Advancing Quantum Simulations with Machine Learning and Graph Theory

Maksim Kulichenko
Theoretical Division, T-1, Los Alamos National Laboratory

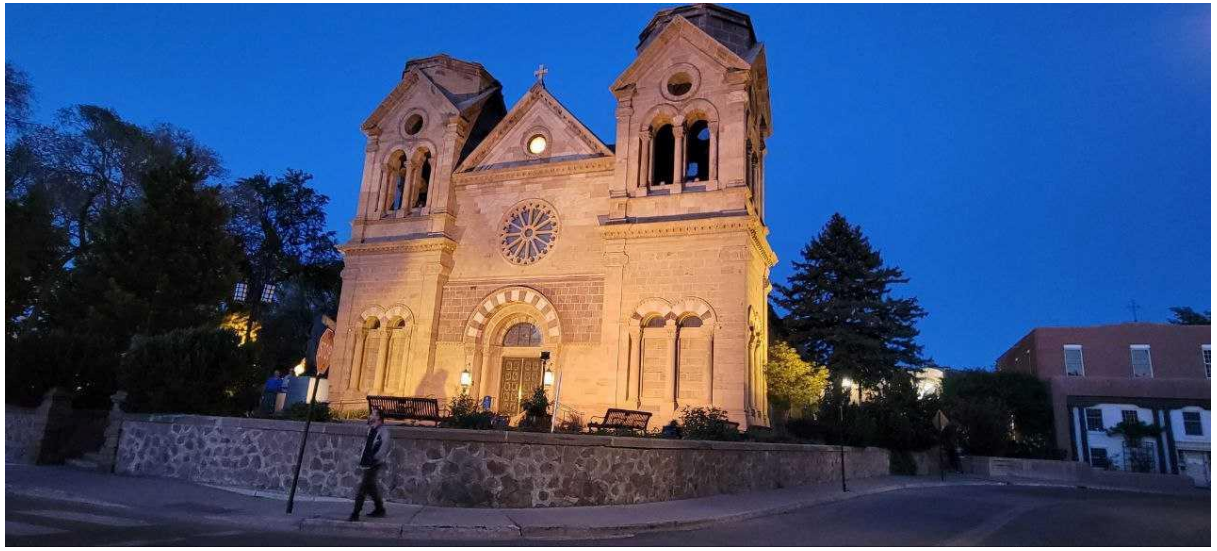
Los Alamos NATIONAL LABORATORY

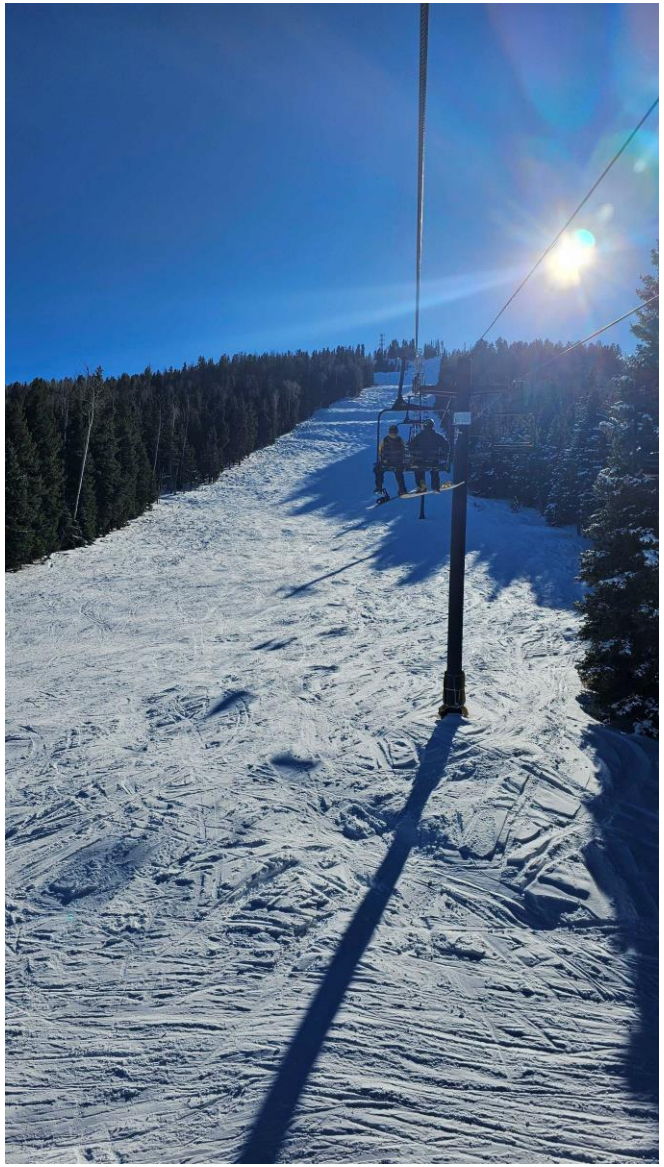
September 4, 2024

LA-UR-24-28802


T-1
K-6

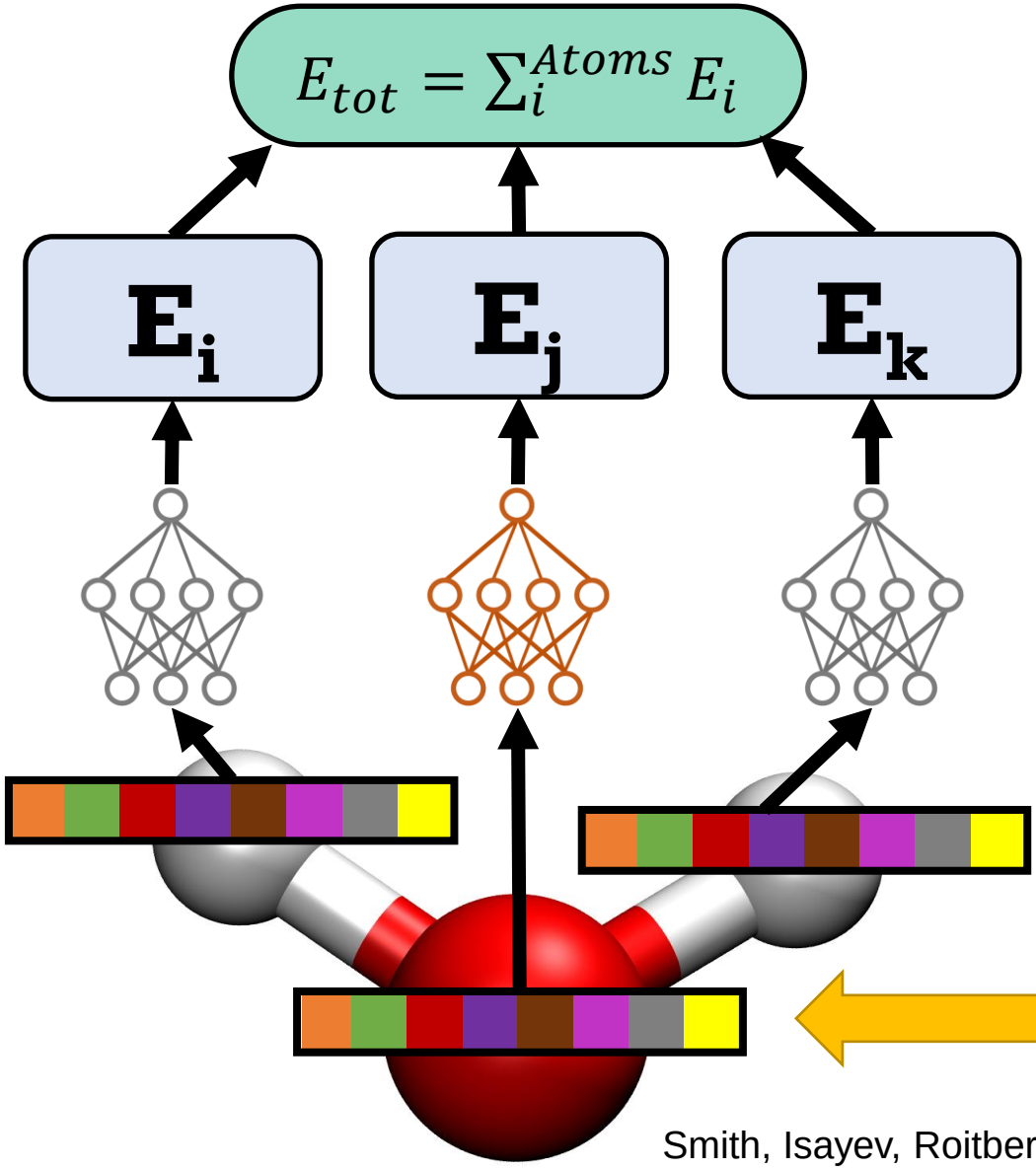
Northern New Mexico Area





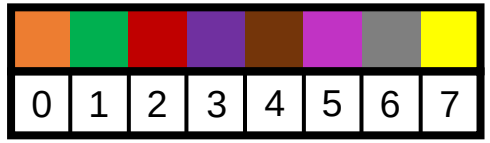
ML to Replace QM

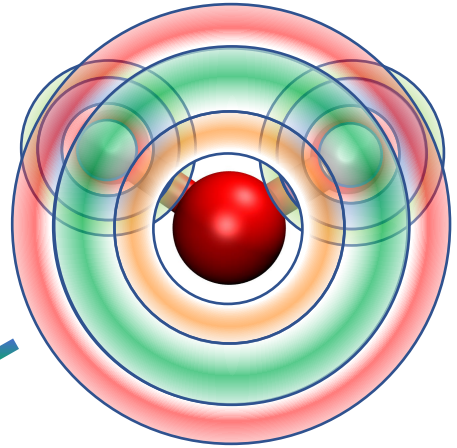
The ANI-type Neural Network potential



- High speed
- High accuracy (when applied within a domain of training data)

- Molecular representation via local atomic environment descriptors
- Energy decomposition into atomic contributions



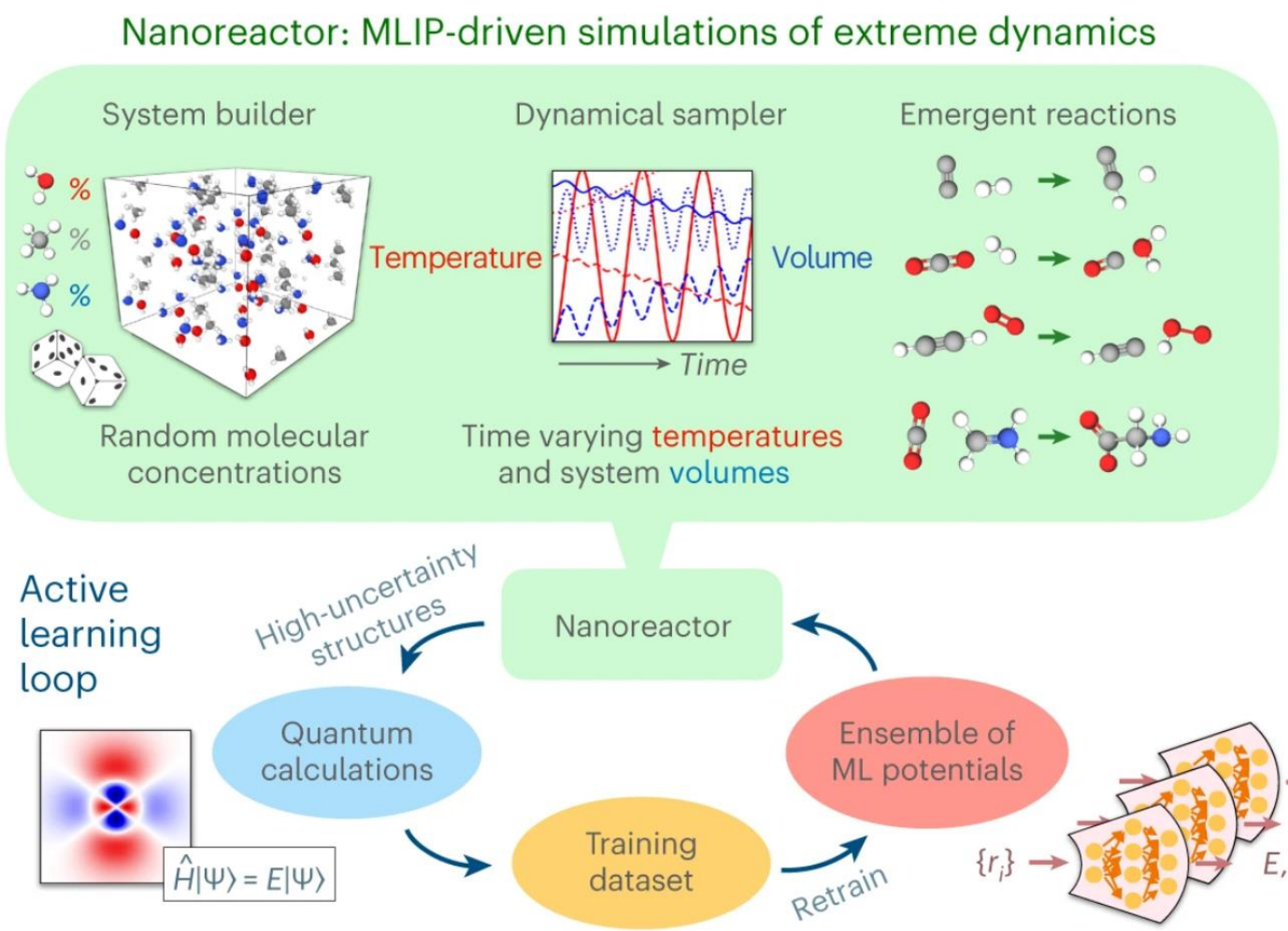


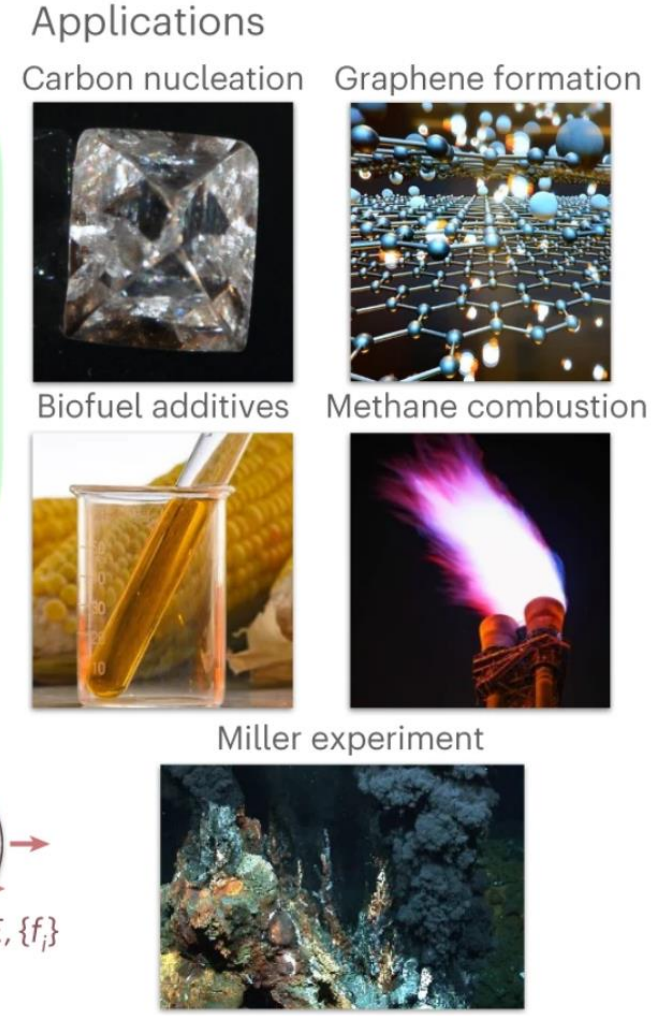
Smith, Isayev, Roitberg, *Chem. Sci.*, 2017, **8**, 3192-3203

ML to Replace QM: Large Scale Simulations

Simple models can simulate complex processes when provided with high-quality data.

ML potential trained to ANI-1xnr dataset:
Active Learning via high-T non-equilibrium condensed phase MD snapshots. 26K datapoints (Boxes of up to 150 atoms)





Zhang, S. *et al*. *Nat. Chem.* **16**, 727–734 (2024)

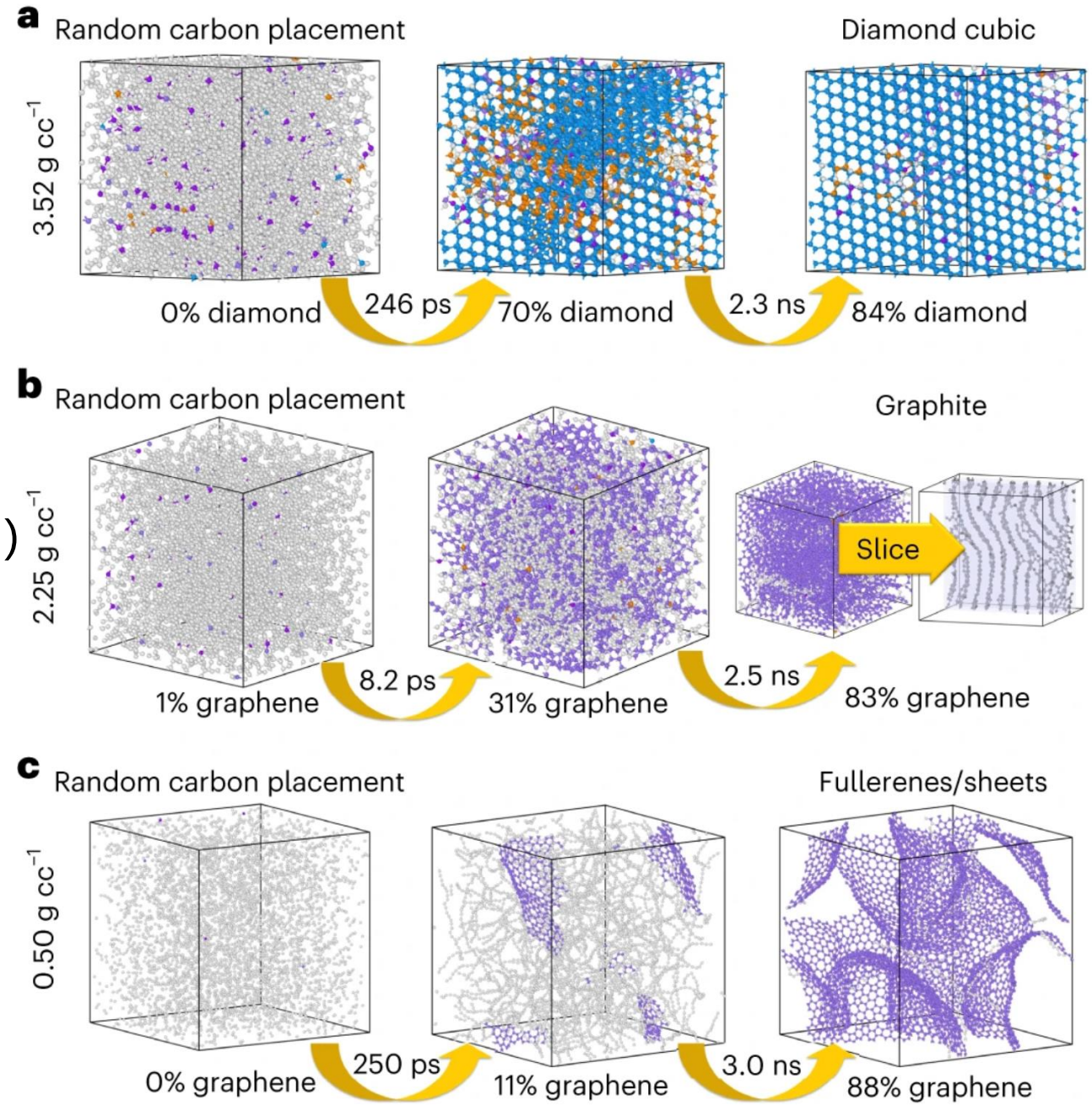
Carbon solid-phase nucleation simulation

- Initial conditions: 5K atoms at different box lengths, 0.5 fs timestep
- ANI-1xnr produces the correct structure of carbon for the respective density

high (3.52 g cc^{-1})

medium (2.25 g cc^{-1})

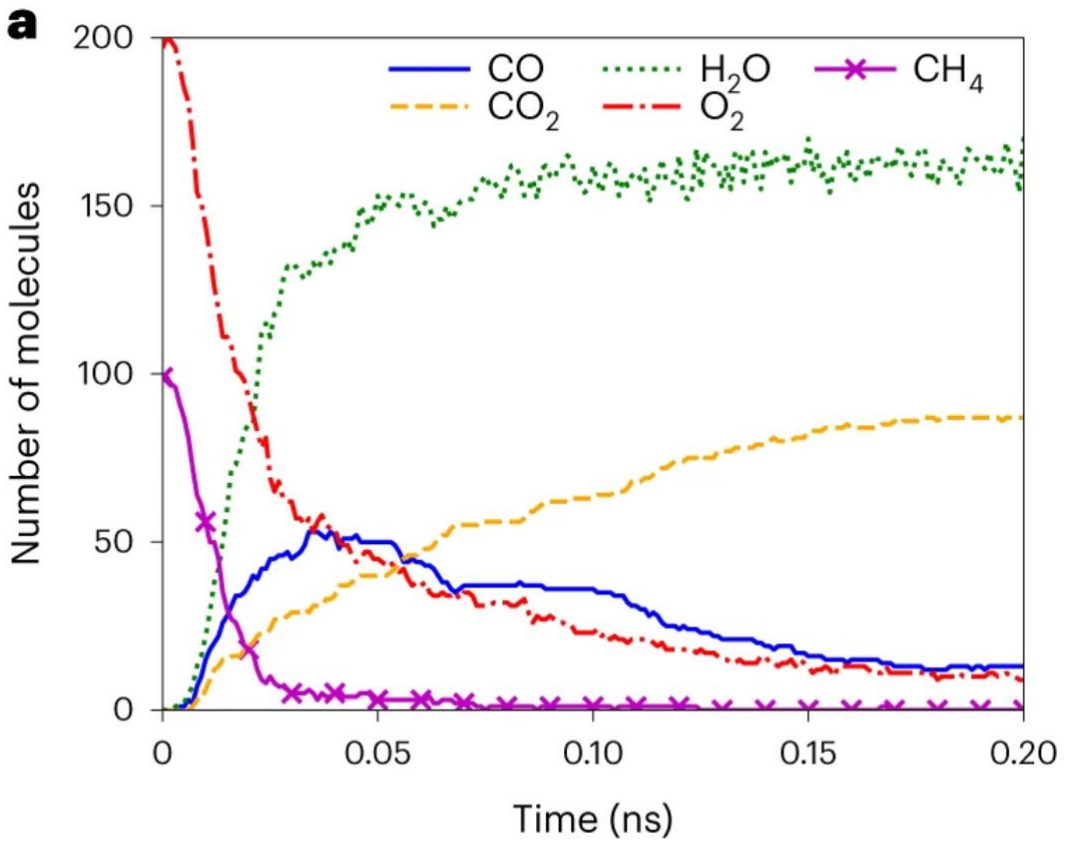
low (0.50 g cc^{-1})

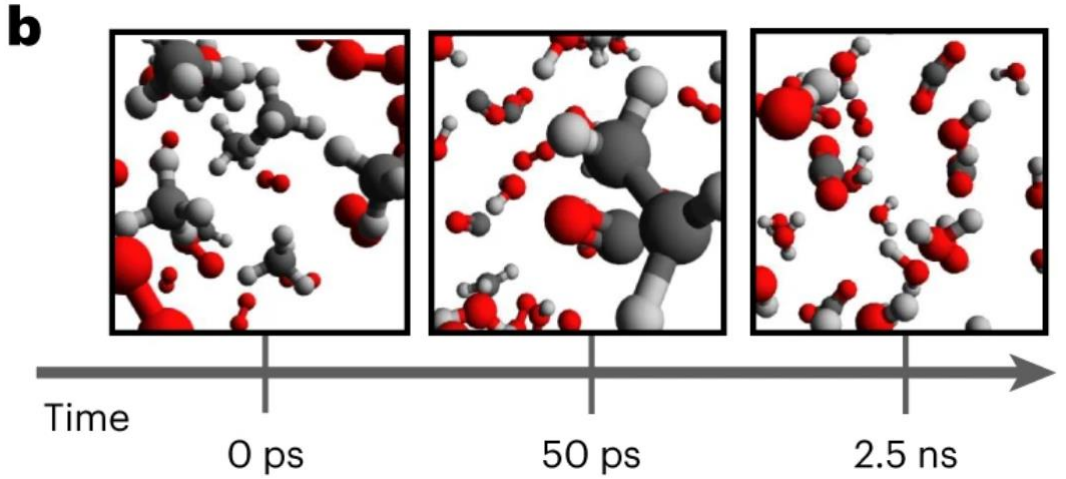


Zhang, S. *et al*. *Nat. Chem.* **16**, 727–734 (2024)

Methane combustion

- Initial conditions: 100 CH_4 and 200 O_2 molecules, 0.1fs timestep
- ANI-1xnr potential correctly produces major products and species profiles
- Reaction rates depend on the reference DFT level of theory

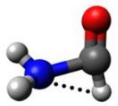




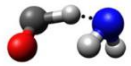
Miller Experiment (Life Formation)

- Initial conditions:16 H_2, 14 H_2O, 14 CO, 14 NH_3 and 14 CH_4
- 0.25 fs timestep
- Glycine is formed!

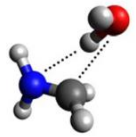
$CO + NH_3 \rightarrow CH_3NO$



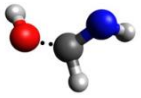
$CO + NH_3 \rightarrow \dot{C}HO + \dot{N}H_2$



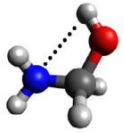
$CH_6\dot{N}O \rightarrow \dot{C}H_4N + H_2O$



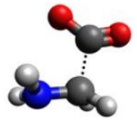
$CH_3NO \rightarrow \dot{C}H_2N + \dot{O}H$



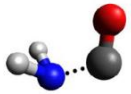
$CH_2O + NH_3 \rightarrow CH_5NO$



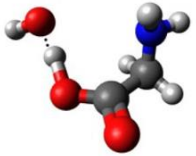
$\dot{C}H_4N + CO_2 \rightarrow C_2H_4NO\dot{O}$

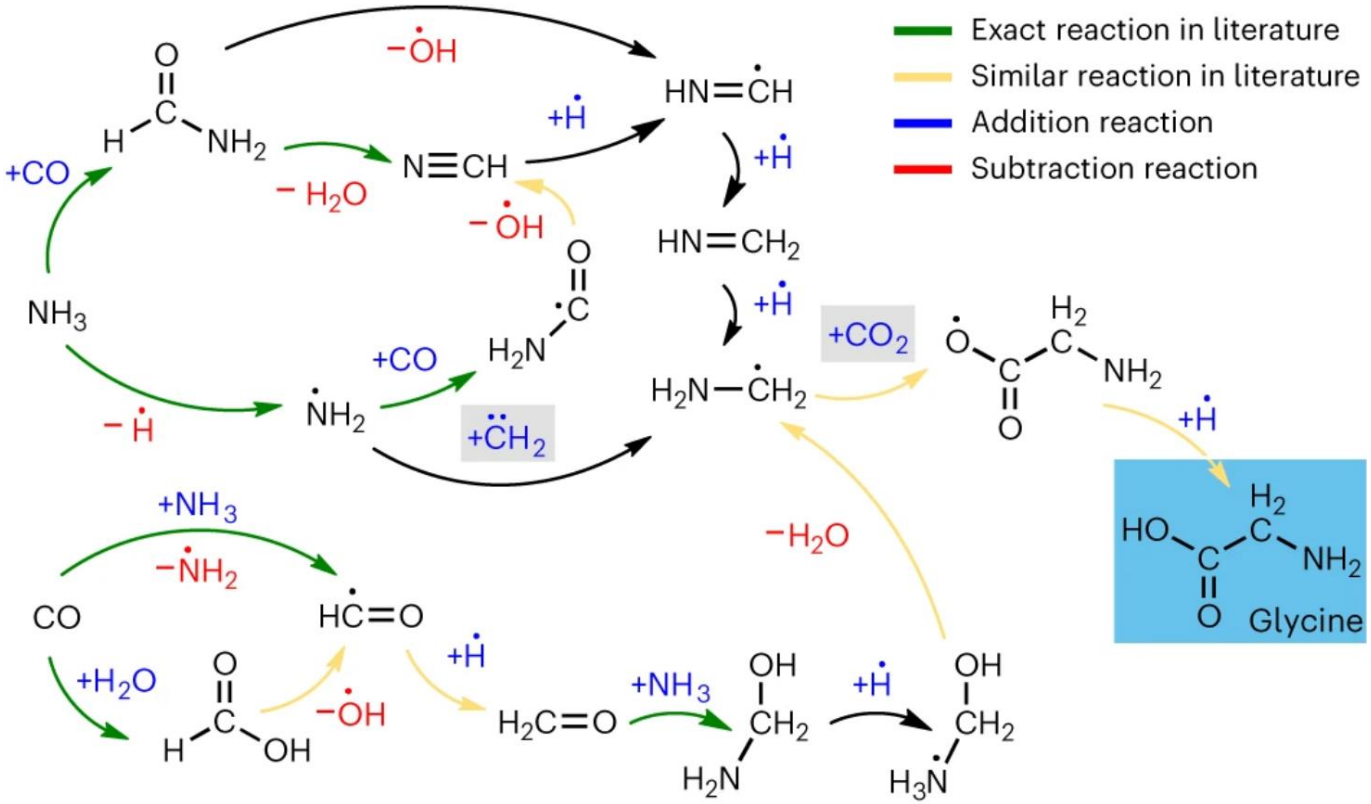


$CO + \dot{N}H_2 \rightarrow \dot{C}H_2NO$



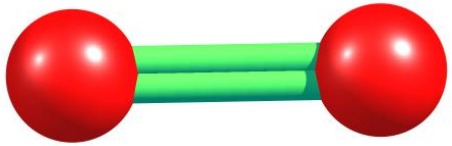
$C_2H_4NO\dot{O} + H_2O \rightarrow C_2H_5NO_2 + \dot{O}H$





ML potentials limitations: no electronic structure information

O_2

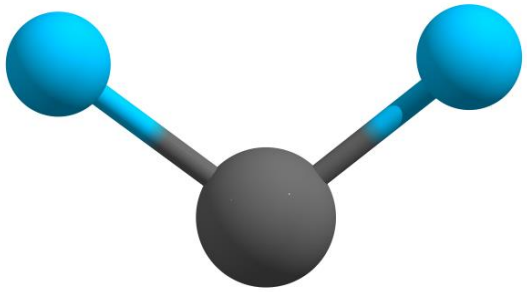


Singlet: 1 eV

↕

Triplet: 0.0 eV

CH_2



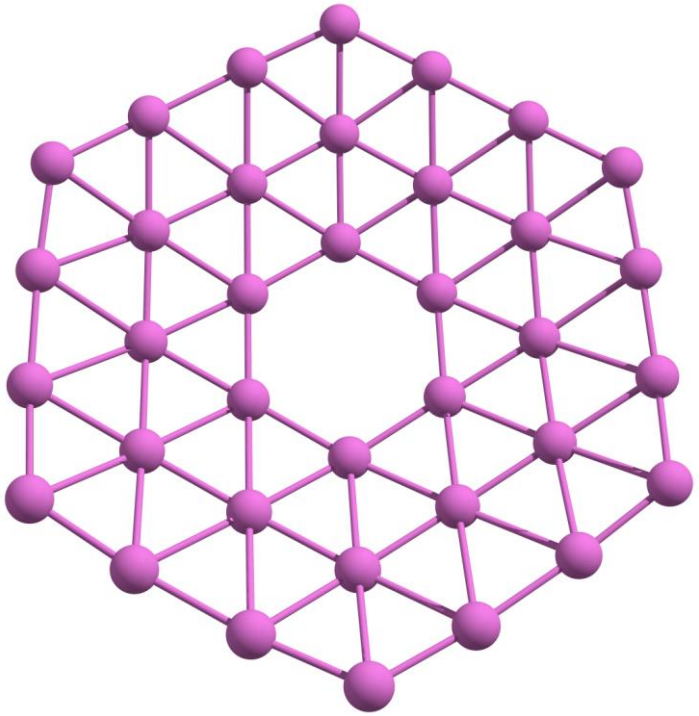
Singlet: 0.5 eV

↕

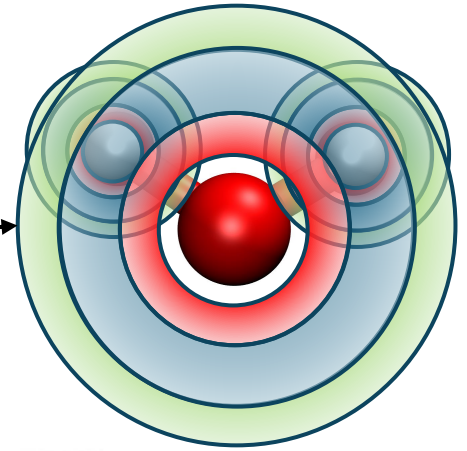
Triplet: 0.0 eV

ML potentials limitations: locality

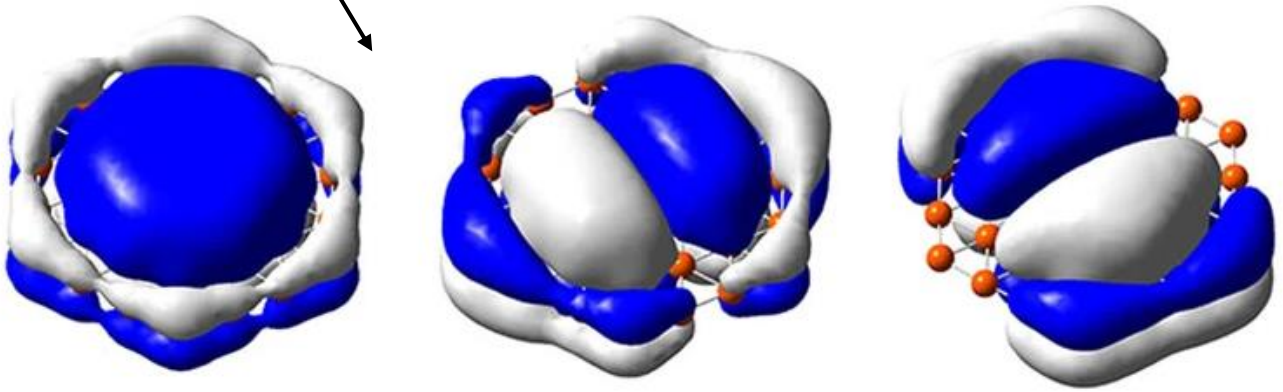
B_{36}



Local descriptors!



Delocalized electrons



- B_{36} has fully delocalized chemical bonds
- Aromatic (like benzene) → planar

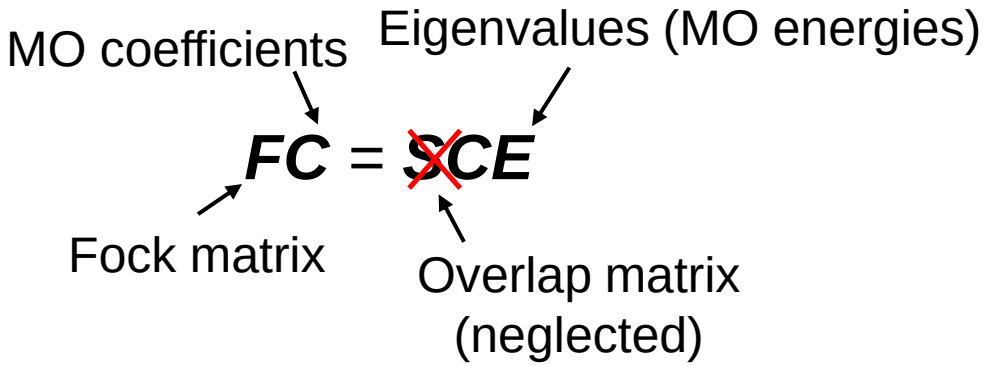
❖ Non-local electronic effects often dictate structural stability

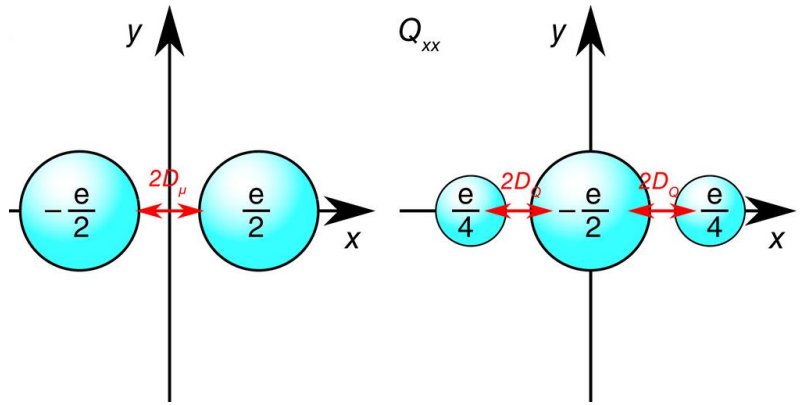
Wang, *et al*. *Nat Commun* 5, 3113 (2014).

ML to Assist QM
Semiempirical Quantum Mechanics (SEQM)

Neglect of Diatomic Differential Overlap (NDDO):

- Based on Hartree-Fock formalism
- Overlap matrix is neglected
- 3c- and 4c-2e integrals are neglected
- 2c-2e integrals are approximated by multipoles
- Valence shell electrons only, minimal basis set
- 1c-1e and 1c-2e integrals are replaced by **static fitted parameters**





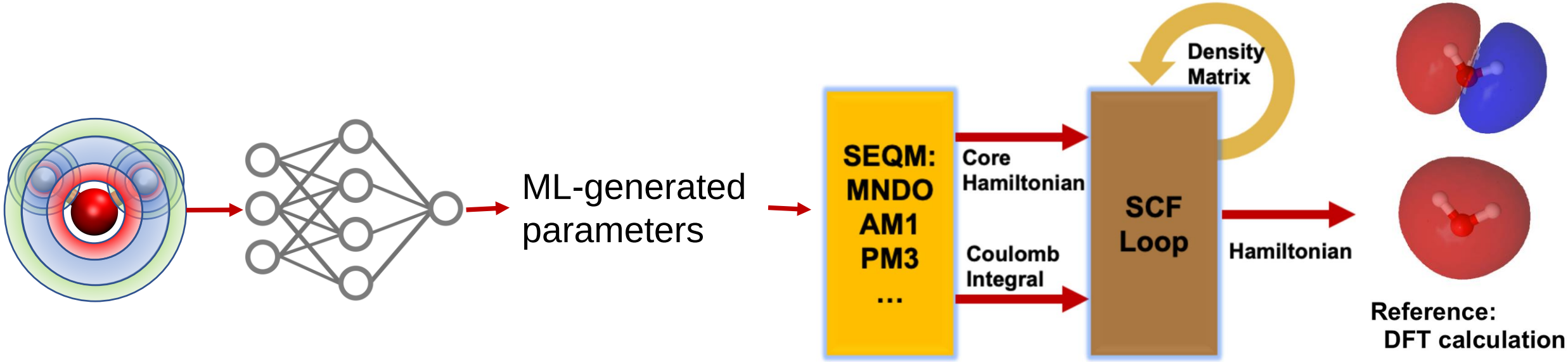
- The accuracy is compromised
- The scaling is reduced from quartic $O(N^4)$ to cubic $O(N^3)$, lower prefactor
- **Electronic structure formalism is retained (charges, multiplicity, non-local effects)**

SEQM Limitation: Accuracy

- Static parameters in Hamiltonian compromise accuracy

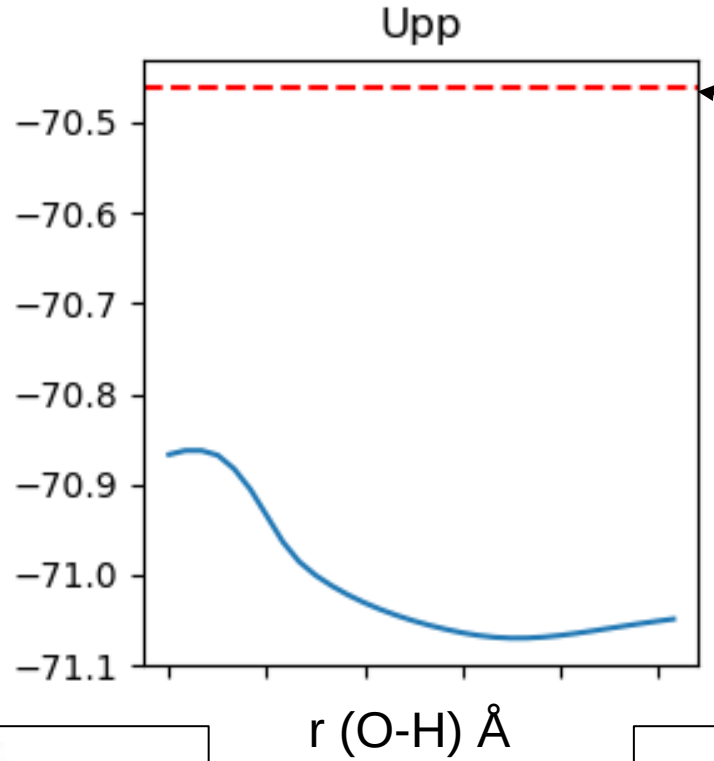
Improving accuracy of SEQM: ML + SEQM

- Replacement of static parameters in semi-empirical Hamiltonian with dynamically responsive
- ML model generates parameter values based on spatial descriptors



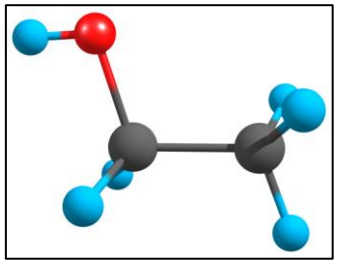
G. Zhou, et. al. *PNAS*, 119, e2120333119 (2022)

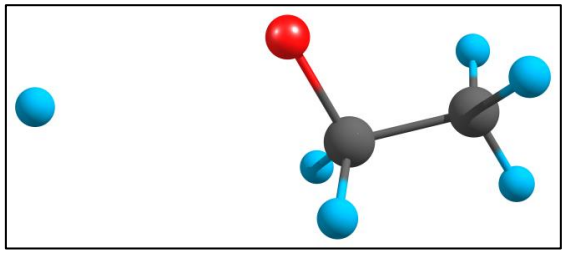
U_{pp}: *p*-AO one-electron one-center integral for O atom



Static parameter

Atomic environment responsive





ML to Assist QM
Semiempirical Quantum Mechanics (SEQM)

PySEQM: PyTorch-Based Semi-Empirical Quantum Mechanics

Capabilities:

- Neglect of Diatomic Differential Overlap models (PM6, PM3, AM1, MNDO)
- Built-in interfaces for ML re-parametrization
- GPU-accelerated simulations
- Multiple parallel simulations via a batch mode
- Forces and errors calculation via PyTorch backpropagation

https://github.com/lanl/PYSEQM.git

G. Zhou, et. al. *PNAS*, 2022, 119, e2120333119
M. Kulichenko et al.J. Chem. Theory Comput. 2023, 19, 11, 3209–3222

ML to Assist QM: Training SEQM to Non-Equilibrium Data Data

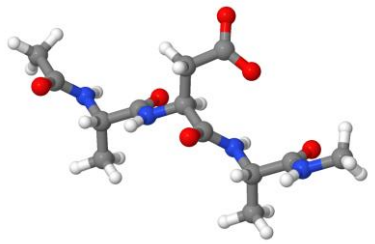
ANI-1x training set

Trained on 600K structures

Trained on 60K structures

RMSE of Atomization energies
Test set: 250 Tripeptides (DFT reference data)

Method	SEQM (PM3)	ML (HIPNN)	ML+SEQM
RMSE [kcal/mol]/atom	0.6	1.1	**0.2**



- ML+SEQM achieves higher accuracy than pure ML and SEQM
- ML+SEQM needs less training data than pure ML
- **In fact, ML+SEQM can't have too much data**
  - **Training is expensive because of error backpropagation through SCF loop.**

N. Lubbers, et. al. *J. Chem. Phys*. 148, 241715 (2018)
G. Zhou, et. al. *PNAS*, 119, e2120333119 (2022)

ML to Assist QM: Training SEQM to Reactive Data

Robert Stanton et al. Unpublished

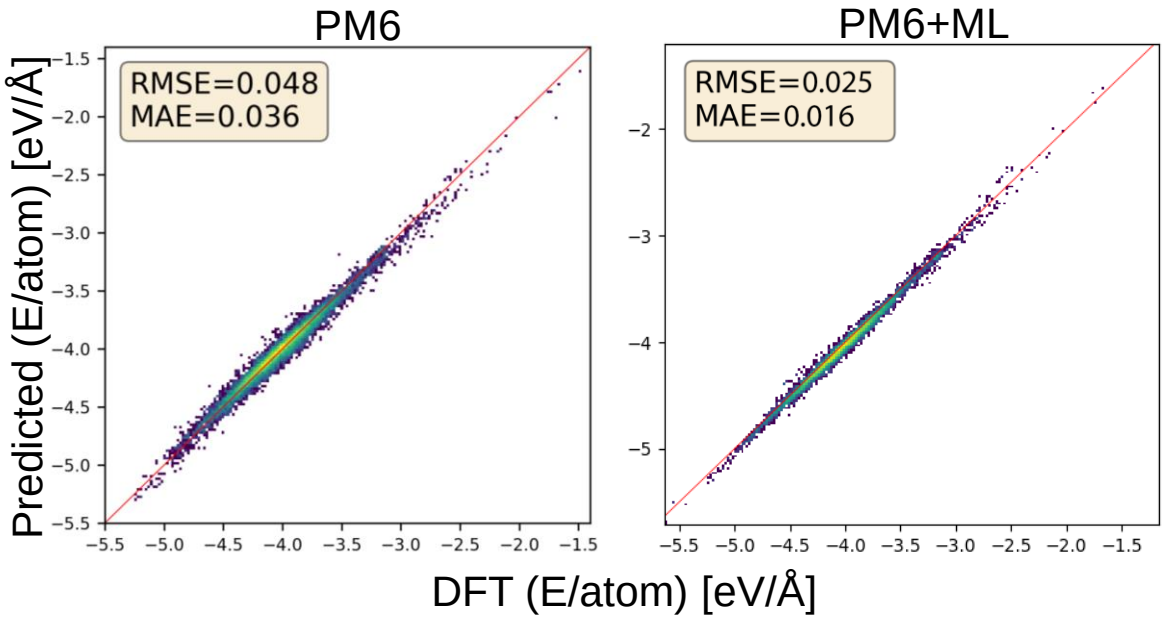
Training data: 140k broken-symmetry open shell singlets and doublets

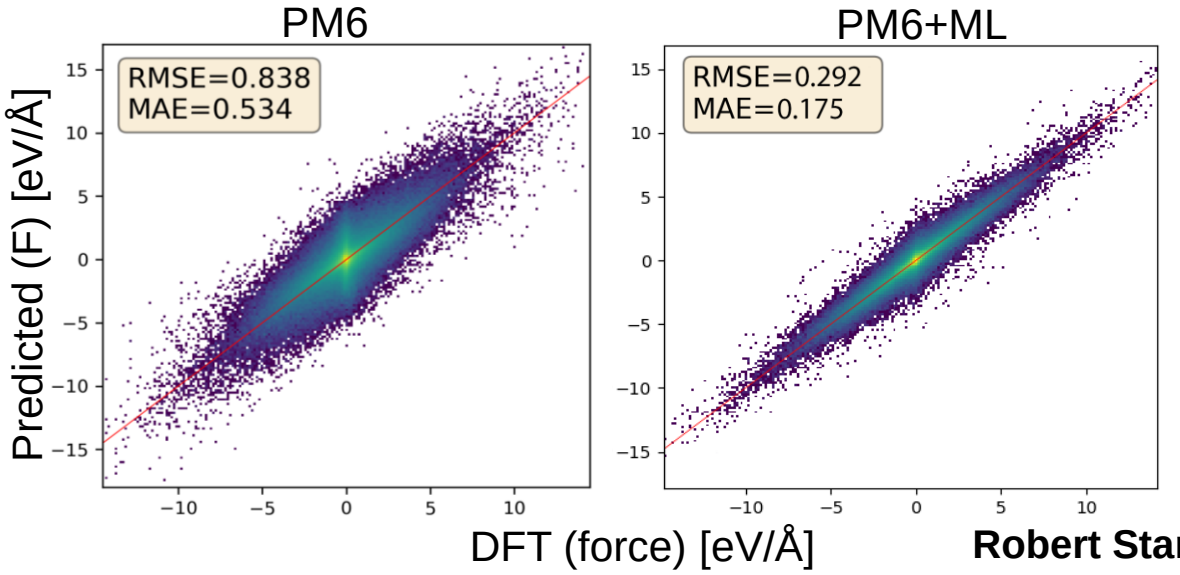
- 60% Transition1x[1] – reaction pathways
- 10% BSE49[2] – bond dissociation paths
- 30% ANI-1x[3] – non-equilibrium geometries

1) M. Schreiner et al., Sci Data 2022, 9, 779
2) V.K. Prasad et al., Sci Data 2021, 8, 300
3) J.S. Smith et al., Sci Data 2020, 7, 134

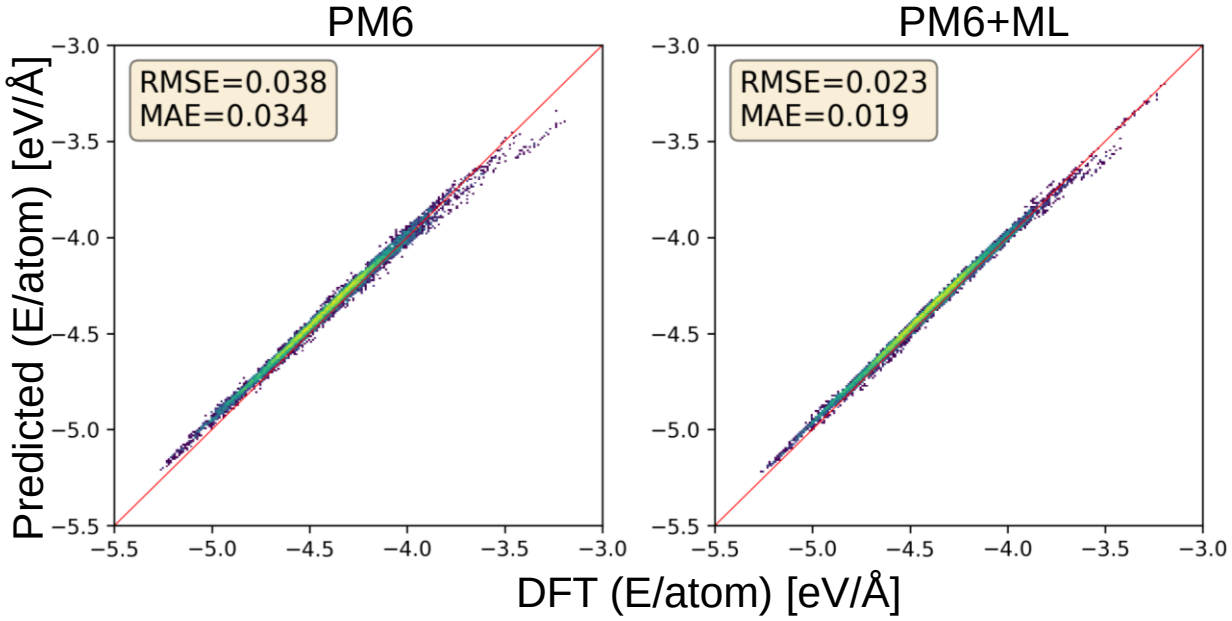
ML to Assist QM: Training SEQM to Reactive Data

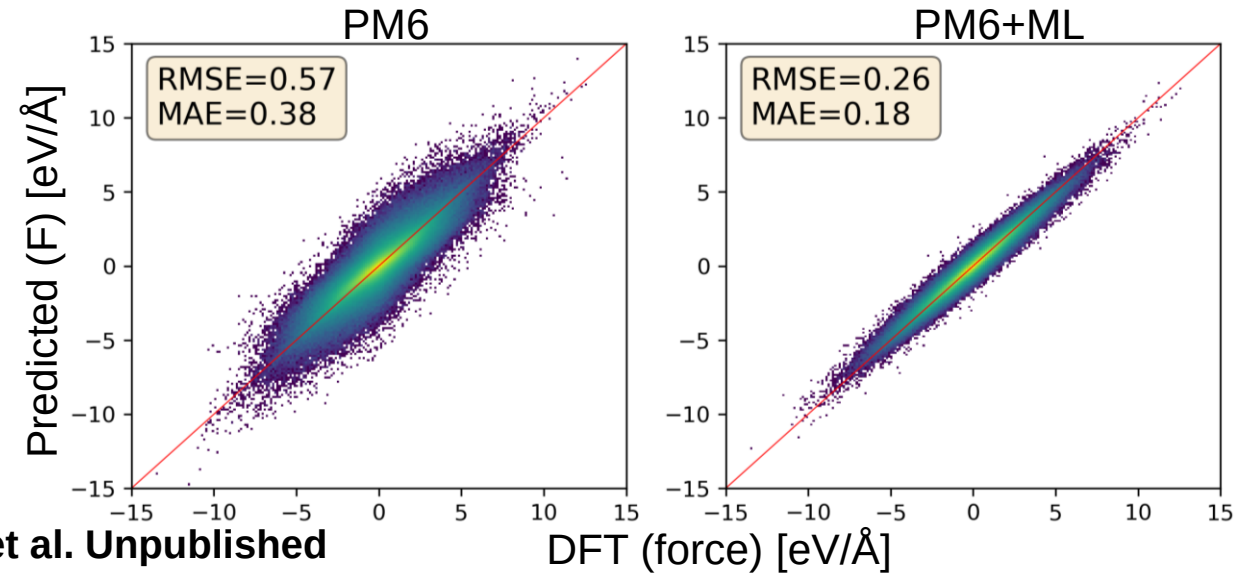
Held-out Test Set (14k, reactive)





Drug Bank subset (13k, real drug molecules)

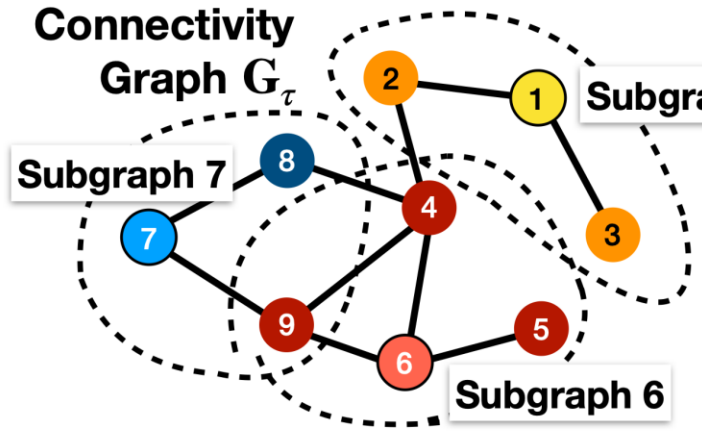




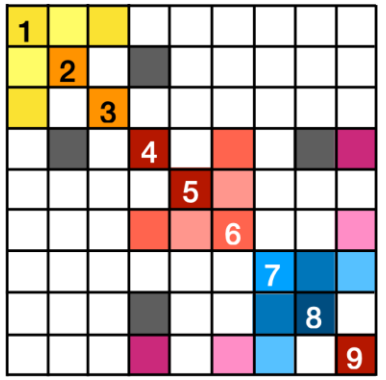
Robert Stanton et al. Unpublished

Graph-based Electronic Structure Theory

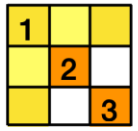
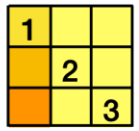
Non-overlapping cores
Overlapping halos

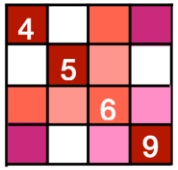
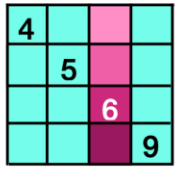


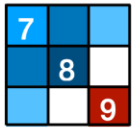
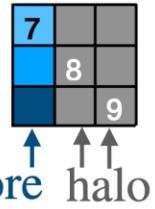
$\mathbf{X} =$



Initial Density Matrix

$\mathbf{x}^{(1)} =$  $f(\mathbf{x}^{(1)}) =$ 

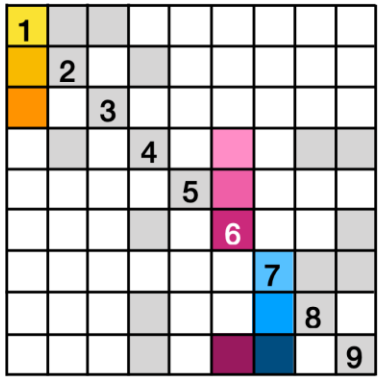
$\mathbf{x}^{(6)} =$  $f(\mathbf{x}^{(6)}) =$ 

$\mathbf{x}^{(7)} =$  $f(\mathbf{x}^{(7)}) =$ 

Initial Density SubMatrices

Updated Density SubMatrices

$f_{G_\tau}(\mathbf{X}) =$



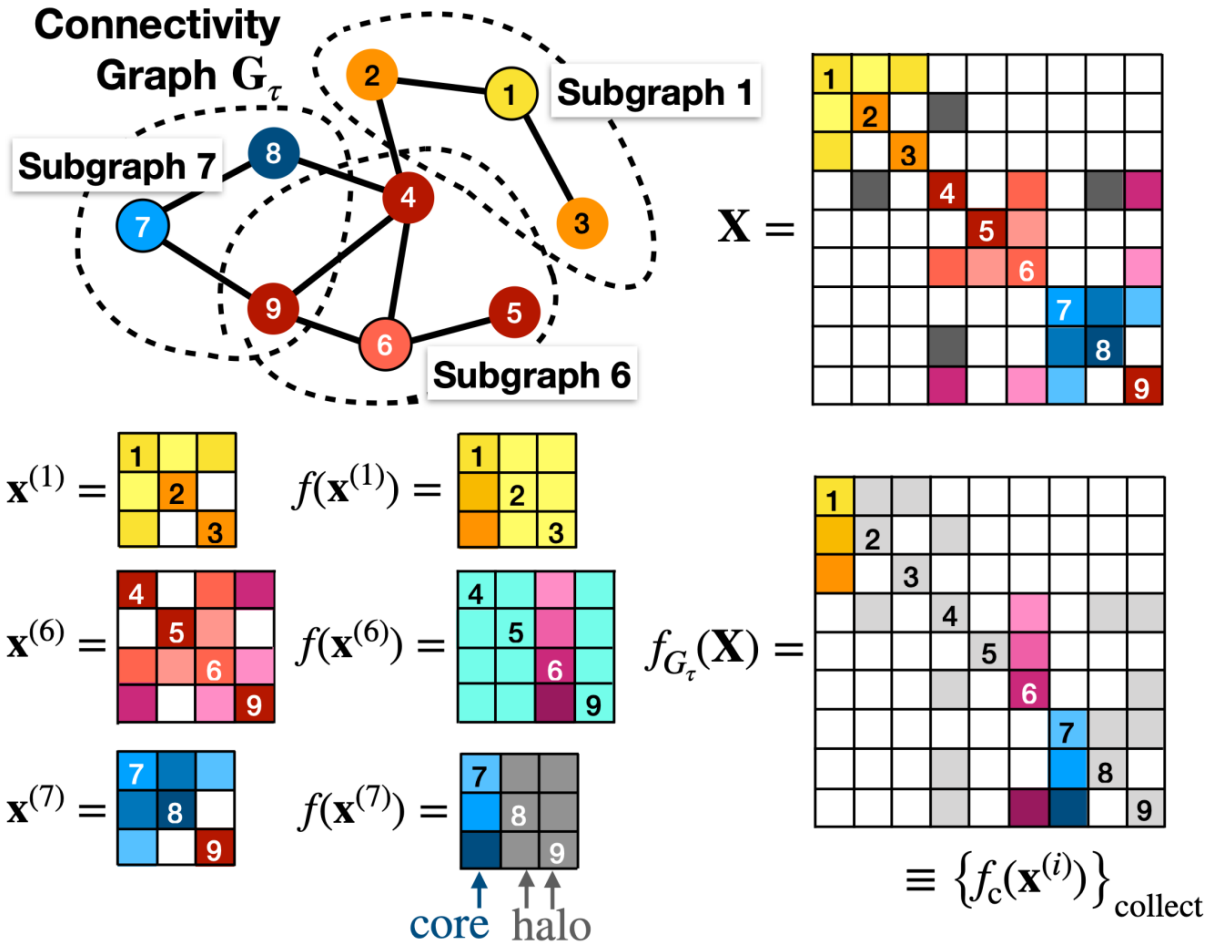
$\equiv \{f_c(\mathbf{x}^{(i)})\}_{\text{collect}}$

Updated Density Matrix Reconstructed from Subgraphs

C. F. A. Negre *et al*. J. Chem. Phys. 21 February 2023; 158 (7): 074108

Graph-based Electronic Structure Theory

- Graph partitioning is adaptive: a new graph is constructed at each MD step or DM optimization step
- Overlap matrix or distances for initialization



Subgraphs are not isolated!

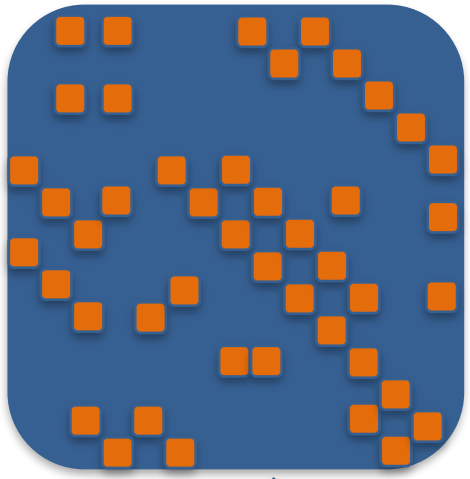
A full (non-local) Coulomb summation is performed when constructing Fock SubMatrices.

Subgraphs "feel" each other.

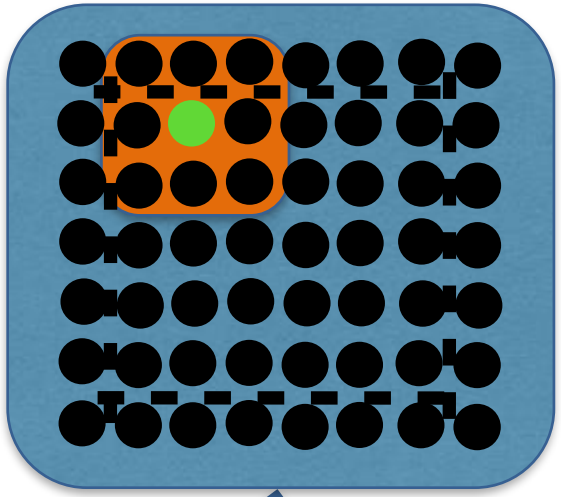
But the eigenvalue problem is solved for a subgraph only.

Linear Scaling Electronic Structure Theory

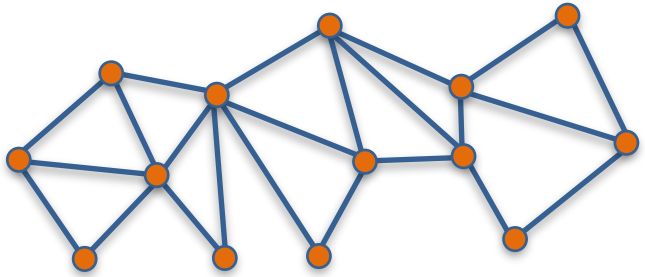
Sparse Matrix Algebra



Divide and Conquer



Graph Theory



a) Simple error control

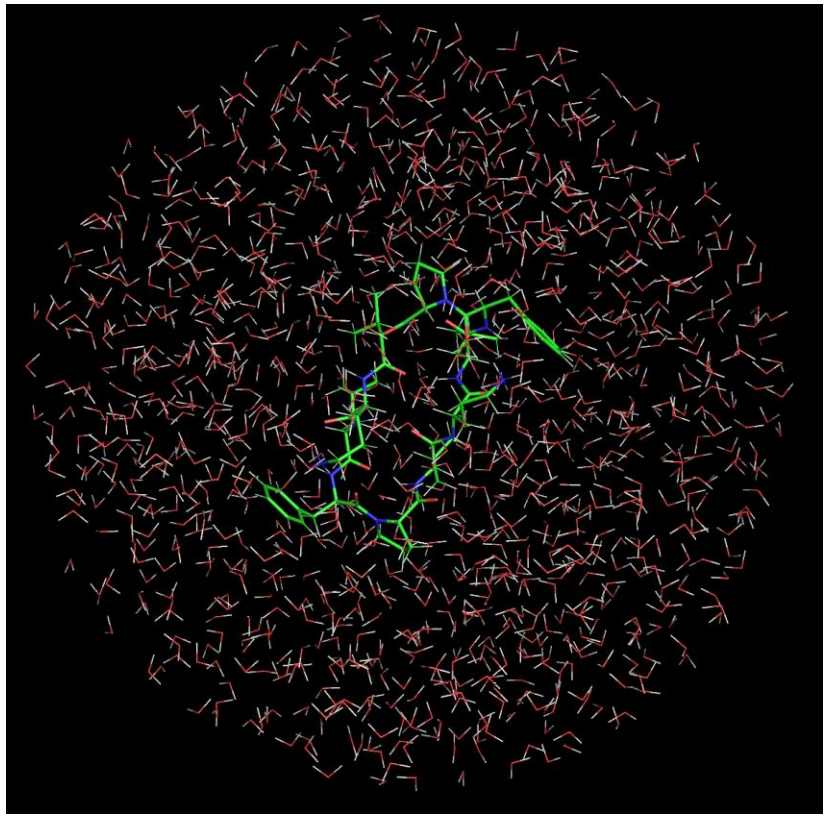
b) Efficient parallelism on distributed memory

c) Small overhead

AMNN et al. *"Graph-based linear scaling electronic structure theory"*, J. Chem. Phys. 144, 234101 (2016)

Density Matrix Optimization of Solvated Gramicidin S

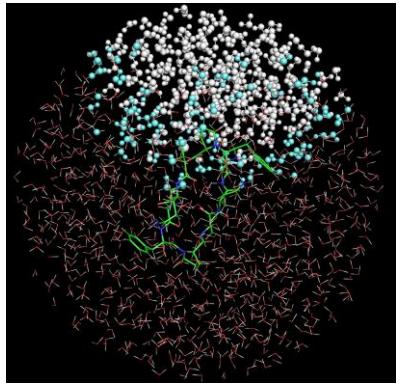
Gramicidin S solvated in water
3605 atoms



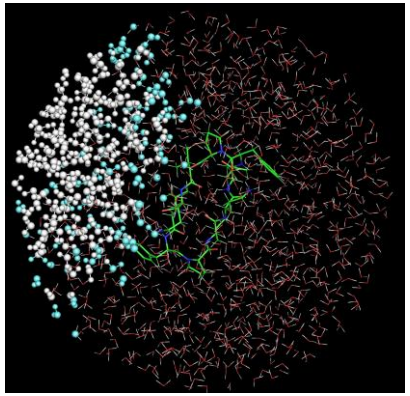
(Built via PACKMOL)

7 subgraphs

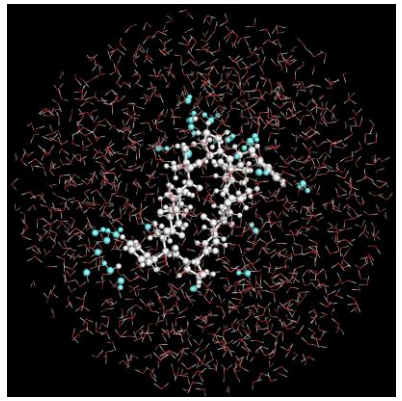
CORE: 636
HALO: 183



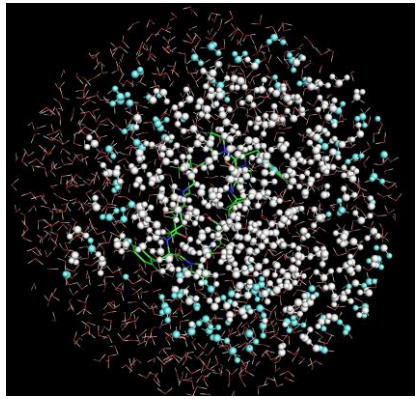
CORE: 443
HALO: 165



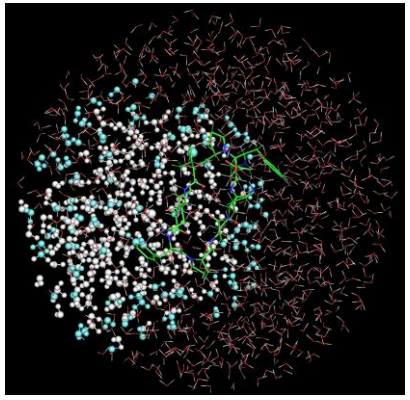
CORE: 180
HALO: 46



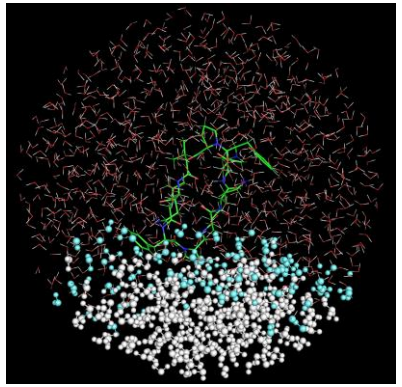
CORE: 624
HALO: 178



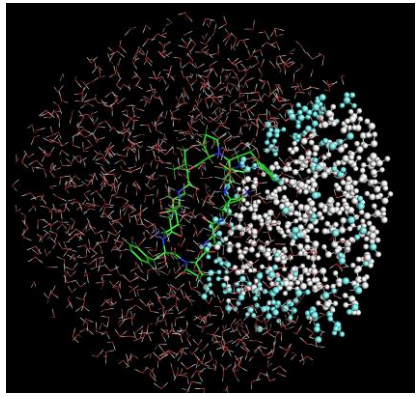
CORE: 556
HALO: 218



CORE: 639
HALO: 188

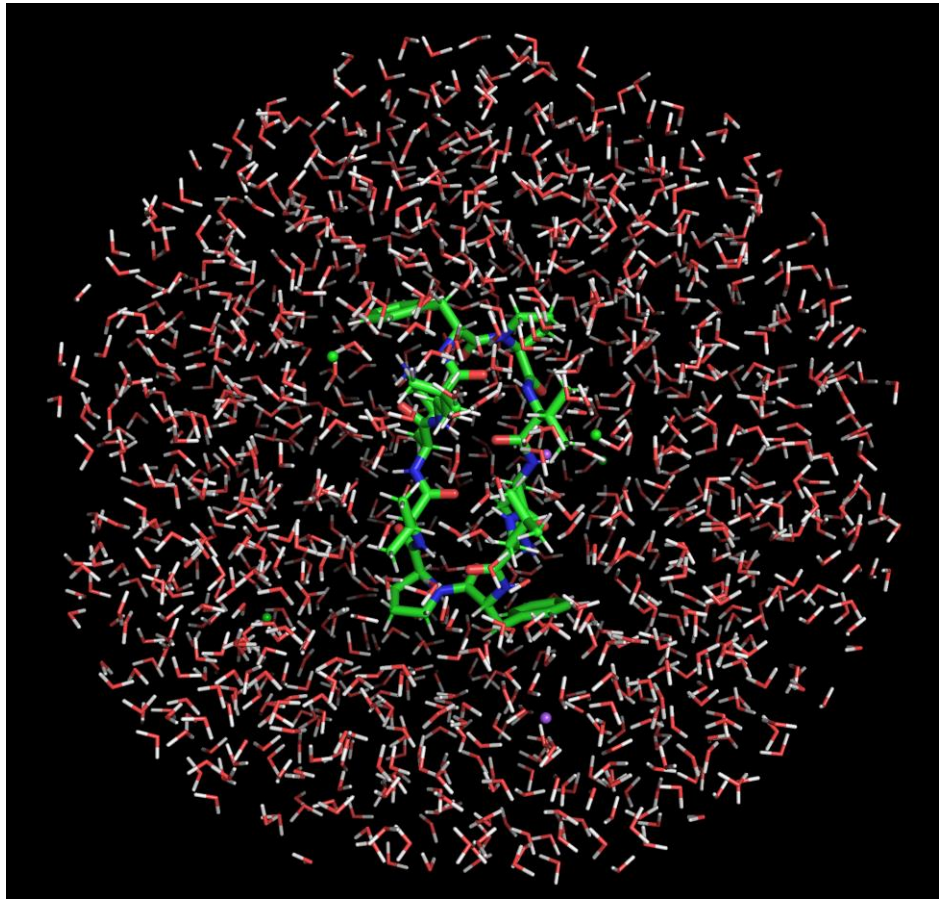


CORE: 527
HALO: 203



Density Matrix Optimization of solvated Gramicidin S

Intel(R) Xeon(R) Gold 5218R CPU @ 2.10GHz



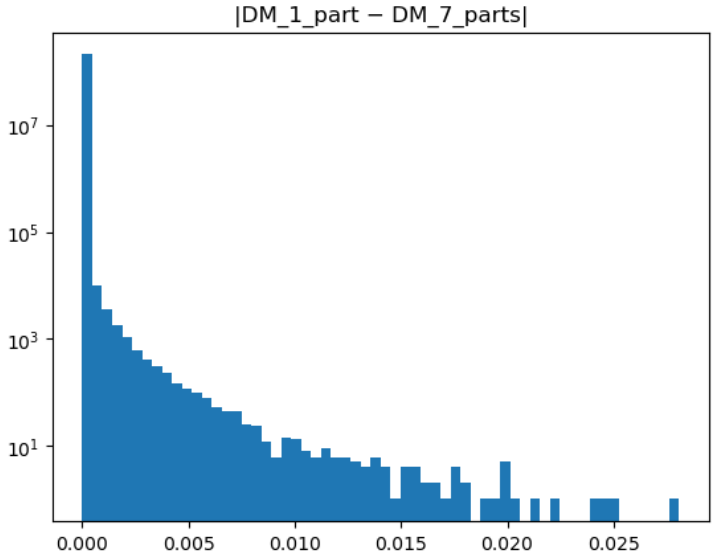
3605 atoms

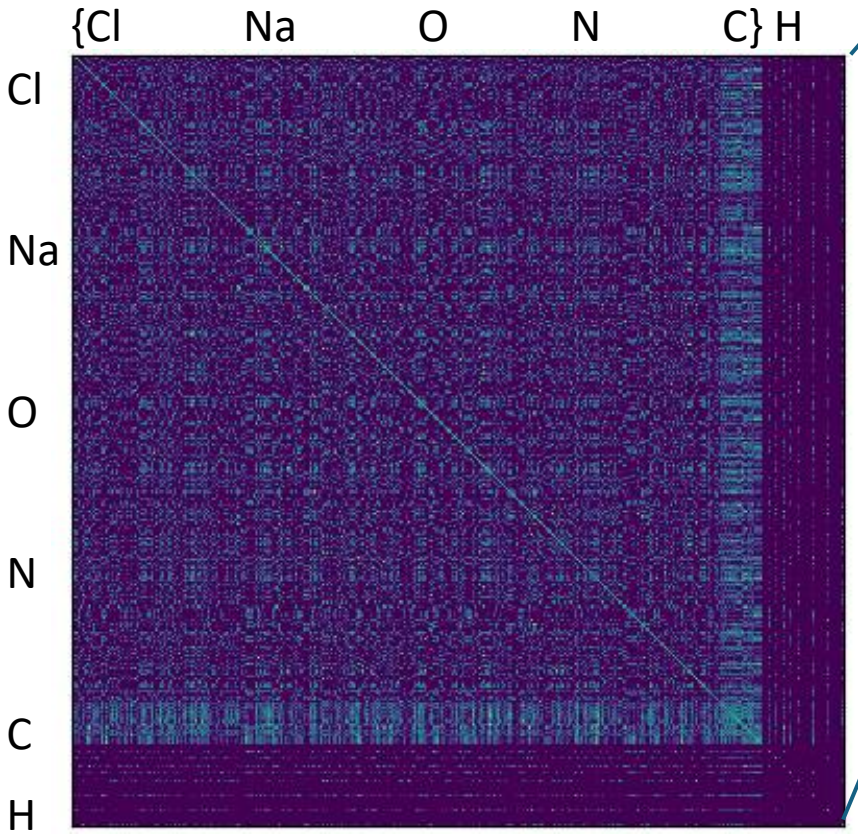
1 part:	7 parts:
Etot: -378831.186 eV	Etot: -378831.191 eV

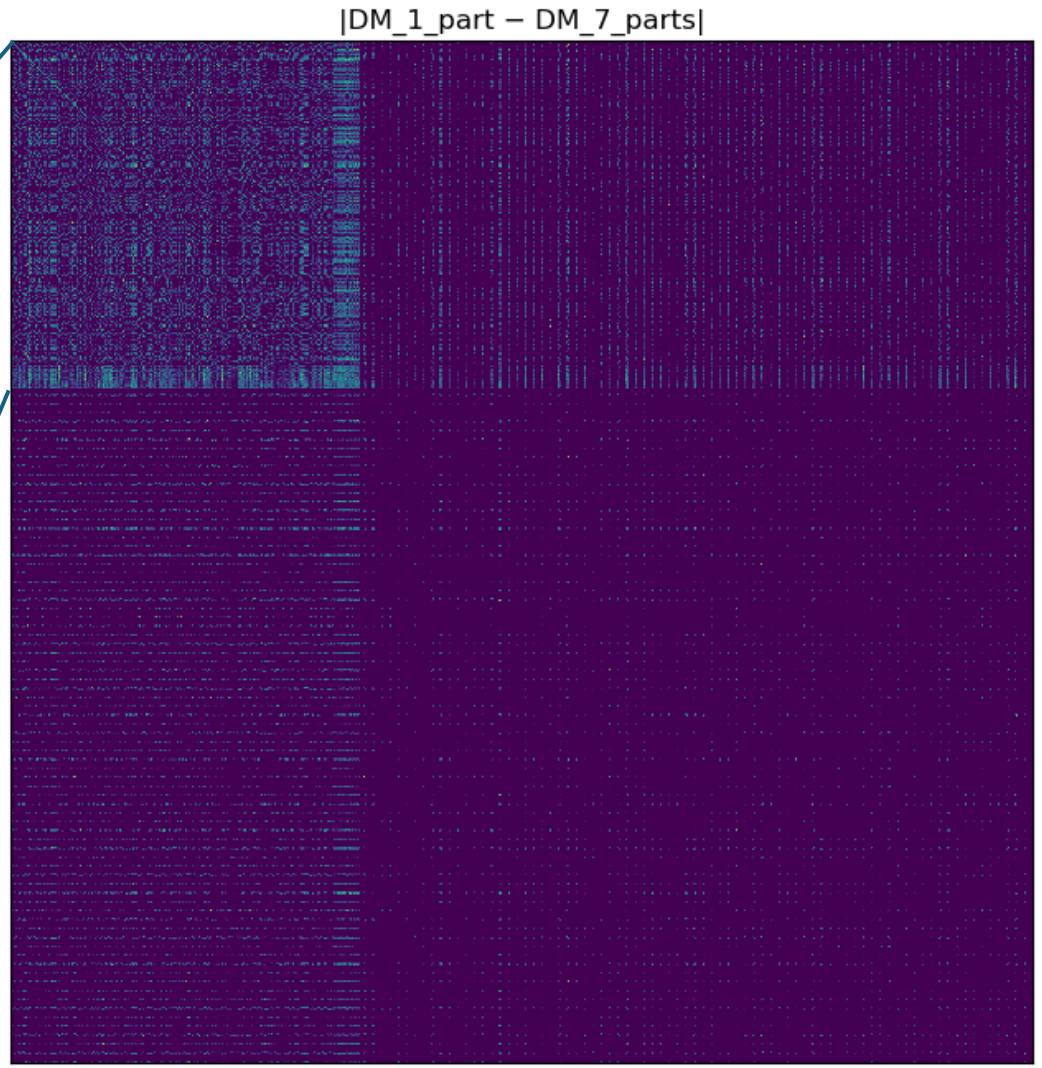
ΔEtot = 0.005 eV

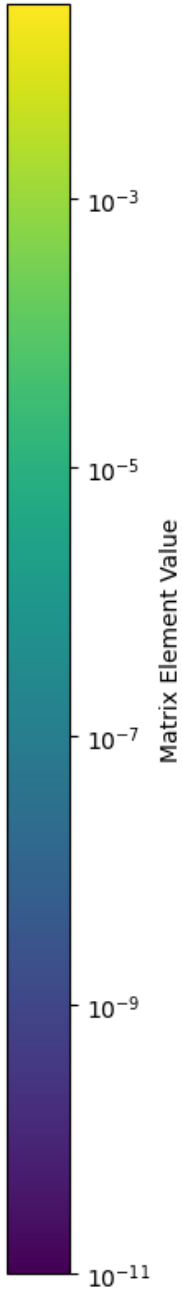
11min 37s	2min 11s

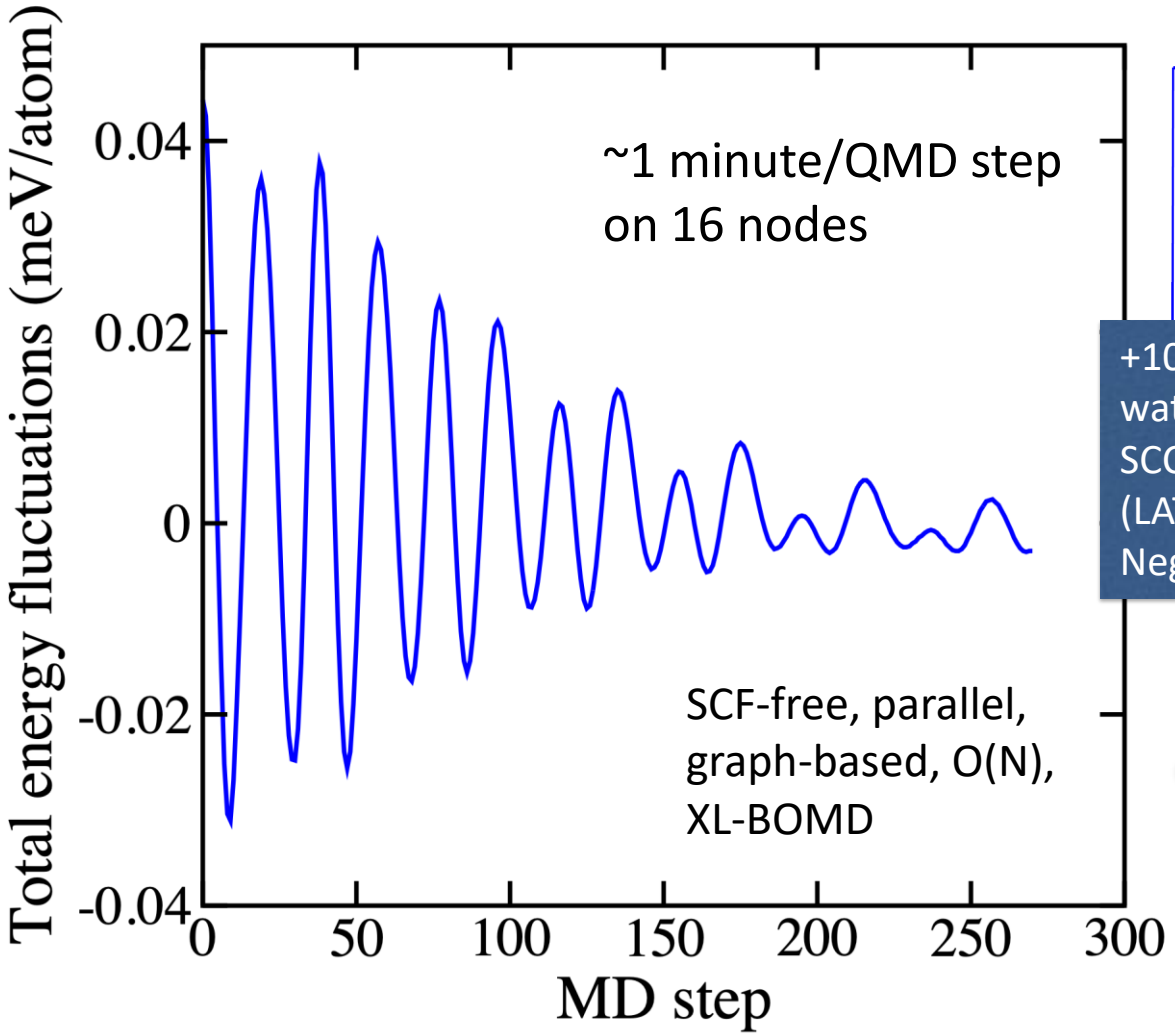
Density Matrix Comparison
Whole System vs Subgraphs

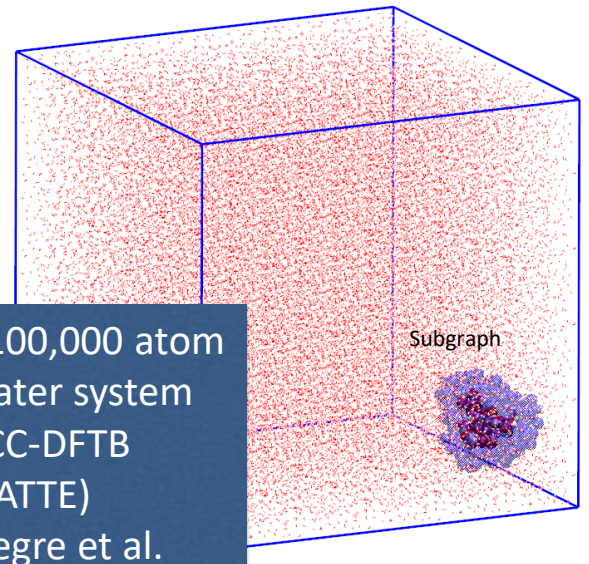




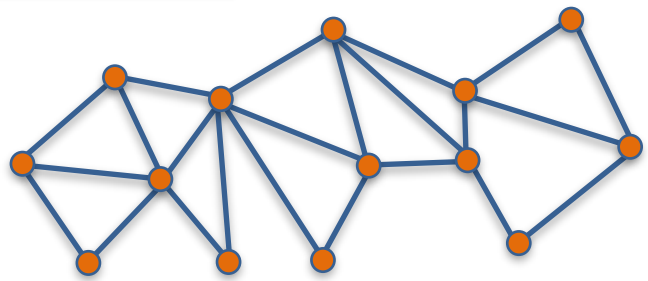






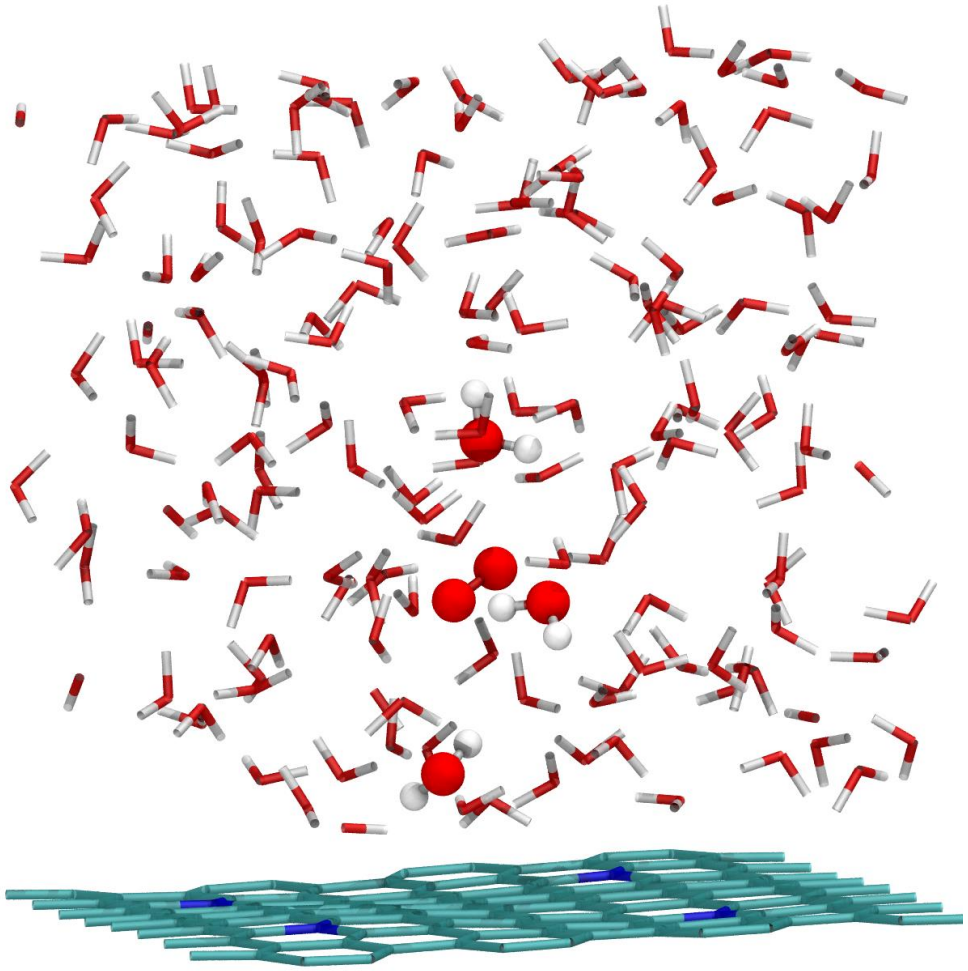


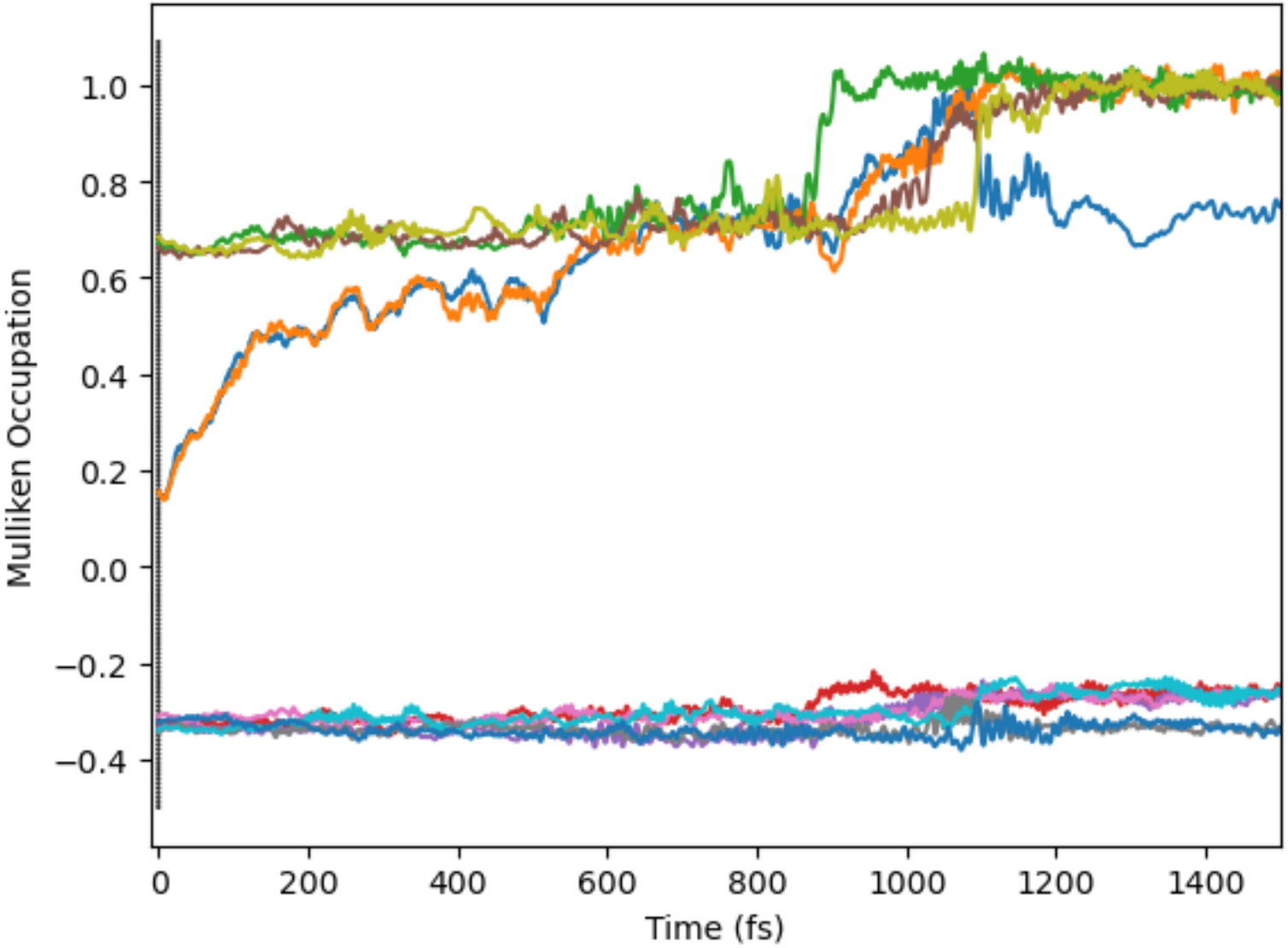
+100,000 atom water system SCC-DFTB (LATTE) Negre et al.



Graph Theory

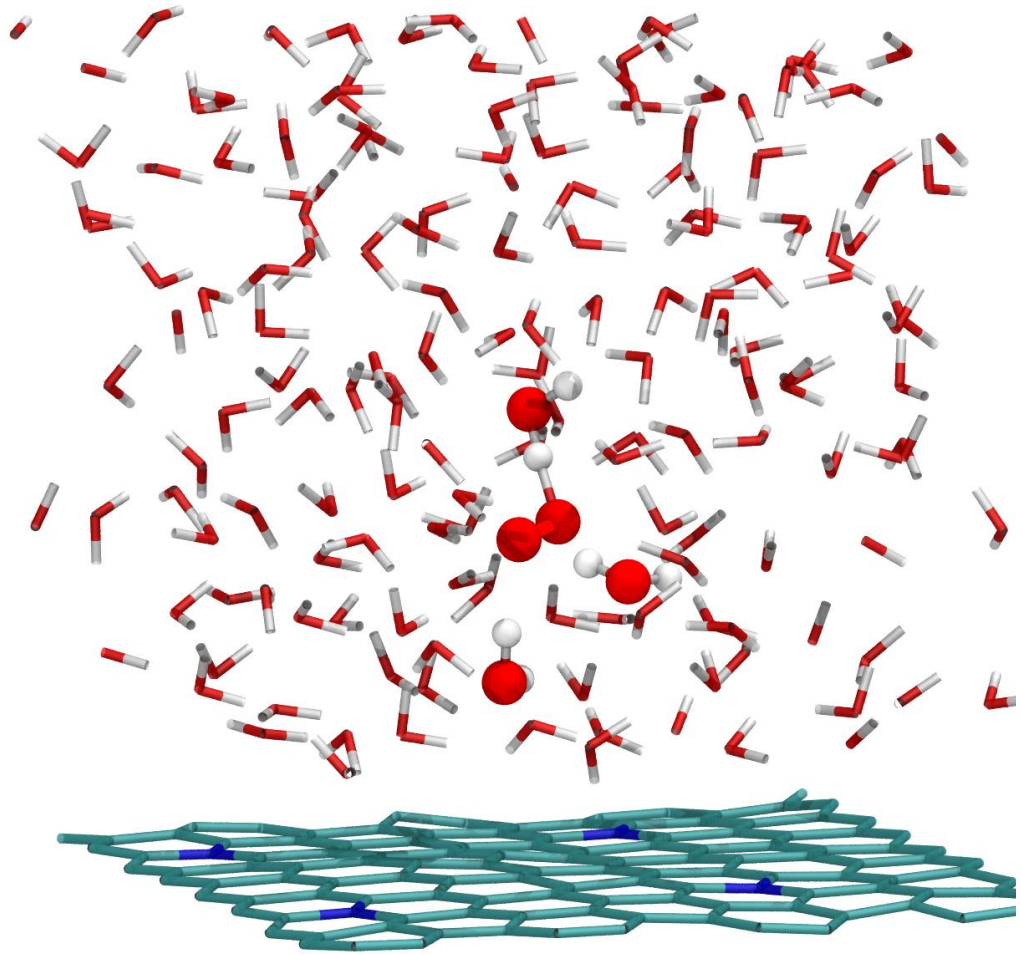
Oxygen Reduction Reaction

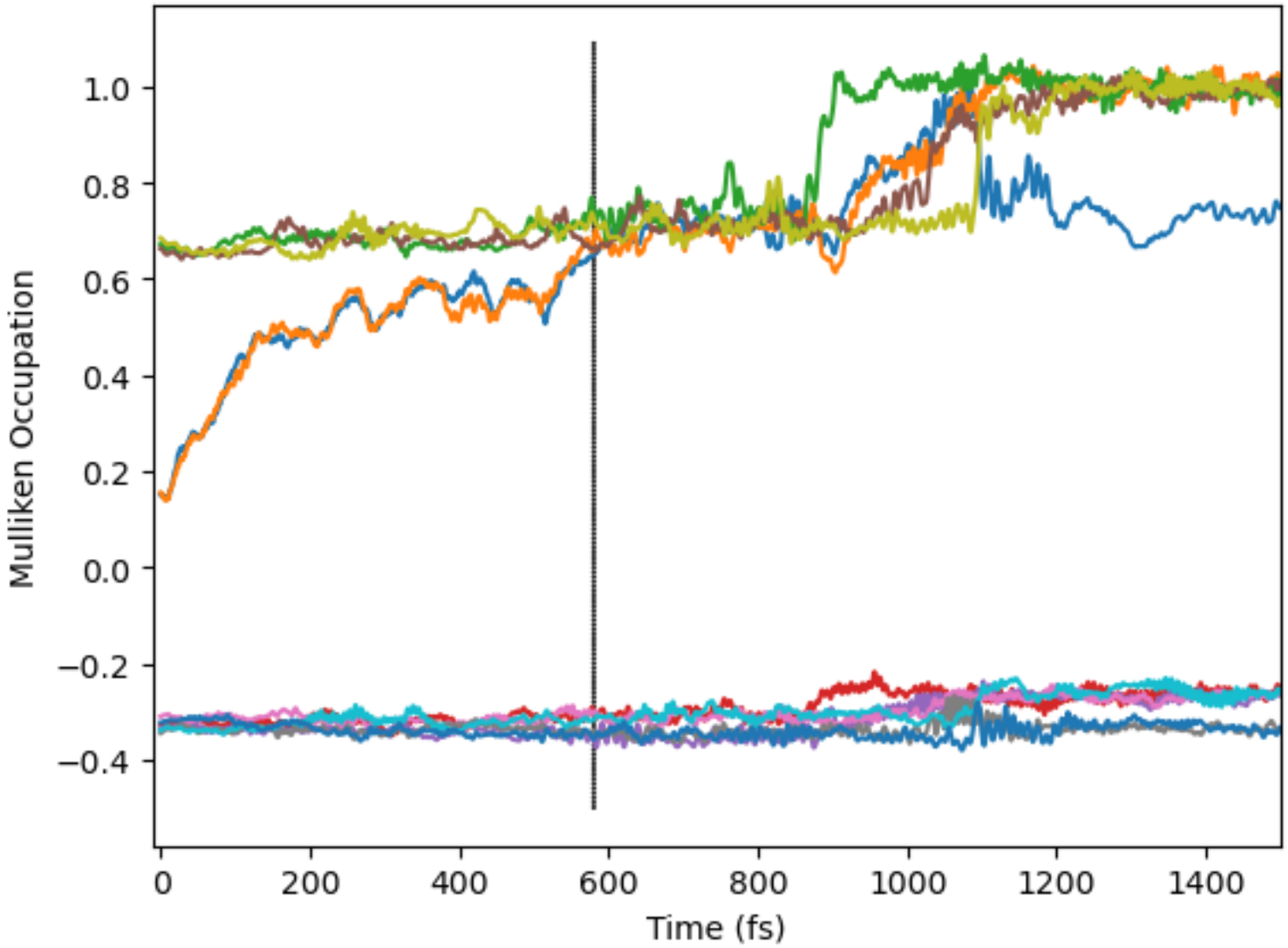


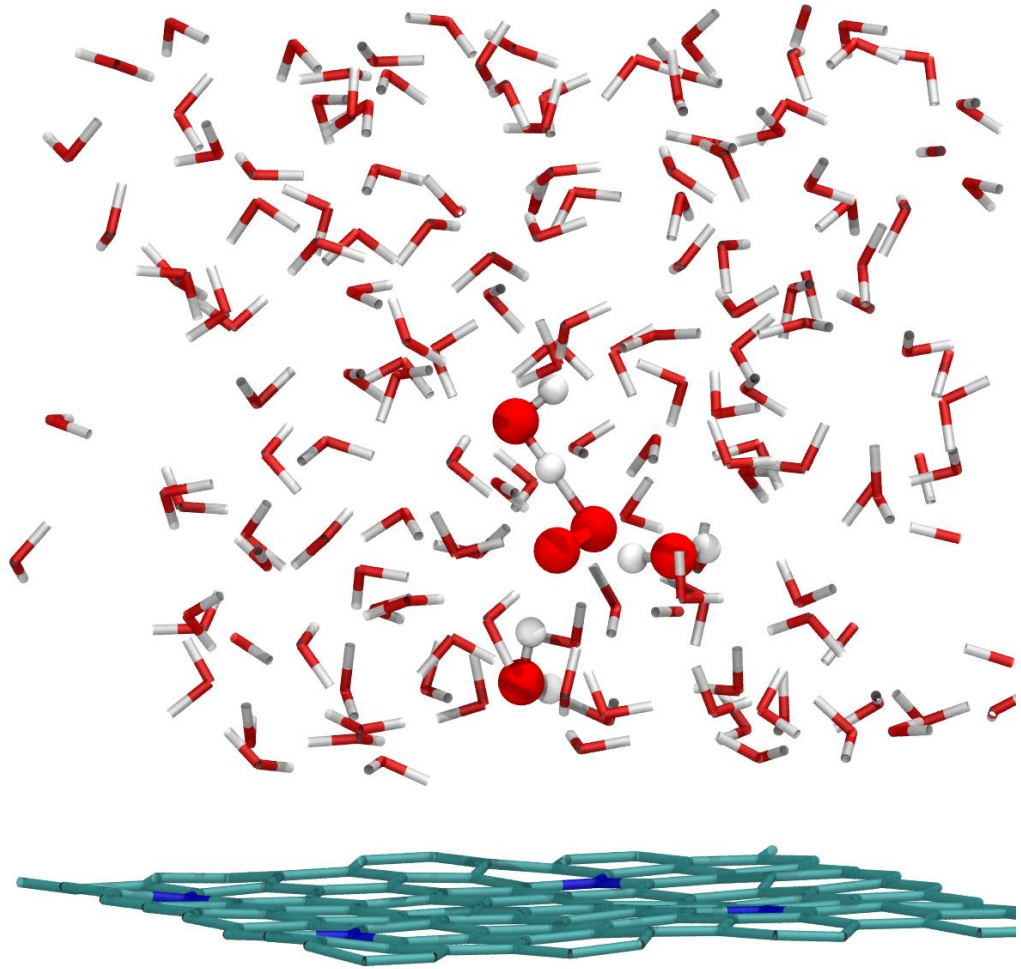


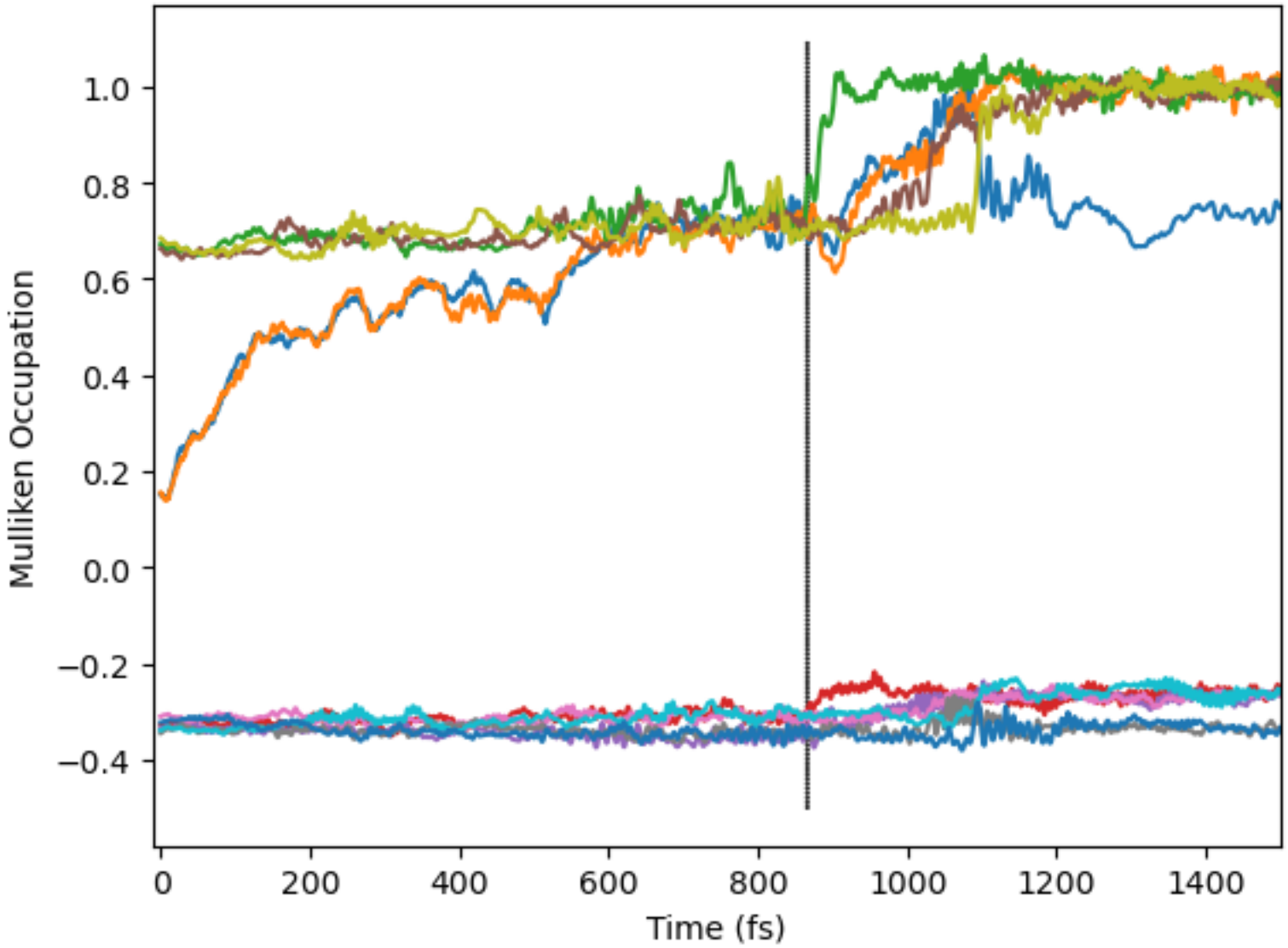
Fundamental cathode reaction in fuel cells/electrochemistry

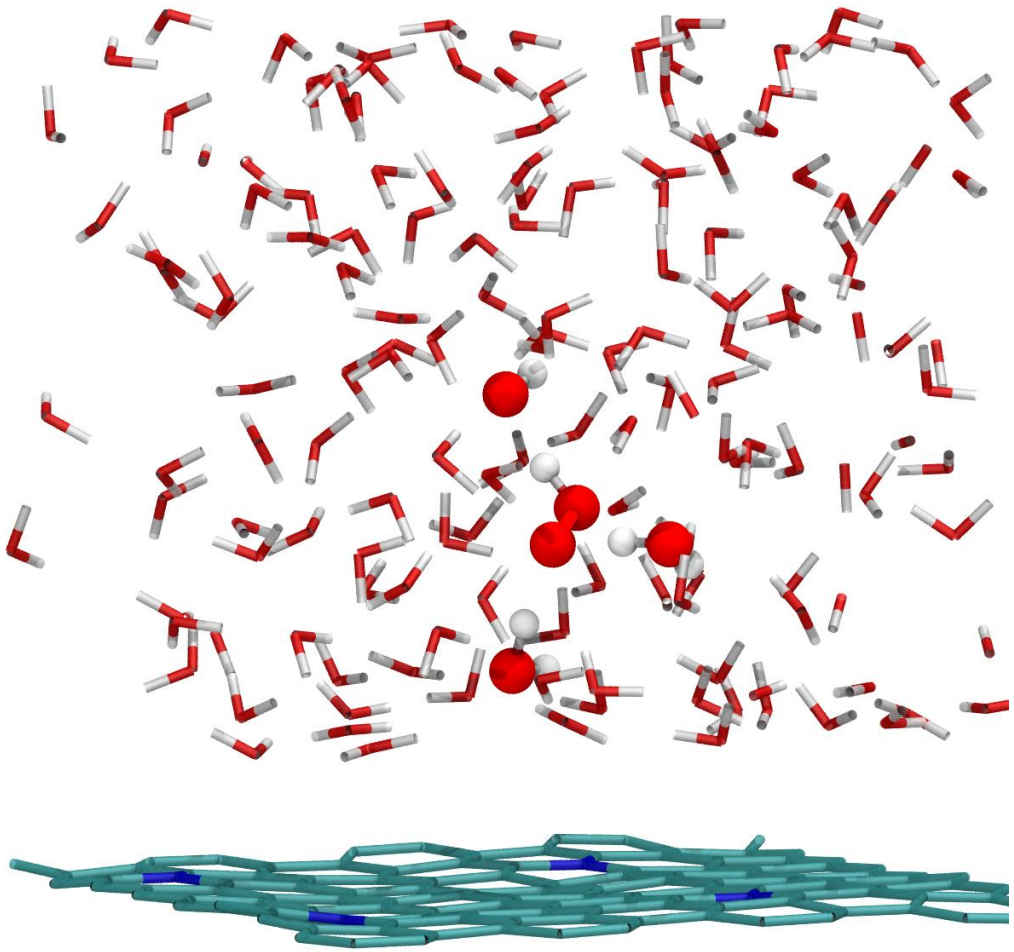
Rae C. Grove et al. Unpublished

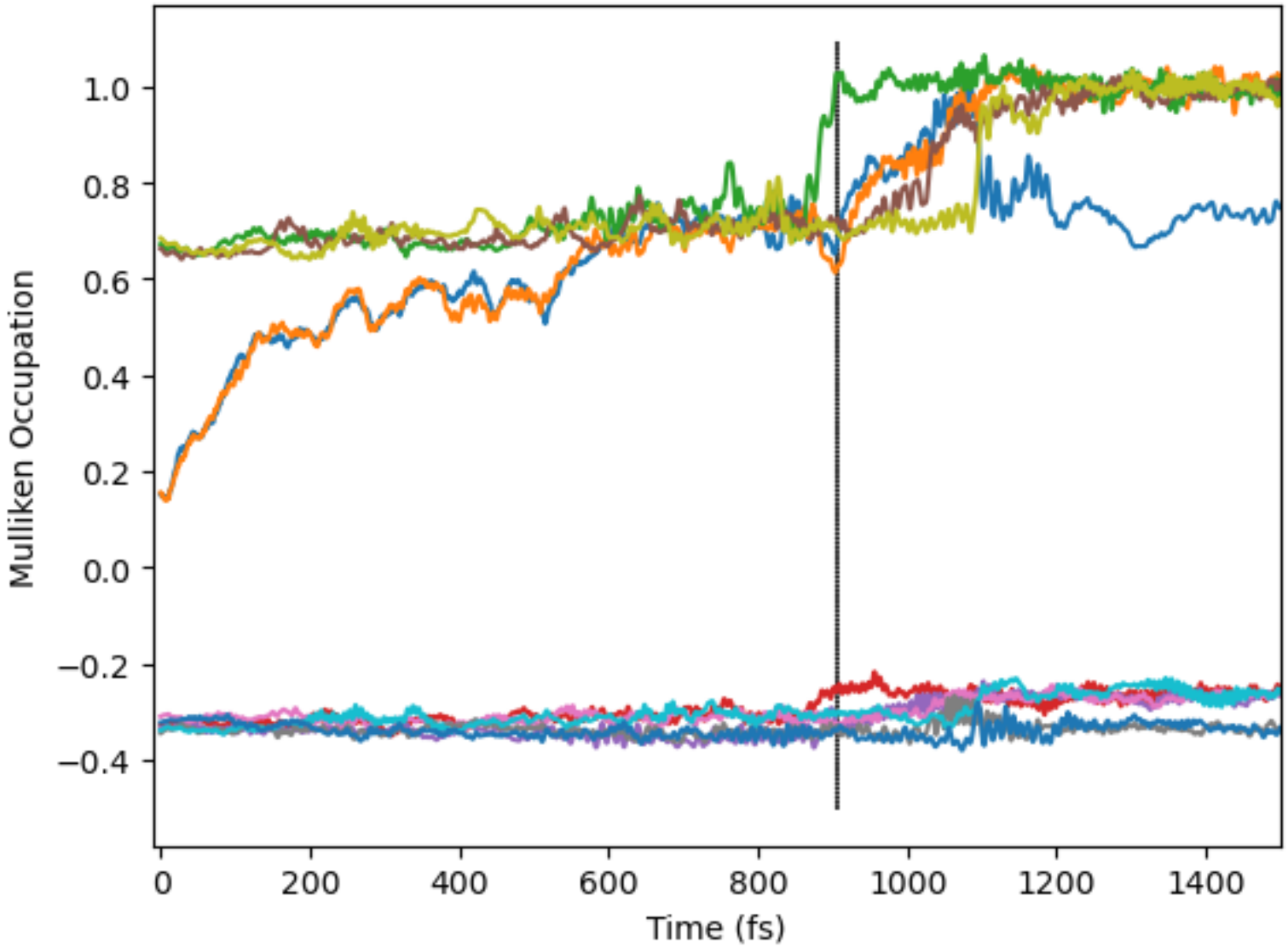


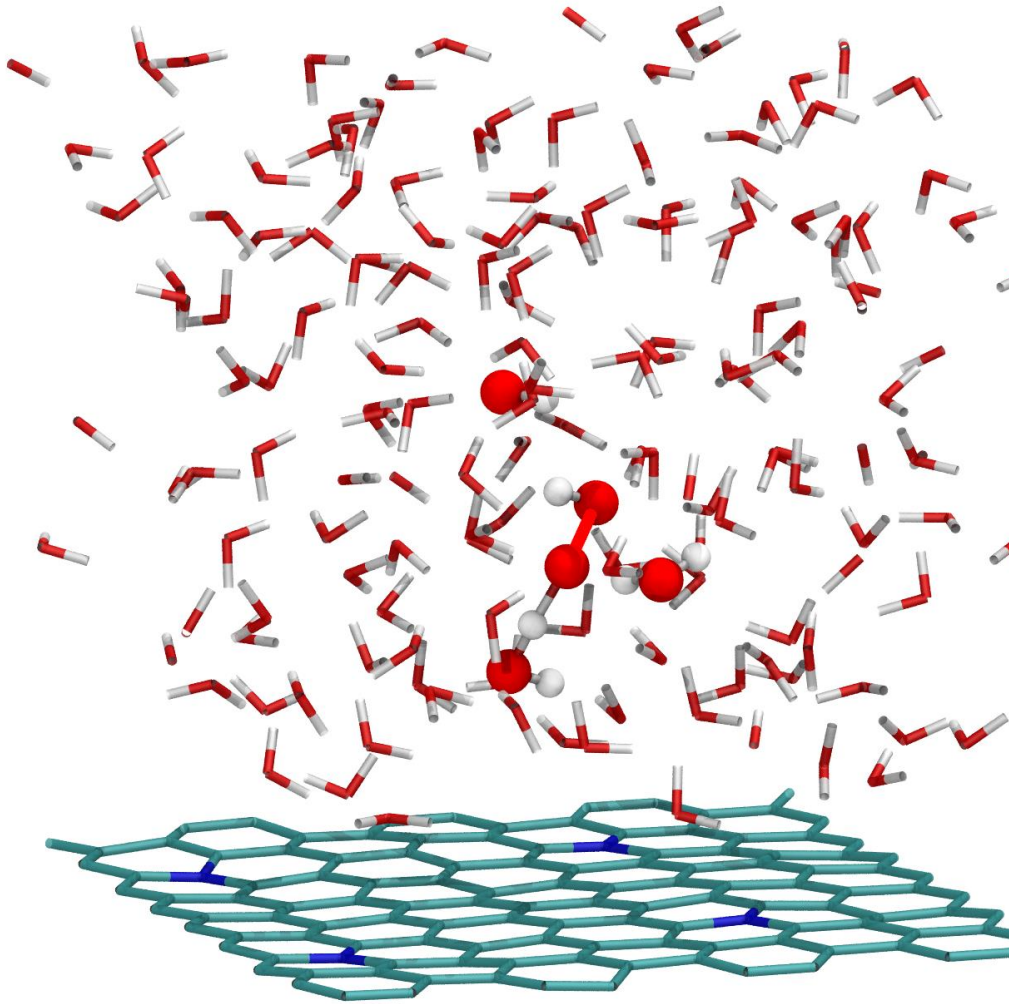

Mulliken Occupation
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
0
200
400
600
800
1000
1200
1400
Time (fs)

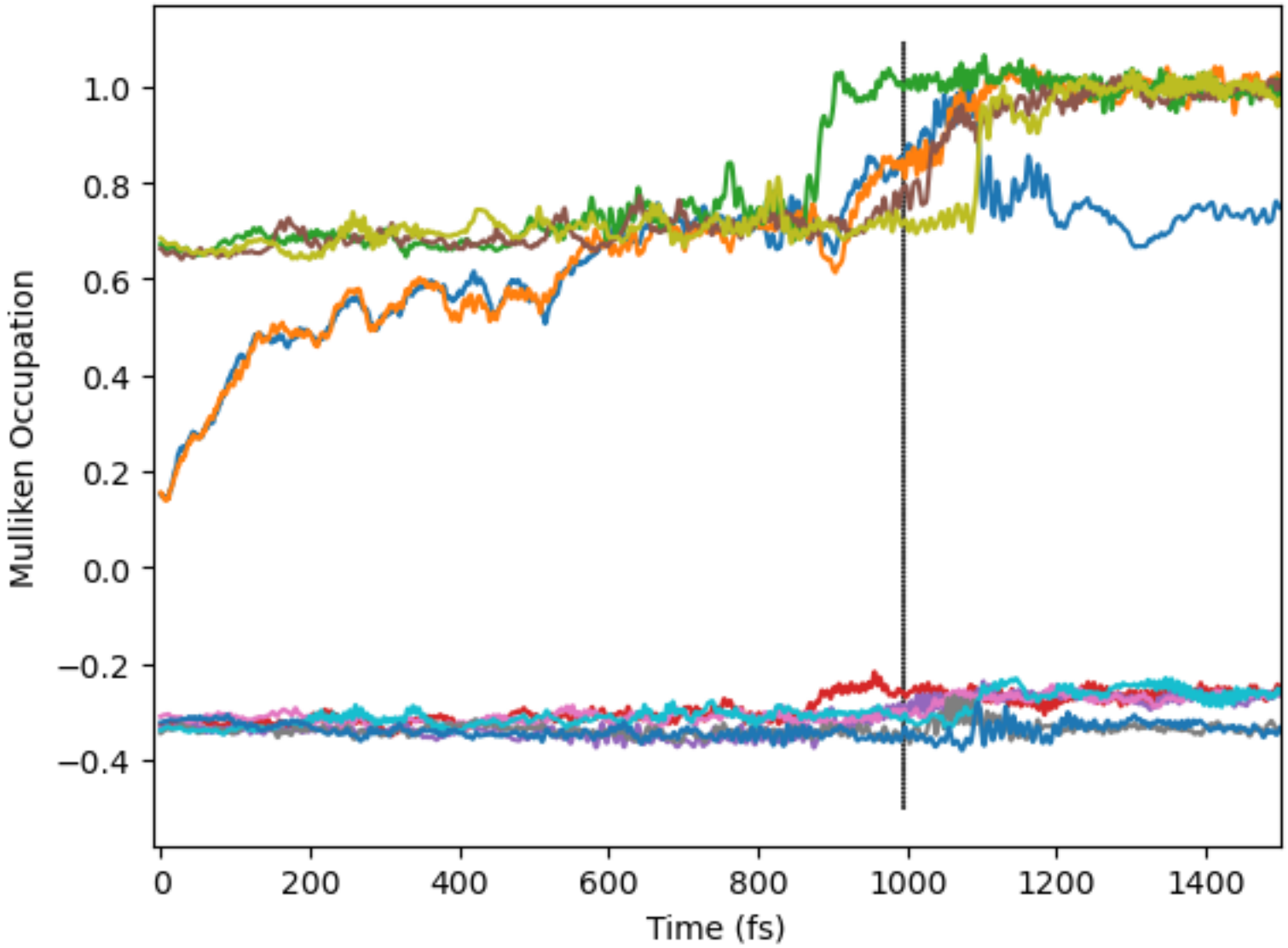


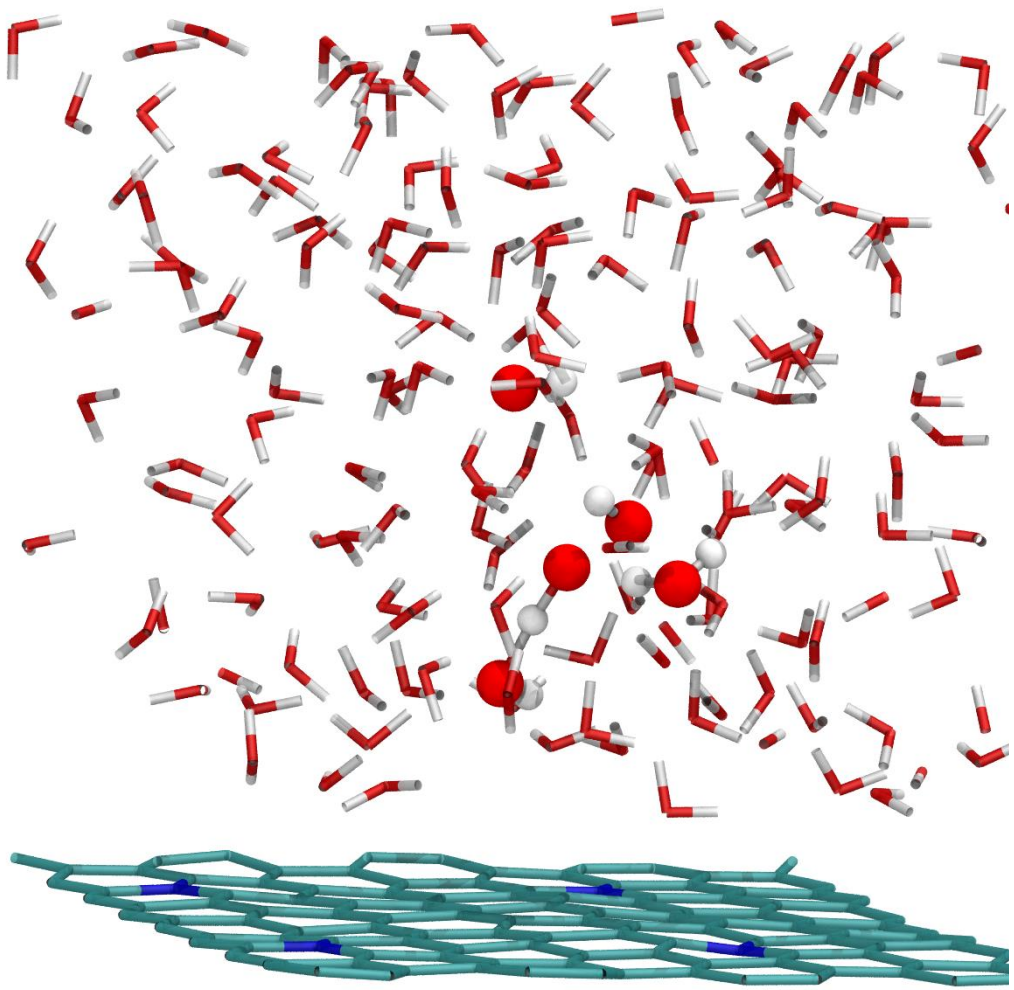

Mulliken Occupation
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
0
200
400
600
800
1000
1200
1400
Time (fs)

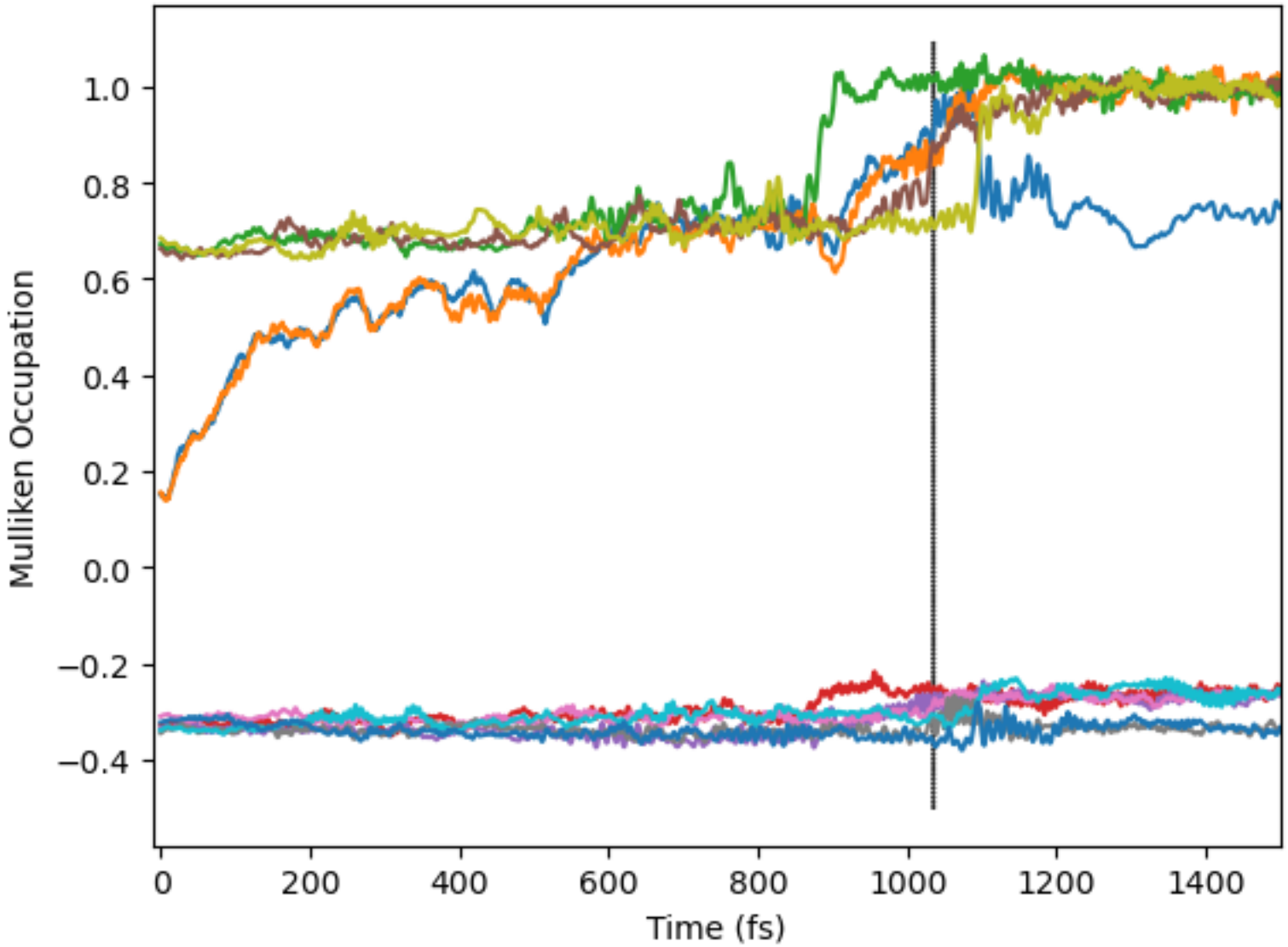


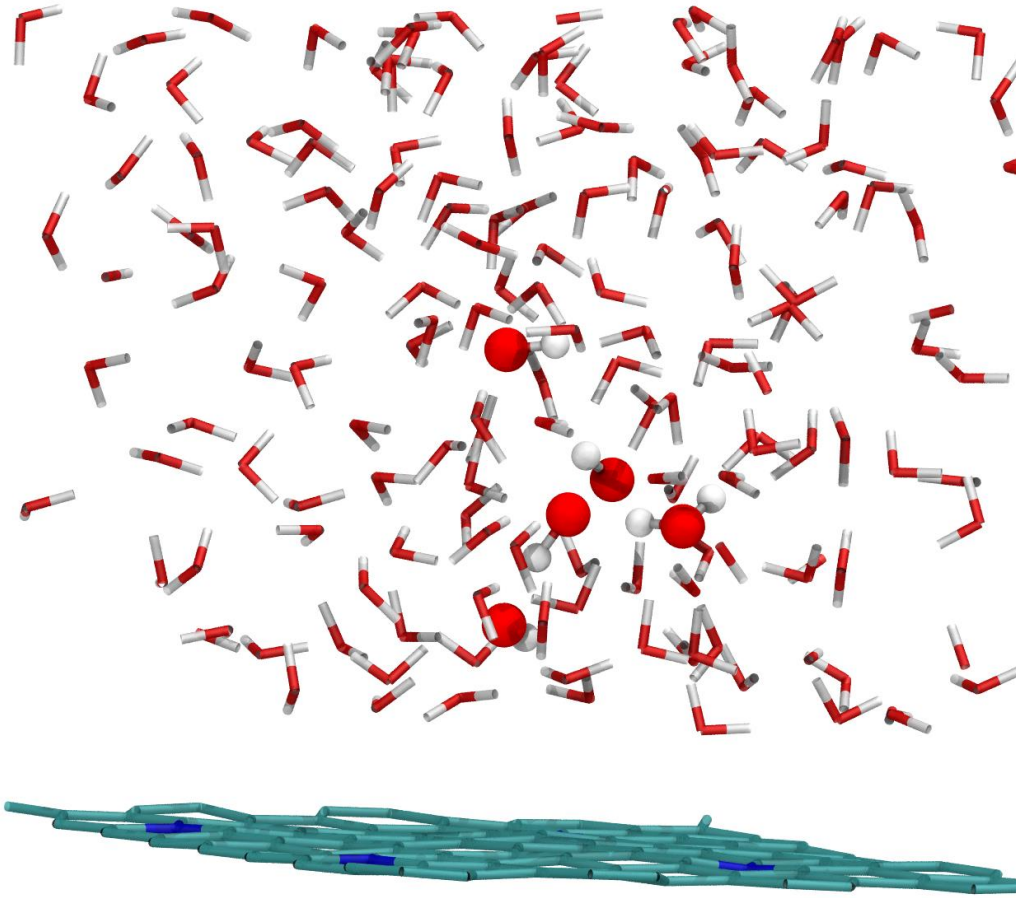

Mulliken Occupation
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
0
200
400
600
800
1000
1200
1400
Time (fs)

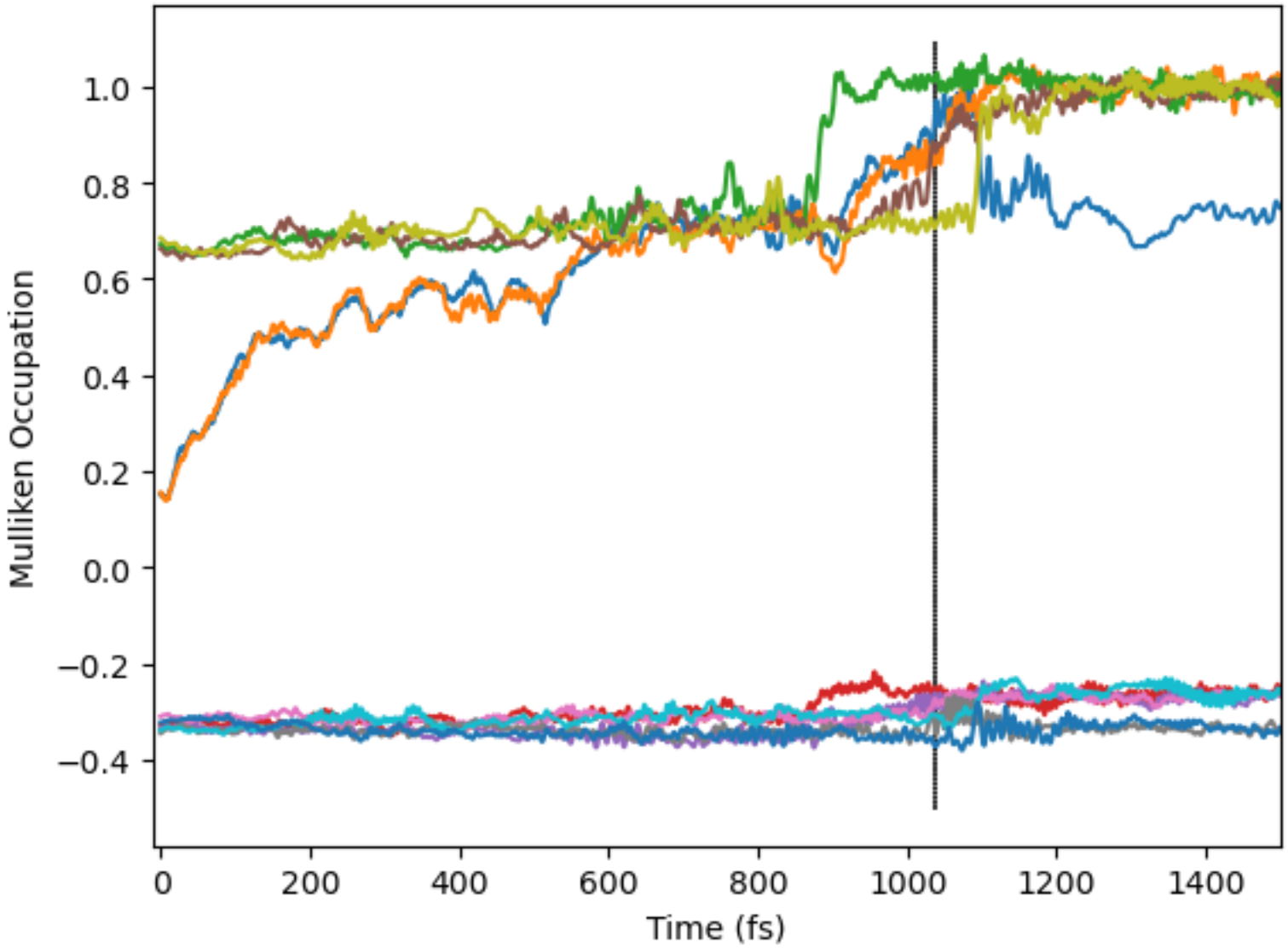


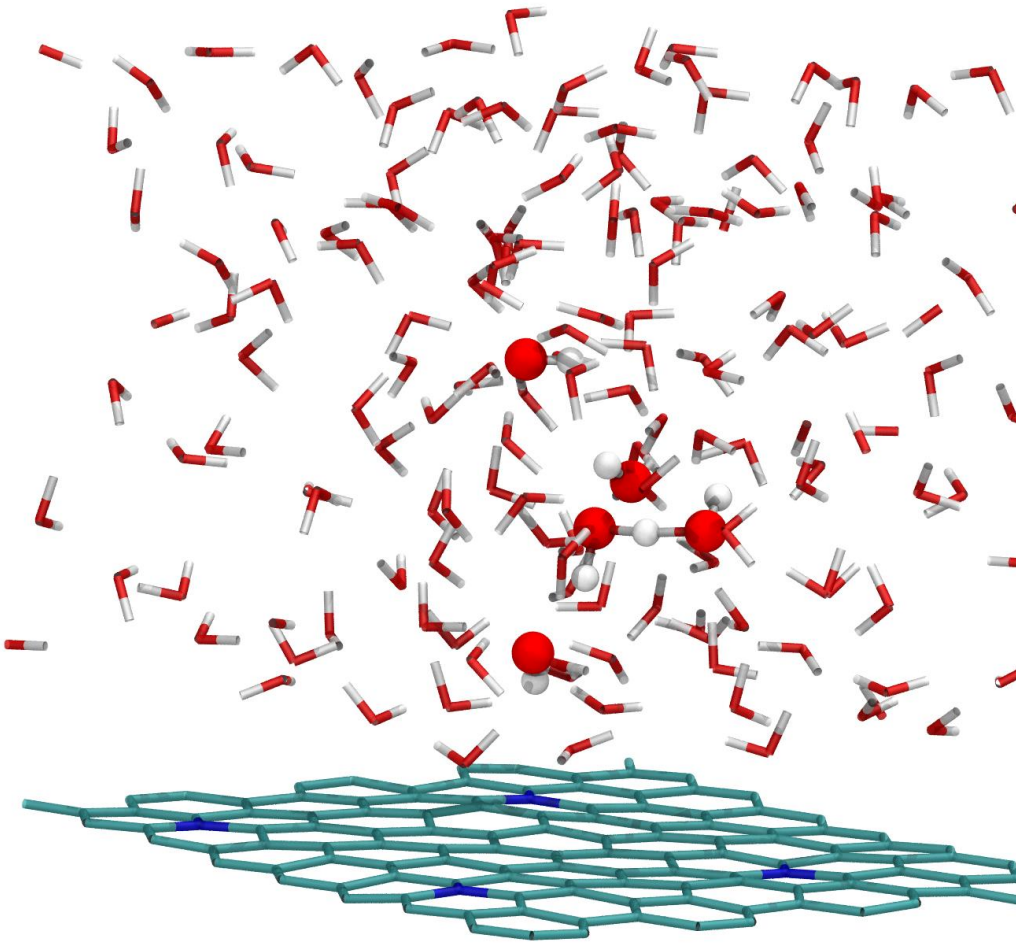

Mulliken Occupation
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
0
200
400
600
800
1000
1200
1400
Time (fs)

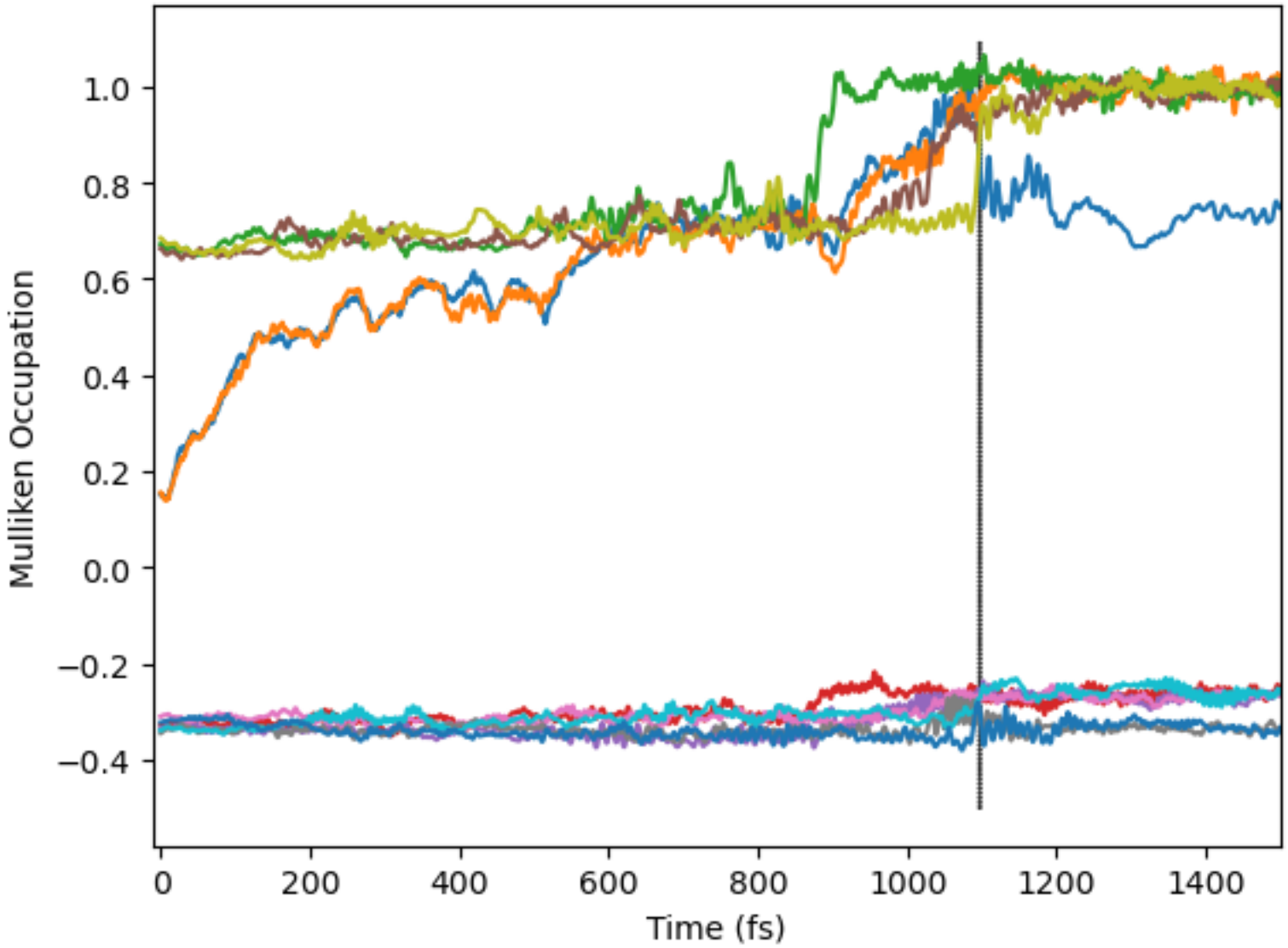


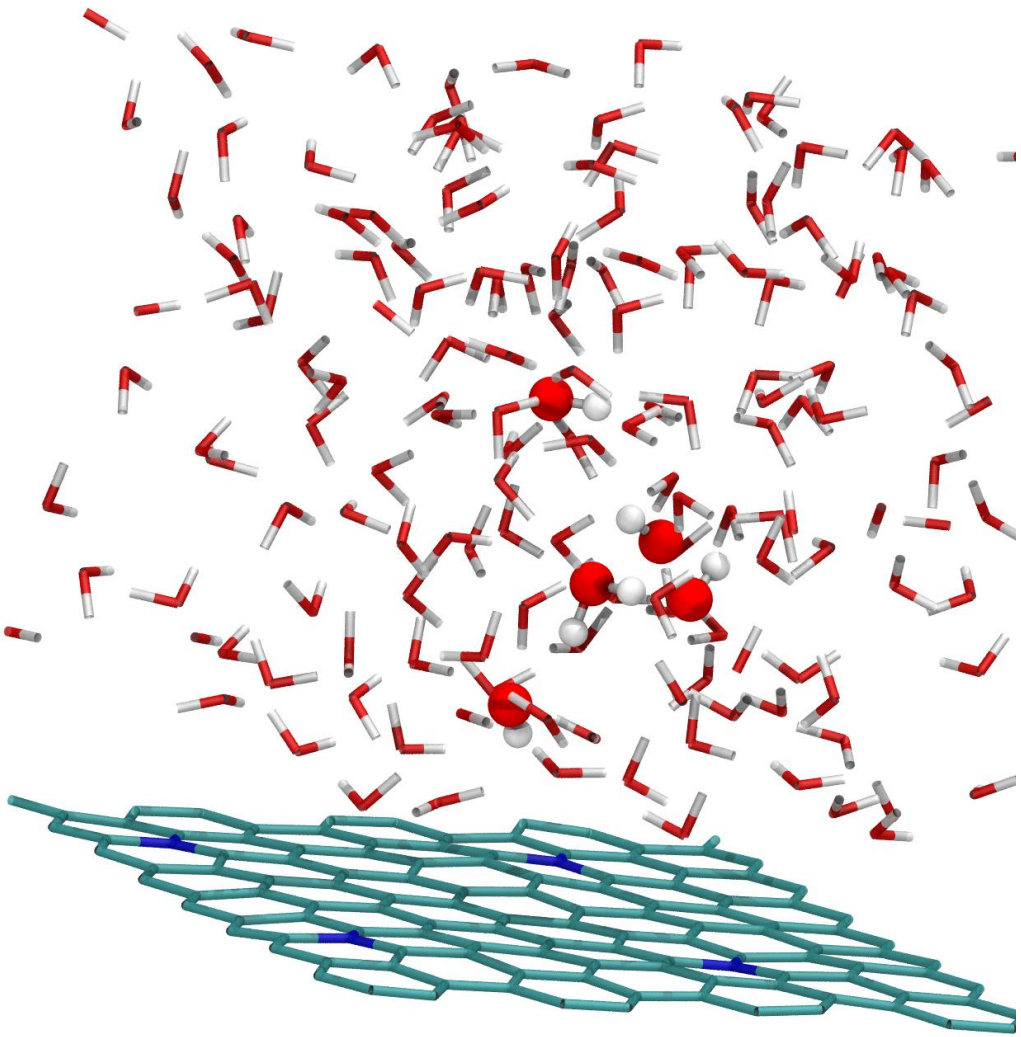

Mulliken Occupation
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
0
200
400
600
800
1000
1200
1400
Time (fs)

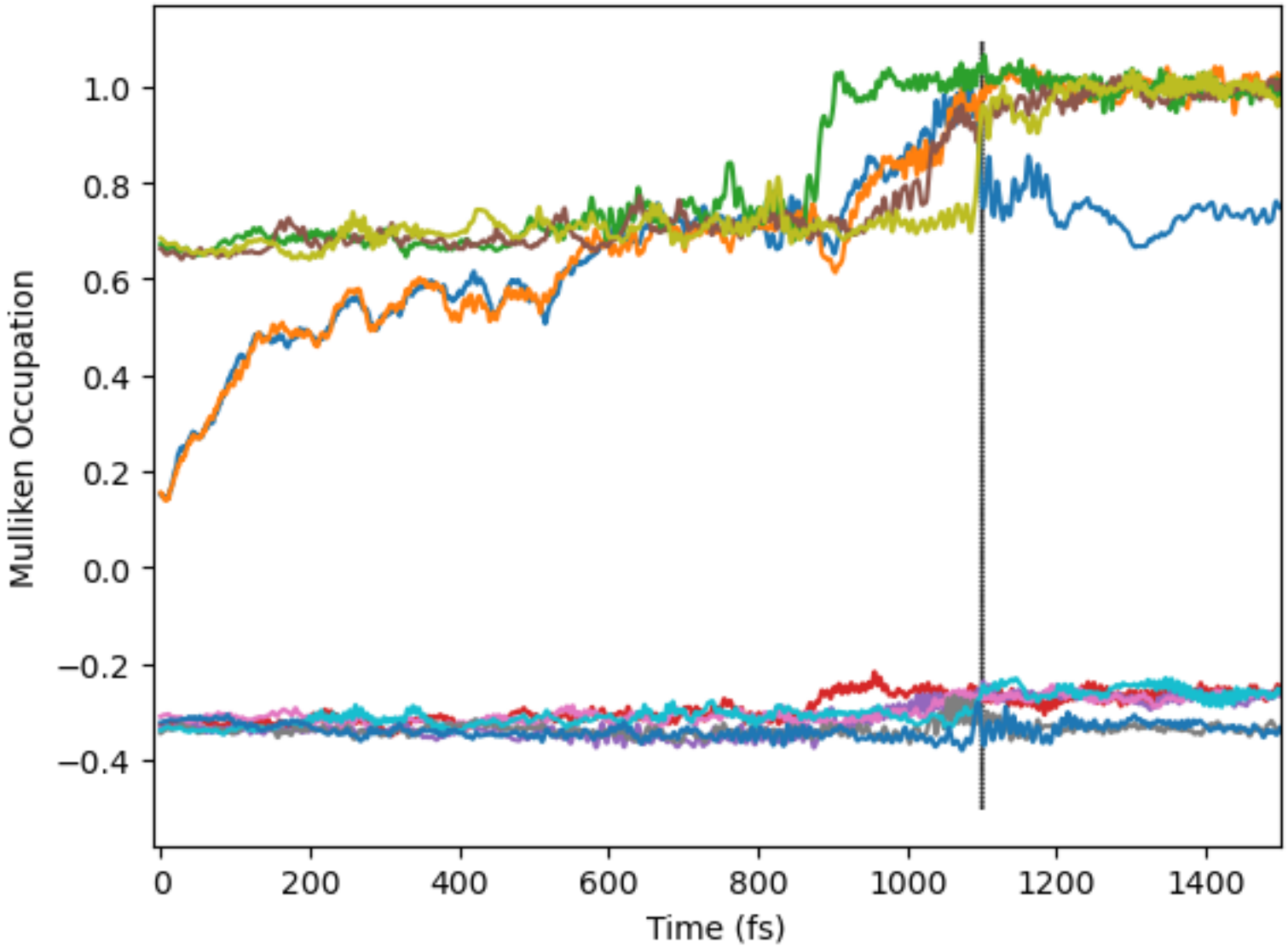


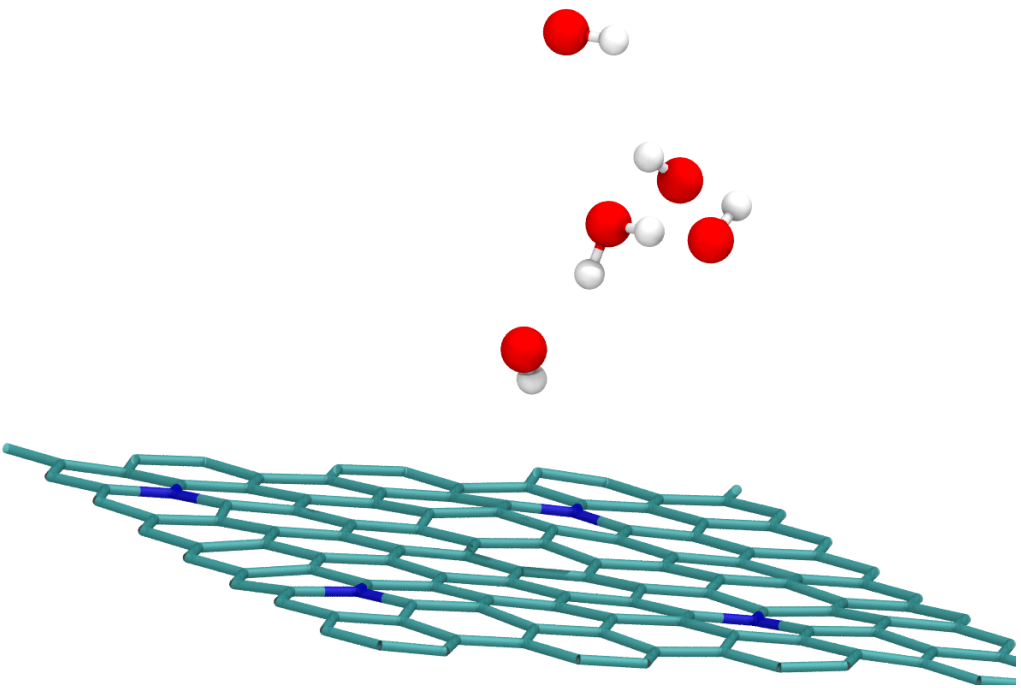

Mulliken Occupation
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
0
200
400
600
800
1000
1200
1400
Time (fs)

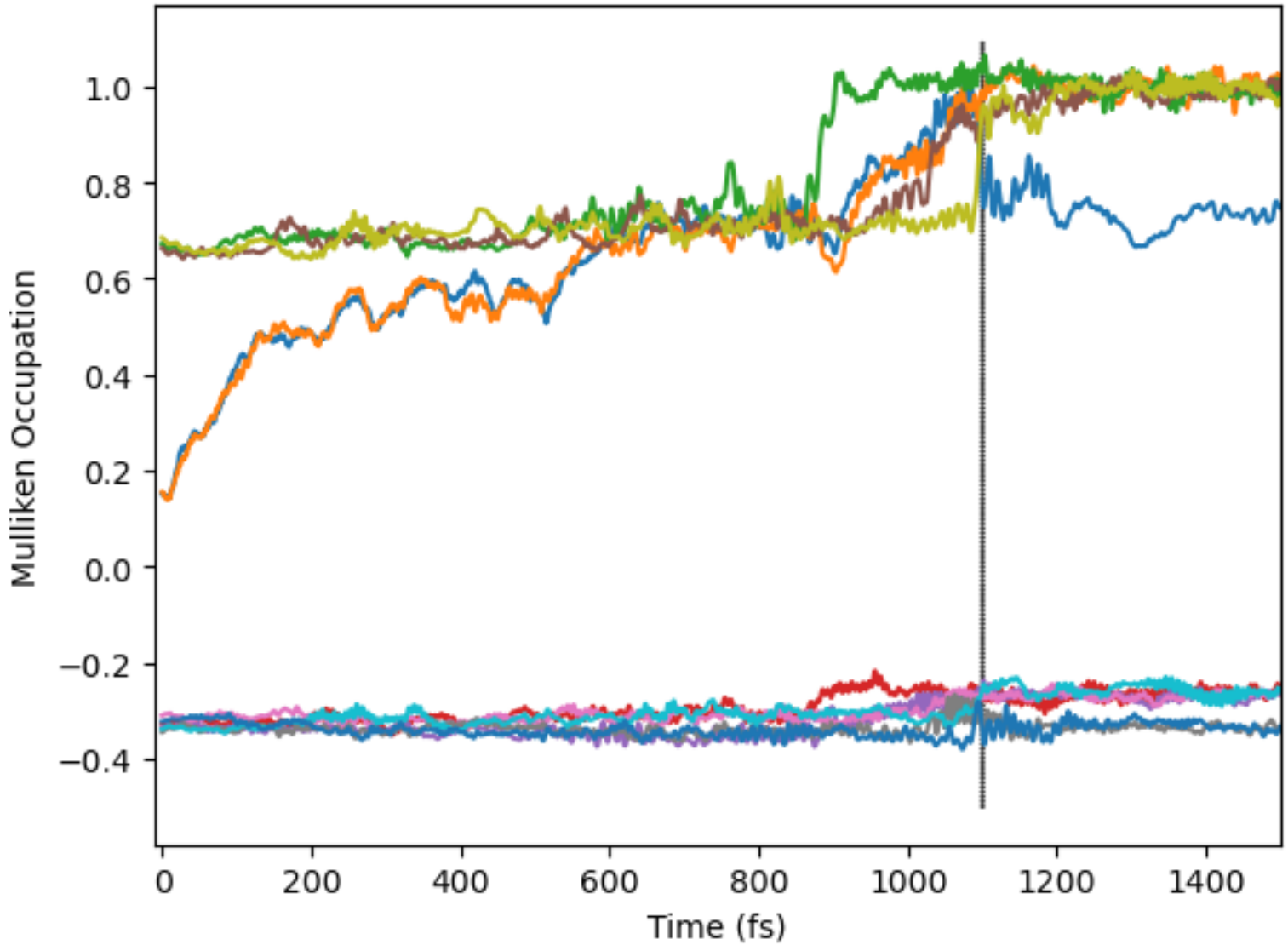



Mulliken Occupation
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
0
200
400
600
800
1000
1200
1400
Time (fs)




Mulliken Occupation
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
0
200
400
600
800
1000
1200
1400
Time (fs)



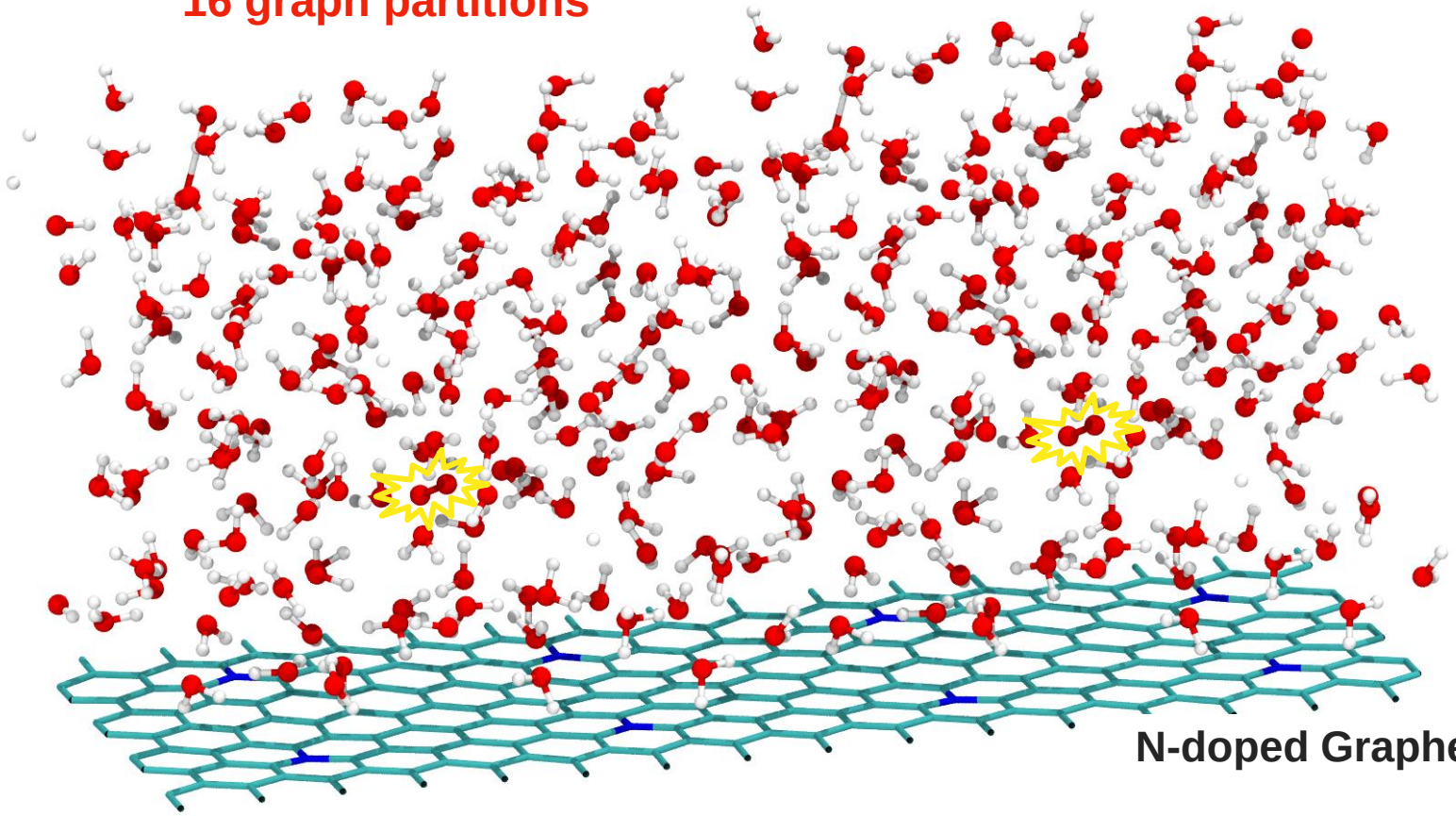

Mulliken Occupation
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
0
200
400
600
800
1000
1200
1400
Time (fs)

Graph-Based QMD: Oxygen Reduction Reaction

Non-locality is properly described in graph-based QMD

16 graph partitions

>1,000 atom QDM



N-doped Graphene + Water

DFTB (LANL 31 parameters)

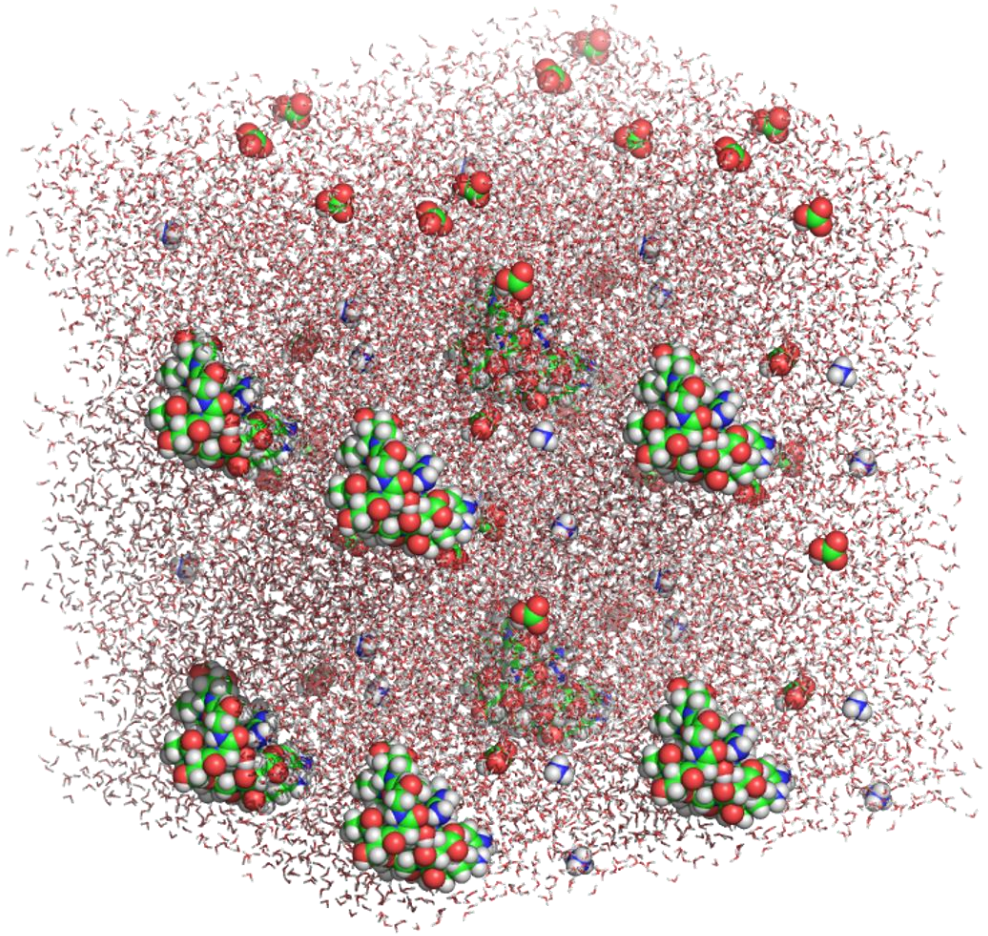
10 ps total, 0.1 fs timesteps

~30 hours

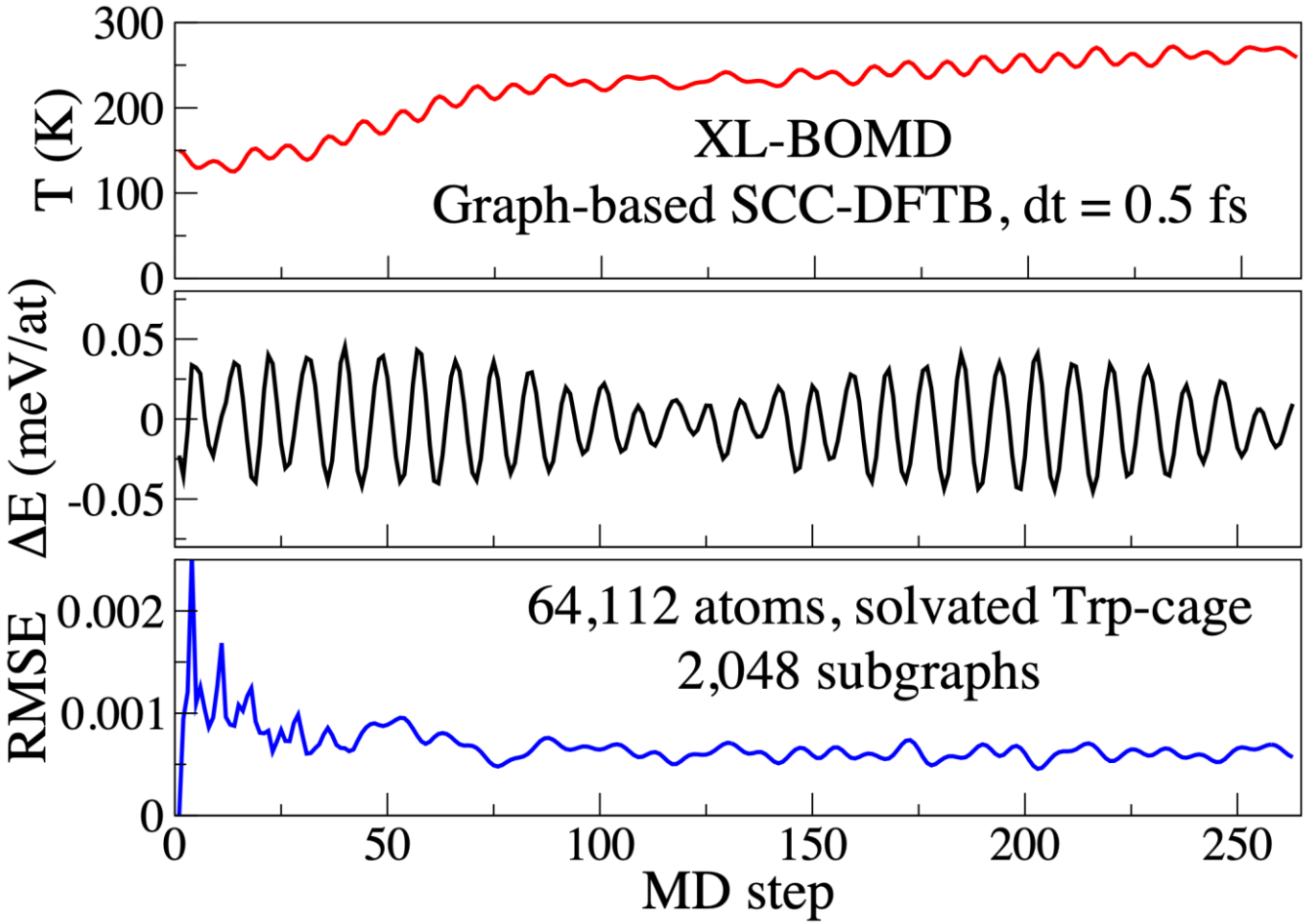
XL-BOMD

Rae C. Grove et al. Unpublished

64K atoms Trp-cage
+ ammonium bicarbonate solution
Partitioned in 2,048 subgraphs
running on 32 CPU nodes



Distributed graph-based shadow QMD for reactive & charge sensitive systems (SCC-DFTB, dt = 0.5 fs)



Negre, Wall, Niklasson, J. Chem. Phys. 158, 074108 (2023), *"Graph-based Quantum-response Theory and Shadow Born-Oppenheimer Molecular Dynamics"*

Summary

- **ML-based interatomic potentials:**
  - \+ Fast
  - \+ Accurate (in the domain of training data chemical space)
  - \- Ignore (usually) electronic structure (charge, multiplicity, non-local effects)
  - \- Require a lot of training data

- **SEQM methods:**
  - \+ Retain electronic structure formalism (charge, multiplicity, non-local effects)
  - ± Faster than *ab initio* but slower than ML
  - \- Compromised accuracy

- **ML+SEQM:**
  - \+ Higher accuracy than just ML or SEQM
  - \+ Require less training data than ML
  - \- Training is more challenging than for pure ML (backpropagation through SCF loop)
    - Special data preparation techniques are needed, e.g. Active Learning
- **Graph-based QMD**
  - \+ Improves scalability of quantum methods

Thank You for Your Attention!

Acknowledgements

LANL T-1

Sergei Tretiak

Ben Nebgen

Anders Niklasson

Kipton Barros

Nick Lubbers

Richard Messerly

Robert Stanton

Rae A. C. Grove

Nikita Fedik

LANL CSS

Nick Lubbers

Ying Wai Li

NVIDIA

Justin Smith

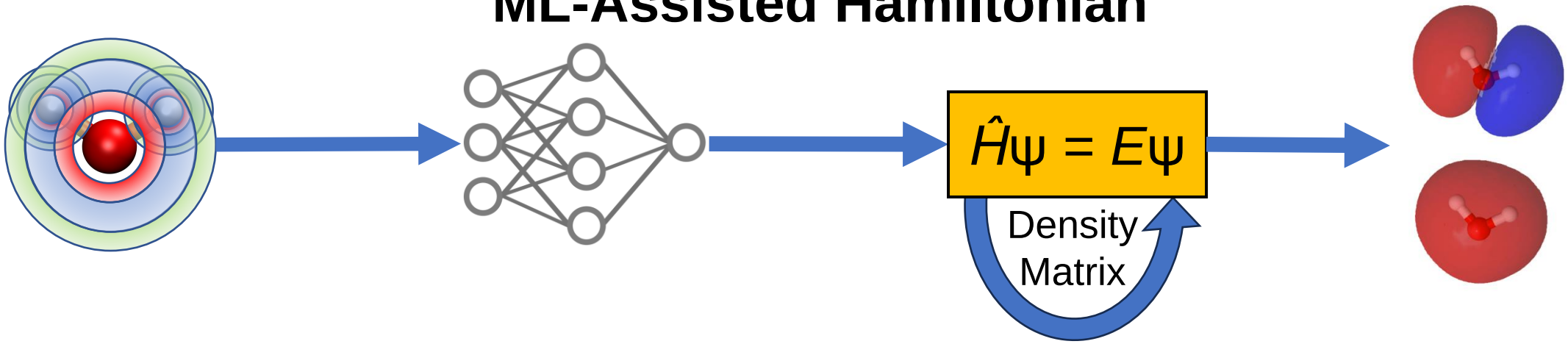
Guoqing Zhou

Center for Nonlinear Studies

LDRD LABORATORY DIRECTED RESEARCH & DEVELOPMENT

9/4/2014 41

ML-Assisted Hamiltonian



Linear Scaling via Graphs