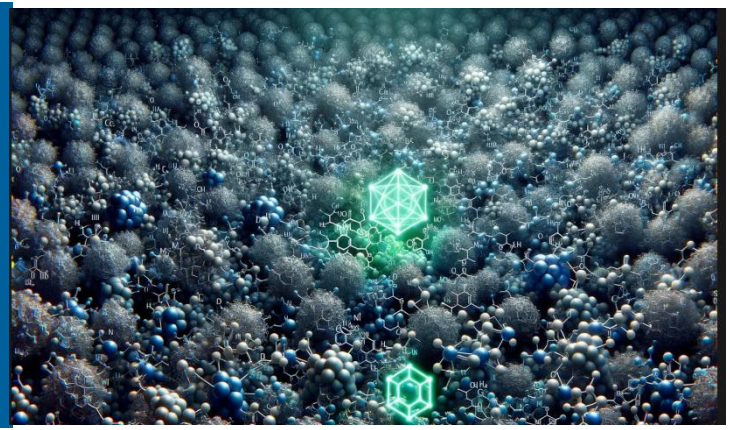
Artificial Intelligence for Materials Science (AIMS)

NIST, July 17th to 18th, 2024

INCREASING AI/ML PREDICTIONS THROUGH DMC-ENHANCED DELTA LEARNING



QMCPACK

A. BENALI

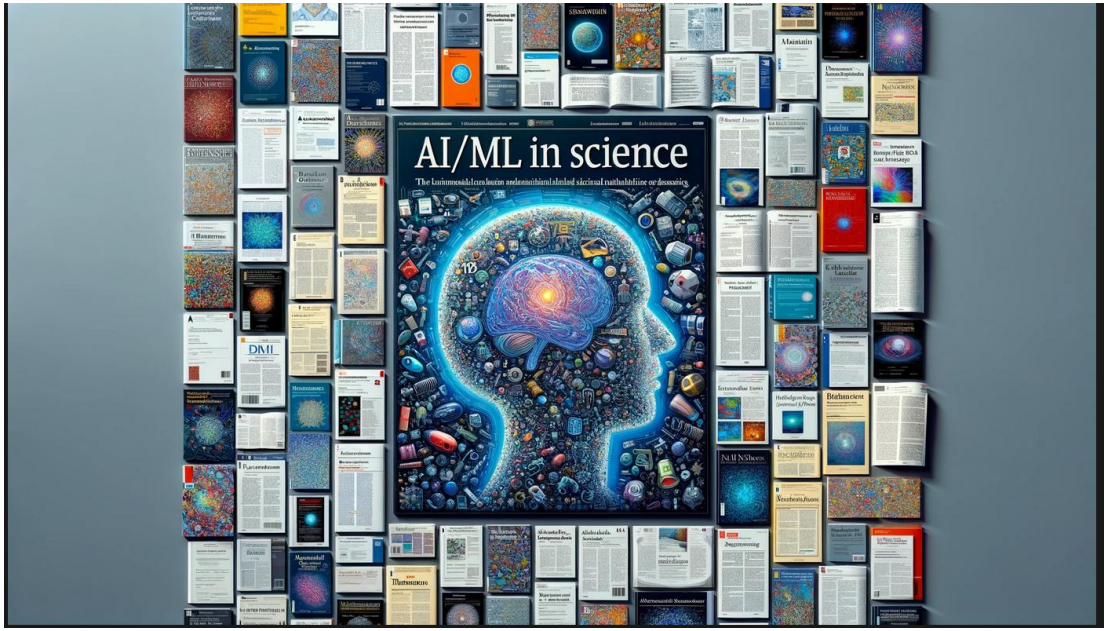
Group Lead CPS
Computational Science Division / Material Science Division
Argonne National Laboratory, Argonne IL, 60622, USA

CENTER FOR PREDICTIVE SIMULATION OF FUNCTIONAL MATERIALS

US-DEPARTMENT OF ENERGY - BES COMPUTATIONAL MATERIALS SCIENCES PROGRAM

CPSFM | Center for Predictive Simulation of Functional Materials

FOUNDATION OF QUANTUM MODEL PRECISION



The predictive accuracy of quantum machine learning (QML) models trained on quantum chemistry data and used for the navigation of chemical compound space (CCS) is inherently limited by the predictive accuracy of the approximations used within the underlying quantum theory.

HOW ACCURATE IS DIFFUSION MONTE CARLO (DMC)

Many Body Method

- Solves explicitly many electrons interactions.
- Accounts explicitly for strong correlations and dispersion forces

Fixed Node Approximation:
We introduce a guiding/trial function $\psi_G(\boldsymbol{R})$ which closely approximates the ground state.

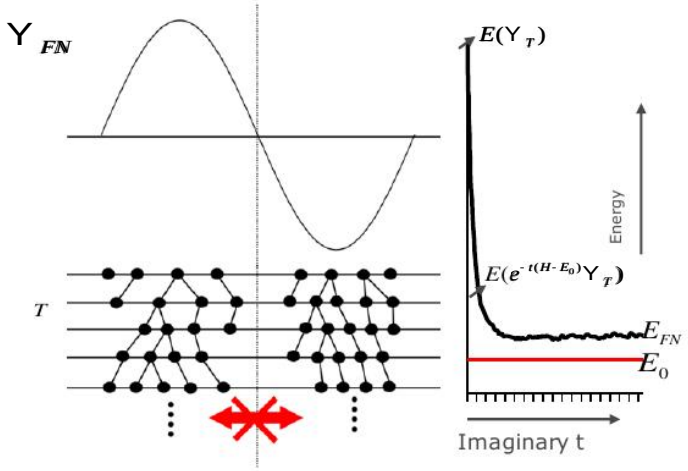
$$f(\boldsymbol{R},\tau) = \psi_G(R)\psi(\boldsymbol{R},\tau)$$

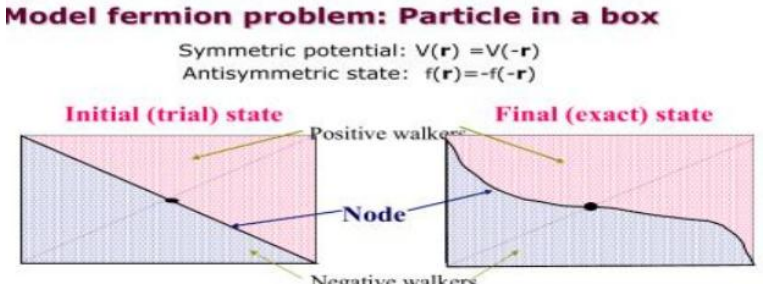
DMC is variational, and upper bound to the exact nodal surface

$$-\frac{\delta f(\boldsymbol{R},\tau)}{\delta\tau} = \left[\sum_{i=1}^{N} -\frac{1}{2}\nabla^2{}_i f(\boldsymbol{R},\tau)\right] - \nabla.\left[\frac{\nabla\psi(\boldsymbol{R})}{\psi(\boldsymbol{R})} f(\boldsymbol{R},\tau)\right] + (E_L(\boldsymbol{R}) - E_T) f(\boldsymbol{R},\tau)$$

Quality of the calculation depends on the quality of nodal surface. Can be systematically improved (by increasing complexity of the guiding function)

electron : Fermion
(anti-symmetric wavefunction)

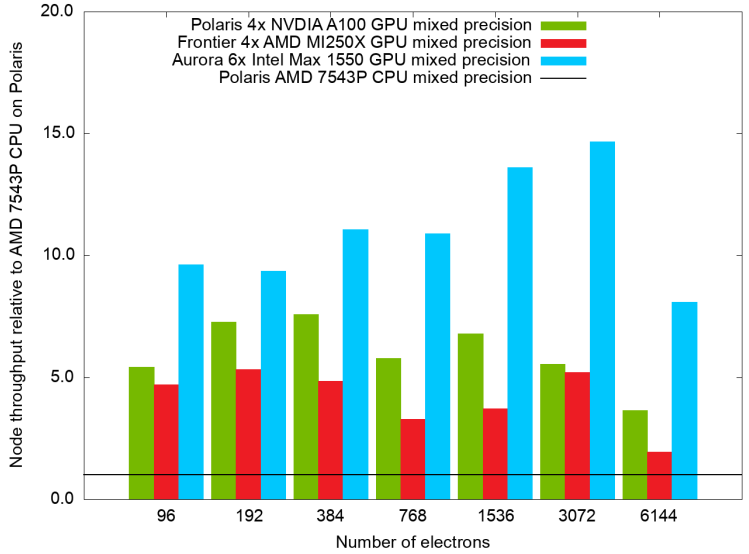




Argonne NATIONAL LABORATORY

QMCPACK

EXASCALE COMPUTING PROJECT



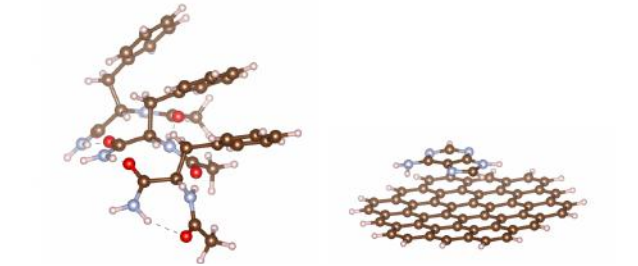
Open source code: www.qmcpack.org
CPU and GPU implementation (github repo)
~400 pages of manual (implementation + theory + exercises)
Google group for support

Batched GPU Code on Aurora (Intel) Achieves 3X speed up Compared to Polaris (NVIDIA) and Frontier (AMD). **15X compared to CPU**

[1] J. Kim et al. "Qmcpack simulation suite." Journal of Physics: Condensed Matter, Volume 30, Number 19 (2018)
[2] PRC. Kent et al. "QMCPACK: Advances in the development (..) " The Journal of chemical physics 152 (17), 174105 (2020)

OAK RIDGE National Laboratory · Argonne NATIONAL LABORATORY · Sandia National Laboratories · Lawrence Livermore National Laboratory · NC STATE UNIVERSITY · Argonne NATIONAL LABORATORY

L7 Benchmark (van der Waals dominated molecules)



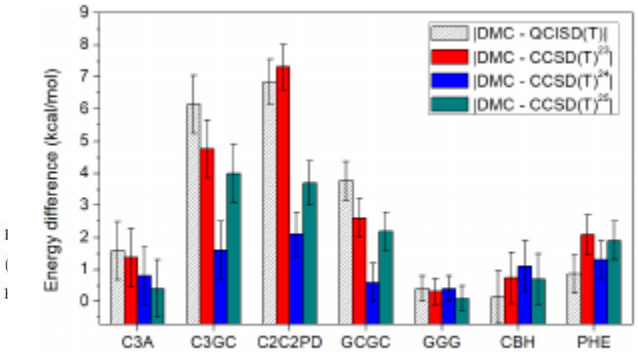


FIG. 1. Binding energy differences in kcal/mol relative to QCISD(T) and CCSD(T).

- DMC in all electrons (up to 5000 MO) at CBS limit.
- CCSD(T) with ecp and approximated CBS.

method	C3A	C3GC	C2C2PD	GCGC	GGG	CBH	PHE
QCISD /CBS	-14.52	-24.79	–	–	-1.30	-9.53	-24.23
QCISD(T) /CBS	-18.19	-31.25	-24.36	-14.37	-2.40	-11.06	-25.76
CCSD(T) [cc-pVDZ-F12][21]	–	–	-19.14	-13.69	-2.36	-11.13	-25.09
CCSD(T)[22] [aug-cc-pVTZ]	–	–	-18.87	–	–	–	–
CCSD(T) / CBS [def2-TZVPP][23]	-17.98	-29.86	-24.81	-13.21	-1.68	-11.64	-22.81
CCSD(T) / CBS [aug-cc-pVDZ][24]	-15.80	-26.70	-19.60	-11.20	-1.60	-9.80	-23.60
CCSD(T) / CBS [mTZVP][25]	-17.0	-29.1	-21.2	-12.8	-1.9	-11.6	-23.0
CCSD(T) / CBS[26]	-	-	-23.32	-13.80	-2.22	-11.48	-25.01
DMC /def2-QZVP	-16.6(9)	-25.1(9)	-17.5(7)	-10.6(6)	-2.0(4)	-10.9(8)	-24.9(6)

	CASINO[1]	QMCPACK[2]	Δ
GGG	-1.5+/- 0.3	-2.0 +/- 0.4	0.5 +/- 0.5
CBH	-11.4 +/- 0.4	-10.9 +/- 0.8	0.5 +/- 0.9
PHE	-26.5 +/- 0.7	-24.9 +/- 0.6	1.6 +/- 0.9
C3A	-15.0 +/- 0.5	-16.6 +/- 0.9	1.6 +/- 1.0
GCGC	-12.3 +/- 0.3	-10.6 +/- 0.6	1.7 +/- 0.7
C3GC	-24.2 +/- 0.7	-25.1 +/- 0.9	0.9 +/- 1.1
C2C2PD	-18.1 +/- 0.4	-17.5 +/- 0.7	0.6 +/- 0.8

[1] “Interactions between Large Molecules: Puzzle for Reference Quantum-Mechanical Methods” Y. S. Al-Hamdani, P. R. Nagy, D. Barton, M. Kallay, J.G. Brandenburg, A. Tkatchenko, Nature Communications volume 12, Article number: 3927 (2021)

[2] “Quantum Monte Carlo benchmarking of large noncovalent complexes in the L7 benchmark set” A. Benali, H. Shin, O. Heinonen, J. Chem. Phys. **153**, 194113 (2020)

Argonne NATIONAL LABORATORY

L7 Benchmark (van der Waals dominated molecules)

Understanding Discrepancies of Wavefunction Theories for Large Molecules

Tobias Schäfer[1*†], Andreas Irmler[1*†], Alejandro Gallo[1] and Andreas Grüneis[1*]

[1]Institute for Theoretical Physics, TU Wien, Wiedner Hauptstraße 8–10/136, Vienna, A-1040, Austria.

Table 2: Comparison of the interaction energy for large molecular complexes in kcal/mol as calculated by different levels of theory. Showcasing partially large discrepancies between CCSD(T) and DMC on the one hand, and an excellent agreement between CCSD(cT) and DMC results for complexes up to the 100-atom scale on the other hand. CCSD(T) and CCSD(cT) results are obtained using our plane wave approach. The calculation and the uncertainty of CCSD(cT)-fit is explained in the supplementary information.

System	CCSD(T)	LNO-CCSD(T) [11]	CCSD(cT)	CCSD(cT)-fit	DMC [11]	DMC [10]
GGG	-1.5 ± 0.5	-2.1 ± 0.2	-1.2 ± 0.5	-1.8 ± 0.2	-1.5(6)	-2.0(8)
GCGC	-13.1 ± 0.5	-13.6 ± 0.4	-12.5 ± 0.5	-12.8 ± 0.5	-12.4(8)	-10.6(12)
C2C2PD	-21.1 ± 0.5	-20.6 ± 0.6	-19.3 ± 0.5	-18.9 ± 0.7	-18.1(8)	-17.5(14)
C3A		-16.5 ± 0.8		-15.3 ± 0.9	-15.0(10)	-16.6(18)
PHE		-25.4 ± 0.2		-25.0 ± 0.2	-26.5(13)	-24.9(12)
C3GC		-28.7 ± 1.0		-26.7 ± 1.1	-24.2(13)	-25.1(18)
C_{60}@[6]CPPA		-41.7 ± 1.7		-35.6 ± 2.0	-31.1(14)	

arXiv:2407.01442v2 [physics.chem-ph] 4 Jul 2024

IS DMC ROBUST ENOUGH TO BE USED WITH QML

Acknowledgments

JCTC Journal of Chemical Theory and Computation

pubs.acs.org/JCTC Article

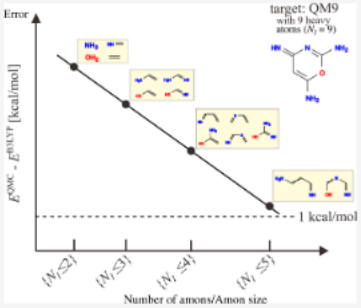
Toward DMC Accuracy Across Chemical Space with Scalable Δ-QML

Bing Huang,* O. Anatole von Lilienfeld,* Jaron T. Krogel,* and Anouar Benali*

Cite This: *J. Chem. Theory Comput.* 2023, 19, 1711−1721 Read Online

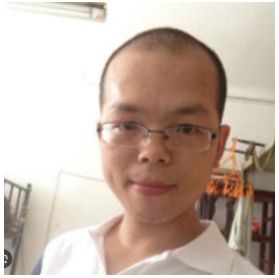
ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: In the past decade, quantum diffusion Monte Carlo (DMC) has been demonstrated to successfully predict the energetics and properties of a wide range of molecules and solids by numerically solving the electronic many-body Schrödinger equation. With $O(N^3)$ scaling with the number of electrons N, DMC has the potential to be a reference method for larger systems that are not accessible to more traditional methods such as CCSD(T). Assessing the accuracy of DMC for smaller molecules becomes the stepping stone in making the method a reference for larger systems. We show that when coupled with quantum machine learning (QML)-based surrogate methods, the computational burden can be alleviated such that quantum Monte Carlo (QMC) shows clear potential to undergird the formation of high-quality descriptions across chemical space. We discuss three crucial approximations necessary to accomplish this: the fixed-node approximation, universal and accurate references for chemical bond dissociation energies, and scalable minimal amons-set-based QML (AQML) models. Numerical evidence presented includes converged DMC results for over 1000 small organic molecules with up to five heavy atoms used as amons and 50 medium-sized organic molecules with nine heavy atoms to validate the AQML predictions. Numerical evidence collected for Δ-AQML models suggests that already modestly sized QMC training data sets of amons suffice to predict total energies with near chemical accuracy throughout chemical space.



O. A. von Lilienfeld[1]

J. T. Krogel[2]



B. Huang[1]

1- University of Toronto
2- Oak Ridge National Laboratory

Argonne NATIONAL LABORATORY

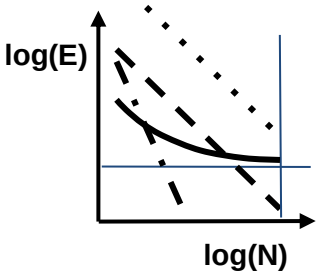
Atom in a Molecule: “AM-on”

Elementary → **building blocks** → **structure** → **effect**

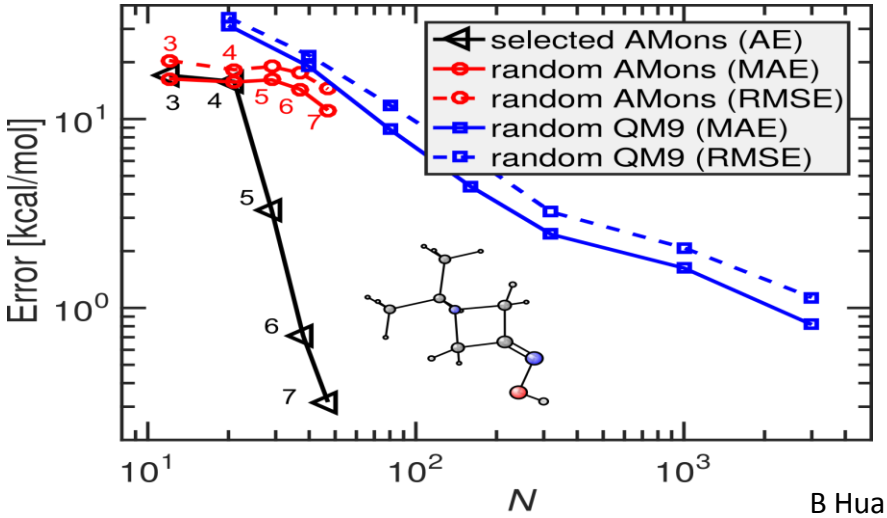
letters → words → sentence → meaning

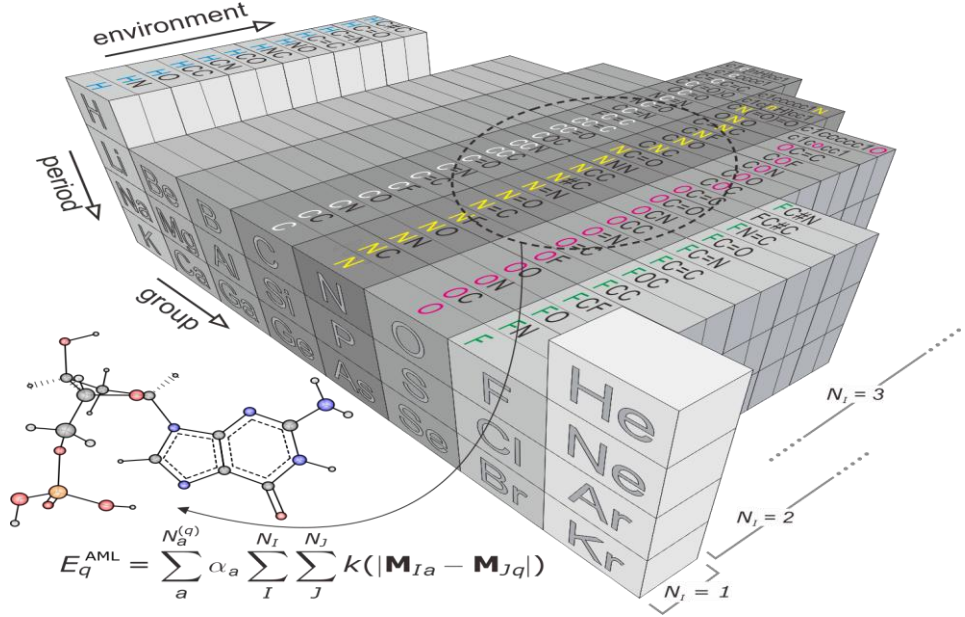
atoms → **AMONS** → molecule → property

Choice/Design of the AMONS is critical for efficiency



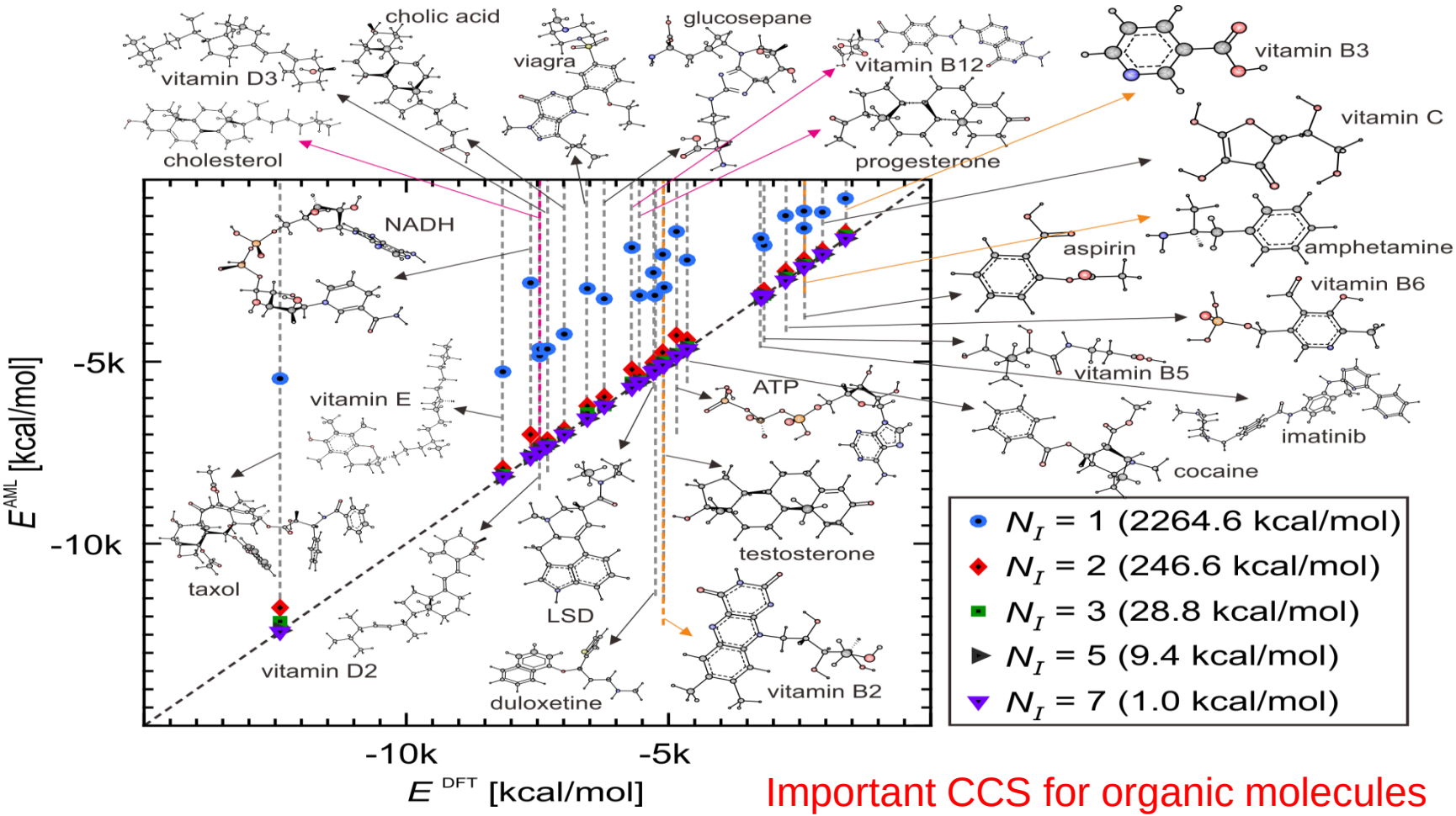
Represent molecules in terms of atoms in fragments





B Huang, OAvL, *Nature Chemistry* (2020), arXiv 1707.04146

Argonne NATIONAL LABORATORY

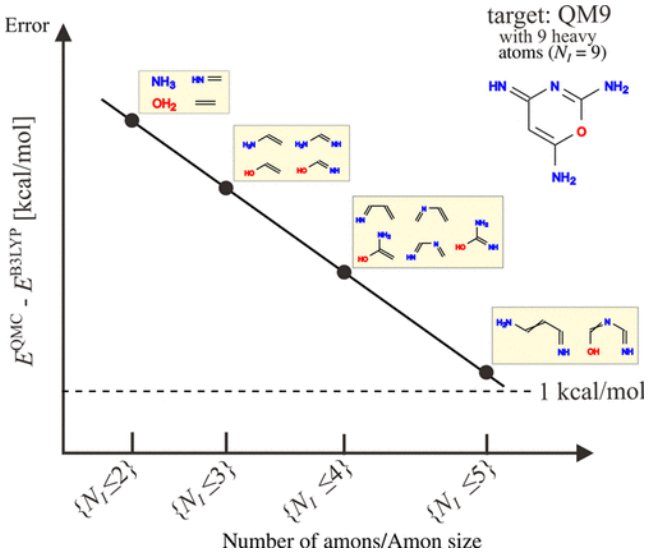


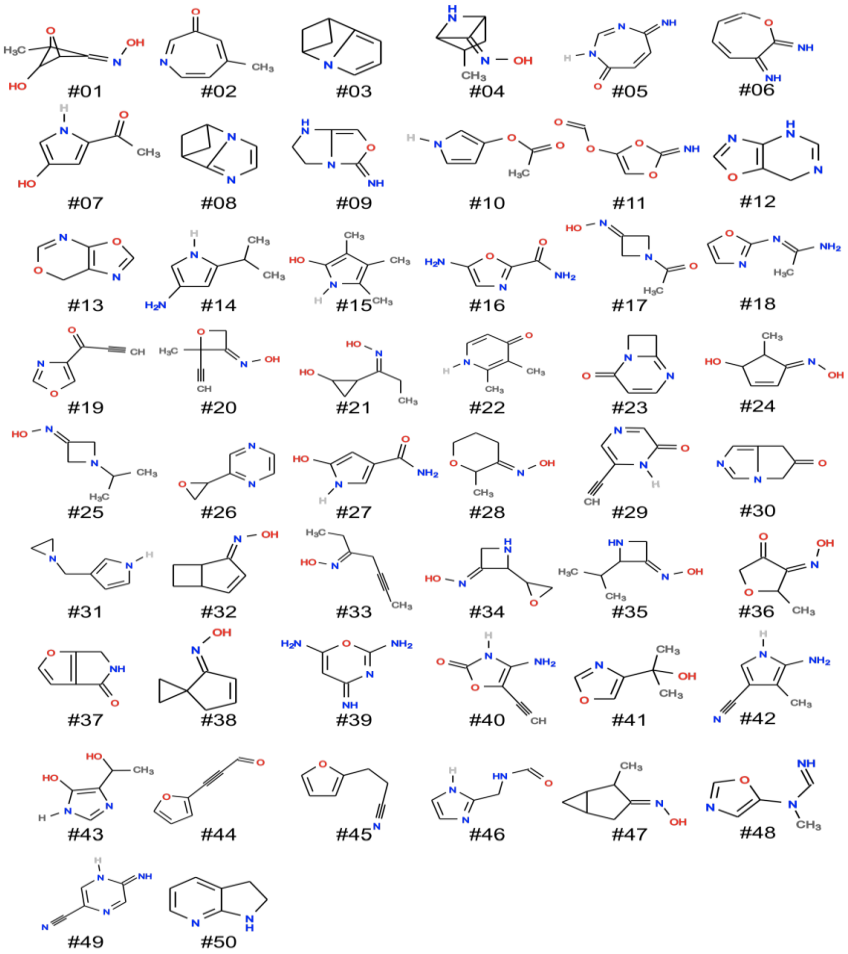
Important CCS for organic molecules

B Huang, OAvL, *Nature Chemistry* (2020), arXiv 1707.04146

Argonne NATIONAL LABORATORY

Amon based Δ-QML on QMC





Converged DMC estimates of atomization energies for 1'175 amons with $N \leq 5$ from amon dictionary …

and for 50 random QM9 molecules with $N = 9$

Huang, von Lilienfeld, Krogel, Benali *JCTC* (2023) https://arxiv.org/abs/2210.06430

Argonne NATIONAL LABORATORY

Amon based Δ-QML on QMC

To reduce magnitude of reference energies to learn, we use "dressed-atom" (DA), and its energy can be obtained through least-squares regression of a simple linear model.

n_A^i is the number of atoms of type A in the i[th] training molecule and ε_A corresponds to the dressed-atom energy of A

$$E_i^{\text{ref}} = \sum_A n_A^i \varepsilon_A^{\text{ref}}$$

Training and test for (single-level) AQML models

$$E_i^{\text{ref}} - \sum_A n_A^i \varepsilon_A^{\text{ref}}$$

We combine ΔML with AQML

$$\Delta E_i^{\text{DMC}-\text{B3LYP}} = \sum_A n_A^i (\varepsilon_A^{\text{DMC}} - \varepsilon_A^{\text{B3LYP}})$$

DMC energy for any query molecule q

$$E_q^{\text{est,DMC}} = E_q^{\text{B3LYP}} + \Delta E_q^{\text{est,DMC}-\text{B3LYP}}$$

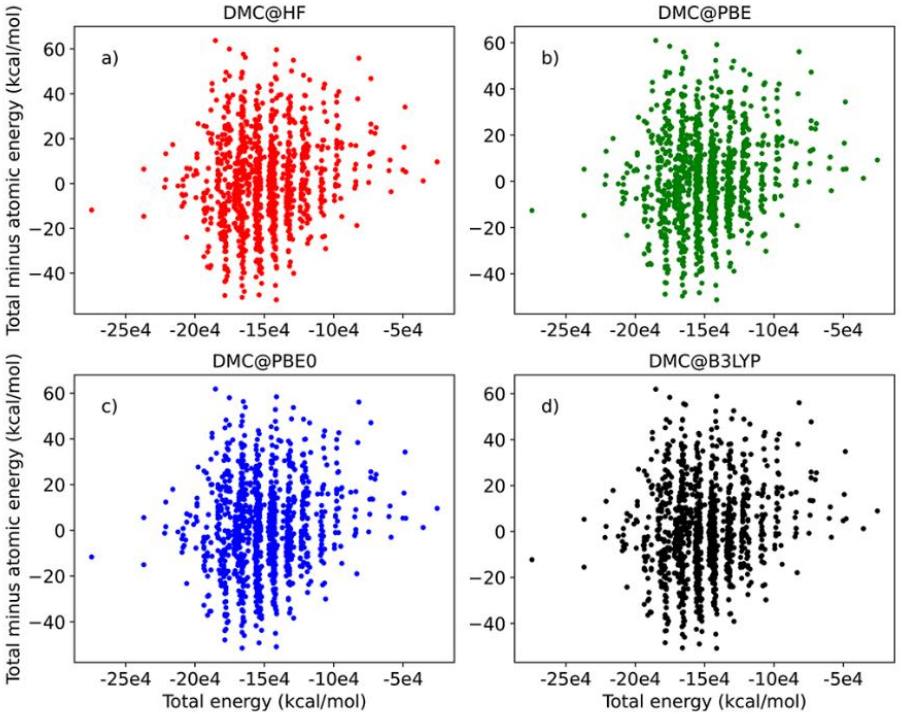
Huang, von Lilienfeld, Krogel, Benali *JCTC* (2023) https://arxiv.org/abs/2210.06430

DMC Nodal surface

Computational details:

- All electrons
- cc-pvQz basis set
- HF, PBE, PBE0 and B3LYP nodal surfaces
- Using qmcpack 3.5.0 (current version 3.18.0)
- Total computer time ~140k node/hours (KNL@ALCF)
- Time to solution: 11 to 40 min/compound (32 nodes 128 threads) – (~15x speedup compared to current GPU implementation)

DMC shows detailed agreement in the **residual energies** regardless of the source of nodes.



Total DMC molecular energies for the $\{N_I \leq 5\}$ set with the self-consistent dressed atomic energies removed $(E_i^{\mathrm{DMC}} - \sum_A n_A^i \epsilon_A^{\mathrm{DMC}})$ plotted vs the total energies.

Huang, von Lilienfeld, Krogel, Benali *JCTC* (2023) https://arxiv.org/abs/2210.06430

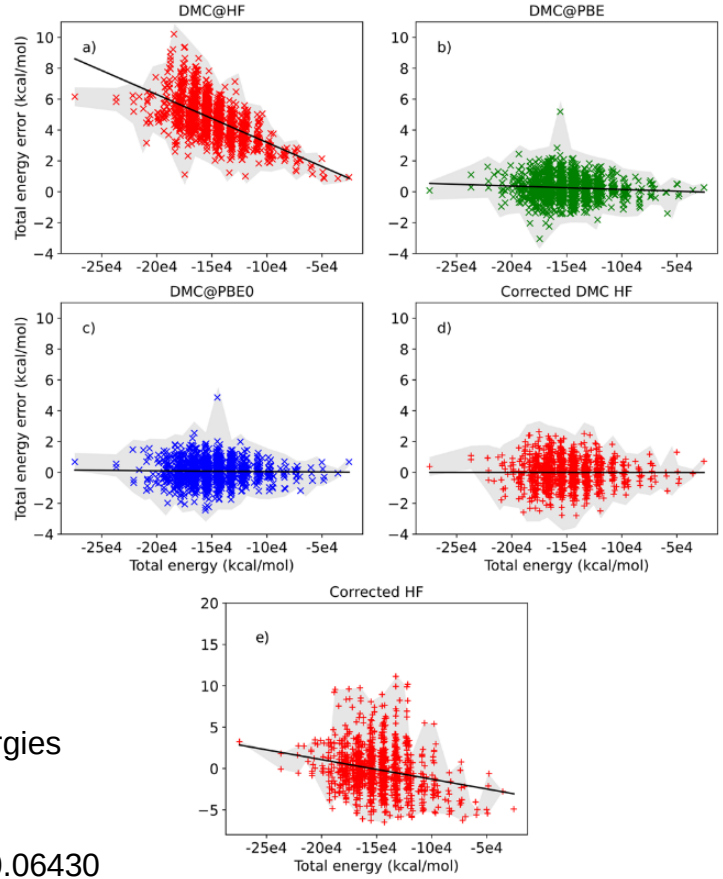
DMC Nodal surface

- Trial wavefunctions from PBE0, PBE, and HF produce energies that are higher than those of DMC@B3LYP by 0.08(2), 0.26(2), and 4.74(2) kcal/mol on average

- Corrected energies: adding the difference between the DMC@B3LYP and DMC@HF dressed atomic energies to the raw DMC@HF total energies (d) and the raw HF total energies (e).

- Small dependence in the nodal surface
- Using dressed atomic energies corrects significantly both HF and DMC@HF

Errors in DMC total energies as measured against the DMC@B3LYP reference for the $\{N_I \leq 5\}$ set.



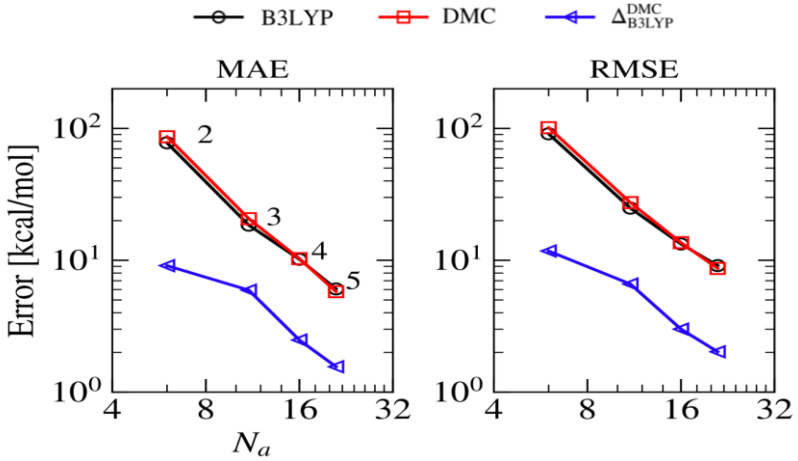
- Shaded region indicates the 1σ statistical uncertainties of the outlying energies
- Linear least-squares fits are shown in black.

Huang, von Lilienfeld, Krogel, Benali *JCTC* (2023) https://arxiv.org/abs/2210.06430

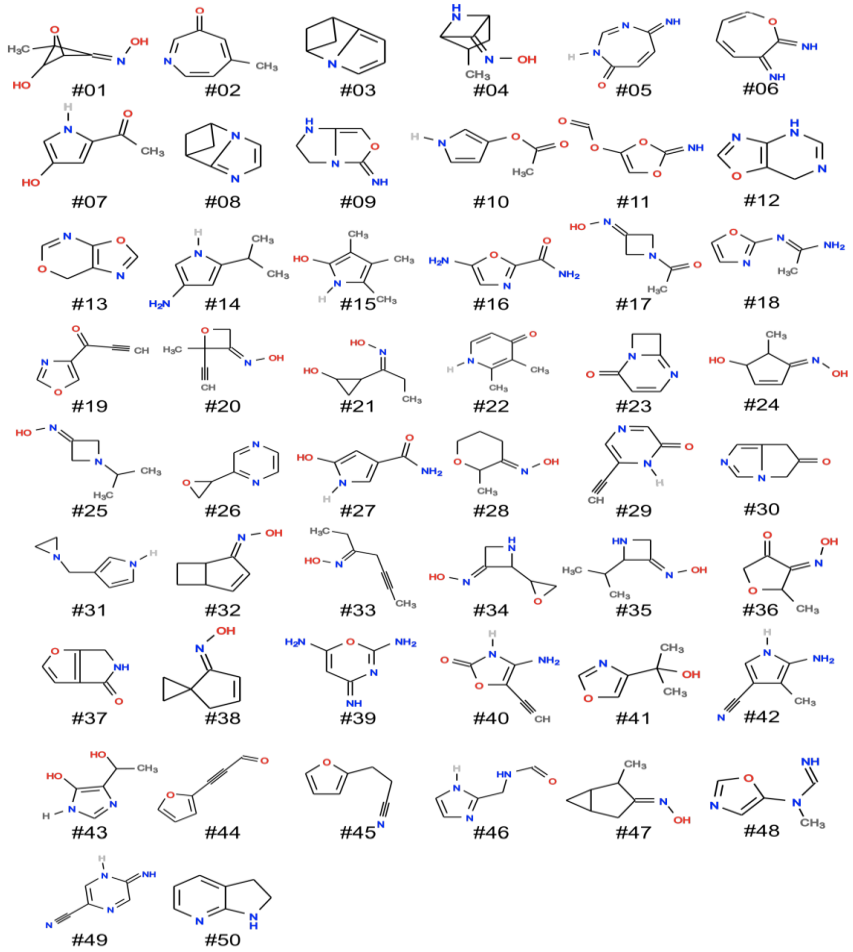
Amon based Δ-QML on QMC

Converged DMC estimates of atomization energies for 1’175 amons with $N \leq 5$ from amon dictionary …

and for 50 random QM9 molecules with $N = 9$



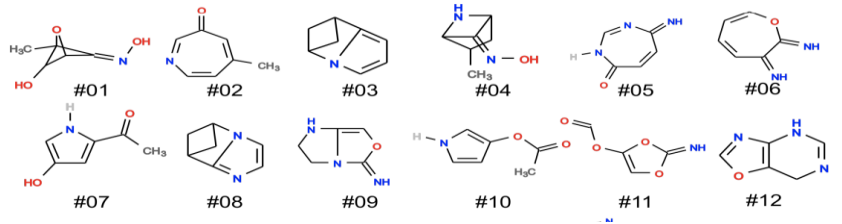
N_a : Number of amons used in the training



Huang, von Lilienfeld, Krogel, Benali *JCTC* (2023) https://arxiv.org/abs/2210.06430

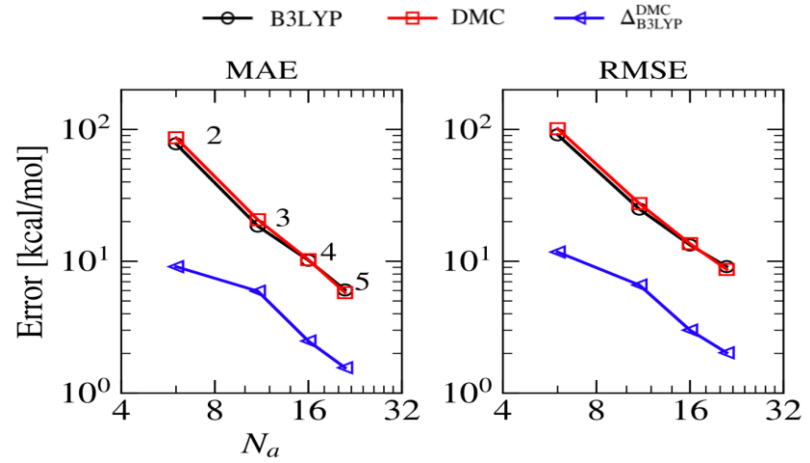
Argonne NATIONAL LABORATORY

Amon based Δ-QML on QMC



Converged DMC estimates of atomization energies for 1’175 amons with $N \leq 5$ from amon dictionary …

and for 50 random QM9 molecules with $N = 9$



N_a : Number of amons used in the training

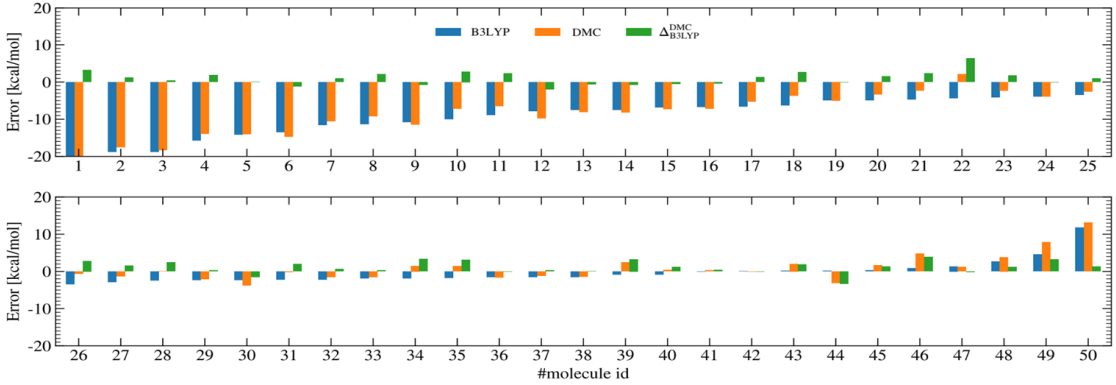
- AQML model leads to a steepening of learning curves, much steeper than QML models based on random sampling.
- Learning curves of DMC-trained AQML almost overlap with that of B3LYP-trained AQML
- Adding reference energies from B3LYP as the baseline to DMC-trained AQML shifts downwards learning curves significantly

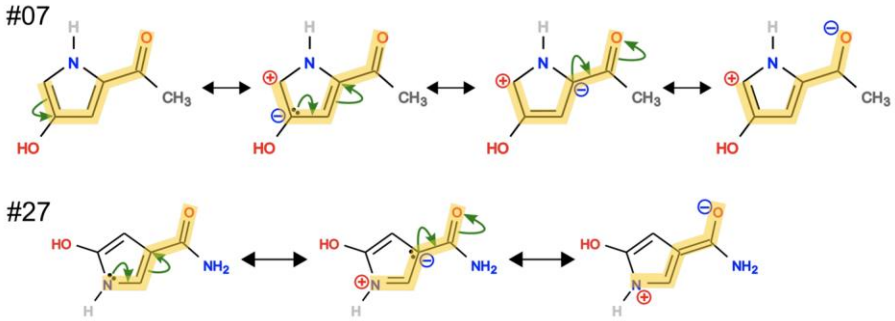
1.6kcal/mol MAE using only 20 DMC Amons selected on the fly.

Huang, von Lilienfeld, Krogel, Benali *JCTC* (2023) https://arxiv.org/abs/2210.06430

Argonne NATIONAL LABORATORY

Amon based Δ-QML on QMC





Conjugation path from drawing molecular resonance forms

- Molecules #21-#47 show absolute error less than 2kcal/mol
- Molecules #1-#20 experience either high internal strain and/or conjugation extending over the whole molecule (Requiring N_l=6)
- Amon data set used to train DMC did not include strained molecules

Increasing the number of heavy atoms is needed to increase accuracy

Huang, von Lilienfeld, Krogel, Benali *JCTC* (2023) https://arxiv.org/abs/2210.06430

Towards Larger Systems ~11k DMC calculations, 50k CPU hours. (~8k on Aurora GPU)

Towards comprehensive coverage of chemical space: Quantum mechanical properties of 836k constitutional and conformational closed shell neutral isomers consisting of HCNOFSiPSClBr

Danish Khan,[1,2] Anouar Benali,[3] Scott Y. H. Kim,[1,2] Guido Falk von Rudorff,[4,5] and O. Anatole von Lilienfeld[6,2,1,7,8,9,10,*]

[1] *Chemical Physics Theory Group, Department of Chemistry, University of Toronto, St. George Campus, Toronto, ON, Canada*
[2] *Vector Institute for Artificial Intelligence, Toronto, ON, M5S 1M1, Canada*
[3] *Computational Science Division, Argonne National Laboratory, Argonne, Illinois 60439, United States*
[4] *Institute of Chemistry, University of Kassel, 34109 Kassel, Germany*
[5] *Center for Interdisciplinary Nanostructure Science and Technology (CINSaT), 34132 Kassel, Germany*
[6] *Department of Materials Science and Engineering, University of Toronto, St. George Campus, Toronto, ON, Canada*
[7] *ML Group, Technische Universität Berlin and Institute for the Foundations of Learning and Data, 10587 Berlin, Germany*
[8] *Berlin Institute for the Foundations of Learning and Data, 10587 Berlin, Germany*
[9] *Department of Physics, University of Toronto, St. George Campus, Toronto, ON, Canada*
[10] *Acceleration Consortium, University of Toronto, Toronto, ON*

ABSTRACT

The Vector-QM24 (VQM24) dataset attempts to more comprehensively cover all possible neutral closed shell small organic and inorganic molecules and their conformers at state of the art level of theory. We have used density functional theory (ωB97X-D3/cc-pVDZ) to optimize 577k conformational isomers corresponding to 258k constitutional isomers. Isomers included contain up to five heavy atoms (non-hydrogen) consisting of *p*-block elements C, N, O, F, Si, P, S, Cl, Br. Single point diffusion quantum Monte Carlo (DMC@PBE0(ccECP/cc-pVQZ)) energies are reported for the subset of the lowest conformers of 10,793 molecules with up to 4 heavy atoms. This dataset has been systematically generated by considering all combinatorially possible stoichiometries, and graphs (according to Lewis rules as implemented in the `SURGE` package), along with all stable conformers identified by GFN2-xTB. Apart from graphs, geometries, rotational constants, and vibrational normal modes, VQM24 includes internal, atomization, electron-electron repulsion, exchange correlation, dispersion, vibrational frequency, Gibbs free, enthalpy, ZPV, molecular orbital energies; as well as entropy, and heat capacities. Electronic properties include multipole moments (dipole, quadrupole, octupole, hexadecapole), electrostatic potentials at nuclei (alchemical potential), Mulliken charges, and molecular wavefunctions. VQM24 represents a highly accurate and unbiased dataset of molecules, ideal for testing and training transferable, scalable, and generative ML models of real quantum systems.

arXiv:2405.05961v2 [physics.chem-ph] 13 May 2024

Argonne NATIONAL LABORATORY

Summary

- DMC is a very accurate method able to reproduce accurately properties of large class of solids and molecules
- Can be used as benchmark method in absence of experimental results or if golden standard method unreliable, unavailable or too expensive.
- Systematically improvable through multideterminant the trial wavefunction
- Optimized for GPUs; High cost in core hours but embarrassingly parallel (ideal for HPC and exascale).
- Black box for ML training sets (in Chemistry)

FUNDING ACKNOWLEDGEMENTS

Supported by the U.S. Department of Energy, Office of Science, Basic Energy Sciences, Materials Sciences and Engineering Division, as part of the Computational Materials Sciences Program.

CPSFM | Center for Predictive Simulation of Functional Materials

QMCPACK A framework for predictive and systematically improvable quantum-mechanics based simulations of materials

ECP EXASCALE COMPUTING PROJECT

An award of computer time was provided by the Innovative and Novel Computational Impact on Theory and Experiment (INCITE) program at the Argonne Leadership Computing Facility, which is a DOE Office of Science User Facility supported under Contract DE-AC02-06CH11357 and the Oak Ridge Leadership Computing Facility, which is a DOE Office of Science User Facility supported under Contract DE-AC05-00OR22725.

U.S. DEPARTMENT OF ENERGY INCITE LEADERSHIP COMPUTING

OAK RIDGE National Laboratory

Lawrence Livermore National Laboratory

Sandia National Laboratories

Argonne NATIONAL LABORATORY

NC STATE UNIVERSITY

BROWN

QUESTIONS