DMRG Quantum Chemistry Dataset for Multi-Reference Machine Learning

Stefan Gugler[1,2] **Nina Glaser**[3,4]
[1]Technische Universität Berlin [2]BIFOLD [3]Niels Bohr Institute [4]University of Copenhagen
stefan.gugler@tu-berlin.de nina.glaser@nbi.ku.dk

Motivation and AI task

As the global population rises, food demand grows in parallel, driving an increased reliance on nitrogen-based fertilizers to sustain agricultural productivity. This necessitates large-scale ammonia production through the Haber–Bosch process, which consumes roughly 2% of the world's energy and contributes about 3% of global CO_2 emissions [Ishaq and Crawford, 2024]. Thus, novel catalysts that operate under milder conditions with higher efficiency are highly sought-after. Machine learning (ML) offers a promising approach to accelerate the discovery of efficient processes for greener chemical reactions and a carbon-neutral future. However, much of catalytic chemistry lies in the multireference (MR) regime, where near-degenerate orbitals and competing spin states challenge single-reference (SR) methods. **Despite this, existing ML datasets are skewed toward closed-shell, near-equilibrium organic molecules, and thus fail to capture the MR challenges central to many catalytic reactions** [Cramer and Truhlar, 2009].

While *exact* MR calculations such as full configuration interaction scale combinatorially and are intractable for realistic catalysts, the density matrix renormalization group (DMRG) algorithm provides a scalable yet highly accurate alternative for treating strongly correlated electronic structures. DMRG leverages a variational treatment of static correlation through a systematic approximation of the full MR wavefunction as a matrix product state (MPS) [White, 1992, Schollwöck, 2011]. The DMRG accuracy is governed by the bond dimension m, ensuring the energy monotonically converges to the exact value. Modern implementations exploit spin and particle-number symmetries to further reduce the computational cost, and by combining DMRG with state-averaged self-consistent field procedures (DMRG-SCF) a balanced description can be obtained also for nearly-degenerate states.

To enable ML-based catalyst design, we propose a 50k-point dataset of DMRG calculations targeting structure–property prediction for molecular systems with pronounced MR character. The primary goal is to learn ground-state energies and low-lying spin-state gaps based on structure, charge, and multiplicity. Secondary targets are correlation-sensitive observables (spin densities, natural-orbital occupations) and entanglement metrics that can guide active-space selection. The dataset is designed to train surrogates that remain predictive when bonds stretch, spin states cross, or multiple electronic configurations compete. This is precisely the regime where current ML models fail, yet it's most critical for advancing catalyst discovery.

Dataset design and DMRG setup

While performing fully converged DMRG calculations for catalytic systems like metalloenzymatic clusters (e.g. the FeMo-cofactor) is infeasible at the 50k-point scale today, the essential MR behavior can be captured already in smaller systems. We will therefore assemble a diverse set of small and intermediate-size molecules that exhibit strong correlation: transition-metal diatomics and triatomics (e.g., FeO, NiO, MnO_2), minimal metal–ligand fragments, organic radicals and diradicals, open-shell ions, and bond-stretch scans. For each species, we will sample spin multiplicities and multiple spatial symmetries. To complete our set, as well as for the validation of future ML models, we will select

Accepted at the AI for Science workshop (NeurIPS 2025)

representative transition metal complexes from the MOR41 dataset for relevant transition metal reactions [Dohm et al., 2018] and its extension to barrier heights in the MOBH35 benchmark database [Iron and Janes, 2019]. This strategy ensures that our dataset covers the MR phenomena that defeat SR models while keeping active spaces small enough for accurate yet affordable DMRG calculations.

The dataset will be generated using a fully automated pipeline: First, geometries are optimized at a hybrid-functional level such as PBE0. Next, active spaces are selected automatically to include all near-degenerate valence shells (metal d and key ligand π/σ) using occupation- and entropy-based heuristics. We then use quantum-information diagnostics from DMRG to guide both active-space choice and the linear basis set ordering on the MPS lattice in the spirit of the `autoCAS` algorithm Stein and Reiher [2019]. By employing a Fiedler-vector ordering based on the mutual-information graph, long-range entanglement is minimized and the bond dimension m required for convergence will be reduced [Legeza and Sólyom, 2003]. For production calculations, state-averaged DMRG-SCF in triple-ζ bases will be performed with SU2 spin adaptation and Abelian point-group symmetries to obtain block-sparse tensors to further reduce the computational cost. To guarantee the accuracy of our single point calculations, we employ the two-site DMRG algorithm with adaptive bond-dimension schedules and explicit monitoring of discarded weights, and enforce tight energy convergence thresholds on the order of 1 mHa.

In addition to ground- and excited-state energies, each calculation will provide 1- and 2-particle reduced density matrices, spin densities, natural-orbital occupations, and entanglement diagnostics. These richer outputs expand the dataset beyond energy labels, enabling ML approaches that learn from electronic structure in real space. Such data can be directly applied to surrogate modeling, active-space selection, or even inverse density functional parametrization [Kanungo et al., 2019], thereby supporting a wide range of workflows in ML-driven catalyst design.

Feasibility, cost, and impact

The data generation workflow will be implemented in Python, relying only on open source packages. For integral generation and mean-field references, we will use pySCF [Sun et al., 2018], and the DMRG-SCF calculations will be performed with block2 [Zhai et al., 2023], a highly optimized DMRG engine. It allows for symmetry-adaptation of the MPS, two-site DMRG sweeps, Fiedler ordering, state-average DMRG-SCF procedures via a readily available pySCF interface, convergence monitoring diagnostics, as well as access to 1- and 2-particle reduced electron densities and related quantities. Importantly, block2 also enables multi-level parallelism over sites, symmetry sectors, and matrix multiplication operations, allowing for efficient single-point calculations [Zhai and Chan, 2021].

Indicative timings range from one CPU-hour for small systems to several CPU-days for transition-metal fragments (see Appendix A). Aggregated across the proposed molecular mix, the 50k-point dataset will require approximately one million CPU-hours. Using a standard cloud TCO model, this corresponds to a computational cost of around $20k. The full dataset, including density matrices, is expected to require single-digit terabytes of storage.

The proposed data set provides what the field lacks: standardized, high-fidelity results in the omnipresent MR regime. Fully exploiting the information contained in DMRG results, our data set contributes more than just highly accurate correlated energies. The availability of reduced density matrices, orbital entropies, and mutual information enables training objectives that directly penalize electronic delocalization errors and reward correct entanglement structure signals absent from common density functional theory based datasets. By training models on DMRG references, we expect improved robustness at dissociation limits, spin crossings, and symmetry breaking. In downstream science, models trained on this corpus will accelerate screening of metal–ligand motifs and open-shell intermediates, and will provide correlation-aware embeddings for hybrid QM/ML treatments of larger active sites. While exhaustive DMRG over full metalloclusters remains a future goal, the proposed dataset captures the physics that matters for catalysis and provides a realistic foundation on which AI models can learn to capture effects of strongly correlated electrons in novel catalytic candidates.

References

Haris Ishaq and Curran Crawford. Review of ammonia production and utilization: Enabling clean energy transition and net-zero climate targets. *Energy Conversion and Management*, 300:117869, 2024. ISSN 0196-8904. doi: https://doi.org/10.1016/j.enconman.2023.117869.

Christopher J. Cramer and Donald G. Truhlar. Density functional theory for transition metals and transition metal chemistry. *Physical Chemistry Chemical Physics*, 11(46):10757–10816, 2009. doi: 10.1039/B907148B.

Steven R. White. Density matrix formulation for quantum renormalization groups. *Physical Review Letters*, 69(19):2863–2866, 1992. doi: 10.1103/PhysRevLett.69.2863.

Ulrich Schollwöck. The density-matrix renormalization group in the age of matrix product states. *Annals of Physics*, 326(1):96–192, 2011. doi: 10.1016/j.aop.2010.09.012.

Sebastian Dohm, Andreas Hansen, Marc Steinmetz, Stefan Grimme, and Marek P. Checinski. Comprehensive thermochemical benchmark set of realistic closed-shell metal organic reactions. *Journal of Chemical Theory and Computation*, 14(5):2596–2608, 2018. doi: 10.1021/acs.jctc.7b01183.

Mark A. Iron and Trevor Janes. Evaluating transition metal barrier heights with the latest density functional theory exchange–correlation functionals: The mobh35 benchmark database. *The Journal of Physical Chemistry A*, 123(17):3761–3781, 2019. doi: 10.1021/acs.jpca.9b01546.

Christopher J. Stein and Markus Reiher. <scp>autocas</scp>: A program for fully automated multiconfigurational calculations. *Journal of Computational Chemistry*, 40(25):2216–2226, 2019. doi: 10.1002/jcc.25869.

Örs Legeza and Jenö Sólyom. Optimizing the density-matrix renormalization group method using quantum information entropy. *Physical Review B*, 68:195116, 2003. doi: 10.1103/PhysRevB.68.195116.

Bikash Kanungo, Paul Zimmerman, and Vikram Gavini. Exact exchange-correlation potentials from ground-state electron densities. *Nature Communications*, 10, 10 2019. doi: 10.1038/s41467-019-12467-0.

Qiming Sun, Timothy C. Berkelbach, Nick S. Blunt, George H. Booth, Sheng Guo, Zhendong Li, Junzi Liu, James McClain, Elvira R. Sayfutyarova, Sandeep Sharma, Sebastian Wouters, and Garnet Kin-Lic Chan. PySCF: the python-based simulations of chemistry framework. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 8(1):e1340, 2018. doi: 10.1002/wcms.1340.

Huanchen Zhai, Henrik R. Larsson, Seunghoon Lee, Zhi-Hao Cui, Tianyu Zhu, Chong Sun, Linqing Peng, Ruojing Peng, Ke Liao, Johannes Tölle, Junjie Yang, Shuoxue Li, and Garnet Kin-Lic Chan. Block2: A comprehensive open source framework to develop and apply state-of-the-art dmrg algorithms in electronic structure and beyond. *The Journal of Chemical Physics*, 159(23):234801, 2023. doi: 10.1063/5.0180424.

Huanchen Zhai and Garnet Kin-Lic Chan. Low communication high performance ab initio density matrix renormalization group algorithms. *The Journal of Chemical Physics*, 154(22):224116, 2021. doi: 10.1063/5.0050902.

A Computational cost details

Indicative timings for representative molecules demonstrate the project's feasibility, with costs ranging from one CPU-hour for small systems to several days for transition-metal fragments (Table 1). Aggregated across the proposed molecular mix, the 50k-point dataset requires approximately one million CPU-hours. Based on a standard Total Cost of Ownership (TCO) model at $0.02 per CPU-hour, this corresponds to a computational cost of around $20k. The full dataset, including reduced density matrices and diagnostics, is expected to require single-digit terabytes of storage. This scalable approach is tractable for a high-throughput dataset, unlike calculations on large clusters like the FeMo-cofactor, which are computationally prohibitive.

Table 1: Indicative DMRG-SCF timings for representative systems. Timings are per single-point calculation on multi-core hardware and are used to estimate the overall project cost.

System	**Active Space**	**Time per point (CPU-h)**	**50k-point Corpus**
N_2 (dissociation)	CAS(10e, 8o)	1	Feasible
O_2 (open-shell)	CAS(12e, 10o)	3	Feasible
FeO (diatomic)	CAS(16e, 12o)	24	Feasible
Cu_2O_2 (minimal core)	CAS(16e, 14o)	96	Feasible
FeMo-cofactor	CAS(54e, 54o)	>1000	Infeasible