High Performance Computing Lecture

"Adaptive Machine Learning Framework to Accelerate *Ab Initio* Molecular Dynamics"

2016/10/14

Kota Goto

Background

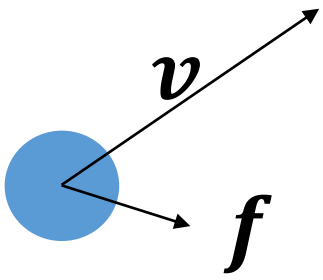
Molecular Dynamics (MD)

- MD simulation calculate whole atoms movement, and get atoms trajectory.

- MD simulation is popular and powerful tool for analyzing proteins, DNA, and other materials.

- Long time simulation is required, but its computational cost is high. (e.g. 100ns/day)

$$\boldsymbol{f} = m\boldsymbol{a}$$



Background

Quantum Mechanics (QM)

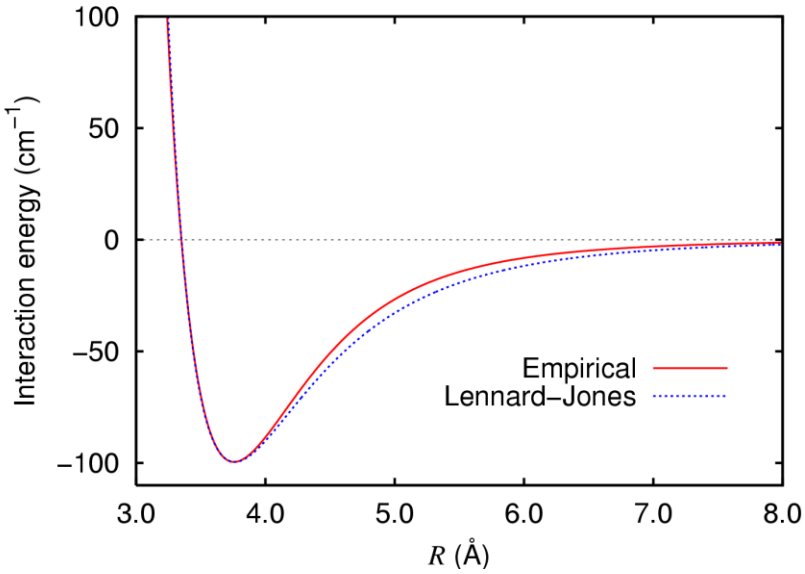
- QM consider effects of electron.
- QM simulation can calculate chemical reactions.
- High accuracy than MD simulation,
  and very high Computational cost.
- Base theory : density functional theory(DFT).

Background

Force field

- Method of approximation to explain forces.

- Experimental model, … etc.
  (e.g. Lennard-Jones potential)



https://upload.wikimedia.org/wikipedia/commons/9/93/Argon_dimer_potential_and_Lennard-Jones.png

Background

Ab Initio MD

- Potential energy and forces are obtained using QM.

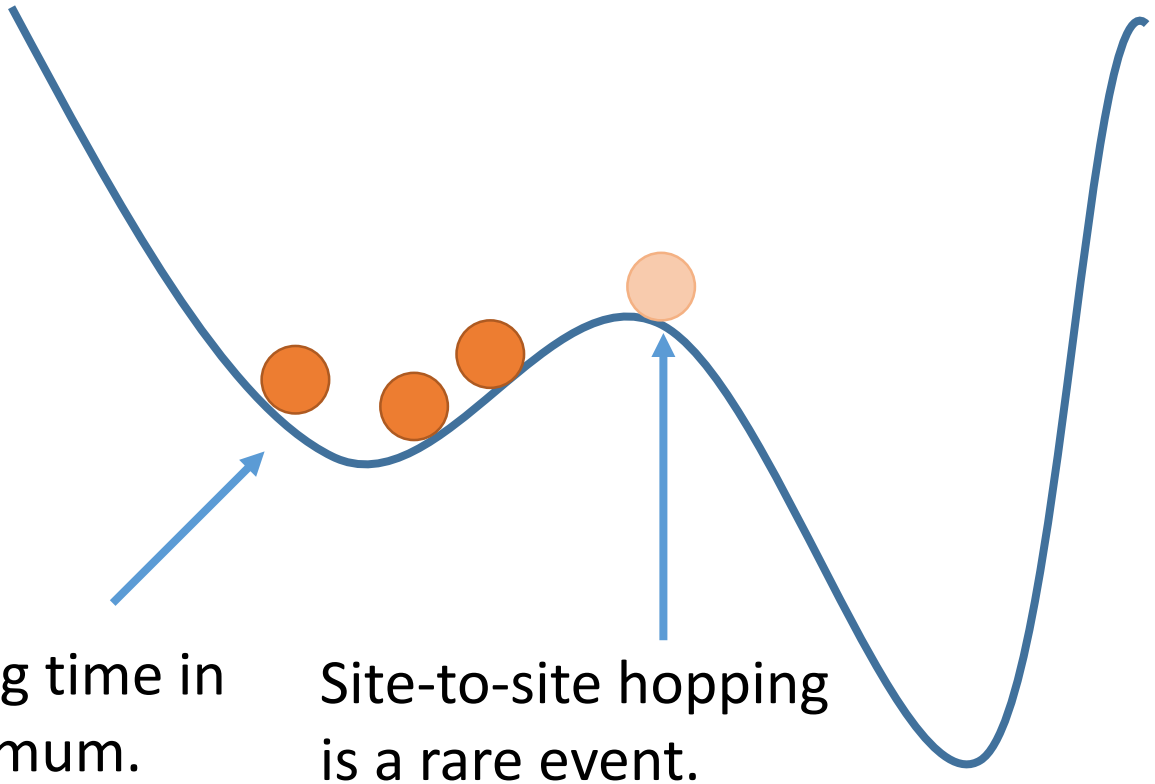
- QM calculation is very expensive and bottleneck of *ab initio* MD.

Selected papers

V. Botu, R. ramprasad,
"**Adaptive Machine Learning Framework to Accelerate *Ab Initio* Molecular Dynamics** ",
Int. J. Quantum Chem., 2015, 115,1075-1083.

- Ab initio Molecular Dynamics with Machine Learning
- Machine Learning model predicts energy and force fast.

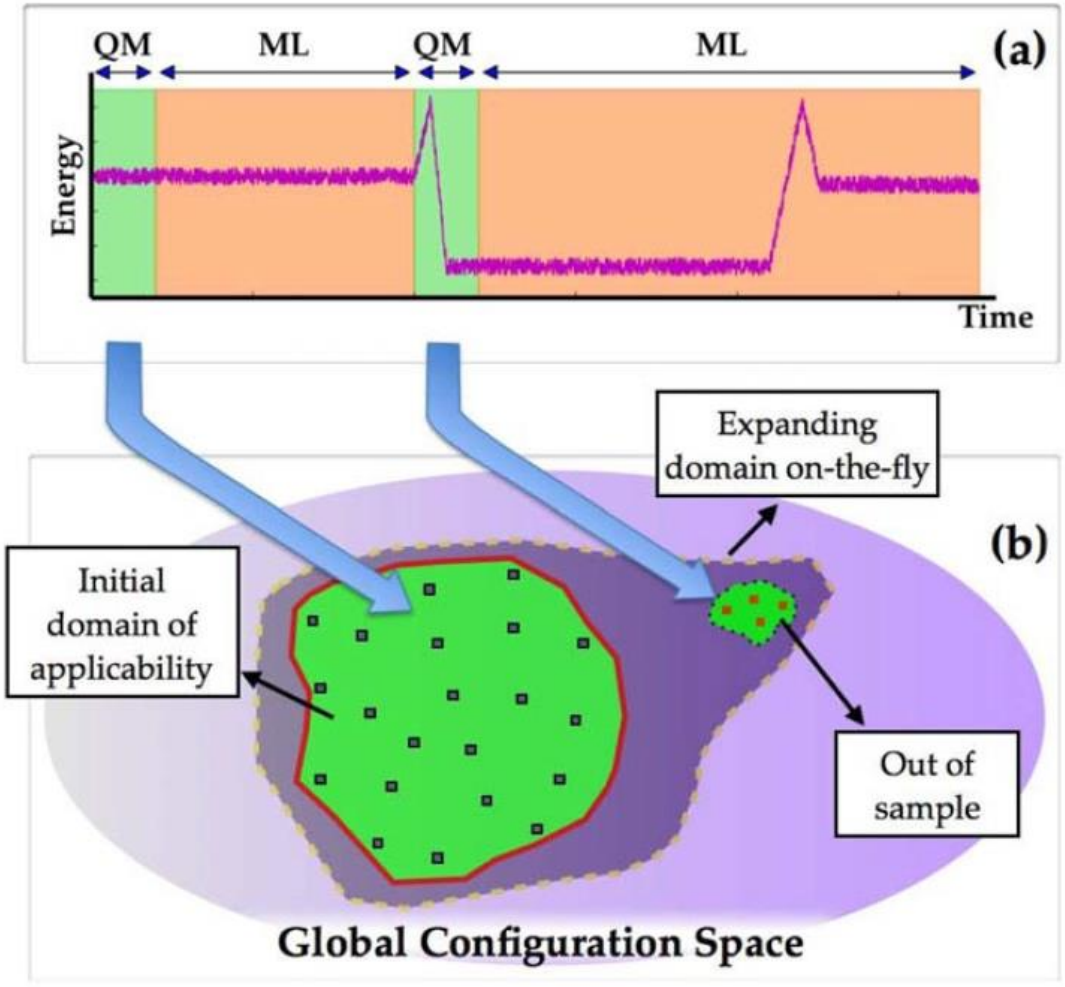
Introduction



Spend long time in local minimum.

Site-to-site hopping is a rare event.

Introduction



- Similar configuration (parameter) will have similar properties (energies, forces).
- This energies and forces can predict using Machine Learning(ML).

Figure 1. a) A typical MD energy trajectory, with the green and orange regions identifying the quantum mechanical (QM) and machine learning (ML) phases, respectively, of the adaptive learning framework. b) Expansion of the domain of applicability on-the-fly, if and when new configurations are visited. c) A flowchart of the adaptive learning framework. The green and orange arrows indicate the use of QM or ML models.

Introduction

- Similar configuration
  → Use ML properties.

- Dissimilar configuration
  → Use QM method and train.

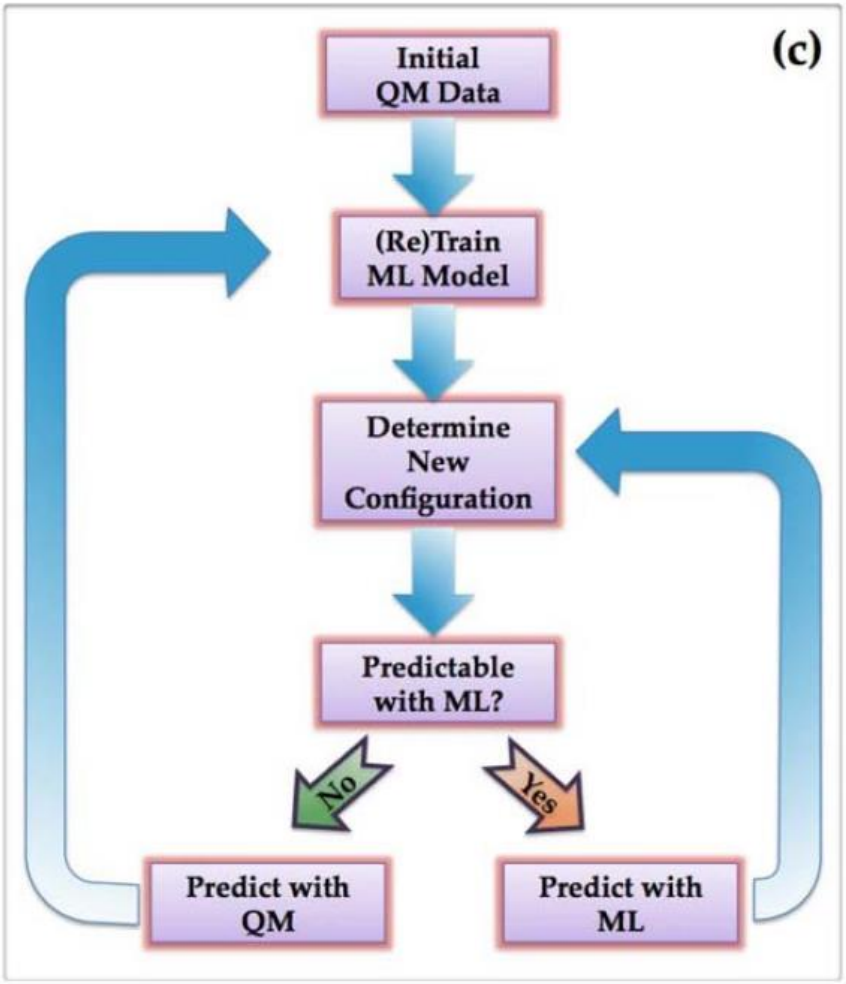


Figure 1. a) A typical MD energy trajectory, with the green and orange regions identifying the quantum mechanical (QM) and machine learning (ML) phases, respectively, of the adaptive learning framework. b) Expansion of the domain of applicability on-the-fly, if and when new configurations are visited. c) A flowchart of the adaptive learning framework. The green and orange arrows indicate the use of QM or ML models.

Method & Models

- Learning Model
  - Configuration → Feature vector

- Configuration decision engine
  - Is it Predictable Configuration?

Learning Model

Generate learning model for energy and each atom's force.

η: length of fingerprint

- Fingerprint for energy:

$$C(\eta) = \frac{1}{N}\sum_{i}^{N} A_i(\eta) \tag{5}$$

- Fingerprint for force : $\boldsymbol{V}_i(\eta) = \{V_i^x(\eta), V_i^y(\eta), V_i^z(\eta)\}$

$$V_i^k(\eta) = \sum_{j \neq i} \frac{r_{ij}^k}{r_{ij}} \mathrm{e}^{-\left(\frac{r_{ij}}{\eta}\right)^2} f(r_{ij}), \quad k \in \{x, y, z\} \tag{6}$$

Fingerprint include information of distance to near atoms.

Learning Model

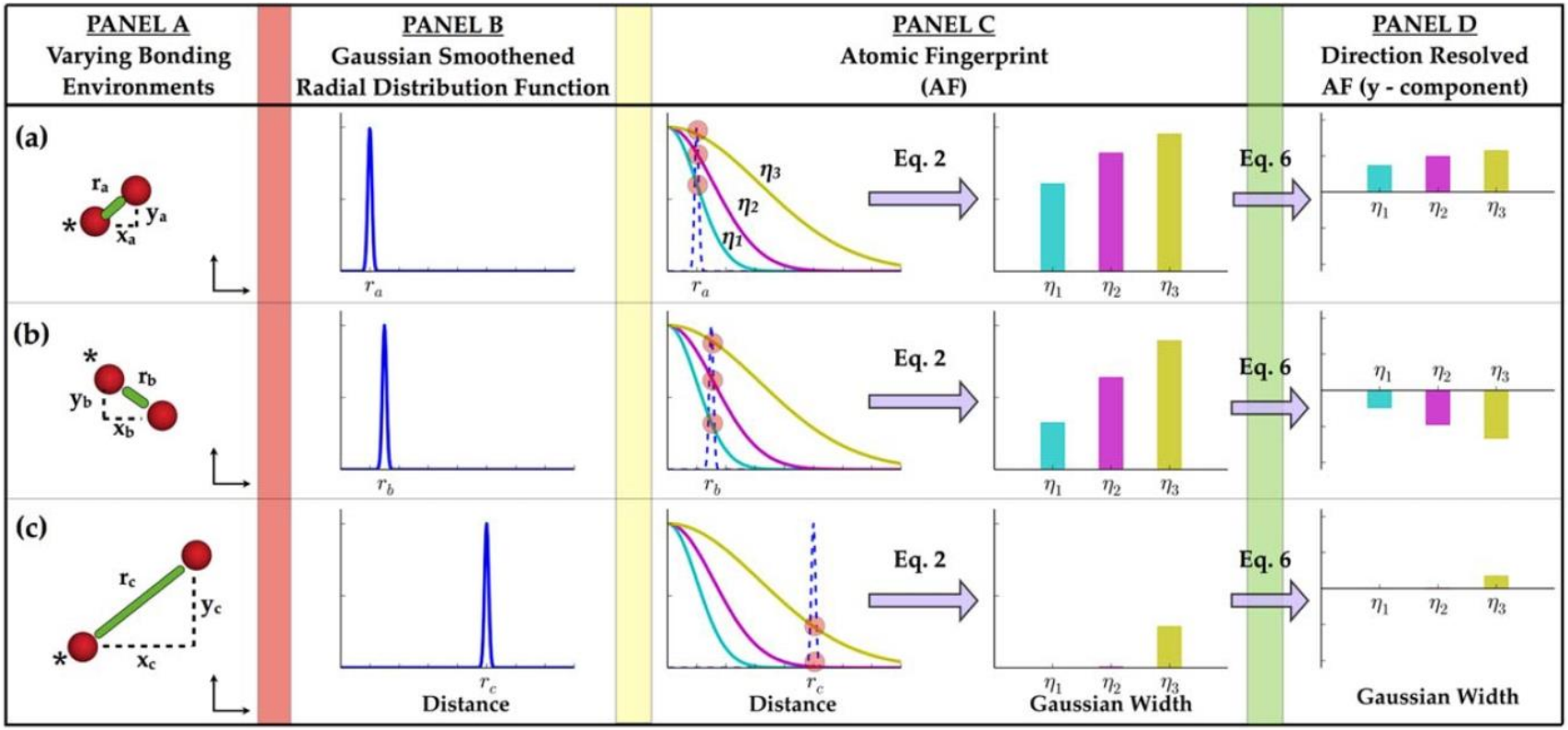


Figure 2. Panel A: A homonuclear diatomic molecule displaying three different bond lengths. Panel B: The corresponding Gaussian smoothened radial distribution function (RDF) for each of the bonding environments. Panel C: Transformation of the RDF using Gaussian functions on an eta-grid as indicated by the colored lines, into an atomic fingerprint. Panel D: The *y*-component of the direction resolved atomic fingerprint of an atom in the three bonding environments. The fingerprints generated are for the atom indicated by * in Panel A.

Learning Model

Kernel ridge regression

$$P_u = \sum_v \alpha_v e^{-\frac{1}{2}\left(\frac{|d_{uv}|}{\sigma}\right)^2} \qquad (7)$$

- 5-fold cross-validation
- σ is determined by cross-validation

Decision Engine

- If all finger print components is within range of already trained finger print.

- More complex decision engine can be developed. (not attempted in this paper.)

Predictable range (normalized)

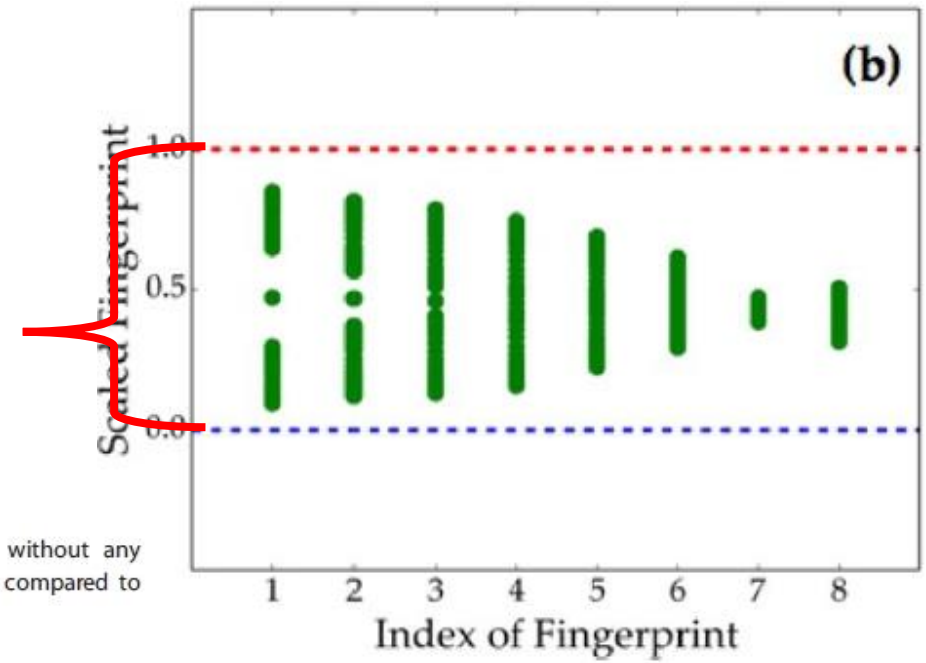


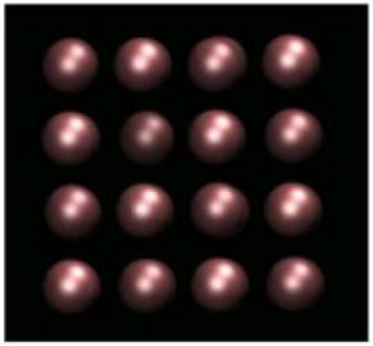
Figure 8. a) Parity plot showing accurate force prediction without any retraining, and b) Direction resolved atomic fingerprint range compared to the training dataset of the force model.

Datasets (MD)

Simulating system

(i) defect-free bulk Aluminum
(ii) Bulk Aluminum containing vacancy
(iii) Clean (111) Aluminum surface
(iv) (111) Aluminum surface with adatom

(ii) system



(111) : crystal orientation

Datasets (ML)

- Energies:
  2,000 configure for each system.

- Forces:
  32,000 configure for (i),(ii) system.
  64,000 configure for (iii),(iv) system.

Training data select randomly,
The remaining is test data.

Results

Evaluation

- Error vs. Fingerprint vector size
- Error vs. Training data size
- Prediction accuracy ML vs. QM
- Retraining

Results: Fingerprint size

Required accuracy

Energy : $< 1 \frac{meV}{atom}$

Force : $< 0.05 \frac{eV}{\text{Å}}$

- Increasing fingerprint length achieve accurate model.
- Error levels converging well below numerical DFT noise.

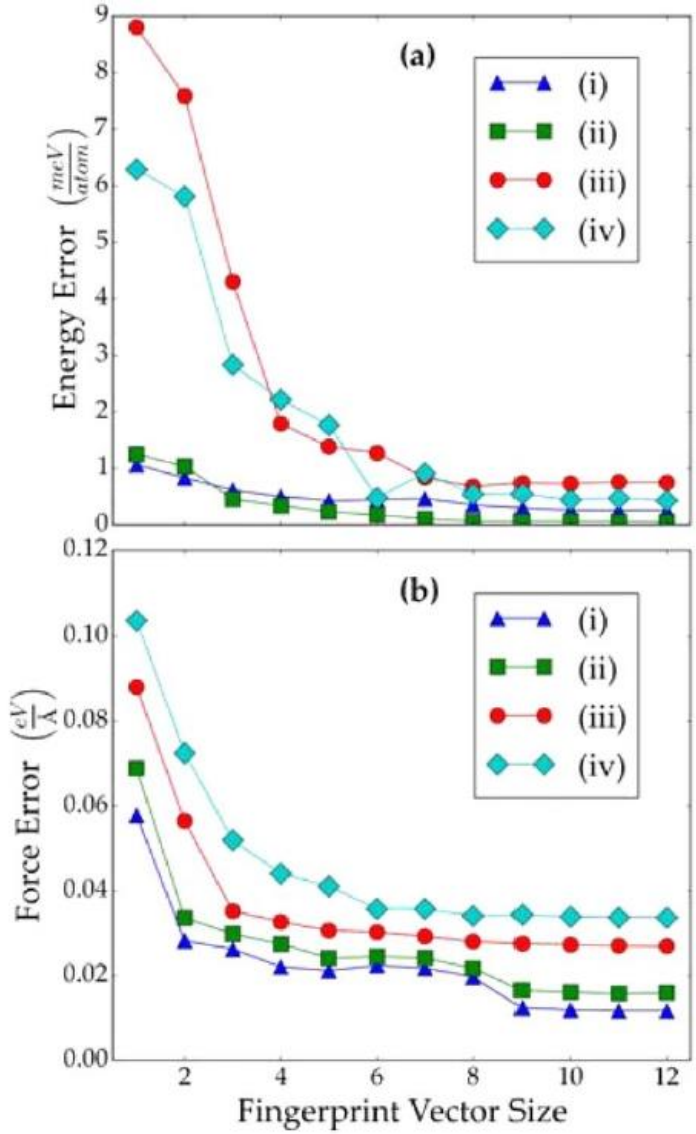


Figure 3. Energy a) and force b) error versus length of fingerprint size for (i) defect-free bulk Al, (ii) bulk Al containing a vacancy, (iii) a clean (111) Al surface, and (iv) the (111) surface with an Al adatom.

Results: Training data size

Fingerprint size = 8

- Training cost scales as $O(N^3)$. Optimizing data size is important for accuracy and acceleration.

- (i),(ii) : > 25 configure (energy)
  > 50 configure (force)

  (iii),(iv): > 50 configure (energy)
  >200 configure (force)

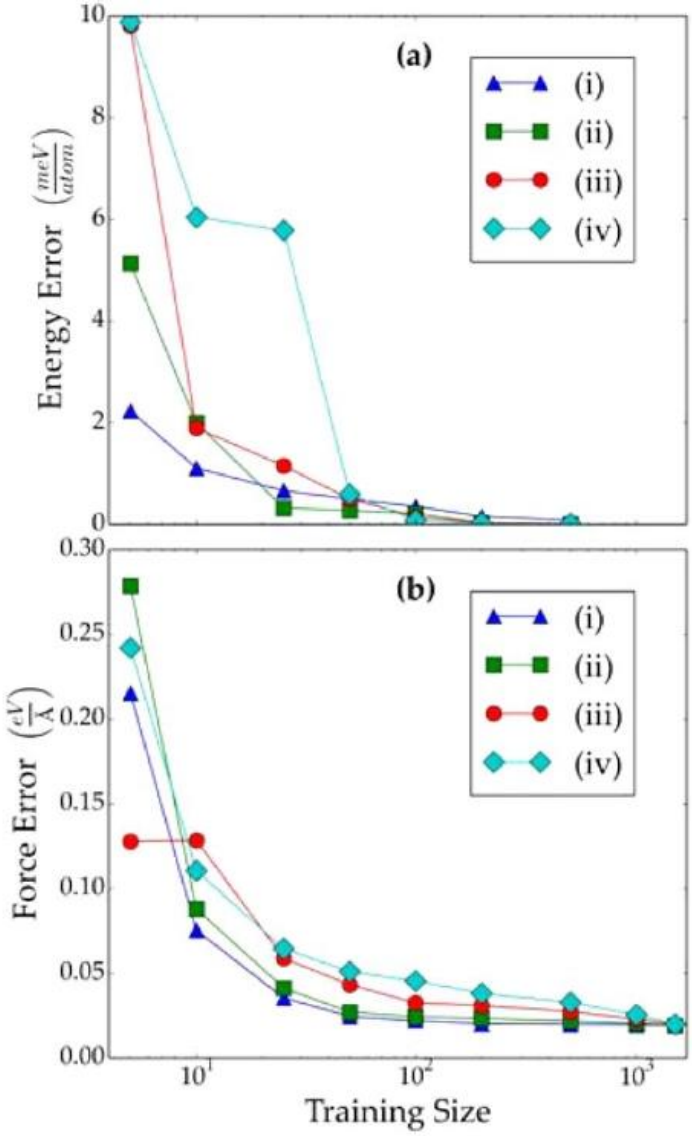


Figure 4. Energy a) and force b) error versus training size for (i) defect-free bulk Al, (ii) bulk Al containing a vacancy, (iii) a clean (111) Al surface, and (iv) the (111) surface with an Al adatom.

Results: ML vs. QM

Training Data

System	# configure for Energy	# configure for Force
(i),(ii)	100	100
(iii),(iv)	100	750

Table 1. Mean absolute error in energy and force predictions of the four cases.

Case	Energy ($\frac{meV}{atom}$)	Force ($\frac{eV}{Å}$)
(i) Defect-free bulk Al	**0.04** (0.03)	**0.02** (0.02)
(ii) Bulk Al w. vacancy	**0.06** (0.02)	**0.02** (0.02)
(iii) Clean (111) Al surface	**0.16** (0.08)	**0.03** (0.02)
(iv) (111) Surface w. adatom	**0.22** (0.07)	**0.03** (0.03)

Test error in bold and training error in brackets.

Energy : $< 1 \frac{meV}{atom}$

Force : $< 0.05 \frac{eV}{Å}$

Results: ML vs. QM

Accuracy

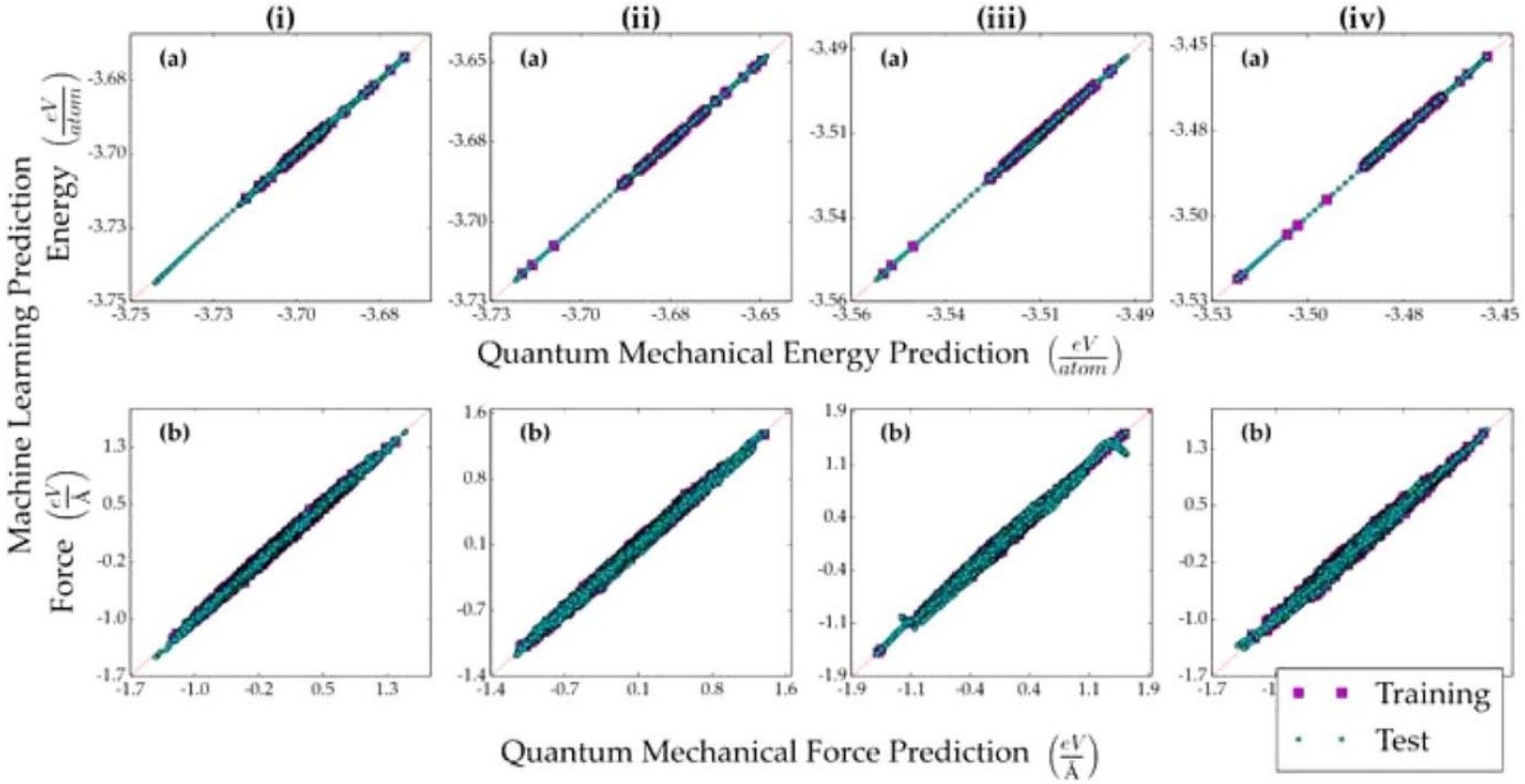


Figure 5. Parity plot for (i) defect-free bulk Al, (ii) bulk Al containing a vacancy, (iii) a clean (111) Al surface, and (iv) the (111) surface with an Al adatom, with energy a) and force b) predictions in the top and bottom rows, respectively. An eight component fingerprint, with 100 training configurations for the energy models and 100 [for (i) and (ii)] and 750 [for (iii) and (iv)] training configurations for the force models were used.

Results: ML vs. QM

Speed

- Each prediction takes roughly a millisecond
- DFT (numerical method) takes 45min on 16 core machine.

"speed up on order of 10^6"

Discussion of trade off ?? (accuracy vs. speed)

Results: Retraining

Unpredictable configure : Transition state (TS)

Vacancy migration

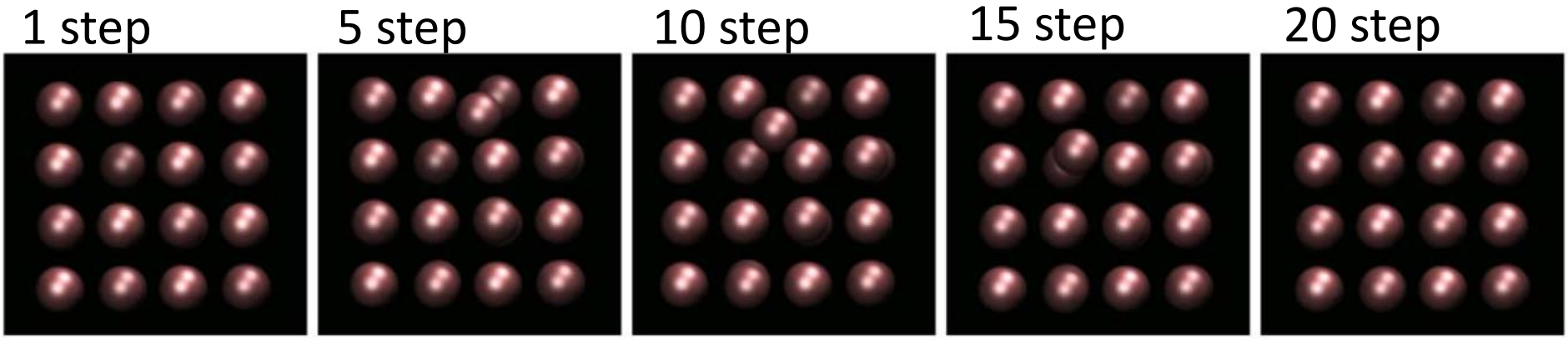


Figure 6. Vacancy migration within bulk Al. The structures shown correspond to steps 1, 5, 10, 15, and 20 along the 20-step trajectory.

Results: Retraining(Energy)

★ : Added to training data

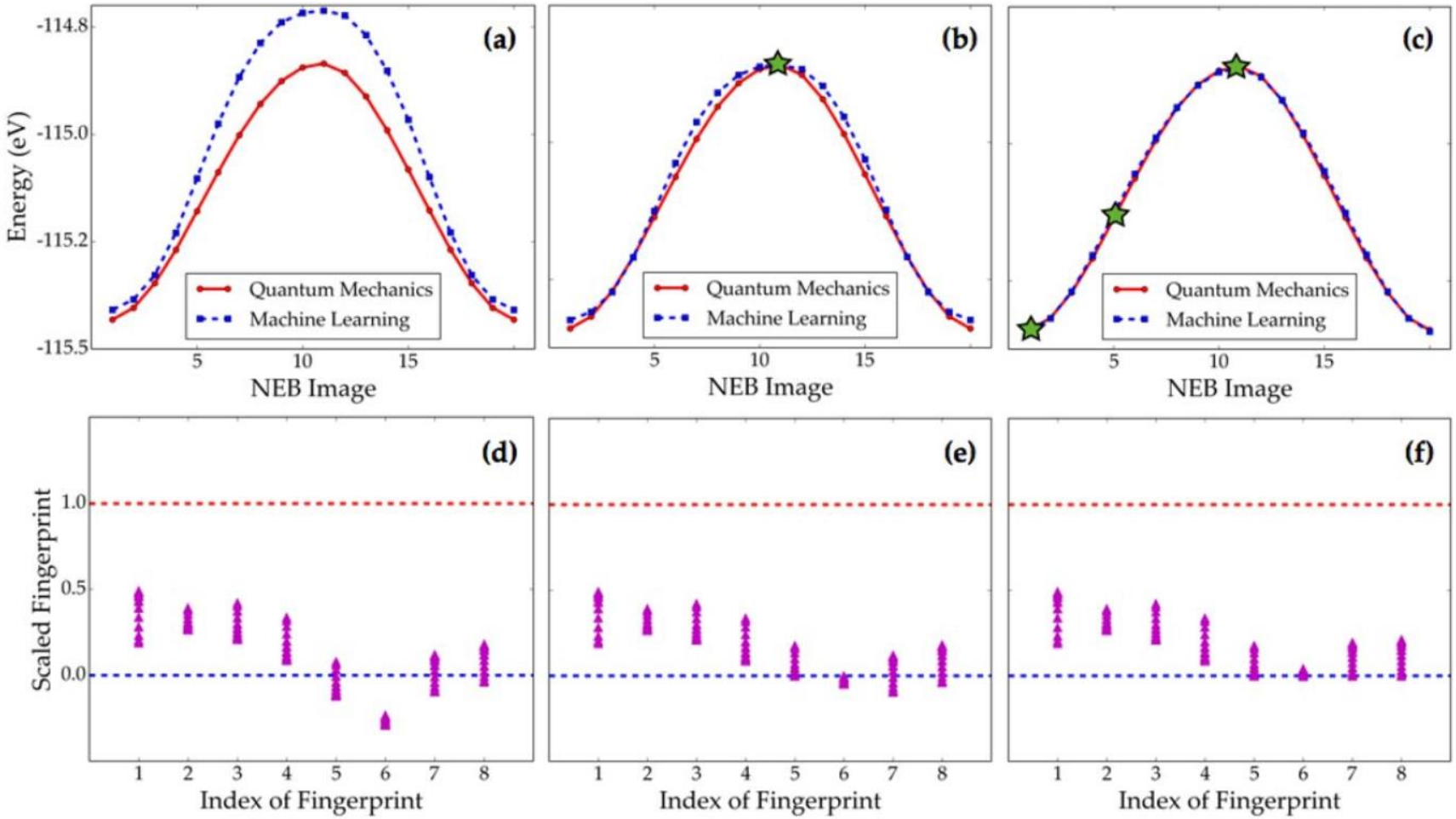


Figure 7. QM and ML energy, a)–c), and the range of crystal fingerprint components with respect to the training dataset, d)–f), of each image along the vacancy migration trajectory. a) and d) with no retraining, b) and e) with the TS added to training and c) and f) with TS and image 1 and 5 added to the training. ⋆ indicates the configurations added during retraining.

Results: Retraining(Force)

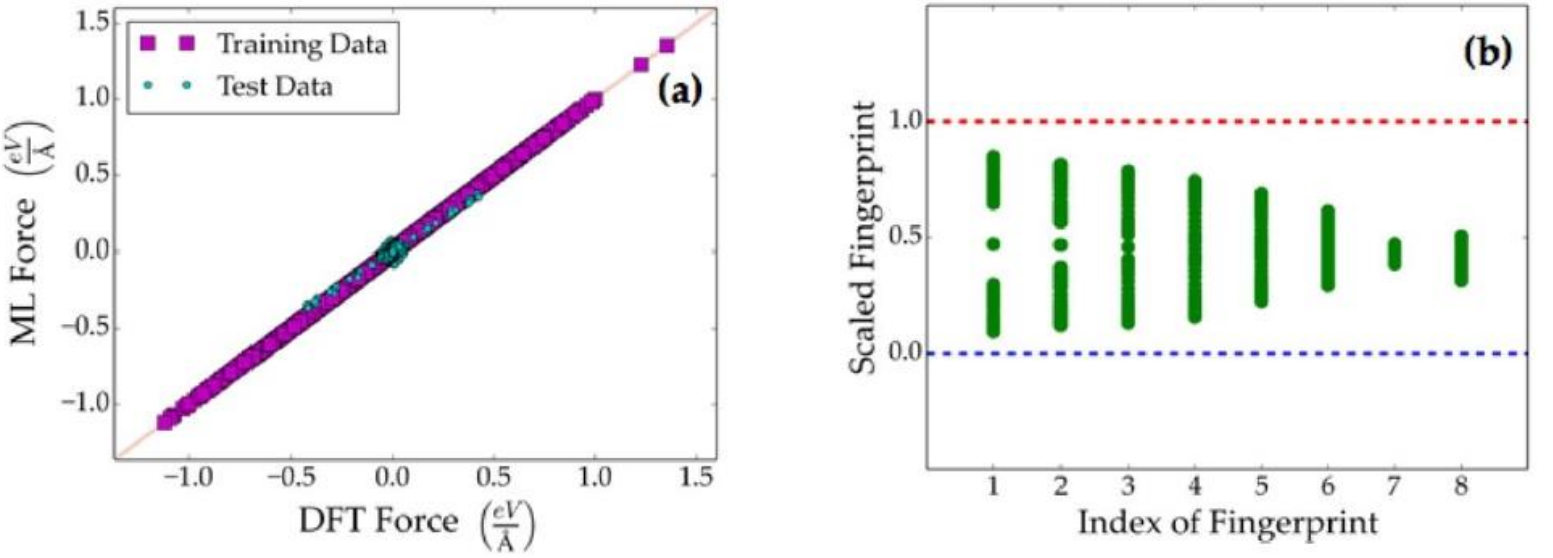


Figure 8. a) Parity plot showing accurate force prediction without any retraining, and b) Direction resolved atomic fingerprint range compared to the training dataset of the force model.

Forces are accurately predicted.

Conclusions

- The authors supposed ab initio MD scheme with ML.
- ML scheme learns previously visited configuration.
- This adaptive strategy is applicable to nonmetallic system.