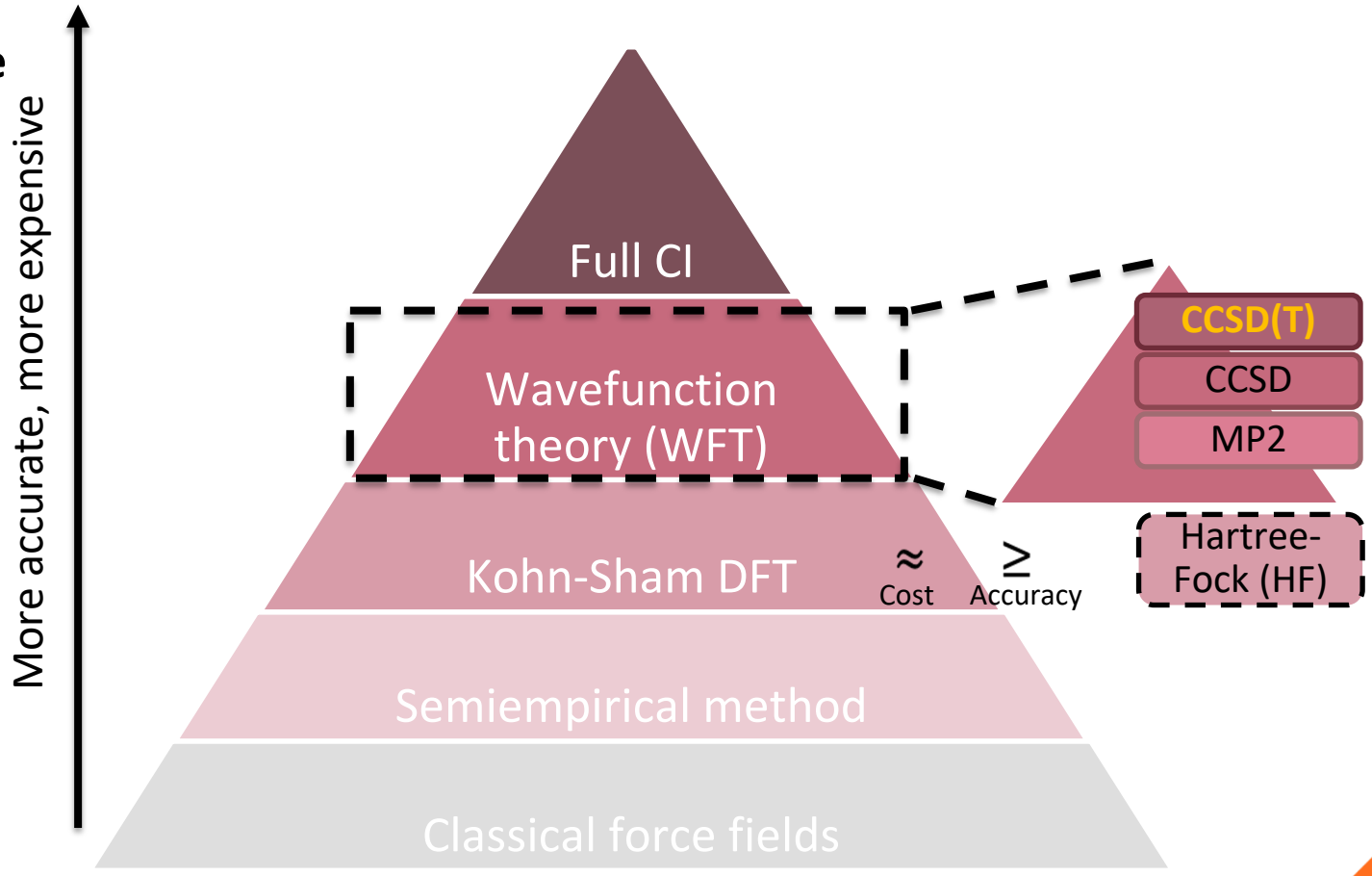
few e^-

~20 Heavy atoms

~200 Heavy atoms

Up to 10^6 atoms

10^8 atoms

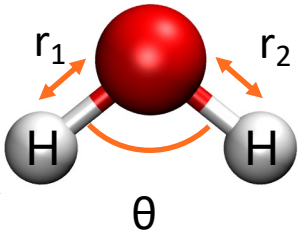


Caltech

ML for quantum chemistry: Many promising trials

Usual strategy
ML with atom-based features

Usual goal: DFT accuracy at force field cost



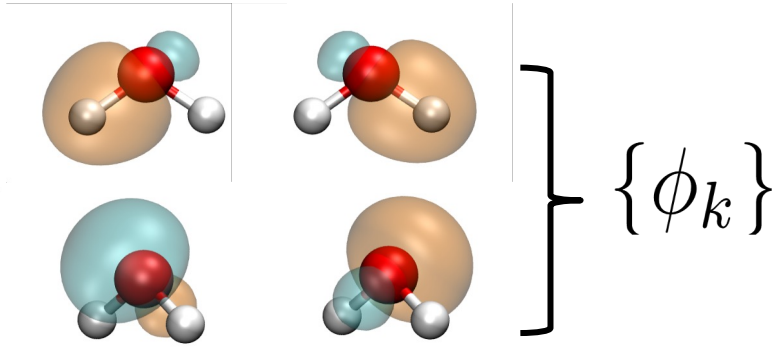
- Atom types: O, H
- Bond types: O-H
- Bond angles: θ
- Bond lengths: r_1, r_2
- Fragment motifs
- …

Less direct: Complicated mapping

Less transferable: Model not applicable to other atom/bond types not seen

Our strategy
*ML with **MO-based** features* from HF

A more modest goal: Wavefunction theory accuracy ***at HF cost using MO***



Direct: Easier mapping

Transferable: Across atom types & diverse chemistries

Physical driven problem formulation & feature design

Advantageous factorization

$$E_{total} = E_{HF} + E_c \text{ and } E_c = \sum_{ij}^{occ} \epsilon_{ij}$$

Nesbet's theorem: For any given post-HF method, the correlation energy can be written as a sum of pair energies

$$\epsilon_{ij} = \epsilon\left(\{\phi_k\}^{ij}\right) \approx \epsilon\left(\mathbf{f}_{ij}\right)$$

A single function of the MOs (independent of i,j)

List of MO-based features associated with orbitals i, j.

Learning labels: Pair energies ϵ_{ij}

Feature design

Desired properties

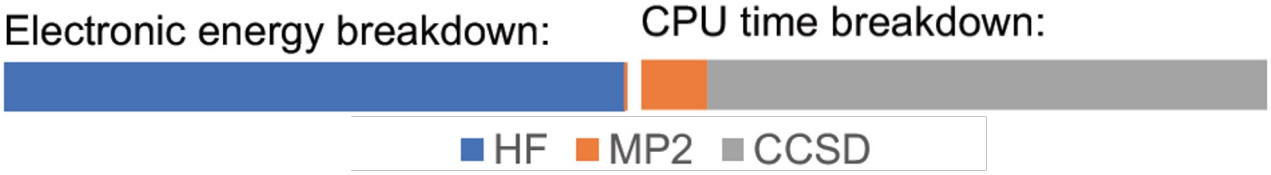
- First principles
- Uniqueness
- ***Consistent ordering***
- ***Size-consistency (improved feature)***

Invariances

- Translational
- Rotational
- Atom index permutation
- Orbital index permutation

$$E_{MP2} = -\frac{1}{4}\sum_{ij}^{occ}\sum_{ab}^{virt} \frac{|\langle ab||ij\rangle|^2}{\mathbf{F}_{aa} + \mathbf{F}_{bb} - \mathbf{F}_{ii} - \mathbf{F}_{jj}}$$

Learning features: Energy matrix elements $\mathbf{f}_{ij}$ in Fock matrix **F,** Coulomb matrix **J,** exchange matrix **K** for each MO computed in HF



MOB-ML for small molecules

Symmetrized feature set

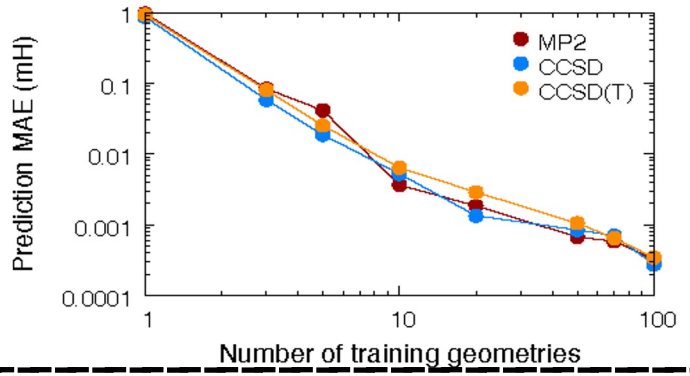
- Energy matrix elements from HF
- Symmetrize orbitals i & j for f_{ij}: $f_{ij}=f_{ji}$ to preserve $\epsilon_{ij}^{ML} = \epsilon_{ji}^{ML}$
- Order other MOs by centroid distances
- Fixed feature length & filled in zeros for non-existing elements

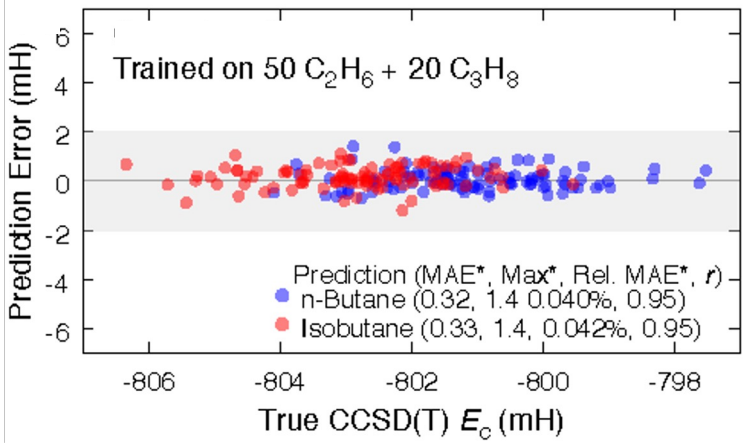
$$|\hat{i}\rangle = \frac{1}{\sqrt{2}}(|i\rangle + |j\rangle)$$

$$|\hat{j}\rangle = \frac{1}{\sqrt{2}}(|i\rangle - |j\rangle)$$

Transfer across different configurations

- Full Gaussian process regression
- Train & test on thermalized water structures
- Same accuracy for all different theories
- Chemical accuracy (~2 mH) with one structure





Transfer across a family of molecules

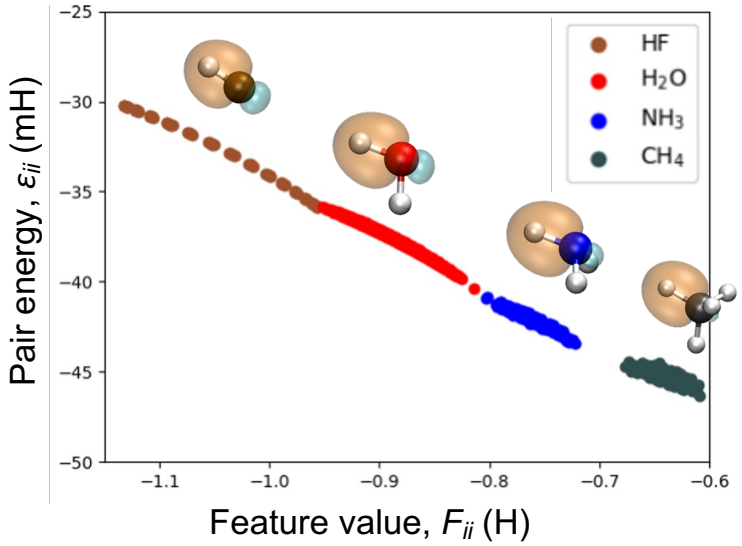
- Full Gaussian process regression
- Train on small alkanes & test larger alkanes
- All the relative errors are smaller than chemical accuracy (~2 mH)

Caltech

Understand the chemical space by MOB representation

Observation from quantum computation

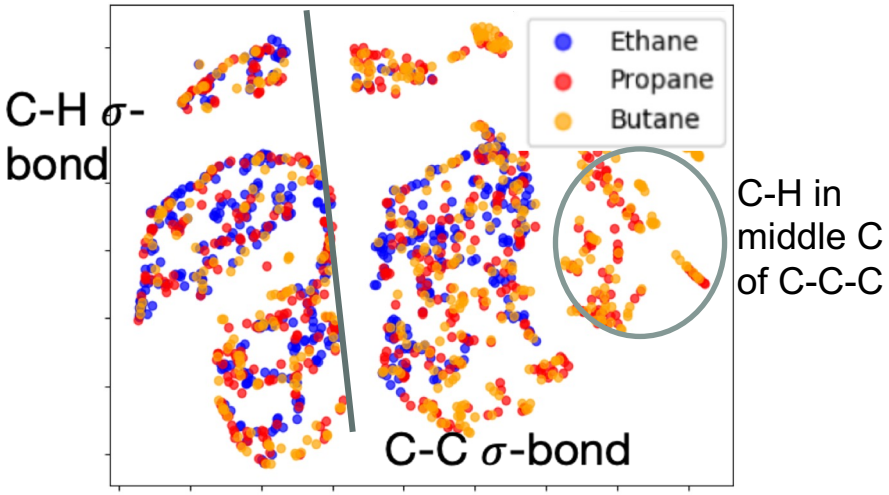
σ-bond orbital pair energy



- Local linearity of MOs
- Same type of MOs in different molecules have similar trend vs features

Chemical space by MOB features

T-SNE visualization of alkane



- Chemical space form clusters
- Conservative MOB features across molecules

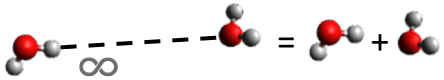
- Supervised/Unsupervised clustering the chemical space
- Clustering + local GP may further scale MOB-ML

Caltech

Improved MOB features and scalable GPR

Improved feature set (Husch et al, 2021)

- Keep all previous operations, except:
- Improved (Consistent) feature/orbital ordering
  - Centroid distance → Est. contribution to MP2/MP3 energy
- Ensure size consistency: correct scaling with distance
  - No interaction (ϵ_{ij}=0) in long distance
  - ϵ_{ij} insensitive to far-away MO k



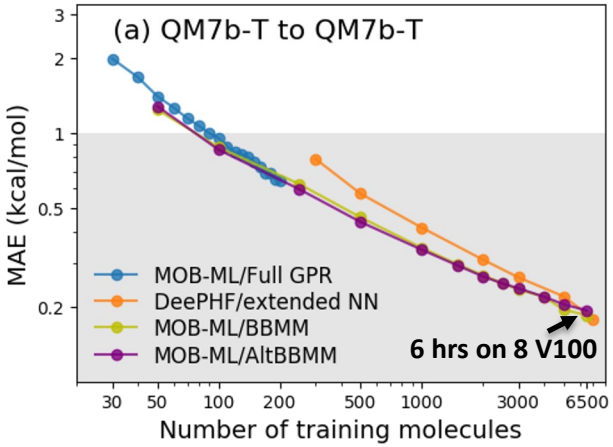
Size consistency: $E^{AB} = E^{A} + E^{B}$

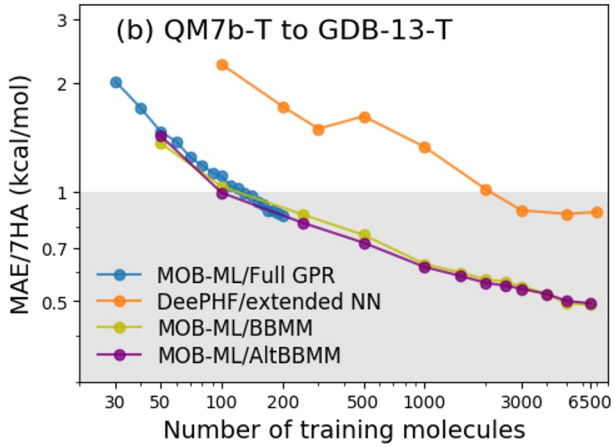
$$M_{ij} \sim \frac{1}{1+r_{ij}^{6}} M_{ij}$$

$$M_{ik} \sim \frac{1}{r_{ik}^{3}}$$

Extend to big data regime by scalable exact GP (Sun et al, 2021)

- Approximated GP not working well
- AltBBMM: An alternative implementation of scalable GP algorithm (BBMM)
  - O(N^2) time & O(N) memory
  - Training on over 1 million ϵ_{ij} with 6 hrs (AltBBMM)





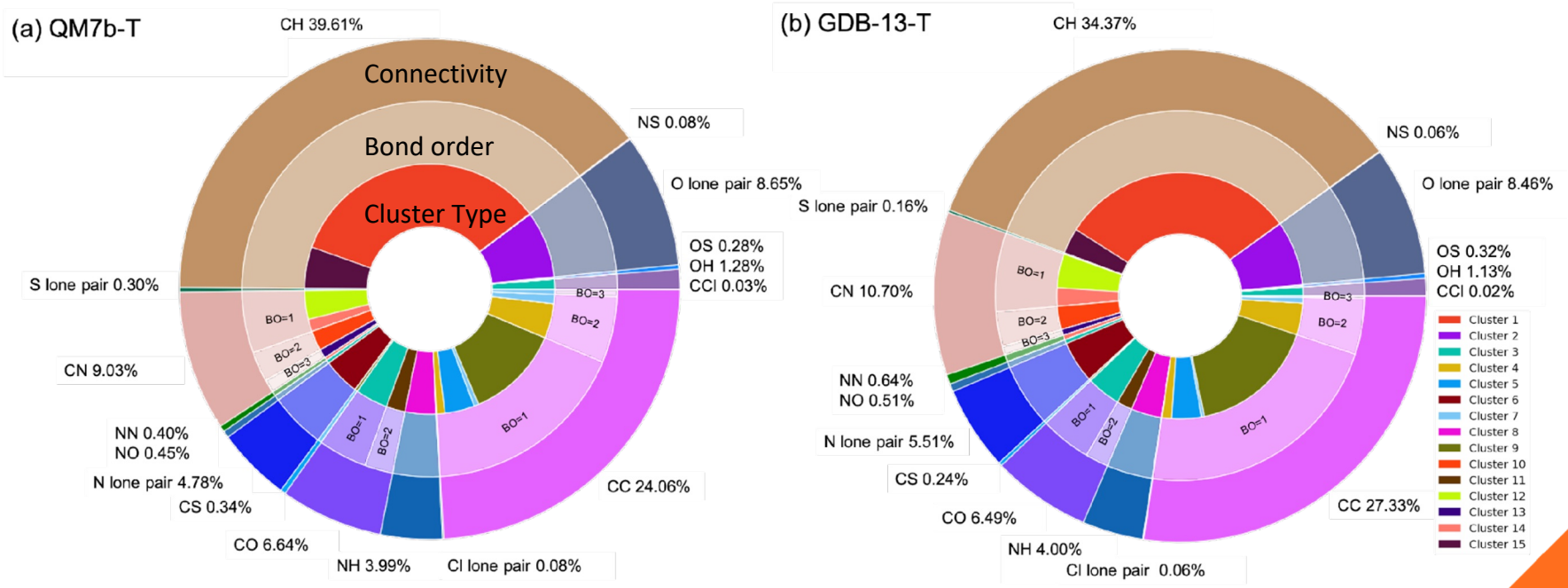
All collected with improved features

Caltech

Unsupervised clustering for organic chemical spaces

Gaussian Mixture model (GMM) clustering

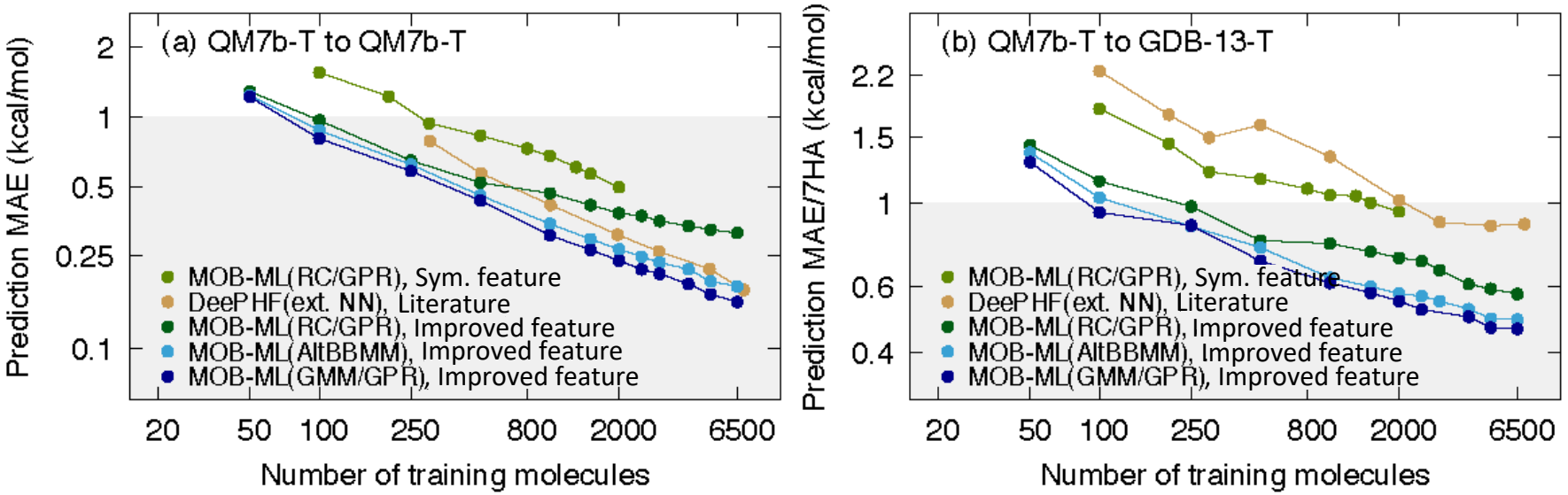
- **Improved MOB feature enables unsupervised clustering**
  - No additional classifier
  - Blackbox: Auto detections of number of clusters by Bayesian information criteria
  - Agree well w/ chemical intuition of MO types



Unsupervised clustering for organic chemical spaces

Gaussian Mixture model (GMM) clustering

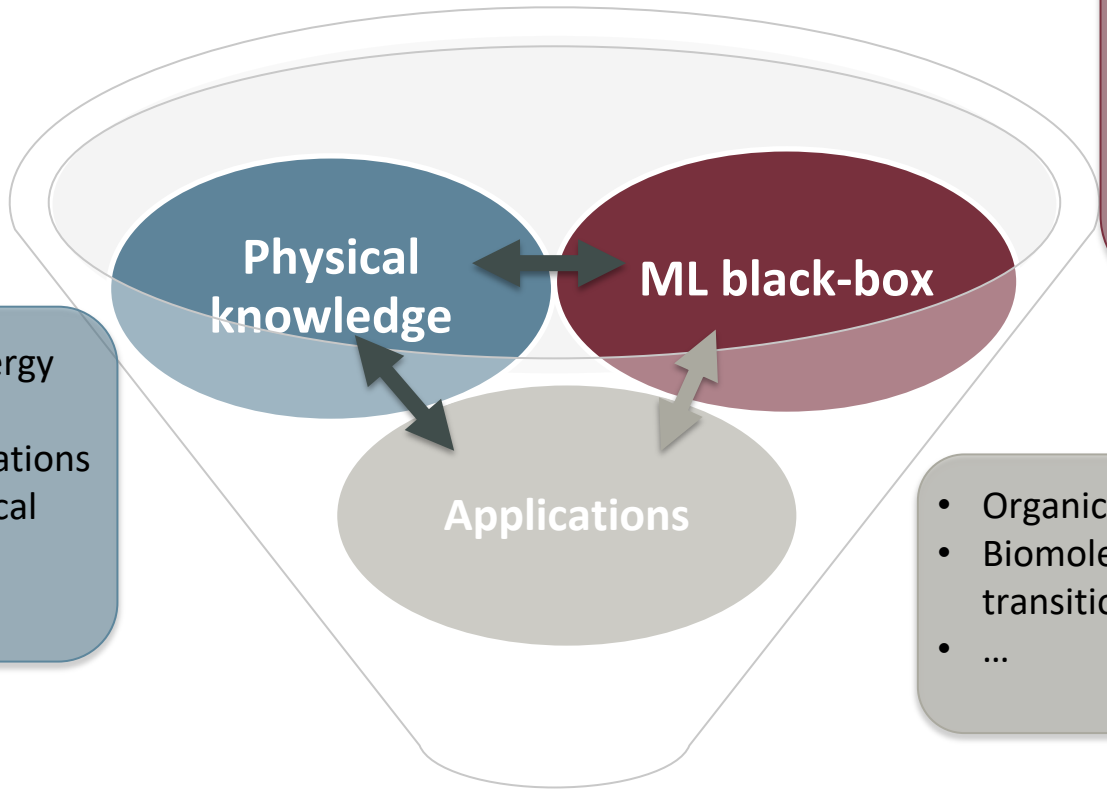
- **Cluster + local GP provides STOA accuracy**
  - Further bridge with regression: Local scalable GP for each cluster
  - Reducing & scale up MOB-ML training
  - No/little loss of transferability with clustering + local GP
  - Most accurate results from GMM/GPR



Conclusion

Research goal: Merging maximum physical knowledge into ML to create a general tool for chemical applications

- Regressions for molecular energy
- Supervised/Unsupervised clustering for chemical space exploration
- Works for both small & big data regimes



- Decompose energy as pair energies
- MOB representations preserve chemical info & satisfy physical limits

- Organic chemistry
- Biomolecule interaction & transition metal complex
- …

MOB-ML: Physical driven, black-box, interpretable, accurate and transferable molecular modelling method with HF costs

Acknowledgement

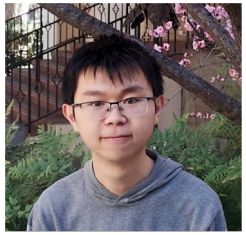
MOB-ML developers

Prof. Thomas F. Miller III

Dr. Matt Welborn

Dr. Tamara Husch

Jiace Sun



Dr. Sebastian Lee

Dr. J. Emiliano Deustua

My committee and Miller group members

Funding from

- U.S. Army Research Laboratory (W911NF-12-2-0023)
- U.S. Department of Energy (DE-SC0019390),
- Caltech DeLogi Fund
- Camille and Henry Dreyfus Foundation (Award ML-20-196).
- NERSC, a DOE Office of Science User Facility supported by the DOE Office of Science under contract DE-AC02-05CH11231.

Q & A