

Boosting applications in nuclear physics via recent advances in tensor network state methods

Synergies among physics, chemistry, math and computer science

Örs Legeza

Strongly Correlated Systems “Lendület” Research Group
Wigner Research Centre for Physics, Budapest, Hungary

Institute for Advanced Study, Technical University of Munich, Germany

Parmenides Foundation, Pöcking, Germany

Workshop on Tensor Networks and (Quantum) Machine Learning
for High-Energy Physics

CERN, 11.04.2024

in collaboration with

- ▶ Andor Menczer, Gergely Barcza, Szilárd Szalay, Imre Hagymási, Adam Ganyecz, Mihály Máté, Andras Olasz, Tamás Mosoni
- ▶ Florian Gebhard, Reinhard M. Noack, Georg Ehlers
- ▶ Frank Verstraete, Klaas Gunst, S. Wooters, D. van Neck
- ▶ Jens Eisert, Christian Krumnow, Edoardo Fertitta, Beate Paulus
- ▶ Reinhold Schneider, Max Pfeffer
- ▶ Libor Veis, Jiri Pittner, Jiri Brabec, Andrej Antalik, Jan Brandejs
- ▶ Frank Neese, Ali Alavi, Peter Saalfrank
- ▶ Gero Friesecke, Mi-Song Dupuy, Benedikt Graszwald
- ▶ Alexander Tichai, Achim Schwenk, Takai Miyagi
- ▶ Simen Kvaal, Fabian Faulstich, Andre Laestadius
- ▶ Moca Pascu, Miklós Werner, Kornel Kapas, Gergely Zaránd
- ▶ Uli Schollwöck, Martin Grundner, Sam Mardazad, Christian Schilling
- ▶ Karol Kowalski, Sotiris Xantheas, ... (incomplete list)

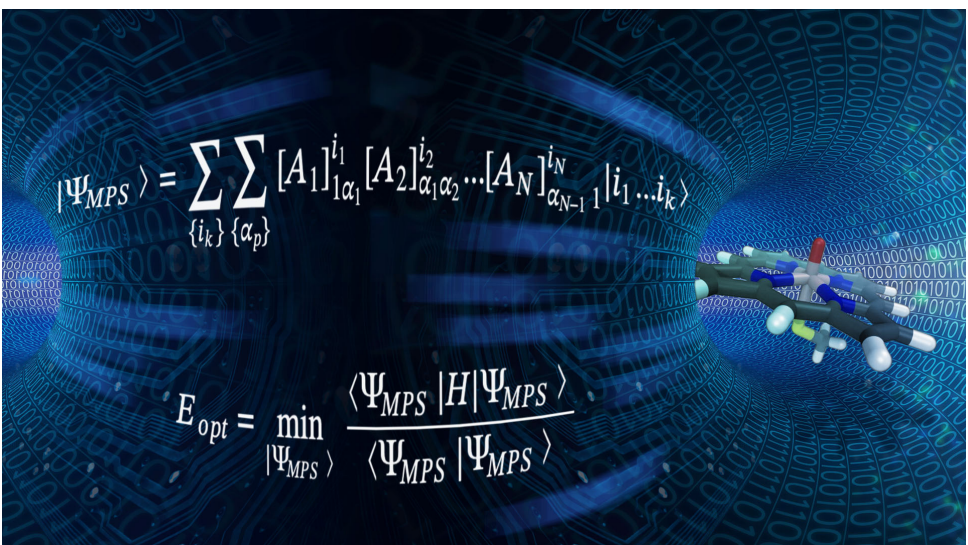
in collaboration with

- ▶ more than 30 research groups worldwide from condensed matter physics, quantum chemistry, nuclear physics, quantum information theory, applied mathematics and computer science
- ▶ High-Performance Computing Center Stuttgart, Germany
- ▶ Pacific Northwest National Laboratory (PNNL), USA
- ▶ National Energy Research Scientific Computing Center (NERSC), USA

Our computer program package is used by more than 30 research groups worldwide for more than two decades.

Recently there is also an interest by industrial partners.

- ▶ NVIDIA, USA
- ▶ AMD, USA
- ▶ SandboxAQ, USA (Google startup)
- ▶ Riverlane LTD, UK
- ▶ Furukawa Electric Institute of Technology, Japan
- ▶ Dynaflex LTD, Hungary

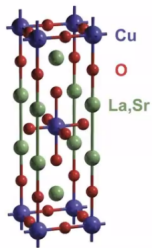

$$|\Psi_{MPS}\rangle = \sum_{\{i_k\}} \sum_{\{\alpha_p\}} [A_1]_{1\alpha_1}^{i_1} [A_2]_{\alpha_1\alpha_2}^{i_2} \dots [A_N]_{\alpha_{N-1}1}^{i_N} |i_1 \dots i_N\rangle$$

$$E_{opt} = \min_{|\Psi_{MPS}\rangle} \frac{\langle \Psi_{MPS} | H | \Psi_{MPS} \rangle}{\langle \Psi_{MPS} | \Psi_{MPS} \rangle}$$

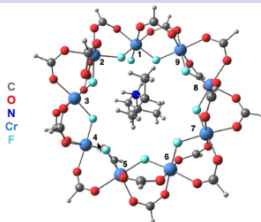
<https://www.pnnl.gov/news-media/collaboration-speeds-complex-chemical-modeling>

<https://www.newswise.com/articles/collaboration-speeds-complex-chemical-modeling>

Strong correlations between electrons used by nature and in new technologies

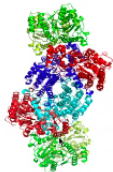
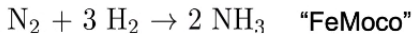


High T_c superconductors

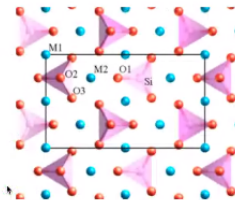


Lee, Small & Head-Gordon, *JCP*, 2018, 149, 244121

Single molecular magnets (SMM)



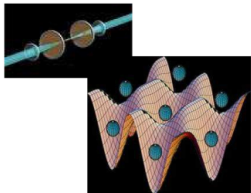
Nitrogen fixation



Battery technology

Experimental realizations: optical lattices

Numerical simulations: model systems



Atoms (represented as blue spheres) pictured in a 2D-optical lattice potential

Potential depth of the optical lattice can be tuned.

Periodicity of the optical lattice can be tuned.

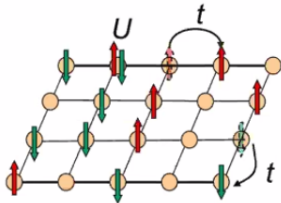
Hubbard model: lattice model of interacting electron system

$$H = t \sum_{\langle i,j \rangle, \sigma} c_{i,\sigma}^\dagger c_{j,\sigma} + \frac{U}{2} \sum_{\sigma \neq \sigma'} \sum_i n_{i,\sigma} n_{i,\sigma'}$$

t hopping amplitude

U on-site Coulomb interaction

$\sigma \in \uparrow, \downarrow$ spin index



Classical or quantum computers?

TNS/DMRG provide state-of-the-art results in many fields

$$\mathcal{H} = \sum_{ij\alpha\beta} T_{ij}^{\alpha\beta} c_{i\alpha}^\dagger c_{j\beta} + \frac{1}{2} \sum_{ijkl\alpha\beta\gamma\delta} V_{ijkl}^{\alpha\beta\gamma\delta} c_{i\alpha}^\dagger c_{j\beta}^\dagger c_{k\gamma} c_{l\delta} + \dots,$$

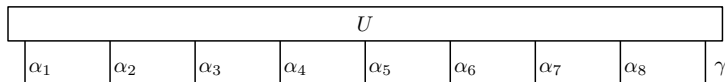
- ▶ T_{ij} kinetic and on-mode terms, V_{ijkl} two-particle scatterings
 - ▶ We consider usually lattice models in real space (DMRG)
 - ▶ In quantum chemistry modes are electron orbitals (QC-DMRG)
 - ▶ In UHF QC spin-dependent interactions (UHF-QCDMRG)
 - ▶ In relativistic quantum chemistry modes are spinors (4c-DMRG)
 - ▶ In nuclear problems modes are proton/neutron orbitals (JDMRG)
 - ▶ In k-space modes are momentum eigenstates (k-DMRG)
 - ▶ For particles in confined potential modes \rightarrow Hermite polynomials
 - ▶ **Major aim: to obtain the desired eigenstates of \mathcal{H} .**
- Symmetries: Abelian and non-Abelian quantum numbers, double groups, complex integrals, quaternion sym. etc
 - # of block states: 1 000 – 60 000. Size of Hilbert space up to 10^8 .
 - In ab initio DMRG the CAS size is: 100 electrons on 100 orbitals.
 - 1-BRDM and 2-BRDM, finite temperature, dynamics
 - Massively parallel implementations CPU/GPU \rightarrow exascale on HPC

Tensor product approximation

State vector of a quantum system in the discrete tensor product spaces

$$|\Psi_\gamma\rangle = \sum_{\alpha_1=1}^{q_1} \dots \sum_{\alpha_d=1}^{q_d} U(\alpha_1, \dots, \alpha_d, \gamma) |\alpha_1\rangle \otimes \dots \otimes |\alpha_d\rangle \in \bigotimes_{i=1}^d \Lambda_i := \bigotimes_{i=1}^d \mathbf{C}^{q_i},$$

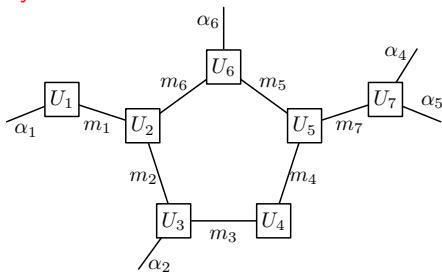
where $\text{span}\{|\alpha_i\rangle : \alpha_i = 1, \dots, q_i\} = \Lambda_i = \mathbf{C}^{q_i}$ and $\gamma = 1, \dots, m$.



$\dim \mathcal{H}_d = \mathcal{O}(q^d)$ Curse of dimensionality!

We seek to reduce computational costs by parametrizing the tensors in some data-sparse representation.

A general tensor network representation of a tensor of order 5.



Matrix product state (MPS) representation / DMRG / TT

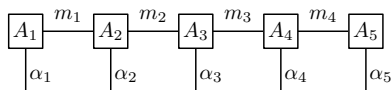
Affleck, Kennedy, Lieb Tagasaki (87); Fannes, Nachtergale, Werner (91), White (92),

Römmer & Ostlund (94), Vidal (03), Verstraete (04), Oseledets & Tyrtshnikov, (09)

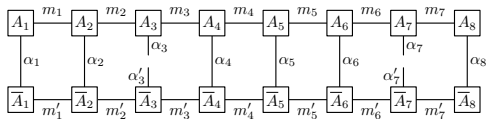
The tensor U is given elementwise as

$$U(\alpha_1, \dots, \alpha_d) = \sum_{m_1=1}^{r_1} \dots \sum_{m_{d-1}=1}^{r_{d-1}} A_1(\alpha_1, m_1) A_2(m_1, \alpha_2, m_2) \dots A_d(m_{d-1}, \alpha_d).$$

We get d component tensors of order 2 or 3. **Scaling:** m^3 .



Calculation of ρ_{ij} corresponds to the contraction of the network except at modes i and j .



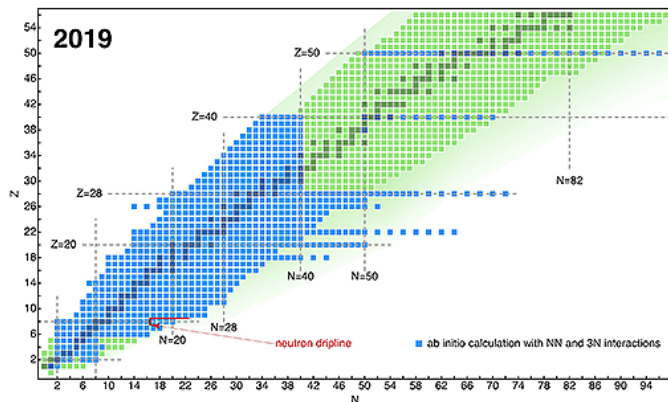
von Neumann quantum information entropy, $s = - \sum_{\alpha} \lambda_{\alpha}^2 \ln \lambda_{\alpha}^2$.

Mutual information, $I = s_i + s_j - s_{ij}$.

Ö.L & Sólyom, (03), Rissler, Noack, White (06)

Nuclear physics: modes are proton/neutron orbitals (JDMRG)

Dukelsky, Papenbrock, Pittel (2003), Ö.L., Veis, Dukelsky, Poves (2015)



$$H = \sum_{\alpha} \varepsilon_{\alpha} c_{\alpha}^{\dagger} c_{\alpha} - \frac{1}{2} \sum_{\alpha\beta\gamma\delta} V_{\alpha\beta\gamma\delta} c_{\alpha}^{\dagger} c_{\beta}^{\dagger} c_{\delta} c_{\gamma},$$

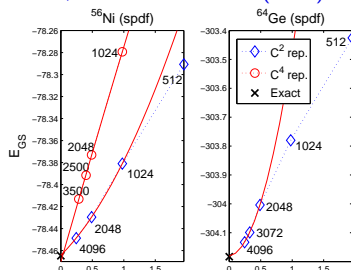
- ▶ where c_{α}^{\dagger} and c_{α} creates and annihilates a particle with quantum numbers $\alpha = (n, l, j, m, \tau_z)$. $j \geq 1/2$, Isospin,
- ▶ no-core shell models
- ▶ effective Hamiltonian including parts of 3-body interactions

Nuclear shell DMRG: ^{64}Ge pf+g9/2, $\alpha = (n, l, j, m, \tau_z)$.

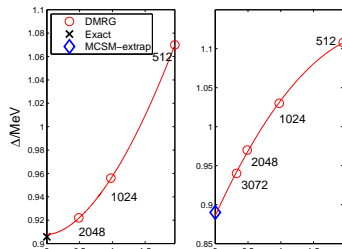
Ö.L, L. Veis, J. Dukelsky, A. Poves (2015)

Particle-hole (phDMRG) J. Dukelsky, S. Pittel, S. Dimitrova, M. Stoitsov (2002)

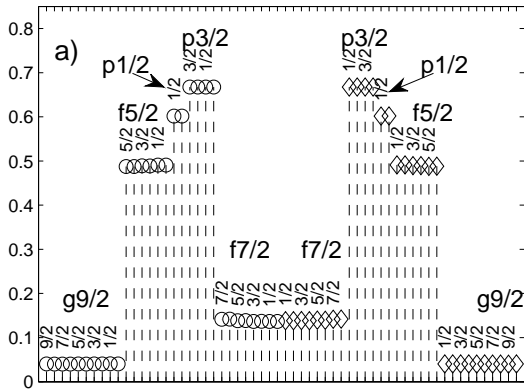
j-coupling scheme (JDMRG) T. Papenbrock, D.J. Dean (2005), B. Thakur, S. Pittel, and N. Sandulescu (2008)



The DMRG ground state energy for ^{56}Ni and ^{64}Ge in the pf-shell and ^{64}Ge are shown as a function of $1/M$. Diamonds (DMRG), Crosses (Exact Diagonalization), Circle (MCSM).



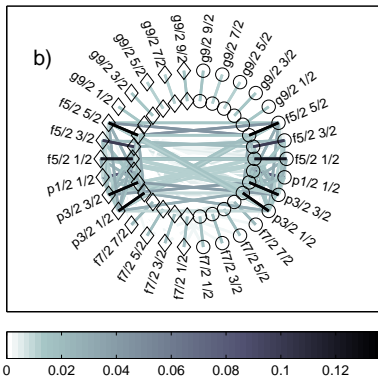
The energy gap between the 0 and 2^+ states for ^{64}Ge with pf and pf+g9/2 valence space as a function of $1/M$.



Single-site entropy obtained with DMRG for ^{64}Ge in the $pf+g9/2$ -shell.

Circles and diamonds label proton and neutron orbitals respectively.

One-mode entanglement ($\max(s_i) = \ln 2 = 0.69$)
 Multireference problem (strongly correlated)



The mutual information matrix elements obtained with DMRG for ^{64}Ge in the $pf+g9/2$ -shell. on a ladder topology with time-reversed pairs in the rungs.

Circles and diamonds label proton and neutron orbitals respectively, and sites are denoted by l, j, m with $+m$ outside the ladder and $-m$ inside.

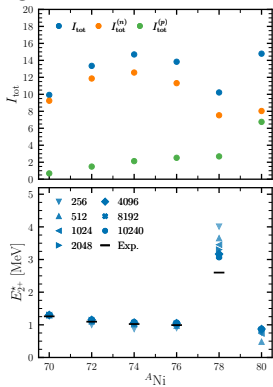
The mutual information is approximately equal for the $p1/2$, $p3/2$ and $f5/2$ orbitals, and independent of their j_z projections, due to strong $T=1$ proton-proton and neutron-neutron pairing coherence.

Proton-neutron $T=1$ pairing correlations between time reversed and charge conjugated states for the $p1/2, p3/2$ and $f5/2$ orbitals.

Significant correlation between proton-neutron maximally aligned states for the $p3/2$ and $f5/2$ orbitals, which could be related to $J = 2j_z$ pairing and/or quadrupole-quadrupole in the $T=0$ channel.

In-medium similarity renormalization group & DMRG

- Starting with two-nucleon (NN) and three-nucleon (3N) interactions, the VS-IMSRG generates a valence-space-decoupled Hamiltonian restricted to an active space of limited size.
- Many-body operators of higher particle rank are truncated at the normal-ordered two-body level, defining the IMSRG(2) truncation.
- Neutron-rich nickel isotopes attract significant experimental attention, e.g., with the recent discovery of the **doubly magic** nature of ^{78}Ni



- Neutron, proton, and total entropies (top) and 2^+ excitation energies (bottom) along even-mass nickel isotopes.

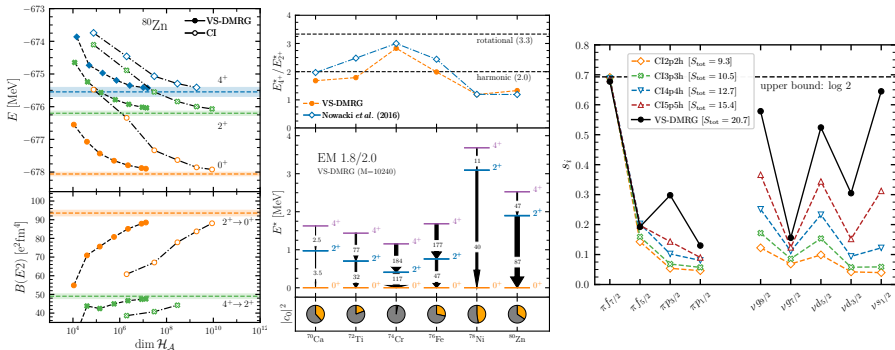
- Entropies are calculated at bond dimension $M = 10240$ whereas for the excitation energies the bond dimension was varied between $M = 256 - 10240$.

Tichai, Knecht, Kruppa, Ö.L., Moca, Schwenk, Werner, Zarand (2022)

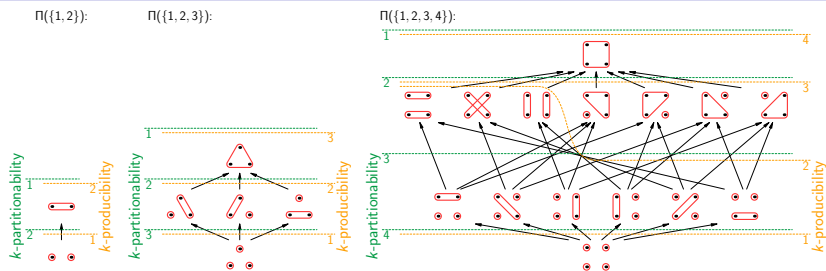
- Current work: deformation, clustering

Low-lying spectroscopy of $N = 50$ isotones

- Nowacki et al observed a rapid transition from single-particle-like excitations in ^{78}Ni to collective rotational excitations in ^{74}Cr .
- Rotational structures can be approximately extracted by comparing the level spacings of the low-lying spectrum to that of a rigid rotor, $E^*(J) \sim J(J+1)$. For a perfect rotor $E_{4+}^*/E_{2+}^* = 3.33$
- The calculated B(E2) values show a maximum for ^{74}Cr for the $2^+ \rightarrow 0^+$ transition that is characteristic for collective rotational excitations and a signature of nuclear deformation.



Multiorbital correlations Sz. Szalay (2015)



- ▶ partitions of the system:
 $\xi = \{X_1, X_2, \dots, X_{|\xi|}\} \equiv X_1 | X_2 | \dots | X_{|\xi|} \in \Pi(L)$
- ▶ refinement: $v \preceq \xi$ def.: $\forall Y \in v, \exists X \in \xi : Y \subseteq X$
- ▶ ξ -**correlation** (ξ -mutual information):

$$C_\xi(\rho) = \min_{\sigma \in \mathcal{D}_{\xi\text{-uncorr}}} D(\rho || \sigma) = \sum_{X \in \xi} S(\rho_X) - S(\rho)$$

- ▶ multipartite monotonicity: $v \preceq \xi \Leftrightarrow C_v \geq C_\xi$

k -partitionability-correlation and k' -producibility correlation:

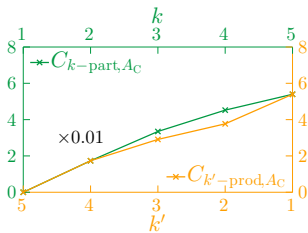
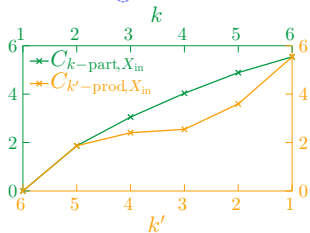
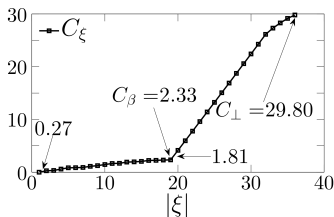
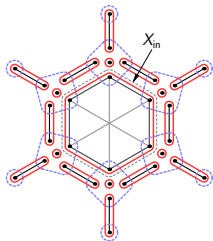
$$C_{k\text{-part}}(\rho_L) = C_{\mu_k}(\rho_L) = \min_{|\mu| > k} C_\mu(\rho_L), \quad C_{k'\text{-prod}}(\rho_L) = C_{\nu_{k'}}(\rho_L) = \min_{\forall N \in \nu: |N| < k'} C_\nu(\rho_L)$$

Example (aromatic system): C_6H_6 (benzene)

Szalay, Barcza, Szilvási, Veis, Ö.L (2017)

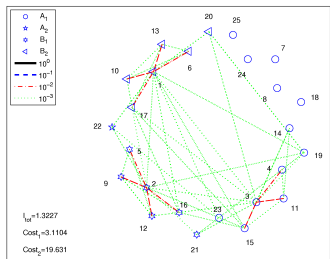
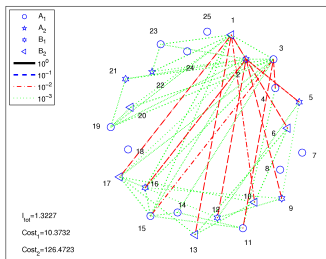
“atoms”: $\alpha = A_1|A_2|\dots|A_{|\alpha|}$, “bonds”: $\beta = B_1|B_2|\dots|B_{|\beta|}$

$$\sum_{A \in \alpha} C_{\perp, A}(\varrho_A) + C_{\alpha}(\varrho_M) = \sum_{B \in \beta} C_{\perp, B}(\varrho_B) + C_{\beta}(\varrho_M) = C_{\perp}(\varrho_M)$$



Tensor topology optimization: $\sum_{ij} l_{ij} \times d_{ij}^\eta$ (Ex. LiF 6/25)

Energetical ordering (MPS) $d_{ij} = |i - j|$ Entanglement localization (MPS)



- ▶ Reordering orbitals by minimizing the entanglement distance:

$$\hat{I}_{\text{dist}} = \sum_{i,j} l_{i,j} \times |i - j|^\eta,$$

- ▶ Apply spectral graph theory: Fiedler vector $x = (x_1, \dots, x_N)$ is the solution that minimizes $F(x) = x^\dagger Lx = \sum_{ij} l_{i,j} (x_i - x_j)^2$, with

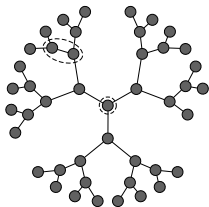
$$\sum_i x_i = 0 \text{ and } \sum_i x_i^2 = 1, \text{ and the graph Laplacian is } L = D - I \text{ with } D_{i,i} = \sum_j l_{i,j}.$$

The second eigenvector of the Laplacian is the Fiedler vector.

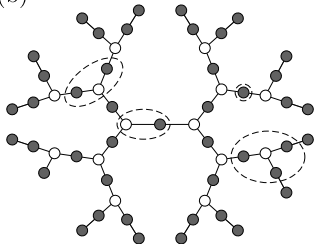
ÖL, Sólyom (2003), Barcza, ÖL, Marti, Reiher (2011)

T3NS a new tensor format Gunst, Verstraete, Wooters, Ö.L., van Neck (2018)

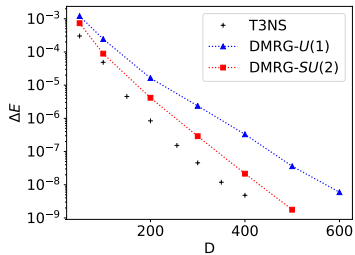
(a)



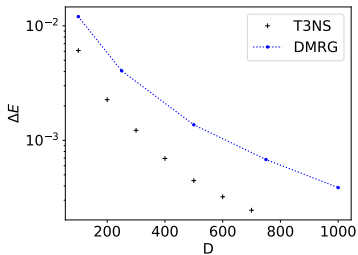
(b)



LiF



N_2



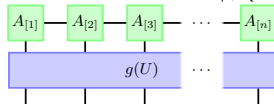
Redefinition of the fermionic modes by a linear transformation

Krumnow, Veis, Ö.L., Eisert, 2014-2016

- Linear transformations of a set of fermionic annihilation operators $\{c_i\}$ to a new set $\{d_j\}$ satisfying the canonical anti-commutation relations:

$$c_i = \sum_{j=1}^{Np} U_{i,j} d_j, \quad p \text{ denotes the number of different fermion species}$$

- Under this change of basis a state vector $|\psi(U)\rangle = G(U)|\psi(\mathbb{1})\rangle$



- Denoting the Hamiltonian written in terms of the transformed modes by $H(U) = G(U)^\dagger H G(U)$, we are interested in the solutions of

$$(U_{\text{opt}}, |\psi_{\text{opt}}\rangle) = \underset{|\psi\rangle \in \mathcal{M}_{D_{\text{max}}}}{\text{argmin}}_{U \in U(Np)}, \langle \psi | H(U) | \psi \rangle.$$

- The global basis change is composed of local unitaries solutions of

$$U_{\text{opt}}^{\text{loc}} = \underset{U \in V}{\text{argmin}} f_j(|\psi(\mathbb{1}_j \oplus U \oplus \mathbb{1}_{N-j-2})\rangle),$$

cost function $f_j^{(1)}(|\psi\rangle) = \|\Sigma_\psi^j\|_1$ where Σ_ψ^j denotes the Schmidt spectrum of $|\psi\rangle$ for a bipartiting cut between sites j and $j+1$.

Global fermionic mode optimization via swap gates

Friesecke, Werner, Kapas, Menczer, ÖL(2024)

- Finding an optimal representation of a quantum many body wave function, i.e., a parametrization with the minimum number of parameters for a given error margin is a task of utmost importance in modern quantum physics and chemistry

$$\Psi = \sum_{\mu_1, \dots, \mu_N=0}^1 C(\mu_1, \dots, \mu_N) \Phi_{\mu_1, \dots, \mu_N}$$

- Computational complexity \sim block entropy area (BEA)

$$B\alpha(C) = \sum_{\ell=1}^{N-1} S_\alpha(\rho_{1,2,\dots,\ell})$$

$$\rho_{1,2,\dots,\ell}(\mu_1, \dots, \mu_\ell; \mu'_1, \dots, \mu'_\ell) = \sum_{\mu_{\ell+1}, \dots, \mu_N} C(\mu_1, \dots, \mu_N) C^*(\mu'_1, \dots, \mu'_\ell, \mu_{\ell+1}, \dots, \mu_N)$$

$S_\alpha(\rho) = \frac{1}{1-\alpha} \ln(\text{Tr } \rho^\alpha)$ is the Rényi entropy for some $0 < \alpha < 1$.

- The important feature needed is concavity, so that density operators are favoured whose eigenvalues are either very large or very small.

Under a single particle unitary mode transformation $U \in U(N)$

- New modes $\varphi'_i = \sum_j U_{ij}\varphi_j$, and $C' = G(U)^\dagger C$ where $G(U)$ is a unitary transformation on the space of many-body coefficient tensors.
- For time reversal symmetric case, C and the φ_i are real-valued and $U \in O(N)$ or, discarding an immaterial overall sign factor, $U \in SO(N)$.
- U can be parametrized as $U = e^A U_*$ with U_* an arbitrary fixed matrix in $SO(N)$ and A real and skew-symmetric, the parametrization being unique for U close to U_* .
- Thus stationarity of a scalar function f on $SO(N)$ at U_* is equivalent to

$$0 = \left. \frac{d}{dt} \right|_{t=0} f(e^{tA} U_*) = \text{Tr} \frac{\partial f}{\partial U}(U_*) U_*^T A^T, \quad \forall A^T = -A$$

that is to say $\frac{\partial f}{\partial U}(U_*) U_*^T$ symmetric.

- **Reduction to pairwise rotations:** To achieve stationarity minimize f over all pairwise rotations $U_{ij}(\theta)$ given by $e^{\theta E_{ij}}$, $E_{ij} = e_i e_j^T - e_j e_i^T$, where e_i is the unit vector of \mathbb{R}^N whose i -th component is 1 and whose other components are zero. This corresponds to the mode transformation $\varphi'_i = \cos \theta \varphi_i + \sin \theta \varphi_j$, $\varphi'_j = -\sin \theta \varphi_i + \cos \theta \varphi_j$ which leaves all other modes the same.

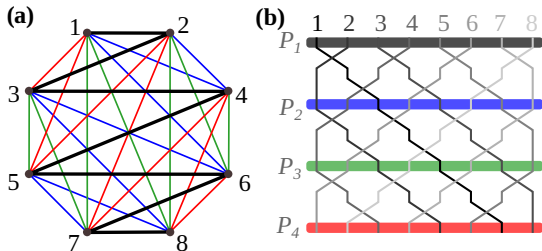
Reduction to permutations and nearest neighbor rotations

- Set of all pairwise rotations $U_{ij}(\theta)$ can be realized by $N/2$ global re-orderings of the orbitals and the $N-1$ nearest-neighbor rotations for each ordering.
- if $\tau_1, \dots, \tau_{N/2}$ are the specific permutations such that any pair of orbitals become nearest neighbours under one of these permutations (that is, for all $i < j$ there exist ν and ℓ such that $\{\tau_\nu(i), \tau_\nu(j)\} = \{\ell, \ell + 1\}$), then

$$U_{ij}(\theta) = \tau_\nu^{-1} U_{\ell, \ell+1}(\pm\theta) \tau_\nu$$

with '+' if $\tau_\nu(i) < \tau_\nu(j)$ and '-' otherwise.

Swap gates controlled permutations: The optimal set of permutations $\tau_1, \dots, \tau_{N/2}$ can be generated by Walecki's method (1882):



- The modes are placed in a zig-zag line to the vertices of a regular polygon with N vertices

Local mode optimization and block entropy area

Krumnow, Veis, ÖL, Eisert (2015-2016)

- Consecutive permutations are easily generated by two layers of nearest neighbor swap operations placed in a checkerboard pattern ,i.e., full forward mode optimization sweep with fixed, but alternating angles of $\pi/2$ and 0 and a backward sweep in a reversed order.
- To avoid truncation of the wavefunction, the bond dimension has to be increased by a factor of q .

Local mode optimization and block entropy area: $H(U) = G(U)^\dagger H G(U)$ is constructed iteratively from two-mode unitary operators by optimizing $\theta_{l,l+1}$ while sweeping through the network for $l = 1 \dots N - 1$

$$U_{\text{opt}}^{\text{loc}} = \operatorname{argmin}_{U \in V} f_l(|\psi(\mathbb{1}_l \oplus U \oplus \mathbb{1}_{N-l-2})\rangle),$$

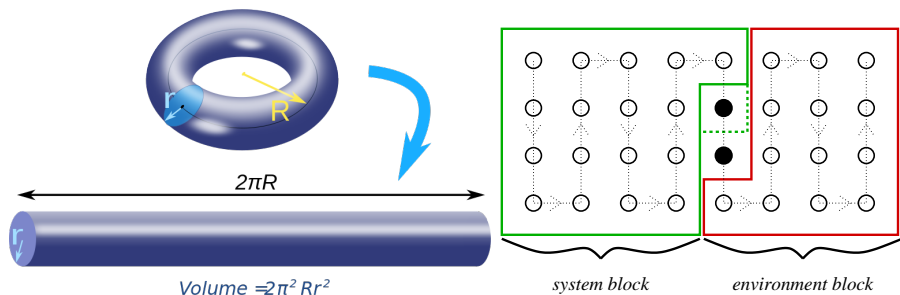
- At each micro-iteration step, f_l , i.e., the half-Rényi block entropy $S_{1/2}(\rho_{\{1,2,\dots,l\}})$ is minimized by a two-mode rotation (**disentangler**)

Locality of the BEA: Nearest neighbor rotations by $\theta_{l,l+1}$ change only the block entropy measured when the cut is at mode l , while all other block entropies remain invariant.

2d spinless fermions with PBCs Krumnow, Veis, Eisert, Ö.L (2019-2021)

$$H = \sum_{\langle i,j \rangle} c_i^\dagger c_j + \sum_{\langle i,j \rangle} V n_i n_j,$$

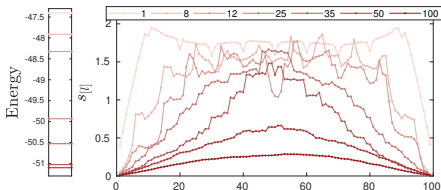
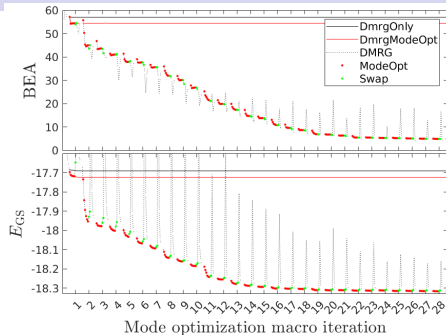
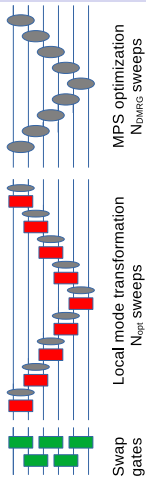
$$U = \begin{pmatrix} \exp(i\theta_1)\cos(\theta_2) & \exp(i\theta_1)\sin(\theta_2) \\ -\exp(-i\theta_1)\sin(\theta_2) & \exp(-i\theta_1)\cos(\theta_2) \end{pmatrix}$$



- ▶ Optimization on the mps manifold and on the Grassman manifold
- ▶ Hamiltonian becomes long ranged

$$H = \sum_{i,j=1}^n t_{i,j} c_i^\dagger c_j + \sum_{i,j,k,l=1}^n v_{i,j,k,l} c_i^\dagger c_j^\dagger c_k c_l,$$

2d-spinless Hubbard, $N = 6 \times 6$ and 10×10 with $D = 80$

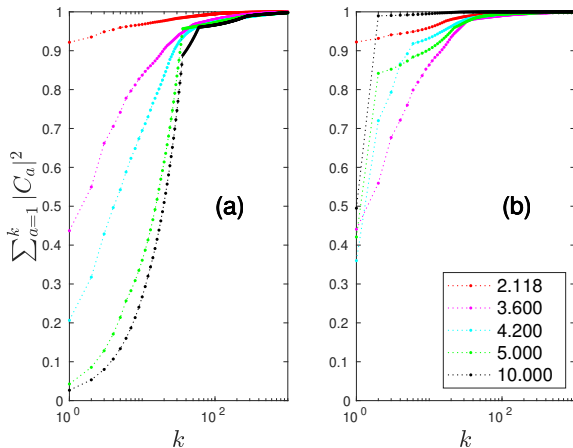


Consistency: Perturbation induced by the swap layers can be eliminated by subsequent application of several DMRG sweeps via nearest neighbor mode optimization, once the algorithm has found the stationary solution for both energy and BEA.

Compressing multireference character via fermionic mode optimization

Máté, Petrov, Szalay, ÖL (2022)

- Example: stretching nitrogen dimer in the full space CAS(6,14)
- single→multi reference problem



(a) Sum of the square of the absolute values of the 1000 largest CI coefficients for various bond lengths. (b) for the optimized MOs.

- Example: stretching nitrogen dimer in the full space CAS(14,28)

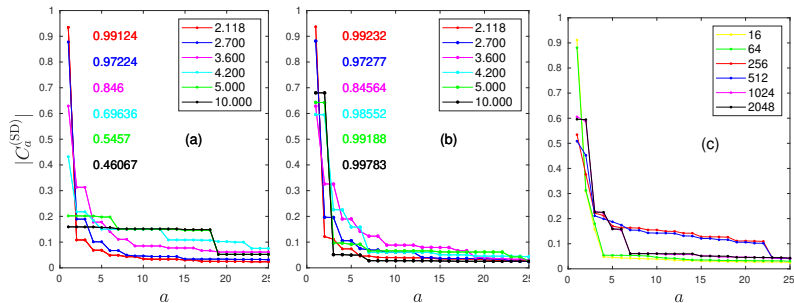
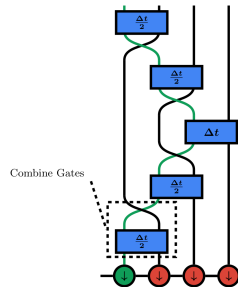
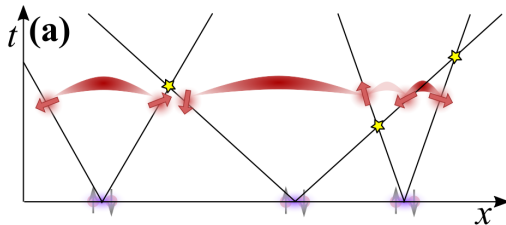


Figure: (a) Absolute values of the first 25 largest CI coefficients including single and double excitation levels for various bond lengths, extracted from the MPS wave function, obtained by the DMRG algorithm, with a bond dimension $D = 4096$. The inscribed numbers are the norm squares of the wave function component corresponding to single and double excitations for the various bond lengths.

(b) Similar to (a), but for the optimized MOs with $D_{\text{opt}} = 512$.

(c) Convergence of the absolute values of the first 25 largest CI coefficients including single and double excitation levels for $r = 4.200a_0$ for the optimized MOs, as a function of the bond dimension D .

Long time evolution Krumnow, Eisert, Ö.L. (2019)



- ▶ At time $t = 0$ we perturb the system.
- ▶ After the quench the quasiparticles collide with each other. , the scattering events are denoted by stars. World lines of the excited quasiparticle pairs. Entanglement structure of the gas is indicated by arched red stripes. At $t=0$ singlet pairs with zero total momentum are excited.
- ▶ There are different time-evolution methods for MPS which are currently in use to solve the time-dependent Schrödinger equation (TDSE).
- ▶ application of $\hat{U}(\delta_t) = e^{-i\delta_t \hat{H}}$, i.e. , $|\psi(t)\rangle \rightarrow |\psi(t + \delta_t)\rangle$

Coupled cluster method with single and double excitations tailored by matrix product state wave functions

L. Veis, A. Antalik, F. Neese, Ö.L., J. Pittner (2016)

- ▶ Formally single reference theory, Fermi vacuum is a single determinant
- ▶ **Split-amplitude ansatz**

$$\Psi_{\text{TCC}} = e^{\mathcal{T}} \Psi_{\text{ref}} = e^{\mathcal{T}^{\text{ext}} + \mathcal{T}^{\text{CAS}}} \Psi_{\text{ref}}$$

▶ \mathcal{T}^{CAS}

- ▶ amplitudes extracted from DMRG (CASCI) calculation
- ▶ frozen during CC calculation
- ▶ account for static correlation

▶ \mathcal{T}^{ext}

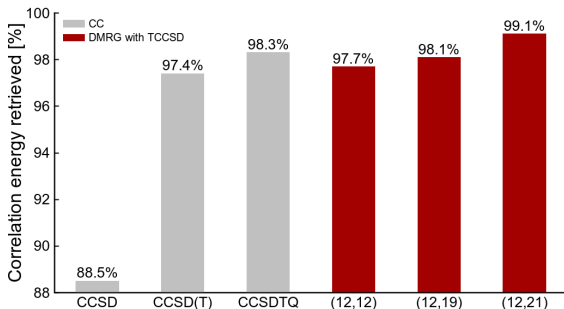
- ▶ determined through the usual CC
- ▶ account for dynamic correlation

$$\begin{aligned} \Psi_{\text{TCCSD}} &= e^{(\mathcal{T}_1^{\text{ext}} + \mathcal{T}_2^{\text{ext}})} e^{(\mathcal{T}_1^{\text{CAS}} + \mathcal{T}_2^{\text{CAS}})} \Psi_{\text{ref}} \\ &\approx e^{(\mathcal{T}_1^{\text{ext}} + \mathcal{T}_2^{\text{ext}})} \Psi_{\text{CASCI}} \end{aligned}$$

- ▶ Requires **only small modifications** of the CC code

Chromium dimer – correlation energies

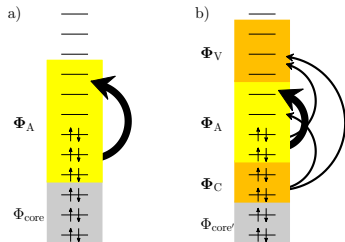
- ▶ Single-point calculation at 1.5 Å
- ▶ One-particle basis: RHF with Ahlrichs' SV basis set → (48e,42o)
- ▶ DMRG space selected based on $S^{(1)}$ profile
- ▶ DMRG performed with DBSS ($\epsilon_{\text{tr}} \approx 10^{-7}$)
- ▶ Extrapolated DMRG by Olivares-Amaya et al. JCP 142, 034102, 2015 serves as a FCI benchmark



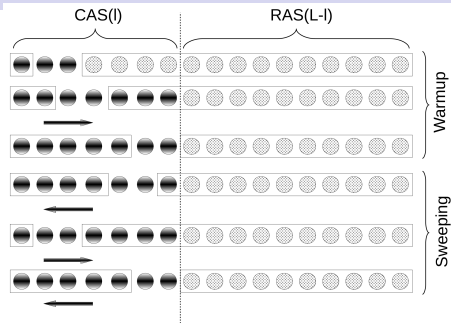
Error analysis: Faulstich, Laestadius, ÖL, Schneider, Kvaal: Quadratic error bound for a given CAS-EXT split.

Extensions: Jiri Pittner on similarity transformed TCCSD, Andrej Antalík on LPNO-TCCSD and Jan Brandejs on 4c-DMRG-TCCSD.

Restricted active space DMRG Barcza, Werner, Zaránd, Ö.L., Szilvási (2021)

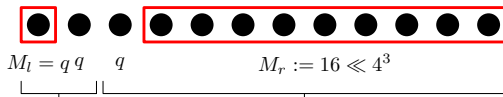


Schematic illustration of the CAS and RAS concepts.



DMRG-RAS scheme

- In the RAS scheme, in addition to active orbitals some virtual (V) and core (C) orbitals can also be excited with restrictions: the maximal number of particle excitations in these orbitals is r .
- Implementation through the dynamically extended active space (DEAS) procedure. [ÖL, J. Sólyom, 2003](#), (similar appr. by [Larsson et al 2022](#))



Rigorous mathematical analysis of the error dependence

Friesecke, Barcza, Ö.L. (2022)

N-electron Hilbert space for the DMRG-RAS method:

$$\mathcal{H}(\ell, k) = \mathcal{H}_{\text{CAS}}(\ell) \oplus \mathcal{H}_{\text{RAS}}(L - \ell, k)$$

$$E^0(\ell, k) = \min_{\Psi \in \mathcal{H}(\ell, k): \langle \Psi, \Psi \rangle = 1} \langle \Psi, H\Psi \rangle,$$

partitioning of the full Hamiltonian into a reference Hamiltonian associated with the CAS energy and a remainder:

$$\begin{aligned} H &= H_0 + H' \text{ with} \\ H_0 &= PHP + (E_0 + \Delta)Q \\ H' &= H - PHP - (E_0 + \Delta)Q \end{aligned}$$

where P is the projector of \mathcal{H} onto the CAS Hilbert space \mathcal{H}_{CAS} , $Q = I - P$ is the projector onto the RAS Hilbert space, E_0 is the CAS ground state energy, i.e.

$$E_0 = E_{\text{CAS}}^0(\ell),$$

and $\Delta > 0$ is a parameter to be chosen later.

This partitioning has the following desirable features:

- (i) The CAS ground state energy is the ground state energy of H_0
- (ii) The operator $H_0 - E_0$ is invertible on the orthogonal complement of the ground state of H_0 , yielding well-defined perturbation corrections at all orders;
- (iii) the first order perturbation correction $E^{(1)} = \langle \Psi_0 | H' | \Psi_0 \rangle$ vanishes regardless of the choice of the orbitals and parameters such as ℓ and Δ ;
- (iv) H_0 does not couple the CAS and RAS Hilbert spaces, with all the coupling contained in H' .

The latter property is evident by re-writing

$$H - PHP = \underbrace{QHP}_{H_{\text{CAS} \rightarrow \text{RAS}}} + \underbrace{PHQ}_{H_{\text{RAS} \rightarrow \text{CAS}}} + \underbrace{QHQ}_{H_{\text{RAS} \rightarrow \text{RAS}}}$$

(where the first term maps CAS to RAS, the second one, RAS to CAS, and the last one, RAS to itself), and makes transparent that DMRG-RAS can be considered an *embedding method*.

Look at the ground state energy $E_\lambda(\ell, k)$ of $H_0 + \lambda H'$ on $\mathcal{H}(\ell, k)$ for small $\lambda > 0$. We focus on the standard choice $k = 2$.

Assuming the ground state of H_0 is nondegenerate and denoting it by Ψ_0 ,

$$E_\lambda^{\text{FCI}} = E_0 + \lambda E^{(1)} + \lambda^2 E^{(2)} + \lambda^3 E^{(3)} + O(\lambda^4) \text{ as } \lambda \rightarrow 0.$$

Working through the algebra we get for the overall error scaling

$$e_\lambda^{\text{RAS}} = E_\lambda(\ell, 2) - E_\lambda^{\text{FCI}} = O(\lambda^4) \text{ as } \lambda \rightarrow 0.$$

On the other hand, since $E^{(1)} = 0$ and $E^{(2)} < 0$, the pure CAS error satisfies

$$e_\lambda^{\text{CAS}} = E_0 - E_\lambda^{\text{FCI}} = \Omega(\lambda^2) \text{ as } \lambda \rightarrow 0.$$

In the **weak coupling limit** this gives a scaling law which relates the CAS and DMRG-RAS error:

$$e_\lambda^{\text{RAS}} = O\left((e_\lambda^{\text{CAS}})^2\right) \text{ as } \lambda \rightarrow 0.$$

In general: the optimal Δ is the *expected value of the spectral gap between the pure RAS eigenvalues and the CAS ground state* with respect to the particular RAS state $\Psi' / \|\Psi'\|$.

New extrapolation procedure: DMRG-RAS-X

DMRG-RAS is a fully self-consistent method, and therefore capable of capturing more than the guaranteed perturbation contributions, thus we observe a scaling law of the semi-empirical form

$$E^0(\ell, 2) - E^{\text{FCI}} = a(E_{\text{CAS}}^0(\ell) - E^{\text{FCI}})^p \text{ for some } p > 1.$$

To predict the exponent p , the prefactor a and the offset E_{FCI} from $E_{\text{CAS}}^0(\ell)$ and $E^0(\ell, 2) = E_{\text{RAS}}(\ell)$ is achieved by minimizing the mean squared regression error of RAS versus CAS error in a log log plot,

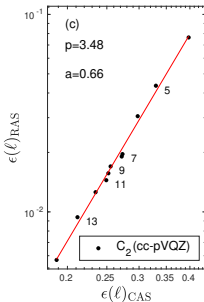
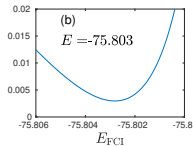
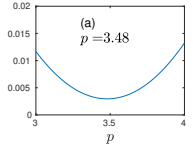
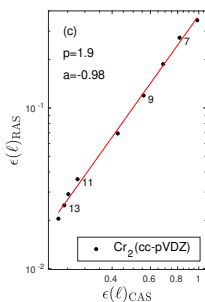
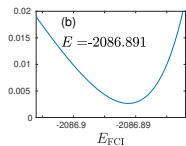
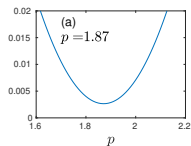
$$\text{MSE} = \frac{1}{n} \sum_{\ell} \left(y_{\ell} - (p \cdot x_{\ell} + \log a) \right)^2$$

where n is the number of datapoints and

$$x_{\ell} = \log(E_{\text{CAS}}(\ell) - E_{\text{FCI}}), \quad y_{\ell} = \log(E_{\text{RAS}}(\ell) - E_{\text{FCI}}).$$

This reduces to a minimization over the single free variable E_{FCI} , thus the predicted FCI energy is

$$E_{\text{RAS-X}} = \arg \min_{E^{\text{FCI}}} \text{MSE},$$



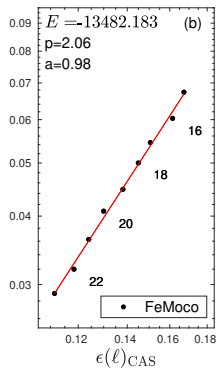
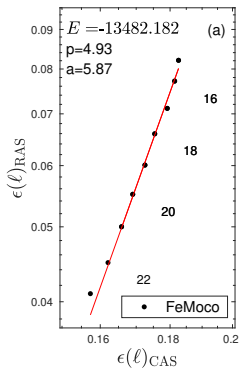
CAS(12,68)

CAS(8,108)

- we first perform our extrapolation scheme yielding the predicted parameters shown in (a) and (b).
- Next, using the predicted full-CI energy, $E_{RAS-X} = -2086.891$, we show in (c) that the linear scaling on double logarithmic axes for different ℓ values is recovered, as expected.
- $E^0(17, 2) = -2086.8769$ is already below the CCSDTQ by 8×10^{-3} ,
- the extrapolated energy is between $E_{RAS-X} = 2086.884$ ($p_{RAS-X} = 2.06$) and $E_{RAS-X} = 2086.891$ ($p_{RAS-X} = 1.88$) using the first 12 or 14 data points in the fit, leading to an **error estimate of the order of 10^{-3}** .

Method	Ground state energy
i-FCIQMC-RDME	-13482.17495(4)
i-FCIQMC-PT2	-13482.17845(40)
sHCI-VAR	-13482.16043
sHCI-PT2	-13482.17338
DMRG	-13482.17681
DMRG(D=8192)	-13482.1718
DMRG(D=10240,NO)	-13482.1754
RAS(23)	-13482.1421
RAS(23,NO)	-13482.1544

Non-extrapolated ground state energies obtained by various methods for the **FeMoco** in **CAS(54,54)** orbital space.



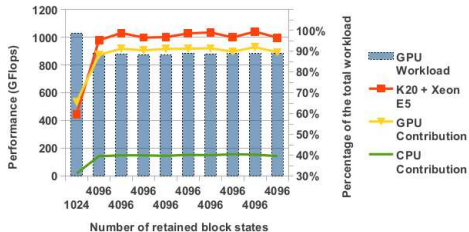
(a) Result of the DMRG-RAS-X for the FeMoco for the model space taken from Ref. Reiher(2007).

(b) The same but for the natural orbital basis.

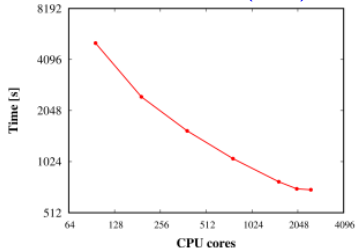
Produced on CPU-GPU for less than one day
 Friesecke, Barcza, ÖL (2023)

Towards exascale computations on supercomputers

GPU: MPS and TNS
on kilo-processor architectures:
Nemes, Barcza, Nagy, Ö.L., Szolgay, 2014



Massive parallelization
Brabec, Brandejs, Kowalski
Xantheas, Ö.L., Veis (2020)



(a) Davidson procedure

FeMoco cluster
[CAS(113,76)]



Centralized scheduling: unideal society

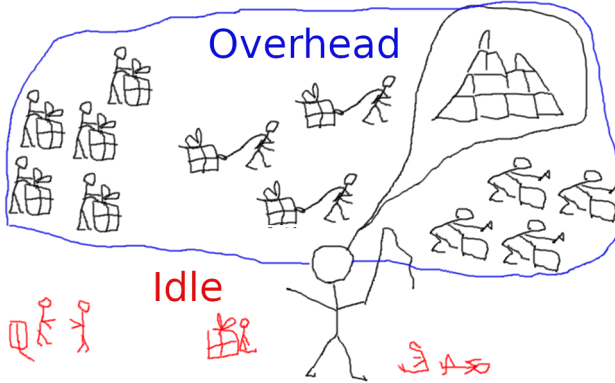
- Set of workers to generate tasks → Workers are threads
- Set of workers to transfer tasks → Transfer: IO communication
- Set of workers to execute tasks → CPU, GPU, FPGA units



- ▶ Central scheduler has to organize the full workflow, measure complexity of tasks, distribute tasks, check execution etc
- ▶ Central scheduler envisions the global aim & wants to accomplish it
- ▶ **Tasks: several millions of independent tensor and matrix operations**

Centralized scheduling: Huge overhead, units can be idle

- Central scheduler performs lot of measurements, estimations, communication to rearrange tasks and workers → huge overhead



- ▶ Central scheduler cannot see everything in a given moment → workers can be idle
- ▶ Too much workload on scheduler → inefficient scheduling, tasks can pile up partially

Self motivated workers → ideal "team-like" society

- Central unit: Contractor, contract book (only meta-data communicated, boolean-like bookkeeping flags)
- Everybody is motivated to achieve global aim

Tasks



Transfer



Task creators

Contract book

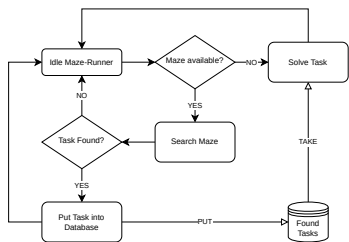


Executors

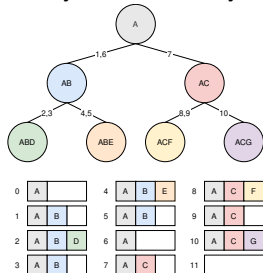
~~Idle~~

~~Overhead~~

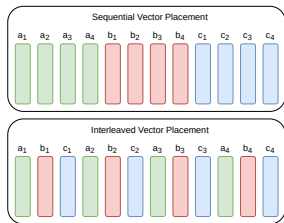
Life Cycle of a Maze-Runner Thread.



Graph theory based memory management



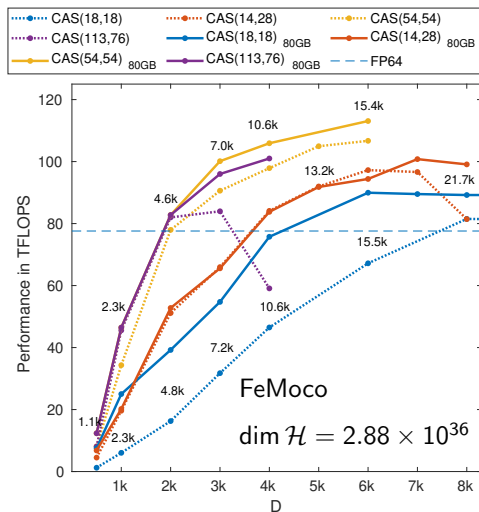
Strided Batched operations via data localization



Execution via hierarchy of tasks

Hashing				By groups				By tasks			
A	B	C	D	A	B	C	D	A	B	C	D
0	1	2	3	0	1	2	3	0	2	2	0
0	1	2	3	0	1	2	3	1	3	3	1
0	1		3	0	1		3	2	0		2
2	1			0	1			3	1		
0				0				0			
2				0				1			

Boosting performance via AI accelerators. Wall time: $D^3 \rightarrow D$



- A factor of 40 speedup compared to a single node with 128 cores

→ flexible scaling

- 116 TFLOPS > 76 TFLOPS of the FP64 limit of NVIDIA → utilization of highly specialized tensor core units (TCU)

- Power consumption reduced by a factor of 5 to 8

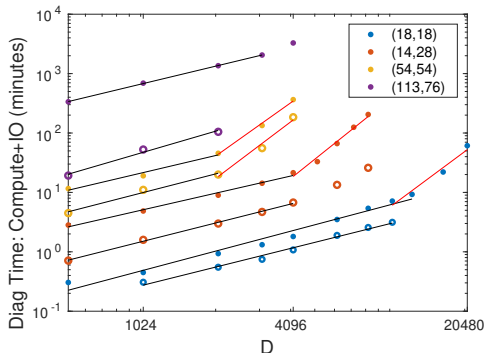
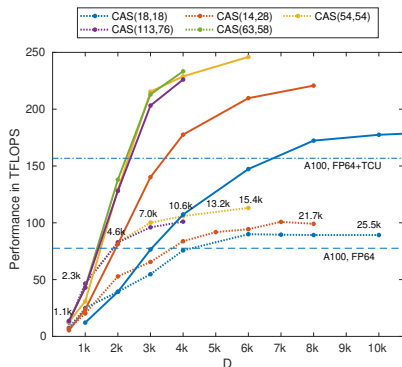
→ Green DMRG

- NVIDIA DGX H100 and Grace Hopper GH200: Testing performance up to ~ 250 TFLOPS in collab with NVIDIA and SandboxAQ

M. van Damme, A. Menczer, M. Ganahl, J. Hammond, Ö.L

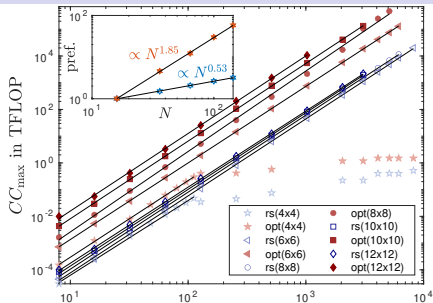
- Combination of our MPI and GPU kernels: → petascale computing.

Quarter petaflops on a single node $\sim 10000\times$ speedup

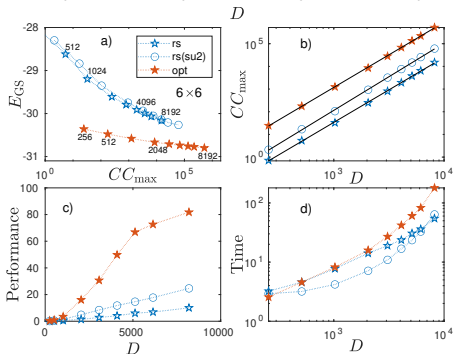


- NVIDIA DGX H100 and Grace Hopper GH200: Testing performance up to ~ 250 TFLOPS in collab with NVIDIA and SandboxAQ M. van Damme, A. Menczer, M. Ganahl, J. Hammond, S. Xantheas, ÖL
- New model to utilize NVIDIA D2D links. A. Menczer ÖL (unpublished 2023)
- Combination of our MPI and GPU kernels: full replacement of *boost library*, asynchronous IO, multiNode-multiGPU
→ **petascale computing**. A. Menczer ÖL (unpublished 2023-2024)

Maximum computational complexity for 2D $t - t' - V$ model



- The solid lines are first-order polynomial fits leading to exponents $\nu \simeq 3 \pm 0.2$
- inset: scaling of the prefactor as a function of system size N with fitted exponents 0.53 and 1.85 for the real space and for the optimized basis, respectively.



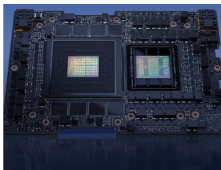
- Half-filled 6×6 Hubbard model at $U = 4$ on a torus geometry
- Performance in TFLOPS
- Time in minutes

Our TNS/DMRG code will be used as one of the benchmarks



NVIDIA GH200 Grace Hopper Superchip

The breakthrough accelerated CPU for large-scale AI and high-performance computing (HPC) applications.



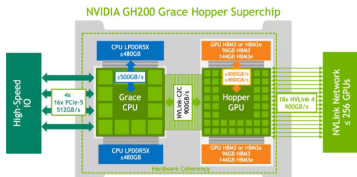
The World's Most Versatile Computing Platform

The NVIDIA Grace Hopper™ architecture brings together the groundbreaking performance of the NVIDIA Hopper™ GPU with the versatility of the NVIDIA Grace™ CPU in a single superchip, connected with the high-bandwidth, memory-coherent NVIDIA® NVLink® Chip-2-Chip (C2C) interconnect.

NVIDIA NVLink-C2C is a memory-coherent, high-bandwidth, and low-latency interconnect for superchips. The heart of the GH200 Grace Hopper Superchip, it delivers up to 900 gigabytes per second (GB/s) of total bandwidth, which is 7X higher than PCIe Gen5 lanes commonly used in accelerated systems. NVLink-C2C enables applications to oversubscribe the GPU's memory and directly utilize NVIDIA Grace CPU's memory at high bandwidth. With up to 480GB of LPDDR5X CPU memory per GH200 Grace Hopper Superchip, the GPU has direct access to 7X more fast memory than HMB3 or almost 8X more fast memory with HBM3e. GH200 can be easily deployed in standard servers to run a variety of inference, data analytics, and other compute and memory-intensive workloads. GH200 can also be combined with the NVIDIA NVLink Switch System, with all GPU threads running on up to 256 NVLink-connected GPUs and able to access up to 144 terabytes (TB) of memory at high bandwidth.

Key Features

- > 72-core NVIDIA Grace CPU
- > NVIDIA H100 Tensor Core GPU
- > Up to 480GB of LPDDR5X memory with error-correction code (ECC)
- > Supports 96GB of HBM3 or 144GB of HBM3e
- > Up to 624GB of fast-access memory
- > NVLink-C2C: 900GB/s of coherent memory



Outlook: TNS in nuclear structure theory

Suggested/required tasks to be completed for nuclear structure theory

- ▶ Long time evolution via mode optimization + BUG integrator (basis-update & Galerkin)
- ▶ High spin $SU(2)$ symmetries
- ▶ Higher dimensional tensor network topologies, like TTNS, T3NS
- ▶ DMRG-TCC
- ▶ DMRG-RAS-X alternative to IMSRG
- ▶ Multipartite entanglement
- ▶ DMRG-SCF like basis optimization
- ▶ Particle entanglement and clustering
- ▶ Dissipative quantum systems/Limbladian
- ▶ 3-body interactions + massive parallelization

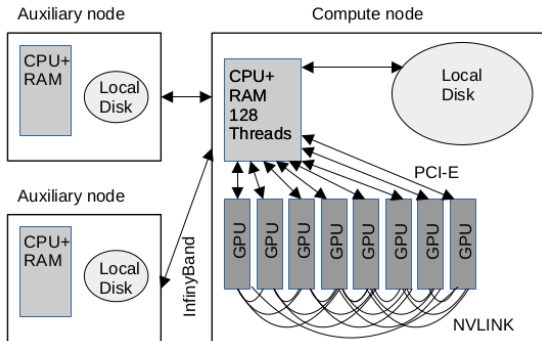
Conclusion

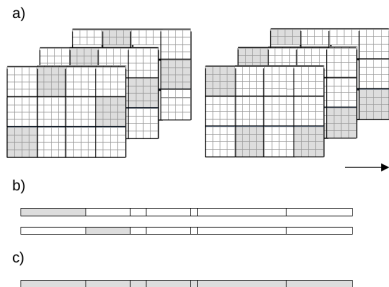
- ▶ Tensor topologies together with proper basis representations are important for efficient data sparse representation of the wavefunction
- ▶ Global mode transformation: MPS/TNS based black-box tool to improve basis
- ▶ Long time evolution with adaptive mode transformation is a promising direction in collab. Eisert, Lubich
- ▶ Combination of TNS with other (conventional) methods can exploit benefits of the individual methods
- ▶ Massive Parallelization MPI and NVIDIA/AMD → exascale computation
- ▶ → Simulation of realistic material properties in collab. Riverlane, Furukawa

Supports: Lendület grant of the Hungarian Academy of Sciences, the Hungarian National Research, Development and Innovation Office TKP2021-NVA-04, Hungarian Quantum Technology National Excellence Program, Quantum Information National Laboratory of Hungary, European Research Area(ERA), Alexander von Humboldt Foundation (Germany), Hans Fischer Senior

For discussions: Cost optimized TNS Menczer, ÖL 2024 in prep.

- H100 costs 100 USD/hour on Google Cloud
- Schematic plot of hardware topology illustrating the various communication channels (arrows), such as host to host (H2H), host to device (H2D) and device to host (D2H), and device to device (D2D), i.e., InfiniBand, PCI-E, and NVLink, accordingly.
- The compute node is a very powerful and expensive unit surrounded by one or more cheap auxiliary nodes with minimal computational capacity, but with substantial amount of RAM





a) Schematic plot of quantum number based block sparse representation of matrices and tensors. Shaded area indicates used sectors that are processed in computation and data serialization.

b) Skeleton of serialized data segments used during disk IO save procedure or MPI based communication. Only one segment is filled with data in a given time (shaded region) and transferred immediately to storage media or to another node. This requires only a small additional memory used for buffering.

c) Skeleton of serialized data segments filled completely with data when asynchronous save IO procedure is utilized. This leads to substantial increase in peak memory to store redundant data that is saved to storage media in parallel to subsequent computation tasks.