HPC+AI: pushing molecular dynamics with ab initio accuracy to 100 million atoms

Lin Lin

Department of Mathematics, UC Berkeley;

Lawrence Berkeley National Laboratory

New York Scientific Data Summit 2020,

arXiv: 2005.00223

EFFICIENT LONG-RANGE CONVOLUTIONS FOR POINT CLOUDS

A PREPRINT

Yifan Peng
Zhiyuan College
Shanghai Jiao Tong University
Shanghai 200240, China
pyf04142017@sjtu.edu.cn

Lin Lin
Department of Mathematics
University of California, Berkeley
Berkeley, CA 94720
linlin@berkeley.edu

Lexing Ying
Department of Mathematics
Stanford University
Stanford, CA
lexing@stanford.edu

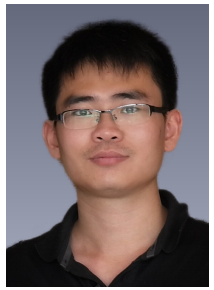
Leonardo Zepeda-Núñez
Department of Mathematics
University of Wisconsin-Madison
Madison WI 53706
zepedanunez@wisc.edu

October 13, 2020

ABSTRACT

The efficient treatment of long-range interactions for point clouds is a challenging problem in many scientific machine learning applications. To extract global information, one usually needs a large window size, a large number of layers, and/or a large number of channels. This can often significantly increase the computational cost. In this work, we present a novel neural network layer that directly incorporates long-range information for a point cloud. This layer, dubbed the long-range convolutional (LRC)-layer, leverages the convolutional theorem coupled with the non-uniform Fourier transform. In a nutshell, the LRC-layer mollifies the point cloud to an adequately sized regular grid, computes its Fourier transform, multiplies the result by a set of trainable Fourier multipliers, computes the inverse Fourier transform, and finally interpolates the result back to the point cloud. The resulting global all-to-all convolution operation can be performed in nearly-linear time asymptotically with respect to the number of input points. The LRC-layer is a particularly powerful tool when combined with local convolution as together they offer efficient and seamless treatment of both short and long range interactions. We showcase this framework by introducing a neural network architecture that combines LRC-layers with short-range convolutional layers to accurately learn the energy and force associated with a N-body potential. We also exploit the induced two-level decomposition and propose an efficient strategy to train the combined architecture with a reduced number of samples.

Our team

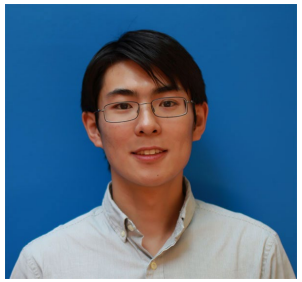


Weile Jia (Berkeley)

Han Wang (CAEP)

Mohan Chen (Peking)

Denghui Lu (Peking)



Lin Lin (Berkeley)

Weinan E (Princeton)

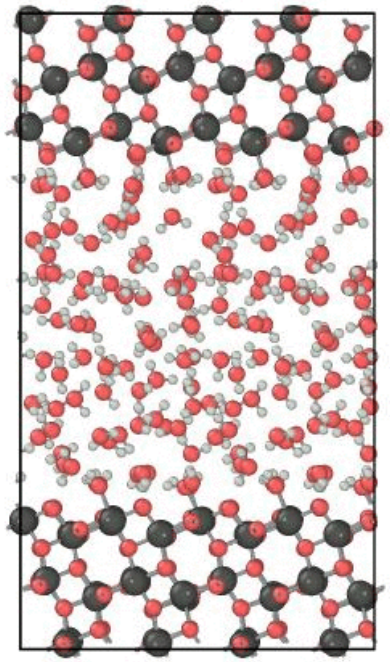
Roberto Car (Princeton)

Linfeng Zhang (Princeton)

Molecular dynamics

$$E = E\left(R_1, R_2, \ldots, R_M\right)$$

$$M_I \frac{d^2 R_I}{dt^2} = F_I = -\nabla_{R_I} E\left(R_1, R_2, \ldots, R_M\right)$$



Example: water on TiO2 surface

MD and COVID-19

Two main approaches

Calculate energy E and force F_I:

- Computing on the fly using quantum mechanics (e.g. Kohn-Sham density functional theory). Accurate but expensive: *ab initio* molecular dynamics (AIMD)

$$E = \langle \Psi_0 | H_e^{KS} | \Psi_0 \rangle$$

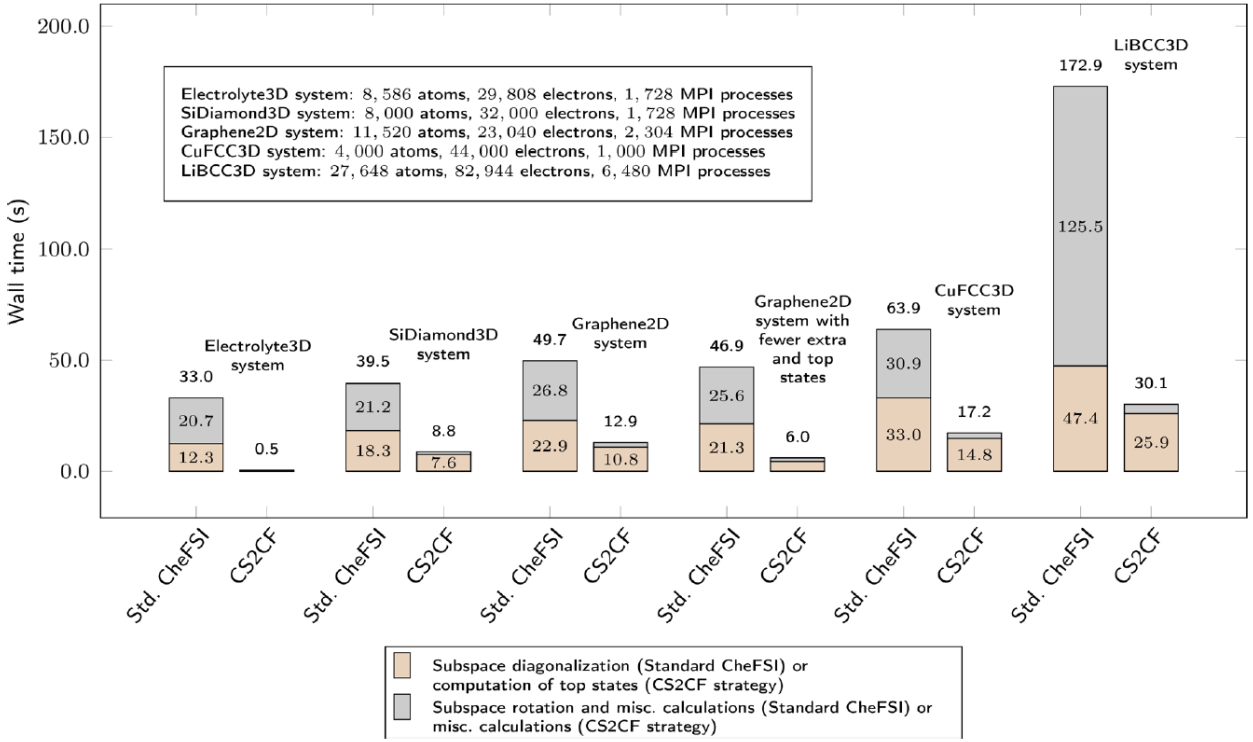
Routinely done for hundreds of atoms, 1 picosecond (10^{-12} s) or less per day

- Empirical potentials: efficient but maybe much less accurate, e.g. TIP3P for water

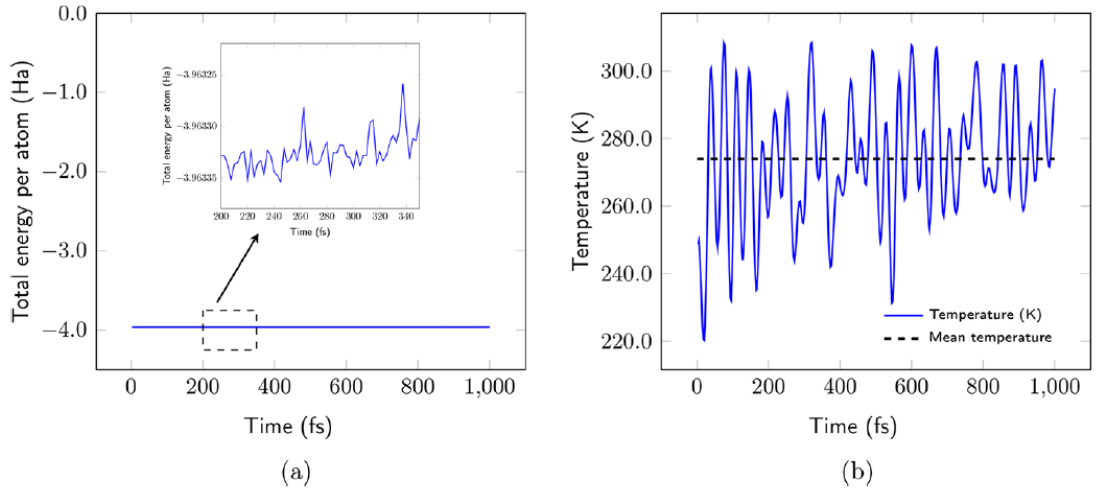
$$E = \sum_i \sum_j \frac{k_C q_i q_j}{r_{ij}} + \frac{A}{r_{00}^{12}} - \frac{B}{r_{00}^{6}}$$

Routinely done for millions to billions of atoms, nanosecond (10^{-9} s) to microsecond (10^{-6} s) per day

Pushing the limit of *ab initio* molecular dynamics with reduced scaling algorithms and supercomputers



CS2CF: Two-level Chebyshev filter based complementary subspace method



AIMD simulation of 8000 Si atoms (32000 electrons) for 1 ps (10^{-12}s)

Total number of CPU cores: 34560.

28 hour wall clock time

Nearly 1 million CPU hours.

[**L.**, Lu, Ying and E, J. Comput. Phys. 2012]

[Banerjee, **L.**, Suryanarayana, Yang, Pask, J. Chem. Theory Comput. 2018]

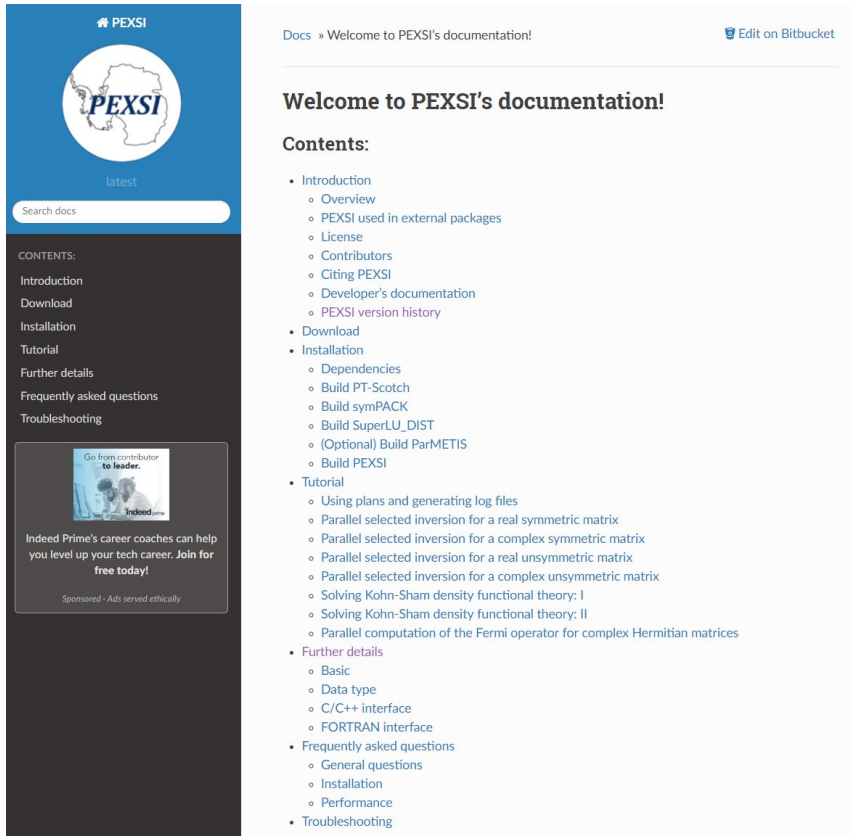
Pole expansion and selected inversion (PEXSI)

- At most $O(N^2)$ scaling (insulators, semiconductor and metals). Standard method scales as $O(N^3)$

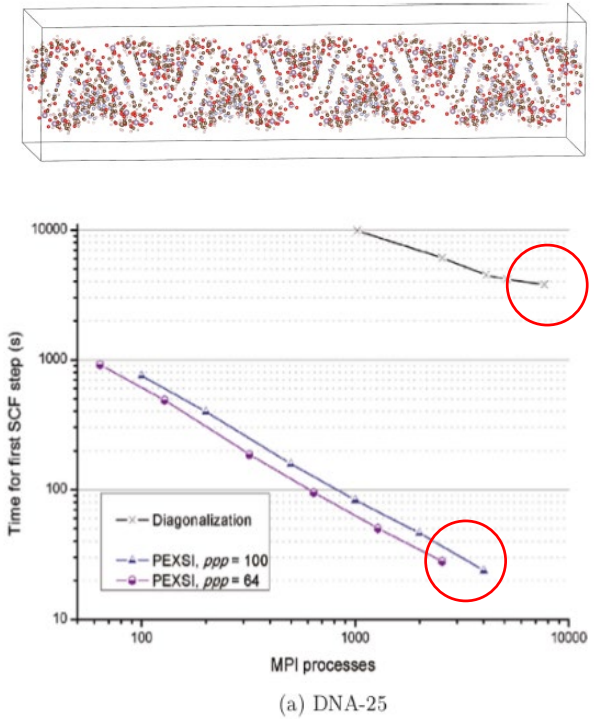
- Integrated with a number of community electronic structure software packages
- Solve systems > 10000 atoms.
- Efficiently use 10,000-100,000 cores.
- BigDFT
- CP2K
- DFTB+
- DGDFT
- FHI-aims
- QuantumWise AtK
- SIESTA

- "Electronic structure infrastructure" (ELSI) https://wordpress.elsi-interchange.org/

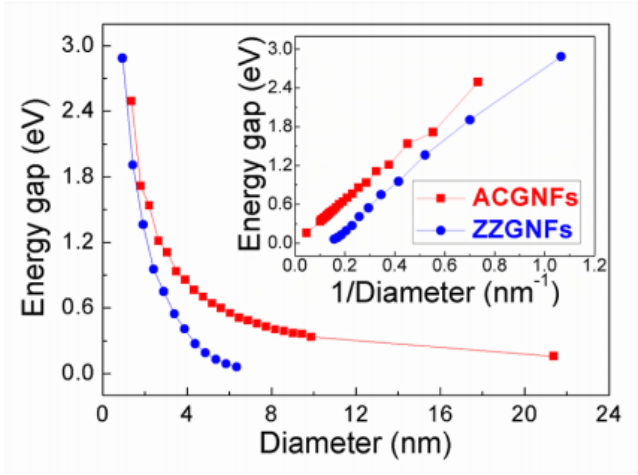
http://www.pexsi.org/



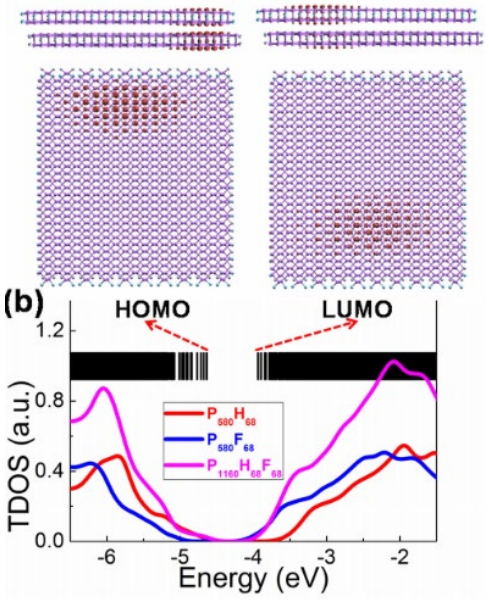
Solving quantum mechanics with ~ 10000 atoms: Pole expansion and selected inversion (PEXSI)



Large scale DNA calculation (20000 atoms)



Electronic structure of large scale graphene nanoflake (10000 atoms)

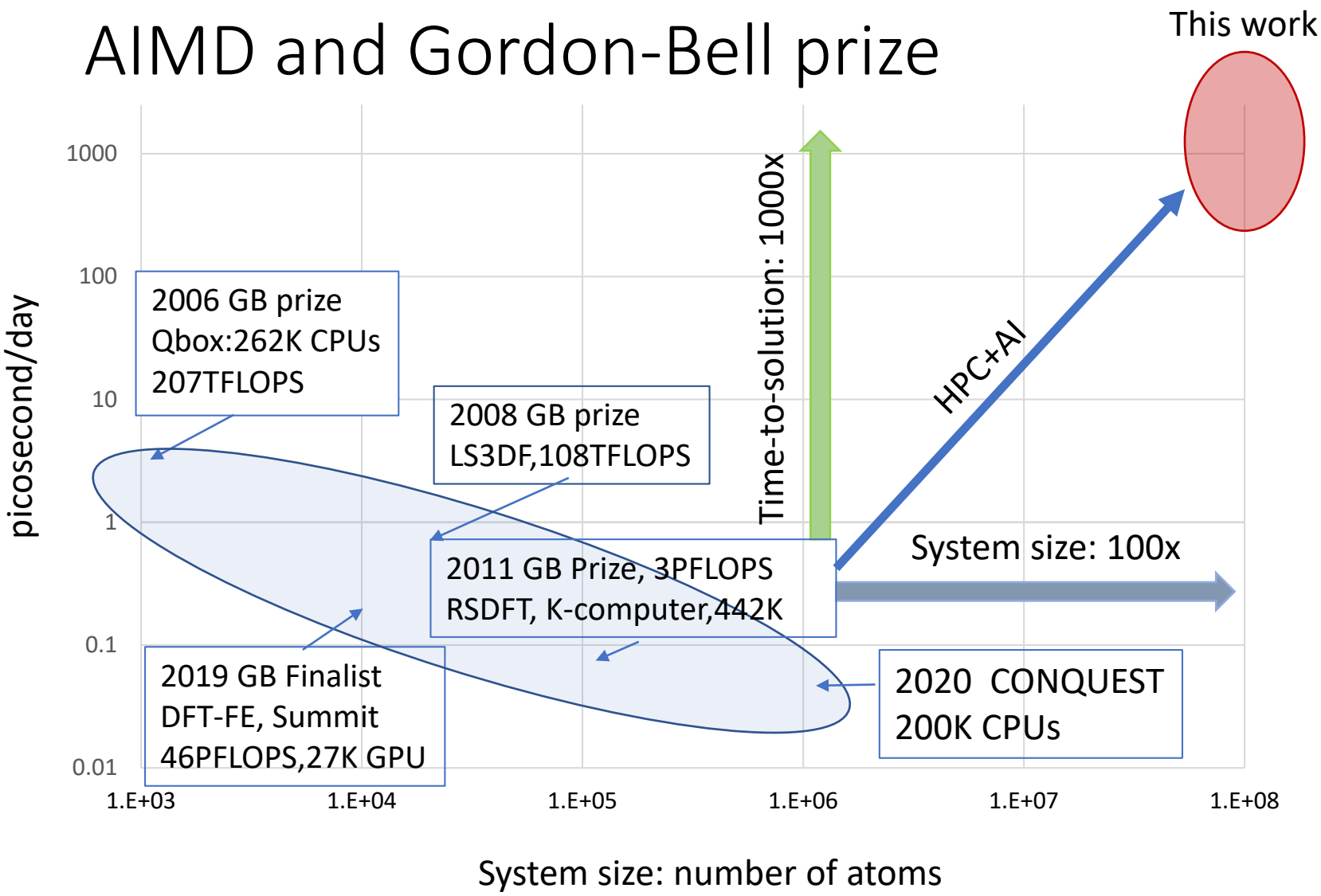


Predict large scale phospherene nanoflake (PNF) heterojunction as new candidates of solar cells (9000 atoms)

[**L.**, Lu, Ying, Car and E, Commun. Math. Sci. 2009]
[**L.**, Garcia, Huhs, Yang, JPCM 2014]
[Hu, **L.**, Yang, Yang, J. Chem. Phys. 2014]

[Hu, **L.** and Yang, Phys. Chem. Chem. Phys. 2015]
[Hu, **L.**, Yang, Dai and Yang, Nano Lett., 2016]

AIMD and Gordon-Bell prize



Compared to state-of-the-art

>10^3 times faster

>10^2 times bigger

DP: 91PFLOPS (45% of the peak)

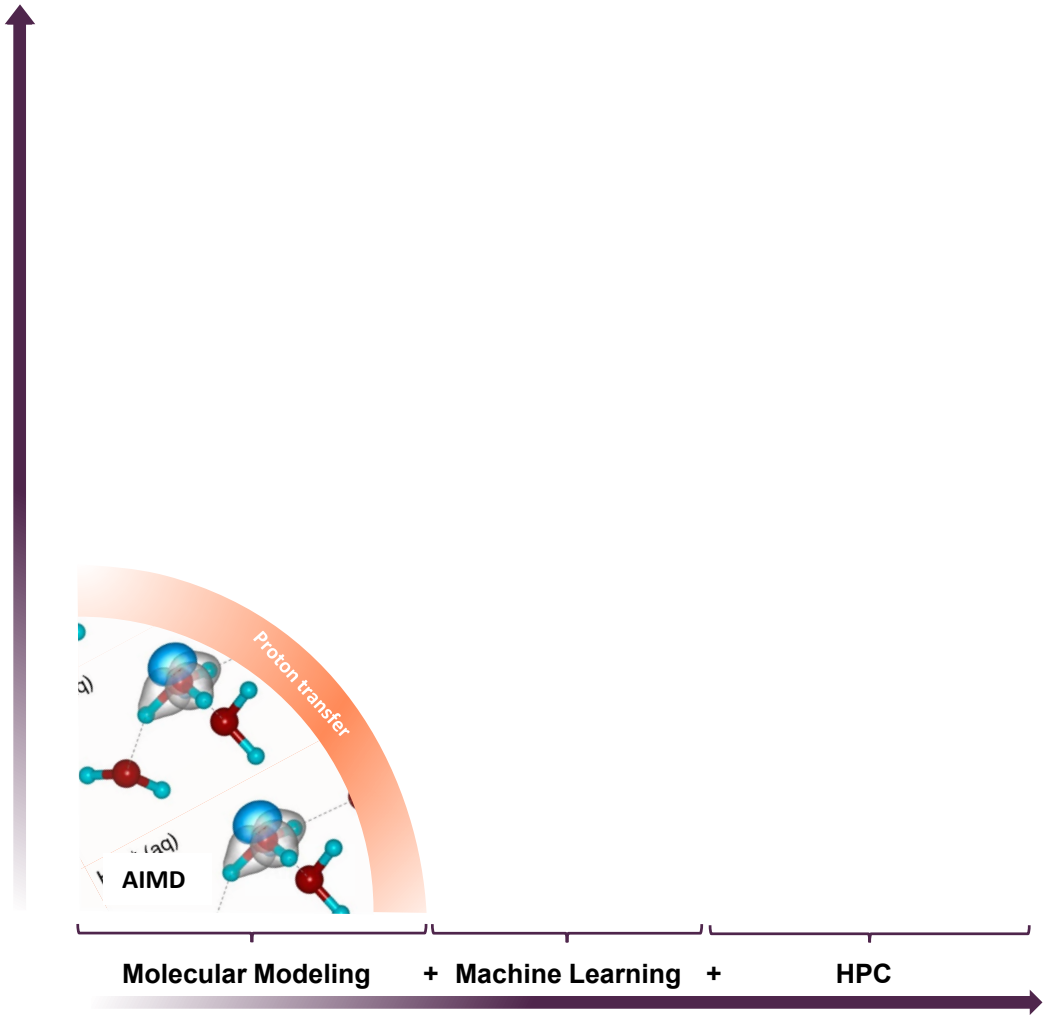
Mixed-SP: 162 PFLOPS

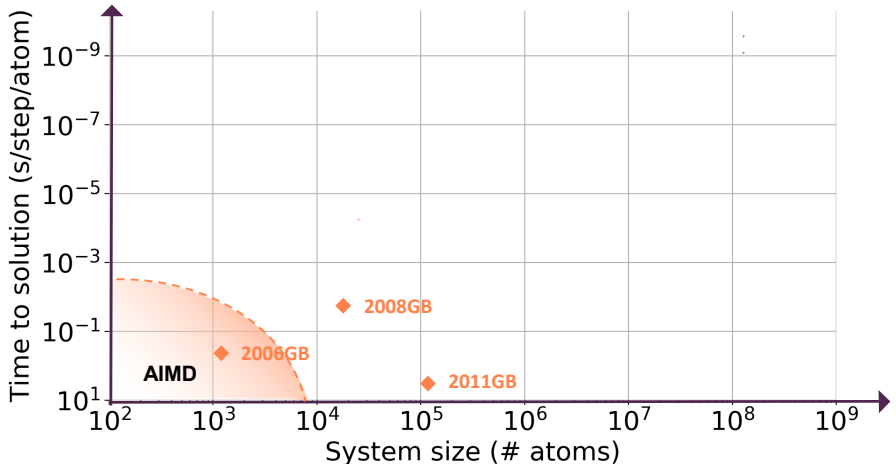
Mixed-HP: 275 PFLOPS

SC20 Gordon Bell Prize Finalist

arXiv: 2005.00223

Ab initio Molecular Dynamics (AIMD): Solving DFT "on-the-fly"





Advantages: General and accurate
Limitations: Time and size scales

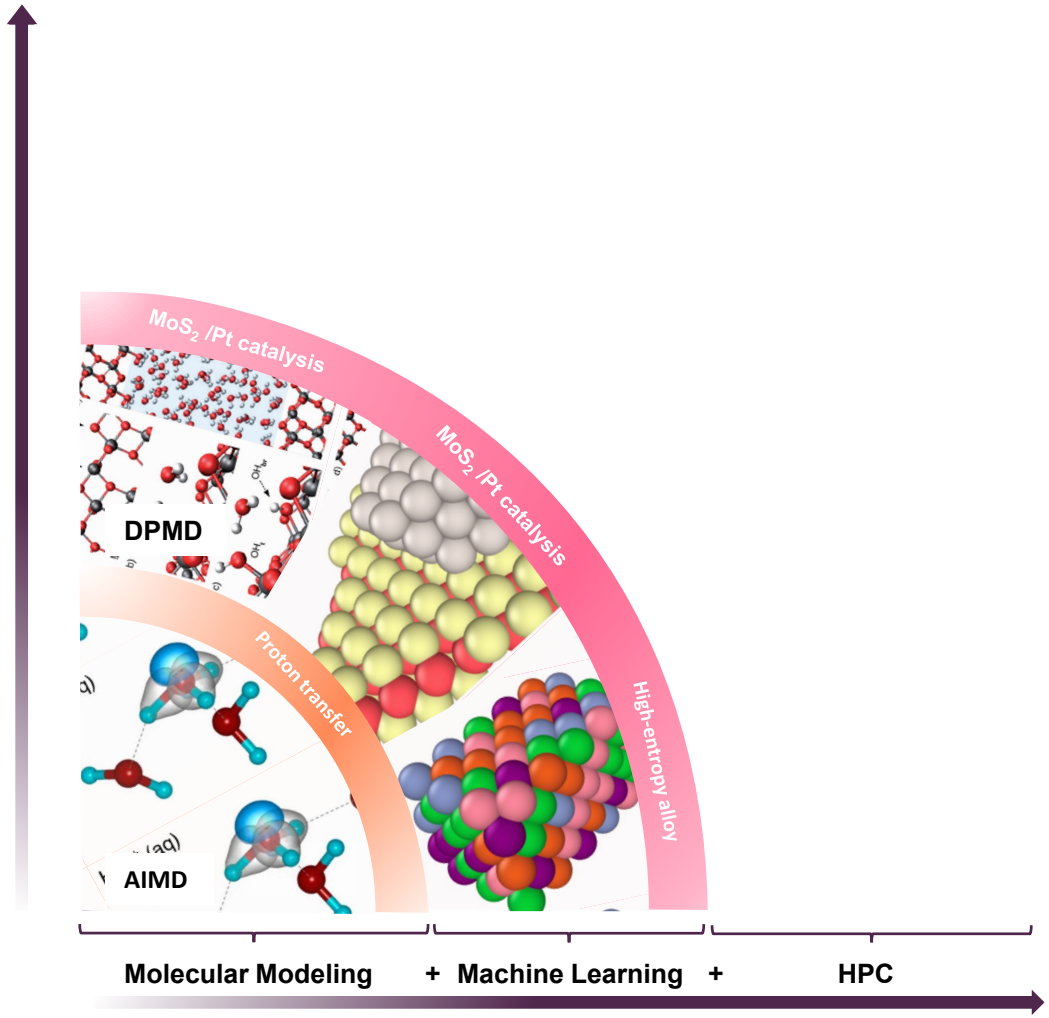
Density Functional Theory (DFT) solver

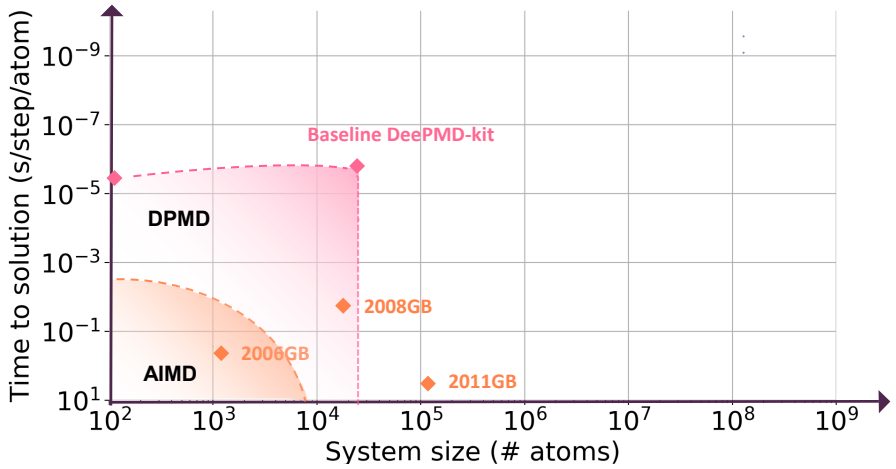
Energy
Force
Virial

Coordinate
Type
Cell tensor

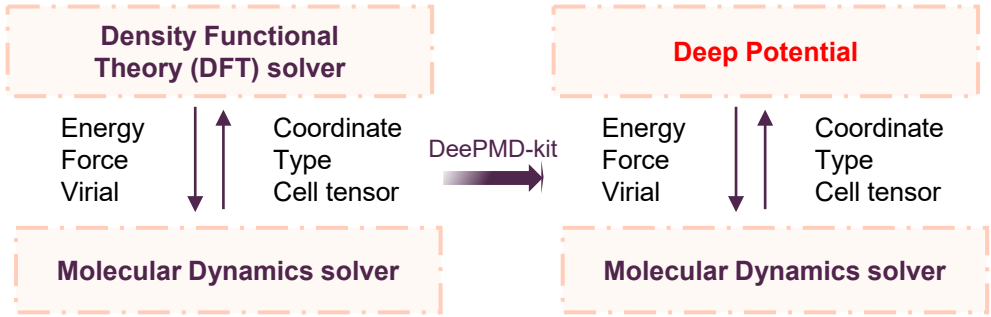
Molecular Dynamics solver

Deep Potential Molecular Dynamics (DPMD): boosting AIMD with ML



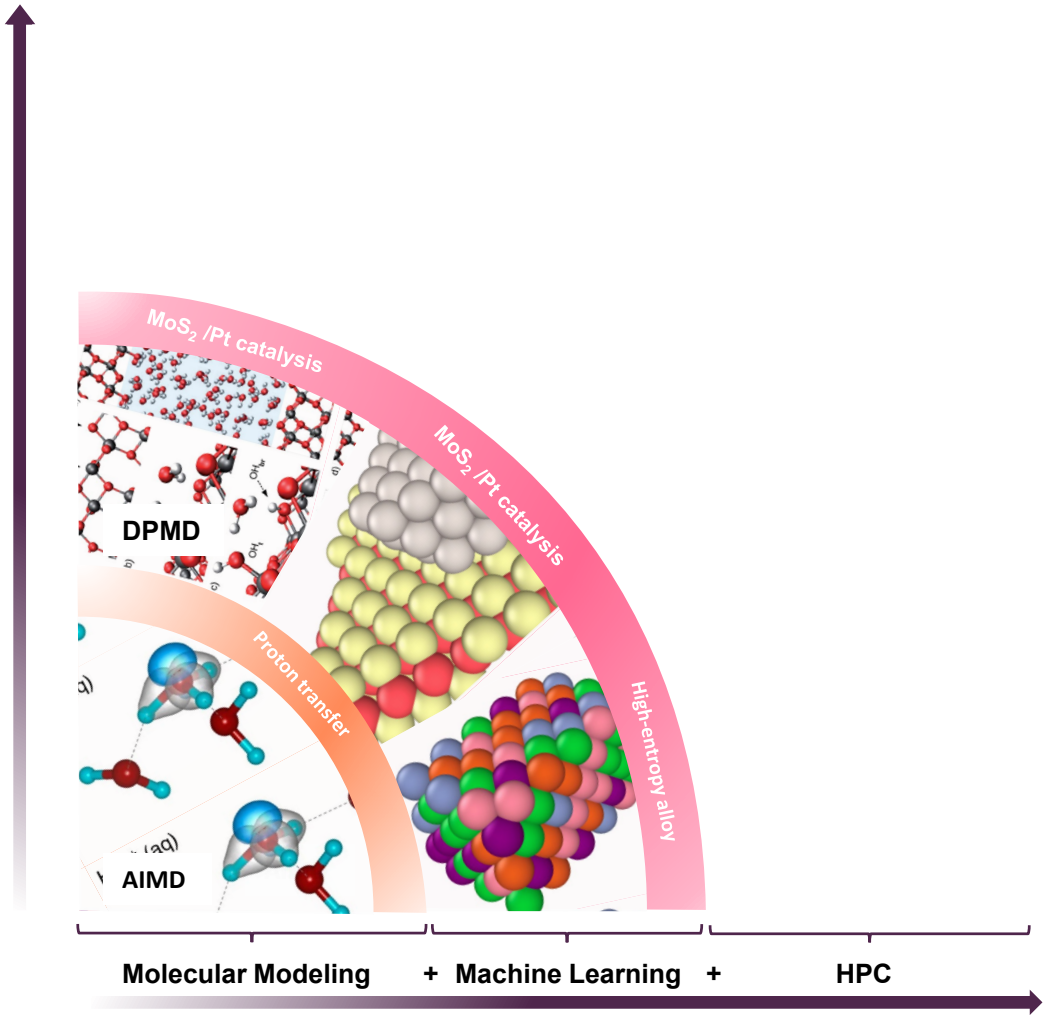


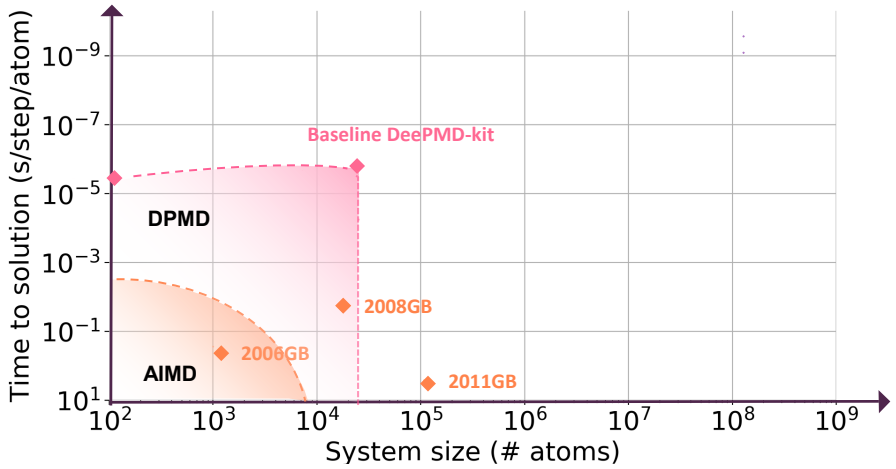
Deep Potential:
Physical requirements + machine learning



Zhang et al, Phys. Rev. Lett. 2018;
Zhang et al, NeurIPS 2018.

Deep Potential Molecular Dynamics (DPMD): boosting AIMD with ML

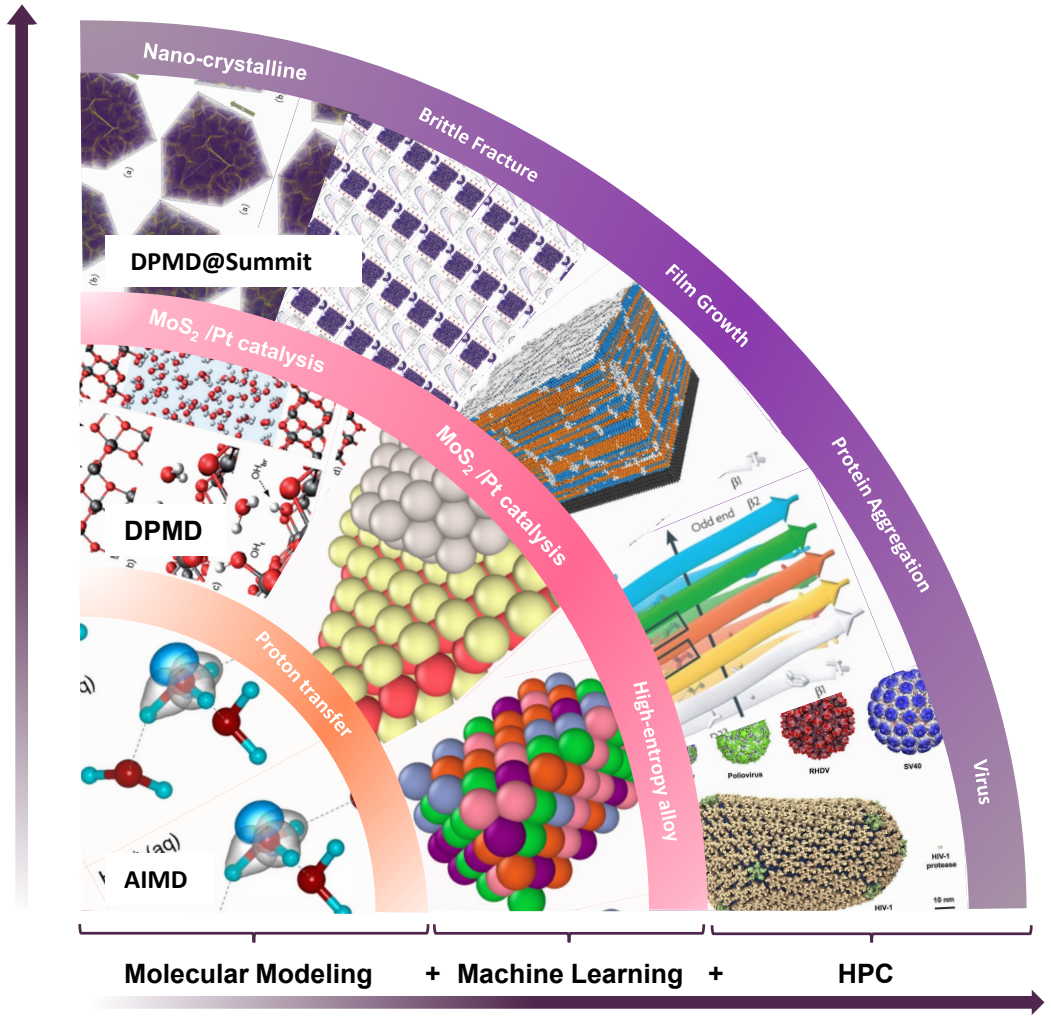


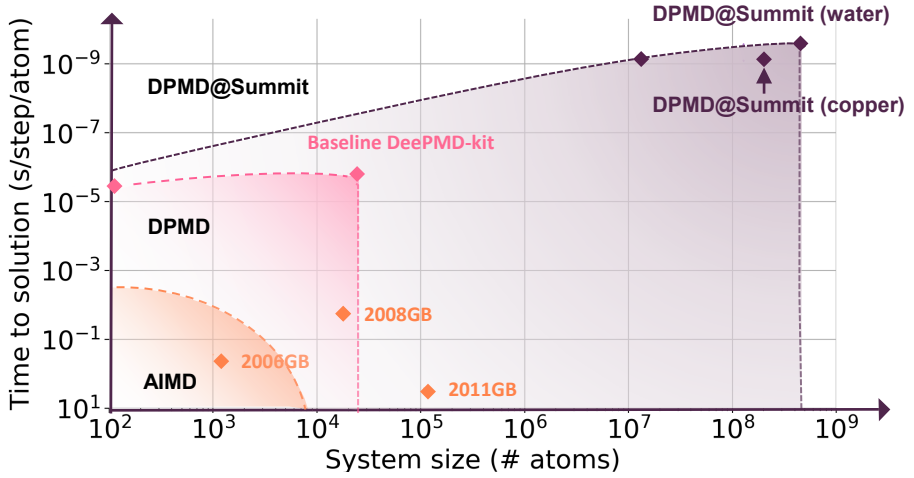


Time and size scales required by important Problems

Problem	Time span [ns]	System size [#atom]
Droplet coalescence	~10	~1e+8
Dynamic fracture	~0.1	~1e+8
Strength of nanocrystalline metal	~0.01	~1e+6
Heterogeneous aqueous interfaces	~100	~1e+6

This work: molecular modeling + machine learning + HPC





Time and size scales required by important Problems

Problem	Time span [ns]	System size [#atom]
Droplet coalescence	~10	~1e+8
Dynamic fracture	~0.1	~1e+8
Strength of nanocrystalline metal	~0.01	~1e+6
Heterogeneous aqueous interfaces	~100	~1e+6

Method: Deep Potential Molecular Dynamics

Machine learning:
Representing high-dimensional functions

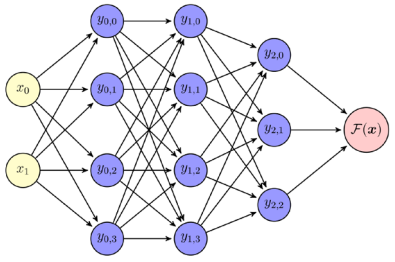
Physical principles:
Extensive property; Symmetry invariance

Representation:

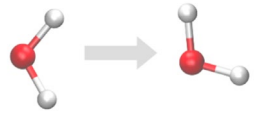
$$\mathcal{F}(\mathbf{x},\omega) = \mathcal{L}^{\text{out}} \circ \mathcal{L}^{2} \circ \mathcal{L}^{1} \circ \mathcal{L}^{0}(\mathbf{x})$$
$$\mathcal{L}^{\text{i}}(\mathbf{y}) = \tanh(\mathbf{W}^{i} \cdot \mathbf{y} + \mathbf{b}^{i}),\ i = 0,1,2$$
$$\mathcal{L}^{\text{out}}(\mathbf{y}) = \mathbf{W} \cdot \mathbf{y} + \mathbf{b}$$
$$\omega = \{\mathbf{W}^{i}, \mathbf{b}^{i}, \mathbf{W}, \mathbf{b} \mid i = 0,1,2\}$$

Optimization/Training:

$$\min_{\omega} \sum_{d\in\mathcal{D}} ||\mathcal{F}(\mathbf{x}_d,\omega) - f(\mathbf{x}_d)||$$



Translational 

Rotational 

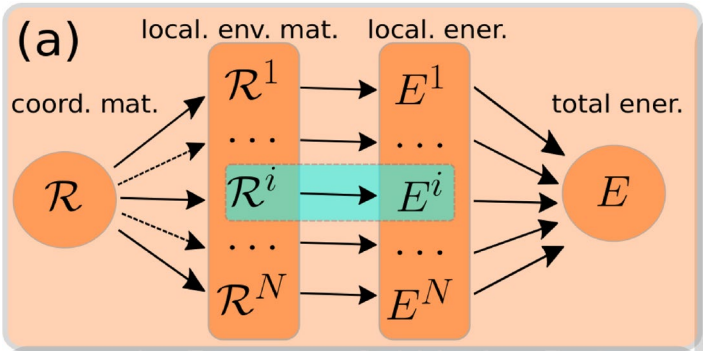
Permutational 

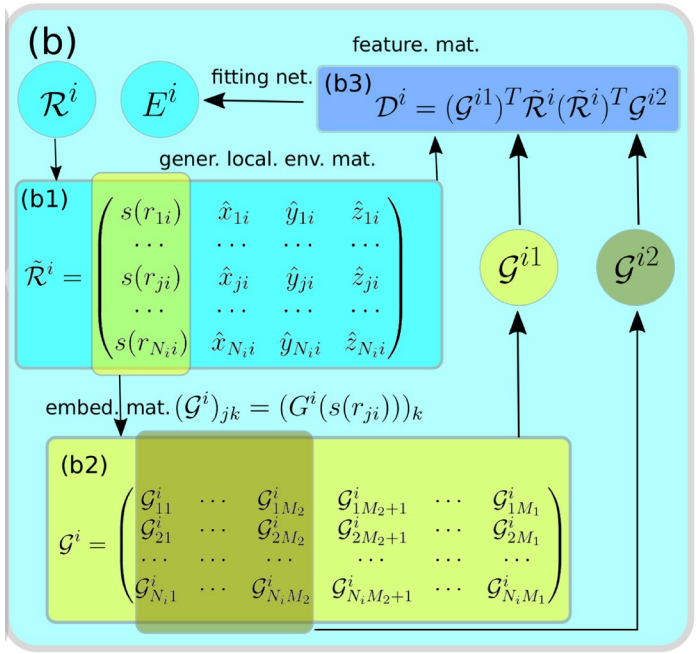
$$E(r_0, r_1, r_2) = E(r_0+s, r_1+s, r_2+s)$$

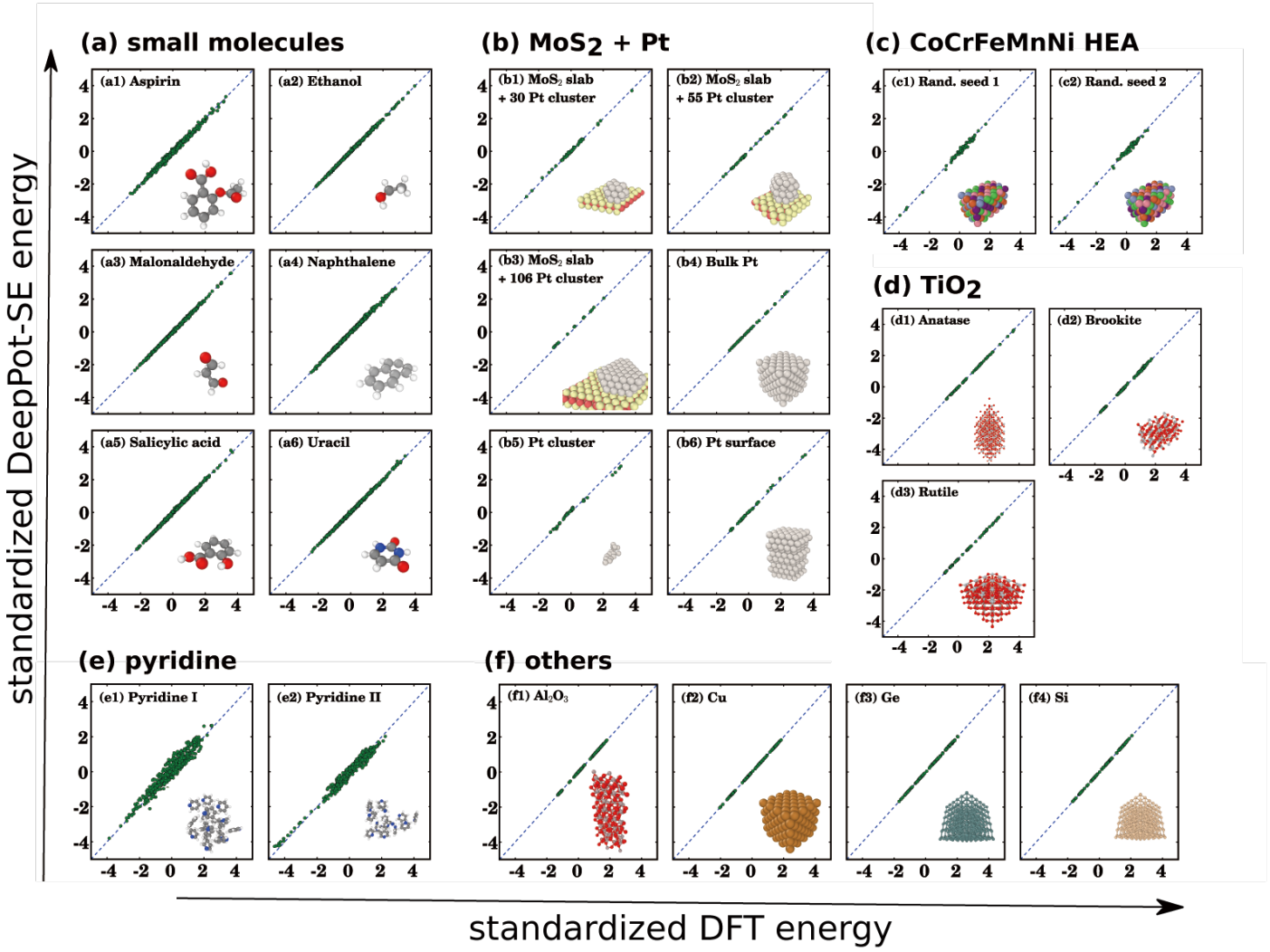
$$E(r_0, r_1, r_2) = E(Ur_0, Ur_1, Ur_2)$$

$$E(r_0, r_1, r_2) = E(r_0, r_2, r_1)$$

Method: Deep Potential Molecular Dynamics

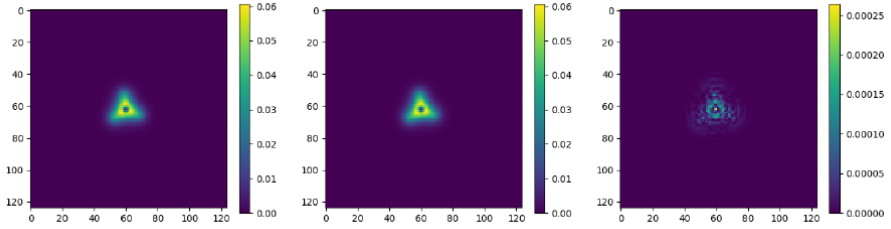




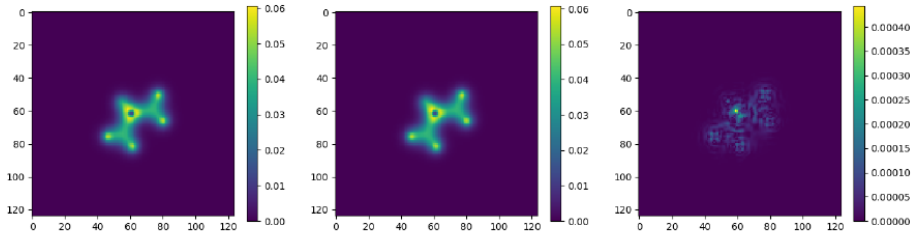


Machine learning of the electron density

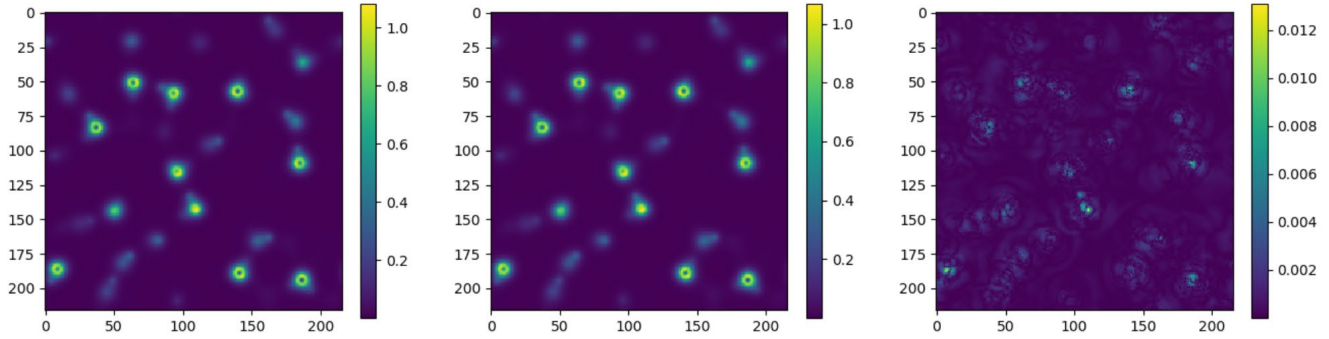
ethane

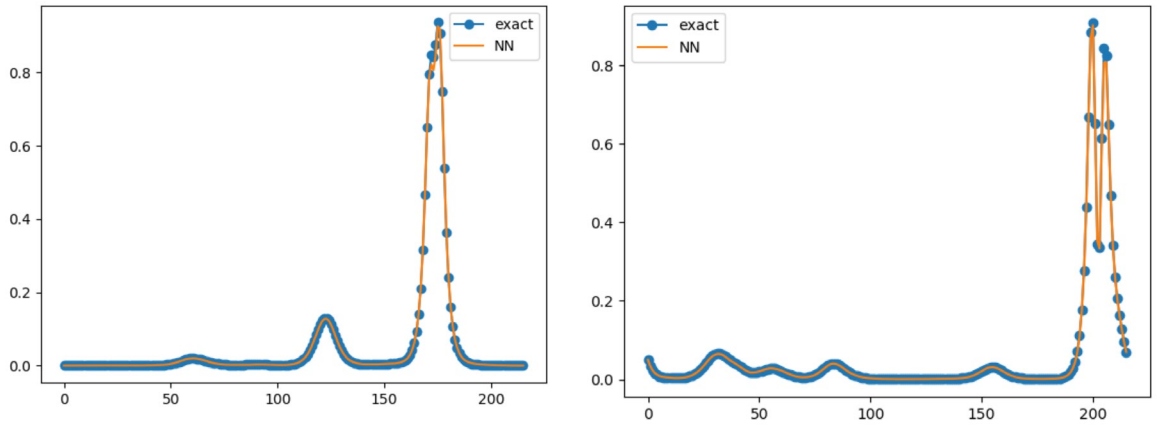


isobutene



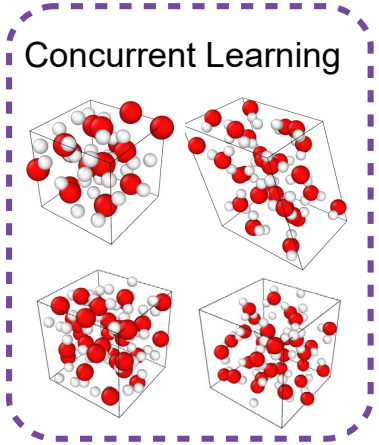
water 256 molecules



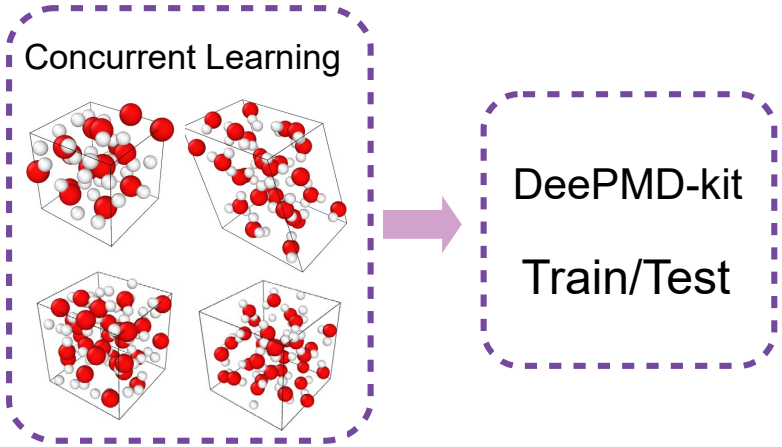


[Zepeda-Nunez, Chen, Zhang, Jia, Zhang, **L.**, 1912.00775]

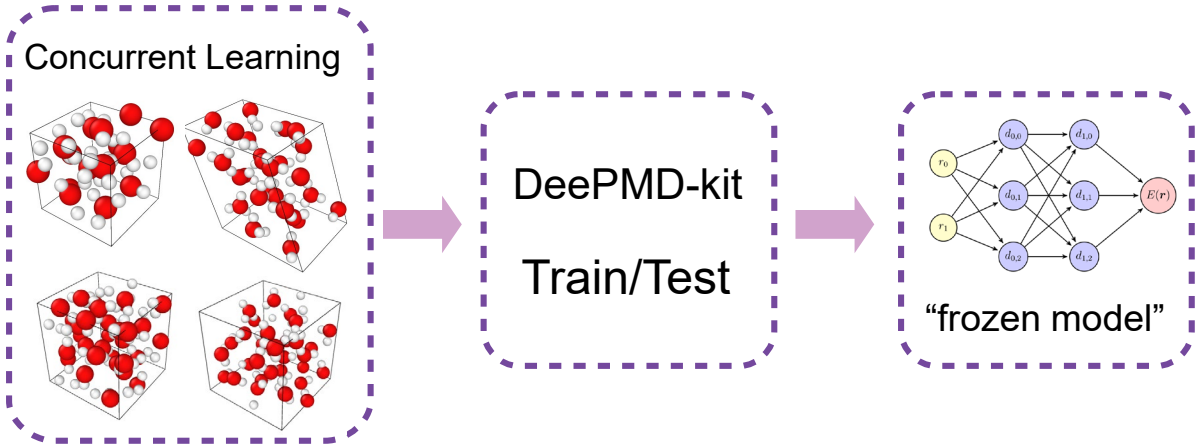
Method: Deep Potential Molecular Dynamics



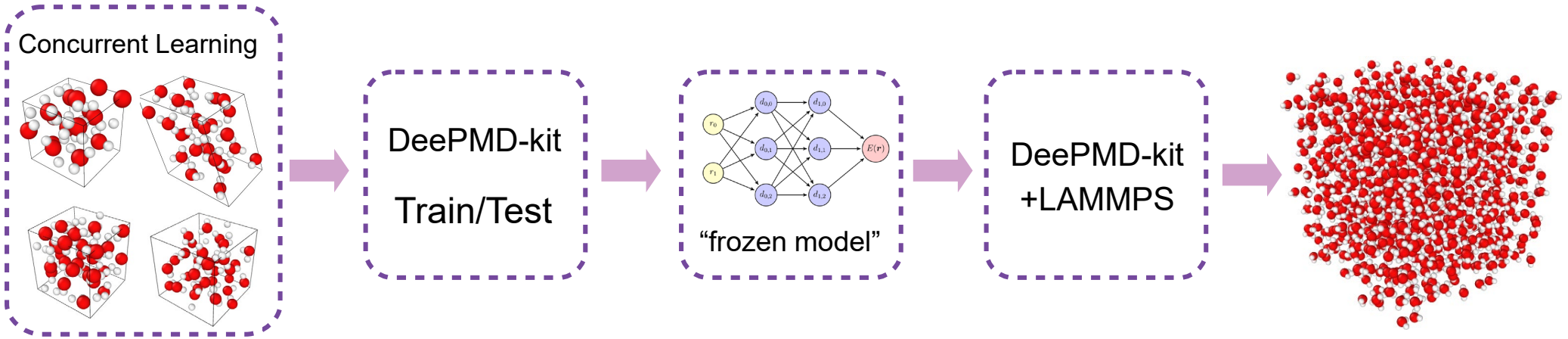
Method: Deep Potential Molecular Dynamics



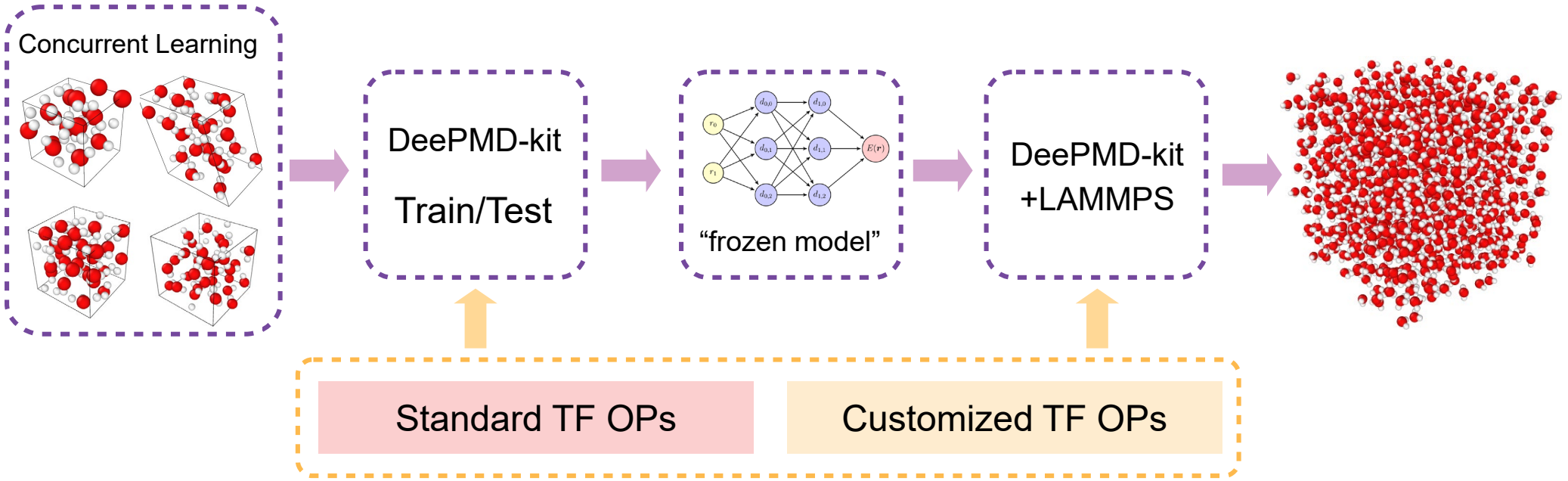
Method: Deep Potential Molecular Dynamics



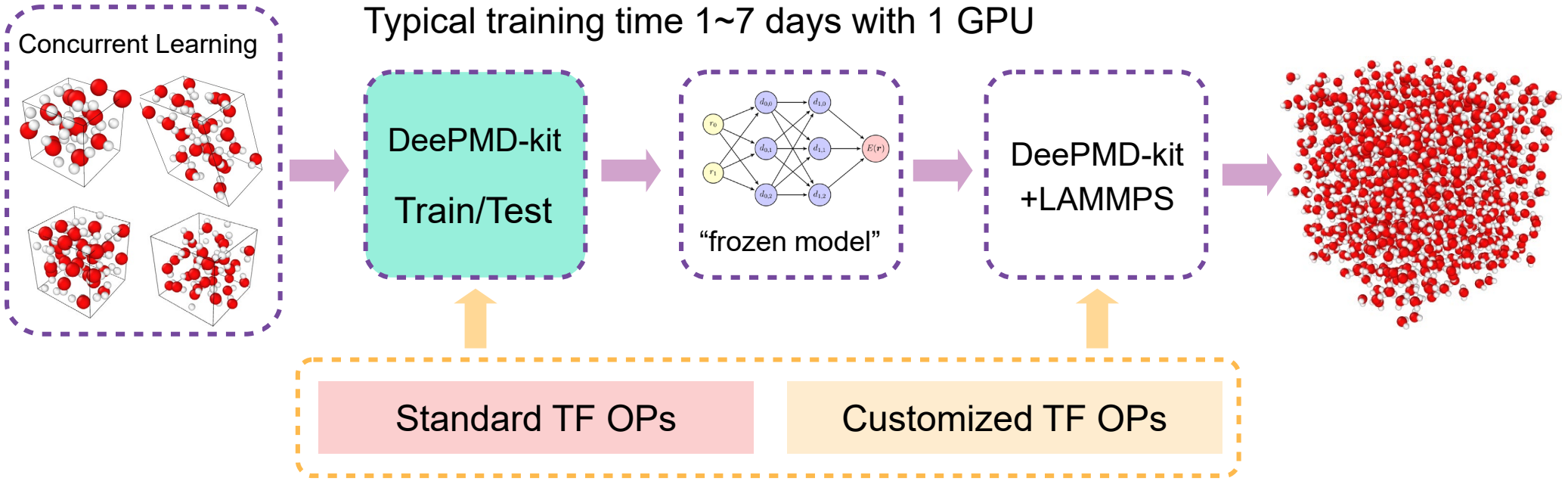
Method: Deep Potential Molecular Dynamics



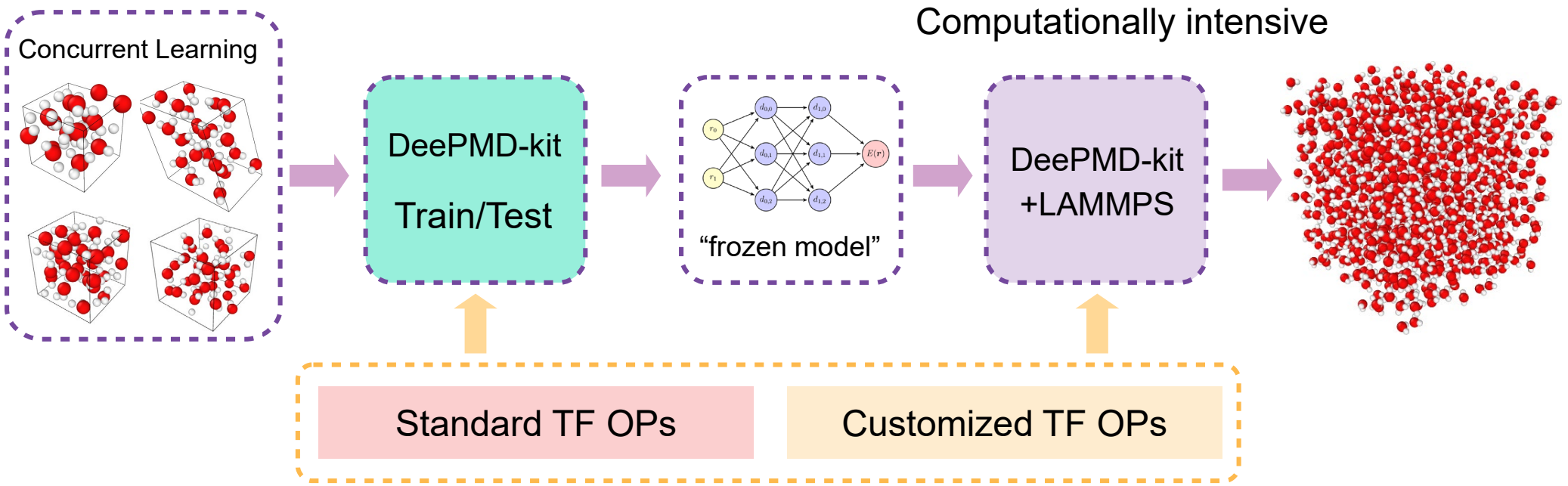
Method: Deep Potential Molecular Dynamics



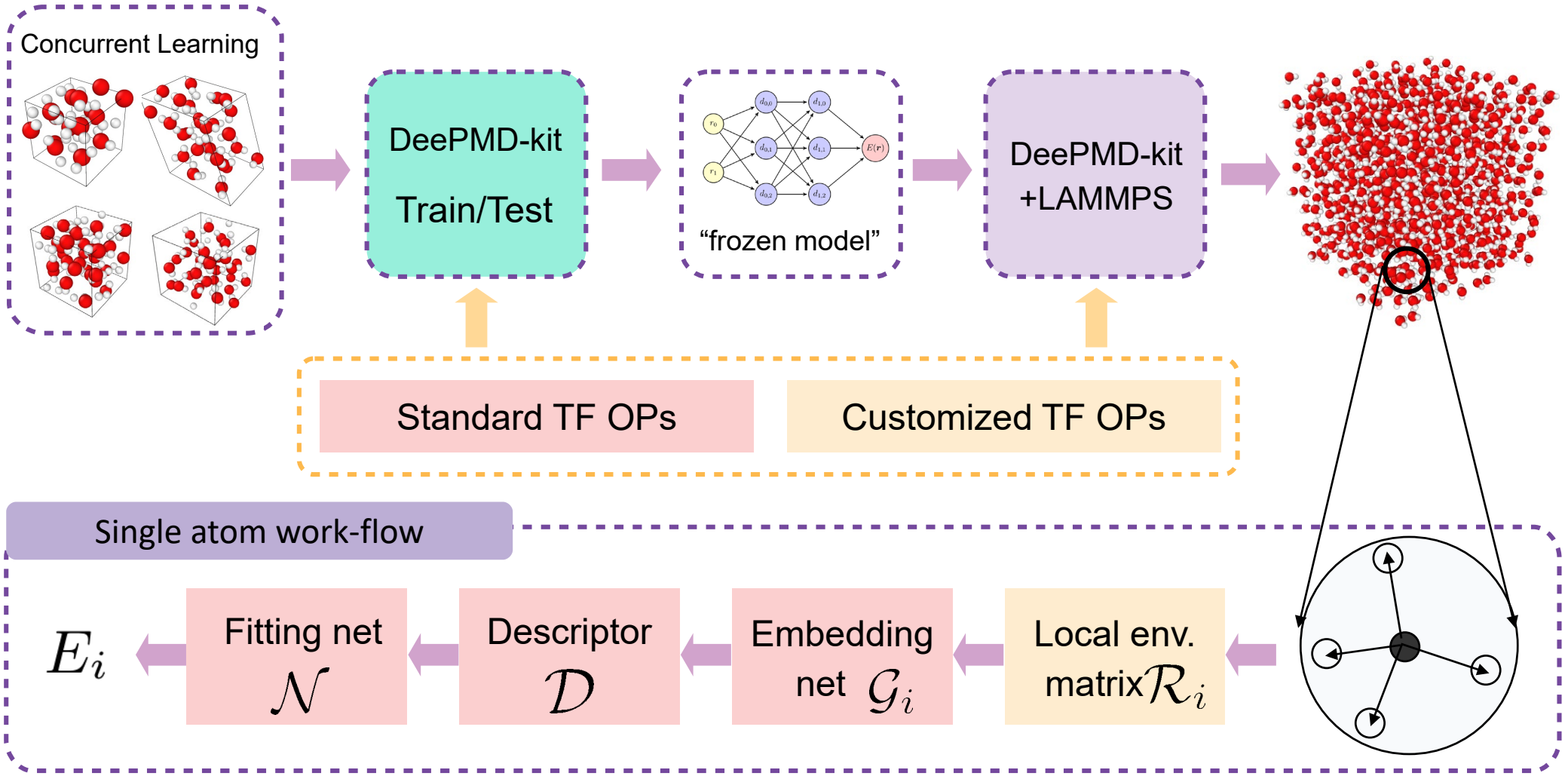
Method: Deep Potential Molecular Dynamics



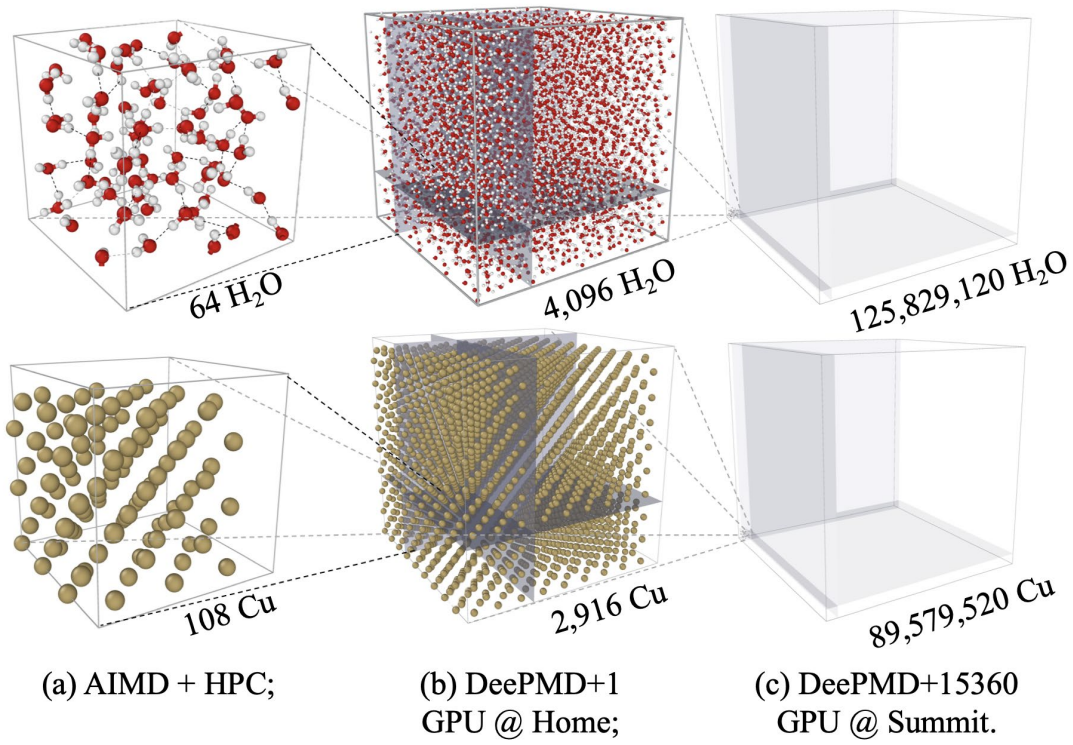
Method: Deep Potential Molecular Dynamics



Method: Deep Potential Molecular Dynamics



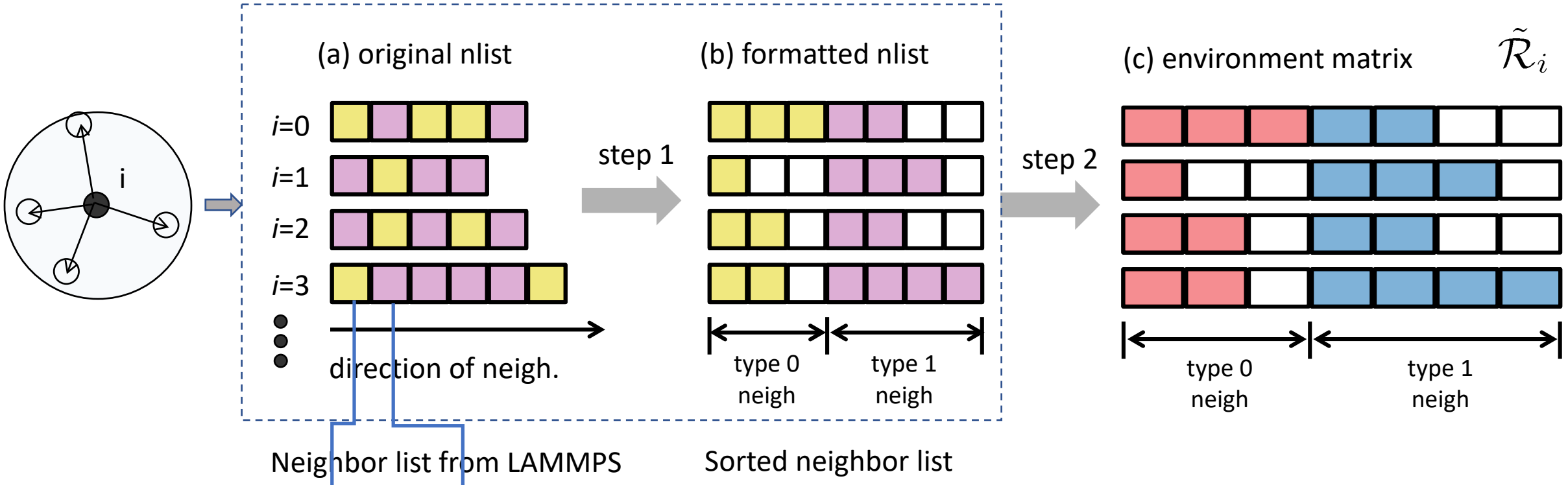
Physical systems



- Single GPU:
  - 8192 H_2O molecule (24576 atoms)
  - 4860 copper atoms
- Scaling:
  - Strong scaling:
    - Copper: 15,925,248 atoms
    - Water: 12,779,520 atoms
  - Weak scaling:
    - Copper: each GPU holds 4656 atoms
    - Water: each GPU holds 24834 atoms

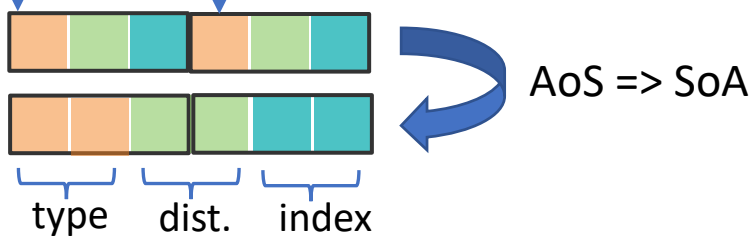
500 MD steps are simulated in the tests.

Customized TensorFlow operators



1 Naïve CUDA kernel

2

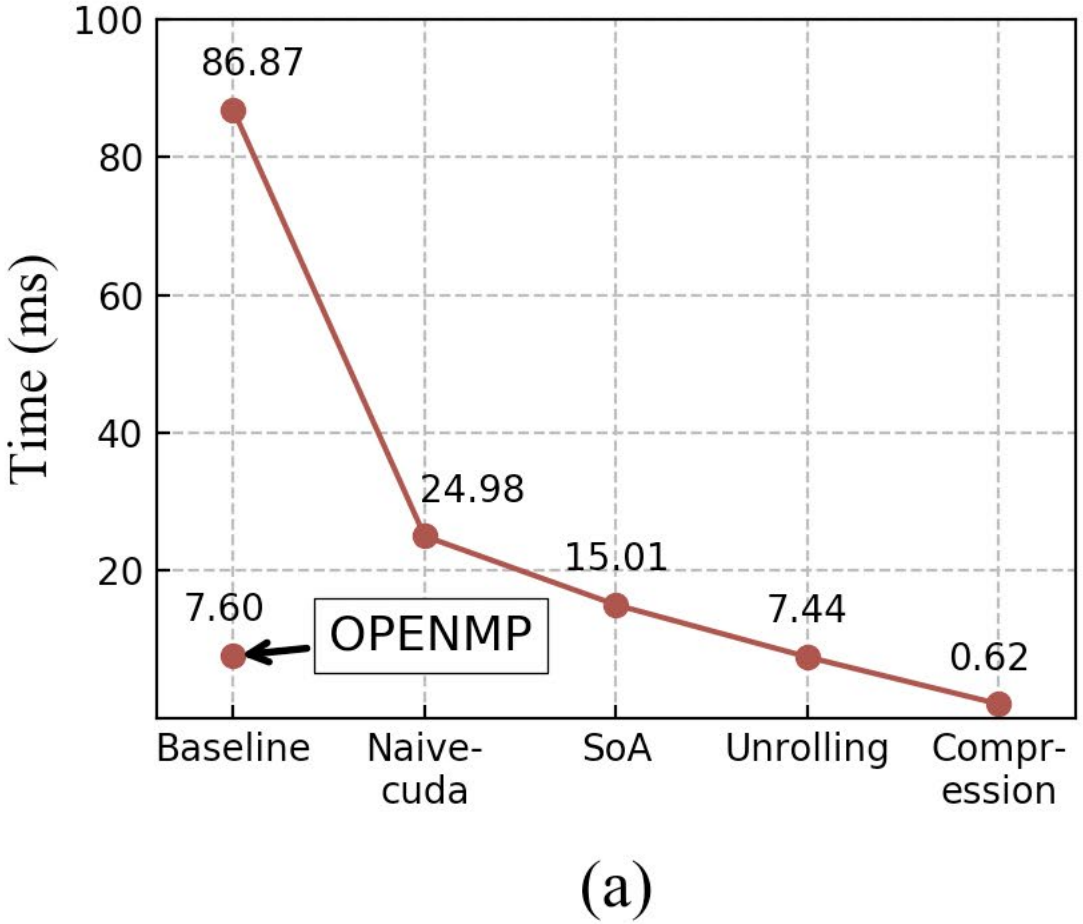


3

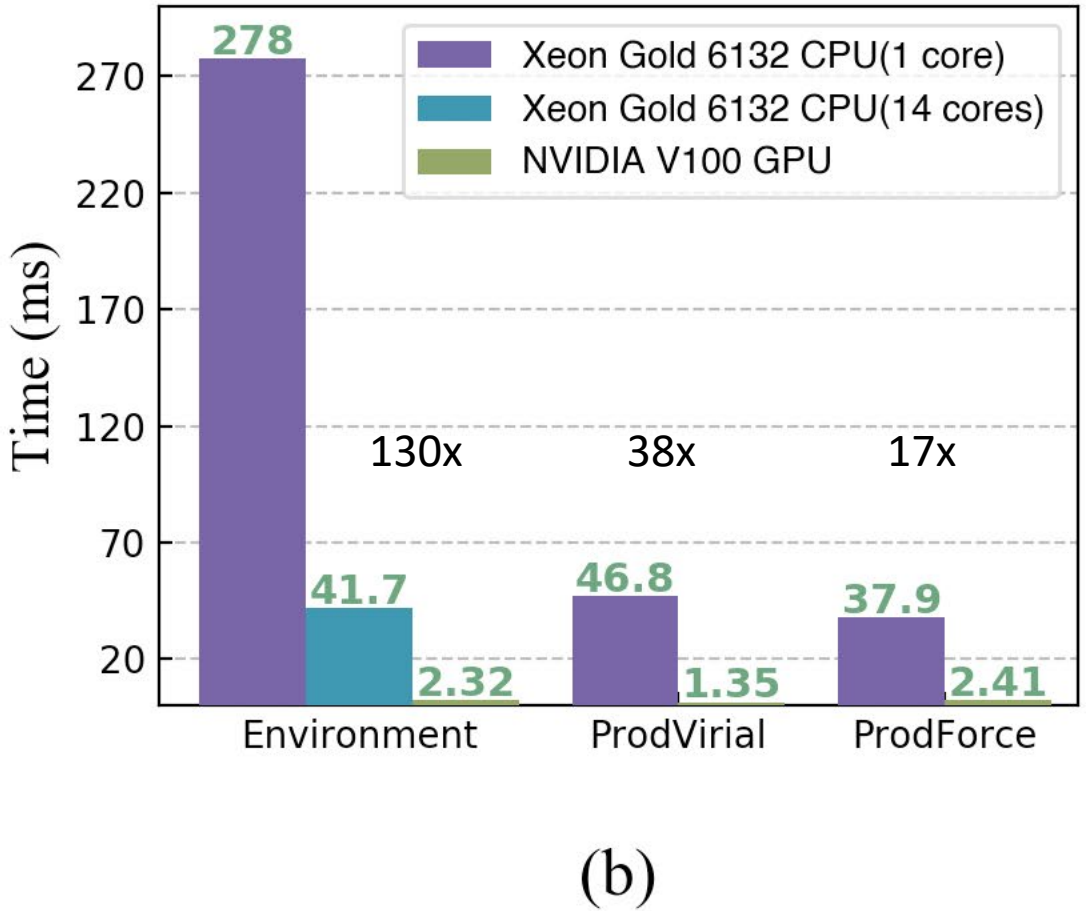
type dist. index

$$\alpha(j) \times 10^{16} + \lfloor |r_{ij}| \times 10^{8} \rfloor \times 10^{6} + j.$$

Customized TensorFlow operators

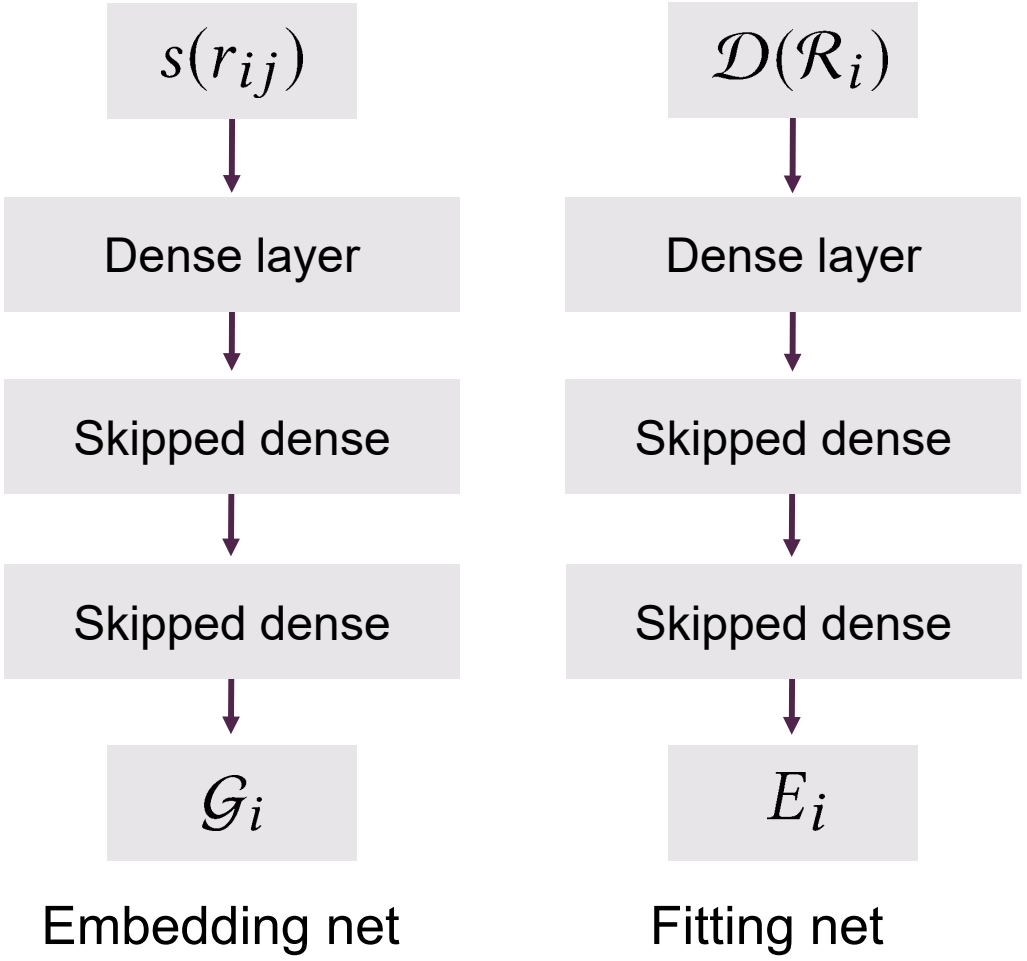


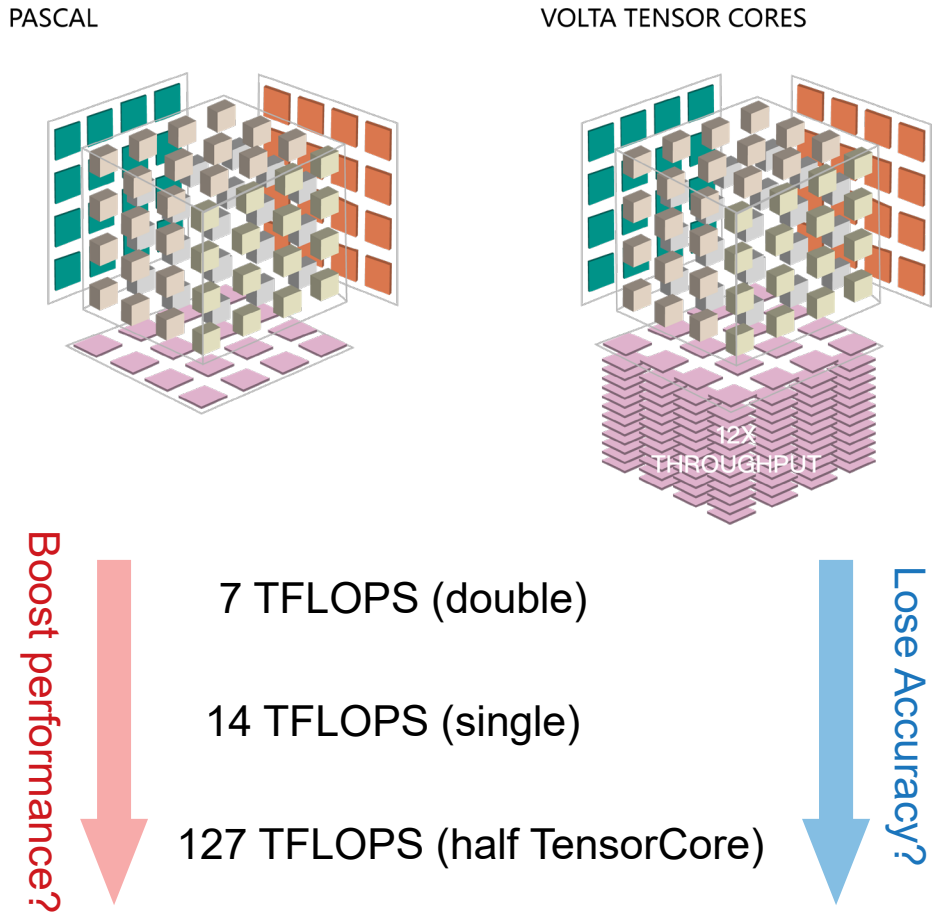
Speedup of formatting neighbor list



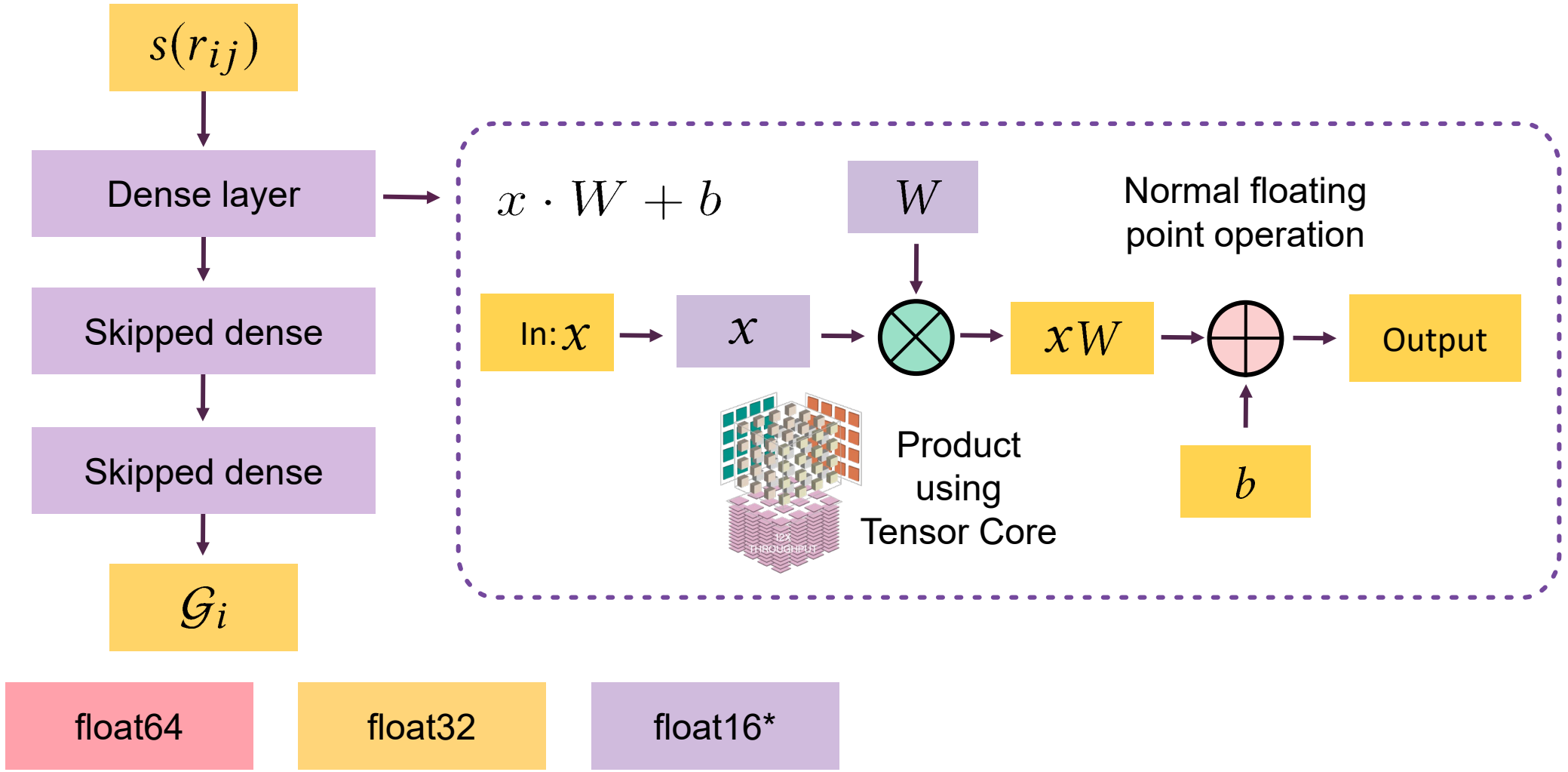
Speedup of all the customized TensorFlow operators

Mixed precision

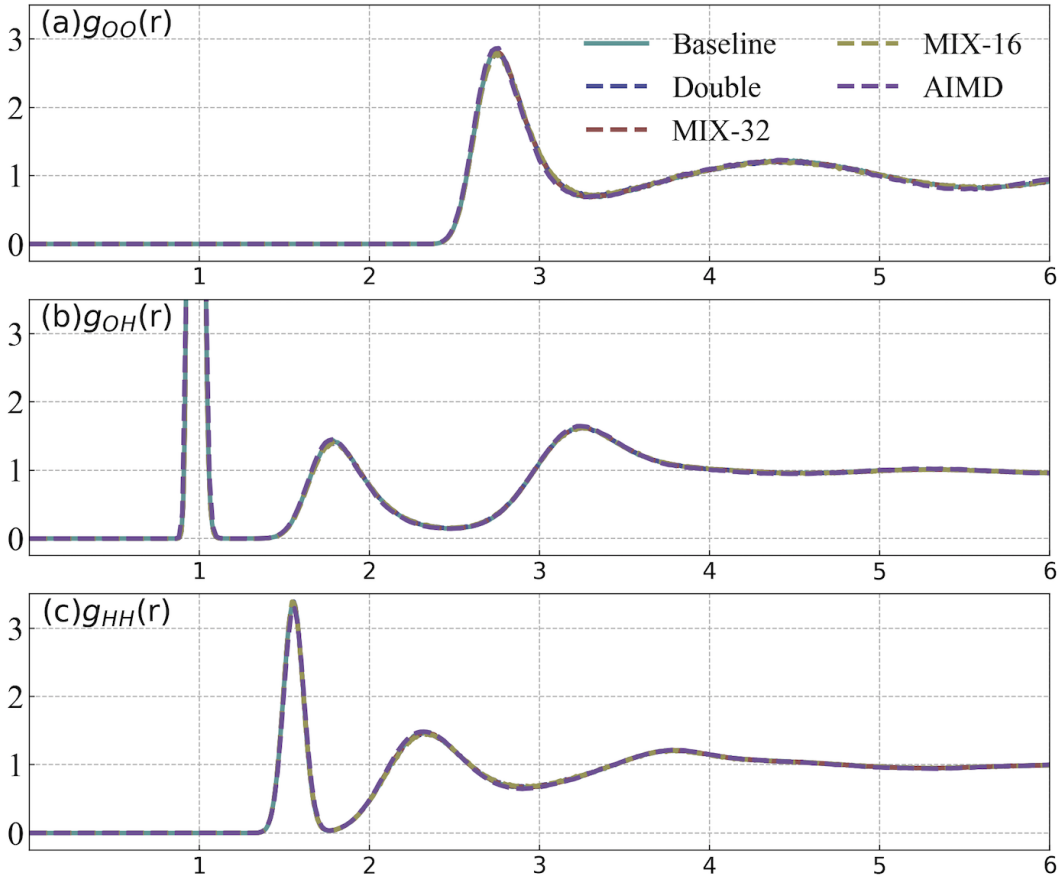




Mixed precision



Mixed precision: accuracy



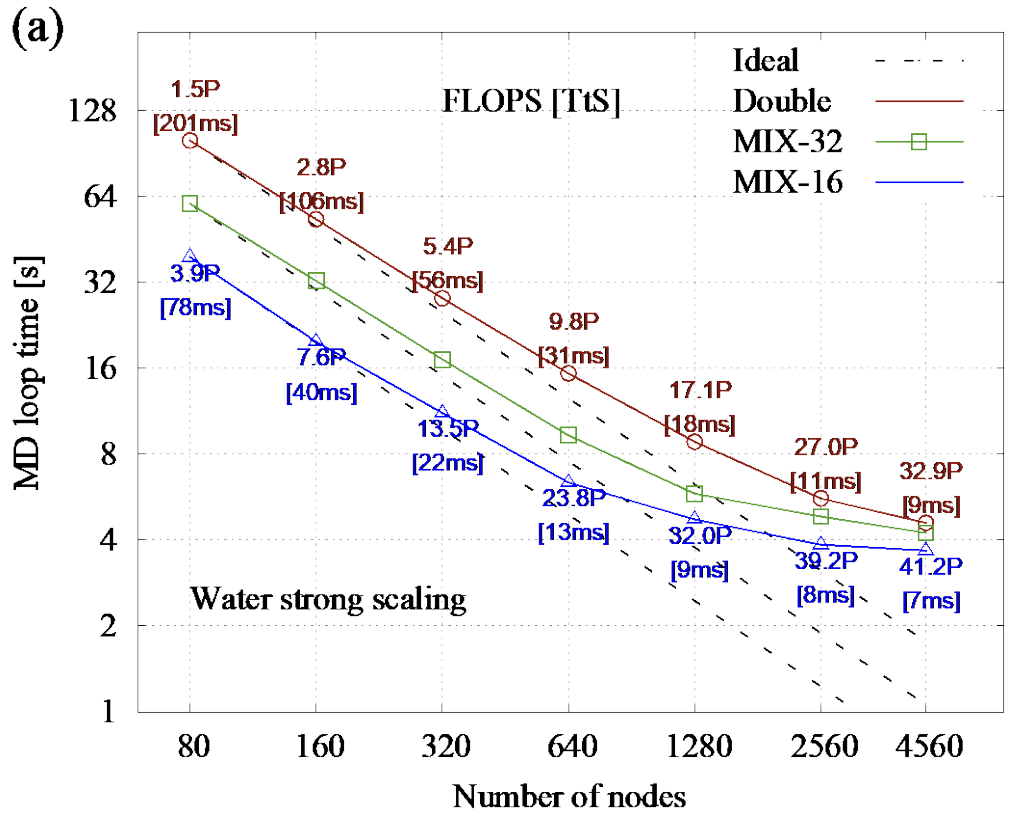
Radial distribution functions $g_{OO}(r)$, $g_{OH}(r)$, and $g_{HH}(r)$ of liquid water at ambient conditions, calculated by AIMD and four DeePMD-kit implementations: baseline, optimized double, MIX-32, and MIX-16

Testing error of 3 different precisions

Precision	Error in energy [eV/molecule]	Error in force [eV/Å]
Double	1.2×10^{-3}	3.7×10^{-2}
MIX-32	1.2×10^{-3}	3.7×10^{-2}
MIX-16	3.6×10^{-3}	3.8×10^{-2}

- Mixed precision can achieve excellent accuracy
- Accuracy of MIX-32 is same as Double

Strong scaling (I)

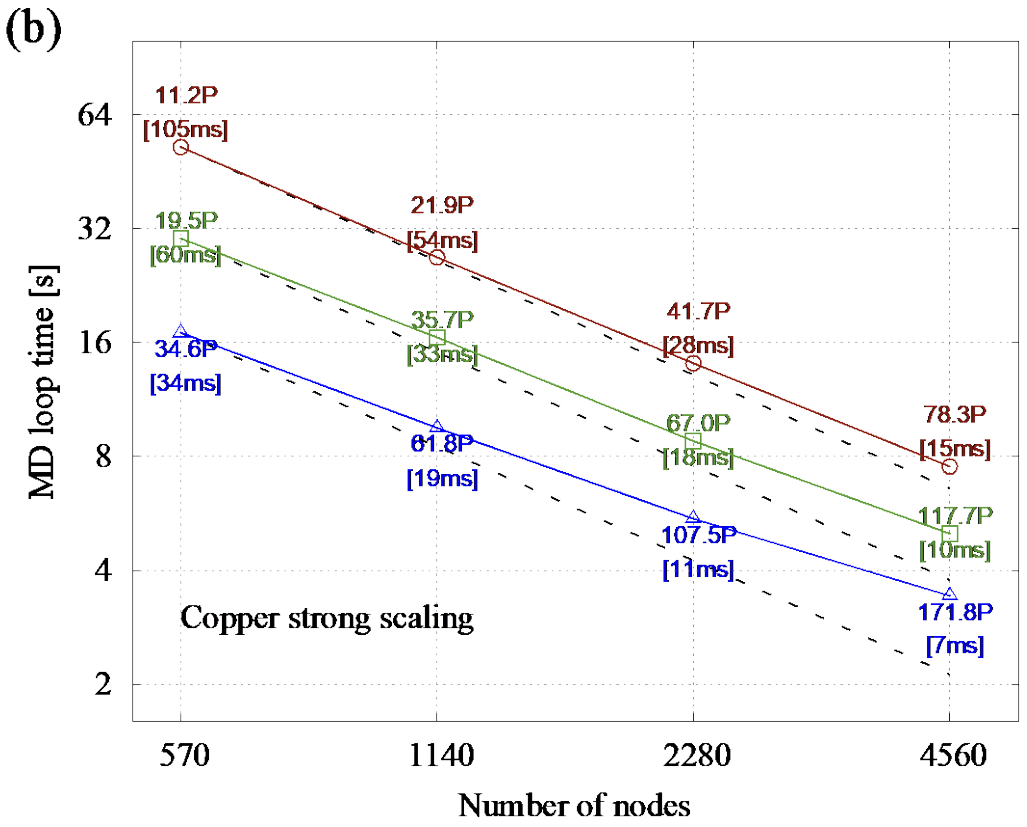


Water: 12,779,520 atoms

#Nodes	80	160	320	640	1280	2560	4560
#GPUs	480	960	1920	3840	7680	15360	27360
#atoms	26624	13312	6656	3328	1664	832	467
#ghosts	25275	17014	11408	7839	5553	3930	3037
MD time	100.4	53.2	28.1	15.4	8.8	5.6	4.6
Efficiency	1.00	0.94	0.89	0.82	0.71	0.56	0.38
PFLOPS	1.51	2.84	5.37	9.84	17.09	26.98	32.90
%of Peak	42.90	40.45	38.26	35.07	30.44	24.03	16.45

Average number of atoms (per GPU), average ghost region size (per GPU), and double precision FLOPS for the 12,779,520 atoms water system.

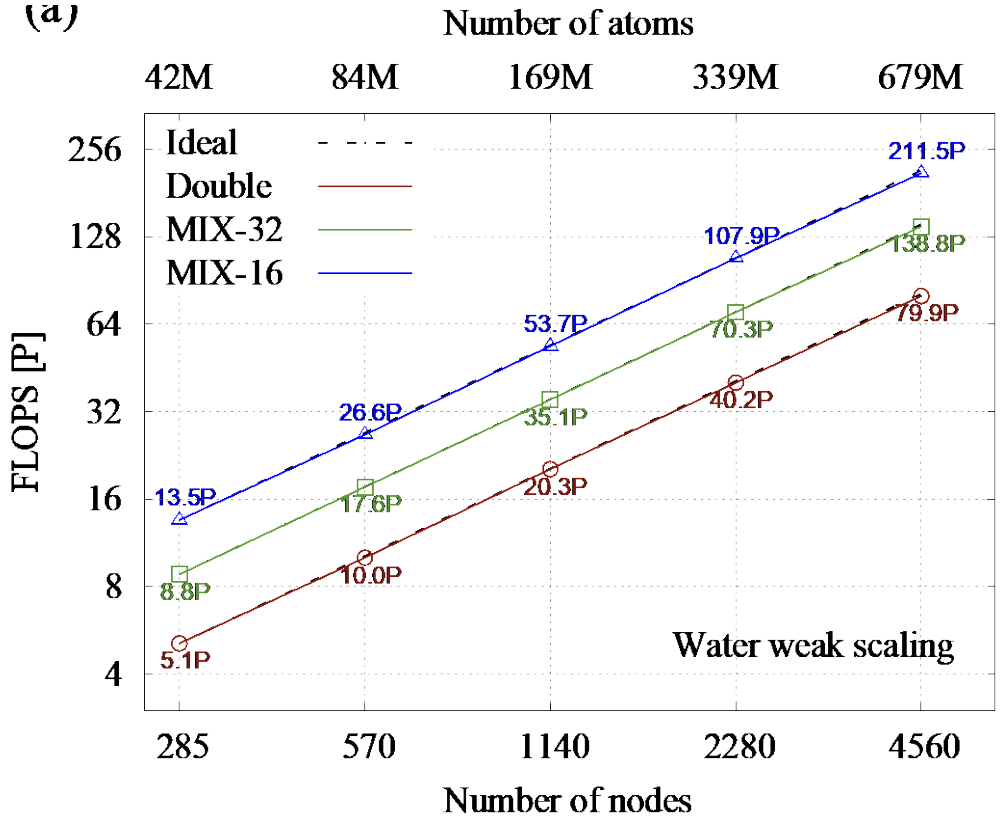
Strong scaling (II)



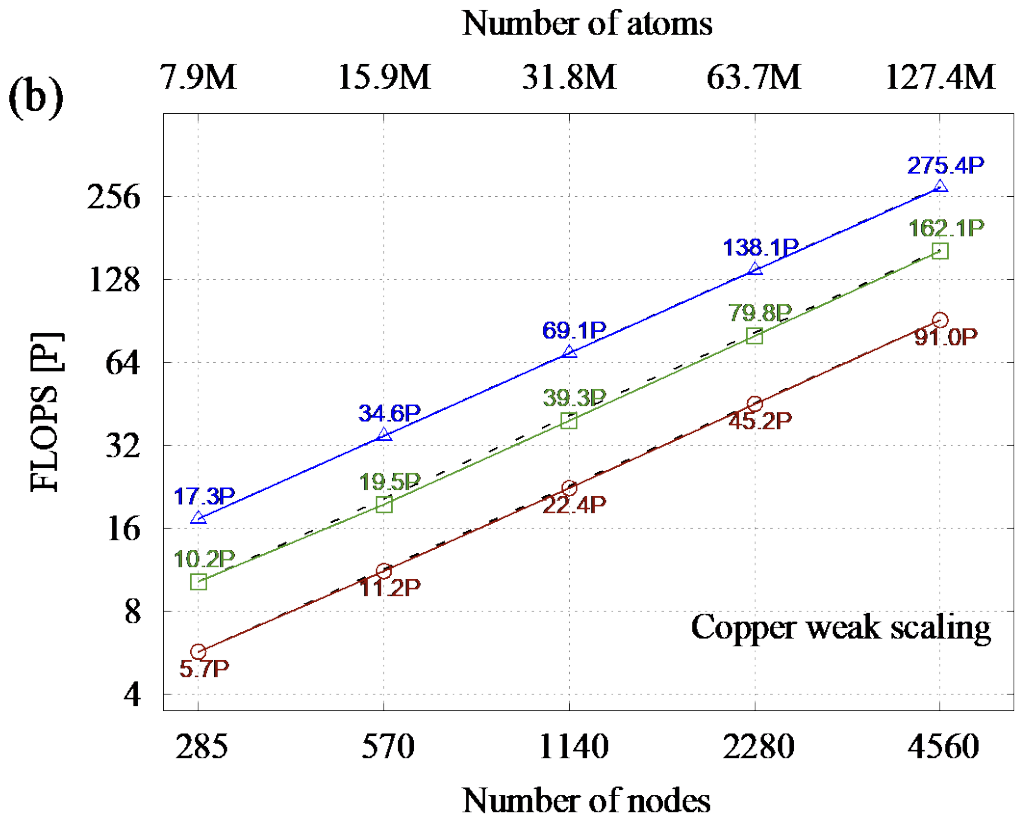
Copper system: 15,925,248 atoms

- Peak performance:
  - 78.3/117/171 PFLOPS for double/MIX-32/MIX-16
  - Parallel efficiency: 87%/72%/62% using 4560 nodes compared to 570 nodes.
  - Double precision scales better because of memory usage
  - MIX-16 is 3x faster on 570 nodes, and 2.2x faster on 4560 nodes.

Weak Scaling: Water and Copper

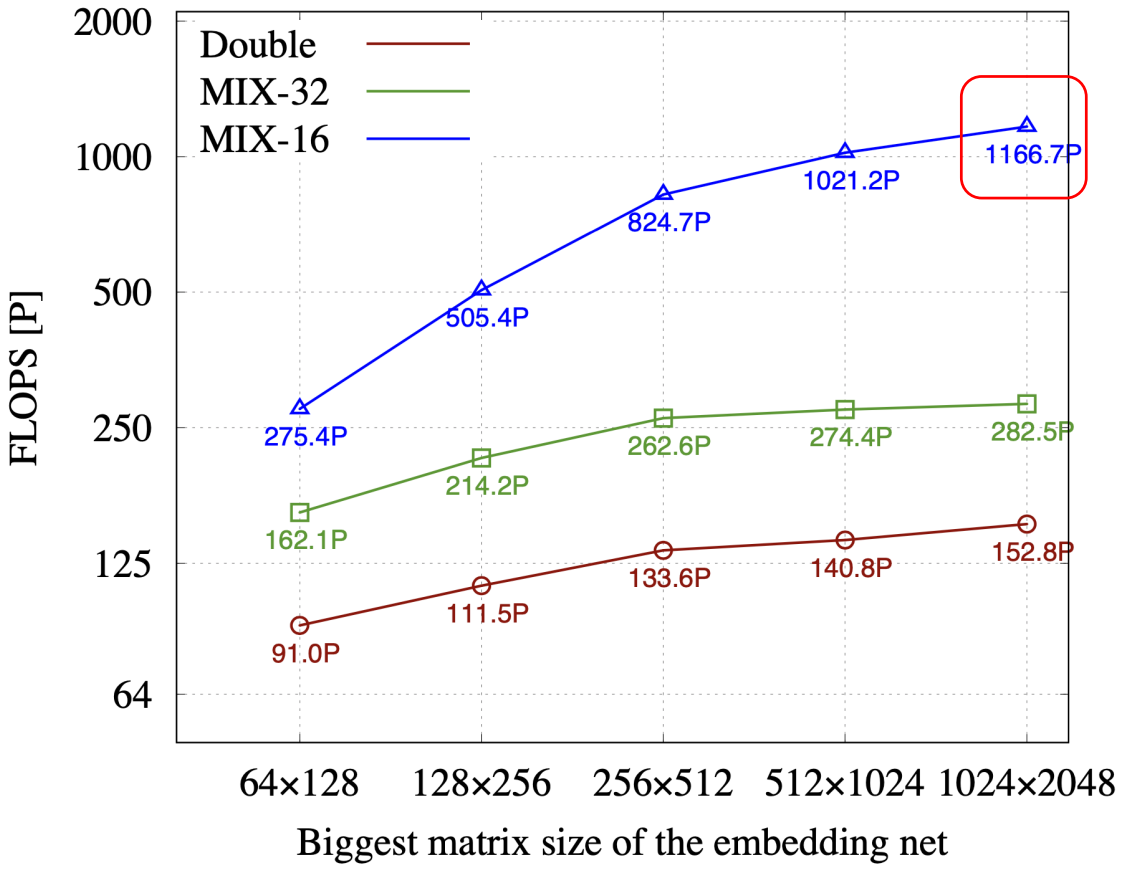


Water: weak scaling from 285 to 4560 nodes
#atoms ranges from 42M to 679M



Copper: weak scaling from 285 to 4560 nodes
#atoms ranges from 7.9M to 127.4M

What if using bigger network?



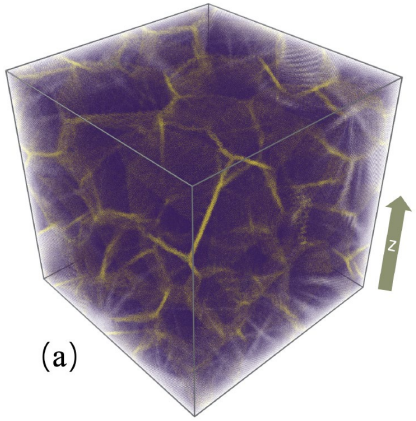
Peak performance of on Summit when using different embedding net size

- Better performance comes with bigger matrix size.
- DeePMD-kit can reach **1.1EFLOPS** with 1024x2048 matrix
- 32x64x128 is enough in terms of accuracy.

MgAlCu System, trained 800w steps				
embedding-net	fitting-net	number of systems	Energy L2err/Natoms	Force L2err
8-16-32	240-240-240	7430	5.909221e-03 eV	5.301006e-02 eV/A
16-32-64	240-240-240	7430	5.393266e-03 eV	5.162859e-02 eV/A
24-48-96	240-240-240	7430	5.272818e-03 eV	5.068834e-02 eV/A
32-64-128	240-240-240	7430	5.071967e-03 eV	5.085757e-02 eV/A

- Computation is bound by hardware FLOP/Byte ratio:
  - V100 GPU, FP-64: 7TFLOPS÷900GB/s =7.8FLOP/Byte
  - V100 GPU, FP-32: 14TFLOPS÷900GB/s =15.5FLOP/Byte
  - V100 GPU, HP-16: 120TFLOPS÷900GB/s =133FLOP/Byte
  - Fujitsu A64FX CPU: 13.51TFLOPS ÷ 1024GB/s= 13.2FLOP/Byte

Application: nanocrystalline copper

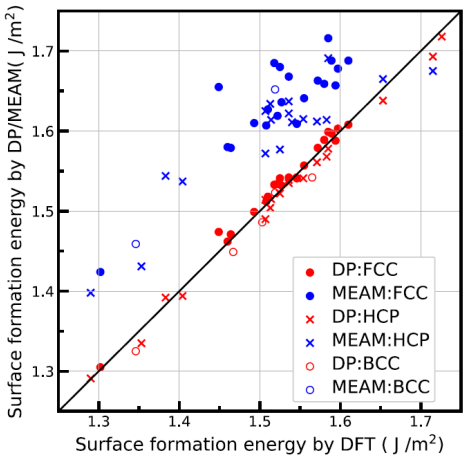


- The strength and hardness of metals can be enhanced by refining their grains into the nanometer scale [1][2]
- MD provides microscopic insights into the underlying mechanism

- A 50 × 50 × 50 nm^3 cube with more than 10 million atoms
- 64 randomly oriented crystals with 15-nm averaged grain diameter
- Purple: copper atoms (face-centered-cubic structure); Yellow: grain boundaries

Recent experimental works on nanocrystallines:
[1] Science **360**, 526-530 (2018). [2] Nature **545**, 80 (2017)

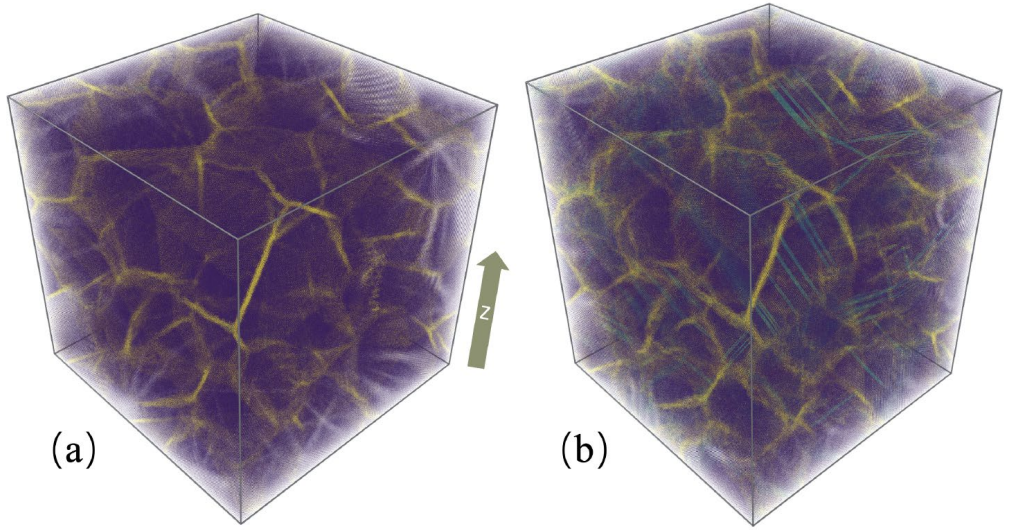
A DP model with DFT accuracy provides more accurate properties for copper than widely used empirical models (MEAM)



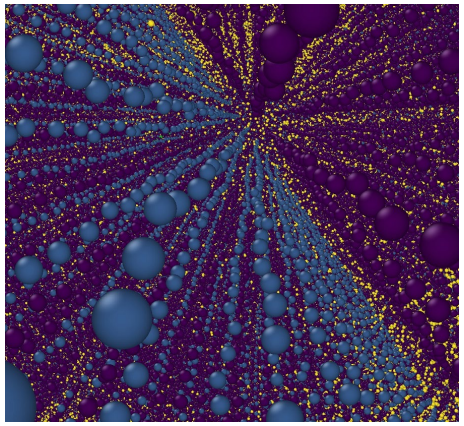
Methods	Stacking fault energy (mJ/m^2)
EXP	41
DFT	38.08
DP	36(2)
MEAM	72.7

Comp. Phys. Comm. 253, 107206 (2020)

Application: nanocrystalline copper



Purple: copper atoms; Yellow: grain boundaries; Cyan: dislocations

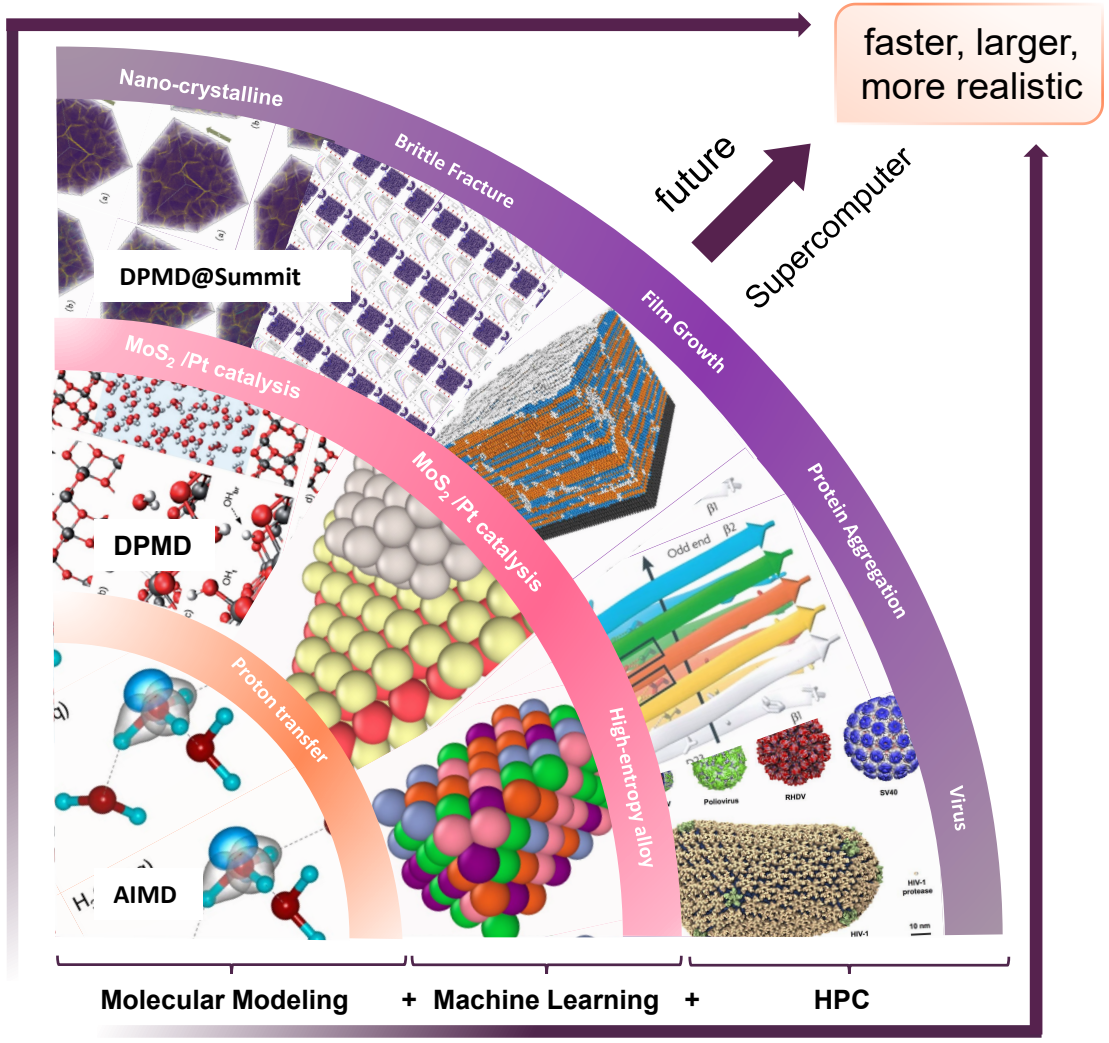


- DPMD simulates the **elongation** of the nanocrystalline copper along the ***z*** direction (deformation: 10%) to yield the strength of nanocrystalline
- DPMD parameters: 50,000 steps at 300 K with a time-step of 0.5 fs (strain rate of $5\times10^{8}s^{-1}$); NPT ensemble

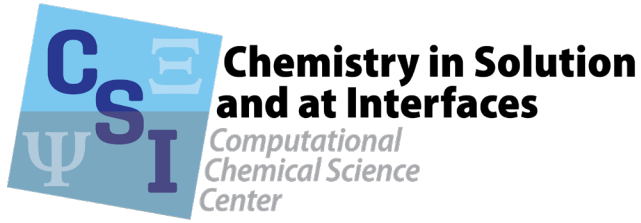
The origins of strength in nanocrystalline is governed by the movements of grain boundaries and dislocations, which can be simulated and analyzed by DPMD.

Conclusion

- **HPC + AI + Physical models: a new paradigm**
  - **>1000x** time-to-solution, **>100x** system size
  - on exa-scale machine: **billions** of atoms
  - Physics-based neural network design
  - AI-specific hardwares in HPC+AI applications
- **Applications**
  - **Materials:** alloy, battery, semiconductor, etc.
  - **Chemistry:** catalysis, combustion, etc.
  - **Biology:** drug design, protein folding, etc.
- **Hardware/Software co-design**
  - New demand from HPC + AI + Physics applications



Acknowledgement



Lin-Wang Wang

Chao Yang

Jiduan Liu

Junqi Yin

Bronson Messer

Thank you for your attention!