30-01-2024

Targeting Real chemical accuracy at the EXascale

D5.4 – Datasets made available for benchmarking and ML modelling

Version 1.0

GA no 952165

Dissemination Level

☒ PU: Public
□ PP: Restricted to other programme participants (including the Commission)
□ RE: Restricted to a group specified by the consortium (including the Commission)
□ CO: Confidential, only for members of the consortium (including the Commission)

Document Information

Project Title	Targeting Real Chemical accuracy at the EXascale
Project Acronym	TREX
Grant Agreement No	952165
Instrument	Call: H2020-INFRAEDI-2019-1
Topic	INFRAEDI-05-2020 Centres of Excellence in EXascale computing
Start Date of Project	01-10-2020
Duration of Project	36 Months
Project Website	https://trex-coe.eu/

Deliverable Number	D5.4
Deliverable title	D5.4 – Datasets made available for benchmarking and ML modelling
Due Date	M40 – 31-01-2024 (from GA)
Actual Submission Date	30-01-2024

Work Package	WP5 - Demonstrations
Lead Author (Org)	Kasia Pernal (TUL)
Contributing Author(s) (Org)	Michele Casula (CNRS), Matthias Rupp (LIST), Michal Hapka (TUL)
Reviewers (Org)	Mariella Ippolito (CINECA), Axel Auweter (Megware), Jan Beerens (UT)
Version	1.0

Dissemination level	PU
Nature	Report
Draft / final	Final
No. of pages including cover	20

TREX: Targeting Real Chemical Accuracy at the Exascale project has received funding from the European Union Horizon 2020 research and innovation programme under Grant Agreement No. 952165.

Disclaimer

TREX: Targeting Real Chemical Accuracy at the Exascale project has received funding from the European Union Horizon 2020 research and innovation program under Grant Agreement No. 952165

The content of this document does not represent the opinion of the European Union, and the European Union is not responsible for any use that might be made of such content.

Versioning and contribution history

Version	Date	Authors	Notes
1.0	30-01-2024	Kasia Pernal (TUL)	First Official Release

Abbreviations

CAS	Complete Active Space
CBS	Complete Basis Set
CC	Coupled Cluster
DFT	Density Functional Theory
LRDMC	Lattice Regularized Diffusion Monte Carlo
MD	Molecular Dynamics
ML	Machine Learning
MLP	Machine-Learning Potential
nCP	Non-Casimir-Polder terms
NVT	Number of Particles, Volume, Temperature
PES	Potential Energy Surface
PIMD	Path Integral Molecular Dynamics
QMC	Quantum Monte Carlo
SAPT	Symmetry-Adapted Perturbation Theory
VMC	Variational Monte Carlo
WP	Work Package

Table of Contents

Contents

Table of Figures

TREX: Targeting Real Chemical Accuracy at the Exascale project has received funding from the European Union Horizon 2020 research and innovation programme under Grant Agreement No. 952165.

Table of Tables

TREX: Targeting Real Chemical Accuracy at the Exascale project has received funding from the European Union Horizon 2020 research and innovation programme under Grant Agreement No. 952165.

Executive summary

This report documents the creation of four datasets for modelling and benchmarking computational methods. They are dedicated to further investigations of systems, which have been used for demonstrations in work package (WP) 5, i.e. hydrogen under pressure, protonated water hexamer and molecular interactions in excited-state organic dimers. The datasets have been developed in the groups of Michele Casula (CNRS), Matthias Rupp (LIST), Kasia Pernal (TUL), and Michal Hapka (University of Warsaw). The description of each dataset presented in the report includes: a motivation for its creation, description of computational protocols used to generate data in the set, description of data included in the set, and presentation of the data structure.

1 Introduction

Molecular dynamics simulation methods are one of the main tools for computational chemistry and physics. Their scope of applications is still limited by the computational costs and accuracy. The former bottleneck is largely mitigated if machine learning potentials (MLP), which are orders of magnitudes faster to evaluate than the ab initio potentials, are employed. To assure sufficient accuracy in simulations ML models must be trained on benchmarking datasets. Data for the latter are obtained from techniques that are more accurate, yet of higher computational complexity than ML methods: Quantum Monte Carlo (QMC) or Density Functional Theory (DFT). There is thus a need for creating datasets for systems of interest for ML modelling.

In WP5 of the TREX project we focus on applications of QMC and *ab initio* methods for systems of importance for technological progress and relevance to energy storage and conversion: hydrogen under pressure, clusters of water molecules, and molecular interactions in electronically excited complexes. Applicability of machine learning methods or empirically parameterized low-computational-cost density functionals to such systems would allow longer simulations to be performed for larger systems and obtaining results for more realistic systems. To make this possible, datasets for MLPs and benchmarking must be available.

The purpose of this document is to report on creation of datasets for ML training and MLP benchmarking for molecular potentials of hydrogen under pressure (section 2), protonated water hexamer (section 3), and for molecular interaction energies for organic dimers in excited states (section 4). The accurate data has been obtained from QMC, DFT, and Symmetry-Adapted perturbation theory (SAPT) methods. In section 5 all datasets are listed, and their content and purposes are briefly summarized.

2 Hydrogen under pressure

Despite its simplicity, hydrogen in condensed phases shows rich phase diagrams as a function of pressure and temperature, which has challenged our physical understanding. While the atomic unit is exactly solvable, many-body systems made by an infinite arrangement of hydrogen atoms are particularly hard to study, given the complexity of interatomic interactions, electronic correlation effects present in the macroscopic systems, and nuclear quantum effects due to the light mass of the hydrogen nuclei. Collective excitations are supposed to lead to exotic phenomena, such as high-temperature superconductivity, superfluidity, liquid-liquid phase transitions. Therefore, it is highly desirable to derive an accurate interatomic potential, possibly by machine learning methods, based on accurate QMC data, to accelerate the understanding of the proposed phases and the discovery of new ones, at extreme conditions.

2.1 Geometries, energies, forces, and stresses at DFT level of theory

Molecular dynamics simulations of atomistic systems are a cornerstone of computational physics, chemistry, and materials science, but are limited by either accuracy or computational cost. This trade-off is dominated by the method used to compute the potential energy surface whose gradient determines the forces that propagate atoms in the simulation. Typical choices include classical force fields, characterized by fixed functional forms, and quantum-mechanical (ab initio) approaches such as density functional theory (DFT) and quantum Monte Carlo (QMC) methods.

Force fields are efficient but limited in accuracy, transferability, and by parametrization effort. Ab initio approaches are accurate and transferable but have high computational costs. Consequently, dynamics simulations are limited in either the system sizes and time scales or the phenomena (e.g., bond breaking and formation) that can be modelled.

Machine-learning interatomic potentials (MLPs) [1] are data-driven approximations of ab initio potential energy surfaces that exploit correlations between atom positions and the resulting potential energy. Essentially, a flexible functional form such as a neural network is parametrized on a training dataset consisting of ab initio calculations. The resulting model is then used to calculate the forces during the simulation. MLPs are typically several orders of magnitude faster to evaluate than the ab initio reference method. This greatly accelerates dynamics simulations and thus enables studying larger systems, longer time scales, and otherwise inaccessible phenomena for DFT. For a QMC reference, an MLP would often enable running such simulations at all.

In the TREX project's WPs 4 and 5, we have worked towards enabling MLPs trained on QMC reference data [2]. Much of this work is based on delta-learning [3], that is, machine-learning the difference between two ab initio reference methods, here DFT and QMC. We then showed that it is harder to learn the DFT baseline than it is to learn the difference between DFT and QMC, most likely due to the latter being smoother than the former. We also found that many state-of-the-art MLPs failed to learn the DFT hydrogen under pressure accurately enough to correctly simulate a liquid-liquid phase transition between atomic and molecular liquid hydrogen.

Consequently, we have created a dataset of hydrogen under (very high) pressure at the DFT level of theory to use as a benchmark for MLPs, the h-llpt-24 dataset. Motivation and details of this dataset are briefly described in the following. We start with the necessity for such a benchmarking dataset.

The aim of successful MLP molecular simulations is to emulate ab initio molecular dynamics. More specifically, the discrepancies in derived macroscopic quantities, such as radial distribution functions and diffusion coefficients, should be as small as possible. The MLP molecular dynamics should also reproduce monotony, limits, and asymptotic behaviour of such quantities. This is necessary to ensure a consistent description of physical phenomena such as proton exchange and phase transitions. Validation and performance assessment of MLPs should therefore base on molecular dynamics simulations.

In practice, MLP performance is evaluated as the average error in force predictions on a test set (data not used during training but from the same distribution as the training data), without involving any simulations. However, force accuracy on a test set is necessary but insufficient for successful molecular dynamics simulations. Consequently, MLPs with reported state-of-the-art test-set accuracy often fail catastrophically in actual simulations.

The inadequacy of test-set force errors as a performance indicator for molecular dynamics simulations is increasingly being recognized in the community. Uptake of superior assessment methods based on MLP molecular dynamics simulations, however, is slow, possibly because they require more human and computational effort as well as expert domain knowledge of the simulated system and the derived properties.

Therefore, the h-llpt-24 dataset is a benchmark for assessment of MLPs that is straightforward to use and does not require expert domain knowledge. After plugging in a trained MLP, provided scripts run molecular dynamics simulations, compute derived properties (including pressure curves, radial distribution functions, diffusion coefficients, and stable molecular fractions), analyze observed physical phenomena (here, a first-order liquid-liquid phase transition), and provide publication-ready figures, tables, and summary statistics.

For this challenging benchmark system, we provide geometries, lattice vectors, energies, forces, stresses, Wigner-Seitz radii, and temperatures for 8568 configurations of 128 hydrogen atoms each, using periodic boundary conditions. The configurations are decorrelated snapshots from 6 * 6 * 17 = 612 molecular dynamics reference simulations in the NVT ensemble at the DFT/PBE level of theory that cover the temperature and pressure regime of the liquid-liquid phase transition, six different temperatures and 17 different mass densities (Figure 1), with six replicas per temperature/density combination.

The data set is split into a training set of 7140 configurations and a test set of 1428 configurations. The test set configurations are always sampled from one of the six repetitions, that is, from different molecular dynamics simulations than the training data. Training and test data are thus sampled independently from the same distribution, supporting the assumption of machine-learning models that data are independent and identically distributed. To facilitate training of MLPs, we further provide a split of the training data into a proper training set and a validation set, following the same approach. The validation set can be used as a hold-out set for early stopping or hyper-parameter optimization, but its use is optional. The benchmark also provides Python scripts for automatically evaluating an MLP by running dynamics simulations with the MLP and extracting observables (see above).

The h-llpt-24 benchmark is still a work in progress but will be submitted before the end of the TREX project. Journal article and benchmark, including data and code, will be made freely available via open access and as open source. Preliminary information about the planned publication: Thomas Bischoff, Bastian Jäckl, Matthias Rupp: Hydrogen under Pressure as a Benchmark for Machine-Learning Potentials, 2024.

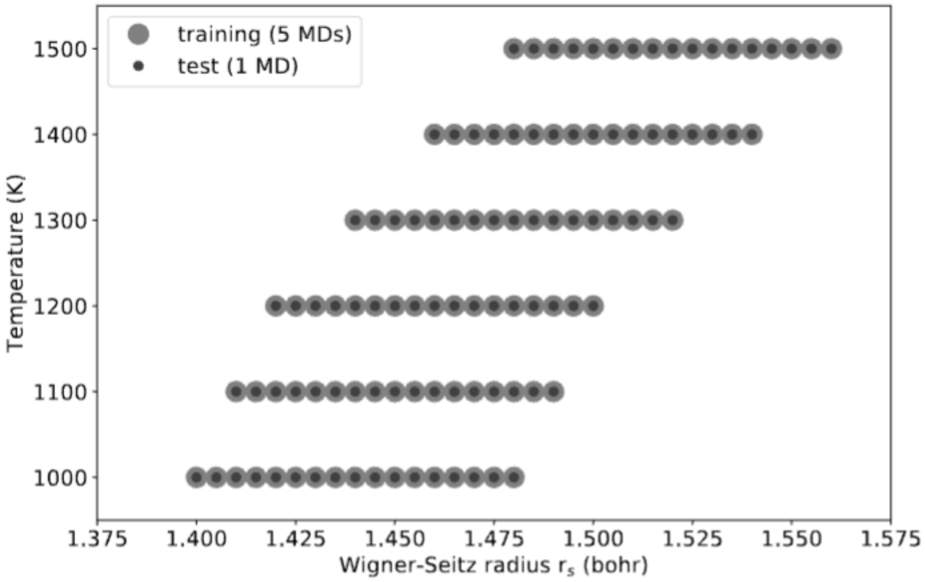

Figure 1: The h-llpt-24 data set overview. Each disk represents a combination of Wigner-Seitz radius and temperature for which six DFT molecular dynamics simulations were performed, five for training of the MLP and one for testing.

2.2 Configurations and energies at QMC level of theory

We provide here a database of crystalline hydrogen configurations, for which we determine the variational Monte Carlo (VMC) and the lattice regularized diffusion Monte Carlo (LRDMC) energies. The geometries are obtained by relaxing the internal coordinates at the given symmetry and at various volumes: 1.416 Å^3, 1.259 Å^3, 1.187 Å^3, 1.116 Å^3, 1.067 Å^3, values expressed per hydrogen atom. These volumes correspond approximatively to pressures: 350 GPa, 450 GPa, 500 GPa, 550 GPa and 650 GPa. The pressure versus volume relation upon which this correspondence is made is obtained for the C2/c-24 symmetry by using the BLYP functional. The correspondence is only approximate in the other cases (other functional or other symmetry). All calculations are performed at fixed volume, however it is useful to express these volumes using indicative pressures, to locate more easily the configurations in the p-T phase diagram.

Beside the C2/c-24 symmetry (molecular phase III), we took also into account the Cmca-12 (molecular phase VI), the Cmca-4 (molecular), the P62/c-24 (molecular), and the Cs-IV (atomic) symmetries. To each symmetry and volume (pressure), we can associate up to 3 different geometries, corresponding to the "classical harmonic" geometry obtained by relaxing the internal coordinates at the BLYP functional level keeping the nuclei classical, the "SSCHA" geometry obtained by relaxing the internal coordinates using the stochastic self-consistent harmonic approximation (SSCHA) treating the hydrogen nuclei as quantum particles, and finally the "deuterium" geometry obtained by the relaxation done with SSCHA for hydrogen nuclei replaced by deuterium (isotope substitution). On top of that, we provide the QMC energies for different supercell sizes, being the QMC calculations affected by finite-size effects. Thus, the full body of data contains several different hydrogen arrangements, possibly useful as training set to construct new hydrogen ML potentials or as benchmarking reference.

The variational Monte Carlo wave function used in our calculations has been extensively described in [2]. The energies reported in the database are yielded by twist averaging (analogous to k-point sampling in DFT), in order to reduce finite-size effects. The k-mesh used for the different structures giving converged DFT energies is reported in Table 1for the unit cell.

Table 1: k-mesh used for different crystalline hydrogen configurations

Symmetry	*k-mesh*
C2/c-24	12x12x6
P62/c-24	12x12x6
Cmca-12	12x12x12
Cs-IV-2	48x48x48
Cmca-4	36x24x24

For the supercells, the k-mesh is reduced for each lattice vector direction by a factor corresponding to the number of replicas taken in the supercell along the corresponding direction.

The lattice regularized diffusion Monte Carlo calculations are performed with a lattice space of 0.25 a_0. The energies are extrapolated with respect to the population size, but not with respect to the lattice space. Therefore, in the database, the LRDMC energies do not correspond to the converged fixed-node energies, but they are slightly underestimated, the lattice space extrapolation converging from below. However, the energy differences between our LRDMC values are unbiased, as verified in [2].

The files in the database are written in the extended xyz format. Their filename reveals their content. The convention used for the filename is as follows:

hydrogen_c2c24_harm_P350GPa_N96_lrdmc.xyz

The first entry can be hydrogen or deuterium and indicates the mass used in the SSCHA quantum treatment to relax the geometry. The second entry is the symmetry of the crystalline configuration. For Cmca24 we adopted the shortcut cmca, while the P62/c-24 symmetry has been indicated by hyphite. The third entry can be harm or sscha, the former if the relaxation is for classical nuclei (infinite mass), the latter is SSCHA has been used with hydrogen mass. For deuterium, this entry is not specified, because in this case SSCHA is meant to be used with deuterium mass. The fourth entry is the pressure value (equivalent to the corresponding volume reported above). The fifth entry is the number of hydrogen atoms in the supercell. The sixth and final entry indicates whether the energy provided inside the file refers to vmc or lrdmc calculations. Beside the energy, we report also the statistical error bar in a new field defined in the extended xyz format and starting by “error=”.

The units are Angstrom for lengths and Hartree for energies.

3 Protonated water hexamer

The protonated water hexamer is the smallest cluster comprising both Zundel and Eigen limiting structures of the hydrated proton in water. For this reason, it is one of the most widely studied and highly paradigmatic protonated water clusters, belonging to the series where more and more water molecules are gathered around the proton charge defect. Due to its rather compact size, the protonated hexamer is the ideal system where highly correlated methods are still affordable, while allowing for a non-trivial behavior of the solvated proton dynamics. The interplay between water supramolecular interactions, nuclear quantum effects and thermal excitations makes the solvated proton dynamics, and more generally, proton transfer, challenging problems that still need to be fully addressed with the necessary accuracy. One of the possible long-term perspectives is to predict the proton transfer behavior in complex biochemical processes relevant in living systems, with calculations accelerated by machine learning potentials derived from accurate QMC simulations.

Configurations, forces, and energies

We provide a database of trajectories produced by classical molecular dynamics (MD) simulations of the protonated water cluster, with nuclear forces computed at the VMC level of theory. These trajectories belong to the body of data published in [4]. In that work, we carried out both classical and path integral molecular dynamics (PIMD), in order to consider the quantum nature of the nuclei, described as ring polymers, according to the quantum-to-classical isomorphism. Here, we present the trajectories of classical simulations, which bear enough information for future ML analysis. Indeed, both classical and path integral simulations sample the same potential energy surface (PES), which is the main target of ML potentials derivations. Classical simulations have a more direct access to the PES, because our PIMD code stores forces and energies averaged over the particles' rings, thus blurred around the centroid position of the quantum particles.

The classical molecular dynamics is carried out by a Langevin dynamics (LD), where the noise coming from the QMC stochastic sampling contributes to the thermal excitations in a controlled way, thanks to the algorithm developed in [5]. A time step of 1 fs is used for all temperatures. The nuclear forces are computed at the VMC level, by optimizing at each LD iteration a Jastrow correlated geminal wave function, described extensively in [4]. The explicitly correlated treatment of the electronic problem yields a PES of quality comparable to coupled cluster (CCSD-CCSD(T)) theories, with a full resolution of energy and forces at a milder computational cost.

The database is made of files whose names are of the kind:

prot_hex_100k.xyz

where the temperature is explicitly reported. The extended xyz format described in Sec. 2.2 is adopted also here, with the same units. The number of configurations corresponds to the number of LD iterations, because all steps are listed. In addition to the information provided for the QMC hydrogen database, here we give also the nuclear forces acting on each nucleus, in Hartree/Angstrom units. The cluster geometry is written according to the following convention: [O, H, H, H, O, H, H, O, H, H, O, H, H, O, H, H, O, H, H], where O (H) represents the oxygen (hydrogen) position, and the two subsequent H positions belong to the H_2O unit led by the preceding O. The only exception is the first set of coordinates, where the third H after the first O is the coordinates of the proton defect, bridging the first two H_2O molecules, belonging to the cluster core.

The typical stochastic error of the total energies reported in our dataset is about 3mH, while the one of forces is about 6mH/Å, irrespective of the force component.

4 Organic molecular dimers in electronically excited states

Molecular interactions in electronically excited molecular complexes play a crucial role in fundamental processes of charge and energy transfer. Accounting for molecular interactions is therefore of importance in designing nanostructures with high phosphorescence quantum yields, or optoelectronic devices such as organic light emitting diodes. Accurate description of noncovalent interactions in excited-state molecular complexes is more demanding than that of ground states. A reliable computational method must account not only for weak intermolecular forces, including long-range correlation (dispersion interaction energy), but also for relatively strong correlation effects in excited states. DFT methods, nowadays widely used in ground state molecular interaction modelling, are no longer adequate for interactions in excited states. One of the reasons is that semiempirical corrections for the dispersion energy, developed for ground states, may be not reliable. Coupled cluster (CC) response theories are, in principle, a viable alternative, but due to their high computational cost, their applicability is limited to small systems.

In [6] and [7] we formulated a framework to describe the dispersion energy in electronically excited van der Waals complexes. We computed both benchmark dispersion energies and benchmark total interaction energies for a number of organic molecular dimers with n-π* or π-π* excitons localized on one of the monomers. The dimers are presented in Figure 2, where n-π* complexes include: peptide-water, peptide-methylamine, acetic acid-pentane, acetamide-pentane, peptide-pentane, while π-π* complexes are: benzene-water, benzene-methanol, benzene-methylamine, pyridine-water, pyridine-methanol, pyridine-methylamine, benzene-cyclopentane, benzene-neopentane.

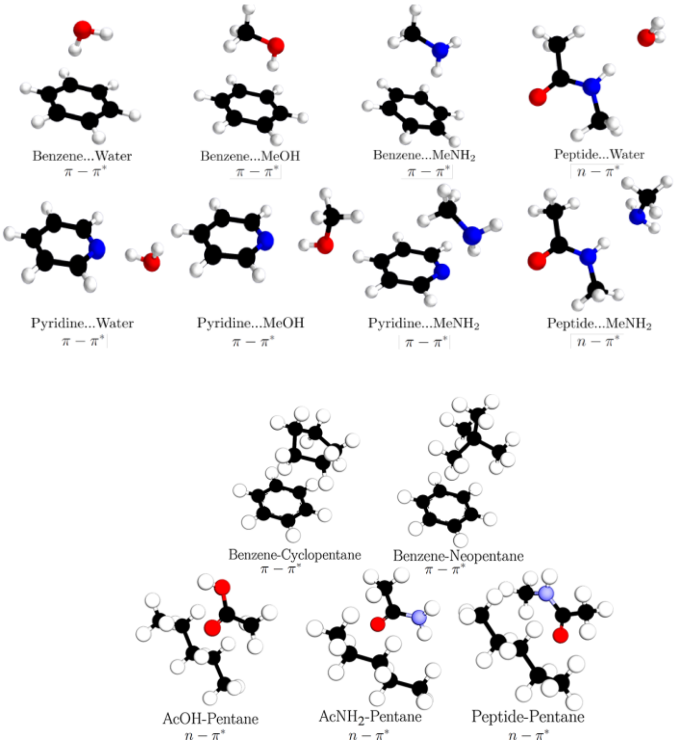

Figure 2 : Structures of n-π and π-π* complexes included in the data set.*

4.1 Dispersion interaction energies

Dispersion interaction energy is the main binding component of molecular interaction in van der Waals species. Molecular simulations, including molecular dynamics, must account for dispersion to yield reliable predictions Most density functional theory approximations, based on semilocal functionals, are generically flawed when it comes to accurate prediction of dispersion interactions. To cure this deficiency various strategies have been proposed, in most of them dispersion energy is added in a form of empirically developed corrections. The latter have been parameterized based on accurate benchmarks for ground states. Until now, benchmark *ab initio* dispersion energies data for excited states have not been available.

In [6] and [7] we have developed theoretical and computational methods dedicated to dispersion energy calculations in excited state molecular dimers. We have shown that dispersion energy interaction in excited states includes special terms related to negative transitions in density response function. Interestingly, these terms may be of the positive sign. Since they do not appear in the usual Casimir-Polder formula for dispersion interaction, we have called them non-Casimir-Polder (nCP) terms. Approximate computational methodology has been developed and actual values of the nCP terms have been presented in [6].

In [7] we have developed a novel method for direct computation of the dispersion energy, based on Cholesky decomposition of Coulomb integrals and expansion of density-density linear functions of monomers in orders of the correlation-coupling-constant. The resulting computational algorithm enables computation of the second-order dispersion interaction with a relatively modest computational cost, scaling with the fifth power of the system size. These developments have allowed us to compute dispersion energy values for all complexes depicted in Figure 2 which are presented in Table 2.

Let us emphasize that these are first dispersion energy values for excited states obtained from an *ab initio* method, i.e. a method free of empirical parameters. They can be used for benchmarking other methods capable of capturing dispersion forces in excited states. More importantly, they can be used to train empirical dispersion corrections for DFT. For the latter, it is known that dispersion energy correction should also include a repulsive component, which in the Symmetry-Adapted perturbation theory corresponds to a second-order exchange-dispersion energy. It dumps dispersion energy at short-range of the inter-monomer interaction. An expression for the exchange-dispersion energy in terms of one- and two-electron reduced density matrices has been developed in [8]. Its efficient implementation, exploiting Cholesky decomposition of integrals, which would lower the computational scaling of exchange-dispersion term by one order of magnitude (similarly to what has been achieved for dispersion energy in [7]), is still not available. Thus, in [7] we have proposed a computational procedure for the exchange-dispersion term in a large basis set. It consists in scaling the corresponding term obtained in a small basis set. The scaling factor is obtained as a ratio of the dispersion energy computed in large and small basis sets.

The benchmarks values of dispersion E_{disp}, exchange-dispersion $E_{exch\text{-}disp}$, their sum E_{DISP}, and total interaction energies (computed from multiconfigurational SAPT method [9]) are presented in Table 2. It is evident that the dispersion energy constitutes a dominating part of the total interaction and it is a substantial binding factor of complexes in excited states. Respective benchmark data is compiled in *intexcit* dataset with entries in the following order: system/E_{disp}/$E_{exch\text{-}disp}$/E_{DISP}/nCP.

Table 2: Dispersion energy E_{disp}, exchange-dispersion $E_{exch-disp}$, their sum E_{DISP}, and the total interaction energy E_{int} (computed from multiconfigurational SAPT method) for molecular complexes shown in Figure 2. Data from [6] and [7]. All values in kcal/mol. aug-cc-pVTZ basis set has been used.

Dimer	E_{disp}	$E_{exch-disp}$	E_{DISP}	E_{int}
benzene-water	-2.88	0.33	-2.55	-2.51
benzene-MeOH	-4.63	0.52	-4.11	-3.25
benzene-$MeNH_2$	-4.62	0.54	-4.08	-2.62
pyridine-water	-4.05	0.84	-3.21	-6.91
pyridine-MeOH	-4.95	0.99	-3.96	-7.44
pyridine-$MeNH_2$	-5.01	0.66	-4.35	-3.82
peptide-water	-2.93	0.46	-2.47	-4.36
peptide-$MeNH_2$	-5.78	1.10	-4.68	-6.40
benzene-cyclopentane	-6.89	0.82	-6.07	-3.63
benzene-neopentane	-5.38	0.61	-4.77	-2.95
AcOH-pentane	-5.61	0.56	-5.05	-2.82
$AcNH_2$-pentane	-6.66	0.81	-5.85	-3.61
peptide-pentane	-8.10	0.88	-7.22	-4.26

4.2 Molecular interaction energies

As already mentioned, existing electronic structure methods struggle with description of molecular complexes in excited states as many electron correlation effects must be considered simultaneously. Coupled cluster (CC) methods with only singles and doubles may not be sufficiently accurate, while CC with triples is limited to small and medium-sized systems. For dimers in Figure 2, reliable CC interaction energies, obtained by combining the CCSD(T)/CBS (CBS - complete basis set) description of the ground state with excitation energies calculated at the EOM-CCSD level of theory have been available. We have used these results to validate the accuracy of the recently proposed, novel multireference methods. They are based on the multiconfigurational wavefunction description of the interacting molecules and are less demanding computationally than the CCSD(T) method, which allows for applications to larger systems. In [6] we have evaluated the performance of a number of methods (Figure 3). Three approaches were identified as the most reliable, providing new benchmark values for interaction energies for several model systems (Figure 2) including n-π* complexes: peptide-water, peptide-methylamine, and π-π* complexes: benzene-water, benzene-methanol, benzene-methylamine, pyridine-water, pyridine-methanol, pyridine-methylamine. The methods recommended for computing interaction energies in excited state dimers are:

- SAPT(CAS): a reduced-density matrix-based second-order symmetry-adapted perturbation method developed for multiconfiguration wavefunction description of monomers [9]; total interaction energy is given as a sum of electrostatic (E_{elst}), exchange E_{exch}, induction E_{ind}, exchange-induction ($E_{exch-ind}$), dispersion (E_{disp}), exchange-dispersion ($E_{exch-disp}$), and higher-order induction ($\delta_{HF/CAS}$) terms
- CAS+DISP: a sum of supermolecular complete active space (CAS) self-consistent field interaction energies and a dispersion energy computed within the second-order symmetry-adapted perturbation method
- lrAC0-CAS: a sum of supermolecular complete active space (CAS) self-consistent field interaction energies with electronic interaction restricted to long-range with short-range exchange-correlation density functional and a long-range correlation energy

The SAPT(CAS) method provided not only benchmark values for the total interaction energies but also their components in the second-order perturbation theories. They can be used to train ML models for electrostatic, dispersion and induction energies or to benchmark future energy decomposition analysis methods.

For three water complexes we also include benchmark values of SAPT(CAS) energy components extrapolated to the complete basis set (CBS) limit, both for ground and excited electronic states. The electrostatic (E_{elst}), exchange (E_{exch}), induction (E_{ind} and $E_{exch\text{-}ind}$), and dispersion (E_{disp} and $E_{exch\text{-}disp}$) terms together with relevant corrections ($\delta_{HF/CAS}$ terms which approximate higher-order induction effects) are presented in Table 3.

Table 3: Benchmark SAPT(CAS) results for ground (GS)- and excited (ES)-state dimers: peptide-water, pyridine-water and benzene-water. E_{elst}, E_{exch}, E_{ind}, $E_{exch\text{-}ind}$ energies are computed in the aug-cc-pVQZ basis set. The E_{disp} and $E_{exch\text{-}disp}$ terms are extrapolated from aug-cc-pVTZ and aug-cc-pVQZ results according to the two-point scheme of Halkier et al [10]. First-order effects beyond the S^2 approximation (S^2_{corr}) are estimated as the difference between E^{10}_{exch} and $E^{10}_{exch}(S^2)$ components at the SAPT0 level of theory. The δ_{CAS} correction for higher-order induction effects in excited states is obtained by scaling of the the δ_{HF} term: $\delta_{CAS} = \delta_{HF} * E_{ind}(ES)/E_{ind}(GS)$. All values in kcal/mol.

Ground State	E_{elst}	E_{exch}	E_{ind}	$E_{exch\text{-}ind}$	E_{disp}	$E_{exch\text{-}disp}$	δ_{HF}	S^2_{corr}	E_{int}
peptide-water	-6.64	5.77	-2.23	1.19	-3.25	0.54	-0.68	0.03	-5.28
pyridine-water	-11.27	11.11	-5.34	3.14	-4.43	0.92	-1.56	0.13	-7.32
benzene-water	-2.63	3.00	-1.25	0.68	-3.11	0.39	-0.34	0.01	-3.26
Excited State	E_{elst}	E_{exch}	E_{ind}	$E_{exch\text{-}ind}$	E_{disp}	$E_{exch\text{-}disp}$	δ_{CAS}	S^2_{corr}	E_{int}
peptide-water	-5.97	5.72	-2.10	1.15	-3.26	0.54	-0.64	0.03	-4.51
pyridine-water	-11.31	11.13	-5.33	3.14	-4.40	0.91	-1.55	0.13	-7.28
benzene-water	-1.81	2.66	-1.15	0.64	-2.94	0.34	-0.32	0.01	-2.57

Benchmark data described in this section is collected in the *intexcit* dataset with the following entries:

- interaction energies in ES: system/CAS/CAS+DISP/IrAC0-CAS/SAPT(CAS)
- SAPT interaction energy components: system/E_{elst}/E_{exch}/E_{ind}/$E_{exch\text{-}ind}$/E_{disp}/$E_{exch\text{-}disp}$/δ_{CAS}
- SAPT energies extrapolated to CBS: system/E_{elst}/E_{exch}/E_{ind}/$E_{exch\text{-}ind}$/E_{disp}/$E_{exch\text{-}disp}$/δ_{CAS}/S2/E_{int}

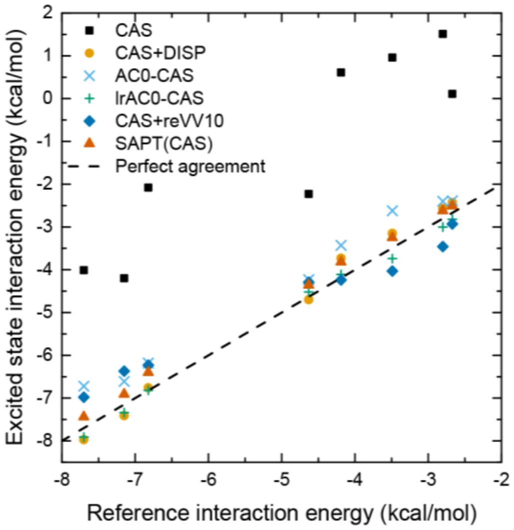

Figure 3: Correlation plots for interaction energies. "Reference interaction energy" values are estimations of EOM-CCSD(T) energies. Data from [6].

5 Summary and outcomes

Four datasets have been made available for benchmarking and ML modelling of: hydrogen under pressure, protonated water hexamer, and molecular and dispersion interactions of organic molecules in excited states. The basis sets have been created by using open-source, high-performance, interoperable flagship QMC codes and libraries developed by the TREX consortium. The basis sets will be used to train ML models and critically assess performance of MLPs. They may be also used to parameterize density functionals or DFT-dispersion interaction corrections to extend their scope of applicability to excited states.

The following datasets have been presented in this report:

(1) *h-llpt-24* data set for hydrogen under high pressure: geometries, lattice vectors, energies, forces, stresses, Wigner-Seitz radii, configurations of 128 hydrogen atoms at the DFT level of theory; purpose: benchmark for MLPs; developed in the group of Matthias Rupp (LIST)

(2) *hydrogen-harm* data set of crystalline hydrogen configurations: energies at VMC and LRDMS level; purpose: benchmark for MLP; developed in the group of Michele Casula (CNRS)

(3) *prot-hex* data set for protonated water hexamer: trajectories from classical molecular dynamics with nuclear forces at VMC level of theory; purpose: ML modelling; developed in the group of Michele Casula (CNRS)

(4) *intexcit* data sets for a set of organic molecular complexes in lowest excited states: dispersion interaction energies, interaction energies, components of SAPT interaction energies at the CAS wavefunction level; purpose: benchmarking *ab initio* methods and density functional dispersion correction modelling; developed by Kasia Pernal (TUL) and Michal Hapka (University of Warsaw)

Datasets (2)-(4) are publicly available at: https://zenodo.org/records/10547300

In addition to potential applications of the datasets listed above, in the broad-term perspective their availability will contribute to developing faster computational methods allowing study of exotic phenomena in different phases of hydrogen, solvated proton dynamic in water clusters including proton transfer in complex biochemical processes, studying light emitting excimers paving the way to developing novel light emitting organic materials.

The following papers are related to this report and acknowledge TREX funding:

- M.R. Jangrouei, A. Krzeminska, M. Hapka, E. Pastorczak, K. Pernal, Dispersion Interactions in Exciton-Localized States. Theory and Applications to π − π* and n − π* Excited States, Journal of Chemical Theory and Computation 18, 3497 (2022).
- M. Hapka, A. Krzeminska, M. Modrzejewski, M. Przybytek, and K. Pernal, Efficient Calculation of the Dispersion Energy for Multireference Systems with Cholesky Decomposition: Application to Excited-State Interactions, Journal of Physical Chemistry Letters 14, 6895 (2023).
- L. Monacelli, M. Casula, K. Nakano, S. Sorella, F. Mauri, Quantum phase diagram of high-pressure hydrogen, Nature Physics **19**, 845 (2023).
- F. Mouhat, M. Peria, T. Morresi, R. Vuilleumier, A.M. Saitta, M Casula, Thermal dependence of the hydrated proton and optimal proton transfer in the protonated water hexamer, Nature Communications **14**, 6930 (2023).

References

[1] O. T. Unke, S. Chmiela, H.E. Sauceda, M. Gastegger, I. Poltavsky, K. T. Schütt, A. Tkatchenko, and K.-R. Müller, Machine Learning Force Fields, Chemical Review **121**, 10142 (2021).

[2] L. Monacelli, M. Casula, K. Nakano, S. Sorella, F. Mauri, Quantum phase diagram of high-pressure hydrogen, Nature Physics **19**, 845 (2023).

[3] R. Ramakrishnan, P. O. Dral, M. Rupp, and O. A. von Lilienfeld, Big Data Meets Quantum Chemistry Approximations: The Δ-Machine Learning Approach, Journal of Chemical Theory and Computation **11**, 2087 (2015).

[4] F. Mouhat, M. Peria, T. Morresi, R. Vuilleumier, A.M. Saitta, M Casula, Thermal dependence of the hydrated proton and optimal proton transfer in the protonated water hexamer, Nature Communications **14**, 6930 (2023).

[5] C. Attaccalite and S. Sorella, Stable Liquid Hydrogen at High Pressure by a Novel Ab Initio Molecular-Dynamics Calculation, Physical Review Letters **100**, 114501 (2008).

[6] M.R. Jangrouei, A. Krzeminska, M. Hapka, E. Pastorczak, K. Pernal, Dispersion Interactions in Exciton-Localized States. Theory and Applications to $\pi - \pi^*$ and $n - \pi^*$ Excited States, Journal of Chemical Theory and Computation **18**, 3497 (2022).

[7] M. Hapka, A. Krzeminska, M. Modrzejewski, M. Przybytek, and K. Pernal, Efficient Calculation of the Dispersion Energy for Multireference Systems with Cholesky Decomposition: Application to Excited-State Interactions, Journal of Physical Chemistry Letters **14**, 6895 (2023).

[8] M. Hapka, M Przybytek, K. Pernal, Second-Order Exchange- Dispersion Energy Based on a Multireference Description of Monomers, Journal of Chemical Theory and Computation **15**, 6712 (2019).

[9] M. Hapka, M. Przybytek, K. Pernal, Symmetry-Adapted Perturbation Theory Based on Multiconfigurational Wave Function Description of Monomers, Journal of Chemical Theory and Computation **17**, 5538 (2021).

[10] A. Halkier, T.Helgaker, P. Jørgensen, W. Klopper, J. Olsen, Basis-set convergence of the energy in molecular Hartree–Fock calculations, Chemical Physics Letters **302**, 437 (1999).