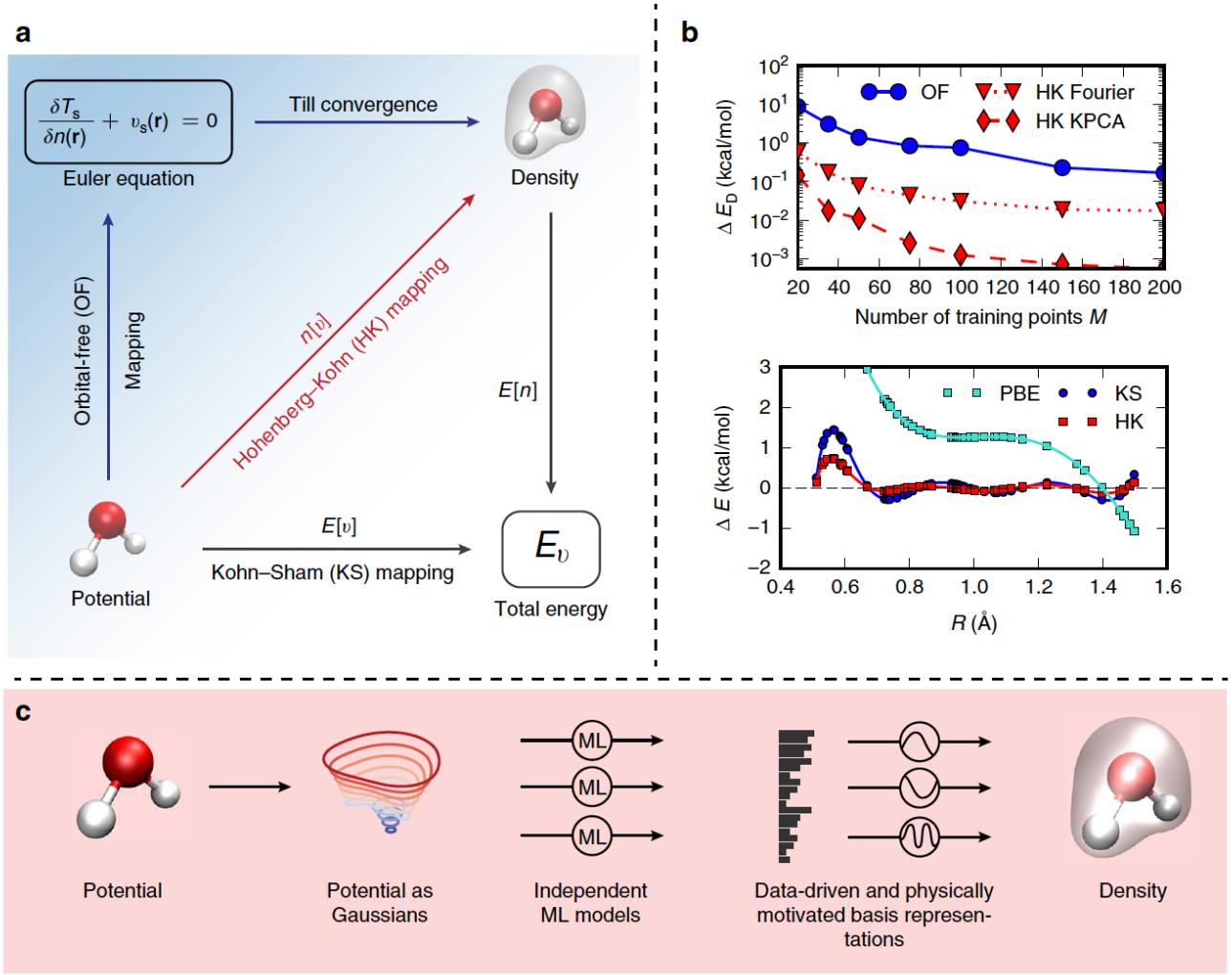
Machine Learning in Computational Chemistry

Shao Group Meeting

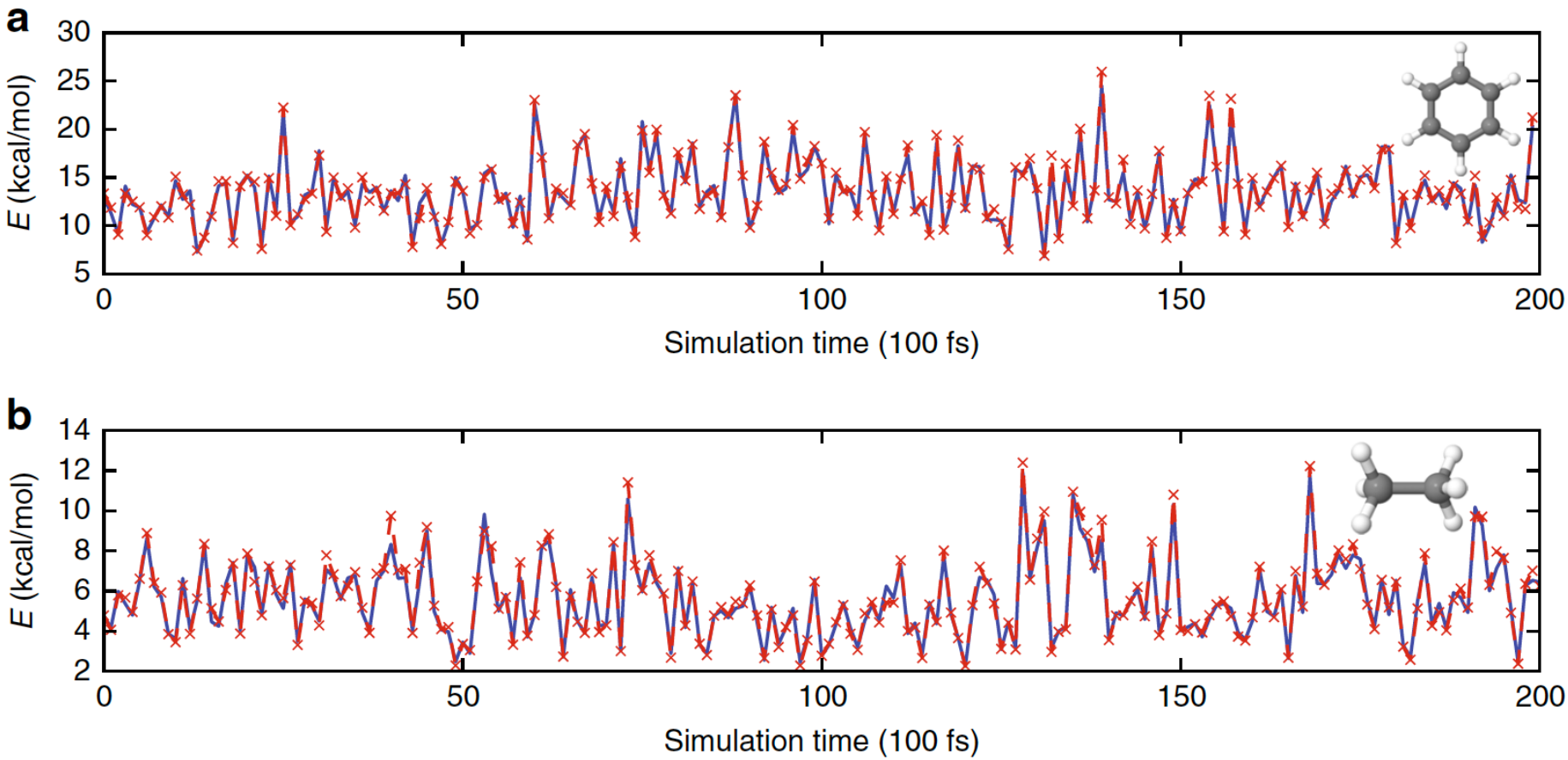
University of Oklahoma

November 14, 2017

1. ML Prediction of Total Energy and Components
   - Total HK DFT Energies, Mueller (2017)
   - Total DFT Energy, Roitberg (2017)
   - 1-Body, 2-Body, and 3-Body Energies, Parkhill (2016)
2. ML Prediction of Energy Differences
   - E(aiQM) - E(SQM), Weitao Yang (2016, 2017)
   - E(CCSD(T)) - E (MP2), Shaw (2017).
   - ΔE(CC2) - ΔE (TDDFT), von Lilenfeld (2015)
   - Deformation Energy and Vibrational Analysis, Thiel (2017)
3. ML Optimization of Energy Functions
   - AMOEBA Model, Roux (2017)

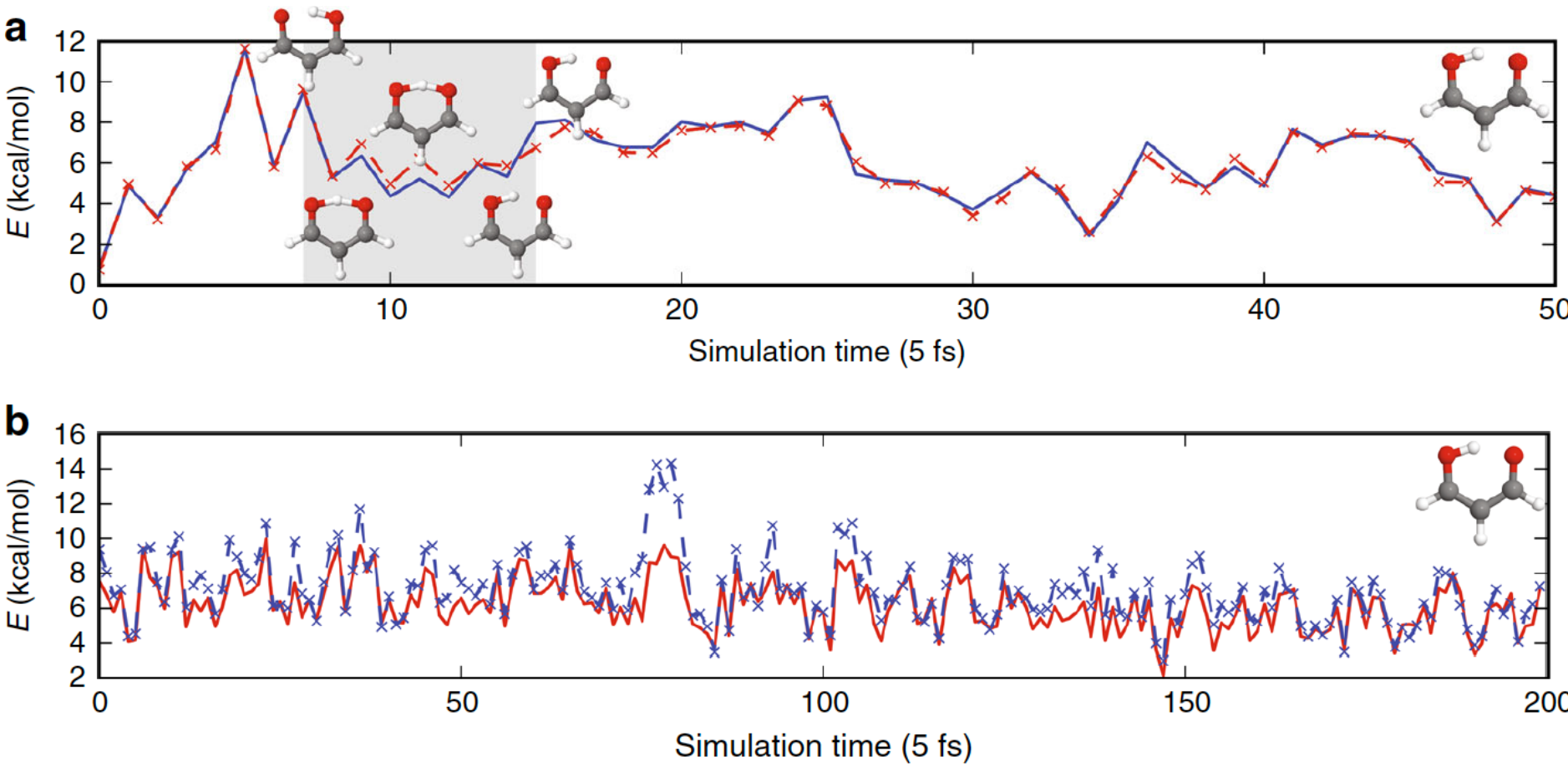


Brockherde, Vogt, Li, Tuckerman, Burke and Mueller, Nat. Commun. 8, 872 (2017).



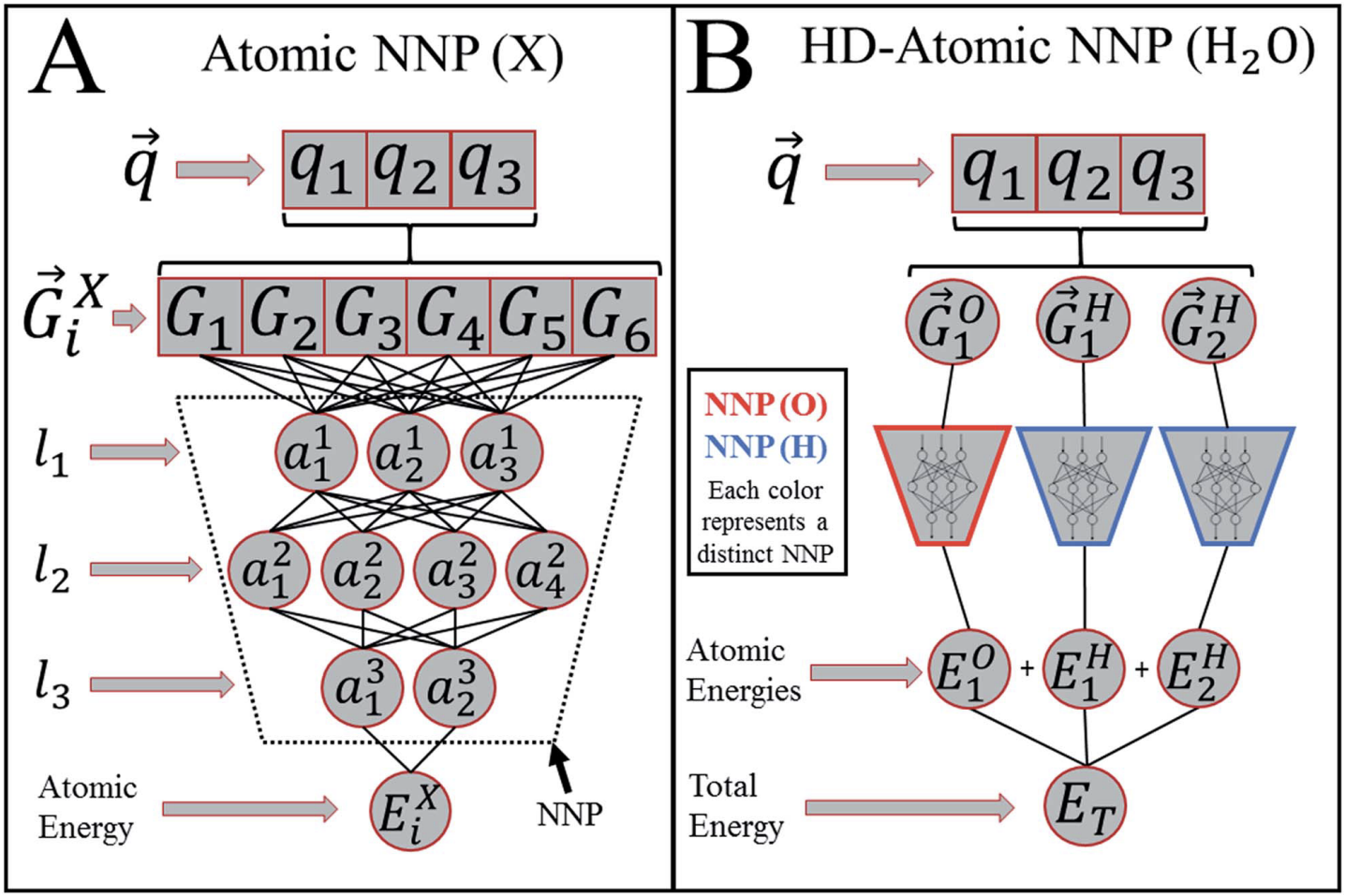
Blue: PBE; Red: ML-HK

Brockherde, Vogt, Li, Tuckerman, Burke and Mueller, Nat. Commun. 8, 872 (2017).



Blue: PBE; Red: ML-HK

Brockherde, Vogt, Li, Tuckerman, Burke and Mueller, Nat. Commun. 8, 872 (2017).



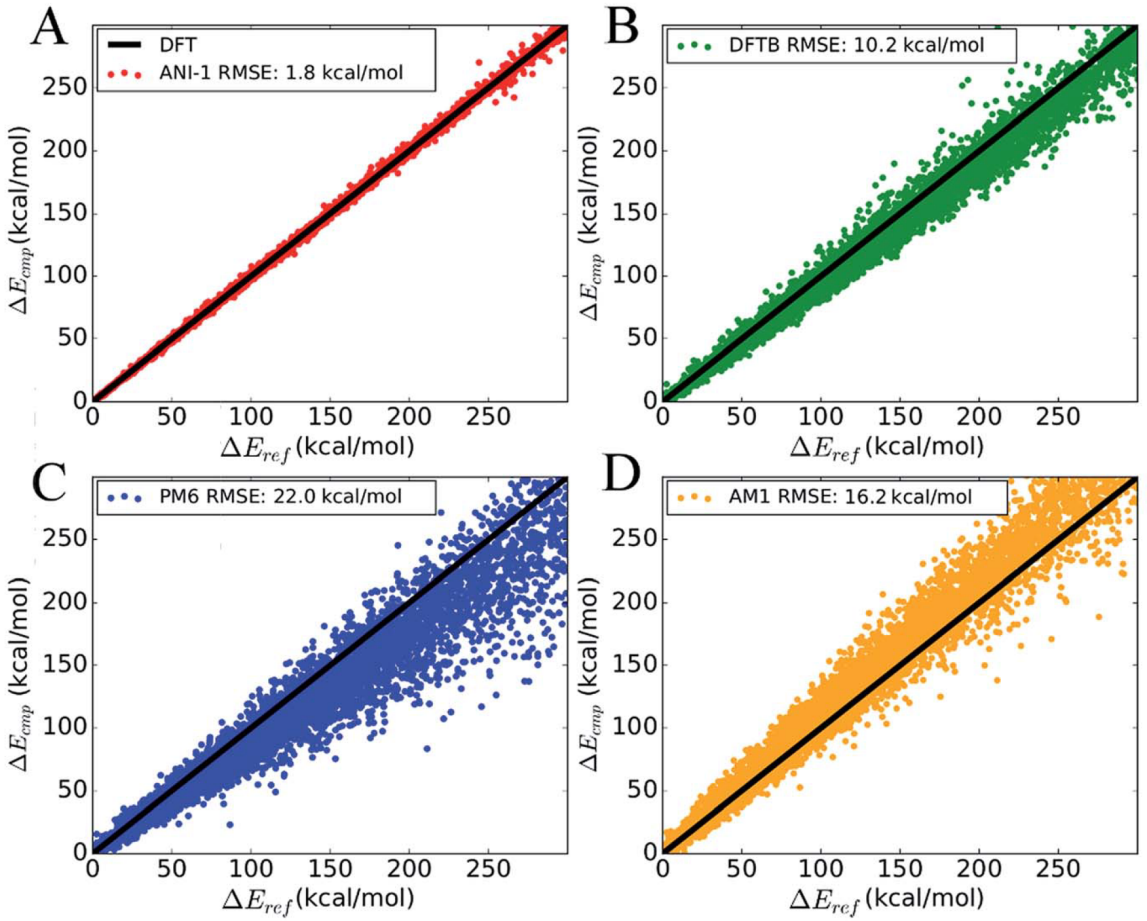
Smith, Isayev, and Roitberg, Chem. Sci., 8, 3192 (2017).

- Working equations

$$
\begin{aligned}
E_{total} &= \sum_i E_i \\
G_i^R &= \sum_{j \neq i}^{N} e^{-\eta(R_{ij}-R_s)^2} f_c(R_{ij}) \\
G_m^{\text{mod}} &= 2^{1-\xi} \sum_{j,k \neq i} \left[1 + cos(\theta_{ijk} - \theta_s)\right]^{\xi} e^{-\eta\left(\frac{R_{ij}+R_{jk}}{2} - R_s\right)^2} f_c(R_{ij}) f_c(R_{ik}) \\
f_c(R) &= \begin{cases} 0.5cos\left(\pi R/Rc\right) + 0.5, & R \leq R_c \\ 0 \end{cases}
\end{aligned}
$$

- Training set: 57951 molecules. Up to 8 “heavy atoms” (C, N, O) from GDB-11 database
- $>$ 3N-6 structures for each molecule. $\sim$ 17.2 million structures in total.
- Target data: ωB97X/6-31G* energies.
- Number of nodes: 768: 128 : 128 : 64 : 1

Smith, Isayev, and Roitberg, Chem. Sci., 8, 3192 (2017).



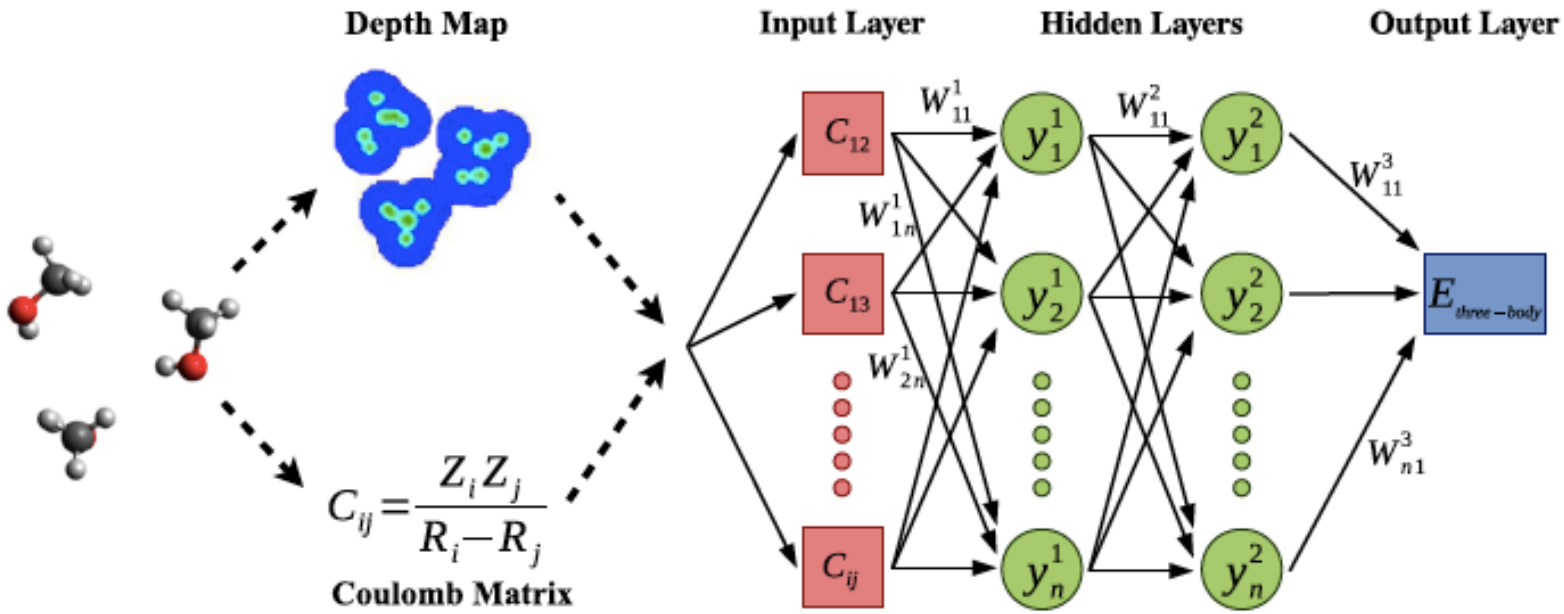
Smith, Isayev, and Roitberg, Chem. Sci., 8, 3192 (2017).

- Many-body expansion

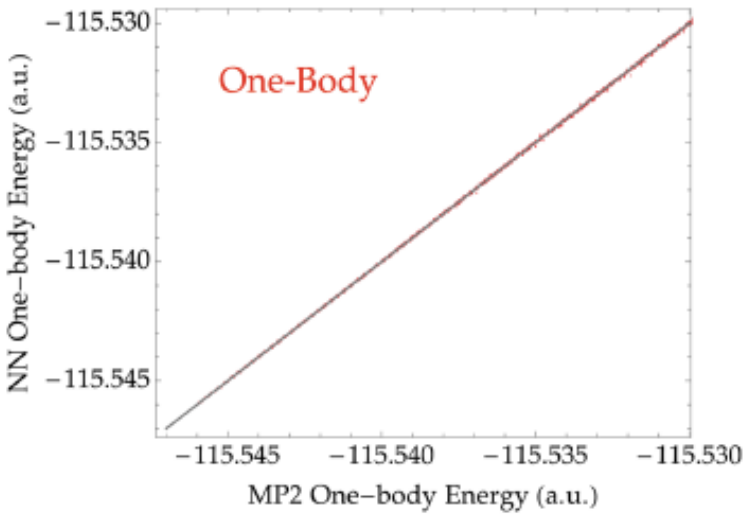
$$
\begin{aligned}
E_{total} &= \sum_i E_i + \sum_{i<j} \Delta E_{ij} + \sum_{i<j<k} \Delta E_{ijk} \\
\Delta E_{ij} &= E_{ij} - E_i - E_j \\
\Delta E_{ijk} &= E_{i,j,k} - \Delta E_{ij} - \Delta E_{ik} - \Delta_{jk} - E_i - E_j - E_k
\end{aligned}
$$

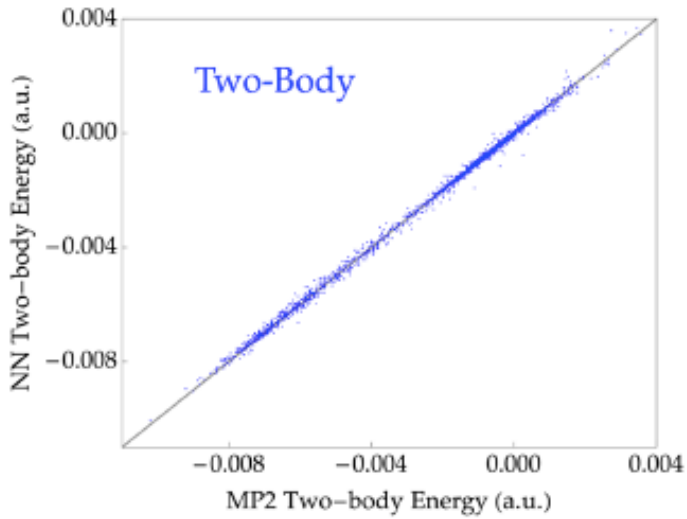
- Machine learning
  - 1-body: one hidden layer NN, 50 000 neurons, 844 800 samples
  - 2-body: two hidden layers, 10 000 + 5000 neurons, 74240 samples
  - 3-body: three hidden layers, 1000 + 2000 + 2000 Neurons, 36 864 samples
  - samples from MD trajectory of 108 methanol molecules
  - Permutational invariance is considered.

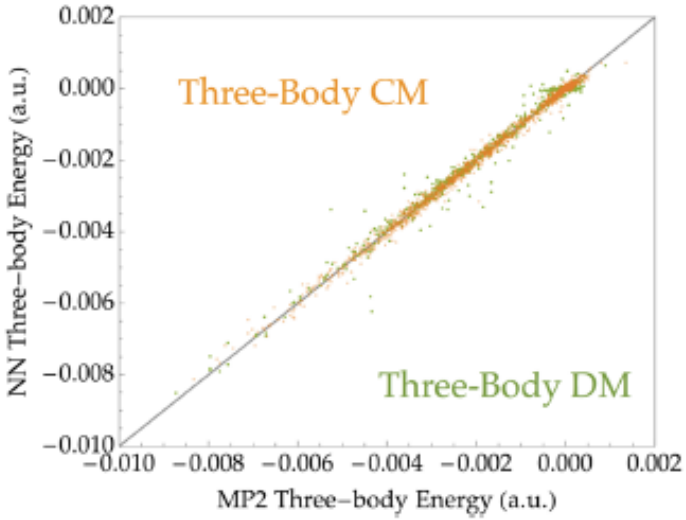
Yao, Herr, Parkhill, JCP, 146, 014106 (2016).

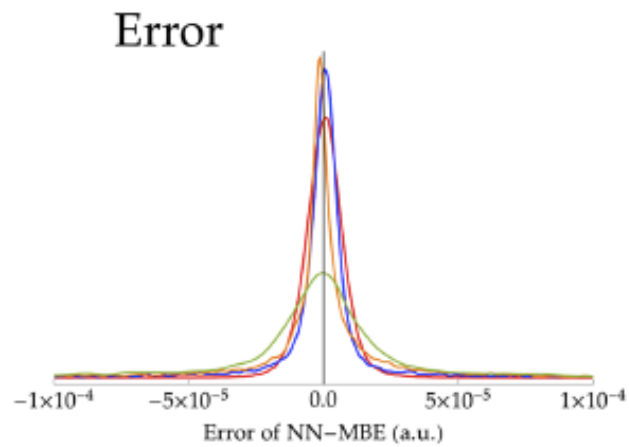


Yao, Herr, Parkhill, JCP, 146, 014106 (2016).





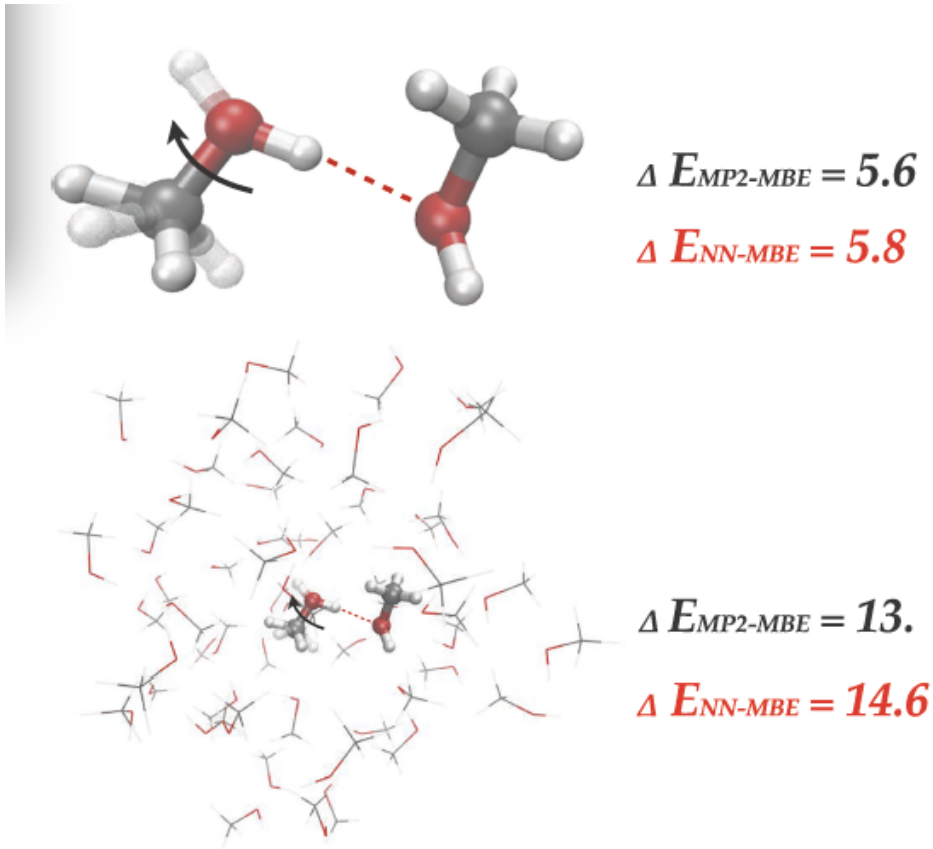




Yao, Herr, Parkhill, JCP, 146, 014106 (2016).

Table: Errors in kcal/mol

Error	1-body	2-body	3-body (CM)	3-body (DM)
MSE	0.0002	0.0006	-0.0007	-0.0025
MAE	0.0038	0.0098	0.0126	0.0244



1. ML Prediction of Total Energy and Components
   - Total HK DFT Energies, Mueller (2017)
   - Total DFT Energy, Roitberg (2017)
   - 1-Body, 2-Body, and 3-Body Energies, Parkhill (2016)

2. ML Prediction of Energy Differences
   - E(aiQM) - E(SQM), Weitao Yang (2016, 2017)
   - E(CCSD(T)) - E (MP2), Shaw (2017).
   - ΔE(CC2) - ΔE (TDDFT), von Lilenfeld (2015)
   - Deformation Energy and Vibrational Analysis, Thiel (2017)

3. ML Optimization of Energy Functions
   - AMOEBA Model, Roux (2017)

- Working equations are

$$E_i = \sum_{j=1}^{L} w_{ij} \tanh\left(w_{ij1}G_i^1 + w_{ij2}G_i^2 + w_{ij3}Q_i + b_{ij}\right) + b_i \quad (1)$$

$$G_i^1 = \sum_{j\neq i}^{N} e^{-\eta(R_{ij}-R_s)^2} f_c(R_{ij}) \quad (2)$$

$$G_i^2 = 2^{1-\xi} \sum_{j,k\neq i}^{N} (1 \pm cos\theta_{ijk})^{\xi} e^{-\eta(R_{ij}^2+R_{jk}^2+R_{ik}^2} f_c(R_{ij}) f_c(R_{jk}) f_c(R_{ik}) \quad (3)$$

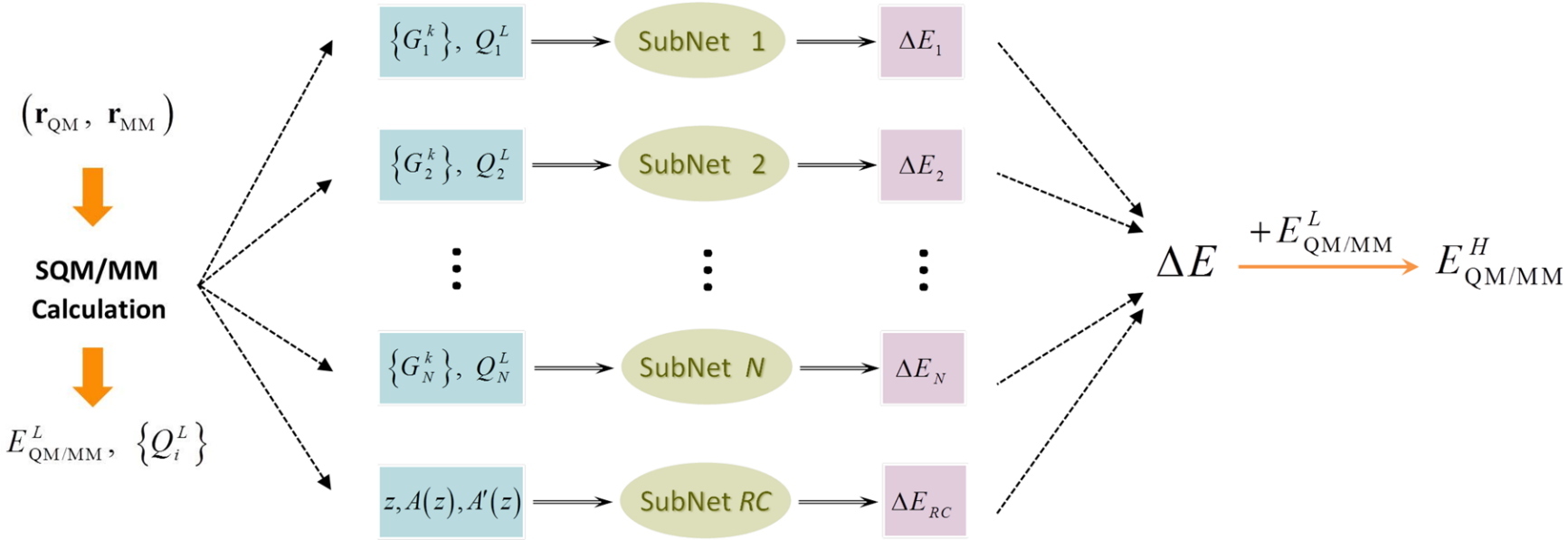
$$f_c(R) = \begin{cases} 0.5cos\left(\pi R/Rc\right) + 0.5, & R \leq R_c \\ 0 \end{cases} \quad (4)$$

$$\Delta E_{RC} = \sum_{j=1}^{L} w'_j \tanh\left(w'_{j1}z + w'_{j2}A(z) + w'_{j3}\frac{\partial A(z)}{\partial z} + b'_j\right) + b' \quad (5)$$

$$\Delta E = \sum_{i}^{N} \Delta E_i + \Delta E_{RC} \quad (6)$$

- Parameters: R_c = 6Å; R_s = 0; η = 0.09 - 1.80; ξ = 0.09-1.80.
- Size of training set: 20 (S_N2); 20 (glycine in water); 30 (Claisen rearrangement)

- Work flow



- The final free energy profile is computed with

$$\Delta A^H_{z_1 \to z_2} = \Delta A^L_{z_1 \to z_2} + \Delta A^{L \to H}(z_2) - \Delta A^{L \to H}(z_1) \quad (7)$$

$$\Delta A^{L \to H}(z) = -\beta^{-1} \ln \left\langle e^{-\beta(E^H - E^L)} \right\rangle_z \quad (8)$$

- Shen, Wu, Yang, JCTC, 12, 4934 (2016).

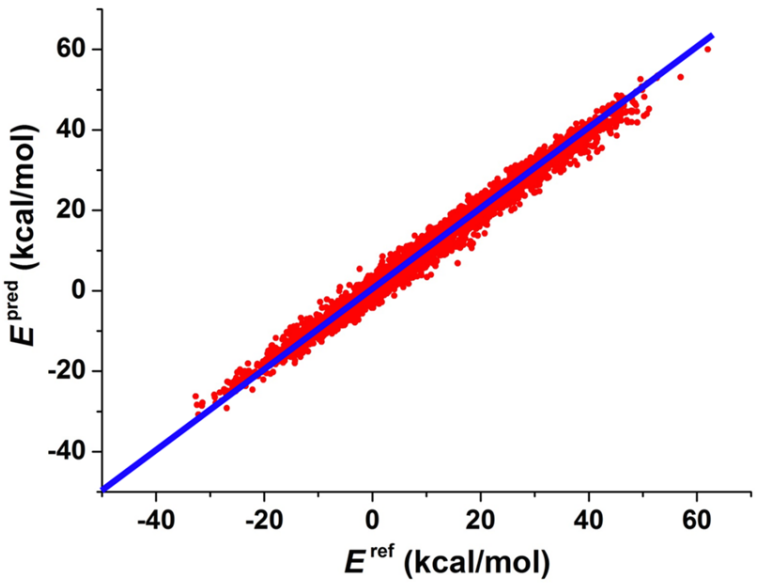
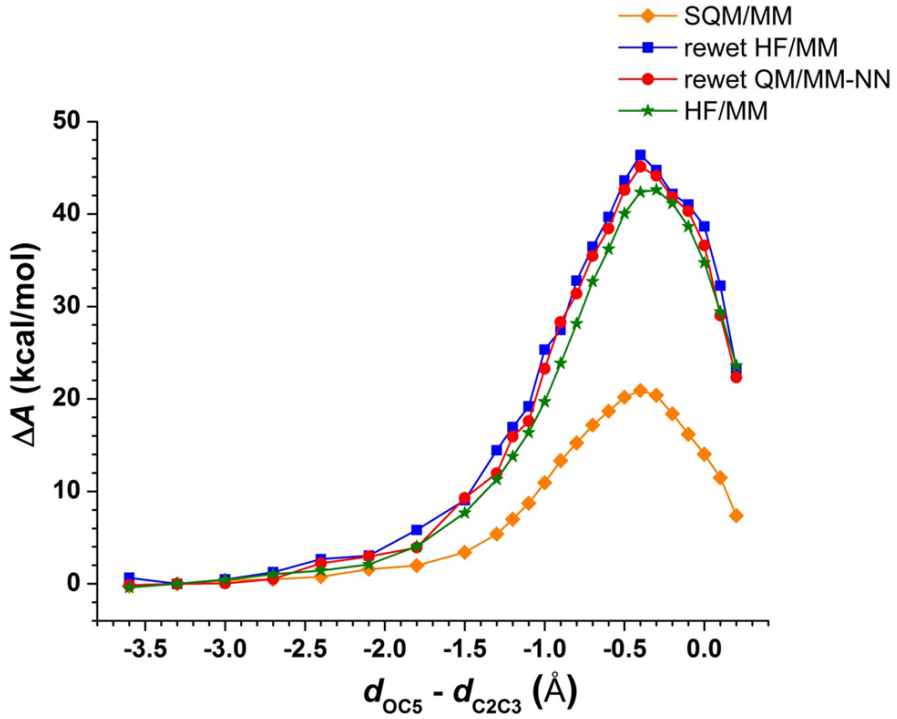


Table 3. Root Mean Squared Errors (kcal/mol) of Training and Testing Sets for Claisen Rearrangement Reaction of AVE with Q^2 Values (in Parentheses)[b]

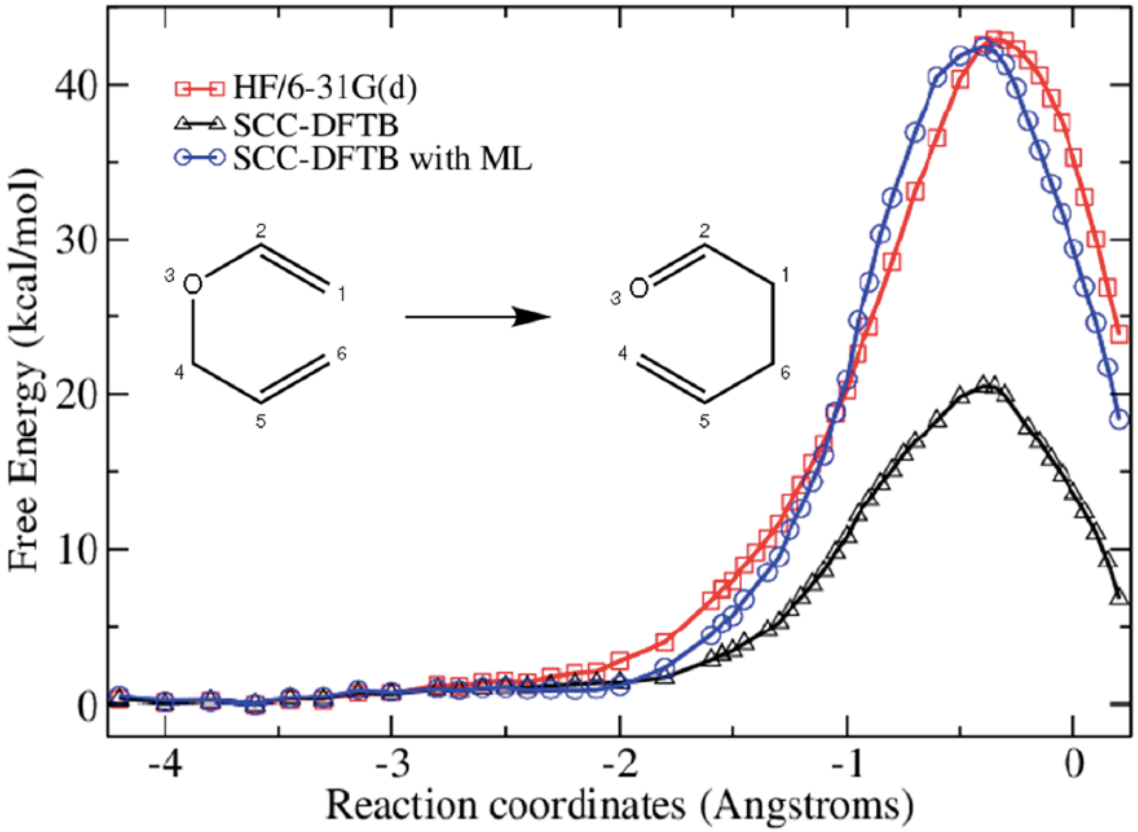
	training set	testing set
20[a]	2.08	2.36 (0.950)
30	2.05	2.21 (0.955)
40	2.00	2.22 (0.955)
50	1.95	2.21 (0.955)



- 1ns of SQM/MM simulation per window
- 2000 frames are collected per window
- QM/MM-NN 66 times faster than “rewet HF/MM”

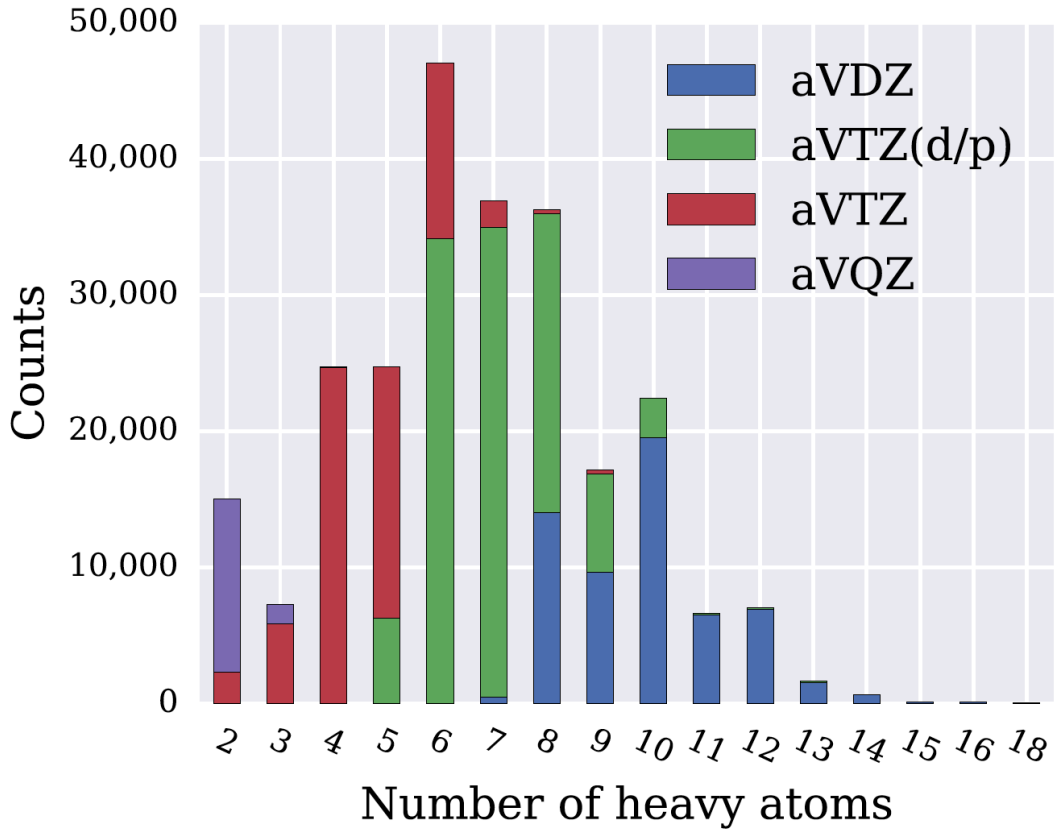
Shen, Wu, Yang, JCTC, 12, 4934 (2016).

- Step 1: Perform 1ns SCC-DFTB/MM simulation for each US window. Collect 200 frames.
- Step 2: Compute the HF/MM force for 200 frames
- Step 3: use ML to "learn" corrections to the force along the reaction coordinate
- Step 4: Rerun the simulation with corrected SCC-DFTB/MM force.
- Step 5. Use WHAM to get the free energy profile.



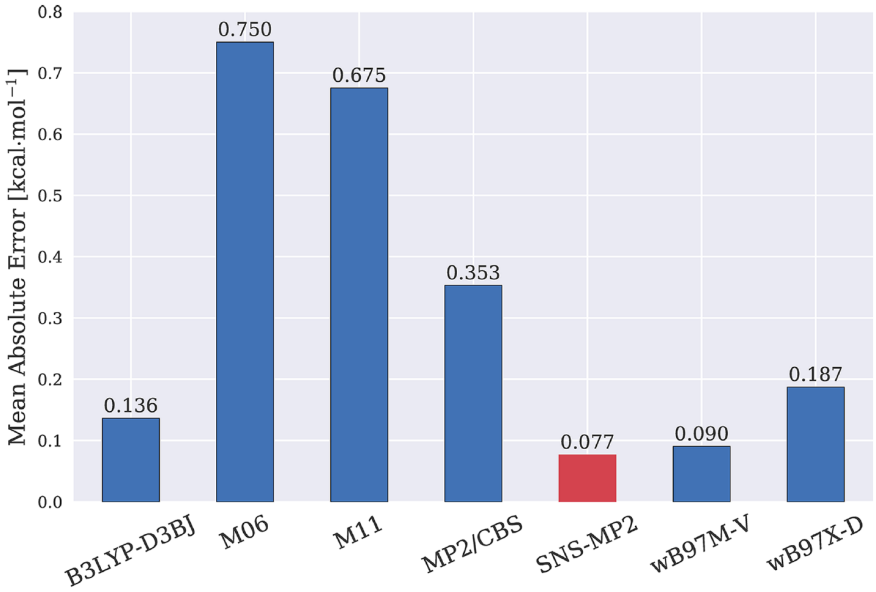
Wu, Shen, Yang, JCP, 147, 161732 (2017).

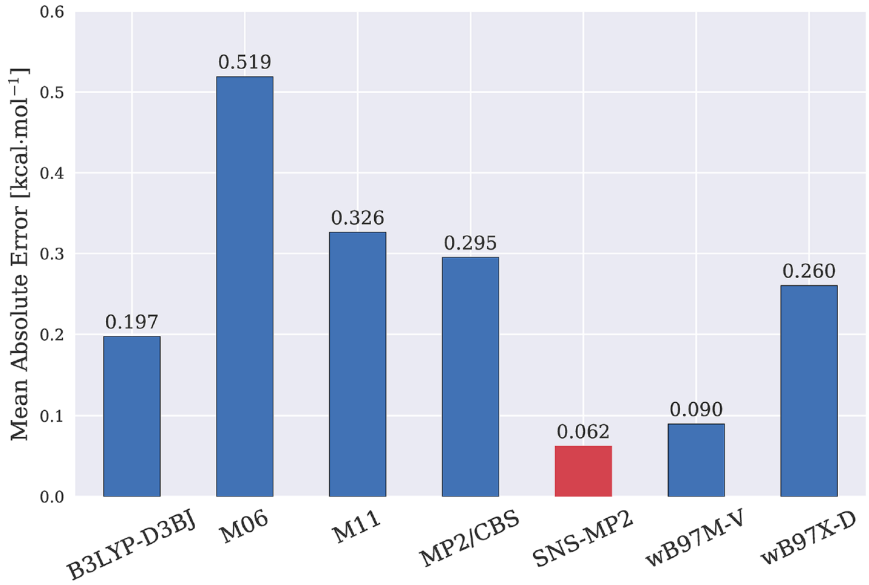
- 3149 different chemical systems; 3—2486 configurations per system
- Target data: CCSD(T)/CBS-quality interaction energies using composite calculations
- Input data: energies, density matrix overlap



McGibbon, · · · , Klepeis, and Shaw, JCP, 147, 161725 (2017).

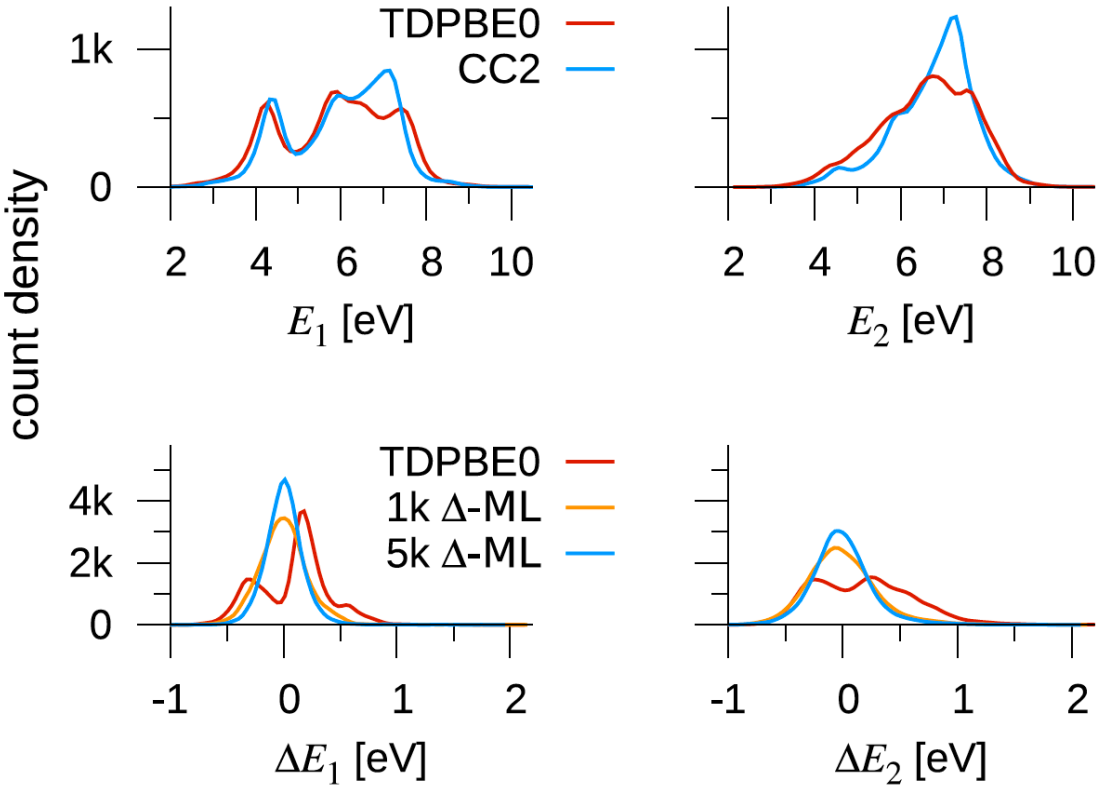
- Results for S66x8 (left) and T80 (right)





McGibbon, · · · , Klepeis, and Shaw, JCP, 147, 161725 (2017).

- The prediction is $\Delta E_i^{CC2}(\mathbf{d}_q) \sim \Delta E_i^{TDDFT}(\mathbf{d}_q) + \sum_i^{N_{\text{training}}} C_{it} e^{-|\mathbf{d}_q - \mathbf{d}_t|/\sigma}$.
- 20,000 organic molecules
- Input data: coulomb matrix (with extra zero elements for small molecules); bag-of-bonds



Ramakrishnan, Hartmann, Tapavicza and von Lilenfeld, JCP, 143, 084111 (2017).
See also Aspuru-Guzik et al, Chem. Sci. 7, 5139 (2016). Error < 0.01eV.

- CH_3Cl molecule, 44819 geometries
- descriptor: 10 atom-atom distances
- Errors in 166 vibrational energy levels up to 5000 cm^{-1} and 3606 vibrational energy levels up to 10000 cm^{-1}

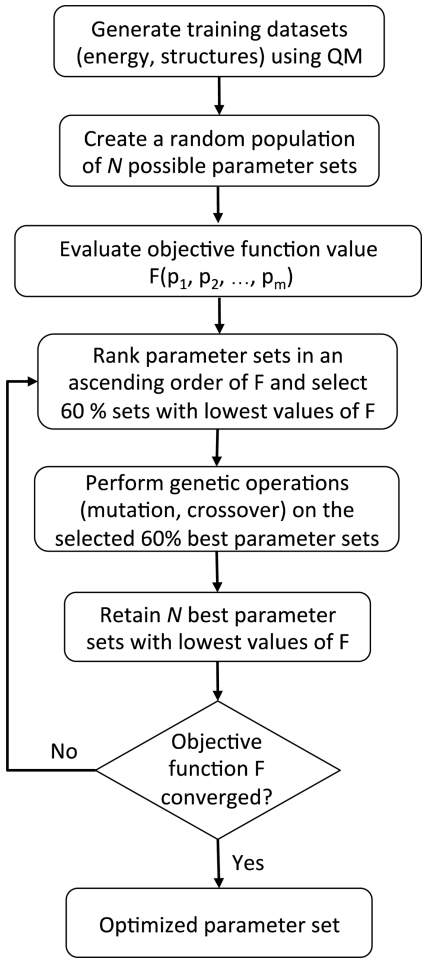
	50%-ML	r50%-ML	25%-ML	10%-ML	s10%-ML
ML model PESs					
RMSE (<5 000 cm^{-1})	0.02	0.10	0.09	1.61	0.14
MAD (<5 000 cm^{-1})	0.01	0.08	0.07	1.29	0.10
RMSE (<10 000 cm^{-1})	0.04	0.18	0.16	1.75	0.28
MAD (<10 000 cm^{-1})	0.03	0.15	0.12	1.44	0.21
Training set PESs					
RMSE (<5 000 cm^{-1})	0.06	0.08	0.12	0.30	0.12
MAD (<5 000 cm^{-1})	0.05	0.06	0.10	0.25	0.10
RMSE (<10 000 cm^{-1})	0.12	0.14	0.19	0.74	0.32
MAD (<10 000 cm^{-1})	0.11	0.09	0.14	0.61	0.24

Dral, Yurchenko and Thiel, 146, 244108 (2017).

1. ML Prediction of Total Energy and Components
   - Total HK DFT Energies, Mueller (2017)
   - Total DFT Energy, Roitberg (2017)
   - 1-Body, 2-Body, and 3-Body Energies, Parkhill (2016)

2. ML Prediction of Energy Differences
   - E(aiQM) - E(SQM), Weitao Yang (2016, 2017)
   - E(CCSD(T)) - E (MP2), Shaw (2017).
   - ΔE(CC2) - ΔE (TDDFT), von Lilenfeld (2015)
   - Deformation Energy and Vibrational Analysis, Thiel (2017)

3. ML Optimization of Energy Functions
   - AMOEBA Model, Roux (2017)



		atom types		
	parameters	O	H_O	C
E_{ESP}	monopole (q)	−1.0–0	0–0.4	0–0.3
	dipole (μ_x, μ_y, μ_z)	(0–0.5, 0.0, 0–0.4)	(−0.14–0, 0.0, −0.4–0)	(−0.4–0, 0.0, 0–0.9)
	quadrupole			
	$\begin{bmatrix} Q_{xx} & * & * \\ Q_{yx} & Q_{yy} & * \\ Q_{zx} & Q_{zy} & * \end{bmatrix}$	$\begin{bmatrix} 0–0.6 & * & * \\ 0.0 & -0.8–0 & * \\ 0–0.6 & 0.0 & * \end{bmatrix}$	$\begin{bmatrix} 0–0.28 & * & * \\ 0.0 & 0–0.06 & * \\ -0.2–0 & 0.0 & * \end{bmatrix}$	$\begin{bmatrix} 0–0.06 & * & * \\ 0.0 & -0.11–0 & * \\ -0.6–0 & 0.0 & * \end{bmatrix}$
	polarizability (α)	0.5–1.0	0.2–0.7	1.0–1.6
	Thole's factor (a)	0.3–0.5	0.3–0.5	0.3–0.5

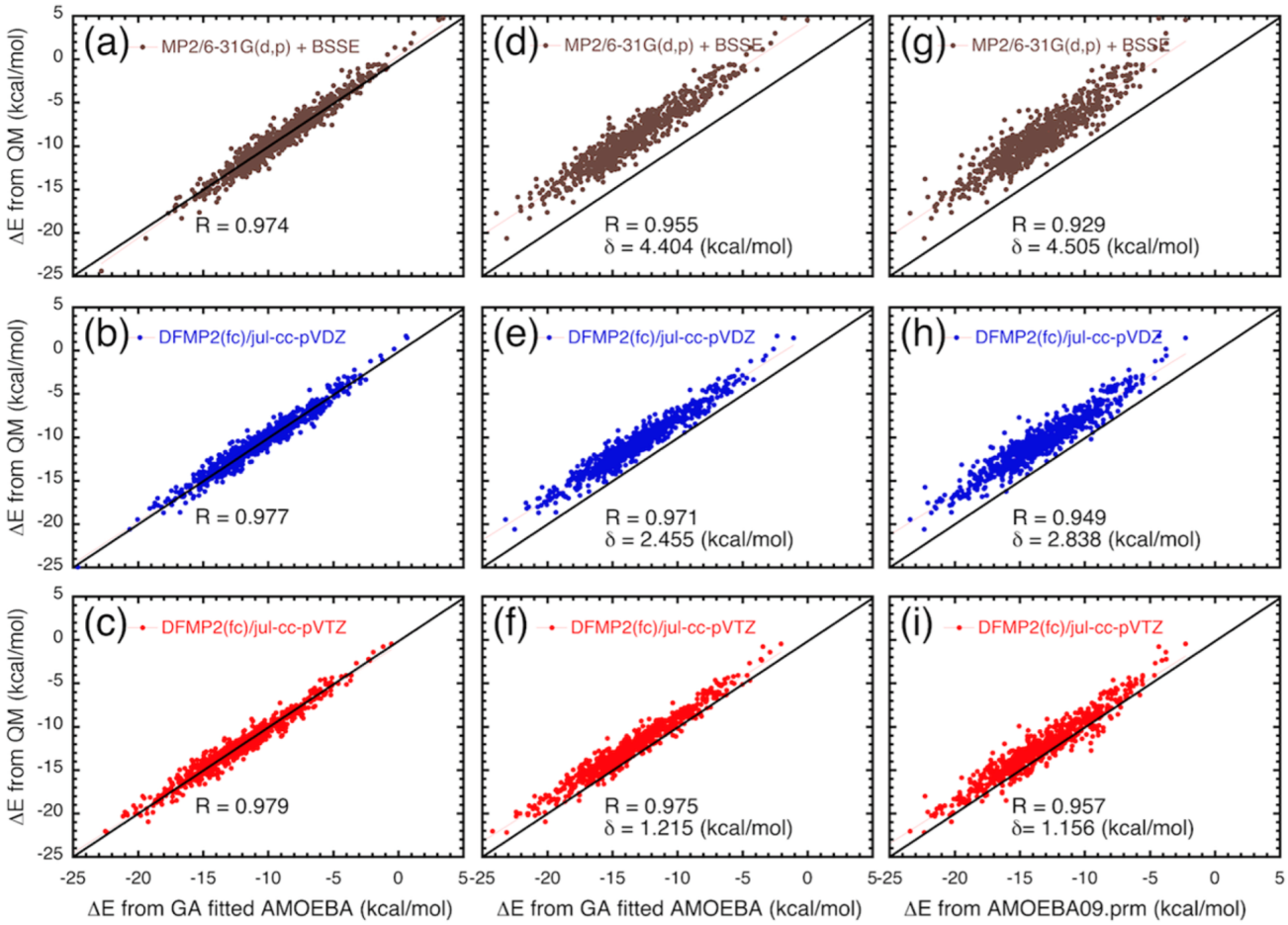
		atom types	
	parameters	O	H_O
E_{vdW}	minimum energy depth (ε_{min})	3.5–4.0	2.5–3.0
	minimum energy distance (R_{min})	0.12–0.15	0.001–0.0015
	H reduction factor (λ)	N/A	0.9–0.95
	energy offset (δ) 0.05–8.0		

Table 2. Number Distribution of the Extracted Clusters (9 Molecules, 11 Molecules, and 13 Molecules) from Different Liquid MD Systems

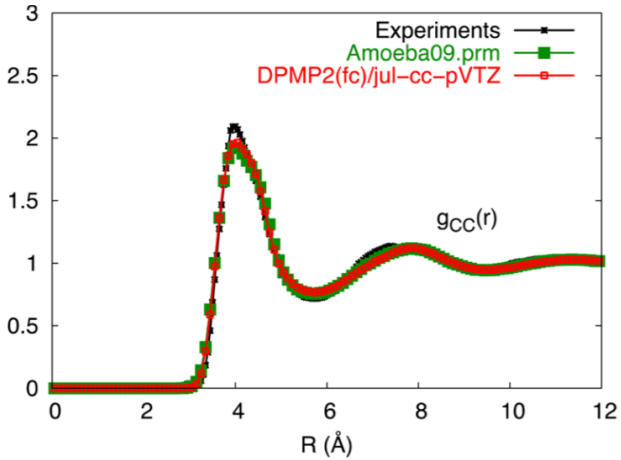
MD simulations	9 methanol	11 methanol	13 methanol
ρ = 0.933 g/mL, T = 193.15 K, p = 1.000 atm	372	157	94
ρ = 0.794 g/mL, T = 293.15 K, p = 1.200 atm	502	N/A	N/A
ρ = 0.627 g/mL, T = 393.15 K, p = 6.293 atm	125	N/A	N/A

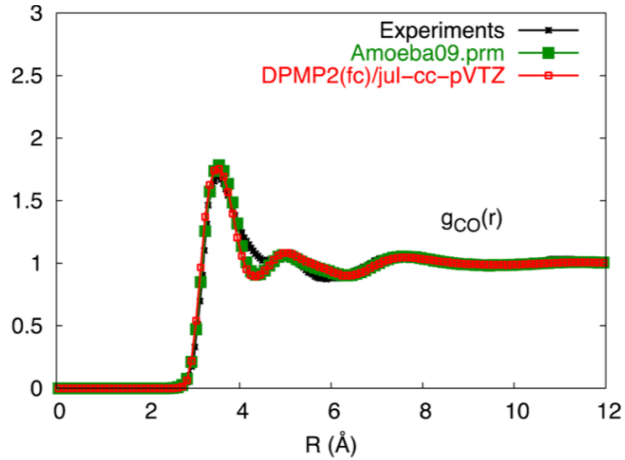
Li, Li, Pickard, · · · , Roux, Brooks, Roux, JCTC, 13, 4492 (2017).

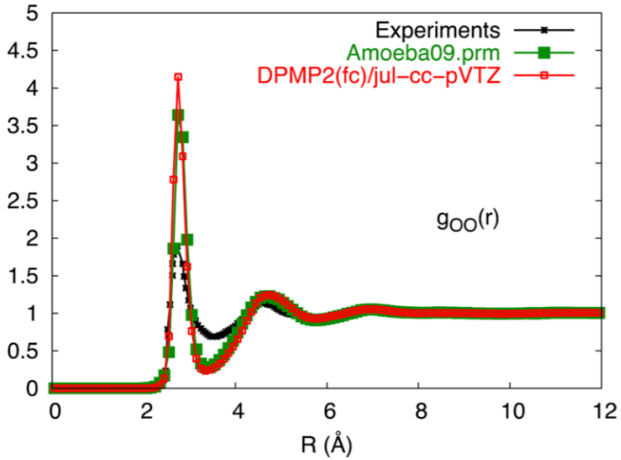
- Interaction energy of 1250 methanol clusters

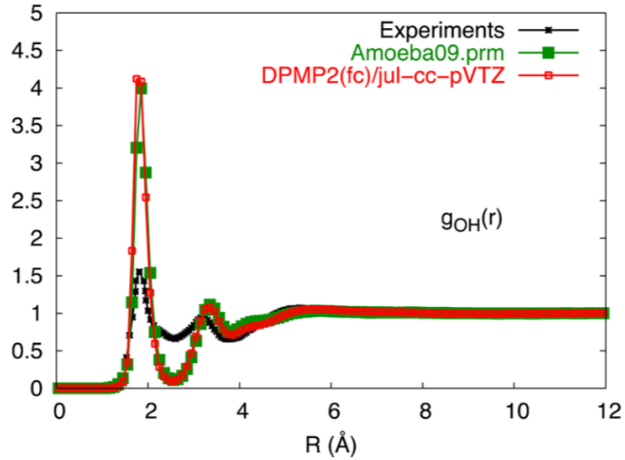


Li, Li, Pickard, · · · , Roux, Brooks, Roux, JCTC, 13, 4492 (2017).









		MD						
		MP2/6-31G(d,p)		DFMP2(fc)/jul-cc-pVDZ		DFMP2(fc)/jul-cc-pVTZ		
property	exp.	$\delta = 0$	$\delta \neq 0$	$\delta = 0$	$\delta \neq 0$	$\delta = 0$	$\delta \neq 0$	AMOEBA
ρ (g/mL)	0.786	0.401	0.759	0.568	0.763	0.686	0.781	0.774
ΔH_{vap} (kcal/mol)	8.95	6.54	9.44	7.23	9.43	8.58	8.98	9.17

Li, Li, Pickard, · · · , Roux, Brooks, Roux, JCTC, 13, 4492 (2017).