Breaking the Million-Electron and 1 EFLOP/s (FP64) Barriers

Biomolecular-Scale *Ab Initio* Molecular Dynamics Using MP2 Potentials

GIUSEPPE M. J. BARCA

Associate Professor of Digital Innovation & HPC
School of Computing, University of Melbourne
Head of Research, QDX Technologies

QDX

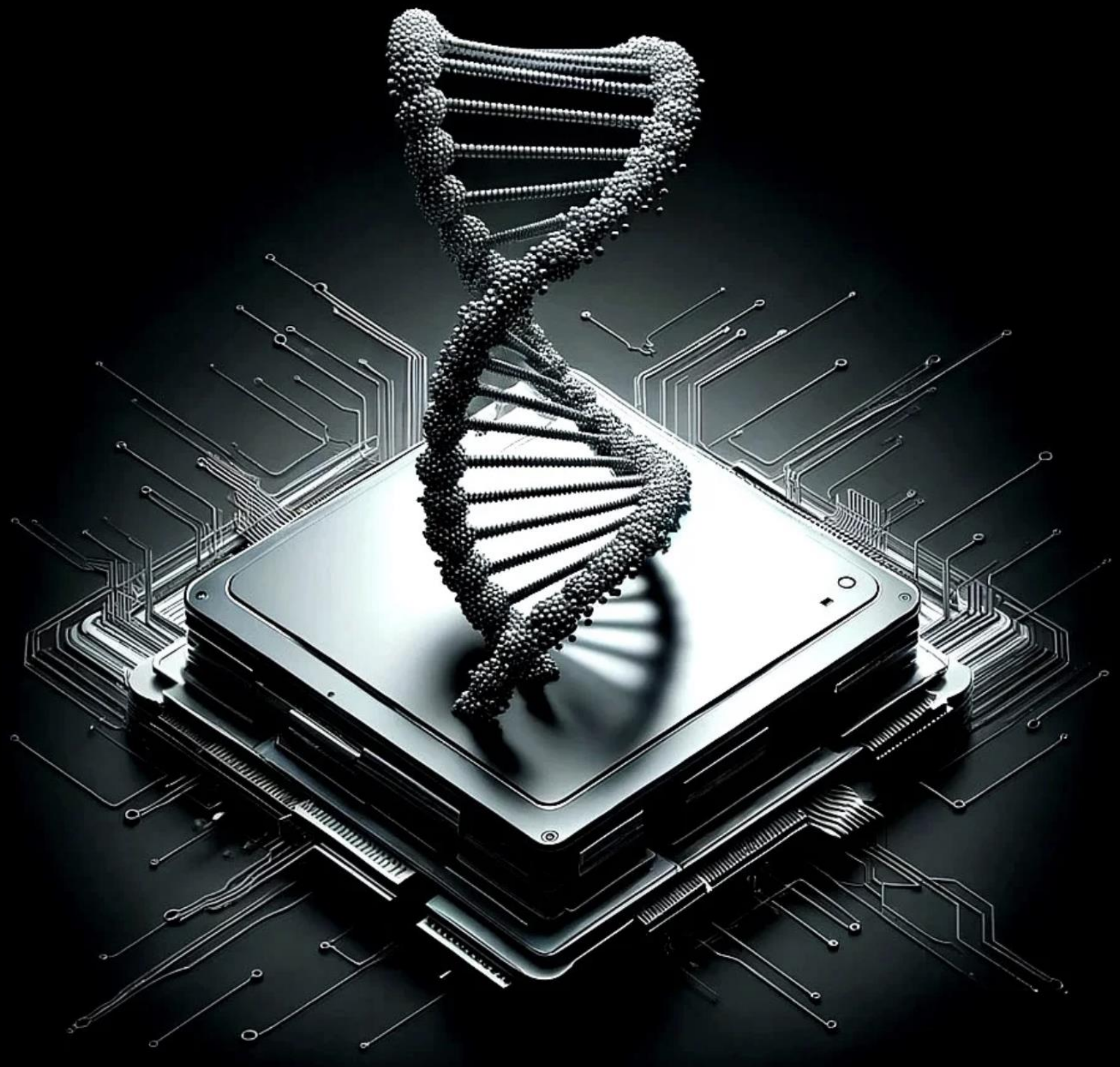
POSTERA CRESCAM LAUDE

THE UNIVERSITY OF
MELBOURNE

THE

BARCA GROUP

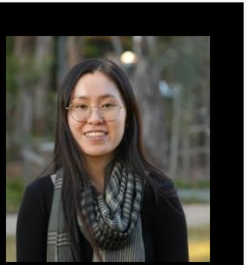
HIGH-PERFORMANCE COMPUTING, AI & DIGITAL CHEMISTRY



www.barcagrp.com

Ryan Stocks

J. Galvez-Vallejo



Fiona Yu

C. Snowdon

Barca Group in Computer Science & QDX

E. Palethorpe

J. Kurzak
AMD

D. Bykov
ORNL

Prof. Peter Gill

President of the World Association for Theoretical and Computational Chemists (WATOC)

Asia-Pacific

Prof. Frank Neese

Director of Max Plank Institute for Molecular Theory and Spectroscopy

Europe

Prof. Laura Gagliardi

Director of the Catalyst Design for Decarbonization Center, Editor in Chief of the Journal of Chemical Theory and Computation of the American Chemical Society

USA

Developing a single new drug **takes 10–15 years,** costs up to **$2.6 billion,** and **passes clinical trials** only **12% of the time.**

80% of disease-driving proteins are “undruggable” with non-covalent therapeutics, leaving diseases like Alzheimer’s, cancers, and multidrug-resistant infections largely incurable.

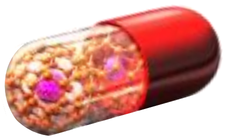
These diseases affect 75 million people globally, causing over 13.5 million deaths annually.

This is equivalent to 90+ Hiroshima, every year.

Covalent inhibitors, which form irreversible bonds with proteins**, can target “undruggable” proteins.**

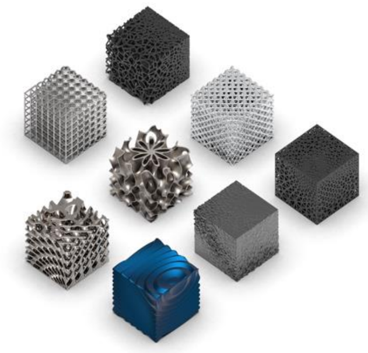
This work paves the way for **the first accurate *in silico* software to design and model covalent binders**.

SOME OUTSTANDING SCIENCE & TECHNOLOGY CHALLENGES



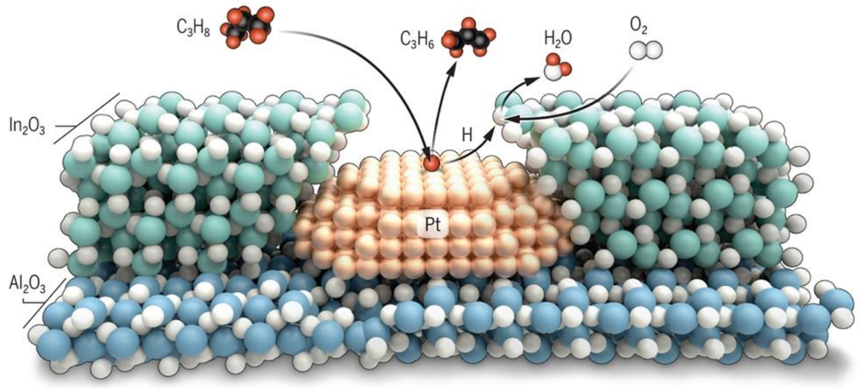
Biology, medicine, biochemistry

☞ Drug design and drug binding, biological interfaces, enzymatic catalysis



Nanomaterial engineering

☞ energy generation (batteries, hydrogen storage), drug delivery systems, purification membranes, biosensors, opto- and nano-electronics, exfoliation, and many others.



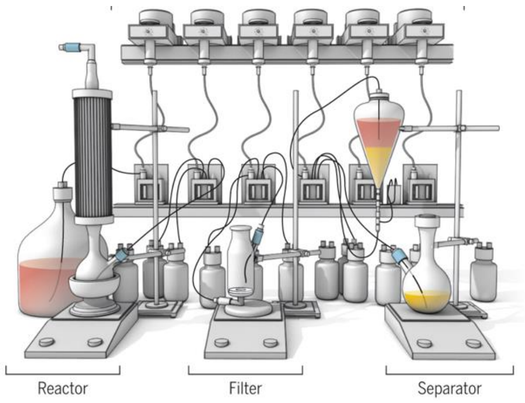
Heterogenous catalysis

☞ Second generation biofuels (biomass conversion), liquid phase catalysis, green catalysis, production of high-added-value (fine) chemicals.

Computer Simulations

☞ Fast and inaccurate

☞ Accurate but too slow



Physical Experiments

☞ Expensive and slow

☞ Not available, unreliable

MOLECULAR DYNAMICS WITH CLASSICAL POTENTIALS

Classical Potentials

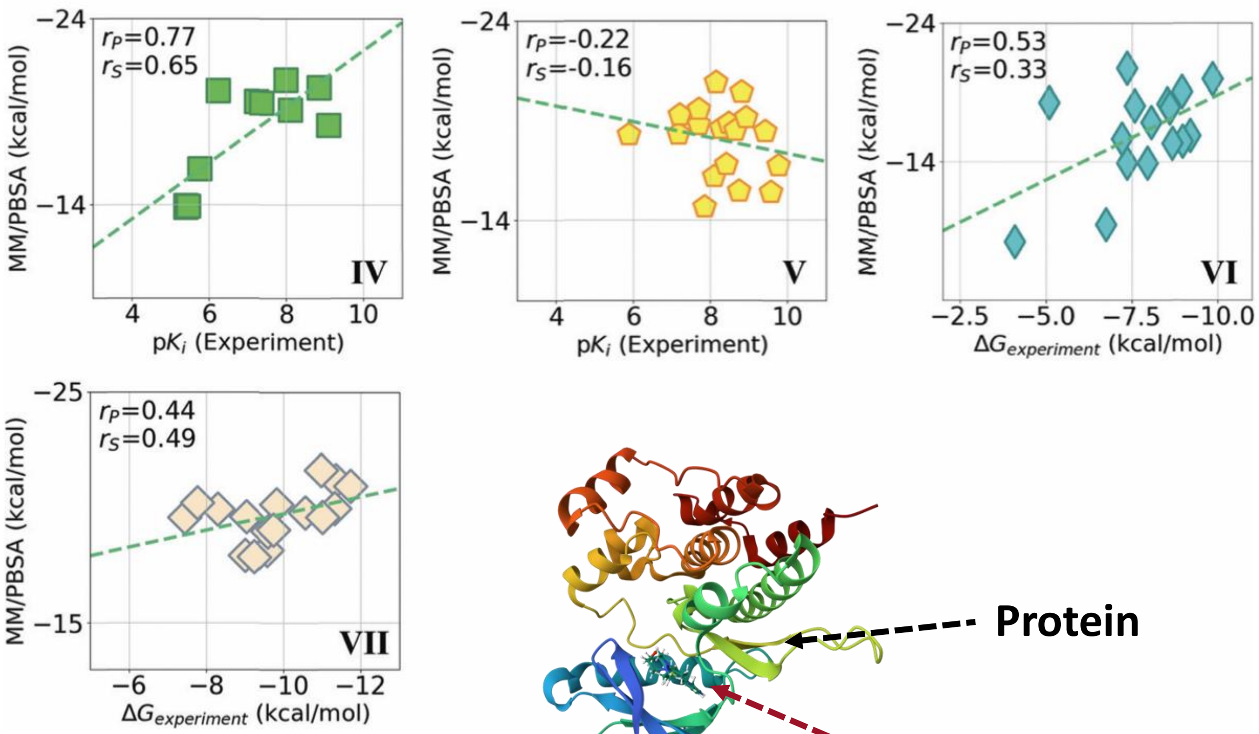
Atoms are treated as classical particles (no electrons). Use empirical, parameterized models (*e.g.,* ball and spring) for molecular interactions.

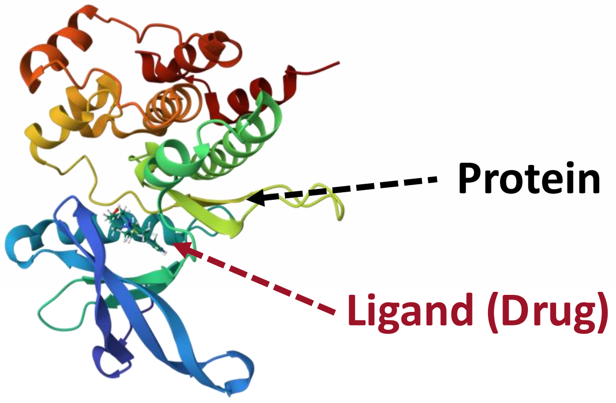
ADVANTAGES

- **Fast and Scalable**: Suitable for very large systems (e.g., proteins, membranes) and long-time-scale simulations.
- **Wide Range of Tools**: Mature field with extensive libraries.

DISADVANTAGES

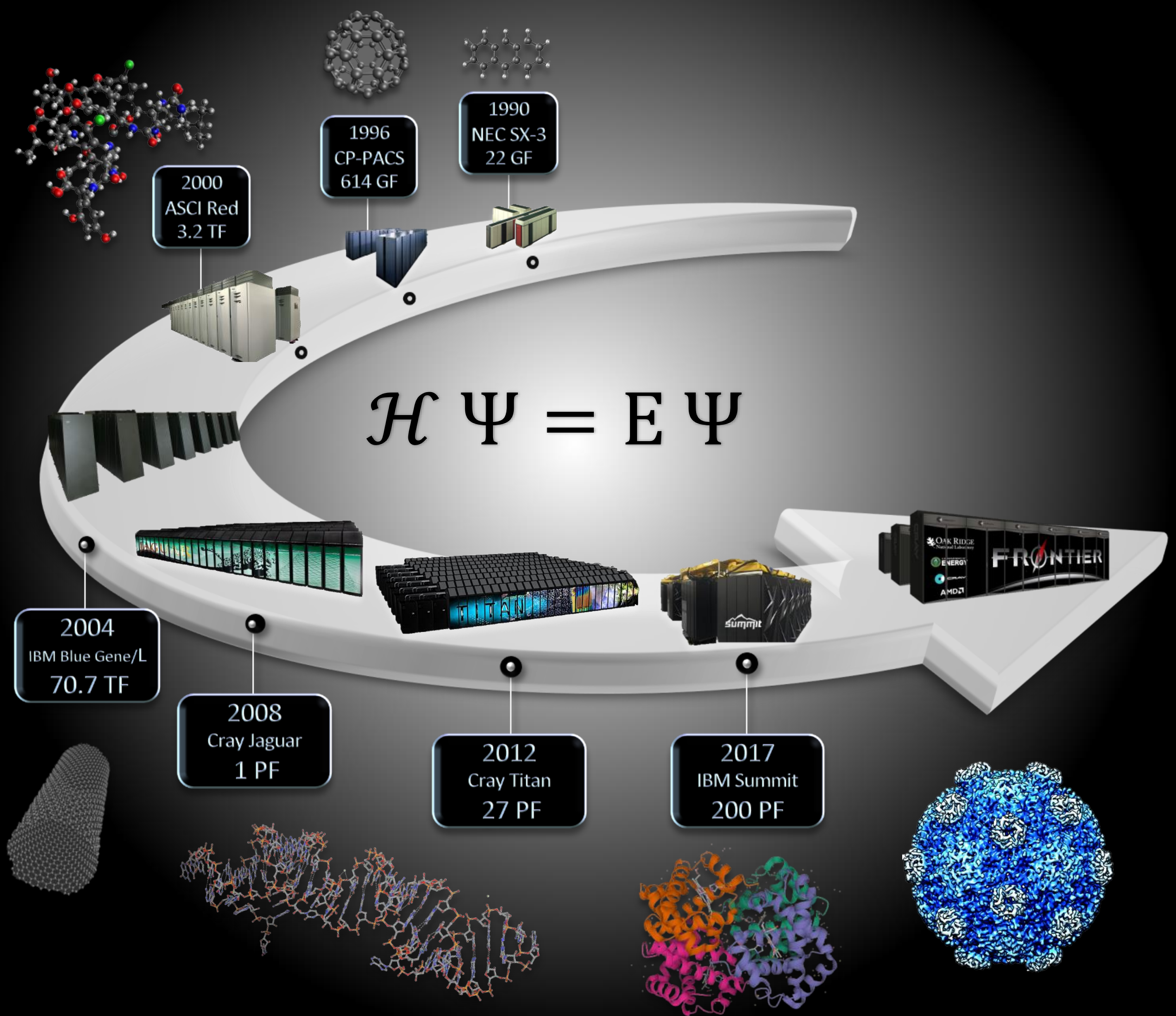
- **Lack of Physics Details**: Cannot describe electronic effects, such as charge transfer or bond breaking/forming (no reactions).
- **Limited Accuracy**: Cannot accurately model H-bonds, dispersion forces, and other non-covalent interactions that play a key role in biomolecular systems' energetics.
- **Limited Transferability**: Parameters typically do not transfer well between different molecular environments.





- **Correlation with experiments can be quite poor**
- **Not sufficiently accurate and reliable for drug discovery**

Maier, R., et al., *Phys. Chem. Chem. Phys.*, 2022, 24, 14525



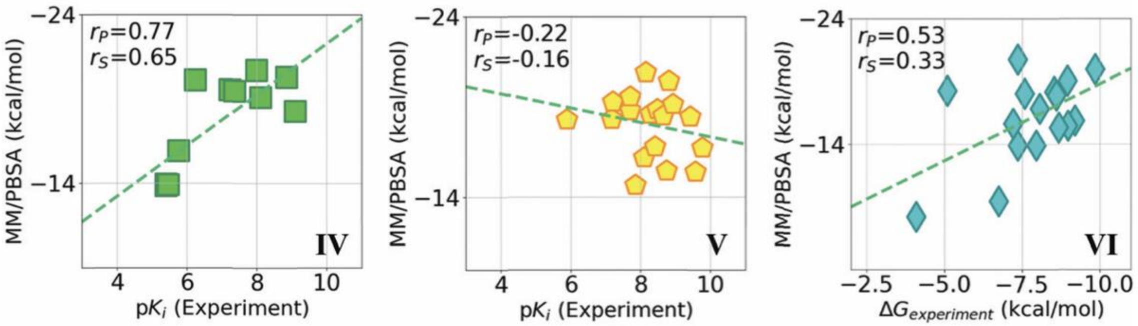
"The underlying physical laws necessary for the mathematical theory of a large part of physics and the whole of chemistry are thus completely known,
and the difficulty is only that the exact application of these laws leads to equations much too complicated to be soluble."
P. A. M. Dirac, 1929.

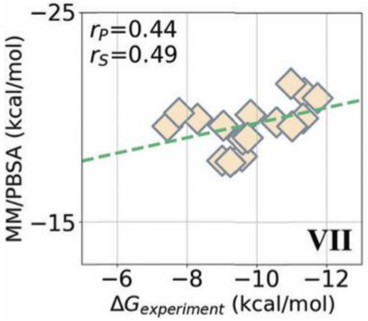
Ab initio quantum chemistry methods solve the Schrödinger equation from first principles (e.g. MP2), **without relying on empirical parameters** (no DFT).

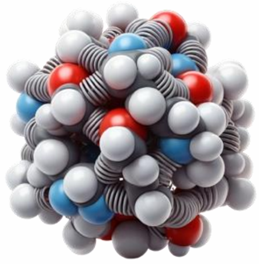
Can provide an **accuracy that rivals physical experiments,** though at a high computational cost.

ACCURACY OF QUANTUM VERSUS CLASSICAL POTENTIALS

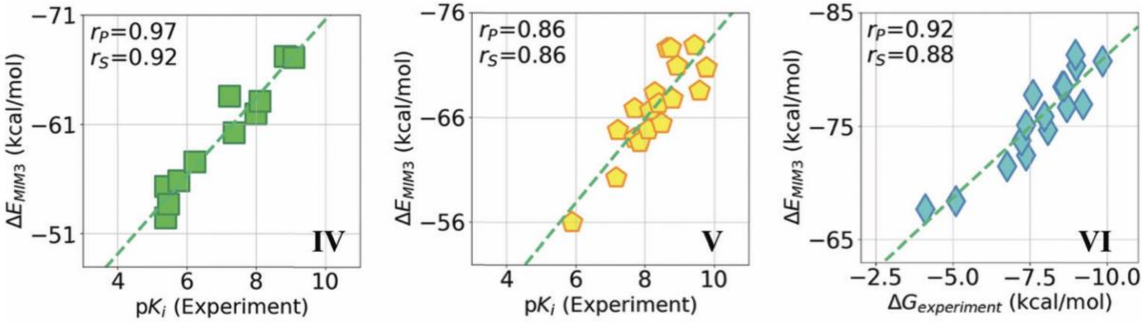
Classical Potentials

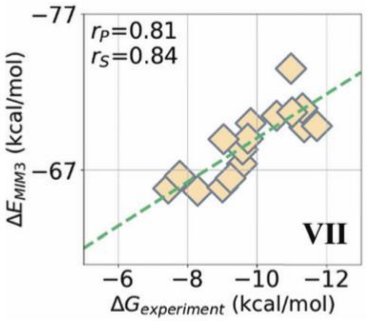


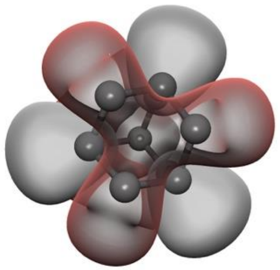




Quantum Potentials







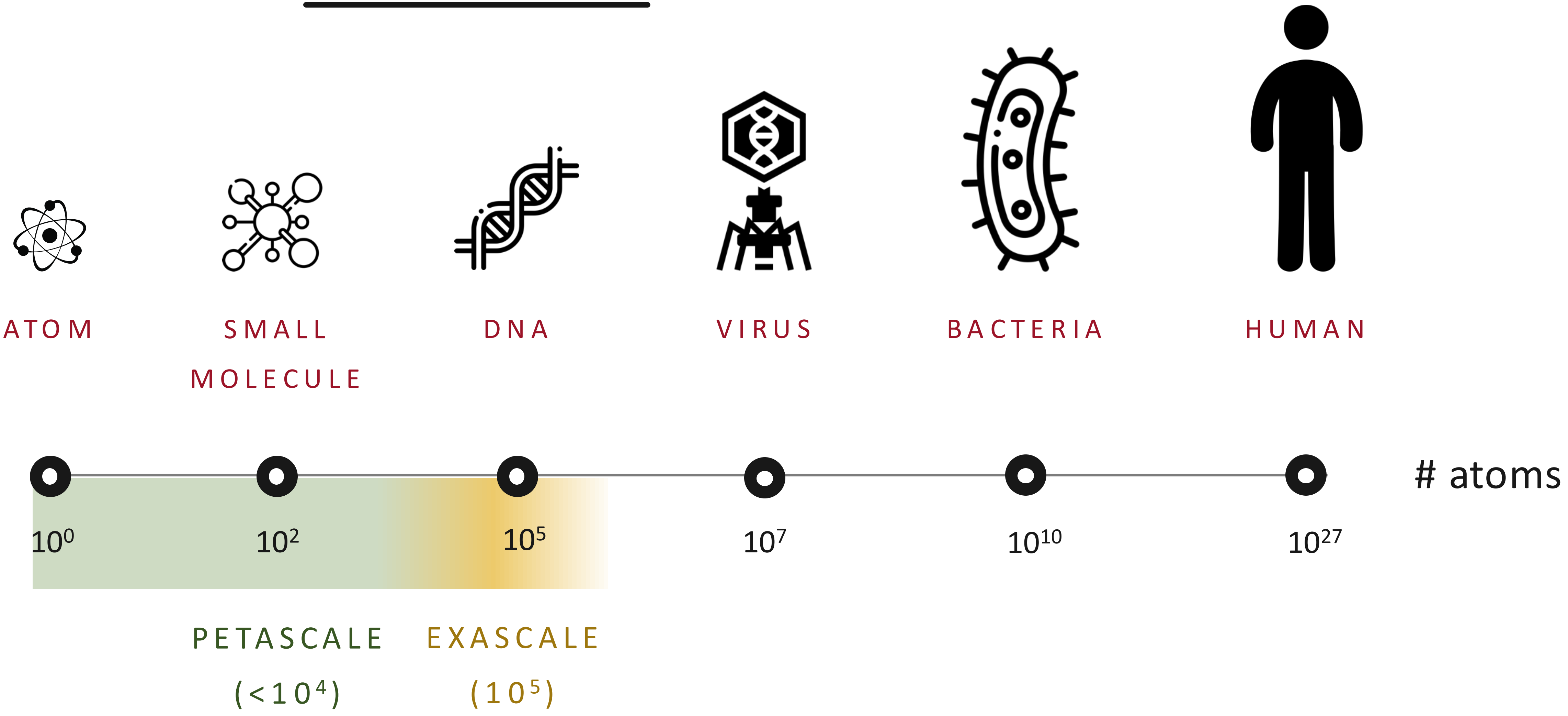
- **Poor correlation with experiment** is the result of **inaccurate physics** models
- Longer simulations do not improve correlation in this case

- Using **quantum mechanical potentials** results in much **better alignment with experiments**
- Some **improvements** are not only quantitative but also qualitative, representing **fundamentally different** and enhanced outcomes (*e.g.*, as shown in the yellow data points)
- **Can model bond breaking and formation**

Maier, R., et al., *Phys. Chem. Chem. Phys.*, 2022, 24, 14525

QUANTUM CHEMICAL CALCULATIONS

IN ATOMS



ATOM — 10^0

SMALL MOLECULE — 10^2

DNA — 10^5

VIRUS — 10^7

BACTERIA — 10^{10}

HUMAN — 10^{27}

# atoms

PETASCALE ($<10^4$)

EXASCALE (10^5)

CHALLENGES

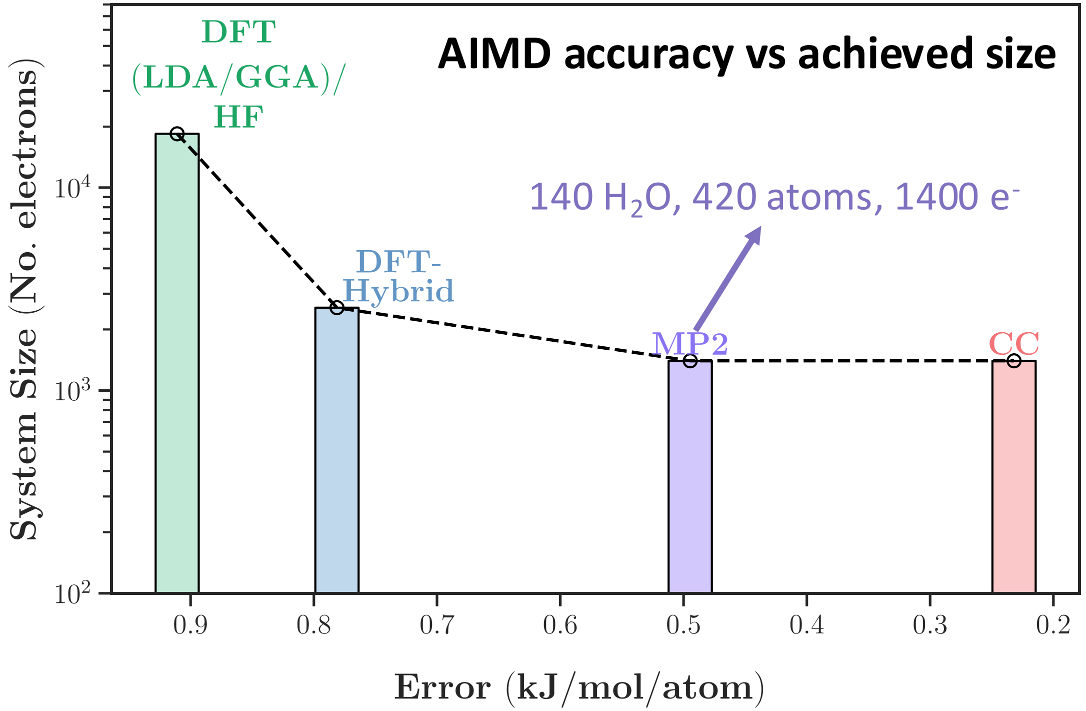
Scalability

The amount of **computation required to solve** (accurately enough) **the Schrodinger equation scales as a high power of the number of atoms**, *N*, within a molecular system.

Method	Scaling (time complexity)	Accuracy
Hartree-Fock, Local DFT	$\mathcal{O}(N^3)$	Qualitative
Hybrid DFT GGA, Meta-GGA	$\mathcal{O}(N^3)$	Not always accurate, can be predictive
PT2-based (Scaled MP2, Double-Hybrids)	$\mathcal{O}(N^5)$	Accurate, predictive with some flaws
CCSD(T)	$\mathcal{O}(N^7)$	Very accurate, predictive

Accuracy

Accurate modelling of biomolecular system behavior requires **quantum mechanical accuracy beyond hybrid DFT.**

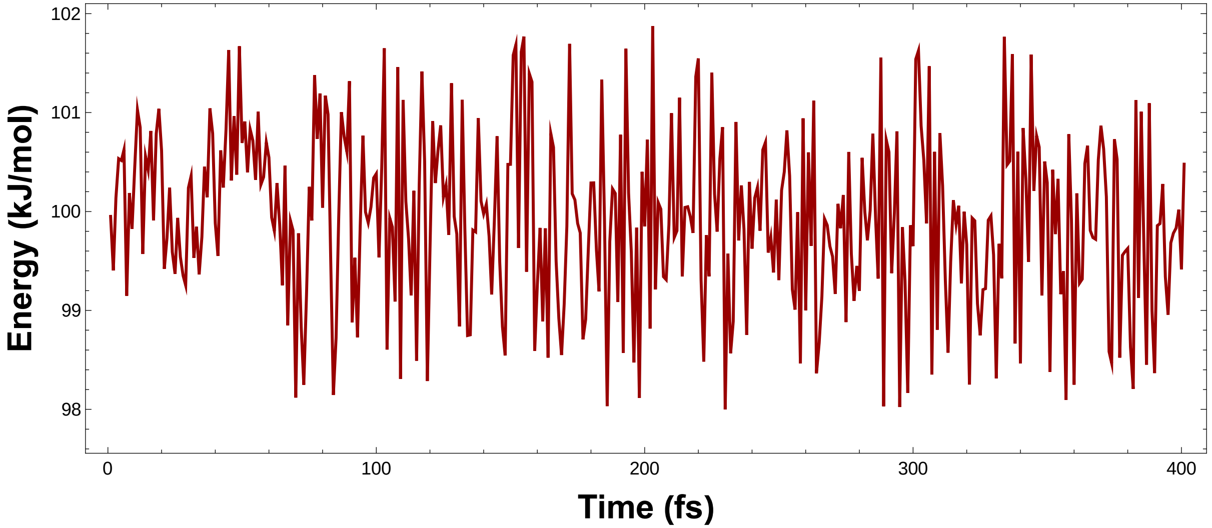


- **Hybrid DFT struggles** with the accurate modelling of **non-covalent interactions** which play a critical role in biomolecular systems.

CHALLENGES

Time Evolution

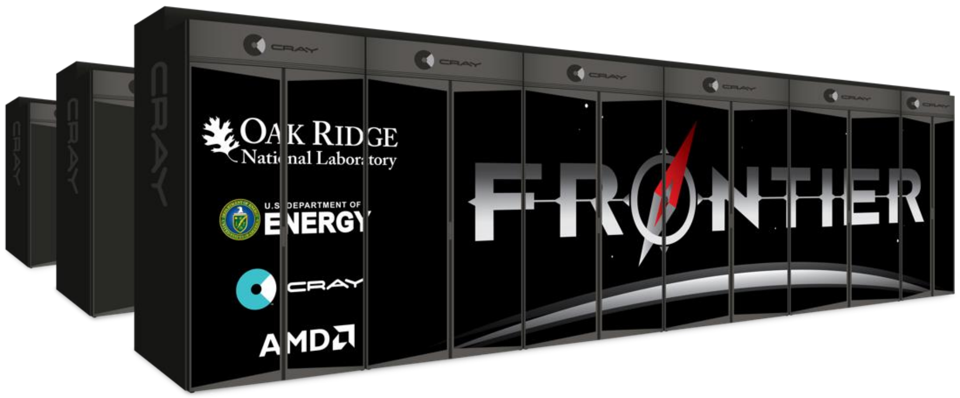
Static energy calculations have limited predictive power. **Dynamic simulations** (time-dependent) **are typically required** to obtain **statistically meaningful** predictions of **macroscopic properties**.



- ☞ Requires complex quantum mechanical gradients
- ☞ Can require many timesteps

Computational Efficiency

Inability of many quantum chemistry methods and algorithms **to use efficiently novel massively parallel processors and computer architectures**.



▻ 14k cores/GPU, 4 GPUs/node, 9408 nodes

- ☞ Most quantum chemistry codes run at 0.1-10% of R-Peak
- ☞ Most quantum chemistry codes are not ported to GPU

THE PATH TO EXTREME-SCALE QUANTUM CHEMISTRY

To devise quantum chemistry **methods**, **algorithms** and **implementations** that

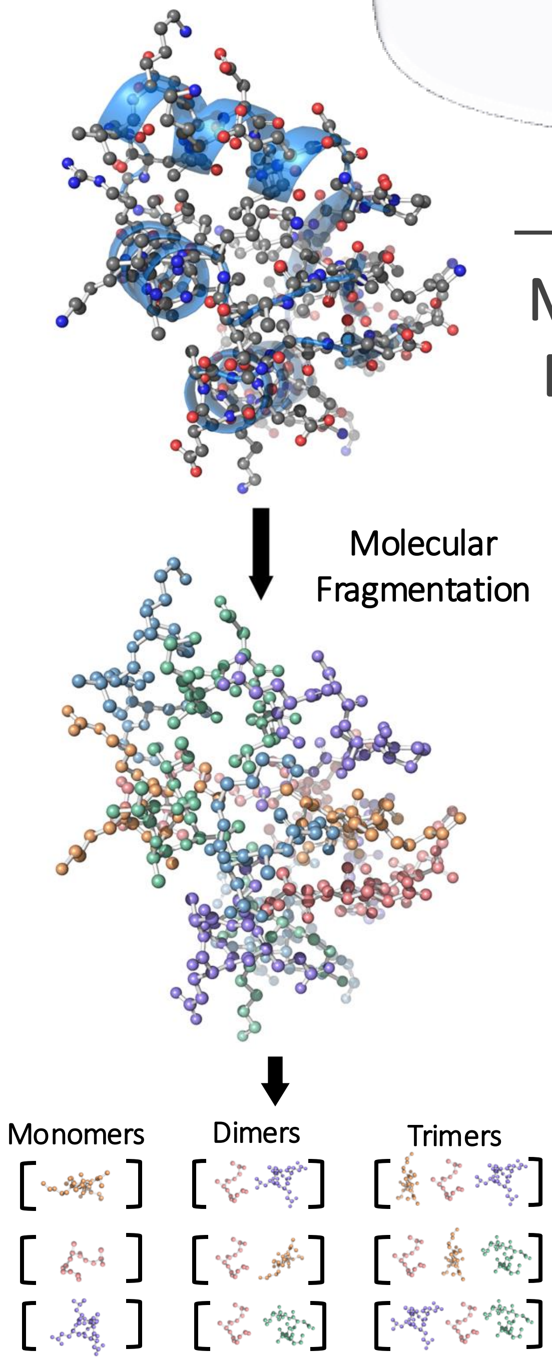
1. Have a **reduced computational complexity**, while retaining the required accuracy.
2. Are designed to **efficiently exploit** the **computational capabilities** of **throughput-oriented massively parallel hardware**.

The Extreme-scale Electronic Structure System

EXESS

Towards and Extreme-Scale Electronic Structure System, *J. Chem. Phys. 28 July 2023; 159 (4): 044112*

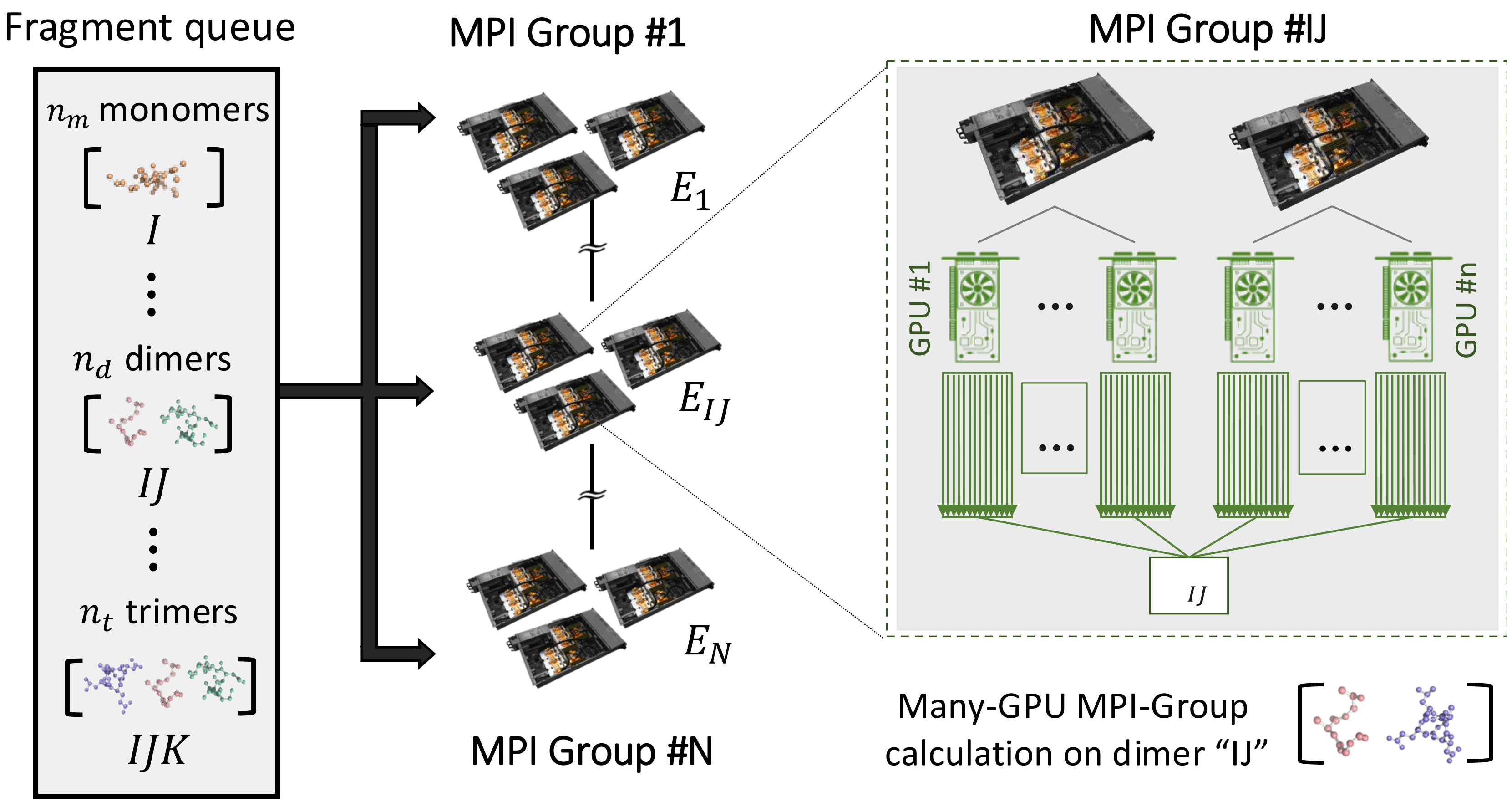
LOWER SCALING & MASSIVE PARALLELISM: FRAGMENTATION METHODS



MANY-BODY EXPANSION

$$E \cong \sum_{I} E_I + \sum_{I<J} (E_{IJ} - E_I - E_J) + \sum_{I<J<K} (E_{IJK} - \cdots) + \cdots$$

$O(n_m)$ scaling

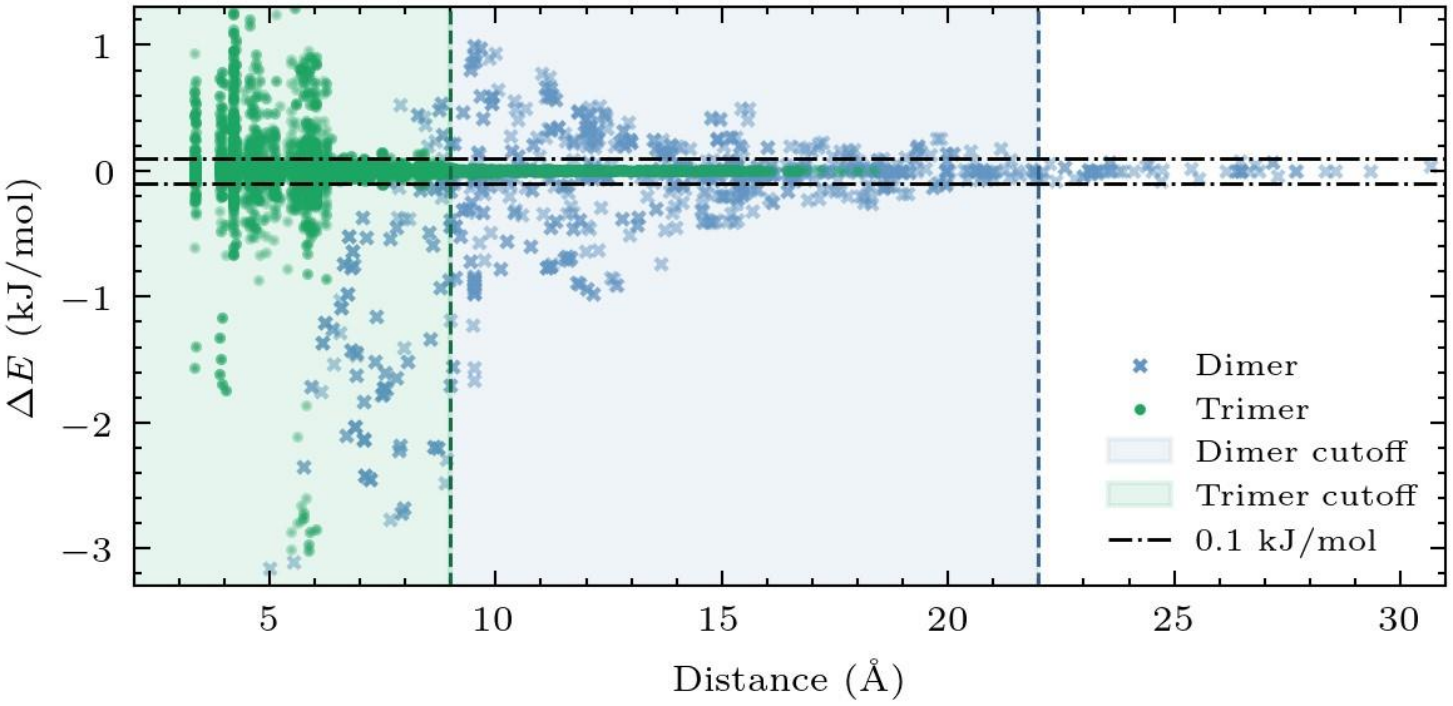


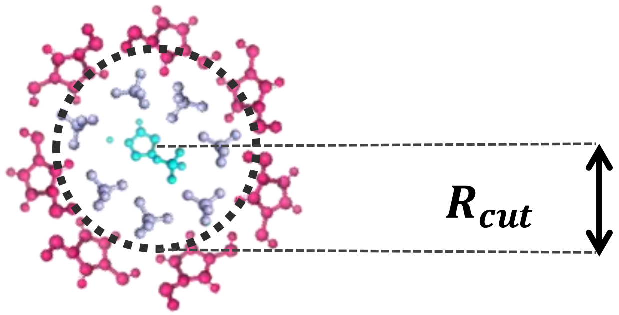
ACHIEVING LINEAR SCALING

MANY-BODY EXPANSION

$$E \cong \sum_{I} E_I + \sum_{\substack{I<J \\ R_{IJ}<R_{cut}}} \Delta E_{IJ} + \sum_{\substack{I<J<K \\ R_{IJ},R_{JK},R_{IK}<R_{cut}}} \Delta E_{IJK}$$

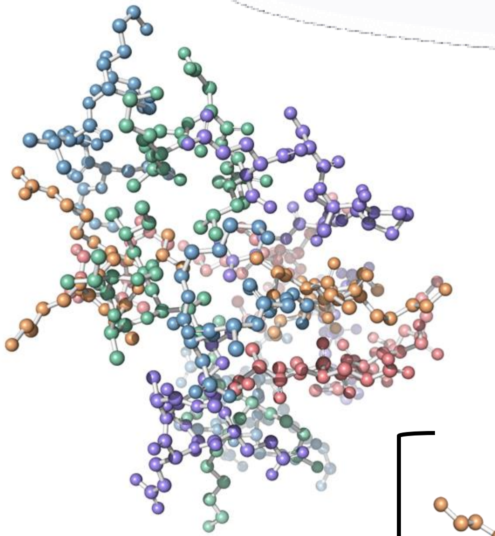
n_m = #monomers





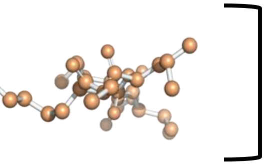
- Each monomer "I" is coupled only with $O(1)$ monomers "J" within $\boldsymbol{R_{cut}}$
- In total only $O(n_m)$ dimers are computed ➛ linear computational complexity
- For sufficiently large $\boldsymbol{R_{cut}}$, no accuracy is lost!

MOLECULAR FRAGMENTATION METHODS: COMPONENTS OF THE ENERGY

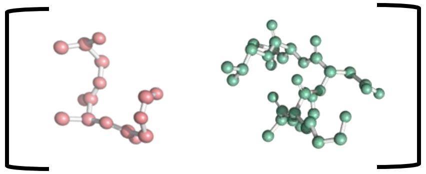


MANY-BODY EXPANSION

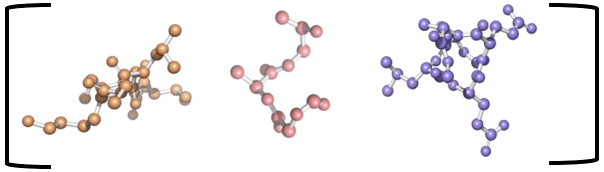
$$E \cong \sum_{I} E_I + \sum_{I<J} \Delta E_{IJ} + \sum_{I<J<K} \Delta E_{IJK}$$



$$E_I = E_I^{HF} + E_I^{MP2}$$



$$\Delta E_{IJ} = E_{IJ}^{HF} + E_{IJ}^{MP2} - E_I - E_J$$



$$\Delta E_{IJK} = E_{IJK}^{HF} + E_{IJK}^{MP2} - \Delta E_{IJ} - \Delta E_{IK} - \Delta E_{JK} \\ -E_I - E_J - E_K$$

HF

$$E_f^{HF} = \frac{1}{2} \sum_{\mu\nu}^{N_f} D_{\mu\nu}^{f} (H_{\mu\nu}^{f} + F_{\mu\nu}^{f})$$

$$f = \{I, IJ, IJK\}$$

FLOPS	*MEM*
$\mathcal{O}(N_f^4)$	$\mathcal{O}(N_f^2)$

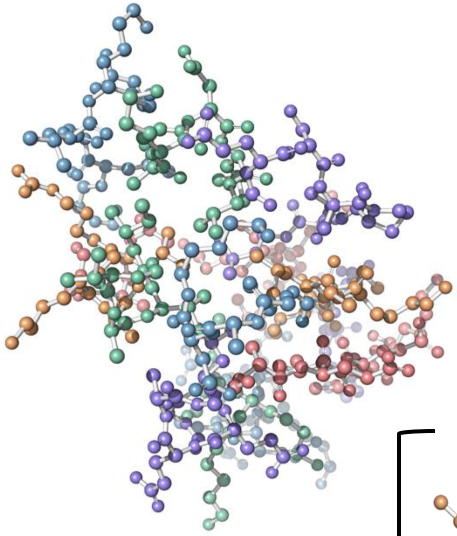
RI-MP2

$$E_f^{MP2} = \sum_{ij}^{N_o^f} \sum_{ab}^{N_v^f} \frac{G_{ia}^{jb}(2G_{ia}^{jb} - G_{ib}^{ja})}{\epsilon_i + \epsilon_j - \epsilon_a - \epsilon_b}$$

$$G_{ia}^{jb} = (ia|jb)$$

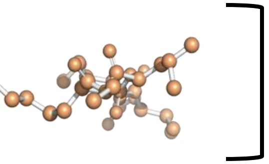
FLOPS	*MEM*
$\mathcal{O}(N_f^5)$	$\mathcal{O}(N_f^3)$

MOLECULAR FRAGMENTATION METHODS: COMPONENTS OF THE ENERGY GRADIENT

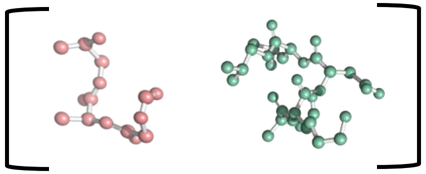


MANY-BODY EXPANSION

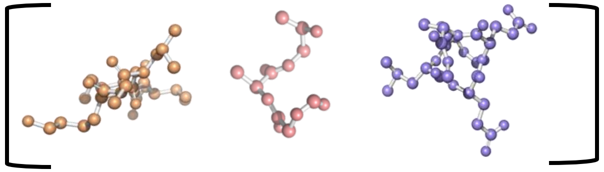
$$\nabla E \cong \sum_{I} \nabla E_I + \sum_{I<J} \nabla \Delta E_{IJ} + \sum_{I<J<K} \nabla \Delta E_{IJk}$$



$$E_I = E_I^{HF} + E_I^{MP2}$$



$$\Delta E_{IJ} = E_{IJ}^{HF} + E_{IJ}^{MP2} - E_I - E_J$$



$$\Delta E_{IJK} = E_{IJK}^{HF} + E_{IJK}^{MP2} - \Delta E_{IJ} - \Delta E_{IK} - \Delta E_{JK} \\ -E_I - E_J - E_K$$

RI-HF

$$\nabla_i E_{HF} \\ = \sum_{\mu\nu} D_{\mu\nu}\, h_{\mu\nu}^{i} - \sum_{\mu\nu} W_{\mu\nu} S_{\mu\nu}^{i} + \sum_{\mu\nu P} Y_{\mu\nu}^{P} (\mu\nu|P)_i - \frac{1}{2} \sum_{PQ} Z_{PQ} J_{PQ}^{i}$$

FLOPS $\mathcal{O}(N_f^4)$

MEM $\mathcal{O}(N_f^2)$

RI-MP2

$$\nabla_i E_{MP2} \\ = 4 \sum_{\mu\nu P} \Gamma_{\mu\nu}^{P} (P|\mu\nu)_i - 2 \sum_{PQ} (P|Q)_i + 2 \sum_{\mu\nu} \{ D_{\mu\nu} F_{\mu\nu}^{i} - W_{\mu\nu} S_{\mu\nu}^{i} \}$$

$\mathcal{O}(N_f^5)$

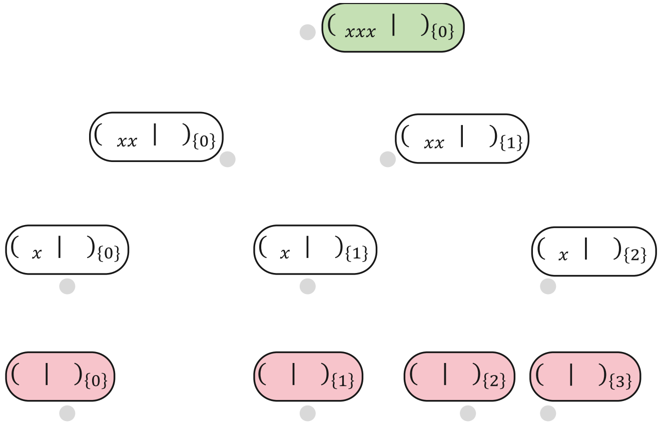
$\mathcal{O}(N_f^3)$

RESOLUTION OF THE IDENTITY (RI) HF AND MP2

4C ERI

$$(\mu\nu|\lambda\sigma) = \iint \frac{\phi_\mu(\boldsymbol{r}_1)\phi_\nu(\boldsymbol{r}_1)\phi_\lambda(\boldsymbol{r}_2)\phi_\sigma(\boldsymbol{r}_2)}{|\boldsymbol{r}_1 - \boldsymbol{r}_2|} d\boldsymbol{r}_1 d\boldsymbol{r}_2$$

- The calculation of **4-centre (4C) electron repulsion integrals (ERI)** can be the source of major computational inefficiencies
- **$O(N_f^4)$ ERIs, too many to be stored**
- Computed **using recursion** in **batches** with **different workloads** depending on the nature of the $\phi_\mu, \phi_\nu, \phi_\lambda, \phi_\sigma$ functions
- Can be **memory-bound with low FLOP rates**



HF BUILD

$$F_{\mu\nu} = \sum_{\lambda\sigma} D_{\lambda\sigma}\left[(\mu\nu|\lambda\sigma) - \frac{1}{2}(\mu\lambda|\nu\sigma)\right]$$

- **Computed each iteration and combined on-the-fly with $D_{\gamma\delta}$** to obtain Fock matrix elements
- **Permutational symmetry** used to save integrals

$$(\mu\nu|\lambda\sigma) = (\nu\mu|\lambda\sigma) = (\nu\mu|\sigma\lambda) = (\sigma\lambda|\nu\mu)$$

- Leads to **scattered memory access** and potential **race conditions** in **parallel Fock matrix updates**.

$F_{\alpha\beta} \leftarrow D_{\gamma\delta}(\alpha\beta|\gamma\delta)$

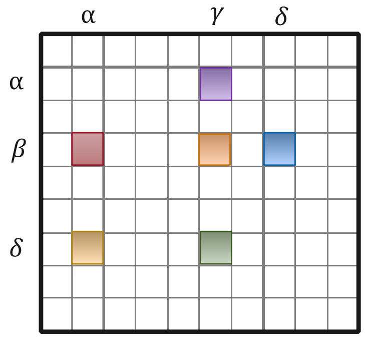
$F_{\alpha\delta} \leftarrow D_{\beta\gamma}(\alpha\beta|\gamma\delta)$

$F_{\beta\delta} \leftarrow D_{\alpha\gamma}(\alpha\beta|\gamma\delta)$

$F_{\gamma\delta} \leftarrow D_{\alpha\beta}(\alpha\beta|\gamma\delta)$

$F_{\alpha\gamma} \leftarrow D_{\beta\delta}(\alpha\beta|\gamma\delta)$

$F_{\beta\gamma} \leftarrow D_{\alpha\delta}(\alpha\beta|\gamma\delta)$



RI-HF & RI-MP2

$$(\mu\nu|\lambda\sigma) \approx (\mu\nu|\lambda\sigma)_{RI} = \sum_P B_{\mu\nu}^P B_{\lambda\sigma}^P$$

$$B_{\lambda\sigma}^P = \sum_Q (\mu\nu|P)(P|Q)^{-1/2}$$

- Compute (**on GPU**) only $\boldsymbol{O(N_f^3)}$ 3C integrals $(\mu\nu|P)$ and $\boldsymbol{O(N_f^2)}$ 2C integrals $(P|Q)$
- **Computed once and stored** on host/device

$$F_{\mu\nu} = \sum_P \sum_{\lambda\sigma} D_{\lambda\sigma}\left[B_{\mu\nu}^P B_{\lambda\sigma}^P - \frac{1}{2} B_{\mu\lambda}^P B_{\nu\sigma}^P\right]$$

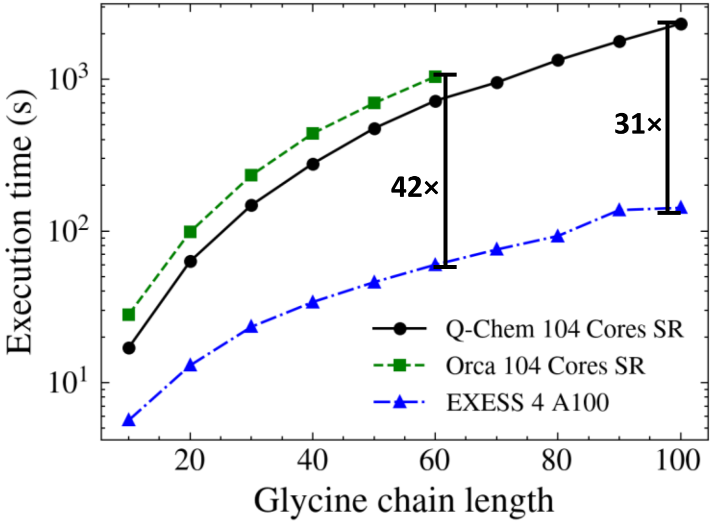
- **Fock build is implemented using DGEMM!**
- **The $O(N_f^5)$ bottleneck of MP2** also becomes a **sequence of DGEMMs!**

$$(ia|jb) \approx (ia|jb)_{RI} = \sum_P B_{ia}^P B_{jb}^P$$

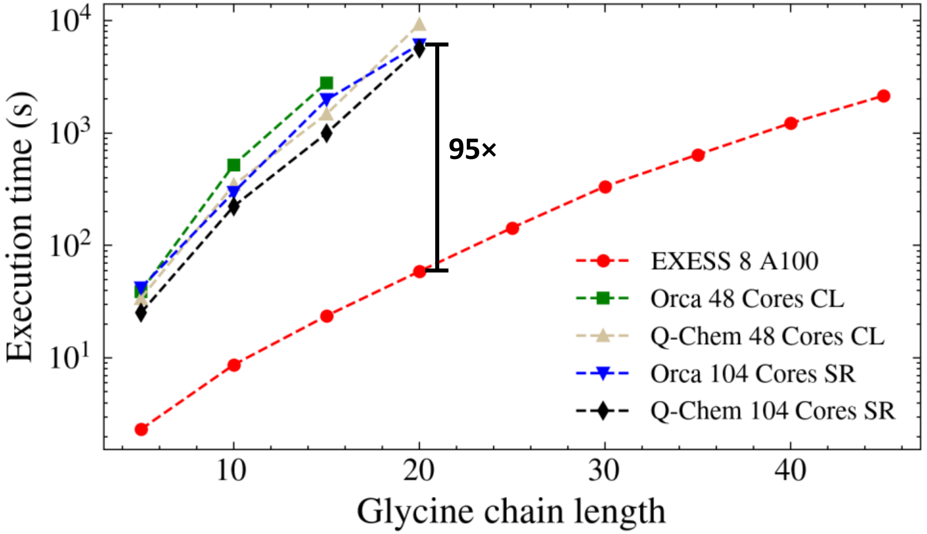
- **Can synergistically re-use tensors between RI-HF and RI-MP2, further reducing inefficiency overheads!**

COMPOUNDING PERFORMANCE

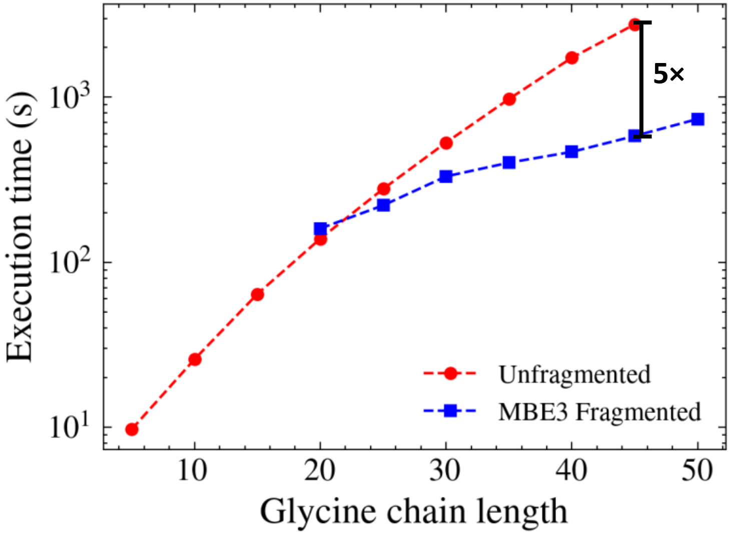
HARTREE-FOCK



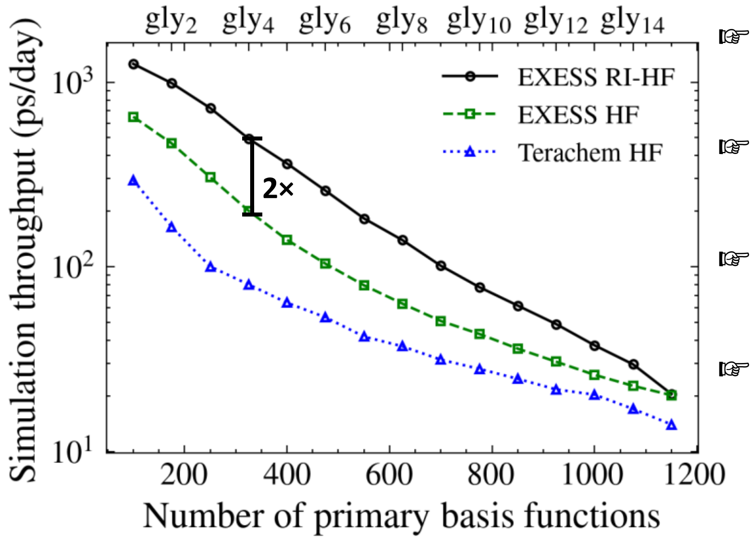
RI-MP2 GRADIENTS



RI-HF + RI-MP2 ENERGY & GRADIENTS



AIMD/RI-HF



- **EXESS HF 31-45× faster, 12-18× more energy efficient** than CPU SOTA
- **EXESS AIMD/RI-HF 2× faster** than EXESS (traditional) HF
- **EXESS RI-MP2 energy and gradients 95× faster, 19× more energy efficient** than CPU SOTA
- **EXESS MBE3/RI-HF+RI-MP2 energy and gradients 5× faster** than unfragmented for Gly$_{45}$
- **RSMD of MBE/RI-HF+RI-MP2 gradients** O(10^{-6}) – **below geometry optimization convergence threshold** for gradients in SOTA is 10^{-4}

Gly$_n$ = polyglycines, SR=Sapph. Rapids, CL=Casc. Lake

FRAGMENTATION ERROR

gly$_n$	Absolute error (Hartree/Bohr)		
	Mean	Max	RMSD
20	4.87E-07	2.54E-05	2.05E-06
25	9.76E-07	3.96E-05	3.56E-06
30	1.44E-06	4.95E-05	5.12E-06
35	1.83E-06	6.36E-05	6.59E-06
40	2.19E-06	7.78E-05	7.92E-06
45	3.15E-06	8.84E-05	9.67E-06

RI-MP2 → Stocks, R., Palethorpe, E. and Barca, G.M.J, 2024. JCTC, 20(6), 2505
RI-HF → Stocks, R., Palethorpe, E. and Barca, G.M.J, 2024. JCTC, 20 (17), 7503
HF → Palethorpe, E., Stocks, R., and Barca, G.M.J, 2024. JCTC, in press

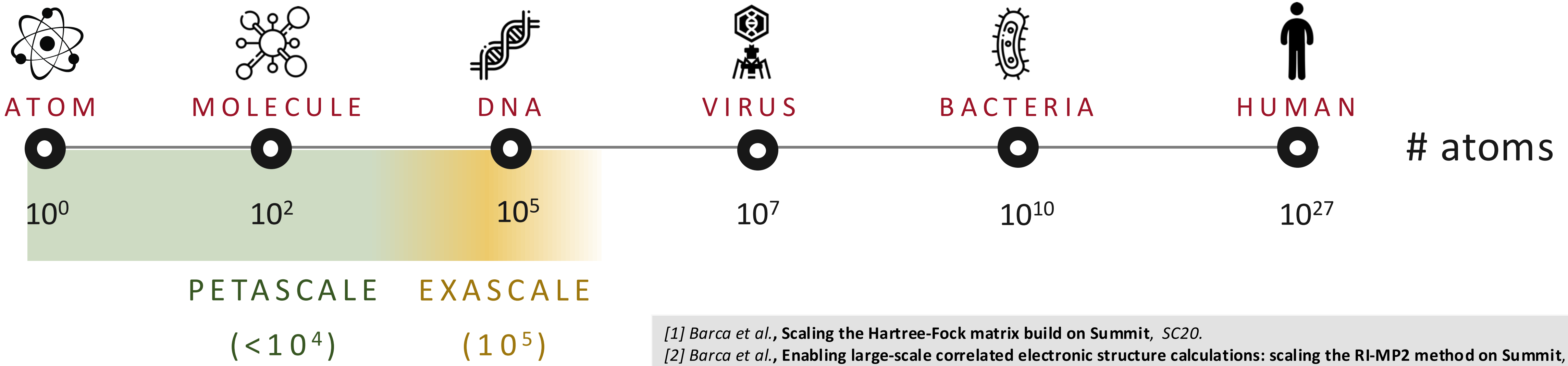
SOME ACHIEVEMENTS

- 1.7 EFLOPS
- 9408 nodes
- 75,776 MI250x GCDs
- #1 in Top500

- 150 PFLOPS
- 4698 nodes
- 27,648 V100 GPUs

OAK RIDGE National Laboratory

- **2020**, using the entire Summit supercomputer for the **largest MBE2/HF calculation [1],** on over **60,000 atoms – previous record 10,000 atoms**
- **2021**, the **largest MBE2/HF+RI-MP2 calculation [2],** on over **45,000 atoms – previous record 2,440 atoms**
- **2022,** the **largest FMO2/HF+RI-MP2 calculation,** on over **145,000 atoms [3]**
- **2023 – rewrote the whole codebase!**



[1] *Barca et al.*, **Scaling the Hartree-Fock matrix build on Summit**, *SC20.*
[2] *Barca et al.*, **Enabling large-scale correlated electronic structure calculations: scaling the RI-MP2 method on Summit**, *SC21*
[3] *Barca et al.*, **Scaling correlated Fragment Molecular Orbital Calculations on Summit**, *SC22*

LARGE SCALE QUANTUM MOLECULAR DYNAMICS USING MP2 POTENTIALS

- 1.7 EFLOPS
- 9408 nodes
- 75,776 MI250x GCDs

- 270 PFLOPS
- 2,668 nodes
- 10,752 GH200

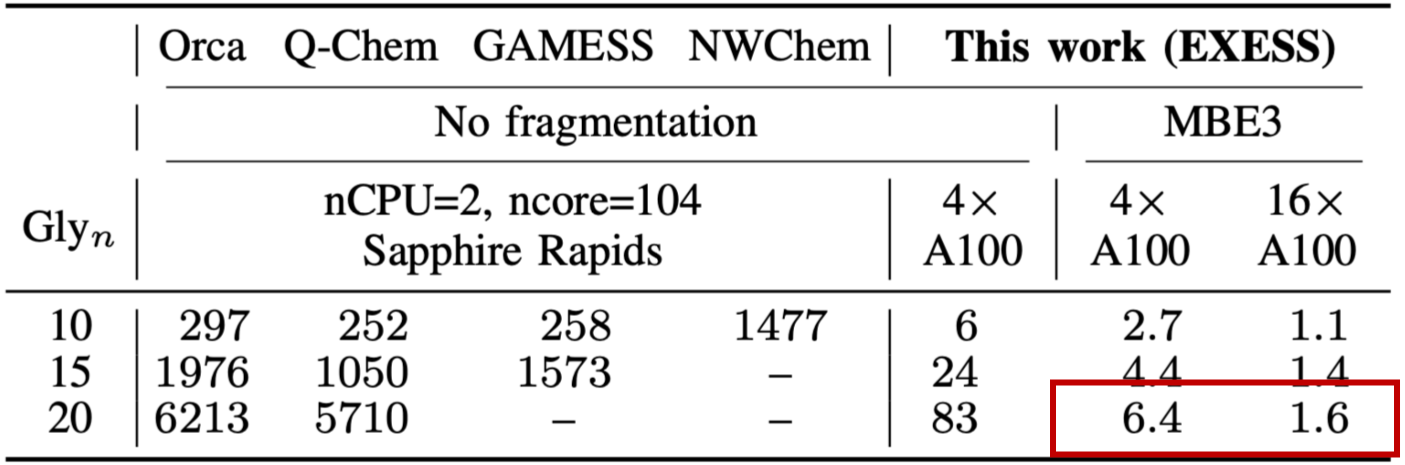
- 113 PFLOPS
- 1,536 nodes
- 6,144 A100

$$m\,\ddot{r}_i = -\nabla_i\langle\Psi|\hat{H}|\Psi\rangle = -\nabla_i(E_{RIHF} + E_{RIMP2})$$

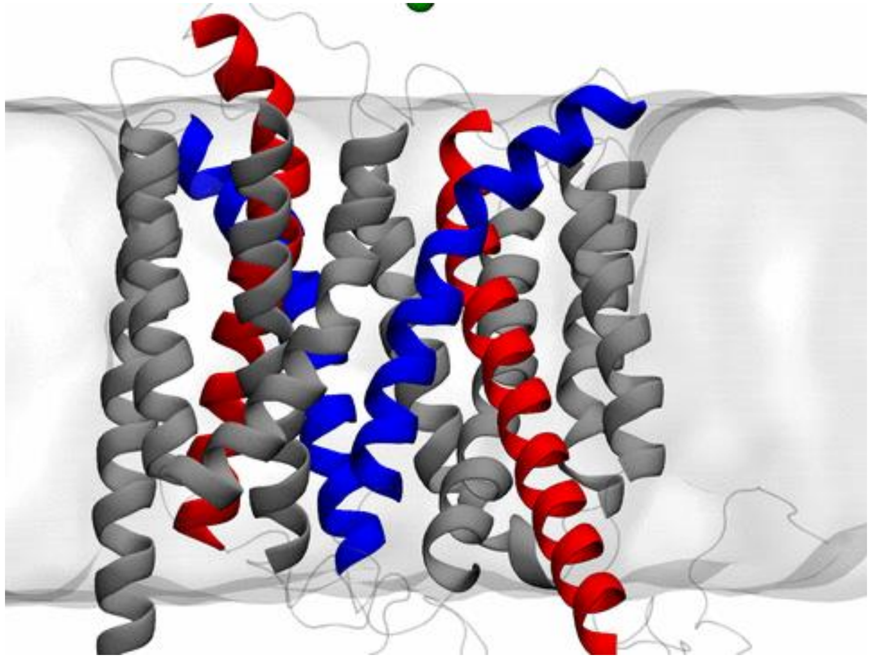
Forces are obtained from quantum mechanics on-the-fly as the MD simulation evolves

Can we simulate the ab initio molecular dynamics of biosystems at the MP2 level?

AIMD/RI-HF+RI-MP2/cc-pVDZ TIMESTEP LATENCY (s)

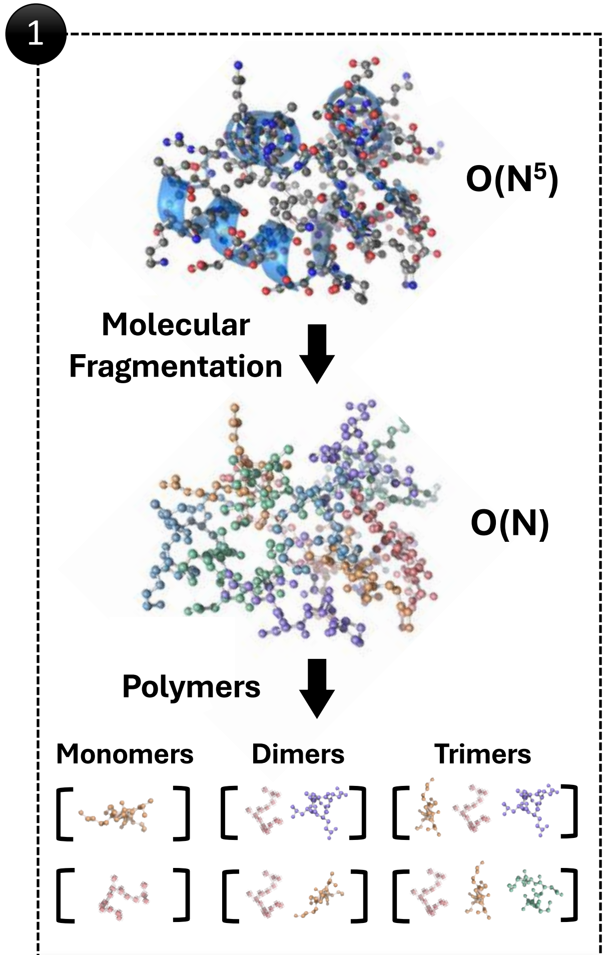


	Orca	Q-Chem	GAMESS	NWChem	**This work (EXESS)**		
	No fragmentation					MBE3	
Gly_n	nCPU=2, ncore=104 Sapphire Rapids				4× A100	4× A100	16× A100
10	297	252	258	1477	6	2.7	1.1
15	1976	1050	1573	–	24	4.4	1.4
20	6213	5710	–	–	83	6.4	1.6



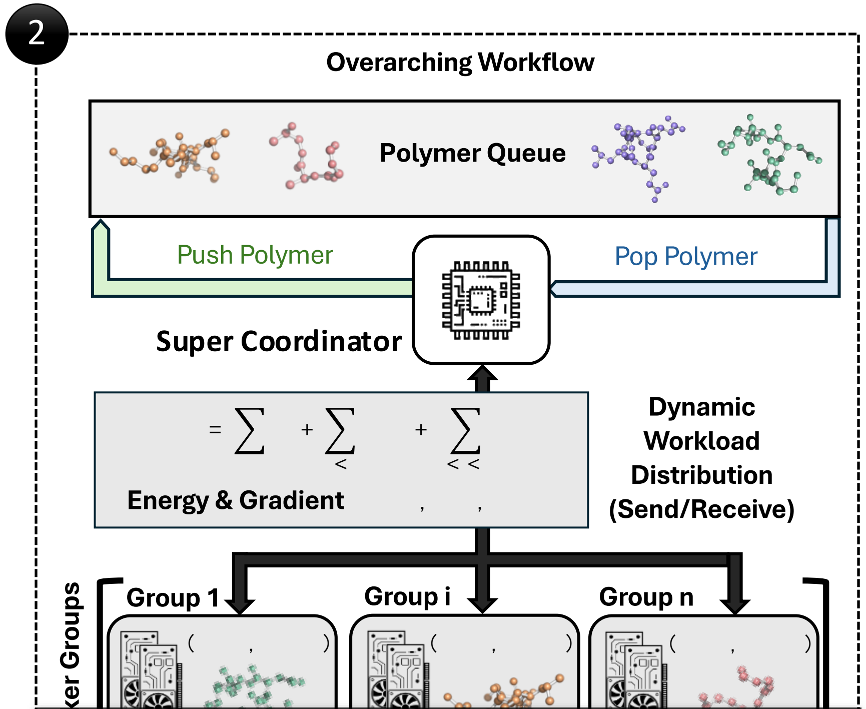
Stocks et al., **Breaking the Million-Electron and 1 EFLOP/s Barriers: Biomolecular-Scale AIMD Using MP2 Potentials,** *SC24*

BIOMOLECULAR-SCALE AIMD with MP2 POTENTIALS



1. Molecular Fragmentation (MBE3)

- Reduce scaling from O(N^5) to O(N)
- Enable globally sparse, locally dense large-scale parallelism

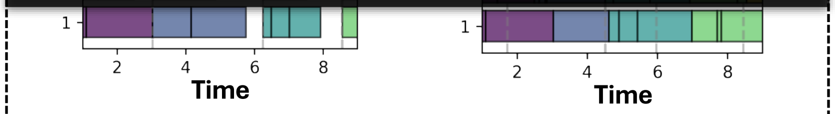


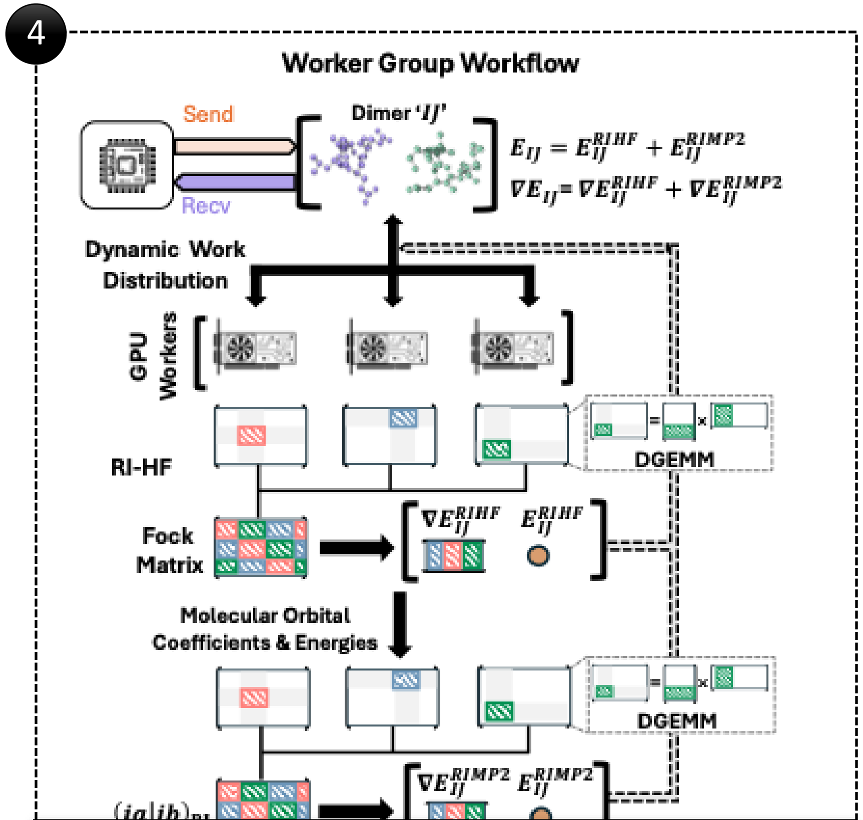
3. Asynchronous AIMD Time Steps

- Eliminate workers synch barrier at each timestep

2. Multi-Layer Distributed GPU Memory & Workload Manager

- Allocate (CPU, GPU) and pin memory across whole distributed system only once
- Efficient and fast re-use of pinned and GPU memory
- Efficient, lightweight workload (fragment) distribution across nodes and within nodes across GPUs




4
Worker Group Workflow
Send
Recv
Dimer 'IJ'
$E_{IJ} = E_{IJ}^{RIHF} + E_{IJ}^{RIMP2}$
$\nabla E_{IJ} = \nabla E_{IJ}^{RIHF} + \nabla E_{IJ}^{RIMP2}$
Dynamic Work Distribution
GPU Workers
RI-HF
DGEMM
Fock Matrix
∇E_{IJ}^{RIHF} E_{IJ}^{RIHF}
Molecular Orbital Coefficients & Energies
DGEMM
$(ia|jb)_{RI}$
∇E_{IJ}^{RIMP2} E_{IJ}^{RIMP2}

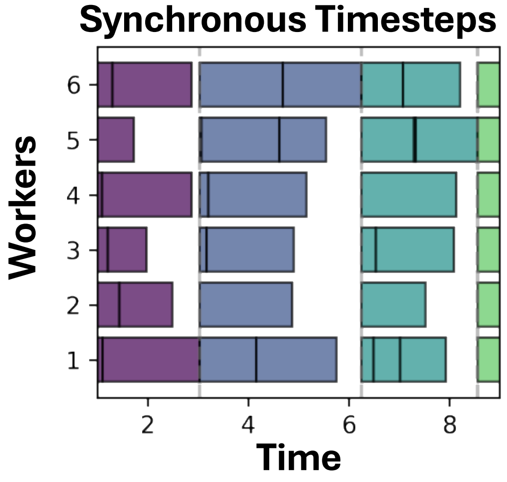
5. Runtime DGEMM Autotuning

- Determines and implements the highest-performance

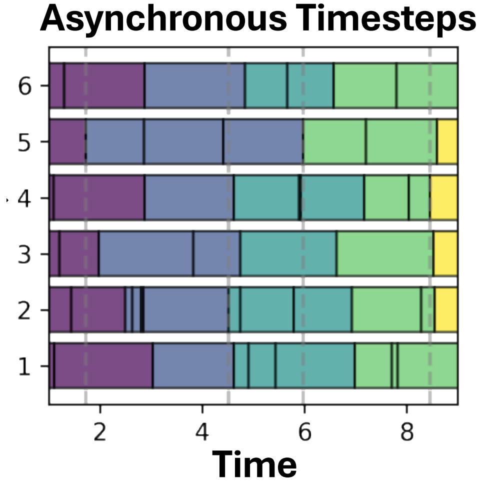
4. Fragment-Level Synergistic RI-HF plus RI-MP2 Algorithm

- Recasts memory- and FLOP-inefficient bottlenecks of traditional HF/MP2 into sequences of matrix multiplications
- Reuses synergistically intermediates between RI-HF and RI-MP2 energy and gradients

ASYNCHRONOUS TIMESTEPS


Synchronous Timesteps

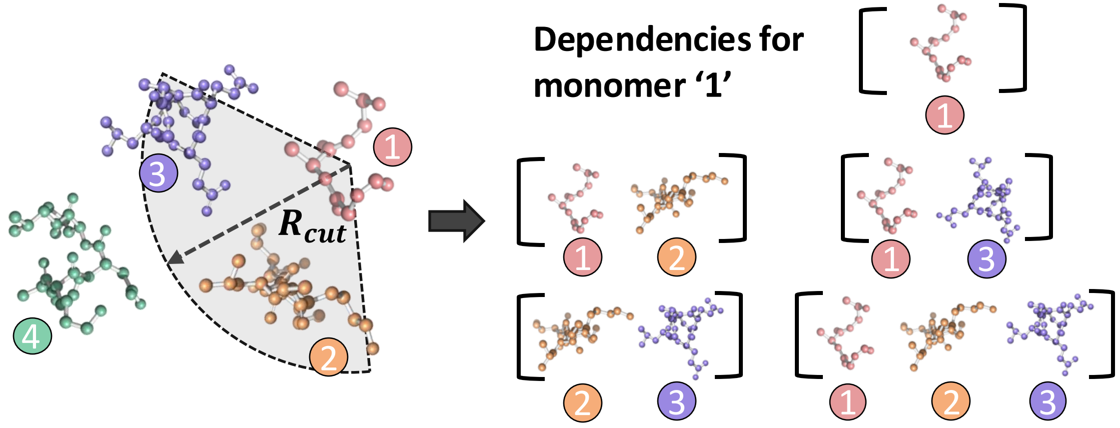
- Forces on different fragments are calculated by different GPUs, creating a **global synchronization point at the end of each timestep**.

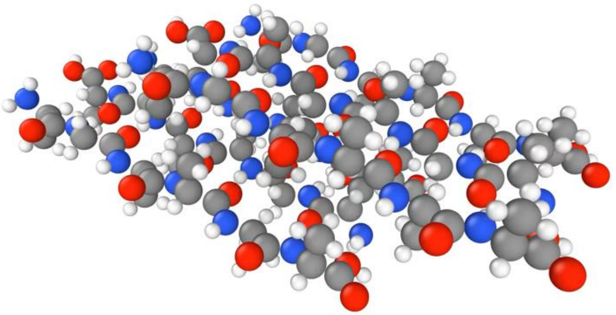

Asynchronous Timesteps

- **Monomers with resolved dependencies** are updated and **moved to the next timestep** pool
- New polymers form from monomers at each timestep and are distributed across system GPUs

- However, all polymers are formed starting from the monomers
- Thus, updates of **positions for the whole molecule**, require **updating only monomers positions** through forces
- **Forces on a given monomer** depend on the quantum **gradient of all polymers including *that* monomer**

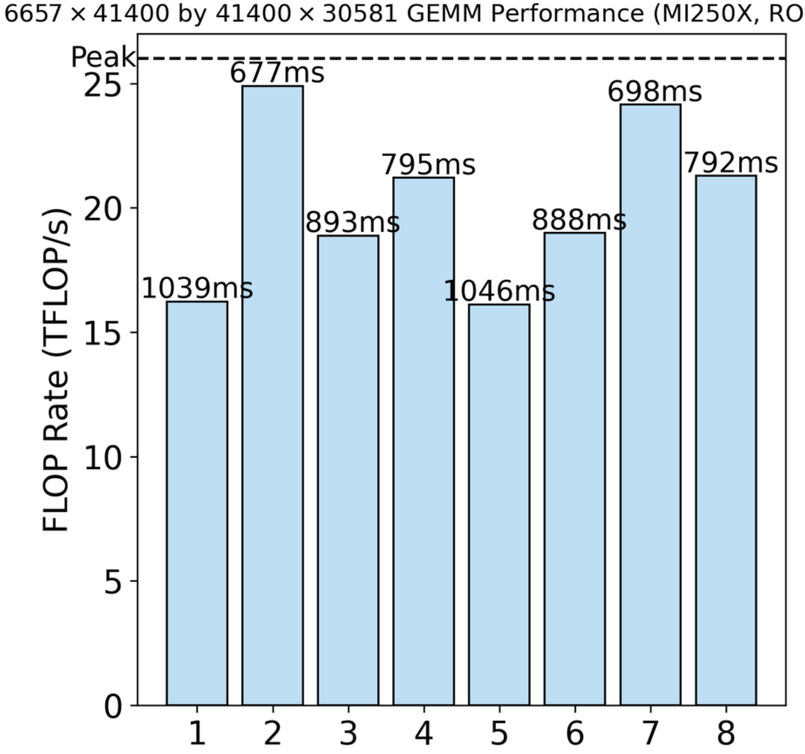
- **Allows to exploit parallelism across timesteps**
- **Global synchronization is eliminated** at each timestep
- 2BEG protein with >5.5k electrons, on 4,098 A100 GPUs, yields **40% speedup from asynchronous timesteps**


Dependencies for monomer '1'



- 1,024 nodes
- 4,098 A100 GPUs

RUN TIME AUTOTUNING (RTAT)

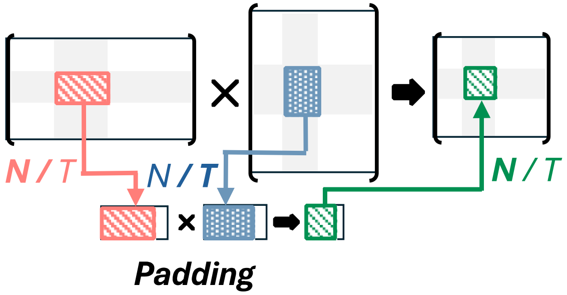


GEMM strategies

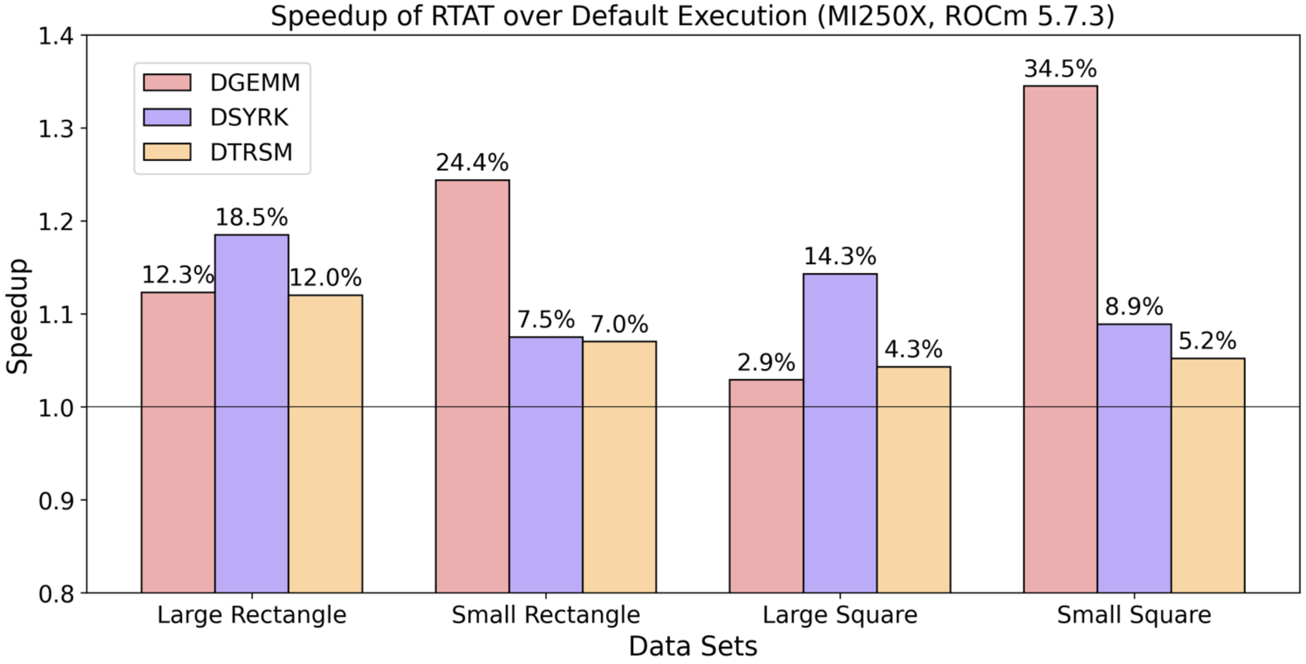
1. C := A^TB
2. X := A^T, C := XB
3. Y := B^T, C := A^TY^T
4. X := A^T, Y := B^T, C := XY^T
5. Z := B^TA, C := Z^T
6. X := A^T, Z := B^TX^T, C := Z^T
7. Y := B^T, Z := YA, C := Z^T
8. X := A^T, Y := B^T, Z := YX^T, C := Z^T

DGEMM runtime autotuning

n	m	k	Optimal
512	512	256	NTN
160	160	512	NNT
256	120	400	NNN



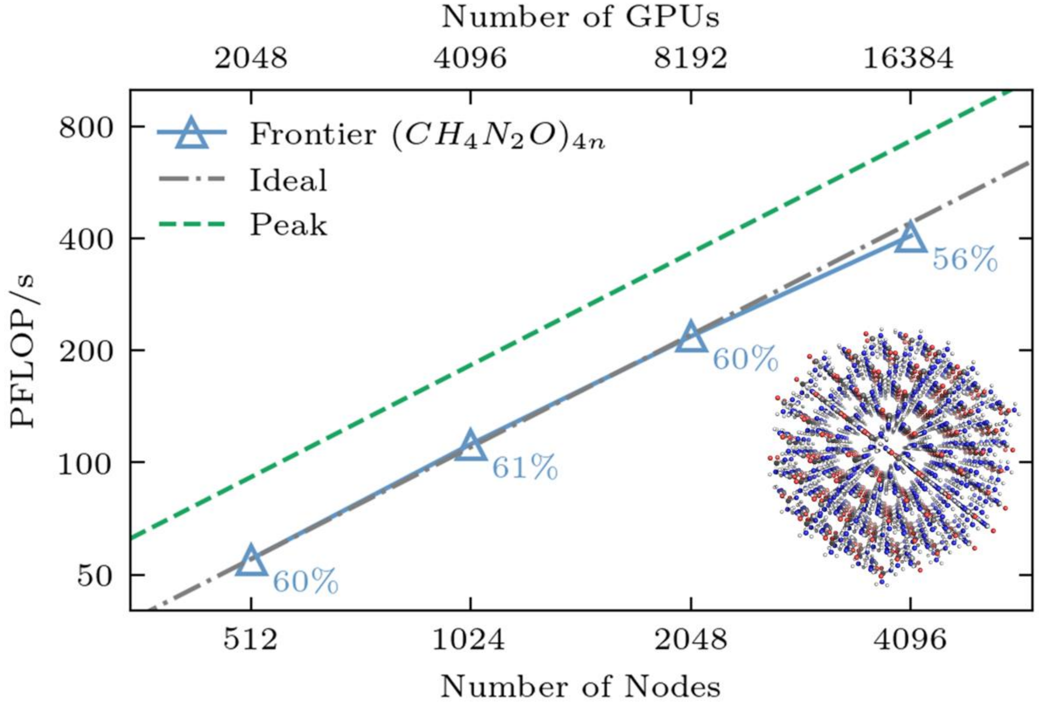
- **RTAT** is a wrapper around BLAS that **automatically experiments** to find the **best execution strategy** for each LA problem.
- Experiments are at **runtime** and ***in situ***; no redundant BLAS calls are performed.



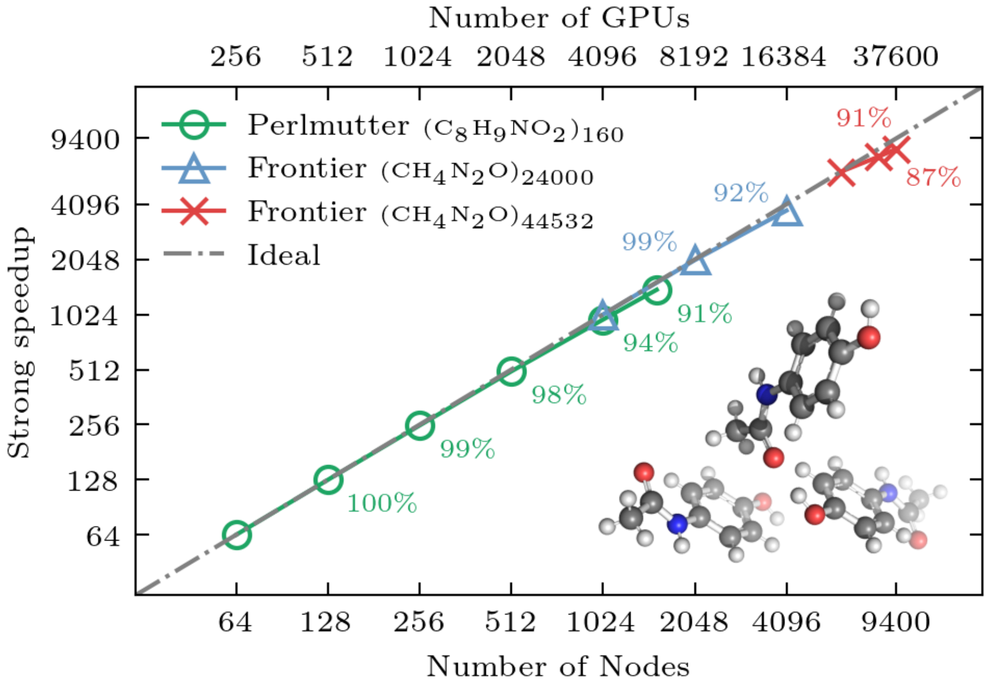
- **It is not straightforward to run DGEMMs at peak on AMD**
- Linear Algebra (LA) calculations can be performed through several different sequences of library calls
- Performance can vary drastically with execution strategy
- **Performance can be improved by finding the correct strategy**

PARALLEL SCALABILITY

WEAK SCALING



STRONG SCALING



- 9,408 nodes
- 75,776 GCDs

- 1,536 nodes
- 6,184 GPUs

- **Molecular Systems:** paracetamol, ibuprofen, and urea crystal structures
- **Weak Scaling**
  - **Percentages are with respect to FP64 R-Peak**
  - With a suitable balance of workload, timestep latency and resources, we can run at 60% of peak!

- **Strong Scaling**
  - Nearly ideal scaling
  - Largest system (×) 232k atoms, 1.024 million electrons, on 9400 nodes,
  - Little loss of parallel efficiency on 9400 nodes due sufficient workload

All calculations done in double-precision at MBE3/RI-HF+RI-MP2/cc-pVDZ level of theory (no frozen core)

PARALLEL SCALABILITY & FLOP PERFORMANCE

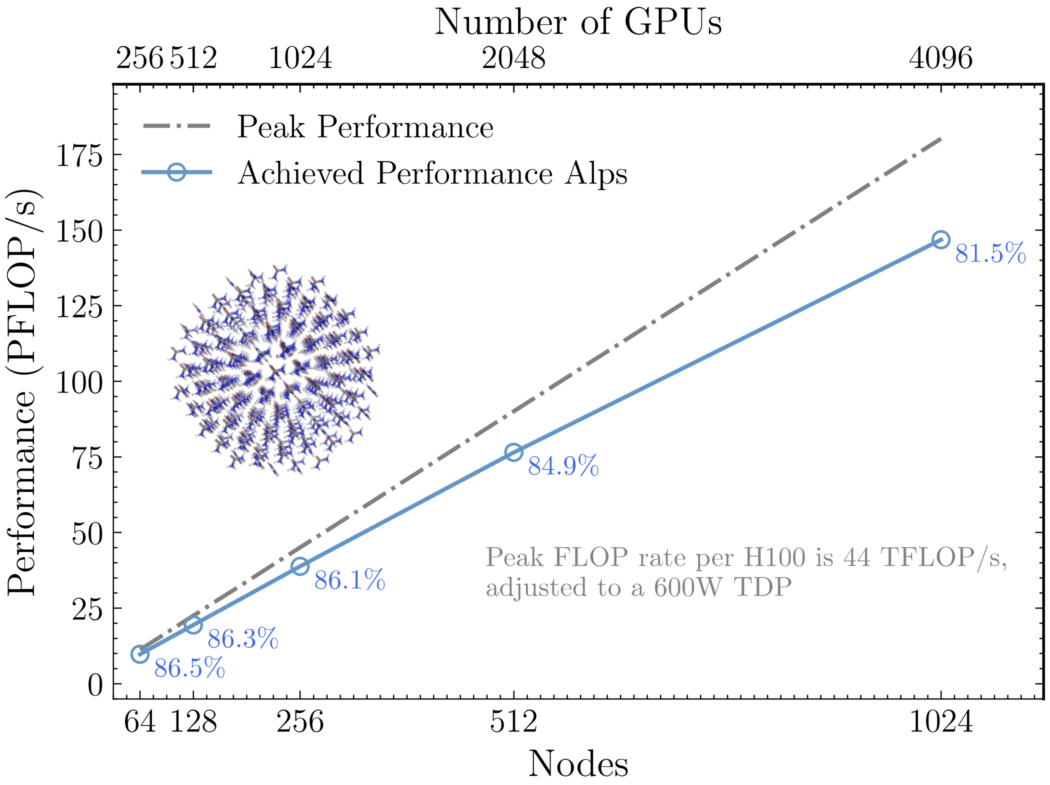
PERFORMANCE MEASURES

☞ FLOP counts obtained counting **only DGEMM FLOPs**, *i.e.*, $2 \times m \times k \times n$, where m, k, n are the matrix dimensions

☞ Provides a **lower-bound** on total **FLOPs**

☞ Runtime measured at the beginning of each time step in addition to rank local timings of every fragment calculation.

☞ FLOP rates obtained dividing FLOP count by wall time for the whole program execution



Peak FLOP rate per H100 is 44 TFLOP/s, adjusted to a 600W TDP

▻ 1,024 nodes

▻ 4,096 GH200

☞ **Molecular Systems:** 2BEG protein and urea crystal structures

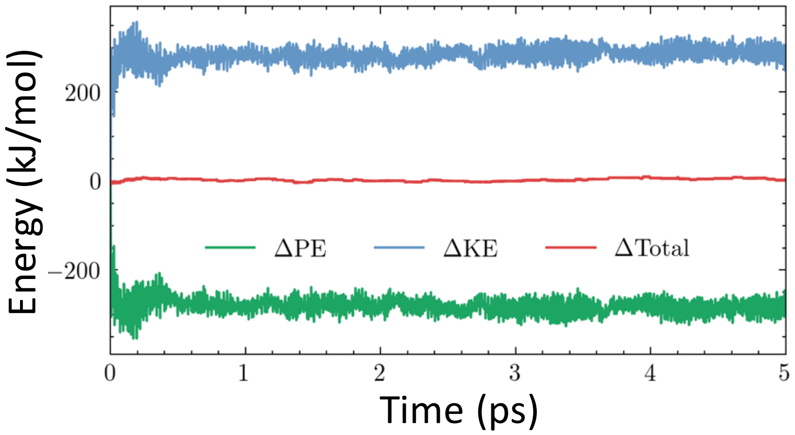
☞ **Percentages** are with respect to **FP64 R-peak**

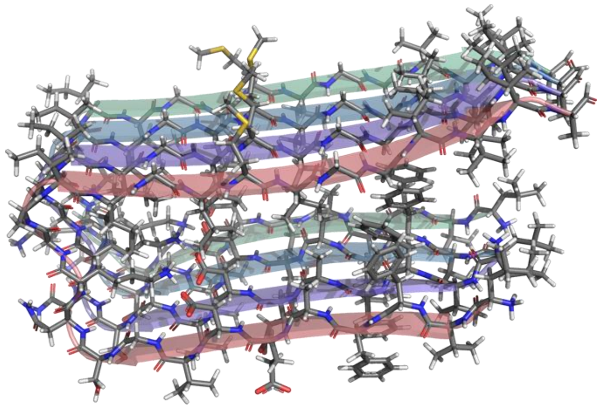
☞ **81.5% of FP64 R-peak on 4,096 GH200**

All calculations done in double-precision at MBE3/RI-HF+RI-MP2/cc-pVDZ level of theory (no frozen core)

RECORD TIME STEP LATENCY IN AIMD

- **Simulate the folding and misfolding processes of amyloid fibrils**, specifically targeting the Aβ (beta-amyloid) fibril PDB ID: 2BEG.
- Aβ fibril formation is a **hallmark of Alzheimer's pathology**, with misfolded fibrils aggregating into plaques that disrupt cellular functions in the brain.
- **Force fields have consistently failed** to capture the complex folding dynamics of Aβ fibrils, primarily due to the process being governed by non-covalent interactions, including hydrogen bonding, π-π stacking, and van der Waals forces.
- 2BEG includes 1,496 atoms and 5,504 electrons, presenting vast computational demands and requiring high-accuracy modelling of electronic effects that influence stability and folding.



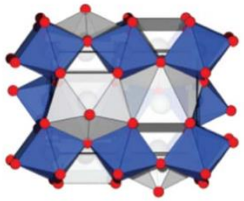


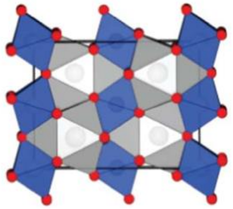
- 1,024 nodes
- 4,096 A100 GPUs
- 3.4 s/timestep (25 ps/day)
- **≫1000× faster than SOTA**

- 1,024 nodes
- 4,096 GH200 Superchips
- **1.03 s/timestep (83.9 ps/day)**
- **≫1000× faster than SOTA**

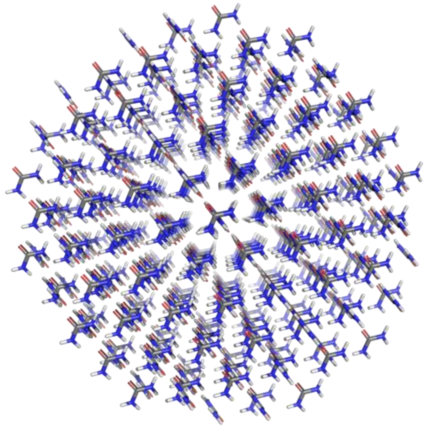
All calculations done in double-precision at MBE3/RI-HF+RI-MP2/cc-pVDZ level of theory (no frozen core)

BREAKING THE MILLION-ELECTRON & EFLOP/s (FP64) BARRIERS





- **Predict polymorphic** (multiple crystalline) **forms of therapeutics and organic compounds**
- Urea and paracetamol chosen for their academic and industrial relevance(pharmaceuticals, cosmetics, and solvent production).
- Both compounds display polymorphism influencing key properties like solubility, dissolution, and drug efficacy.
- **Challenge in Prediction**: Polymorph lattice energies differ by a few kJ/mol—requiring high accuracy.
- **Relevance of Non-Covalent Interactions**: Stability of crystal lattices in these biomolecules is dominated by non-covalent interactions, an area where hybrid DFT methods struggle.



- 9,408 nodes
- 75,776 GCDs

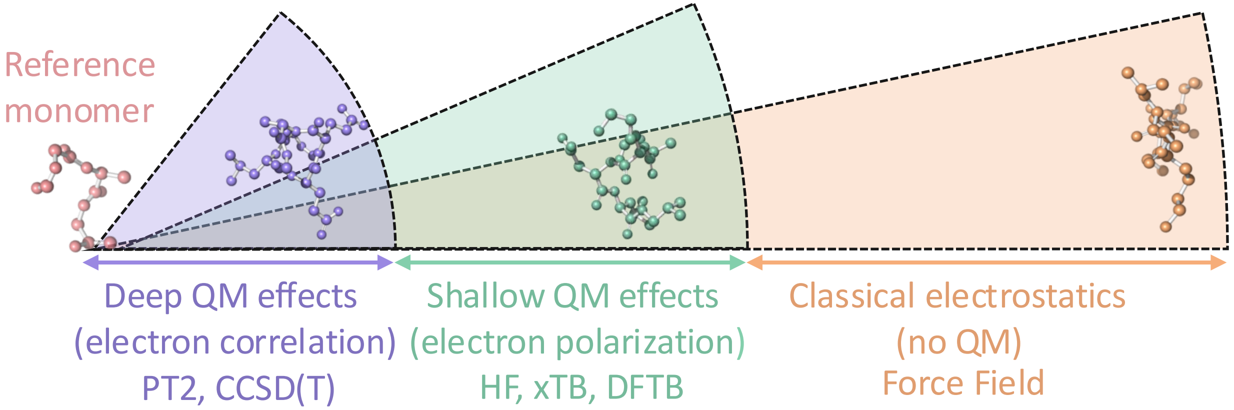
RECORD SIZE & PERFORMANCE

- Largest crystal included **510,832 atoms, 2,043,328 electrons**
- **>1000× larger than SOTA**
- **Using 9400 nodes obtained 1.007 EFLOP/s performance, 59% of FP64 R-Peak**
- **1st time breaking EFLOP/s barrier fully in FP64 (double-precision)**

All calculations done in double-precision at MBE3/RI-HF+RI-MP2/cc-pVDZ level of theory (no frozen core)

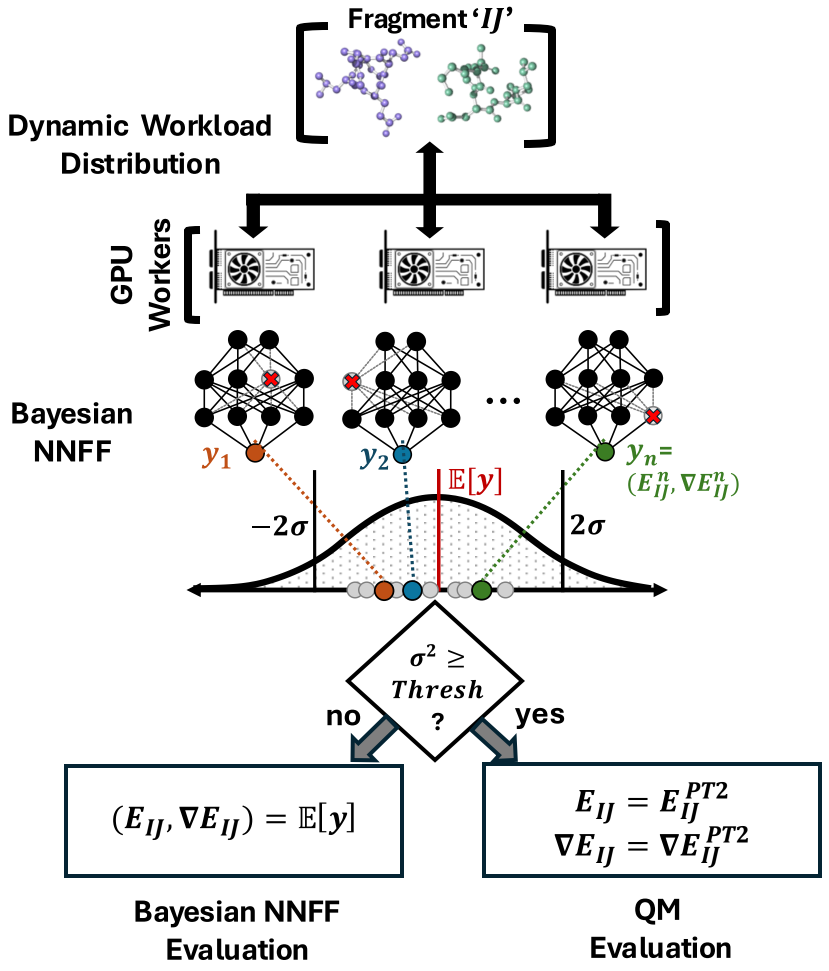
FURTHER ACCELERATION

MULTI-LAYER MOLECULAR MECHANICS



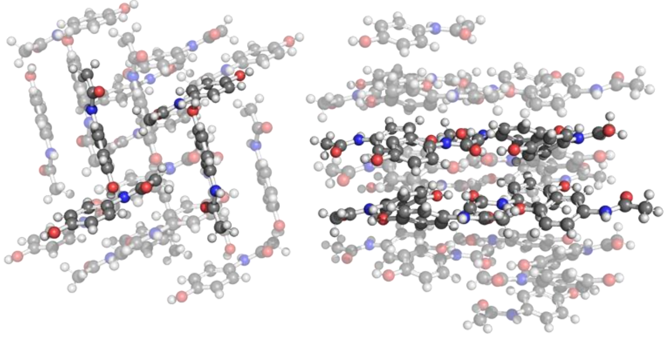
- Current scheme evaluates **all fragment interactions at the MP2 level**
- Large time savings can be obtained by **treating fragment interactions in a multi-layer hierarchical way based on distance**
- Close fragments require higher level of theory, while distant ones can be treated even classically (ONIOM style)
- In development an **adaptive hybrid quantum-AI (QAI) AIMD simulator**
- **Fragments** are treated with **either QM or BNNFFs**, trained on quantum-level data, based on prediction uncertainty.
- **BNNFF can actively learn from QM**, lowering uncertainty and accelerating large/long simulations

ADAPTIVE HYBRID QM/ML

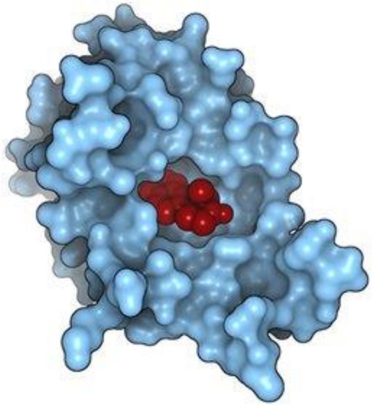


DoE INCITE awarded!

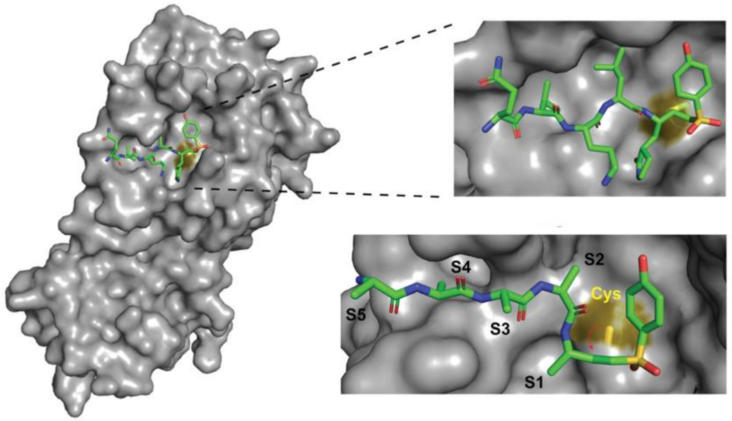
SOME EXCITING APPLICATIONS



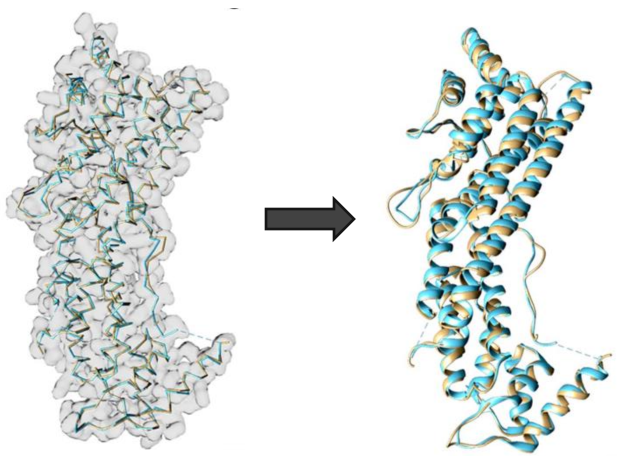
Polymorphism and Crystal Lattice Energies



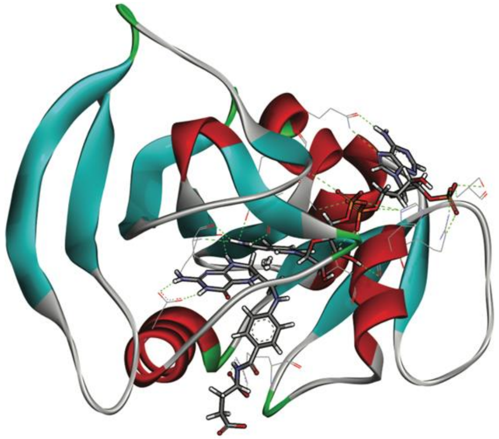
High-Accuracy Design of Non-Covalent Therapeutics



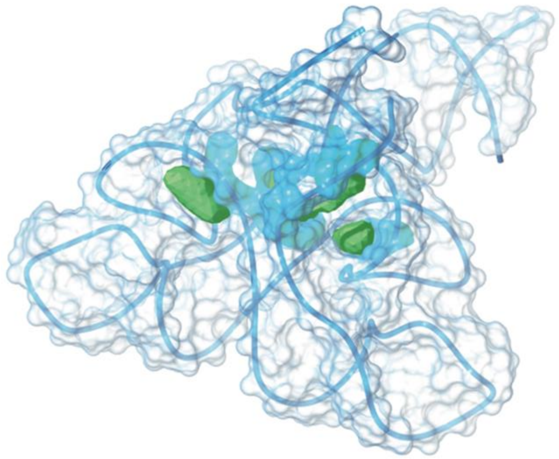
Covalent Therapeutics Reaction Mapping and Design



X-Ray Electron Density Resolution (more accurate Crystal Structures)

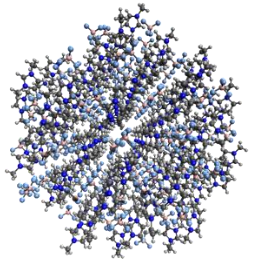


Enzymatic Reaction Mapping & Enzyme Design

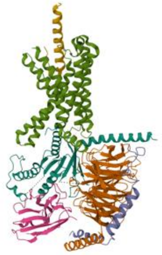


Small Molecule Drug Design Targeting RNA

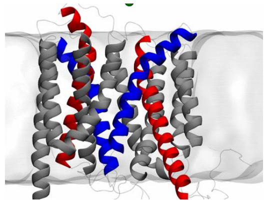
ADDITIONAL CAPABILITIES IN EXESS (ON GPU)



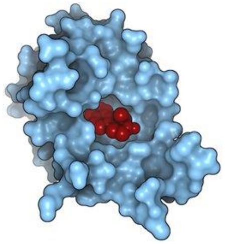
Crystal Lattice Energies



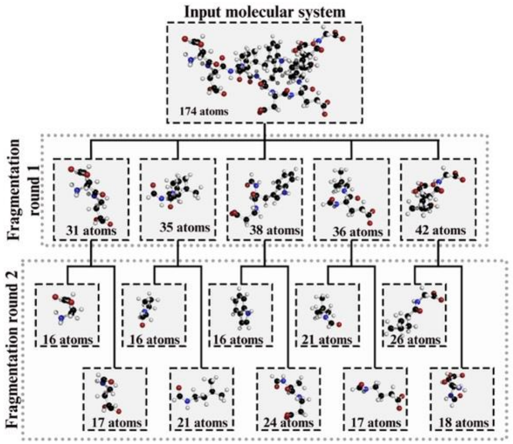
Ligand-Protein Binding Affinities



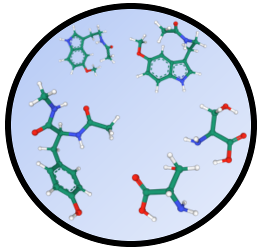
Ab Initio Molecular Dynamics



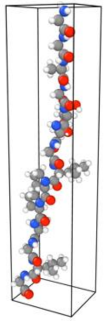
PBSA implicit solvent



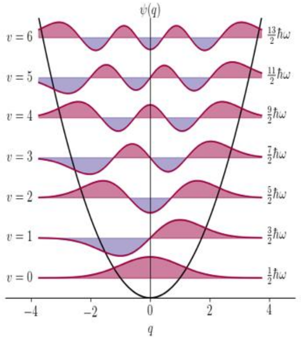
Automatic Molecular Fragmentation



GGA, meta-GGA Hybrid DFT, Regularized MP2



Geometry Optimization



Numerical Hessians

High Angular Momentum HF/DFT
(g functions, for RI already available)

Coupled Cluster [CCSD(T)]

Neural Network Force Fields

Polarizable Continuum Models

QM/MM

Range-separated DFT & Double Hybrids

Analytical Hessians

Transition State Search

Ab Initio Meta-Dynamics

Available

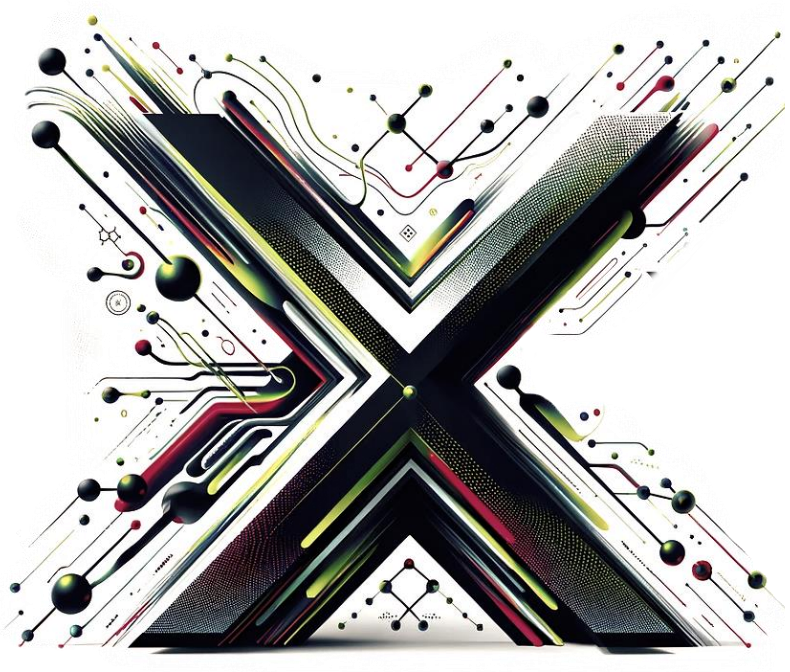
Under Development

EXESS is currently being released — free for academics — on the major HPC platforms

CONCLUDING REMARKS

EXESS

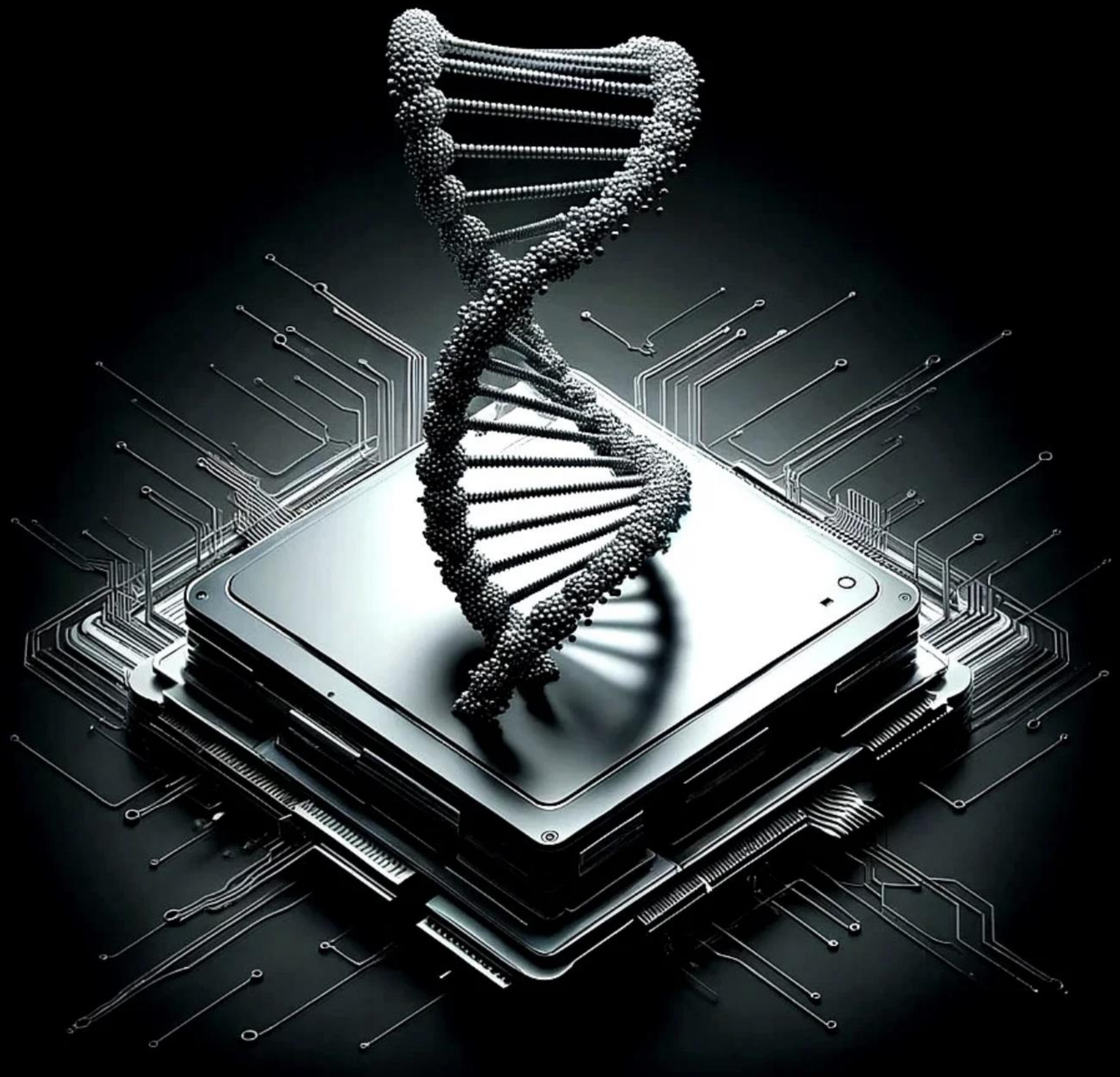
https://exess.qdx.co



- **Quantum Chemistry at Scale**: Performed the largest-ever AIMD simulations using MP2 potentials, modelling systems with up to over 2 million electrons, >1,000× larger than prior state-of-the-art.
- **Record-Breaking Performance**: Achieved 1,006.7 PFLOP/s on Frontier, utilizing 59% of its FP64 R-Peak, and broke the 1 EFLOP/s barrier for the first time.
- **Excellent Scalability**: Near-perfect strong and weak scaling across thousands of GPUs, showcasing the versatility and adaptability of the computational framework for current and future exascale systems.
- **Record Time to Solution**: Achieved a timestep latency of 1.03 s for a >5.5k electron protein using 4,096 GH200s, >1,000× faster than state-of-the-art.
- **Direct Impact on Science and Society** : Enabling to tackle grand challenges in in drug discovery, enzymatic catalysis, and biomolecular science, from polymorphism and Alzheimer’s disease, to the design of covalent therapeutics.
- **Vision for the Future**: This work not only pushes the limits of what is computationally possible but also sets the stage for the next generation of quantum-AI simulations, enhancing capabilities for real-world challenges.
- **Serving the Community**: **EXESS is available free of charge for the academic community!**

THE BARCA GROUP

HIGH-PERFORMANCE COMPUTING, AI & DIGITAL CHEMISTRY



www.barcagrp.com

QDX

EXESS

https://exess.qdx.co

Postdocs
Jorge Galvez-Vallejo

Undergrad Students
Brendan Wilson
Monique Jeacocke

Acknowledgments
Dmytro Bykov (Oak Ridge)
Jakub Kurzak (AMD)

PhD students
Fazeleh Kazemian
Fiona Yu
Calum Snowdon
Joshua Soon
Elise Palethorpe
Ryan Stocks
Yufan Xia

Openings

We are looking for two PhD students and one Postdoc in AI and HPC applied to digital chemistry