Combining state-of-the-art quantum chemistry and machine learning make gold standard potential energy surfaces accessible for medium-sized molecules

Apurba Nandi[*,†,‡,⊥] and Péter R. Nagy[*,¶,§,‖,⊥]

†*Department of Chemistry and Cherry L. Emerson Center for Scientific Computation, Emory University, Atlanta, Georgia 30322, United States*

‡*Department of Physics and Materials Science, University of Luxembourg, L-1511, Luxembourg City, Luxembourg*

¶*Department of Physical Chemistry and Materials Science, Faculty of Chemical Technology and Biotechnology, Budapest University of Technology and Economics, Műegyetem rkp. 3., H-1111 Budapest, Hungary*

§*HUN-REN–BME Quantum Chemistry Research Group, Műegyetem rkp. 3., H-1111 Budapest, Hungary*

‖*MTA-BME Lendület Quantum Chemistry Research Group, Műegyetem rkp. 3., H-1111 Budapest, Hungary*

⊥*The authors contributed equally to this study.*

E-mail: apurba.nandi@emory.edu; nagy.peter@vbk.bme.hu

1 Abstract

Developing full-dimensional machine-learned potentials with the current "gold-standard" coupled-cluster (CC) level is a challenging already for medium-sized molecules due to the high computational cost. Consequently, researchers are often bound to use lower-level electronic structure methods such as density functional theory or second-order Møller–Plesset perturbation theory (MP2). Here, we demonstrate on a representative example that gold-standard potentials can now be effectively constructed for molecules of 15 atoms using off-the-shelf hardware. This is achieved by accelerating the CCSD(T) computations via the accurate and cost-effective frozen natural orbital (FNO) approach. The Δ-machine learning (Δ-ML) approach is employed with the use of permutationally invariant polynomials to fit a full-dimensional PES of the acetylacetone molecule, but any other effective descriptor and ML approach can similarly benefit from the accelerated data generation proposed here. Our benchmarks for the global minima, H-transfer TS, and many high-lying configurations show the excellent agreement of FNO-CCSD(T) results with conventional CCSD(T) while achieving a significant time advantage of about a factor of 30 – 40. The obtained Δ-ML PES shows high fidelity from multiple perspectives including energetic, structural, and vibrational properties. We obtain the symmetric double well H-transfer barrier of 3.15 kcal/mol in excellent agreement with the direct FNO-CCSD(T) barrier of 3.11 kcal/mol as well as with the benchmark CCSD(F12*)(T+)/CBS value of 3.21 kcal/mol. Furthermore, the tunneling splitting due to H-atom transfer is calculated using a 1D double-well potential, providing improved estimates over previous ones obtained using an MP2-based PES. The methodology introduced here represents a significant advancement in the efficient and precise construction of potentials at the CCSD(T) level for molecules above the current limit of 15 atoms.

Keywords:

delta machine learning; potential energy surface; quantum dynamics; tunneling splitting; gold standard quantum chemistry; reduced-cost CCSD(T); natural orbital approach

2 Introduction

The potential energy surface (PES) originating from the Born–Oppenheimer approximation, that is the electronic energy of a molecule or material expressed as a function of its nuclear coordinates, plays a central role in theoretical and computational modeling. At least for relatively small systems and/or cost-efficient approximate models, it is possible to obtain (a required portion of) the PES in real time by repeatedly extracting the energies and gradients (or forces) from electronic structure models for the relevant nuclear configurations. However, the computational expense becomes prohibitive as the complexity of the electronic structure theory and the size of the system increases. An alternative approach constructs a precise analytical representation of the PES fitted to datasets of electronic energies (and/or gradients) that cover the required dimensions of the configuration space. Since the PES establishes a connection between the configuration of atomic nuclei and the forces acting upon them, it is key in numerous fields including reaction dynamics, kinetics, and thermodynamics, vibrational analysis incorporating both harmonic and anharmonic aspects, spectroscopic properties connected to nuclear motion, and structure optimization.

In the past 15 years, significant progress has been made in the enhancement of nonparametric machine learning (ML) approaches used to fit large datasets of electronic energies (and/or gradients) for molecules and molecular clusters. Three commonly utilized techniques for this purpose include permutationally invariant polynomials (PIPs), a linear regression method, as well as neural networks (NN), and Gaussian process regression (GPR), which are both non-linear regression methods.[1–11] Extending these methods to larger molecules of interest in the fields of chemistry, physics and biology presents a significant challenge.

Another major obstacle in the PES development for molecules of increasing size is to retain models of sufficiently high predictive power. Both the number of data points needed for PESs with increasing degrees of freedom and the computational cost of the quantum chemistry models scales steeply with the systems size. In particular, coupled-cluster (CC) models, with well converged wave-function and basis set expansions have proven their accuracy for a wide range of molecular and material properties.[12–14] Especially, the CC method with single and double excitations (CCSD) augmented with perturbative triples correction [CCSD(T)][15] is generally considered as the "gold standard" of quantum chemistry. However, the steep $\mathcal{O}(\mathcal{N}^4)$- and $\mathcal{O}(\mathcal{N}^7)$-scaling data storage and operation count complexity with system size $\mathcal{N}$ hinders the routine application of conventional CCSD(T) for PESs of systems above ca. 10–12 atoms.

Regarding the acceleration of CCSD(T), shared-memory intra-node (Open Multi-Processing, OpenMP) and/or multinode (Message Passing Interface, MPI) parallelization ideas were extensively explored.[16–26] For example, the hybrid MPI/OpenMP CCSD(T) from one of us utilizes all permutational symmetries of the CCSD(T) equations while exhibiting the highest peak-performance utilization reported so far.[26] Still, our record-sized single-point CCSD(T)/quadruple-ζ computation for a 31-atom molecule took almost 3 days on 224 cores, which illustrates that the routine generation of extensive databases, training sets, or PESs remains a bottleneck with conventional CCSD(T), at least for molecules of a few dozen atoms.

Explicitly correlated (F12) CC approaches can also be effective to reduce the basis set requirement of CCSD(T),[27–29] out of which our recent (T+) approach in combination with the CCSD(F12*) model[30] was shown to be particularly effective.[31,32] Alternatively, reduced-cost CCSD(T) approaches, such as the here employed frozen natural orbital (FNO) method (see Sect. 3.1) can be utilized to compress the space spanned by the virtual molecular orbitals (MOs).[33–37] The combination of the FNO approach with our MPI/OpenMP CCSD(T) implementations was especially beneficial to enable basis set limit FNO-CCSD(T) computations

for up to 50 atoms.[32,36]

In addition to the high cost of single-point CCSD(T) computations, the number of configurations required for high-dimensional PES construction also steeply increases with the number of atoms. Therefore, to compute PESs of larger molecules, lower-level electronic structure methods like density functional theory (DFT) and second-order Møller–Plesset perturbation theory (MP2) are employed, while CCSD(T)-level PESs of molecules above 10 atoms have been reported in only a handful of recent studies.[38–48] As an illustration, the symmetric gradient domain machine learning model have been proposed to construct global the force field of flexible molecules up to with 15 atoms at the CCSD(T) level by Tkatchenko and co-workers in 2018.[38] In 2019, Roitberg and co-workers introduced a transfer learning (TL) technique to develop neural network potentials for several organic and drug like molecules approaching CCSD(T) quality correcting a DFT dataset.[40] Meuwly and co-workers employed TL using thousands of local CCSD(T) energies to improve their MP2-based neural network PESs for malonaldehyde, acetoacetaldehyde, and acetylacetone.[41] Recently, Daru, Behler, and Marx reported a high dimensional coupled-cluster level neural network potential for liquid water for condensed phase simulations including nuclear quantum effects via path integral dynamics.[46]

Instances of PESs for chemical reactions involving 6–10 atoms have been documented, which were constructed by fitting tens or even hundreds of thousands of CCSD(T) energy data points.[49–52] In 2016, Bowman and co-workers reported the PES for 10-atom formic acid dimer $(HCOOH)_2$, using 13475 energies at CCSD(T)-F12 and triple-ζ level of theory.[53] This PES was subsequently applied for zero-point energy computation using the diffusion Monte-Carlo (DMC) method and ground-state tunneling splitting for the H-transfer process. The cost-reduction capabilities of local approximations to CCSD(T)[54–57] has also motivated their use in PES development for medium-sized systems.[41,43,44,46] In 2021, a 15-atom acetylacetone PES was developed by Bowman and co-workers using 2151 local approximated CCSD(T)-F12 computations.[43] More recently, Nandi *et al.* utilized an efficient fragment-based molecular

tailoring approach to construct full-dimensional CCSD(T) PESs for 15-atom acetylacetone and tropolone molecules.[44,58] The latter tropolone PES is extensively used for ring-polymer instanton calculations to compute tunneling splittings in agreement with the experimental values.[58]

To advance high-throughput reference data generation required, e.g., for the PES generation protocols of systems of up to about 15–20 atoms, like in the present study, we recommend and employ here the FNO-CCSD(T) model for the first time. We combine the uniquely efficient FNO-CCSD(T) implementation[36,37] of the MRCC program suite[59,60] (see Sect. 3.1) and advanced ML approaches (see Sect. 3.2) to enable routine PES generation at the 15-atom scale on the example of acetylacetone. For that, we first extend the previous FNO benchmarks on, e.g., atomization and reaction energies, as well as molecular interactions[32,36,37,61] to configuration energies required for the PES fitting by assessing the accuracy of FNO-CCSD(T) against approximation free CCSD(T) reference. Then, we recommend accurate and routinely applicable settings suitable for high-throughput FNO-CCSD(T) computations, which we also employ for several hundreds of acetylacetone configurations to develop its CCSD(T)-level PES. Here we use PIP method to fit the coupled-cluster level PES. The PIP approach is well-established and has been successfully applied for small to medium-sized molecules with numerous applications. Next, this PES is utilized for conducting overall fidelity assessments such as geometry optimization, calculating normal mode frequencies, and determining the tunneling splitting for the H-transfer process.

The paper is organized as follows. In the next section, we provide a brief overview of the FNO-CCSD(T) technique utilized for data generation, as well as the PIP approach to fitting the PES employing the Δ-ML method. Then, we present benchmark results for the accuracy of the FNO-CCSD(T) method as well as the newly fitted PES. Finally, the "Conclusions and Outlook" section ends the paper.

3 Methods

3.1 Accelerated coupled cluster methods

The main computational difficulties with CCSD(T) originate from the equations determining the CCSD and (T) amplitudes, which exhibit $\mathcal{O}(\mathcal{N}^6)$- and $\mathcal{O}(\mathcal{N}^7)$-scaling operation count complexity, respectively, and $\mathcal{O}(\mathcal{N}^4)$-scaling data complexity (for both terms). For the targeted 15-atom system, even with some of the well-optimized implementations and 10–20-core processors, the steep scaling of CCSD(T) already requires hours to days of wall-clock time and tens to hundreds of GBs of data with sufficiently large, triple- or quadruple-ζ basis sets.

Recent developments, such as the use of density-fitting (DF),[24–26,62,63] help to reduce the latter storage challenges. For example, our recent integral-direct DF-CCSD(T) implementation eliminated problematic disk I/O and network traffic bottlenecks as well as also enabled us to develop the most data- and memory-economic algorithms for both the CCSD[26] and the (T)[63] parts reported so far.

Considering the generation of extensive benchmark data sets, such as with the large number of configuration energies for the PES fitting here, the overall task already consists of many independent CCSD(T) computations. Thus, parallelization within a single CCSD(T) computation is only useful up to a certain number of cores, where the available processor performance can be effectively utilized (given other limitations, such as memory or data transfer bottlenecks). For instance, it is not efficient to run 128 different single-thread CCSD(T) jobs on the 128-core nodes employed here due to data bandwidth bottlenecks, while providing multiple (ca. 16–32-core) jobs per node with separate memory channels appear to be highly effective.

Therefore, it is worthwhile to reduce the computational cost via the frozen natural orbital (FNO) approach.[33–37] The natural orbitals (NOs) of FNO method are usually obtained as the eigenvectors of a one-particle density matrix, built mostly using a model wavefunction, such as MP1.[33–37] Then, the NOs with occupation numbers below a threshold are kept frozen,

that is, only the most important, $\bar{n}_\mathrm{v}$ number of frozen NOs are taken into account in the remainder of the CCSD(T) computation. The benefit is that the rate-determining steps of both CCSD and (T) scale with the fourth power of the number of virtual orbitals, n_v. Consequently, a theoretical speedup of $(n_\mathrm{v}/\bar{n}_\mathrm{v})^4$ can be anticipated for the most demanding steps of both the CCSD and the (T) parts. Additionally, the memory requirement for the DF integrals and the doubles amplitudes also decreases by a factor of $(n_\mathrm{v}/\bar{n}_\mathrm{v})^2$, while a factor of $(n_\mathrm{v}/\bar{n}_\mathrm{v})^3$ data compression is realized for intermediates required for the (T) algorithm.[26,63,64] Such data compression has an additional benefit for our integral-direct DF-CCSD(T) algorithm,[36] eliminating the need for repeated four-center ERI assembly in the FNO-CCSD(T) computations presented here. Moreover, both in our conventional[26] and FNO-based[36] CCSD(T) codes of the MRCC program suite[59,60] the operation count and the memory requirement are fully optimized by exploiting all permutational symmetries with an unparalleled 50–70% peak performance utilization up to hundreds of cores. In addition, all terms of our DF-CCSD(T) code are hand-optimized,[26,63] which is particularly important for maintaining the high efficiency with compressed FNO-basis sets.[36] Moreover, the negligible disk I/O and network use allow for the execution of a large number of medium-sized CCSD(T) computations simultaneously on the same cluster/node and network file system.

Due to the properties of the MP one-particle density matrix, the FNO approach can be interpreted as the singular value decomposition (SVD) of the MP doubles amplitude tensor, and thus the FNO approximation provides an optimized model basis for the truncated representation of the wavefunction. Analogous to the FNO method, the natural auxiliary function (NAF) approach was introduced to compress the three-center ERIs appearing in DF methods via SVD.[36,65] To that end, the NAFs can be considered as the optimal linear combination of the DF auxiliary functions. Similarly to the case of the FNOs, the compression rate of the NAF basis can be controlled via a single truncation parameter by dropping singular vectors with a singular values below the NAF threshold. The NAF approximation is very robust and its accuracy can be set to approach that of the DF approximation.[36,65]

The combination of the FNO and NAF approaches is beneficial, since after the introduction of the molecule-specific and compressed FNO virtual basis, the number of remaining FNO product densities decreases by a factor of $(n_{\mathrm{v}}/\bar{n}_{\mathrm{v}})^2$. Consequently, a significant DF auxiliary basis compression can be achieved by generating a molecule specifically optimized NAF basis corresponding to only the retained FNO basis.

The error of the FNO and NAF basis compressions can be corrected via an MP2 level energy correction:[36,37]

$$E^{\mathrm{FNO-CCSD(T)}} = E^{\mathrm{CCSD(T)}}_{\mathrm{FNO}} + E^{\mathrm{MP2}}_{\mathrm{MO}} - E^{\mathrm{MP2}}_{\mathrm{FNO}} \,, \tag{1}$$

where the subscripts indicate that the corresponding CCSD(T) or MP2 energies are evaluated using the compressed FNO (and NAF) basis or the complete molecular orbital (MO) (and DF auxiliary) basis.

3.2 Δ-Machine Learning for PES construction

Δ-Machine learning[66–68] is a general method to bring a property, such a PES, trained on an efficient lower-level method close to the accuracy of a higher-level method. Here, we correct an MP2-level PES to the gold standard CCSD(T) level, for which Δ-ML approach was already employed and tested extensively also by some of us.[43,69,70] The underlying theory of the Δ-ML approach can be succinctly summarized by the following equation:

$$V_{LL\rightarrow CC} = V_{LL} + \Delta V_{CC-LL} \tag{2}$$

In this equation, $V_{LL\rightarrow CC}$ represents the corrected potential energy surface, V_{LL} denotes a PES fitted to low-level electronic energies, such as from DFT or MP2, and ΔV_{CC-LL} corresponds to the correction PES which is a fit to the difference in high-level and low-level energies only (i.e. here without gradients). It is worth noting that the variation of ΔV_{CC-LL}, which represents the difference between CCSD(T) and DFT/MP2 energies, is not

as pronounced as that of V_{LL} in relation to nuclear configurations (see below in Fig. 5). Consequently, only a smaller number of high-level electronic energies are sufficient for fitting the correction PES. In our current application to acetylacetone, a total of 430 FNO-CCSD(T) electronic energies were computed to accomplish this corrected potential fitting.

Here, we employ the PIP approach to fit both the V_{LL} and ΔV_{CC-LL} PESs. The theory of permutationally invariant polynomials is well-established and has been presented in several review articles.[1,2,71–73] In terms of a PIP basis, the potential energy, V, can be written in a compact form as

$$V(\boldsymbol{x}) = \sum_{i=1}^{n_p} c_i p_i(\boldsymbol{x}), \tag{3}$$

where c_i are linear coefficients, p_i are PIPs, n_p is the total number of polynomials for a given maximum polynomial order and $\boldsymbol{x}$ are Morse variables. For example, $x_{\alpha\beta}$ is given by $\exp(-r_{\alpha\beta}/\lambda)$, where $r_{\alpha\beta}$ is the internuclear distance between atoms α and β. The range (hyper)parameter, λ, was chosen to be 2 bohr. The linear coefficients are obtained using standard least squares methods for a large data sets of electronic energies (and when available gradients as well) at scattered geometries.

In order to develop a corrected PES, we need to generate a dataset of high- and low-level energies. In this study, we used MP2/aug-cc-pVTZ energies and gradients as low-level data. The low-level PES, V_{LL} is taken from previously reported data by Chen Qu and co-workers[74] which was a fit using a data size of 5454 energies and their corresponding gradients spanning the energy range of 0 - 40 000 cm^{-1}. More details of this V_{LL} PES can be found in Ref. 74. We briefly note that the model and basis set employed for the FNO and NAF corrections in Eq. (1) is MP2 with the same basis set used for the CCSD(T) computation, which are in general independent from the low-level model and basis set choices.

To develop the correction PES, we train ΔV_{CC-LL} on the difference between the FNO-CCSD(T) and MP2 absolute energies (with aug-cc-pVTZ basis) for 430 geometries (provided in the Supplementary Material). The dataset of 430 geometries were sparsely selected from the MP2 dataset of 5454 geometries, which was taken from recently reported data by Nandi

et al.[44] A low-order PIP fit was employed because the difference ΔV_{CC-LL} is not as strongly varying as V_{LL} with respect to the nuclear configuration. We used maximum polynomial order of 2 with permutational symmetry {1,2,5,7} (this symmetry indicates that the transferring "H" atom is treated as distinguishable, remaining seven "H" atoms are permutable with each other, as the two "O" atoms and five "C" atoms) to fit the training data set. This leads to a total of 86 PIP bases such as unknown coefficients. The PIP basis to fit these V_{LL} and ΔV_{CC-LL} PESs were generated using the latest MSA software.[75,76] These coefficients are determined by a standard linear least-square regression method.

4 Computational details

All conventional[26] and reduced-cost[32,36,37] CCSD(T) computations were performed with the 2022 release of the MRCC quantum chemistry program suite, available open-source for academic use.[59,60] Closed-shell references and the frozen core approximation were applied in all correlation calculations. All valence occupied orbitals were kept in the FNO-CCSD(T) computations, only the virtual orbital space was compressed via the FNO approach, which was governed by the `lnoepsv` keyword. For all reduced-cost FNO-CCSD(T) computations the default threshold of the NAF approach was employed irrespective of the FNO threshold, that is 0.05 a.u., set automatically via the `naf_cor` keyword of MRCC.

For the AO basis set, the correlation consistent X-tuple-ζ (aug-)cc-pVXZ (X =D, T, and Q) sets[77,78] were employed with the corresponding DF auxiliary bases, (aug-)cc-pVXZ-RI-JK[79] and (aug-)cc-pVXZ-RI.[80] For the CCSD(F12*)(T+) explicitly correlated calculations the correlation consistent X-tuple-ζ cc-pVXZ-F12 (X = T, Q) AO basis sets[81] and the corresponding cc-pVXZ-F12-OPTRI CABS bases were employed.[82,83] The extrapolations of the HF[84] and the correlation energies[85] towards the complete basis set (CBS) limit were performed separately, according to standard expressions. The extrapolated results will be denoted as CBS($X, X+1$).

The wall-clock time measurements were carried out using 16 cores per job of 64-core AMD Epyc 7H12 CPUs. The employed two-socket nodes were in a national level computer cluster under production use, executing competing tasks of multiple users. The nodes contain about 220 GB accessible memory (allowing 27 GB memory/job) and do not have local disks, only network storage access.

5 Results and Discussions

In this Section we benchmark the computational settings used for the FNO-CCSD(T) PES generation. A few important configurations (see Fig. 1) are considered including the global minima (GM), a low-laying saddle point (TS-I), H-transfer transition state [TS(H)], a high energy structure denoted as A2, as well as 20 additional configurations exhibiting some of the highest uncertainty in comparison to previous results.[43] Then, after obtaining the FNO-CCSD(T) training set for 430 configurations and fitting the PES, the high quality of the PES will be illustrated on various structural and vibrational properties of acetylacetone.

5.1 Accuracy of the FNO-CCSD(T) configuration energies

The FNO approach (including also the NAF approximations) has been extensively benchmarked on, e.g,. atomization, reaction, and non-covalent interaction energies.[32,36,37] These properties are expected to be somewhat more challenging than the present case, as better compensation of FNO errors can be expected for energy differences between configurations of the same molecule. However, since such configuration energy tests are not yet available, we explore the accuracy of the FNO-CCSD(T) approach against the approximation free references. The FNO truncation errors of FNO-CCSD(T) with respect to the FNO approximation free reference are collected in Fig. 2 for the TS(H) and A2 configuration energies.

The previously established[36,37] default ($5\cdot10^{-5}$ or $3\cdot10^{-5}$) and tighter (10^{-5}) FNO thresholds are found to be highly reliable here too. Thus, we also explore more cost-efficient FNO

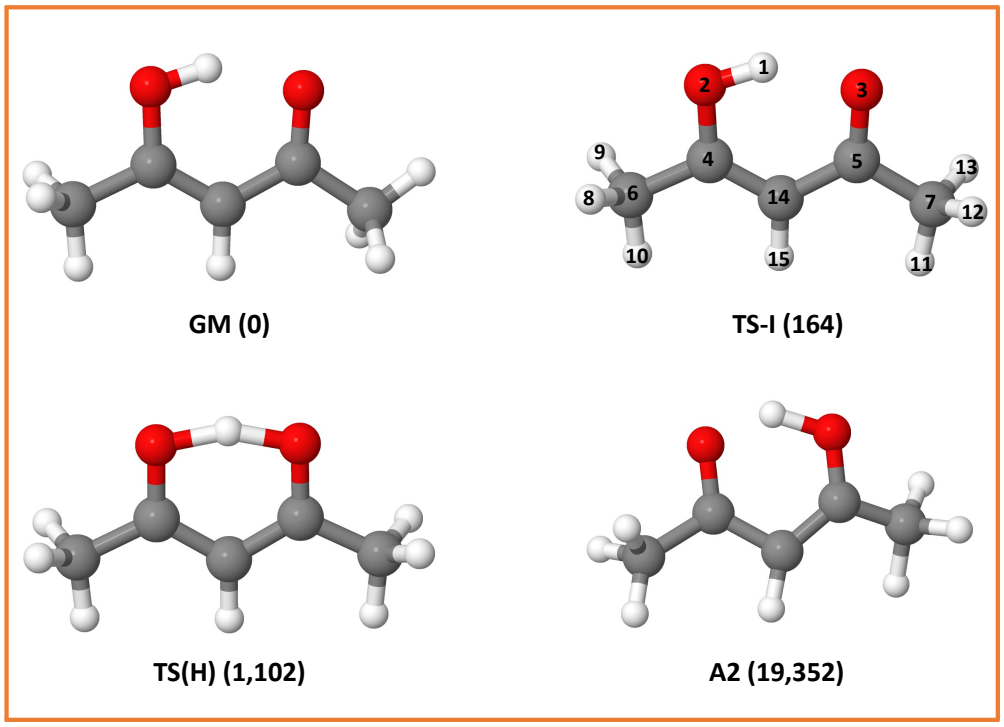

Figure 1: Geometry of global minimum (GM), low-laying saddle point (TS-I), H-transfer saddle point [TS(H)], and a high-energy structure (A2) of acetylacetone and their corresponding electronic energies (cm^{-1}) relative to the global minimum from Δ-ML PES. (Atomic numbering scheme was used to generate PIP bases to fit V_{LL} PES.)

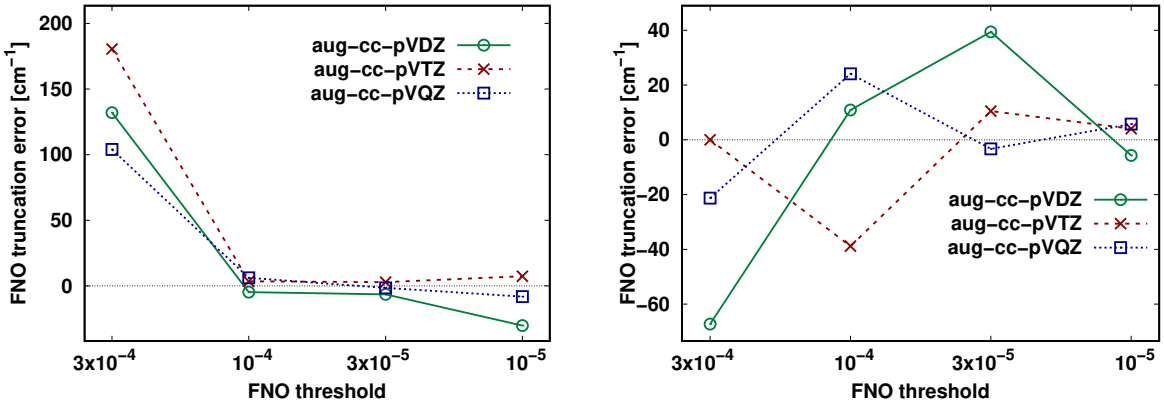

Figure 2: FNO truncation error for the TS(H) barrier height (left) and A2 configuration energy (right) with respect to the FNO threshold with aug-cc-pVXZ basis sets. The results are arranged to show increasing accuracy corresponding to decreasing FNO truncation threshold values on the x axis.

settings of $3 \cdot 10^{-4}$ and 10^{-4}, out of which 10^{-4} performs very well. First, the aug-cc-pVDZ (green circles) FNO errors in Fig. 2 are only moderately higher than those obtained with the practically more relevant aug-cc-pVTZ (red crosses) and aug-cc-pVQZ (blue squares) ones. Moreover, systematic convergence of the (absolute) FNO errors is found with tightening the FNO thresholds. Regarding the aug-cc-pVTZ and aug-cc-pVQZ basis sets, the FNO errors are within the [-0.11,0.07] kcal/mol or [-39,24] cm^{-1} range already with the 10^{-4} threshold for both the barrier height and the A2 configuration energy. It is also satisfactory, that the corresponding (absolute) FNO errors are below 0.03 kcal/mol or 10 cm^{-1} with the one step tighter, $3 \cdot 10^{-5}$ FNO settings.

While the accuracy of the FNO approach is the most important up to the region of TS(H), we also estimate the magnitude of the largest inaccuracies along the entire investigated interval of the PES. To that end, we selected 10 additional configurations where the largest disagreements occurred with the previously obtained local approximated CCSD(T) dataset.[43] All of these 10 configurations turned out to be of very high energy in the range of [33.6,87.9] kcal/mol or ca. [11750,30750] cm^{-1} above the GM. Therefore, we extended this list with 10 configurations selected randomly from the interval between the TS(H) and the configuration at the lower end of the above interval (33.6 kcal/mol).

In Fig. 3 we plot the resulting FNO errors corresponding to the 10^{-4} settings compared to the DF-CCSD(T) reference for these 10+10 configurations [as well as TS(H) and A2]. As expected, the FNO errors somewhat increase with the increasing configuration energy, which appears to be quite systematic for the randomly selected 10 configurations. In this region, the previously found FNO error of ca. 0.1 kcal/mol (35 cm^{-1}) roughly doubles by reaching the 30 kcal/mol (10500 cm^{-1}) configuration energy range. The average error up to this point is 0.15 kcal/mol (52 cm^{-1}), which is excellent considering the low probability of such high-energy configurations. The largest expected uncertainties can be estimated for the other 10 structures in the [33.6,87.9] kcal/mol or [11750,30750] cm^{-1} region. Indeed, here the average error grows to 0.3 kcal/mol (105 cm^{-1}), which is again acceptable considering the

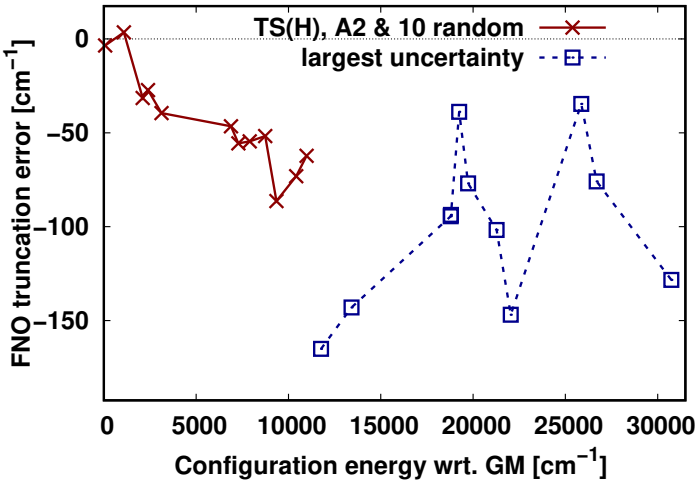

Figure 3: FNO-CCSD(T) configuration energy errors compared to the DF-CCSD(T)/aug-cc-pVTZ reference for all investigated configurations including TS(H), A2, 10 configurations with the highest uncertainty from the [33.6, 87.9] kcal/mol or [11750,30750] cm^{-1} interval, and 10 randomly selected configurations from the [3.1, 33.6] kcal/mol or [1084,11750] cm^{-1} interval.

very low population of these high-energy configurations. All in all, the steepness of the FNO error increase is much lower than the corresponding increase in the configuration energies, and more importantly, very small compared to the exponential decrease in the populations of these configurations.

We also study the FNO-CCSD(T) wall times as a function of the basis set and FNO threshold choice (Fig. 4). Considering the 3–4-fold increase of the wall times stepping from 10^{-4} to $3\cdot10^{-5}$ FNO settings and the satisfactory performance of both, FNO-CCSD(T)/aug-cc-pVTZ with 10^{-4} FNO threshold offers the best accuracy/cost performance for large-scale data generation in PES fitting.

As collected in Table 1, this choice corresponds to a speed up factor of 38 compared to the FNO approximation free CCSD(T)/aug-cc-pVTZ computations. In absolute terms, the FNO-CCSD(T)/aug-cc-pVTZ correlation energy computations with 10^{-4} settings take about 3–4 minutes with an additional 1–2 minutes for the SCF and integral transformation steps. These FNO-CCSD(T)/aug-cc-pVTZ computations require a minimum of 0.8 GB or, without any repeated integral evaluations, only a few GBs of memory. These hardware

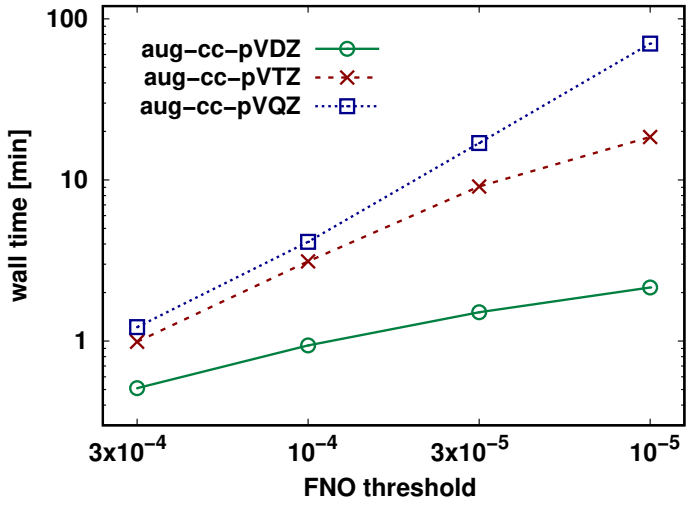

Figure 4: FNO-CCSD(T) correlation energy computation time measurements (in minutes) on 16 cores for the various FNO settings and basis sets.

requirements and the comparable amount of disk and network use make these FNO-CCSD(T) computations especially suitable for high-throughput reference data generation, e.g., for PES fitting or ML training tasks. For example, the 430 points used here for the acetylacetone PES would take less than 2 days on a 16 core machine, which could go down to about 6 hours when using all 128 cores of the here employed dual-socket AMD node. Running 8 FNO-CCSD(T) jobs on such nodes is feasible due to the moderate, few GB memory requirement (Table 1) and limited I/O and network use of the FNO-CCSD(T) implementation [26,36] in the MRCC program suite. [59,60]

Table 1: Wall times, speedups as well as minimal memory requirements (in GB units) without and with the FNO approach using the 10^{-4} threshold and aug-cc-pVXZ basis sets. The maximum utilized memory (in GB units) columns show the memory consumption of the jobs, while the minimal requirement would be sufficient for execution (with some repeated integral evaluation steps). The last column holds the corresponding FNO error in the TS(H) barrier height.

X	Wall time [min]		Speedup	CCSD(T) memory		FNO-CCSD(T) memory		FNO error
	CCSD(T)	FNO-CCSD(T)		min	max used	min	max used	[cm^{-1}]
D	3.7	0.9	3.9	1.0	7.4	0.27	1.3	-4.7
T	119.4	3.1	38.3	10.4	51.2	0.80	6.3	3.5
Q	995.2	4.1	240.9	64.7	107.5	0.99	8.3	6.2

5.2 Fitting and benchmarking the PES

We develop a new full-dimensional PES of acetylacetone at the FNO-CCSD(T)/aug-cc-pVTZ level of theory using the Δ-machine-learning (Δ-ML) approach. To obtain this CCSD(T) PES we add the correction $\Delta V_{\text{CC-LL}}$ to the low-level MP2 PES, V_{LL}. So, the development of this PES can be divided into two parts: low-level PES (V_{LL}) and a correction PES ($\Delta V_{\text{CC-LL}}$). Here we use a previously reported V_{LL} PES which is fit to 5454 energies and their corresponding gradients computed at the MP2/aug-cc-pVTZ level of theory.[74] For this fit, a maximum polynomial order of three was used which led to a total of 6207 PIP basis functions. The symmetry designation for this fit was {1,2,2,2,6,1,1}, meaning that the two oxygens (atoms 2 and 3) are treated as equivalent, as are the two carbons (atoms 4 and 5), the two terminal carbons (atoms 6 and 7), and the six hydrogens (atoms 8–13) on the terminal carbons are treated equivalent too (as labeled in Fig. 1). The remaining 'H' atoms and the central carbon are treated as unique. The weighted average fitting RMS errors for energies and gradients were 49 cm^{-1} and 29 cm^{-1} bohr^{-1}, respectively. More details of this PES can be found in Ref. 74.

To develop a new correction PES, we train $\Delta V_{\text{CC-LL}}$ on the difference between the FNO-CCSD(T) and MP2 absolute energies of 430 geometries. A plot of $\Delta V_{\text{CC-LL}}$ versus the MP2 energies is shown in Fig. 5. Note that we reference $\Delta V_{\text{CC-LL}}$ to the minimum of the difference between the FNO-CCSD(T) and MP2 energies (roughly -23 436 cm^{-1}). The energy range of $\Delta V_{\text{CC-LL}}$ is about 1800 cm^{-1} in Fig. 5, which is much smaller than the MP2 energy range relative to the minimum value (roughly 35 000 cm^{-1}).[74]

Thus, the difference $\Delta V_{\text{CC-LL}}$ is not as strongly varying as V_{LL} with respect to the nuclear configuration, and a low-ordered polynomial can be employed to fit this. We use a maximum polynomial order of 2 with permutational symmetry {1,2,5,7} to fit the dataset which leads to a much smaller number (86) of unknown linear coefficients. We perform both weighted average and unweighted fitting for $\Delta V_{\text{CC-LL}}$. In the process of weighted average fitting, a weight is assigned to each data point based on its energy. The weight is given by

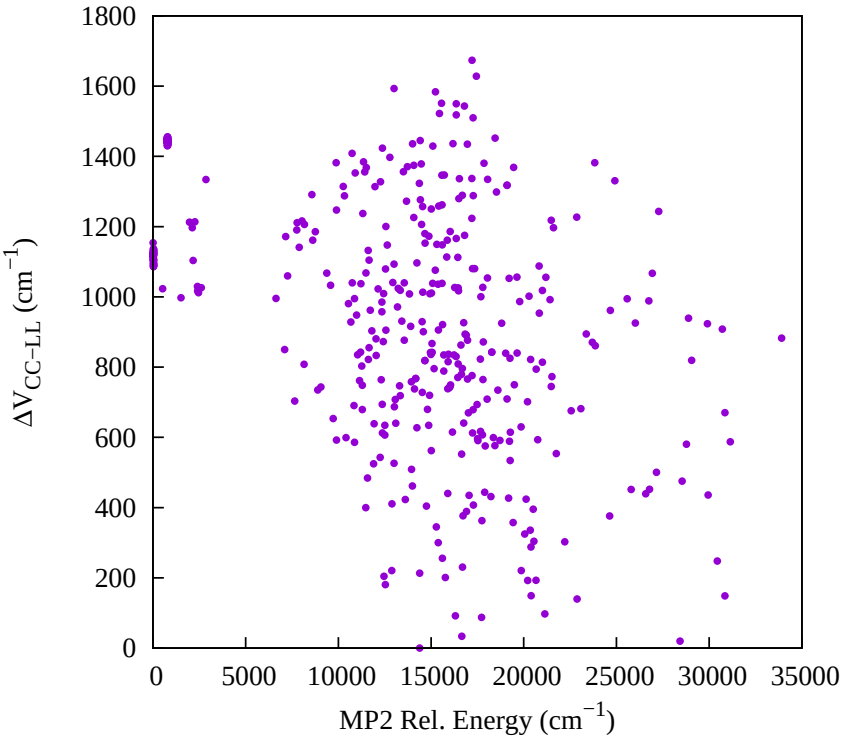

Figure 5: Plot of ΔV_{CC-LL} (relative to the reference value i.e. -23 436 cm^{-1}) vs MP2 energy relative to the global minimum value.

$wt = E_0/(E_0 + dE)$, where dE is the energy relative to the minimum in a.u., and E_0 is the parameter that we could modify. For the unweighted fitting, E_0 is typically set as a large number, such as 10^{10} a.u., resulting in all weights essentially being 1. In this case, we chose E_0 to be 0.02 a.u. to achieve the desired weighted average fitting matching Ref. 74. The RMSEs for the weighted and unweighted fitting are 11 cm^{-1} and 52 cm^{-1}, respectively. To examine the performance of the PES, we use weighted averaged $\Delta V_{CC\text{-}LL}$ fitting only.

We add this correction $\Delta V_{CC\text{-}LL}$ to the low level MP2 PES, V_{LL} to obtain the total CCSD(T) energies. We perform geometry optimization and normal-mode frequency calculations for global minimum (GM) geometry, and two TS geometries to examine the fidelity of this PES. These two TSs are TS-I in which the torsional angle of one methyl rotor is shifted by 60° compared to GM, and the H-transfer saddle point is denoted as TS(H). The structure of these geometries is shown in Fig. 1. We obtain the symmetric double well H-transfer barrier as 1103 cm^{-1} or 3.15 kcal/mol, whereas the CCSD(F12*)(T+)/CBS(T,Q) value is

1124 cm^{-1} or 3.2 kcal/mol, which is an excellent agreement. Note that it is a significant improvement over the MP2-based PES,[74] which leads to a barrier height of 745 cm^{-1} or 2.13 kcal/mol. Additionally, we observe that the TS-I is only 162 cm^{-1} or 0.46 kcal/mol higher in energy compared to the global minimum.

Next, to examine the vibrational frequency predictions of the PES, we perform normal-mode analyses for the global minimum and H-transfer saddle point geometries. A comparison of harmonic mode frequencies for these two structures is shown in Table 2 along with the previously reported results from MP2 PES (V_{LL}) and local CCSD(T)-F12 calculations.[43] For most of the modes in Table 2, the differences between V_{LL} and $V_{LL\rightarrow CC}$ frequencies are small. Nevertheless, for high-frequency modes, this difference is more significant, especially for mode 32 of the global minimum and the imaginary-frequency mode of the TS(H). It is a major improvement over the MP2-based PES.

Another important aspect of this PES is the symmetric double well potential associated with the H-transfer process, which is essential for calculating the ground state tunneling splitting. In this context, we applied an approximate 1D approach to obtain the tunneling splittings following Ref. 86 (see Fig. 6).

In short, we derived a 1D potential, denoted $V(Q_{im})$, which represents the minimum energy pathway as a function of the imaginary-frequency mode (Q_{im}) corresponding to the hydrogen transfer saddle point geometry. This was achieved by optimizing on the $V_{LL\rightarrow CC}$ PES all other coordinates while keeping Q_{im} values fixed, except the methyl rotors, which cannot be described using rectilinear normal coordinates. These are held fixed at the saddle point values all the way along the path. Mostly due to the fixed methyl orientation and partly due to the fitting error, the barrier height of this 1D Q_{im} path is 940 cm^{-1} or 2.69 kcal/mol and it is 147 cm^{-1} lower than the direct FNO-CCSD(T) value or 163 cm^{-1} lower than the barrier on the $V_{LL\rightarrow CC}$ surface. This mass-scaled 1D potential is shown in Fig. 6. This approach is sufficient to test the fidelity of the fitting, while we note that a more-than-1D representation would be required for more accurate barriers and tunneling splitting

Table 2: Harmonic frequencies (in cm^{-1}) of the global minimum and H-transfer saddle point of acetylacetone obtained from the local CCSD(T)-F12, $V_{LL\rightarrow CC}$, and V_{LL} PESs .

	GM			TS(H)		
mode	local CCSD(T)[a]	$V_{LL\rightarrow CC}$	V_{LL}[a]	local CCSD(T)[a]	$V_{LL\rightarrow CC}$	V_{LL}[a]
1	113	97	97	$1278i$	$1044i$	$921i$
2	133	119	119	100	56	53
3	169	155	153	121	61	57
4	197	188	189	165	158	156
5	236	223	229	198	196	198
6	372	360	364	289	281	285
7	392	383	390	412	410	417
8	505	498	507	537	527	531
9	554	564	567	540	535	539
10	643	652	650	578	581	580
11	654	652	656	661	646	645
12	793	795	803	767	739	740
13	919	880	921	781	744	756
14	942	907	936	949	932	953
15	951	926	942	992	977	979
16	1010	1002	1014	1035	1026	1039
17	1040	1037	1048	1037	1032	1043
18	1050	1048	1058	1054	1054	1060
19	1072	1064	1071	1067	1055	1062
20	1192	1187	1200	1195	1175	1189
21	1276	1264	1290	1308	1229	1223
22	1393	1371	1384	1341	1332	1347
23	1405	1391	1399	1406	1397	1409
24	1424	1423	1433	1413	1410	1422
25	1462	1442	1470	1481	1483	1496
26	1480	1481	1494	1487	1483	1496
27	1483	1483	1497	1488	1485	1500
28	1488	1493	1505	1491	1487	1502
29	1502	1500	1512	1569	1543	1567
30	1670	1647	1655	1613	1609	1629
31	1709	1695	1704	1624	1634	1670
32	3047	3019	2855	1904	1709	1685
33	3052	3060	3095	3054	3019	3098
34	3118	3069	3099	3057	3061	3099
35	3122	3156	3178	3130	3163	3190
36	3157	3162	3187	3132	3163	3190
37	3165	3185	3208	3154	3181	3207
38	3220	3193	3218	3156	3182	3208
39	3257	3235	3258	3241	3255	3282

[a] From Ref. 43

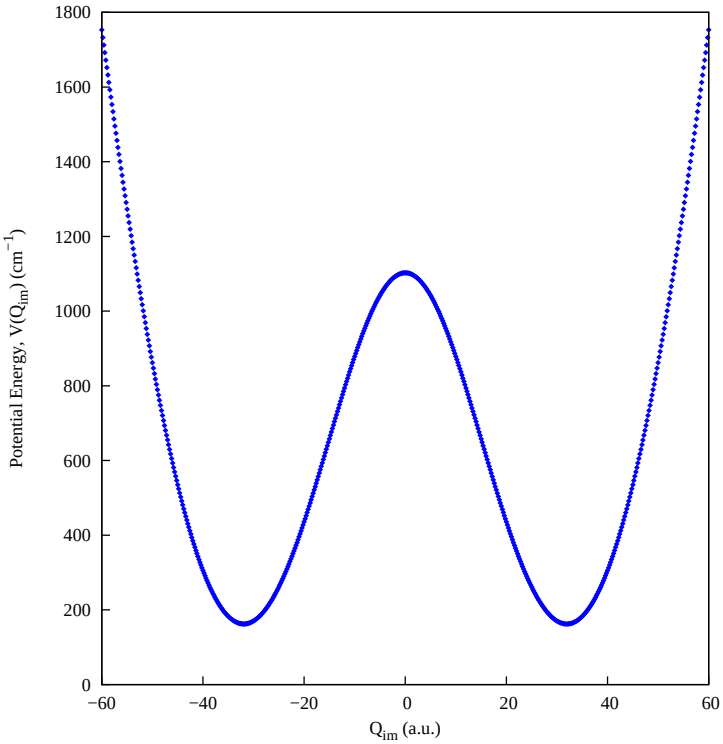

Figure 6: One-dimensional $V(Q_{im})$ path for H transfer in AcAc.

values.

The splittings are obtained through straightforward 1D-DVR calculations,[87] which involve computing the energies of the ground and first excited vibrational states along the $V(Q_{im})$ pathway. It is worth noting that the same method was employed to calculate tunneling splittings for H-atom and D-atom transfer in malonaldehyde.[86] Remarkably, there the 1D-DVR outcomes consistently demonstrated a close correspondence with the rigorous full-dimensional diffusion Monte Carlo splittings, typically falling within a 10 percent margin of agreement (as that 1D potential could represent the barrier from the global minima due to the simpler structure of malonaldehyde). Employing the 1D approach here, we obtain the ground state tunneling splitting of the H-transfer process as 71 cm^{-1} for this barrier height of 940 cm^{-1}, which is a significant improvement compared to the MP2 PES result.[74] In contrast, a previously reported tunneling splitting value of 32 cm^{-1} was obtained with the associate barrier height of 1234 cm^{-1} through rigorous DMC calculation using the local CCSD(T)-F12/cc-pVTZ-F12 PES by Chen *et al.*[43] Given the well-established fact on the

sensitivity of tunneling splitting to the barrier height, we aim to directly compare 1D-DVR tunneling splittings across different acetylacetone PESs, each associated with distinct 1D $V(Q_{im})$ barrier heights (see Table 3).

Table 3: Ground-state tunneling splittings of acetylacetone obtained from 1D $V(Q_{im})$ (with and without morphing) potential with indicative barrier height.

PES	Barrier (cm^{-1})	Splitting (cm^{-1})
MP2 (without morphing)[a]	586	141
MP2 (morphing)[a]	745	104
local CCSD(T) (without morphing)[b]	1055	48
local CCSD(T) (morphing)[c]	1234	38
$V_{LL\rightarrow CC}$ (without morphing)	940	71

[a] From Ref. 74.
[b] From private communication.
[c] From Ref. 43.

For instance, employing the full-dimensional PES at the MP2 level, the 1D-DVR splitting for H-transfer was found to be 141 cm^{-1} with a barrier height of 586 cm^{-1}, which transforms to 104 cm^{-1} after morphing[74] the 1D $V(Q_{im})$ potential with a barrier height of 745 cm^{-1}. Similarly, the tunneling splitting is estimated as 48 cm^{-1} with a barrier height of 1055 cm^{-1} from the local CCSD(T)-F12 PES,[43] and after morphing the 1D potential, it becomes 38 cm^{-1} with a corresponding barrier height of 1234 cm^{-1}.[43]

This analysis highlights the sensitivity of the tunneling splitting value to the barrier height and other PES parameters. While higher-dimensional computations will be the subject of a forthcoming study, our primary objective here was to present a full-dimensional acetylacetone PES using the cost-effective and highly accurate FNO-CCSD(T) data and benchmark its fidelity on various structural and dynamic properties, including the symmetric double well H-transfer motion.

6 Conclusions and outlook

Here, we present the first results with the highly promising combination of accelerated CCSD(T) and advanced Δ-machine learning methods to extend the reach of gold standard-level global PES generation for medium-sized molecules. We show that around a factor of 40 cost-reduction can be achieved by compressing the orbital space used for the CCSD(T) computations via the first application of the frozen natural orbital (FNO) approach in this context. Detailed benchmarks for the 15-atom acetylacetone molecule show that for key, lower-energy configurations, such as the first transition state, the FNO errors are negligible being in the 0.01 kcal/mol or ca. 5 cm^{-1} range. These uncertainties stay around 0.15 (0.3) kcal/mol or ca. 50 (100) cm^{-1} on the average (at the worst cases) even for the very high energy configurations up to ca. 100 kcal/mol or 35000 cm^{-1}, which are required for quantum nuclear computations. This is achieved via a cautious relaxation of usually highly-conservative FNO settings, which can be systematically improved to the conventional CCSD(T) result.

The Δ-ML approach brings significant additional cost-reduction, as it enables the correction of a low-level, cost-effective PES to the CCSD(T) level via an order of magnitude less CCSD(T)-level data points. By also utilizing the permutationally invariant polynomial (PIP) approach tailor-made for PES representation, hundreds (instead of tens of thousands) of CCSD(T) computations were sufficient here for the 15-atom acetylacetone. The combination of FNO method with Mrcc's highly-optimized CCSD(T) implementation, which is developed specifically for such reduced-cost applications, enables FNO-CCSD(T)/aug-cc-pVTZ computations for such 15-atom molecules within ca. 5 minutes on 16 cores. All in all, the FNO-CCSD(T) Δ-ML PES construction exhibited an unprecedented efficiency, enabling the required 430 FNO-CCSD(T)-level computations on a single, relatively accessible 128-core AMD node in around 6 hours. Since the FNO-CCSD(T)-based data set generation is independent from the parameters of the PES representation, any other efficient descriptor and/or ML method combination can similarly benefit from this advancement. The obtained Δ-ML

PES is representative of the expectable accuracy also with other ML methods. It shows high fidelity from multiple perspectives, including energetic, structural, and (an)harmonic vibrational properties, such as the tunneling splitting corresponding to the symmetric double well H-transfer barrier of acetylacetone. In a forthcoming work, this study will be extended to additional CCSD(T) cost-reduction approaches also at the basis set limit, as well as to a broader set of molecules and quantum nuclear properties.

Acknowledgement

The authors thank Prof. Joel Bowman (Emory University) for helpful comments on the manuscript. Financial support from the ERC Starting Grant No. 101076972, "aCCuracy", the National Research, Development, and Innovation Office (NKFIH, Grant No. FK142489 and KKP126451), the János Bolyai Research Scholarship of the Hungarian Academy of Sciences, ÚNKP-23-5-BME-408 New National Excellence Program of the Ministry for Culture and Innovation sourced from the NKFIH fund, and the computing time granted on the Hungarian HPC Infrastructure at NIIF Institute, Hungary are gratefully acknowledged. The research reported in this paper is part of project BME-EGA-02, implemented with the support provided by the Ministry of Innovation and Technology of Hungary from the National Research, Development and Innovation Fund, financed under the TKP2021 funding scheme.

References

(1) Braams, B. J.; Bowman, J. M. Permutationally invariant potential energy surfaces in high dimensionality. *Int. Rev. Phys. Chem.* **2009**, *28*, 577–606.

(2) Qu, C.; Yu, Q.; Bowman, J. M. Permutationally invariant potential energy surfaces. *Annu. Rev. Phys. Chem.* **2018**, *69*, 6.1–6.25.

(3) Behler, J. Neural network potential-energy surfaces in chemistry: a tool for large-scale simulations. *Phys. Chem. Chem. Phys.* **2011**, *13*, 17930–17955.

(4) Behler, J. Constructing high-dimensional neural network potentials: A tutorial review. *Int. J. Quantum Chem.* **2015**, *115*, 1032–1050.

(5) Bartók, A. P.; Csányi, G. Gaussian approximation potentials: A brief tutorial introduction. *Int. J. Quantum Chem.* **2015**, *115*, 1051–1057.

(6) Cui, J.; Krems, R. V. Efficient non-parametric fitting of potential energy surfaces for polyatomic molecules with Gaussian processes. *J. Phys. B: At. Mol. Opt. Phys.* **2016**, *49*.

(7) Jiang, B.; Guo, H. Permutation invariant polynomial neural network approach to fitting potential energy surfaces. *J. Chem. Phys.* **2013**, *139*, 054112.

(8) Jiang, B.; Li, J.; Guo, H. Potential energy surfaces from high fidelity fitting of ab initio points: the permutation invariant polynomial - neural network approach. *Int. Rev. Phys. Chem.* **2016**, *35*, 479–506.

(9) Qu, C.; Yu, Q.; Van Hoozen, B. L.; Bowman, J. M.; Vargas-Hernández, R. A. Assessing Gaussian Process Regression and Permutationally Invariant Polynomial Approaches To Represent High-Dimensional Potential Energy Surfaces. *J. Chem. Theory Comput.* **2018**, *14*, 3381–3396.

(10) Dral, P. O.; Owens, A.; Dral, A.; Csányi, G. Hierarchical machine learning of potential energy surfaces. *J. Chem. Phys.* **2020**, *152*, 204110.

(11) Hou, Y.-F.; Ge, F.; Dral, P. O. Explicit Learning of Derivatives with the KREG and pKREG Models on the Example of Accurate Representation of Molecular Potential Energy Surfaces. *J. Chem. Theory Comput.* **2023**, *19*, 2369.

(12) Crawford, T. D.; Schaefer III, H. F. An Introduction to Coupled Cluster Theory for Computational Chemists. *Rev. Comp. Chem.* **1999**, *14*, 33.

(13) Shavitt, I.; Bartlett, R. J. *Many-Body Methods in Chemistry and Physics: MBPT and Coupled-Cluster Theory*; Cambridge Molecular Science; Cambridge University Press, 2009.

(14) Bartlett, R. J.; Musiał, M. Coupled-cluster theory in quantum chemistry. *Rev. Mod. Phys.* **2007**, *79*, 291.

(15) Raghavachari, K.; Trucks, G. W.; Pople, J. A.; Head-Gordon, M. A fifth-order perturbation comparison of electron correlation theories. *Chem. Phys. Lett.* **1989**, *157*, 479.

(16) Deegan, M. J. O.; Knowles, P. J. Perturbative corrections to account for triple excitations in closed and open shell coupled cluster theories. *Chem. Phys. Lett.* **1994**, *227*, 321.

(17) Kobayashi, R.; Rendell, A. P. A direct coupled cluster algorithm for massively parallel computers. *Chem. Phys. Lett.* **1997**, *265*, 1.

(18) Pitoňák, M.; Aquilante, F.; Hobza, P.; Neogrády, P.; Noga, J.; Urban, M. Parallelized implementation of the CCSD(T) method in MOLCAS using optimized virtual orbitals space and Cholesky decomposed two-electron integrals. *Collect. Czech. Chem. Commun.* **2011**, *76*, 713.

(19) Asadchev, A.; Gordon, M. S. Fast and Flexible Coupled Cluster Implementation. *J. Chem. Theory Comput.* **2013**, *9*, 3385.

(20) Deumens, E.; Lotrich, V. F.; Perera, A.; Ponton, M. J.; Sanders, B. A.; Bartlett, R. J. Software design of ACES III with the super instruction architecture. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2011**, *1*, 895.

(21) Kaliman, I. A.; Krylov, A. I. New algorithm for tensor contractions on multi-core CPUs, GPUs, and accelerators enables CCSD and EOM-CCSD calculations with over 1000 basis functions on a single compute node. *J. Comput. Chem.* **2017**, *38*, 842.

(22) Janowski, T.; Pulay, P. Efficient Parallel Implementation of the CCSD External Exchange Operator and the Perturbative Triples (T) Energy Calculation. *J. Chem. Theory Comput.* **2008**, *4*, 1585.

(23) Solomonik, E.; Matthews, D.; Hammond, J. R.; Stanton, J. F.; Demmel, J. A massively parallel tensor contraction framework for coupled-cluster computations. *J. Parallel Distr. Comput.* **2014**, *74*, 3176.

(24) Peng, C.; Calvin, J. A.; Valeev, E. F. Coupled-cluster singles, doubles and perturbative triples with density fitting approximation for massively parallel heterogeneous platforms. *Int. J. Quantum Chem.* **2019**, *119*, e25894.

(25) Shen, T.; Zhu, Z.; Zhang, I. Y.; Scheffler, M. Massive-parallel Implementation of the Resolution-of-Identity Coupled-cluster Approaches in the Numeric Atom-centered Orbital Framework for Molecular Systems. *J. Chem. Theory Comput.* **2019**, *15*, 4721.

(26) Gyevi-Nagy, L.; Kállay, M.; Nagy, P. R. Integral-direct and parallel implementation of the CCSD(T) method: Algorithmic developments and large-scale applications. *J. Chem. Theory Comput.* **2020**, *16*, 366.

(27) Kutzelnigg, W.; Klopper, W. Wave functions with terms linear in the interelectronic coordinates to take care of the correlation cusp. I. General theory. *J. Chem. Phys.* **1991**, *94*, 1985.

(28) Klopper, W.; Manby, F. R.; S.Ten-no,; Valeev, E. F. R12 methods in explicitly correlated molecular electronic structure theory. *Int. Rev. Phys. Chem.* **2006**, *25*, 427.

(29) Hättig, C.; Klopper, W.; Köhn, A.; Tew, D. P. Explicitly Correlated Electrons in Molecules. *Chem. Rev.* **2012**, *112*, 4.

(30) Hättig, C.; Tew, D. P.; Köhn, A. Accurate and efficient approximations to explicitly correlated coupled-cluster singles and doubles, CCSD-F12. *J. Chem. Phys.* **2010**, *132*, 231102.

(31) Kállay, M.; Horváth, R. A.; Gyevi-Nagy, L.; Nagy, P. R. Size-consistent explicitly correlated triple excitation correction. *J. Chem. Phys.* **2021**, *155*, 034107.

(32) Kállay, M.; Horváth, R. A.; Gyevi-Nagy, L.; Nagy, P. R. Basis set limit CCSD(T) energies for extended molecules via a reduced-cost explicitly correlated approach. *J. Chem. Theory Comput.* **2023**, *19*, 174.

(33) Löwdin, P.-O. Quantum theory of many-particle systems. I. Physical interpretations by means of density matrices, natural spin-orbitals, and convergence problems in the method of configurational interaction. *Phys. Rev.* **1955**, *97*, 1474.

(34) Taube, A. G.; Bartlett, R. J. Frozen natural orbitals: Systematic basis set truncation for coupled-cluster theory. *Collect. Czech. Chem. Commun.* **2005**, *70*, 837.

(35) DePrince, A. E.; Sherrill, C. D. Accurate Noncovalent Interaction Energies Using Truncated Basis Sets Based on Frozen Natural Orbitals. *J. Chem. Theory Comput.* **2013**, *9*, 293.

(36) Gyevi-Nagy, L.; Kállay, M.; Nagy, P. R. Accurate reduced-cost CCSD(T) energies: parallel implementation, benchmarks, and large-scale applications. *J. Chem. Theory Comput.* **2021**, *17*, 860.

(37) Nagy, P. R.; Gyevi-Nagy, L.; Kállay, M. Basis set truncation corrections for improved frozen natural orbital CCSD(T) energies. *Mol. Phys.* **2021**, *119*, e1963495.

(38) Chmiela, S.; Sauceda, H. E.; Müller, K.-R.; Tkatchenko, A. Towards exact molecular dynamics simulations with machine-learned force fields. *Nat. Commun.* **2018**, *9*, 3887.

(39) Sauceda, H. E.; Vassilev-Galindo, V.; Chmiela, S.; Müller, K.-R.; Tkatchenko, A. Dynamical strengthening of covalent and non-covalent molecular interactions by nuclear quantum effects at finite temperature. *Nat. Commun.* **2021**, *12*, 442.

(40) Smith, J. S.; Nebgen, B. T.; Zubatyuk, R.; Lubbers, N.; Devereux, C.; Barros, K.; Tretiak, S.; Isayev, O.; Roitberg, A. E. Approaching coupled cluster accuracy with

a general-purpose neural network potential through transfer learning. *Nat. Commun.* **2019**, *10*, 2903.

(41) Käser, S.; Unke, O. T.; Meuwly, M. Reactive dynamics and spectroscopy of hydrogen transfer from neural network-based reactive potential energy surfaces. *New J. Phys.* **2020**, *22*, 055002.

(42) Young, T. A.; Johnston-Wood, T.; Deringer, V. L.; Duarte, F. A transferable active-learning strategy for reactive molecular force fields. *Chem. Sci.* **2021**, *12*, 10944.

(43) Qu, C.; Houston, P. L.; Conte, R.; Nandi, A.; Bowman, J. M. Breaking the Coupled Cluster Barrier for Machine-Learned Potentials of Large Molecules: The Case of 15-Atom Acetylacetone. *J. Phys. Chem. Lett.* **2021**, *12*, 4902.

(44) Khire, S. S.; Gurav, N. D.; Nandi, A.; Gadre, S. R. Enabling Rapid and Accurate Construction of CCSD(T)-Level Potential Energy Surface of Large Molecules Using Molecular Tailoring Approach. *J. Phys. Chem. A* **2022**, *126*, 1458–1464.

(45) Jensen, A. B.; Kubečka, J.; Schmitz, G.; Christiansen, O.; Elm, J. Massive Assessment of the Binding Energies of Atmospheric Molecular Clusters. *Journal of Chemical Theory and Computation* **2022**, *18*, 7373–7383.

(46) Daru, J.; Forbert, H.; Behler, J.; Marx, D. Coupled Cluster Molecular Dynamics of Condensed Phase Systems Enabled by Machine Learning Potentials: Liquid Water Benchmark. *Phys. Rev. Lett.* **2022**, *129*, 226001.

(47) Yu, Q.; Qu, C.; Houston, P. L.; Conte, R.; Nandi, A.; Bowman, J. M. q-AQUA: A Many-Body CCSD(T) Water Potential, Including Four-Body Interactions, Demonstrates the Quantum Nature of Water from Clusters to the Liquid Phase. *J. Phys. Chem. Lett.* **2022**, *13*, 5068–5074.

(48) Nandi, A.; Qu, C.; Houston, P. L.; Conte, R.; Yu, Q.; Bowman, J. M. A CCSD(T)-Based 4-Body Potential for Water. *J. Phys. Chem. Lett.* **2021**, *12*, 10318–10324.

(49) Fu, Y.-L.; Lu, X.; Han, Y.-C.; Fu, B.; Zhang, D. H.; Bowman, J. M. Collision-induced and complex-mediated roaming dynamics in the H + C_2H_4 → H_2 + C_2H_3 reaction. *Chem. Sci.* **2020**, *11*, 2148–2154.

(50) Lu, D.; Behler, J.; Li, J. Accurate Global Potential Energy Surfaces for the H + CH_3OH Reaction by Neural Network Fitting with Permutation Invariance. *J. Phys. Chem. A* **2020**, *124*, 5737–5745.

(51) Papp, D.; Czakó, G. Rotational Mode-Specificity in the Cl + C_2H_6 → HCl + C_2H_5 Reaction. *J. Phys. Chem. A* **2022**, *126*, 2551–2560.

(52) Gruber, B.; Tajti, V.; Czakó, G. Full-dimensional automated potential energy surface development and dynamics for the OH + C_2H_6 reaction. *J. Chem. Phys.* **2022**, *157*, 074307.

(53) Qu, C.; Bowman, J. M. An ab initio potential energy surface for the formic acid dimer: zero-point energy, selected anharmonic fundamental energies, and ground-state tunneling splitting calculated in relaxed 1–4-mode subspaces. *Phys. Chem. Chem. Phys.* **2016**, *18*, 24835–24840.

(54) Guo, Y.; Riplinger, C.; Becker, U.; Liakos, D. G.; Minenkov, Y.; Cavallo, L.; Neese, F. Communication: An improved linear scaling perturbative triples correction for the domain based local pair-natural orbital based singles and doubles coupled cluster method [DLPNO-CCSD(T)]. *J. Chem. Phys.* **2018**, *148*, 011101.

(55) Ma, Q.; Werner, H.-J. Explicitly correlated local coupled-cluster methods using pair natural orbitals. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2018**, *8*, e1371.

(56) Nagy, P. R.; Samu, G.; Kállay, M. Optimization of the linear-scaling local natural orbital CCSD(T) method: Improved algorithm and benchmark applications. *J. Chem. Theory Comput.* **2018**, *14*, 4193.

(57) Nagy, P. R.; Kállay, M. Approaching the basis set limit of CCSD(T) energies for large

molecules with local natural orbital coupled-cluster methods. *J. Chem. Theory Comput.* **2019**, *15*, 5275.

(58) Nandi, A.; Laude, G.; Khire, S. S.; Gurav, N. D.; Qu, C.; Conte, R.; Yu, Q.; Li, S.; Houston, P. L.; Gadre, S. R.; Richardson, J. O.; Evangelista, F. A.; Bowman, J. M. Ring-Polymer Instanton Tunneling Splittings of Tropolone and Isotopomers using a Δ-Machine Learned CCSD(T) Potential: Theory and Experiment Shake Hands. *J. Am. Chem. Soc.* **2023**, *145*, 9655–9664.

(59) Kállay, M.; Nagy, P. R.; Mester, D.; Rolik, Z.; Samu, G.; Csontos, J.; Csóka, J.; Szabó, P. B.; Gyevi-Nagy, L.; Hégely, B.; Ladjánszki, I.; Szegedy, L.; Ladóczki, B.; Petrov, K.; Farkas, M.; Mezei, P. D.; Ganyecz, Á. The MRCC program system: Accurate quantum chemistry from water to proteins. *J. Chem. Phys.* **2020**, *152*, 074107.

(60) Kállay, M.; Nagy, P. R.; Mester, D.; Gyevi-Nagy, L.; Csóka, J.; Szabó, P. B.; Rolik, Z.; Samu, G.; Csontos, J.; Hégely, B.; Ganyecz, Á.; Ladjánszki, I.; Szegedy, L.; Ladóczki, B.; Petrov, K.; Farkas, M.; Mezei, P. D.; Horváth, R. A. MRCC, *a quantum chemical program suite. See https://www.mrcc.hu/* **Accessed August 1, 2023**,

(61) Nagy, P. R.; Gyevi-Nagy, L.; Lőrincz, B. D.; Kállay, M. Pursuing the basis set limit of CCSD(T) non-covalent interaction energies for medium-sized complexes: case study on the S66 compilation. *Mol. Phys.* **2023**, *121*, e2109526.

(62) DePrince, A. E.; Sherrill, C. D. Accuracy and Efficiency of Coupled-Cluster Theory Using Density Fitting/Cholesky Decomposition, Frozen Natural Orbitals, and a t_1-Transformed Hamiltonian. *J. Chem. Theory Comput.* **2013**, *9*, 2687.

(63) Nagy, P. R.; Kállay, M. Optimization of the linear-scaling local natural orbital CCSD(T) method: Redundancy-free triples correction using Laplace transform. *J. Chem. Phys.* **2017**, *146*, 214106.

(64) Lee, T. J.; Rendell, A. P.; Taylor, P. R. Comparison of the quadratic configuration

interaction and coupled-cluster approaches to electron correlation including the effect of triple excitations. *J. Phys. Chem.* **1990**, *94*, 5463.

(65) Kállay, M. A systematic way for the cost reduction of density fitting methods. *J. Chem. Phys.* **2014**, *141*, 244113.

(66) Liu, Y.; Li, J. Permutation-invariant-polynomial neural-network-based Δ-machine learning approach: A case for the HO_2 self-reaction and its dynamics study. *J. Phys. Chem. Letts.* **2022**, *13*, 4729–4738.

(67) Song, K.; Li, J. The neural network based Δ-machine learning approach efficiently brings the DFT potential energy surface to the CCSD(T) quality: a case for the OH + CH_3OH reaction. *Phys. Chem. Chem. Phys.* **2023**, *25*, 11192–11204.

(68) Li, J.; Li, J. An accurate full-dimensional interaction potential energy surface of $CO_2{}^+N_2$ incorporating Δ-machine learning approach via permutation invariant polynomial-neural network. *Artificial Intelligence Chemistry* **2023**, *1*, 100019.

(69) Nandi, A.; Qu, C.; Houston, P. L.; Conte, R.; Bowman, J. M. Δ-machine learning for potential energy surfaces: A PIP approach to bring a DFT-based PES to CCSD(T) level of theory. *J. Chem. Phys.* **2021**, *154*, 051102.

(70) Bowman, J. M.; Qu, C.; Conte, R.; Nandi, A.; Houston, P. L.; Yu, Q. Δ-Machine Learned Potential Energy Surfaces and Force Fields. *J. Chem. Theory and Comput.* **2023**, *19*, 1–17.

(71) Bowman, J. M.; Braams, B. J.; Carter, S.; Chen, C.; Czakó, G.; Fu, B.; Huang, X.; Kamarchik, E.; Sharma, A. R.; Shepler, B. C.; Wang, Y.; Xie, Z. Ab-initio-based potential energy surfaces for complex molecules and molecular complexes. *J. Phys. Chem. Lett.* **2010**, *1*, 1866–1874.

(72) Xie, Z.; Bowman, J. M. Permutationally Invariant Polynomial Basis for Molecular Energy Surface Fitting via Monomial Symmetrization. *J. Chem. Theory Comput.* **2010**,

6, 26–34.

(73) Bowman, J. M.; Czakó, G.; Fu, B. High-dimensional ab initio potential energy surfaces for reaction dynamics calculations. *Phys. Chem. Chem. Phys.* **2011**, *13*, 8094–8111.

(74) Qu, C.; Conte, R.; Houston, P. L.; Bowman, J. M. Full-dimensional potential energy surface for acetylacetone and tunneling splittings. *Phys. Chem. Chem. Phys.* **2021**, *23*, 7758–7767.

(75) Nandi, A.; Qu, C.; Bowman, J. M. Using Gradients in Permutationally Invariant Polynomial Potential Fitting: A Demonstration for CH_4 Using as Few as 100 Configurations. *J. Chem. Theory Comput.* **2019**, *15*, 2826–2835.

(76) MSA Software with Gradients. `https://github.com/szquchen/MSA-2.0`, 2019; Accessed: 2019-01-20.

(77) Dunning Jr., T. H. Gaussian basis sets for use in correlated molecular calculations. I. The atoms boron through neon and hydrogen. *J. Chem. Phys.* **1989**, *90*, 1007.

(78) Kendall, R. A.; Dunning Jr., T. H.; Harrison, R. J. Electron affinities of the first-row atoms revisited. Systematic basis sets and wave functions. *J. Chem. Phys.* **1992**, *96*, 6796.

(79) Weigend, F. Hartree–Fock Exchange Fitting Basis Sets for H to Rn. *J. Comput. Chem.* **2008**, *29*, 167.

(80) Weigend, F.; Köhn, A.; Hättig, C. Efficient use of the correlation consistent basis sets in resolution of the identity MP2 calculations. *J. Chem. Phys.* **2002**, *116*, 3175.

(81) Peterson, K. A.; Adler, T. B.; Werner, H.-J. Systematically convergent basis sets for explicitly correlated wavefunctions: The atoms H, He, B-Ne, and Al-Ar. *J. Chem. Phys.* **2008**, *128*, 084102.

(82) Yousaf, K. E.; Peterson, K. A. Optimized auxiliary basis sets for explicitly correlated methods. *J. Chem. Phys.* **2008**, *129*, 184108.

(83) Yousaf, K. E.; Peterson, K. A. Optimized complementary auxiliary basis sets for explicitly correlated methods: aug-cc-pVnZ orbital basis sets. *Chem. Phys. Lett.* **2009**, *476*, 303.

(84) Karton, A.; Martin, J. M. L. Comment on: "Estimating the Hartree–Fock limit from finite basis set calculations". *Theor. Chem. Acc.* **2006**, *115*, 330.

(85) Helgaker, T.; Klopper, W.; Koch, H.; Noga, J. Basis-set convergence of correlated calculations on water. *J. Chem. Phys.* **1997**, *106*, 9639.

(86) Wang, Y.; Bowman, J. M. One-dimensional tunneling calculations in the imaginary-frequency, rectilinear saddle-point normal mode. *J. Chem. Phys.* **2008**, *129*, 121103.

(87) Colbert, D. T.; Miller, W. H. A novel discrete variable representation for quantum mechanical reactive scattering via the S-matrix Kohn method. *J. Chem. Phys.* **1992**, *96*, 1982–1991.