npj | computational materials

Article

Published in partnership with the Shanghai Institute of Ceramics of the Chinese Academy of Sciences

https://doi.org/10.1038/s41524-024-01249-y

Coupled cluster finite temperature simulations of periodic materials via machine learning

Basile Herzog [1], Alejandro Gallo [2], Felix Hummel [2], Michael Badawi [1,3], Tomáš Bučko [4,5] ✉, Sébastien Lebègue[1], Andreas Grüneis [2] ✉ & Dario Rocca [1] ✉

Density functional theory is the workhorse of materials simulations. Unfortunately, the quality of results often varies depending on the specific choice of the exchange-correlation functional, which significantly limits the predictive power of this approach. Coupled cluster theory, including single, double, and perturbative triple particle-hole excitation operators, is widely considered the 'gold standard' of quantum chemistry as it can achieve chemical accuracy for non-strongly correlated applications. Because of the high computational cost, the application of coupled cluster theory in materials simulations is rare, and this is particularly true if finite-temperature properties are of interest for which molecular dynamics simulations have to be performed. By combining recent progress in machine learning models with low data requirements for energy surfaces and in the implementation of coupled cluster theory for periodic materials, we show that chemically accurate simulations of materials are practical and could soon become significantly widespread. As an example of this numerical approach, we consider the calculation of the enthalpy of adsorption of CO_2 in a porous material.

In the past few decades, density functional theory (DFT) has become the workhorse of materials simulations and owes its impressive success to its good compromise between accuracy and numerical efficiency[1]. The achievements of DFT come, of course, at a cost: The unknown exchange-correlation function has to be approximated. Very simple approximations derived from the uniform electron gas, such as the local density approximation[1,2], already provide satisfactory accuracy in describing the properties of periodic materials. However, standard DFT functionals are known to fail for certain classes of systems, particularly where weak interactions are important or where the electronic correlation is strong[3,4]; importantly, chemical accuracy is not systematically achieved, and this often limits the predictive power of DFT-based materials simulations. Because of the non-systematic control on the accuracy, it is often difficult to understand if the failure to accurately reproduce or predict experimental results is due to the inadequacy of the particular model under consideration or of the approximations involved in the DFT functional.

Correlated quantum chemical methods based on post-Hartree-Fock (post-HF) approximations are instead systematically improvable and could potentially overcome some of the limitations of DFT for materials simulations. Among those, second-order Møller-Plesset perturbation theory (MP2)[5] and coupled cluster theory[6] have been recently implemented for periodic materials[7–11]. However, their computational cost is significant for most practical applications in materials science and this issue becomes even more dramatic when finite-temperature effects have to be included by performing molecular dynamics (MD) simulations or Monte Carlo sampling. For example, a brute-force computation of the enthalpy of adsorption considered in this work would require billions of CPU hours and hundreds of real-time years to be completed.

In the context of MD simulations, machine learning (ML) is nowadays a well-established tool to achieve larger system sizes and longer time scales[12–15]. This is achieved by decomposing the total energy in atomic contributions and using ML regression models to "fit" the interatomic potential. Still, the ML-accelerated MD typically requires large amounts of data and becomes rapidly challenging for the more expensive approximations. In this work, we show how finite-temperature observables for periodic materials can be evaluated using the 'gold standard' coupled cluster ansatz,

[1]Université de Lorraine and CNRS, LPCT UMR 7019, F-54000 Nancy, France. [2]Institute for Theoretical Physics, TU Wien, Vienna, Austria. [3]Université de Lorraine, CNRS, L2CM, F-57000 Metz, France. [4]Comenius University in Bratislava, Department of Physical and Theoretical Chemistry, Faculty of Natural Sciences, Mlynská Dolina, Ilkovičova 6, SK-84215 Bratislava, Slovakia. [5]Institute of Inorganic Chemistry, Slovak Academy of Sciences, Dúbravská cesta 9, SK-84236 Bratislava, Slovakia. ✉e-mail: tomas.bucko@uniba.sk; andreas.grueneis@tuwien.ac.at; dario.rocca@univ-lorraine.fr

including single, double, and perturbative triple particle-hole excitation operators (CCSD(T)) in combination with machine learning techniques coupled with thermodynamic perturbation theory and Monte Carlo sampling. Within this approach, the computational cost is limited to a small number (a few tens) of single-point energy calculations that are then used to train a data-efficient ML model.

For molecular systems, the application of ML techniques has already been proven to be effective in enhancing the efficiency of CCSD(T) MD simulations[13,16–19]. Very recently, applications to molecular condensed phase systems, specifically to liquid water, have also been considered. In ref. [20], the ML model for periodic water was trained with data produced for finite water clusters using near-linear scaling coupled cluster theory. In ref. [21], CCSD(T) calculations were restricted to very small periodic models based on a box of 16 H_2O molecules, and the ML model was then used to compute radial distribution functions, diffusion coefficients, and vibrational densities of states. To the best of our knowledge, the application of CCSD(T) to finite-temperature simulations of periodic solid materials has not been previously reported in the literature. Our study was challenging for several reasons. First, we dealt with a system containing more than 200 electrons, far more than used in any previous report on ML-assisted MD CCSD(T) simulations. Second, we focused on a measurable thermodynamic quantity, the enthalpy of adsorption, whose prediction imposes high demands on the quality of the ML model. This is because any error in the energy of a configuration affects not only the underlying phase space function used in ensemble averaging but also the statistical weight of that contribution (see Eq. (2)).

This work is based on the combination of two ingredients. The first is an efficient periodic coupled cluster theory implementation. This implementation is based on a plane-wave basis set and finite size and basis set correction techniques that accelerate the convergence to the complete basis set limit and thermodynamic limit significantly[22,23]. Using these techniques, it is possible to obtain well-converged correlation energies at the CCSD(T)-level of theory for periodic solids and surfaces containing more than 100 electrons on modern supercomputers[24–27].

The second fundamental ingredient is an approach that couples machine learning and thermodynamic perturbation theory (TPT)[28], which will be denoted as MLPT[29–32]. Within this approach, an ab initio molecular dynamics simulation is first performed at an affordable level of theory (semi-local DFT), and the statistical distribution is subsequently reweighted to obtain observables at a higher level of theory (e.g. coupled cluster). While this TPT procedure requires, in principle, a large number of single-point calculations at the expensive level of theory, in practice, those can be replaced to a large extent by inexpensive machine-learning predictions. By using efficient machine learning algorithms based on the smooth overlap of atomic positions (SOAP) kernel[33,34] and $\Delta-$ML[35], MLPT requires a limited amount of data to be trained on. For example, the calculation of enthalpies of adsorption at the random phase approximation (RPA) level of theory achieved convergence with as few as 10 single configuration energies[29]. This is particularly important when employing expensive approximations in a finite-temperature context since otherwise, the amount of single-point calculations and the associated computational cost would be too large.

As a specific application of our approach, we consider the calculation of the enthalpy of adsorption of carbon dioxide in protonated chabazite (HChab). The adsorption of molecules in zeolites is fundamental for many applications, including depollution, separation of chemicals, and catalysis[36–38]. In this field, more quantitative and systematically improvable theoretical predictions are instrumental in interpreting experimental findings and predicting new materials. Although the calculations presented in this work are still significantly more expensive than those based on standard density functional theory, our proof-of-principle work paves the way to a more systematic use of highly accurate post-HF methods in materials simulations.

Results

Enthalpy of adsorption from first principles

In this work, we consider the calculation of the enthalpy of adsorption of carbon dioxide in a porous zeolitic material, chabazite. In practice, this quantity is computed as

$$\begin{aligned}\Delta_{\mathrm{ads}}H(\mathrm{M@Z}) &= \Delta_{\mathrm{ads}}U(\mathrm{M@Z}) + \Delta_{\mathrm{ads}}(pV)(\mathrm{M@Z}) \\ &= \langle E(\mathrm{M@Z})\rangle - (\langle E(\mathrm{M})\rangle + \langle E(\mathrm{Z})\rangle) - k_{\mathrm{B}}T,\end{aligned} \tag{1}$$

where $\Delta_{\mathrm{ads}}U$ is the internal energy of adsorption, $\langle E(i)\rangle$ denotes the ensemble average of the total energy of the system i corresponding to a gas phase molecule (M), clean zeolite (Z), and the adsorbed system (M@Z), and the identity $\Delta_{\mathrm{ads}}(pV)(\mathrm{M@Z}) = -k_{\mathrm{B}}T$ is obtained by assuming an ideal gas behavior of M and a negligible change of pV of the zeolite due to adsorption. The value of T is fixed at 300 K in all our simulations. The canonical ensemble energy can be evaluated by directly performing an ab initio molecular dynamics (AIMD) simulation but because of the high computational cost of CCSD(T) and MP2, this approach is impractical at these levels of theory. Based on the plane-wave basis set coupled cluster calculations are performed in several steps involving Hartree–Fock and MP2 theory to obtain corresponding energies and optimized approximate natural orbitals[39]. Once the natural orbitals have been computed, the Cc4s[40] interface to VASP[41,42] is used to compute intermediate quantities[43] that are needed for the subsequent coupled cluster energy calculations, including the corresponding finite size[22] and basis set corrections[23]. In the present calculation, 10 unoccupied approximate natural orbitals per occupied orbital are used for the CCSD calculations, whereas only 5 unoccupied approximate natural orbitals per occupied orbital are employed to evaluate the (T) contribution. A single CCSD(T) calculation for the given structures containing up to 40 atoms took about 10,000 core hours.

Machine learning approach

The large number of high-level calculations required to estimate the enthalpy (or other finite-temperature quantities) can be significantly decreased using machine learning techniques. Specifically, starting from an AIMD trajectory obtained using numerically affordable semi-local DFT with empirical van der Waals corrections (PBE + D2)[44,45], the post-HF ensemble energies are estimated using the MLPT approach trained on a small number of single-point calculations. This approach is described in detail in ref. [29–32], and the two main steps are summarized here:

1. Given a set of configurations $\{\mathbf{R}_i\}_{i=1}^{M}$ from an AIMD trajectory in an NVT ensemble with the PBE + D2 reference Hamiltonian H_0 and potential energy E_0, the ensemble average energy generated by the target Hamiltonian H_1 with potential energy E_1 (MP2 or CC level) can be obtained from thermodynamic perturbation theory by reweighting:

$$\langle E_1\rangle_1 = \frac{\sum_{i=1}^{M} E_1(\mathbf{R}_i)\exp(-\beta\Delta E(\mathbf{R}_i))}{\sum_{i=1}^{M}\exp(-\beta\Delta E(\mathbf{R}_i))}, \tag{2}$$

   where $\Delta E(\mathbf{R})$ denotes the energy difference $E_1(\mathbf{R}) - E_0(\mathbf{R})$ for a specific atomic configuration $\mathbf{R}$. In this work, E_1 denotes either the MP2 or the CCSD(T) target method potential energy. The trajectory obtained with the reference Hamiltonian is called the production trajectory.

2. While the application of Eq. (2) requires a large number of high-level calculations, in practice, those can be largely replaced by inexpensive predictions of a machine learning model. MLPT limits the amount of data required for the training by using efficient algorithms based on the kernel ridge regression with the SOAP kernel[33,34] and $\Delta-$ML[35]. $E_0(\mathbf{R})$ is known, and the evaluation of Eq. (2) requires only the energy difference $\Delta E(\mathbf{R})$.

Since MLPT is based on thermodynamic perturbation theory, a limited overlap between the production and target configurational spaces can lead to inaccurate results. If a suboptimal overlap is suspected, a Monte Carlo (MC) resampling can be performed. This procedure, described in detail in ref. 32, uses Metropolis MC[46] to resample the canonical ensemble at the CCSD(T) and MP2 levels of theory. At each MC step, configurational energies are computed with the production approximation (PBE + D2) and

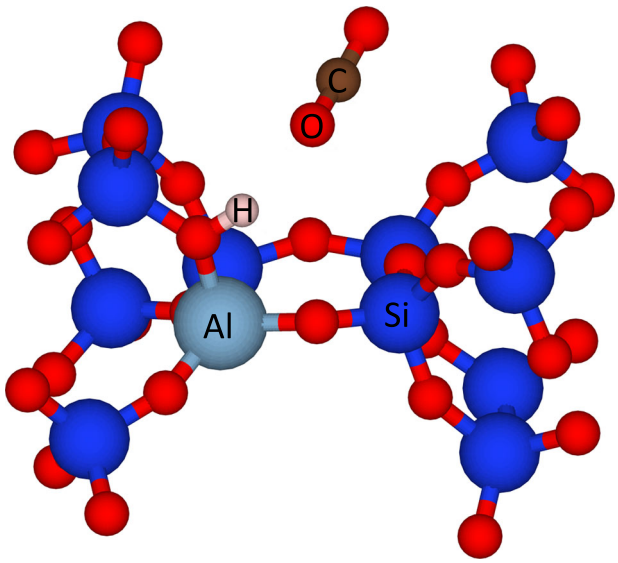

Fig. 1 | Adsorbed system. The unit cell of the system studied in this work is CO_2 in protonated chabazite.

subsequently evaluated at the post-HF level using the same ML model of MLPT. The Metropolis acceptance criterion is applied at the target level of theory, and accordingly, the correct target configurational space is sampled without bias from the starting point.

Calculation of molecular adsorption enthalpies in zeolites

Let us now present and discuss the adsorption enthalpies of CO_2 in protonated chabazite as computed at the MP2 and CCSD(T) levels of theory. The latter approximation is commonly described as the 'gold standard' of quantum chemical simulations and is routinely used to produce reference test sets to benchmark the accuracy of other methods[47,48]. The primitive cell of the model considered here is shown in Fig. 1.

The experimental value of the enthalpy of adsorption of CO_2 in HChab, -8.41 kcal mol^{-1} [49], is used as a reference for the computational results. This experimental estimate is obtained by extrapolating measurements to the zero coverage limit. The errors possibly arising from this procedure are not discussed in ref. [49] and we cannot exactly quantify the uncertainty in the experimental reference.

The computed results are presented in Table 1, where the error bars related to the finite sampling and the ML model are also indicated[29]. The molecular dynamics at the PBE + D2 level leads to an estimate for the adsorption energy, which is more than 1 kcal mol^{-1} below the experimental value, corresponding to a deviation well beyond chemical accuracy. This MD trajectory is used as a starting point for MLPT to obtain post-HF enthalpies. Similarly, the MP2 approximation obtained from MLPT also tends to overbind and leads to results that do not qualitatively differ from PBE + D2. This is not surprising and we believe that this overestimation is caused by the lack of screening of long-ranged correlation effects in MP2 theory. The computational estimate of the enthalpy significantly improves at the CCSD(T) level, which provides a value in excellent agreement with the experiment. This result demonstrates the high accuracy and predictive power of the CCSD(T) approximation also for finite-temperature simulations of materials.

In a previous work, we demonstrated that the RPA also provides accurate enthalpies of adsorption of molecules in zeolites[29]. Specifically, the value for CO_2 in protonated chabazite is -8.01 kcal mol^{-1}. Although the RPA has a diagrammatic structure it is not as straightforward to systematically improve its accuracy as for post-HF methods[50–57]. In practice, the RPA often provides more realistic results starting from a DFT approximation rather than from HF[52], and this starting point dependence makes this approximation less reliable as a general predictive method.

Table 1 | Enthalpy of adsorption of CO_2 in protonated chabazite (kcal mol^{-1}) computed using different target and sampling methods

Target method	Sampling method	Enthalpy (kcal mol^{-1})
PBE + D2	MD	−9.72 ± 0.27
MP2	MLPT	−9.50 ± 0.24
CCSD(T)	MLPT	−8.32 ± 0.28
CCSD(T)	MLMC	−8.09 ± 0.71
Experiment[49]	Adsorption isotherms	−8.41

Discussion

To fully prove the accuracy of the MLPT approach for MP2 and CCSD(T), a crucial point concerns the reliability of the PBE + D2 trajectory used as a starting point for thermodynamic perturbation theory. Specifically, if the target (MP2 or CCSD(T)) configurational space has a small overlap with the production (PBE + D2) configurational space, the results of TPT may be affected by a strong systematic error. As discussed thoroughly in ref. [32] for systems similar to the one considered here, the occurrence of this issue can be identified even if the exact target trajectory is unknown. Thermodynamic perturbation theory is based on the reweighting of the statistics sampled by the production trajectory to obtain the target level statistics (see Eq. (2)); in case of a poor overlap, only a few configurations contribute to the total weight, leading to poor ensemble estimates. In practice, this effect can be measured by the I_w index, as defined in ref. [32]. This index assumes the value of 0.5 in the optimal configuration overlap case and tends to 0 for decreasing overlaps. For the adsorption of molecules in zeolites, it has been shown that even relatively small values of I_w around 0.03–0.05 still allow for reliable MLPT estimates[32]. The reweighting of the trajectories at the MP2 level provides large values for I_w (>0.15), and the corresponding enthalpies in Table 1 should be considered fully reliable. For the CCSD level, a very low I_w value for the adsorbed system (0.008) precludes making any reliable predictions of adsorption enthalpy; for this reason, this level of theory is not discussed here. For the CCSD(T) level of theory, the I_w coefficient is one order of magnitude higher: 0.07 for HChab and 0.05 for the adsorbed system, indicating a better match between the PBE + D2 equilibrium structure as compared to the CCSD level. While these I_w values are likely to be sufficient to confirm the reliability of our results[32], considering the pioneering nature of our work and the lack of any previous finite-temperature benchmark results for periodic CCSD(T), we further investigated the robustness of the MLPT estimate by resampling the CCSD(T) trajectory. This is achieved by performing a Metropolis Monte Carlo (MC) sampling of the canonical ensemble at the CCSD(T) level by replacing the expensive coupled-cluster calculations with the predictions of the same machine learning model previously trained for MLPT. Differently from most machine learning-based MD approaches[12–15], this MLMC approach avoids training on atomic forces, which are not readily available in the current periodic CCSD(T) implementation and would require a significant overhead cost. Since thermodynamic perturbation theory is not used and a new trajectory is instead sampled from scratch, MLMC avoids the starting point bias. The corresponding result for the enthalpy of adsorption, shown in Table 1, differs by only 0.2 kcal mol^{-1} from the MLPT value. In the MLMC case, the error bar is, however, sizeably larger because of the long autocorrelation length of this trajectory (about a factor 10 longer than for the MD trajectory), but this is sufficient to support our conclusion that the PBE + D2 trajectory provides a reliable starting point to compute CCSD(T) ensemble energies. This is also qualitatively confirmed by visualizing the (high-dimensional) geometries sampled by the MD and MC methods with the t-distributed stochastic neighbor embedding (t-SNE) algorithm[58]. As shown in Fig. 2, the PBE + D2 molecular dynamics and the CCSD(T) Monte Carlo trajectories span configurational spaces that overlap well. This figure also demonstrates that the training set provides a rather uniform sampling of the data, as required for a balanced training of the ML model.

To further analyze the overlap between the configurational space of the PBE + D2 functional and of the post-HF methods we consider the structure of the protonated chabazite cage. For this purpose, the radial distribution function of the Si–O pairs has been computed for the PBE+D2 molecular dynamics trajectory and for MP2 and CCSD(T) approaches using MLPT and MLMC. As previously shown in ref. [32], the most spectacular failures of MLPT are encountered when the production approximation predicts equilibrium distances of covalent bonds that differ from the target theory; this translates to very different configurational spaces and fully unreliable perturbative estimates. For protonated chabazite, Fig. 3 clearly shows that the radial distribution functions computed for the Si-O pairs are similar at different levels of theory, and problematic behaviors of MLPT should not be expected.

Finally, it is important to notice that the effects included by MLPT do not correspond to a trivial energy correction. Within a simplified approach, the coupled cluster and MP2 enthalpies could be approximated as

$$\Delta_{ads}H_{CCSD(T)/MP2} \approx \Delta_{ads}H_{PBE+D2} + (\Delta_{ads}E_{CCSD(T)/MP2} - \Delta_{ads}E_{PBE+D2}). \quad (3)$$

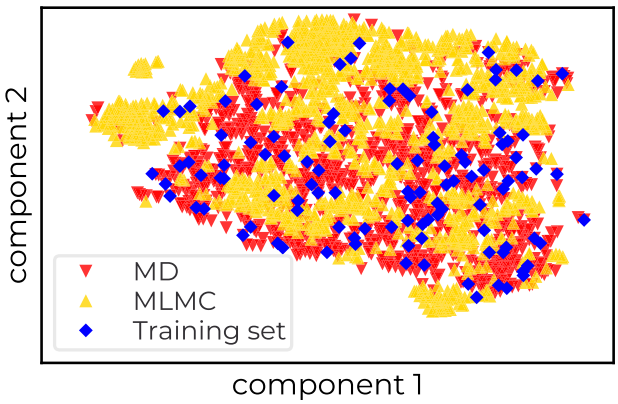

Fig. 2 | Visualization of the configurational spaces. t-SNE representation of the configurational spaces spanned by the PBE + D2 molecular dynamics (MD) trajectory and the CCSD(T) machine learning Monte Carlo (MLMC) trajectory. The configurations included in the training set are also shown to demonstrate that they cover essentially the whole relevant part of the configurational space sampled at the CCSD(T) target level. The axes represent the two components of the t-SNE projection.

In this case $\Delta_{ads}E_{CCSD(T)/MP2} = E_{CCSD(T)/MP2}(M@Z) - E_{CCSD(T)/MP2}(M) - E_{CCSD(T)/MP2}(Z)$ corresponds to the CCSD(T) or MP2 adsorption energies computed using the structures corresponding to potential energy minima determined at the PBE+D2 level; $\Delta_{ads}E_{PBE+D2}$ is analogously defined for PBE + D2. This approximation is efficient since a single post-HF calculation is required for each one of the three systems. However, this "static" approach is based on a strong and not generally valid assumption that the post-HF and DFT approaches produce energy surfaces that are shifted by a constant but otherwise parallel. MLPT instead requires only a reasonable overlap between the configurational spaces of production and target approximations, which can be tested using the I_w index, and "deformation" effects of the energy surface are kept into account by the reweighting in Eq. (2). By applying Eq. (3) we obtain $\Delta_{ads}H_{CCSD(T)} = -8.69$ kcal mol^{-1} and $\Delta_{ads}H_{MP2} = -9.03$ kcal mol^{-1}. The CCSD(T) enthalpy obtained in this way agrees fairly well with the MLPT value. However, within this static correction approach, CCSD(T) and MP2 provide very similar results, while MLPT showed that these results should differ by about 1.2 kcal mol^{-1} (see Table 1). This observation shows that the static correction approach of Eq. (3), while providing reasonable estimates in some cases with fortuitous error cancellations, is not reliable in general and can lead to misconceptions.

In conclusion, we have presented an application of CCSD(T) to compute the enthalpy of adsorption of carbon dioxide in a periodic model of zeolite. Due to the high computational cost, applications of CCSD(T) to periodic materials are so far limited, and direct calculations of finite-temperature observables are unpractical in terms of required computational resources and execution time. Here we showed that these challenges can be overcome by coupling machine learning models requiring small training sets with an efficient implementation of periodic coupled cluster theory. The computed enthalpy of adsorption of carbon dioxide in protonated chabazite was found to be in excellent agreement with the experiment. While still significantly more expensive than approaches based on density functional theory, our pioneering work opens the door to more reliable and predictive simulations of materials in finite-temperature conditions. Future work will be aimed at demonstrating the accuracy of ML-based CCSD(T) in broader classes of problems, including, for example, the computation of free energies of activation, which play a fundamental role in the modeling of catalytic reactions.

Methods

Coupled cluster calculations

The coupled cluster theory calculations are performed using the Cc4s code[40], which is interfaced with the Vienna ab initio simulation package

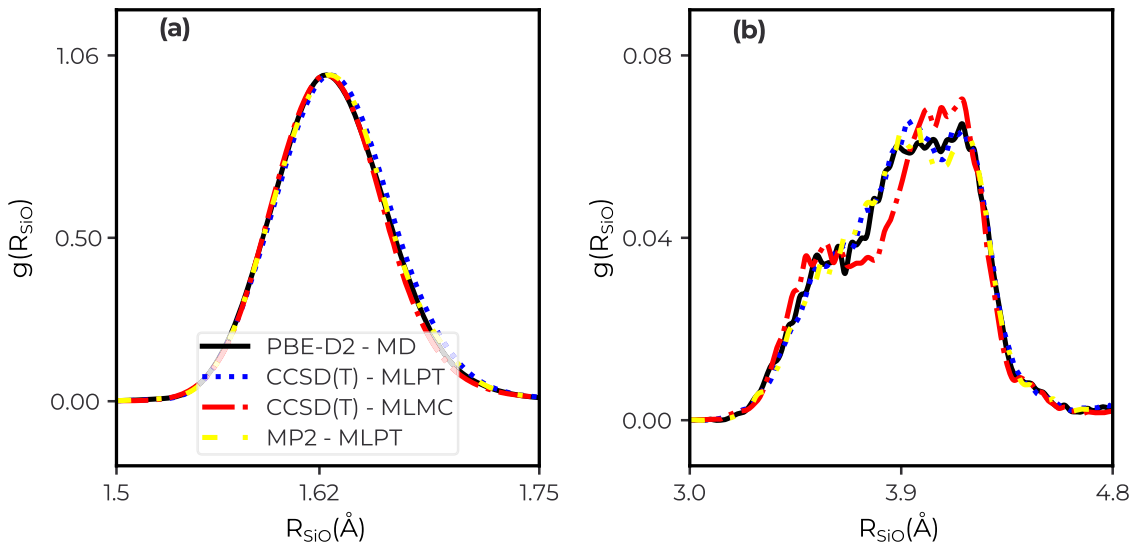

Fig. 3 | Radial distribution functions. First (**a**) and second (**b**) series of peaks of the partial radial distribution function for the Si–O pairs determined at different levels of theory.

(VASP)[41,42]. In ref. [24], all individual steps are described when combined with an embedding approach, which was not necessary for the present system due to its relatively small unit cell containing up to 40 atoms only. The convergence of the CCSD and (T) correlation energy contributions to the molecular adsorption energy was tested on a single configuration, and the details are reported in the Supplementary Information.

Ab initio molecular dynamics

ab initio molecular dynamics simulations based on the PBE + D2[44,45] were performed in the NVT ensemble, and the simulation temperature of 300 K was controlled using the Andersen thermostat[59] with a collision probability of 0.05. Two hundred thousand configurations were sampled with a time-step of 0.5 fs for a total of 100 ps. The first 10 ps of each trajectory were discarded as the equilibration period. The cell parameters were fixed to the values obtained, optimizing the chabazite cell at the PBE level[60]. The VASP electronic structure program was used for all AIMD simulations and single-point calculations within the Γ point approximation. Hydrogen atomic mass was set to 3.0 au.

ML methods

In this work, the training set is based on 100 uncorrelated configurations evenly spaced along the PBE + D2 trajectories and 10 randomly chosen configurations for the test set. The MP2 and CCSD(T) calculations are performed only for those selected geometries. Kernel ridge regression, using the rematch kernel[34] and the smooth overlap of atomic positions (SOAP) descriptor, was used as implemented in the DScribe library[61]. The model is trained to predict the differences between the post-HF and the PBE + D2 energies[35]. Details of the hyperparameter tuning and model accuracy are provided in Supplementary Information.

In the machine learning Monte Carlo resampling, new configurations $\mathbf{x}_{\text{new}}$ are proposed by sampling velocities $\mathbf{v}$ from a Maxwell–Boltzmann distribution and integrating them for a timestep Δt chosen to be 0.5 fs: $\mathbf{x}_{\text{new}} = \mathbf{x}_{\text{old}} + \mathbf{v}\Delta t$. For the adsorbed system, the molecule is additionally subject to a random translation (up to 0.5 Å) and rotation (up to 35°). The new proposed configuration is then accepted or rejected according to the Metropolis criterion based on the energy predicted by the machine learning model.

Data availability

The geometries of the configurations used to train and test the ML models, the corresponding energies, and the input files for the coupled cluster calculations are available at https://github.com/bslhrzg/mlpt-mlmc.

Code availability

The VASP code is copyrighted software and can be obtained from its official website. The CC4S code is available for download at https://gitlab.cc4s.org/cc4s/cc4s. The codes used to perform MLPT and MLMC calculations are available at https://github.com/bslhrzg/mlpt-mlmc.

Received: 28 July 2023; Accepted: 22 March 2024;
Published online: 04 April 2024

References

1. Kohn, W. & Sham, L. J. Self-consistent equations including exchange and correlation effects. *Phys. Rev.* **140**, A1133–A1138 (1965).
2. Perdew, J. P. & Zunger, A. Self-interaction correction to density-functional approximations for many-electron systems. *Phys. Rev. B* **23**, 5048–5079 (1981).
3. Kohn, W., Meir, Y. & Makarov, D. E. van der Waals energies in density functional theory. *Phys. Rev. Lett.* **80**, 4153 (1998).
4. Cohen, A. J., Mori-Sánchez, P. & Yang, W. Challenges for density functional theory. *Chem. Rev.* **112**, 289–320 (2011).
5. Møller, C. & Plesset, M. S. Note on an approximation treatment for many-electron systems. *Phys. Rev.* **46**, 618 (1934).
6. Bartlett, R. J. & Musial, M. Coupled-cluster theory in quantum chemistry. *Rev. Mod. Phys.* **79**, 291 (2007).
7. Pisani, C. et al. Periodic local mp2 method for the study of electronic correlation in crystals: theory and preliminary applications. *J. Comput. Chem.* **29**, 2113–2124 (2008).
8. Marsman, M., Grüneis, A., Paier, J. & Kresse, G. Second-order Møller–Plesset perturbation theory applied to extended systems. I. Within the projector-augmented-wave formalism using a plane wave basis set. *J. Chem. Phys.* **130**, 184103 (2009).
9. Del Ben, M., Hutter, J. & VandeVondele, J. Second-order Møller–Plesset perturbation theory in the condensed phase: an efficient and massively parallel Gaussian and plane waves approach. *J. Chem. Theory Comput.* **8**, 4177–4188 (2012).
10. Booth, G. H., Grüneis, A., Kresse, G. & Alavi, A. Towards an exact description of electronic wavefunctions in real solids. *Nature* **493**, 365 (2013).
11. Dixit, A., Claudot, J., Lebègue, S. & Rocca, D. Communication: a novel implementation to compute mp2 correlation energies without basis set superposition errors and complete basis set extrapolation. *J. Chem. Phys.* **146**, 211102 (2017).
12. Behler, J. & Parrinello, M. Generalized neural-network representation of high-dimensional potential-energy surfaces. *Phys. Rev. Lett.* **98**, 146401 (2007).
13. Chmiela, S. et al. Machine learning of accurate energy-conserving molecular force fields. *Sci. Adv.* **3**, e1603015 (2017).
14. Chmiela, S., Sauceda, H. E., Müller, K.-R. & Tkatchenko, A. Towards exact molecular dynamics simulations with machine-learned force fields. *Nat. Commun.* **9**, 3887 (2018).
15. Zhang, L., Han, J., Wang, H., Car, R. & Weinan, E. Deep potential molecular dynamics: a scalable model with the accuracy of quantum mechanics. *Phys. Rev. Lett.* **120**, 143001 (2018).
16. Schran, C., Brieuc, F. & Marx, D. Converged colored noise path integral molecular dynamics study of the Zundel cation down to ultralow temperatures at coupled cluster accuracy. *J. Chem. Theory Comput.* **14**, 5068–5078 (2018).
17. Sauceda, H. E., Chmiela, S., Poltavsky, I., Müller, K.-R. & Tkatchenko, A. Molecular force fields with gradient-domain machine learning: Construction and application to dynamics of small molecules with coupled cluster forces. *J. Chem. Phys.* **150**, 114102 (2019).
18. Bogojeski, M., Vogt-Maranto, L., Tuckerman, M. E., Müller, K.-R. & Burke, K. Quantum chemical accuracy from density functional approximations via machine learning. *Nat. Commun.* **11**, 5223 (2020).
19. Smith, J. S. et al. The ani-1ccx and ani-1x data sets, coupled-cluster and density functional theory properties for molecules. *Sci. Data* **7**, 134 (2020).
20. 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.* **129**, 226001 (2022).
21. Chen, M. S. et al. Data-efficient machine learning potentials from transfer learning of periodic correlated electronic structure methods: liquid water at AFQMC, CCSD, and CCSD(T) accuracy. *J. Chem. Theory Comput.* **19**, 4510–4519 (2023).
22. Gruber, T., Liao, K., Tsatsoulis, T., Hummel, F. & Grüneis, A. Applying the coupled-cluster ansatz to solids and surfaces in the thermodynamic limit. *Phys. Rev. X* **8**, 021043 (2018).
23. Irmler, A., Gallo, A. & Grüneis, A. Focal-point approach with pair-specific cusp correction for coupled-cluster theory. *J. Chem. Phys.* **154**, 234103 (2021).
24. Schäfer, T., Gallo, A., Irmler, A., Hummel, F. & Grüneis, A. Surface science using coupled cluster theory via local Wannier functions and in-RPA-embedding: the case of water on graphitic carbon nitride. *J. Chem. Phys.* **155**, 244103 (2021).
25. Liao, K., Shen, T., Li, X.-Z., Alavi, A. & Grüneis, A. Structural and electronic properties of solid molecular hydrogen from many-electron theories. *Phys. Rev. B* **103**, 054111 (2021).

26. Liao, K., Li, X.-Z., Alavi, A. & Grüneis, A. A comparative study using state-of-the-art electronic structure theories on solid hydrogen phases under high pressures. *Npj Comput. Mater.* **5**, 1–6 (2019).
27. Tsatsoulis, T., Sakong, S., Groß, A. & Grüneis, A. Reaction energetics of hydrogen on Si (100) surface: a periodic many-electron theory study. *J. Chem. Phys.* **149**, 244105 (2018).
28. Chipot, C. & Pohorille, A. *Free Energy Calculations: Theory and Applications in Chemistry and Biology* (Springer, 2016).
29. Chehaibou, B., Badawi, M., Bucko, T., Bazhirov, T. & Rocca, D. Computing RPA adsorption enthalpies by machine learning thermodynamic perturbation theory. *J. Chem. Theory Comput.* **15**, 6333–6342 (2019).
30. Bucko, T., Gesvandtnerova, M. & Rocca, D. Ab initio calculations of free energy of activation at multiple electronic structure levels made affordable: An effective combination of perturbation theory and machine learning. *J. Chem. Theory Comput.* **16**, 6049–6060 (2020).
31. Gešvandtnerová, M., Rocca, D. & Bučko, T. Methanol carbonylation over acid mordenite: Insights from ab initio molecular dynamics and machine learning thermodynamic perturbation theory. *J. Catal.* **396**, 166–178 (2021).
32. Herzog, B. et al. Assessing the accuracy of machine learning thermodynamic perturbation theory: Density functional theory and beyond. *J. Chem. Theory Comput.* **18**, 1382–1394 (2022).
33. Bartók, A. P., Kondor, R. & Csányi, G. On representing chemical environments. *Phys. Rev. B* **87**, 184115 (2013).
34. De, S., Bartók, A. P., Csányi, G. & Ceriotti, M. Comparing molecules and solids across structural and alchemical space. *Phys. Chem. Chem. Phys.* **18**, 13754–13769 (2016).
35. Ramakrishnan, R., Dral, P. O., Rupp, M. & von Lilienfeld, O. A. Big data meets quantum chemistry approximations: the δ-machine learning approach. *J. Chem. Theory Comput.* **11**, 2087–2096 (2015).
36. Van Speybroeck, V. et al. Advances in theory and their application within the field of zeolite chemistry. *Chem. Soc. Rev.* **44**, 7044–7111 (2015).
37. Grajciar, L. et al. Towards operando computational modeling in heterogeneous catalysis. *Chem. Soc. Rev.* **47**, 8307–8348 (2018).
38. Jablonka, K. M., Ongari, D., Moosavi, S. M. & Smit, B. Big-data science in porous materials: materials genomics and machine learning. *Chem. Rev.* **120**, 8066–8129 (2020).
39. Grüneis, A. et al. Natural orbitals for wave function based correlated calculations using a plane wave basis set. *J. Chem. Theory Comput.* **7**, 2780–2785 (2011).
40. Cc4s user documentation. https://manuals.cc4s.org/user-manual/.
41. Kresse, G. & Furthmüller, J. Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set. *Phys. Rev. B* **54**, 11169–11186 (1996).
42. Kresse, G. & Hafner, J. Norm-conserving and ultrasoft pseudopotentials for first-row and transition elements. *J. Phys. Condens. Matter* **6**, 8245 (1994).
43. Hummel, F., Tsatsoulis, T. & Grüneis, A. Low rank factorization of the coulomb integrals for periodic coupled cluster theory. *J. Chem. Phys.* **146**, 124105 (2017).
44. Grimme, S. Semiempirical GGA-type density functional constructed with a long-range dispersion correction. *J. Comput. Chem.* **27**, 1787–1799 (2006).
45. Bucko, T., Lebegue, S., Gould, T. & Angyan, J. G. Many-body dispersion corrections for periodic systems: an efficient reciprocal space implementation. *J. Phys. Condens. Matter* **28**, 045201 (2016).
46. Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. & Teller, E. Equation of state calculations by fast computing machines. *J. Chem. Phys.* **21**, 1087–1092 (1953).
47. Sinnokrot, M. O., Valeev, E. F. & Sherrill, C. D. Estimates of the ab initio limit for π-π interactions: the benzene dimer. *J. Am. Chem. Soc.* **124**, 10887–10893 (2002).
48. Takatani, T., Hohenstein, E. G., Malagoli, M., Marshall, M. S. & Sherrill, C. D. Basis set consistent revision of the S22 test set of noncovalent interaction energies. *J. Chem. Phys.* **132**, 144104 (2010).
49. Pham, T. D., Liu, Q. & Lobo, R. F. Carbon dioxide and nitrogen adsorption on cation-exchanged SSZ-13 zeolites. *Langmuir* **29**, 832–839 (2013).
50. Furche, F. & van Voorhis, T. Fluctuation-dissipation theorem density-functional theory. *J. Chem. Phys.* **122**, 164106 (2005).
51. Grüneis, A., Marsman, M., Harl, J., Schimka, L. & Kresse, G. Making the random phase approximation to electronic correlation accurate. *J. Chem. Phys.* **131**, 154115 (2009).
52. Ren, X., Tkatchenko, A., Rinke, P. & Scheffler, M. Beyond the random-phase approximation for the electron correlation energy: The importance of single excitations. *Phys. Rev. Lett.* **106**, 153003 (2011).
53. Olsen, T. & Thygesen, K. S. Extending the random-phase approximation for electronic correlation energies: the renormalized adiabatic local density approximation. *Phys. Rev. B* **86**, 081103 (2012).
54. Bates, J. E. & Furche, F. Communication: random phase approximation renormalized many-body perturbation theory. *J. Chem. Phys.* **139**, 171103 (2013).
55. Dixit, A., Ángyán, J. G. & Rocca, D. Improving the accuracy of ground-state correlation energies within a plane-wave basis set: The electron-hole exchange kernel. *J. Chem. Phys.* **145**, 104105 (2016).
56. Hellgren, M., Colonna, N. & de Gironcoli, S. Beyond the random phase approximation with a local exchange vertex. *Phys. Rev. B* **98**, 045117 (2018).
57. Hummel, F., Grüneis, A., Kresse, G. & Ziesche, P. Screened exchange corrections to the random phase approximation from many-body perturbation theory. *J. Chem. Theory Comput.* **15**, 3223–3236 (2019).
58. Maaten, Lvd & Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).
59. Andersen, H. C. Molecular dynamics simulations at constant pressure and/or temperature. *J. Chem. Phys.* **72**, 2384–2393 (1980).
60. Perdew, J. P., Burke, K. & Ernzerhof, M. Generalized gradient approximation made simple. *Phys. Rev. Lett.* **77**, 3865–3868 (1996).
61. Himanen, L. et al. DScribe: library of descriptors for machine learning in materials science. *Comput. Phys. Commun.* **247**, 106949 (2020).

Acknowledgements

The authors would like to warmly thank Mauricio Chagas da Silva for valuable discussions and technical help. This work was supported through the COMETE project (COnception in silico de Matériaux pour l'EnvironnemenT et l'Energie) co-funded by the European Union under the program "FEDER-FSE Lorraine et Massif des Vosges 2014-2020". B.H., S.L., and D.R. acknowledge the financial support of the Agence Nationale de la Recherche under the Lorraine Artificial Intelligence (LOR-AI) project (grant number ANR-20-THIA-0010-01). This work was granted access to the HPC resources of TGCC under the allocations 2022-A0120810433 and 2023-A0140810433 by GENCI. T.B. acknowledges support from the Slovak Research and Development Agency under Contract No. APVV-20-0127 and the grant VEGA 1/0254/24 from the Ministry of Education Research, Development and Youth of the Slovak Republic. A.G. and A.G. thankfully acknowledge support and funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (Grant Agreement No. 715594).

Author contributions

T.B., A.G., and D.R. designed and led the research; D.R. coordinated the collaboration; B.H. performed the research with supervision by D.R.; M.B. provided resources. All the authors contributed to the discussions and writing of the paper.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41524-024-01249-y.

Correspondence and requests for materials should be addressed to Tomáš Bučko, Andreas Grüneis or Dario Rocca.