Pacific Northwest
NATIONAL LABORATORY

Improving Spectroscopic Analysis using Machine Learning from Atomistic Simulations

June 18, 2019

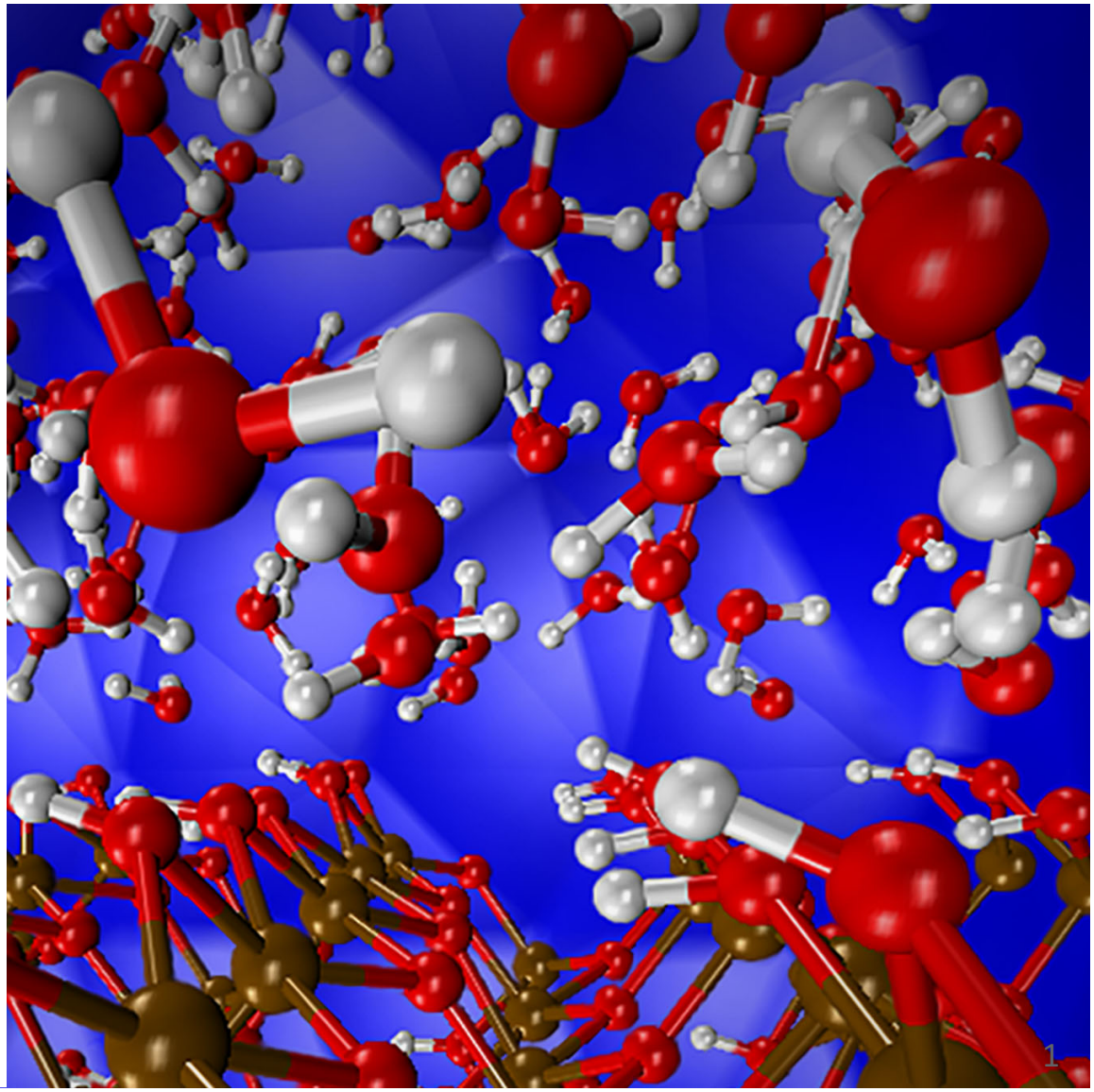
Eric Bylaska (PNNL)

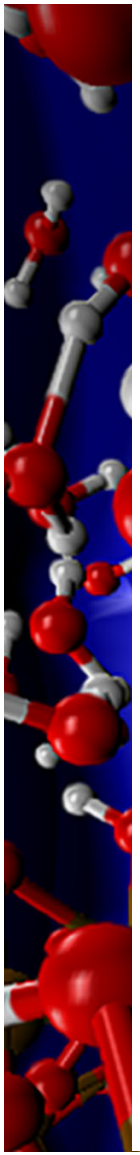
Raymond Atta-Fynn (UTA)

"A man dreams of a miracle and wakes up with loaves of bread"

Erich Maria Remarque

U.S. DEPARTMENT OF ENERGY BATTELLE





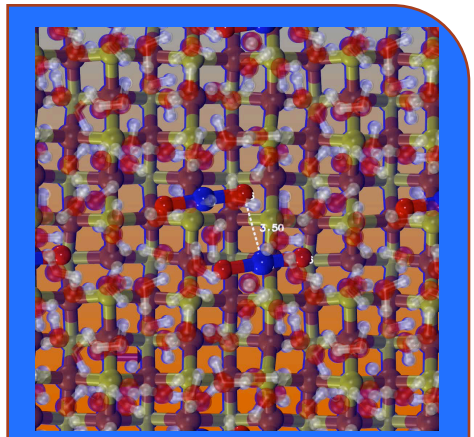
Pacific Northwest
NATIONAL LABORATORY

Outline

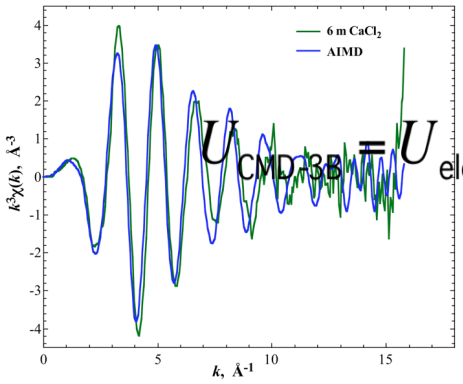
- Molecular Modeling in Geochemistry
- Ab initio Molecular Dynamics - Predictive Model for Molecular Geochemistry
- Solving dx/dt = F(x) faster
- Challenges and Opportunities in Fitting Molecular Dynamics Potentials with ML

Challenges for Molecular Modeling Of Geochemistry/Actinide Chemistry

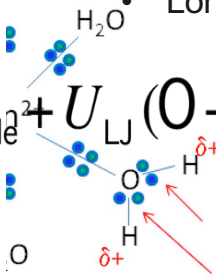
- Most (all) geochemistry problems begin at the nano-scale
  - Large numbers of atoms (1000's) and long simulation times are needed to simulate dilute solutions
- Molecular level experiments need interpretation
- Classical two and three body potentials are often used for these type of simulations
- Unfortunately, these classical potentials are often not very good at polarization, e.g. predicting the hydration shells of many aqueous metal (surface) species, or chemical bond breaking/making
- Ab Initio Molecular Dynamics (AIMD) avoids the use of such potentials but is only practical for ~~100's~~ 1000's of atoms.
- Still need Free energy pathways, PMFs
- Long time? Multiple Instances?



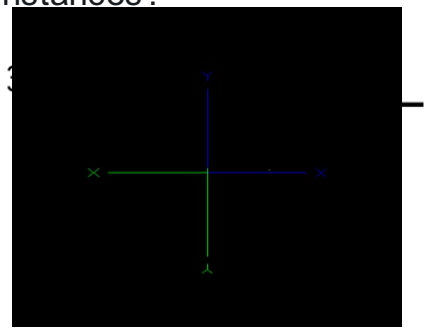
AIMD simulation of U(VI)-U(VI) dimerization on solvated Mackinawite surface(300°K)

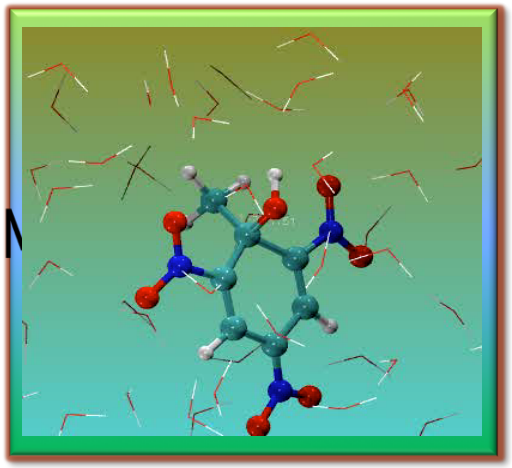


$$U_{CMD-3B} = U_{ele} + U_{LJ}(O-O) + U_{2B}(M$$



Electron pairs in O-H bonds are even closer to oxygen atoms than normal – results in a very large water dipole .



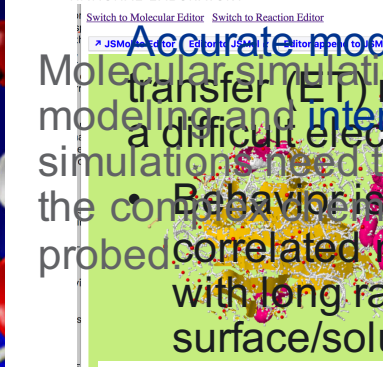


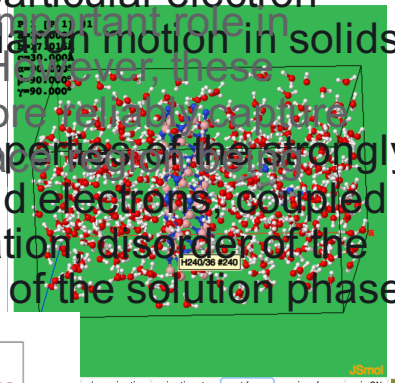
Development of Advanced Molecular Models for Geochemistry

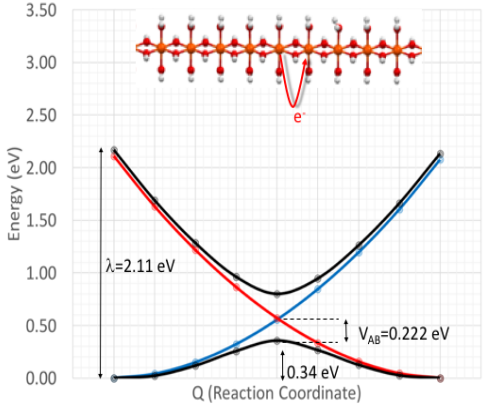
Pacific Northwest NATIONAL LABORATORY

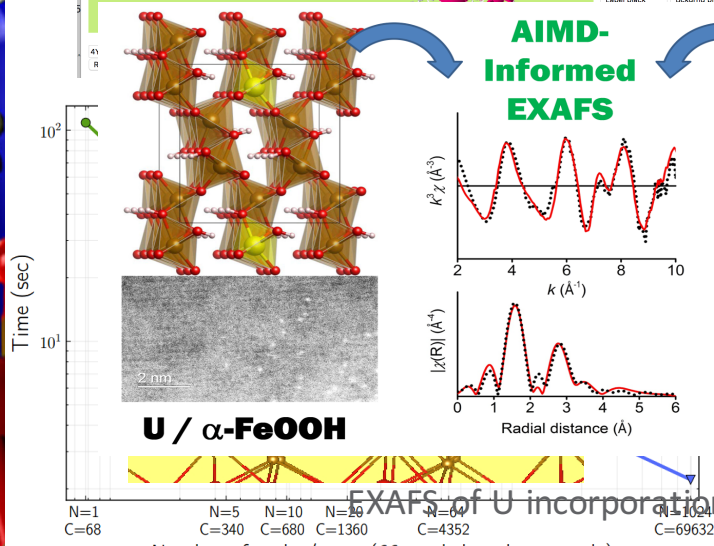
Molecular simulations is playing an increasingly important role in modeling and interpretation of spectroscopies. However, these simulations need to be constantly updated to more realistically capture the complexity in the real and interfacial systems being probed.

- Accurate modeling of redox processes, in particular electron transfer (ET) reactions at interfaces and polaron motion in solids, is a difficult electronic structure problem
  - Behavior is heavily dependent on the properties of the strongly correlated materials containing localized d electrons coupled with long range processes such as hydration, disorder of the surface/solution interface, the interaction of the solution phase

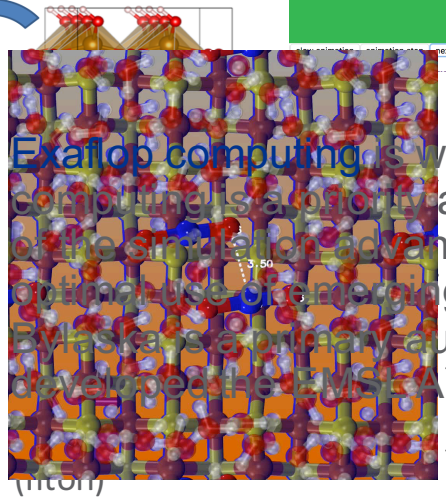




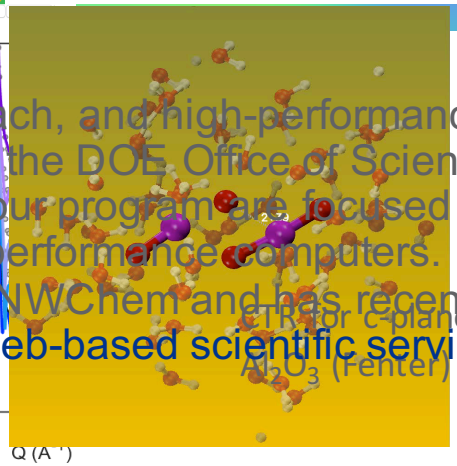


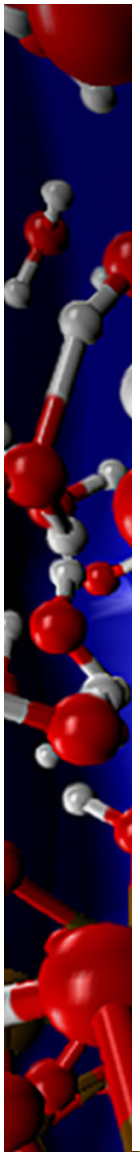


EXAFS of U incorporation (Hfon)



Exaflop computing is within reach, and high-performance computing is a priority area for the DOE Office of Science. All of the simulation advances of our program are focused on the optimal use of emerging high-performance computers. Bylaska is a primary author of NWChem and has recently developed the EMSL Arrows web-based scientific service.

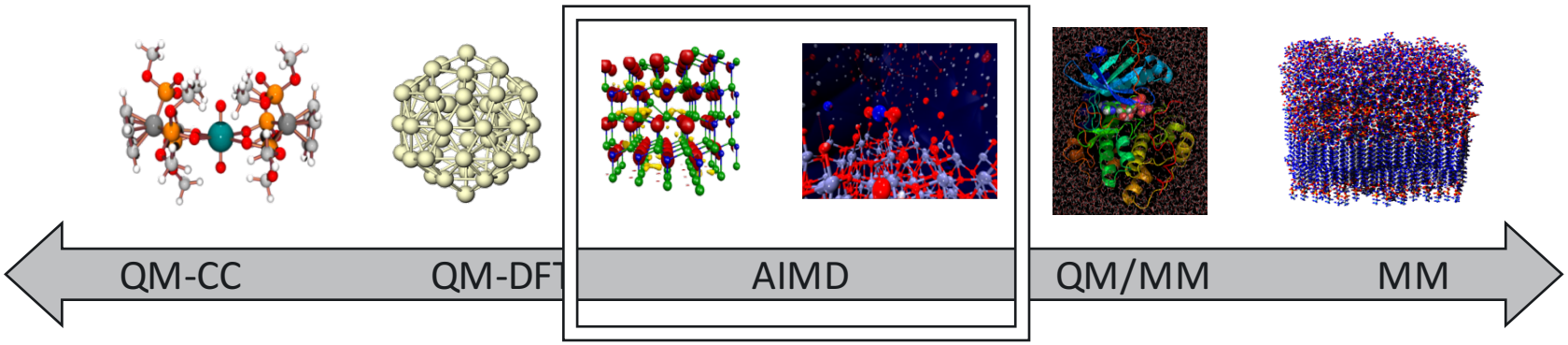




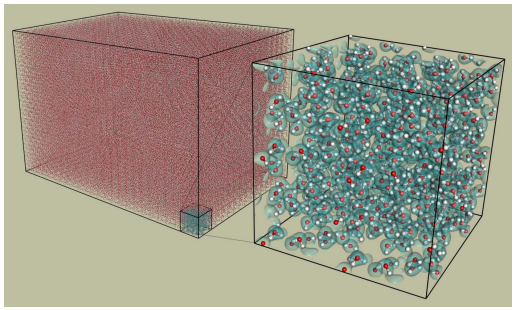
Pacific Northwest
NATIONAL LABORATORY

Ab initio molecular dynamics (AIMD)

Development emphasis is on providing parameter-free predictions applicable to a wide range of temperatures, pressures, and compositions



- 100-1000 atoms, uses plane wave basis
- >>10K atoms likely within 2 years
- Many FFTs and DGEMM operations
- "Meaty": Lots of FLOPs, but also bandwidth sensitive



Pacific Northwest NATIONAL LABORATORY

Basic Features of Ab Initio Molecular Dynamics

DFT Equations → MCSCF

$$H\psi_i = \varepsilon_i \psi_i$$

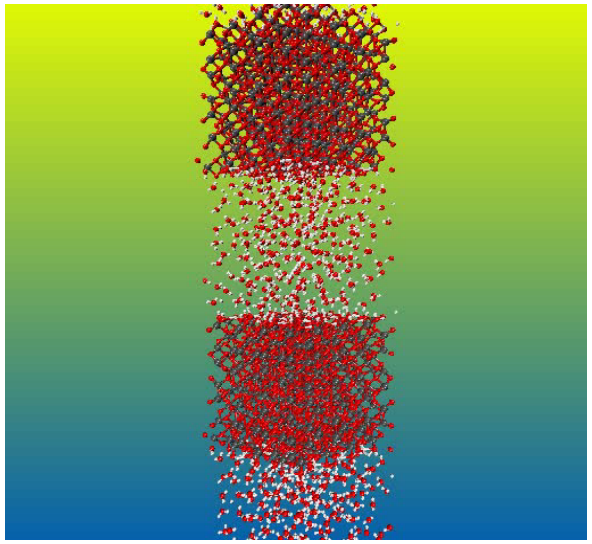
$$H\psi_i(\mathbf{r}) = \begin{pmatrix} -\frac{1}{2}\nabla^2 + V_l(\mathbf{r}) + \hat{V}_{NL} + V_H[\rho](\mathbf{r}) \\ +(1-\alpha)V_x[\rho](\mathbf{r}) + V_c[\rho](\mathbf{r}) \end{pmatrix} \psi_i(\mathbf{r}) - \alpha \sum_j K_{ij}(\mathbf{r})\psi_j(\mathbf{r})$$

CP dynamics: Ion and wavefunction motion coupled. Ground state energy μ=0

$$\mu\ddot{\psi}_i = H\psi_i - \sum_{i=1}^{N_e} \lambda_{ij}\psi_j$$

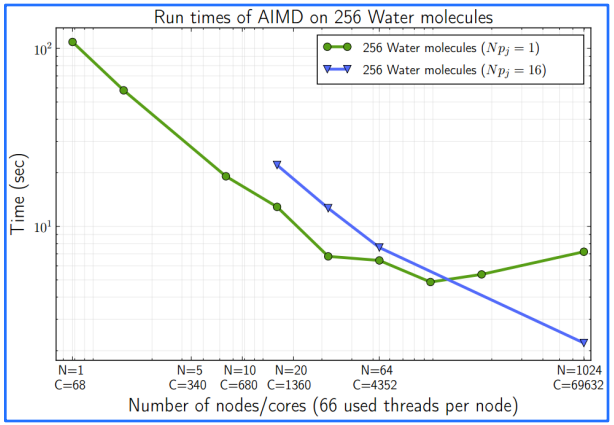
$$M_I \ddot{\mathbf{R}}_I = \mathbf{F}_I \qquad \mathbf{F}_I = \sum_{i=1}^{N_e} \left\langle \psi_i \middle| \frac{\partial H}{\partial \mathbf{R}_I} \middle| \psi_i \right\rangle$$

Want to do this in ~1second per step

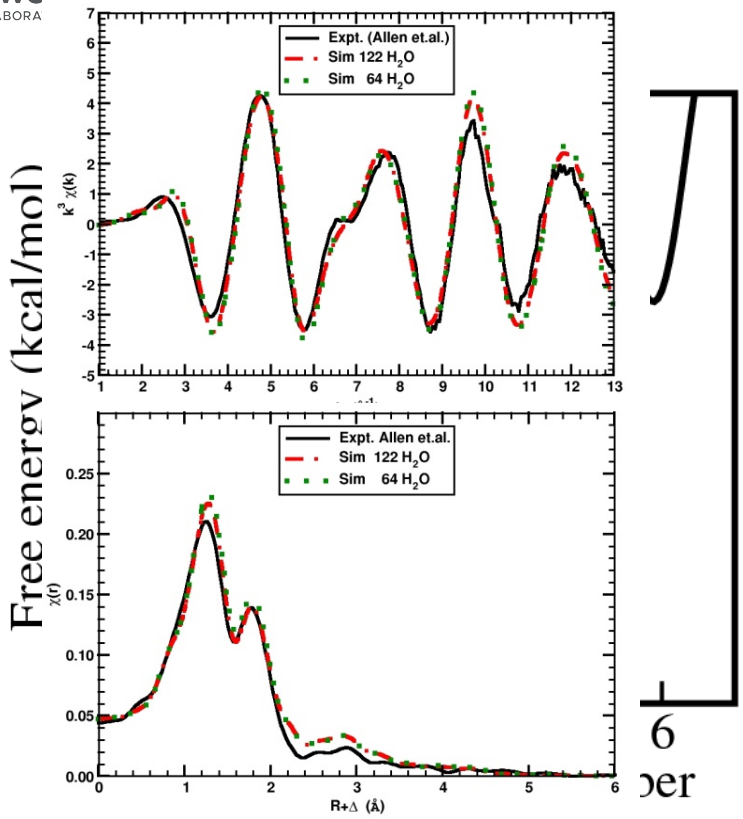


Plane-wave basis sets, pseudopotentials are used to solve PDEs

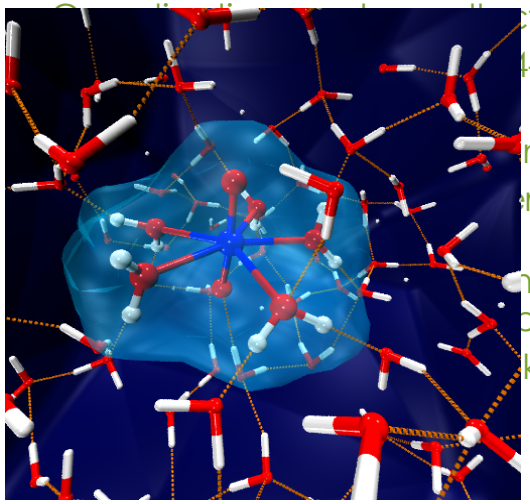
Doable but Expensive



Predictive Model for UO_2^{2+}(aq)

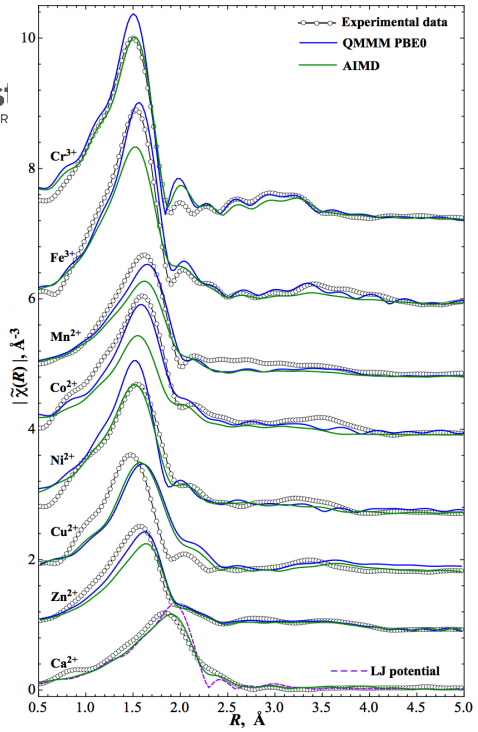


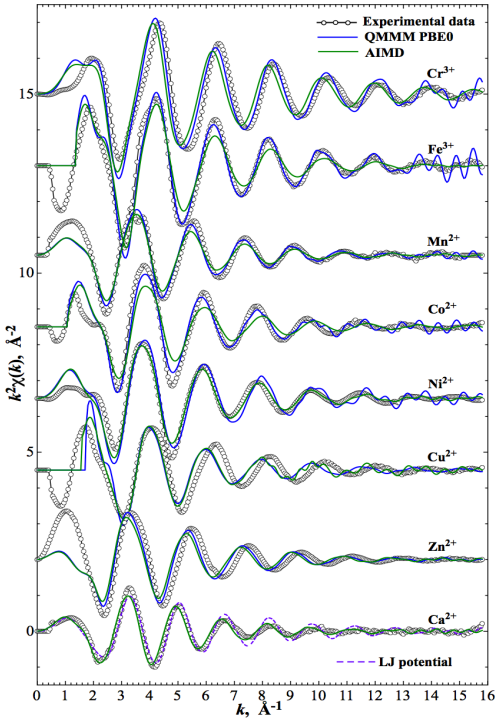
- Good EXAFS agreement but recent HEXS experiments suggest a 4-fold state is energetically nearby
- Results from Metadynamics

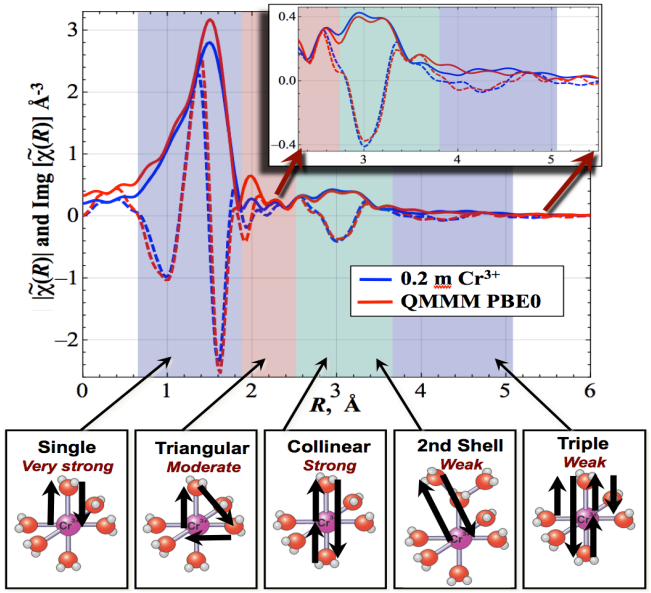


tive variable
-fold state by
mol
r $\Delta A_{5\rightarrow4}^{\ddagger}\approx4.7$
nort lifetime in
5-fold state.
kcal/mol

Pacific Northwest
NATIONAL LABORATORY





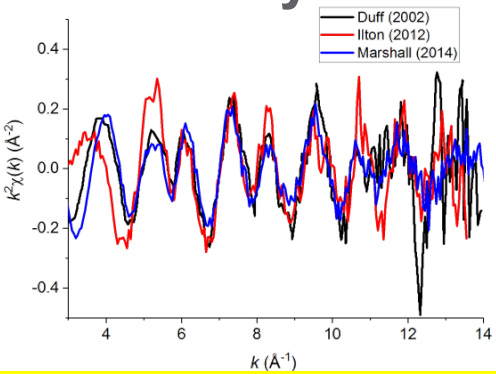


The generally excellent agreement of the 1st principle MD-XAFS simulation with the data. The scans are calculated by a parameter free method which can be implemented more efficiently than the use of empirical interactions suggesting that this **method can be used to interpret more XAFS spectra in more complex environments**.

Octahedral uranyl-like U(VI) incorporated in hematite is accommodated by protonated trans-corner Fe vacancies

Pacific Northwest NATIONAL LABORATORY

Shell by shell



UL_3 EXAFS are nearly identical, but three different interpretations:

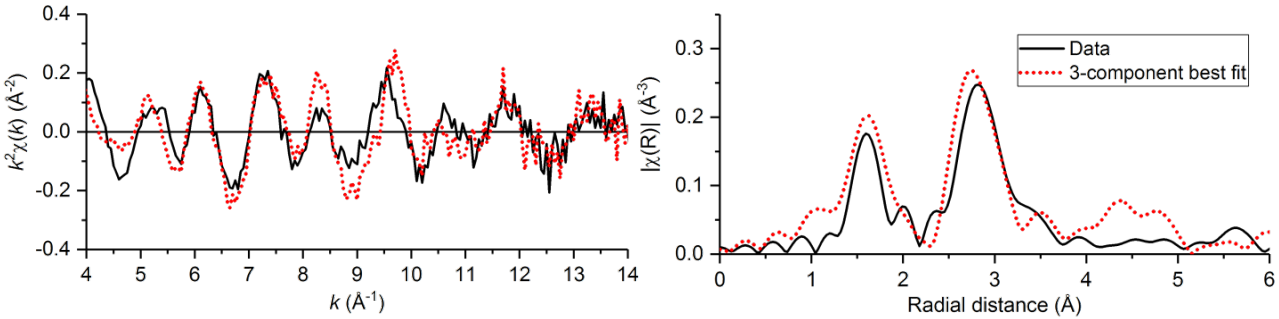
Duff *et al.* (2002)

Ilton *et al.* (2012)

Marshall *et al.* (2014)

All used very high Debye-Waller factors to model the first shell

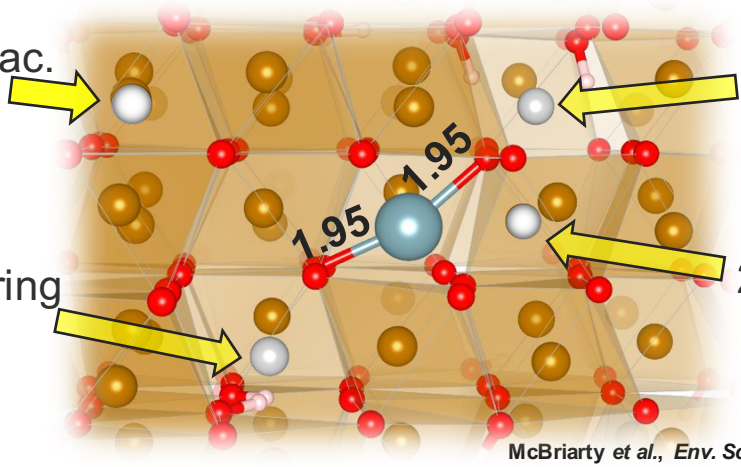
AIMD-informed EXAFS



44% adsorbed and 56% incorporated

30% bulk vac.

corner-sharing vac.

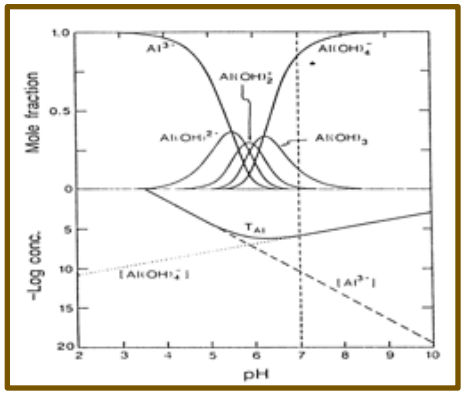


corner-sharing vac.

26% bulk vac.

McBriarty *et al.*, *Env. Sci. Technol.* 52 (2018) 6282
[Experimental data from Marshall et al. 2014]

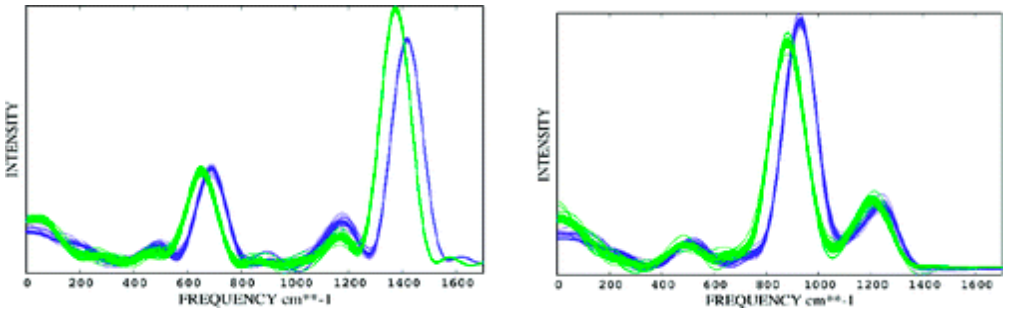
Future Transformative Applications in Geochemistry



Old Way – The diagram for the distribution of aluminum species was determined entirely from fitting thermodynamic data using an **assumed speciation scheme.**

Synergistic AIMD free energy simulations with XAS spectra can be used to determine the solute structure of environmentally important species in solutions as a function of TPX. Used to develop chemically and thermodynamically highly accurate solution models with exceptional extrapolation properties in TPX.

- Strategies to search configuration space must be developed.



AIMD simulations are already find relevant solvent structures in a first principles approach to calculating isotope fractionation based on harmonic frequencies.

- If the simulations were faster, it would be possible to directly calculate the fractionation factors using quantum dynamics.

Other Spectroscopies: Algorithms for IR and Raman spectra interpretation

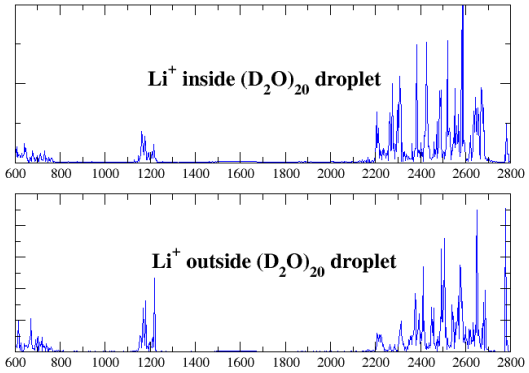
- AIMD simulations have the potential to significantly improve the molecular interpretation of IR and Raman spectroscopies *along recent variants of them where they are combined with instrumentation such as AFM.*
- In principle, AIMD analysis methods can easily be extended to IR and Raman (i.e., AIMD-IR and AIMD-Raman).

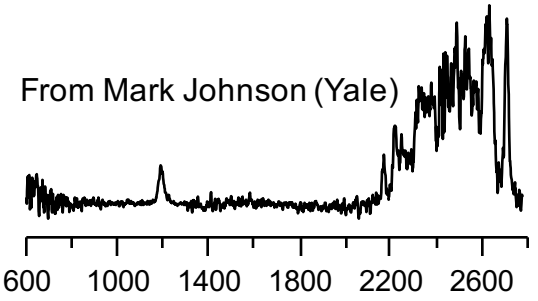
$$Absorption(\omega) = \frac{1}{6cVn(\omega)\varepsilon_0 k_B T}\int_{-\infty}^{\infty} e^{i\omega t}\left\langle \dot{\boldsymbol{P}}(0)\middle|\dot{\boldsymbol{P}}(t)\right\rangle dt$$

$$P_\mu = \frac{2|e|}{|b_\mu|} Im \log\det Q^\mu \quad \text{where} \quad Q^\mu_{i,j} = \left\langle \psi_i \middle| e^{ib_\mu \cdot r} \middle| \psi_j \right\rangle$$

$$\alpha_{\mu\nu} = \frac{2|e|}{|b_\mu|} Im \left[\sum_{i,j}\left(\left\langle \psi_i^{(v)} \middle| e^{ib_\mu \cdot r} \middle| \psi_j \right\rangle + \left\langle \psi_i \middle| e^{-b_\mu \cdot r} \middle| \psi_j^{(v)} \right\rangle\right)[Q^v]^{-1}_{j,i}\right]$$

- Improving standard vibrational analysis
  - Larger and more complex systems
  - Better sampling
  - Longer time scales for single molecule

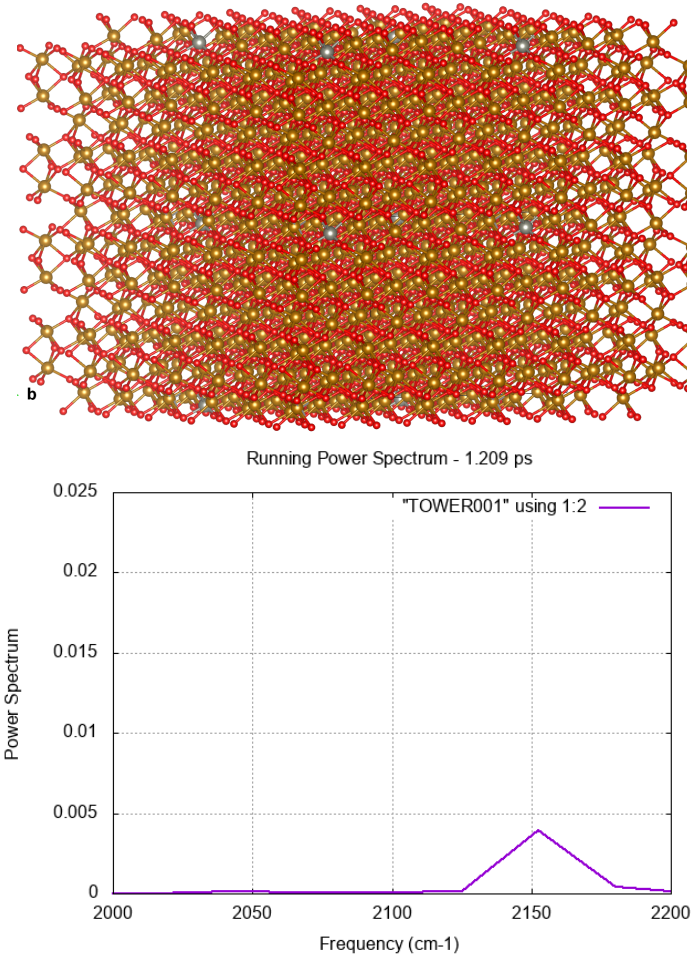


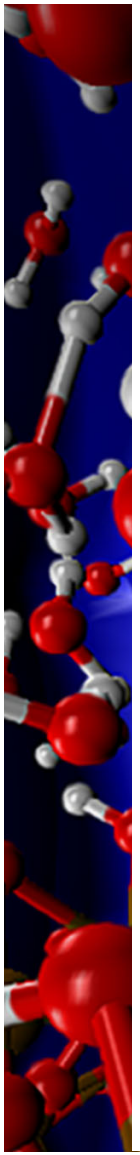


Pacific Northwest NATIONAL LABORATORY

AIMD simulations will need long trajectories to perform autocorrelation functions

- Ab initio molecular dynamics (AIMD) has transformed how spectroscopic measurements from advanced light sources are analyzed, such as Raman, EXAFS, CTR, XANES, etc.
- Advanced HPC algorithm development has made the first-principles analysis of advanced light sources possible for the first time
- However, the computation cost of AIMD is prohibitive for many projects and other possible predictive analysis, e.g. isotope fractionation using quantum dynamics



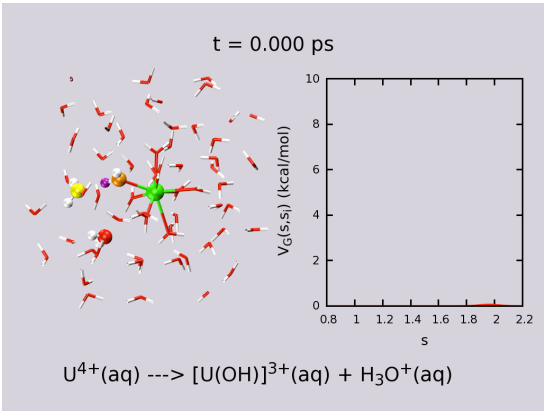


Pacific Northwest NATIONAL LABORATORY

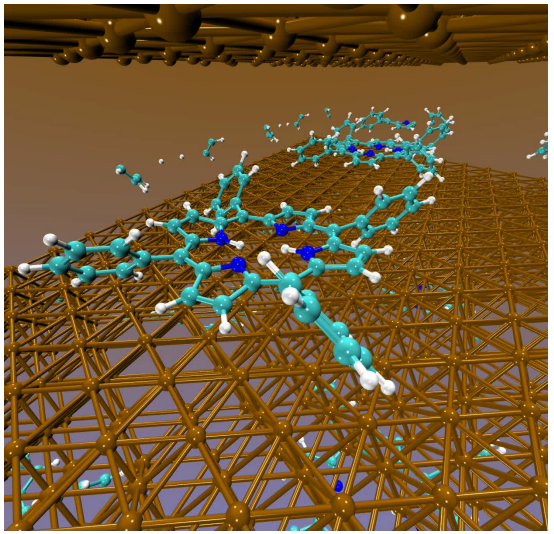
Strong Scaling is Key

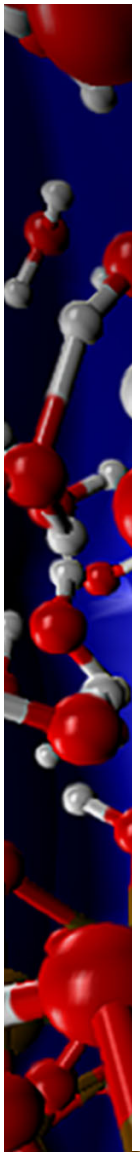
- 20 picoseconds of simulation time ≈ 200,000 steps
  - 1 sec/step = 2-3 days simulation time
  - 10 sec/step = 23 days simulation time
  - 13 sec/step = 70 days simulation time
- Mesoscale phenomena at longer time scales
  - Assume 1 sec/step
  - 100 psec = 10-15 days simulation time
  - 1 nsec = 100 - 150 days simulation time
- Strong scaling required to reduce time per time step as much as possible
  - At least below 1sec/step

Free Energy Simulations



Surface spectroscopies

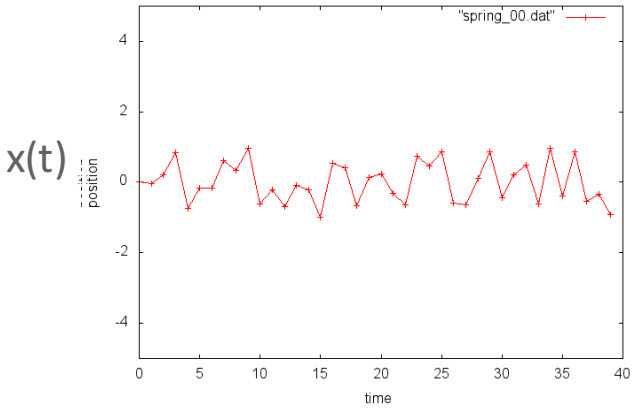




Pacific Northwest NATIONAL LABORATORY

Possible Solution: Parallel in Time

- Increasing the time step (dt) in time integration quickly becomes unstable
- One approach to bridging these temporal scales is the development of algorithms which parallelize over time, i.e. parallel in time algorithms
- The central philosophy of parallel in time integration is to start with a guess for the trajectory over some fixed time interval and then attempt to relax it until it approximates the “true” trajectory.



t = time

Trajectory for a simple spring (K=1,x0=1,v0=0)

Increasing time step

Can this be parallelized????

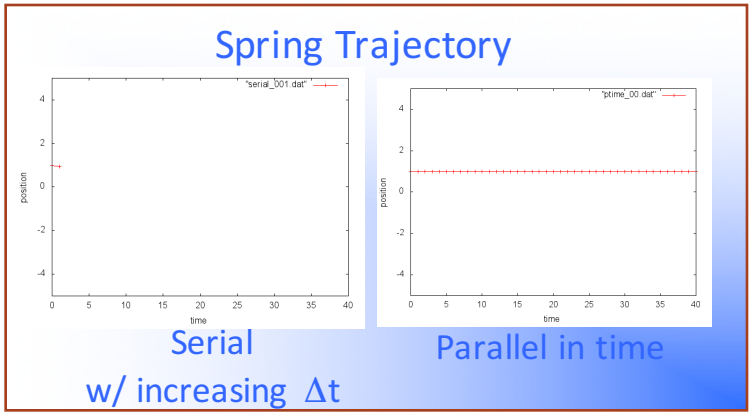
Pacific Northwest NATIONAL LABORATORY

Parallel In Time Algorithms Without Using Approximate Models: Fixed Point Parallel in Time Algorithms

These algorithms transform standard forward substitution time integration solvers, i.e. $x_{i+1} \leftarrow f(x_i)$, into fixed-point root problems

F(X) = 0 or

$$\begin{pmatrix} x_1 - f(x_0) \\ x_2 - f(x_1) \\ \vdots \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \end{pmatrix}$$



Can be solved using a variety of optimization techniques, including preconditioned fixed-point, quasi-Newton, and preconditioned quasi-Newton optimization methods (i.e. solve **F(G-1(X))=0**). These algorithms can be parallelized since the evaluation of the trial root function **F(X)** can be done in parallel. See Bylaska et al. Extending Molecular Simulation Time Scales: Parallel in Time Integrations for High-Level Quantum Chemistry and Complex Force Representations. *J. Chem. Phys.* **2013**, *139*, 074114. DOI: 10.1063/1.4818328.

These algorithms are particularly useful for diffusion based mescoscale models

- Note phase field $\Delta t \sim \Delta X^4$

Pacific Northwest
NATIONAL LABORATORY

Parallel in Time: Fixed Point Iteration

The serial solution to time integration,

$$x_{i+1} = f(x_{i-1}) \quad x_0 = x_0$$

with initial condition is

$$X_{trajectory} = \begin{bmatrix} f(x_0) \\ f(f(x_0)) \\ f(f(f(x_0))) \\ f(f(f(f(x_0)))) \end{bmatrix}$$

Using column vector to store each step in the time iteration from i=1,4

This equation can also be solved by a fixed-point iteration (or more advanced root finding algorithms) over the whole path or trajectory

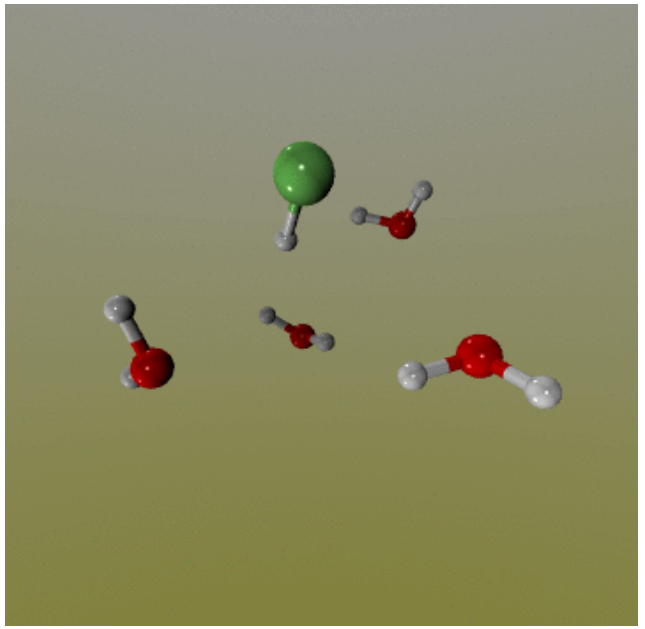
$$X^{(k+1)} = X^{(k)} - F\left(X^{(k)}\right) \quad \text{or}$$

$$\begin{bmatrix} x_1^{(k+1)} \\ x_2^{(k+1)} \\ x_3^{(k+1)} \\ x_4^{(k+1)} \end{bmatrix} = \begin{bmatrix} x_1^{(k)} \\ x_2^{(k)} \\ x_3^{(k)} \\ x_4^{(k)} \end{bmatrix} - \begin{bmatrix} x_1^{(k)} - f(x_0) \\ x_2^{(k)} - f(x_1^{(k)}) \\ x_3^{(k)} - f(x_2^{(k)}) \\ x_4^{(k)} - f(x_3^{(k)}) \end{bmatrix}$$

Parallelized by distributing work over rows

Real Example: HCl+4H_2O MP2 AIMD Simulations

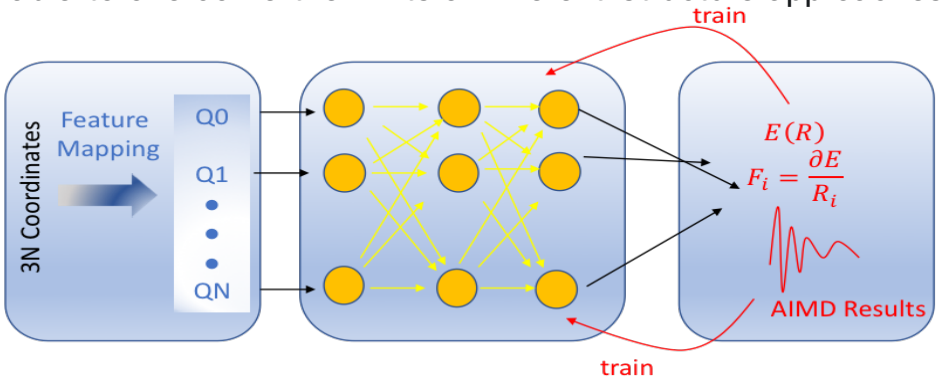
Algorithm	Method	Preconditioner	M	$K^a_{average}$	Speedup	$(M/K_{average})$
QN	MP2/6-31G*		10	2.6	3.7	(3.8)
QN	MP2/6-31G*		20	3.4	5.6	(5.9)
QN	MP2/6-31G*		25	4.0	5.8	(6.3)
QN	MP2/6-31G*		50	6.0	7.7	(8.3)
QN	MP2/6-31G*		100	11.0	8.0	(9.1)
QN	MP2/6-31G*		150	17.0	8.2	(8.8)
PQN	MP2/6-31G*	HF/3-21G	10	2.0	3.7	(5.0)
PQN	MP2/6-31G*	HF/3-21G	20	3.0	4.1	(6.7)
PQN	MP2/6-31G*	HF/3-21G	25	3.0	4.6	(8.3)
PQN	MP2/6-31G*	HF/3-21G	50	4.0	4.8	(12.5)
PQN	MP2/6-31G*	HF/3-21G	100	5.0	4.8	(25.0)
PQN	MP2/6-31G*	HF/3-21G	150	5.0	5.3	(30.0)
QN	MP2/6-311+G*		10	2.5	3.3	(4.0)
QN	MP2/6-311+G*		20	3.5	4.9	(5.7)
QN	MP2/6-311+G*		25	4.0	5.9	(6.3)
QN	MP2/6-311+G*		50	6.0	7.7	(8.3)
QN	MP2/6-311+G*		100	11.0	8.2	(9.1)
QN	MP2/6-311+G*		150	17.0	7.9	(8.8)
PQN	MP2/6-311+G*	HF/3-21G	10	2.0	3.9	(5.0)
PQN	MP2/6-311+G*	HF/3-21G	20	3.0	5.3	(6.7)
PQN	MP2/6-311+G*	HF/3-21G	25	3.0	6.2	(8.3)
PQN	MP2/6-311+G*	HF/3-21G	50	3.5	10.2	(14.3)
PQN	MP2/6-311+G*	HF/3-21G	100	5.0	14.1	(25.0)
PQN	MP2/6-311+G*	HF/3-21G	150	5.0	14.3	(30.0)

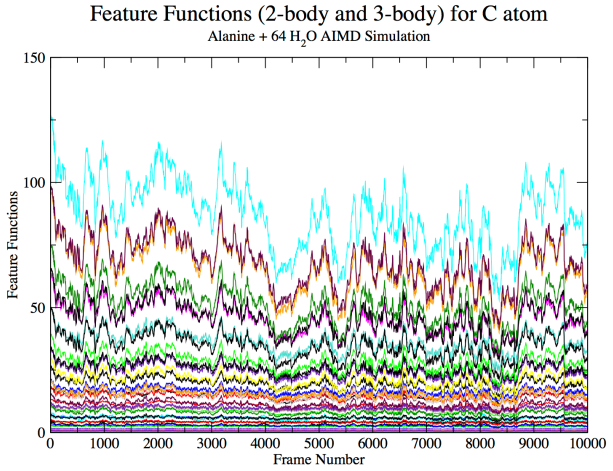


Ideal Speedups of 30 seen, and FAS methods show further promise. However, preconditioners really help!

New machine learning strategies for improving spectroscopic analysis using atomistic simulations

- *This development proposes to speed up the AIMD/spectroscopy analysis* by generating and using machine-learned atomistic potentials on the fly to speed up the sampling used in spectroscopic analysis, while maintaining the accuracy of the full AIMD analysis. In addition, using machine learning to regress AIMD into effective molecular dynamics potentials has the potential to enable quantum dynamics approaches for isotope fractionation that are able to overcome the limits of inherent structure approaches.

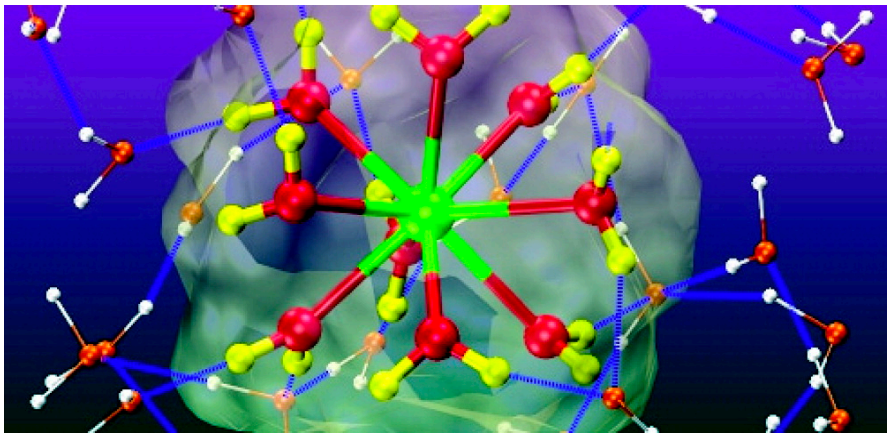




In these approaches, 2-body and 3-body functions are used to define the feature mapping or basis that is input into the feed-forward neural network. These functions are similar to 2- and 3-body molecular dynamics potentials but with varying parameters, e.g.,

$$q_i^{(\eta_\alpha)} = \sum_{\substack{j=1 \\ i \neq j}}^{localatoms} e^{-\eta_\alpha (R_{ij} - R_s)^2} f_c(R_{ij}) \qquad q_i^{(\eta_\alpha)} = 2^{1-\xi} \sum_{\substack{j,k=1 \\ i \neq j, i \neq k, j \neq k}}^{localatoms} (1 - \lambda cos\theta_{ijk})^{\xi} e^{-\eta_\alpha \left(R_{ij}^2 + R_{ik}^2 + R_{jk}^2\right)} f_c(R_{ij})\, f_c(R_{ik})\, f_c(R_{jk}).$$

Traditional Fitting of MD Potentials

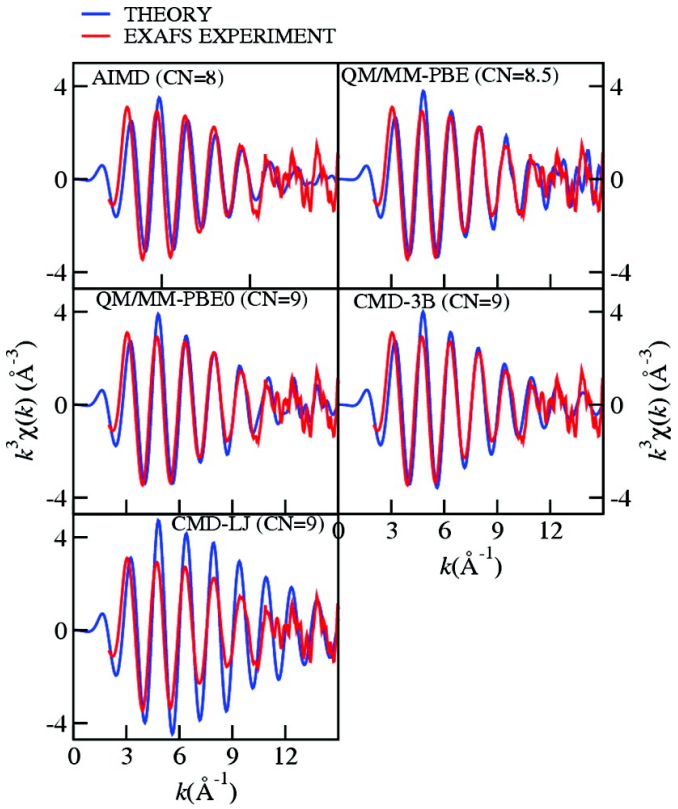


Postulate a potential form

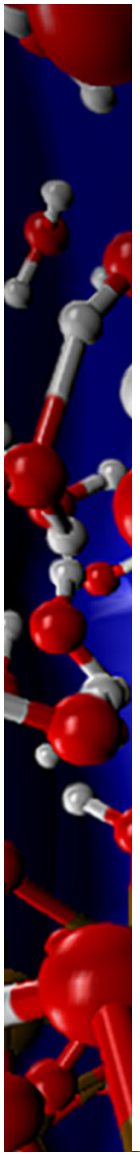
$$U_{2B}(\mathrm{Cm}^{3+}\!-\!\mathrm{H_2O}) =$$

$$\sum_{\alpha=\mathrm{O,H1,H2}} \left[A_{\mathrm{Cm}\alpha} \exp(-B_{\mathrm{Cm}\alpha} r_{\mathrm{Cm}\alpha}) + \frac{C_{\mathrm{Cm}\alpha}}{r_{\mathrm{Cm}\alpha}{}^4} + \frac{D_{\mathrm{Cm}\alpha}}{r_{\mathrm{Cm}\alpha}{}^6} \right]$$

$$U_{3B}(\mathrm{O}-\mathrm{Cm}^{3+}-\mathrm{O}) = \alpha \exp(-\beta r_1 - \beta r_2 - \gamma r_3)$$



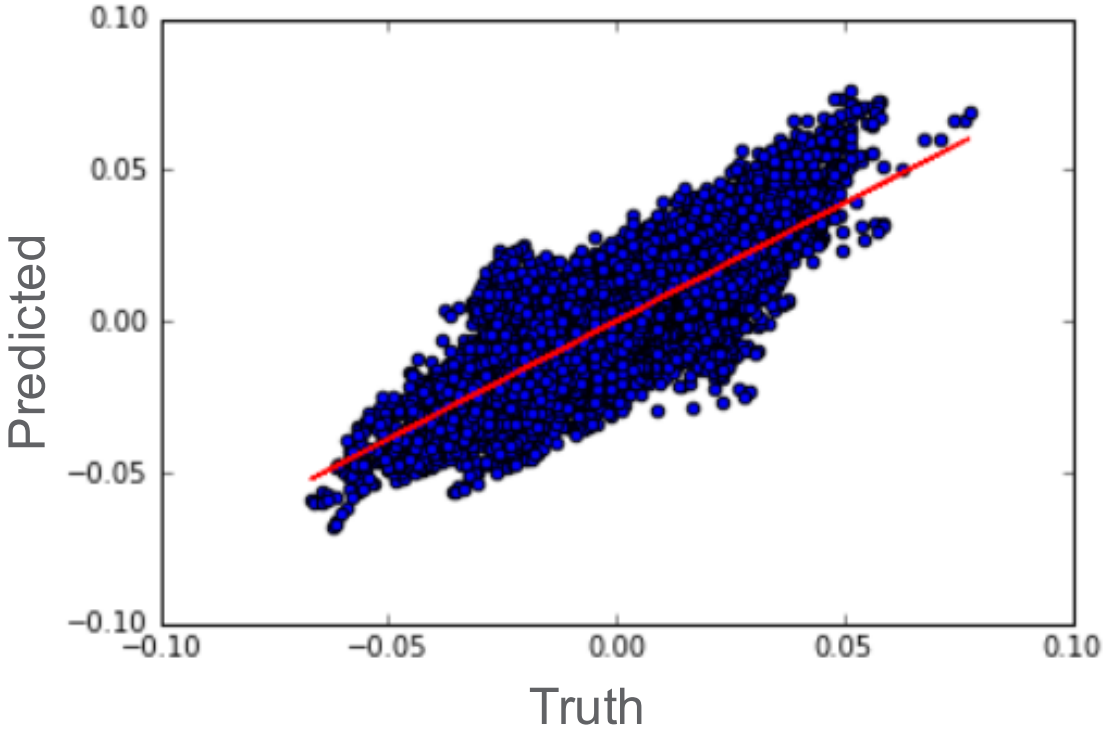
And use non-linear regression (mrqmin) of simulation/experimental data to find parameters



Pacific Northwest NATIONAL LABORATORY

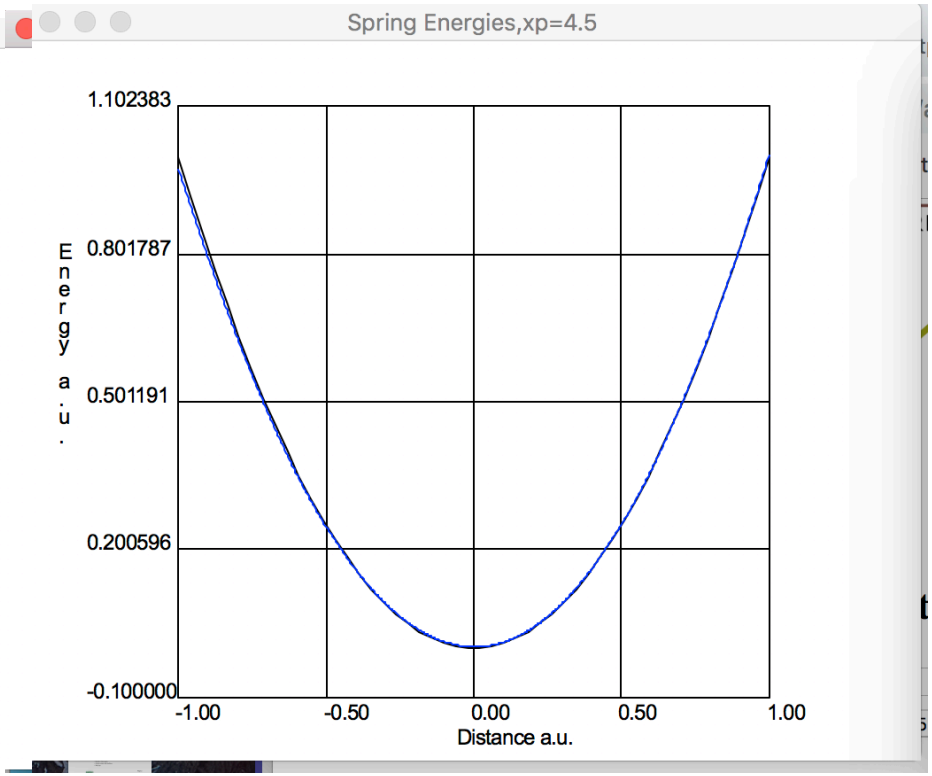
First Attempt – Fitting forces with TensorFLow

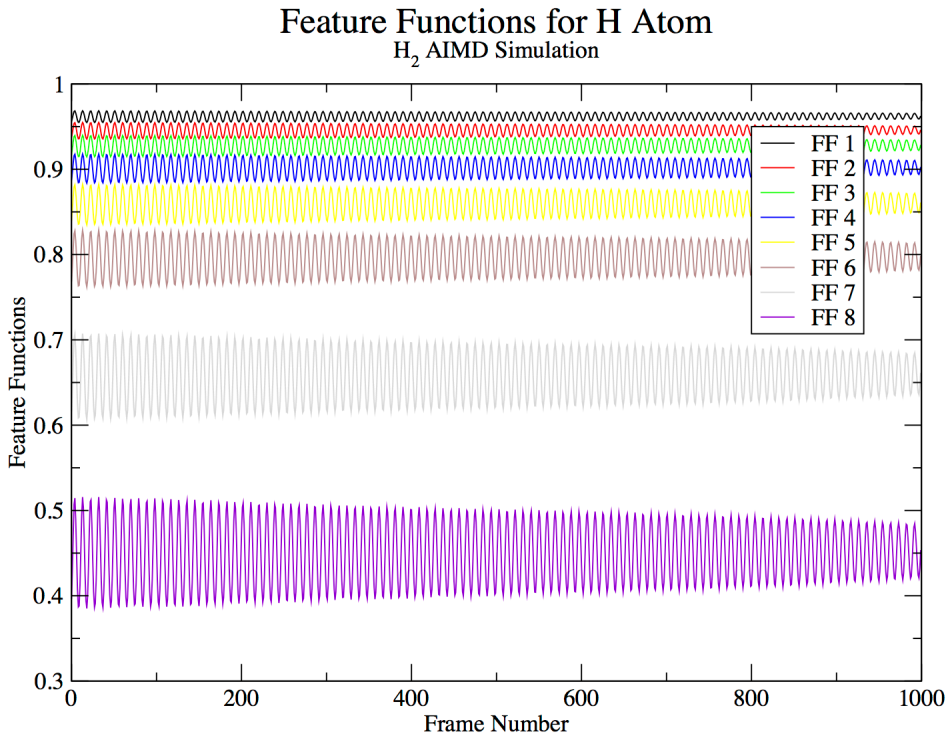
We were able to reduce the mean absolute error w down to 0.006, which represents roughly 5% of the range of the force. However, this error is too large and , additional research into choosing feature functions as well as the use of longer AIMD simulations is required for this method to be truly competitive.



Back to the Basics

- 1x20x40x1 – Adam solver

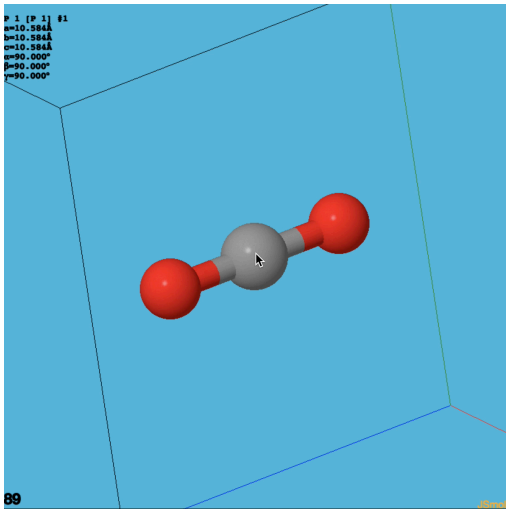


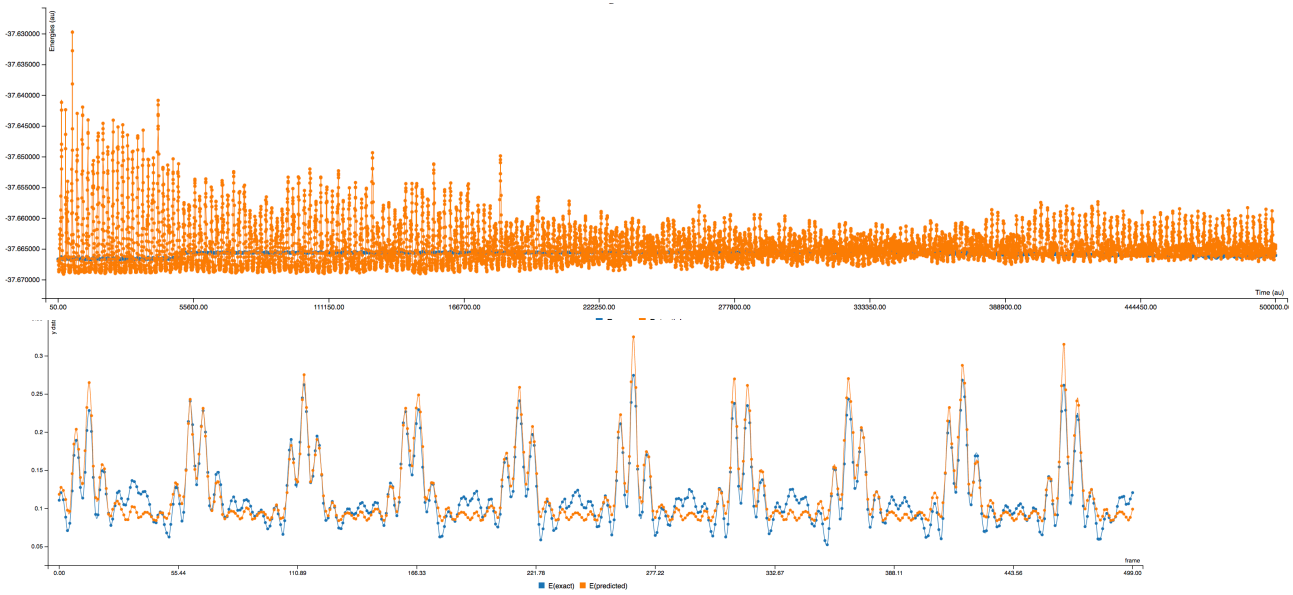


ML Potential for CO_2 Molecule

Pacific Northwest NATIONAL LABORATORY

- 3x60x120x1
- Described by a 3d space
- 10,000 points ~= 21^3





While the ML approach looks feasible and automatable there are several challenges going forward

- High-dimensional spaces
- Charges, dipoles, polarization, bond breaking, …

More model input will probably be needed for this approach to be predictive

- Start with fitted MD potentials and correct with ML

Pacific Northwest NATIONAL LABORATORY

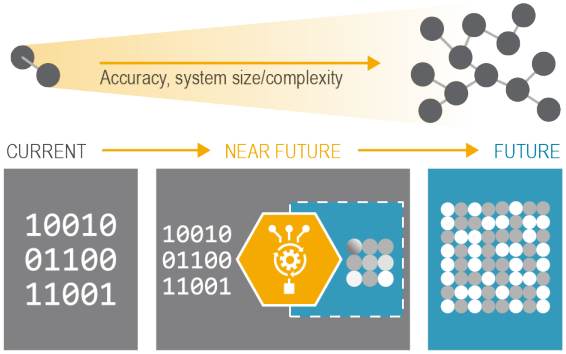
Computing Horizons – The Next Wave

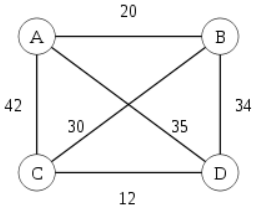
$$e^{-iHt} = \prod_{j=1}^{M} e^{-ih_j P_j t} + O(M^2 \max_{j,k} \|[P_j, P_k]\| \max |h_j|^2 t^2).$$

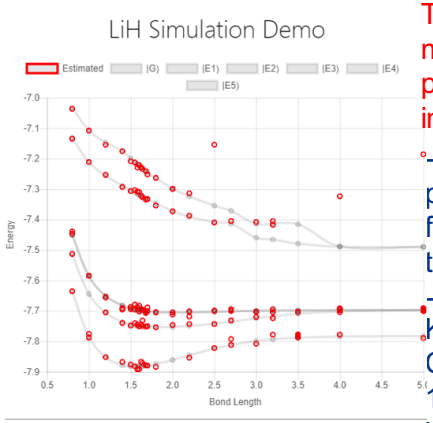
- Classical computing is becoming power bound
- All DOE supercomputers are going to have more GPUs than they know what to do with…..
- Planning for many cycles to be used for ML

ECP EXASCALE COMPUTING PROJECT

Exact quantum chemistry not computable on classical computers

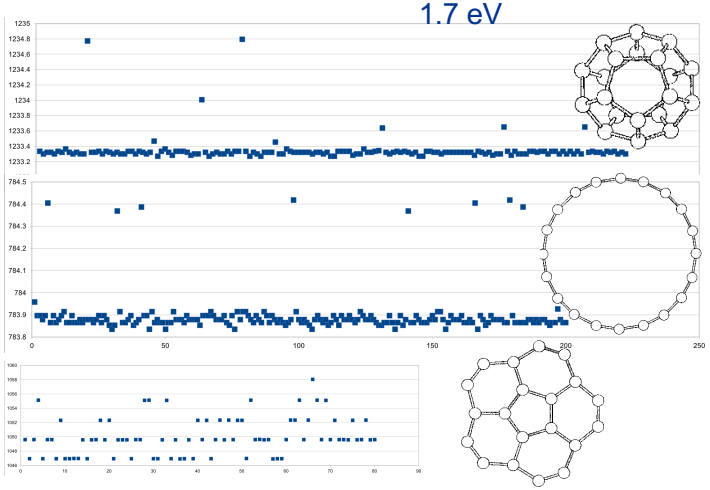


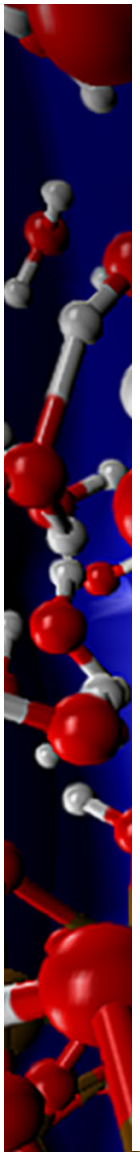




The smaller the molecule in a published paper the greater its importance

- Trotterization predicts the fullerene lower than the ring by 1.75 eV
- Taylor, Bylaska, Kawai, Weare – CCSD(T) early 1990s, Fullerene lower than ring by 1.7 eV



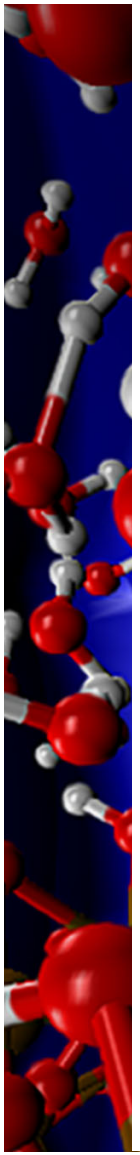


Pacific Northwest
NATIONAL LABORATORY

Conclusions

- Computation chemistry methods are becoming truly predictive, rather then just rationalizing existing knowledge. Synergistic use of AIMD and spectroscopies is already changing many spectroscopies.
- New machine-learning methods for developing MD potentials will support longer dynamical simulations and improved phase sampling methods, which will provide new models of chemical mechanisms in complex brines, defected solids and interfaces.
  - Inverse modeling expertise from the chemistry and condensed matter communities needs to be better incorporated.
- All these developments will be available to the wider geochemistry, chemistry, and materials communities via inclusion in the NWChem (NWChemEx) program or the EMSL Arrows scientific service.

Acknowledgements This work was primarily supported by the DOE BES Geosciences program. Also thanks to DOE BER EMSL, DOE ECP, DOE BES Chemistry. A portion of this research was performed using the NERSC Computing Facility at LBL. EMSL is a national scientific user facility sponsored by the Department of Energy's Office of Biological and Environmental Research located at Pacific Northwest National Laboratory, DE-AC06-76RLO 1830. We also acknowledge EMSL for supporting the development of NWChem. The Pacific Northwest National Laboratory is operated by Battelle Memorial Institute.



Pacific Northwest
NATIONAL LABORATORY

Thank you