Computing RPA adsorption enthalpies by machine learning thermodynamic perturbation theory

Bilal Chehaibou,†‡ Michael Badawi,†‡ Tomáš Bučko,¶§ Timur Bazhirov,∥ and Dario Rocca*,†‡

†Université de Lorraine, LPCT, UMR 7019, 54506 Vandœuvre-lès-Nancy, France
‡CNRS, LPCT, UMR 7019, 54506 Vandœuvre-lès-Nancy, France
¶Department of Physical and Theoretical Chemistry, Faculty of Natural Sciences, Comenius University in Bratislava, Mlynská Dolina, Ilkovičova 6, SK-84215 Bratislava, Slovakia
§Institute of Inorganic Chemistry, Slovak Academy of Sciences, Dúbravská cesta 9, SK-84236 Bratislava, Slovakia
∥Exabyte Inc., San Francisco, California 94103, USA

E-mail: dario.rocca@univ-lorraine.fr

Abstract

Correlated quantum-chemical methods for condensed matter systems, such as the random phase approximation (RPA), hold the promise of reaching a level of accuracy much higher than that of conventional density functional theory approaches. However, the high computational cost of such methods hinders their broad applicability, in particular for finite-temperature molecular dynamics simulations. We propose a method that couples machine learning techniques with thermodynamic perturbation theory to estimate finite-temperature properties using correlated approximations. We apply this
approach to compute the enthalpies of adsorption in zeolites and show that reliable estimates can be obtained by training a machine learning model with as few as 10 RPA energies. This approach paves the way to the broader use of computationally expensive quantum-chemical methods to predict the finite-temperature properties of condensed matter systems.

1 Introduction

Bridging *ab initio* molecular dynamics (AIMD) with correlated quantum-chemical methods could significantly improve the accuracy of materials simulations in realistic finite-temperature conditions. Since it was introduced for the first time, AIMD has been almost exclusively coupled with density functional theory (DFT), which provides a highly efficient numerical scheme to solve the Schrödinger equation. However, the practical approximations used for the exchange-correlation functional are not systematically improvable and are affected by a number of shortcomings. Correlated methods could provide an alternative to DFT in MD simulations possibly reaching the threshold of the chemical accuracy (1 kcal/mol). Traditional quantum chemical methods, such as Møller-Plesset perturbation theory to second order (MP2) or coupled-cluster theory, have been recently implemented for condensed phase applications. In this work we will focus on the random phase approximation (RPA), which is particularly suitable for condensed matter applications and is computationally affordable for systems with a few hundreds of electrons. The RPA is a correlated method that can be derived in the framework of DFT, Green’s functions, and coupled-cluster theory. Successful (zero temperature) applications of this approach include description of properties of liquids, layered materials, bulk solids, and molecules adsorbed on surfaces. The RPA is considered particularly accurate for the description of systems where weak van der Waals (vdW) interactions play a dominant role. In order to improve the description of short range interactions a series of beyond-RPA approaches have been developed based on considerations on the homogeneous electron gas, by using the
adiabatic hybrid functional approximation,\textsuperscript{30} by including single excitations,\textsuperscript{31} or based on exact exchange kernels both in the local\textsuperscript{32–34} and approximate non-local forms.\textsuperscript{35–39}

The main bottleneck in the application of RPA and other correlated methods in AIMD simulations is their prohibitively high computational cost. Accordingly, RPA and MP2 AIMD simulations have been rarely attempted\textsuperscript{17,40} and are by far too expensive to be routinely used in applications that require higher levels of accuracy with respect to traditional density functional theory approximations. In Sec. 3 we will consider examples of computations of adsorption enthalpies in zeolites. Directly performing a RPA AIMD for these systems is estimated to be several tens of millions of CPU hours. In order to compute finite-temperature properties at the RPA level of theory in a simple and numerically efficient way, in this paper we propose the MLPT method, that couples machine learning (ML) techniques with thermodynamic perturbation theory (TPT).

Since the work of Behler and Parrinello (BP) in 2007\textsuperscript{41} the use of ML algorithms in MD simulations is growing in popularity.\textsuperscript{42–45} In the BP approach and its variants, a certain number of configurations of a system is generated by performing a relatively short MD simulation or by properly sampling the configuration space. A ML learning model, usually a neural network (NN), is trained on the energies and/or forces computed for these configurations. A full molecular dynamics (MD) simulation is then performed by using the previously trained neural network. At each step of the MD simulation, instead of directly solving the Schrödinger equation, forces and energies are inexpensively predicted by the ML model and used to generate the configuration in the next step. In this way the timescale accessible by a typical MD run is significantly increased while keeping to a large extent the accuracy of an \textit{ab initio} simulation. However, the training of a ML learning algorithm for a MD simulation typically requires thousands or even tens of thousands of configurations and, accordingly, the BP approach is almost exclusively trained on DFT energies and/or forces. In recent work by Chmiela and coworkers\textsuperscript{43} a method was proposed to decrease the number of configurations by using the symmetries of molecules. In this way highly accurate MD simulations were
performed at the coupled-cluster level of theory. This approach is limited to high-symmetry systems and has never been applied to condensed phase systems, which are the main focus of this work.

Similarly to the BP approach, in this paper we use a ML learning regression model to predict the energy surface from a certain number of training examples. Differently from the BP approach, this model is not used to perform a full MD simulation but it is rather coupled with thermodynamic perturbation theory, whose brute force application also requires an impractically high number of RPA calculations. The corresponding MLPT calculation is executed in the following steps: (1) an AIMD simulation is performed using a numerically affordable DFT functional; (2) a small number $N_{\text{train}}$ of configurations generated by the initial MD is selected and RPA calculations are performed on them; (3) a ML model is trained on the selected configurations to predict the RPA energies for all the remaining configurations; (4) based on TPT, the contributions from the configurations generated at step 1 are reweighted using the predicted energies and RPA ensemble averages are computed. Similar ideas were already proposed in previous work \cite{46,47} but never coupled with machine learning techniques to significantly decrease the required computational time. Specifically, in Ref. 47 the melting temperature of silicon at the RPA level of theory was computed by including a relatively small number (200) of configurations in the perturbative approach. In our previous work on zeolites, \cite{46} the RPA enthalpy of adsorption was computed by using perturbation theory on selected configurations (selPT), which applies TPT on a small number of representative configurations that “span” the full energy probability distribution. The comparison between MLPT and selPT is further discussed in Sec. 3. For the specific objective of this work, the present MLPT approach has the following advantages with respect to BP MD: (1) The ML regression algorithm is exclusively trained on energies; this is useful since most condensed phase implementations of correlated methods do not support the calculation of analytical gradients. (2) The implementation is easier, since the BP method requires also the derivative of the ML model with respect to atomic positions in order to compute forces.
The accuracy of the ML predictions can be enhanced in a straightforward way by using a $\Delta$-machine learning approach, which trains the model on energy differences instead of total energies (see Sec. 2).

In order to establish the accuracy and efficiency of the MLPT methodology we consider the calculation of molecular enthalpies of adsorption in zeolites, which are porous materials used in many applications such as depollution, separation of chemicals, and catalysis. Specifically, four enthalpies are computed by considering the adsorption of CH$_4$ and CO$_2$ in siliceous and protonated chabazite. Results at the RPA level of theory are in excellent agreement with experiments. Importantly, for these systems reliable estimates for the RPA enthalpy of adsorption can be obtained by training the ML model with as few as 10 configurations. The use of a small number of training examples is crucial to limit the number of highly expensive correlated calculations and to make MLPT a practical approach to be routinely used to improve the accuracy of finite-temperature *ab initio* simulations.

The layout of the paper is organized as follows. In the following section, Sec. 2, we discuss thermodynamic perturbation theory and machine learning for materials properties, which are the main ingredients of the MLPT method. In Sec. 3, we present numerical results for molecular enthalpies of adsorption in zeolites. Sec. 4 contains our conclusions.

## 2 Methodological approach

### 2.1 Thermodynamic perturbation theory

Starting from a molecular dynamics driven by a numerically inexpensive approximation, the goal of this work is to use thermodynamic perturbation theory (TPT) to obtain finite-temperature observables at a high level of theory. This methodology is general but for the specific numerical applications of this work the van der Waals corrected PBE+D2 functional will be used as a (cheap) starting point to compute RPA ensemble averages.

A general review of TPT in the context of classical MD can be found e.g. in Refs.
The use of TPT to compute ensemble averages based on correlated methods starting from cheaper DFT approximations have been proposed in Refs. 46,47. Here we will only summarize the main ideas necessary to understand the MLPT method. Molecular dynamics is used to evaluate the finite-temperature expectation values of a certain observable $O$:

$$\langle O \rangle_H = \frac{\int O(q, p) \exp \{-\beta H(q, p)\} dq dp}{\int \exp \{-\beta H(q, p)\} dq dp},$$

(1)

where $\langle \ldots \rangle_H$ is used to compactly denote the average over the canonical ensemble corresponding to the Hamiltonian $H$, $\beta = 1/k_B T$, and $q$ and $p$ denote nuclear positions and momenta, respectively. In Eq. 1 the nuclear motion is driven by the classical Hamiltonian

$$H(q, p) = T(p) + V(q),$$

(2)

where $T(p)$ is the nuclear kinetic energy and $V(q)$ is the potential acting on the nuclei. Here we will focus on AIMD, where the potential $V(q)$ is determined by the solution of the electronic Schrödinger equation within different approximations, usually based on DFT:

$$V(q) = \langle \Psi(q)|\hat{H}_{DFT}|\Psi(q) \rangle + U_{II}(q),$$

(3)

where $U_{II}$ is the repulsive nuclear energy and $\hat{H}_{DFT}$ is the electronic DFT Hamiltonian (not to be confused with the classical Hamiltonian $H$). As explained above, the main goal of this work is to use thermodynamic perturbation theory to compute observables within a sophisticated quantum chemical approximation (e.g. the RPA) starting from a molecular dynamics driven by a computationally “cheap” DFT approximation (e.g. PBE+D2). Below the Hamiltonian $H$ and the corresponding potential $V$ will be associated with the less expensive quantum-mechanical approximation while primed quantities ($H'$ and $V'$) will be
associated with the more expensive approach. By using the identity

$$H'(\mathbf{q}, \mathbf{p}) = H(\mathbf{q}, \mathbf{p}) + V'(\mathbf{q}) - V(\mathbf{q}) = H(\mathbf{q}, \mathbf{p}) + \Delta V(\mathbf{q}),$$  \hspace{1cm} (4)$$

the average of the observable $O(\mathbf{q}, \mathbf{p})$ over the ensemble corresponding to $H'$ can be written as

$$\langle O \rangle_{H'} = \frac{\int O(\mathbf{q}, \mathbf{p}) \exp\{-\beta H'(\mathbf{q}, \mathbf{p})\} \, d\mathbf{q} \, d\mathbf{p}}{\int \exp\{-\beta H'(\mathbf{q}, \mathbf{p})\} \, d\mathbf{q} \, d\mathbf{p}}$$

$$= \frac{\int O(\mathbf{q}, \mathbf{p}) \exp\{-\beta \Delta V(\mathbf{q})\} \exp\{-\beta H(\mathbf{q}, \mathbf{p})\} \, d\mathbf{q} \, d\mathbf{p}}{\int \exp\{-\beta \Delta V(\mathbf{q})\} \exp\{-\beta H(\mathbf{q}, \mathbf{p})\} \, d\mathbf{q} \, d\mathbf{p}}$$

$$= \frac{\langle O \exp\{-\beta \Delta V\} \rangle_H}{\langle \exp\{-\beta \Delta V\} \rangle_H},$$  \hspace{1cm} (5)$$

which provides the main result of TPT used in this work. This equation has important theoretical and practical implications, showing that the ensemble average $\langle O \rangle_{H'}$ can be redefined in terms of an ensemble average corresponding to $H$; this is achieved through the reweighting factor $\exp\{-\beta \Delta V\}$. From a numerical point of view Eq. 5 implies that the statistics sampled by a MD driven by $H$ can be used to obtain finite-temperature expectation values corresponding to $H'$. This is a crucial point since directly performing a MD based on $H'$ would be extremely costly if feasible at all.

From a practical point of view, in order for the perturbative procedure in Eq. 5 to be effective, it is necessary that a MD driven by $H$ would sample a configurational space that is similar to that sampled by an hypothetical $H'$ MD. If the corresponding high probability ensembles are disjoint or only partially superimposed, the perturbative approach will have low accuracy. A detailed discussion of this issue can be found in Refs. 61,62. Following Ref. 46, in this work we suppose that PBE+D2 represents a reliable starting point for RPA.

In this work the TPT formalism will be applied to the specific problem of computing adsorption energies in zeolites at the RPA level of theory. The adsorption enthalpy can be
\[ \Delta_{\text{ads}}H(T) = \langle V'_{S+A} \rangle_{H'} - \left[ \langle V'_{S} \rangle_{H'} + \langle V'_{A} \rangle_{H'} \right] - k_B T, \]  

(6)

where \( S \) denotes the clean substrate (a zeolite in this case) and \( A \) the adsorbate; the values of \( \langle V' \rangle_{H'} \) can be obtained from Eq. 5 for the specific case \( O = V' \). In order to evaluate numerically the ensemble averages in practical numerical calculations, a MD based on \( H \) is first performed and then the ensemble average is computed as

\[ \langle V' \rangle_{H'} \approx \frac{\sum_{i=1}^{N_{\text{tot}}} V'_i \exp \{-\beta (V'_i - V_i)\}}{\sum_{i=1}^{N_{\text{tot}}} \exp \{-\beta (V'_i - V_i)\}} = \frac{\sum_{i=1}^{N_{\text{tot}}} V'_i \exp \{-\beta \Delta V_i\}}{\sum_{i=1}^{N_{\text{tot}}} \exp \{-\beta \Delta V_i\}}, \]  

(7)

where the index \( i \) runs over all the \( N_{\text{tot}} \) configurations generated by the production MD run. Since in the MD simulations considered below \( N_{\text{tot}} = 190000 \), Eq. 7 highlights the limited practical applicability of TPT alone. Indeed, even by considering a smaller subset of uncorrelated configurations, several thousands of calculations of \( V' \) (namely of RPA total energies) are expected to be necessary to reach convergence. This issue can be significantly alleviated by using a ML regression algorithm. Within this approach only few \( V'_i \)'s are explicitly computed and used to train a ML model; all the remaining values up to \( N_{\text{tot}} \) are then predicted at a negligible computational cost.

### 2.2 Machine Learning approach

In order to significantly decrease the number of expensive calculations (i.e. based on the RPA) required to apply thermodynamic perturbation theory, in this work we propose the use of a ML regression model. To this purpose we chose the smooth overlap of atomic positions (SOAP)\(^{63,64}\) as descriptor for the geometric configurations and the kernel ridge regression (KRR)\(^{65}\) as ML algorithm to predict total energies. This choice, that allowed us to reach a satisfactory level of accuracy, was partially driven by a recent comparative study that reported a good performance of SOAP in predicting adsorption energies of hydrogen on nanoclusters.\(^{66}\) Several other descriptors for periodic materials have also been proposed.
in the literature, such as, for example, the many-body tensor representation (MBTR)\textsuperscript{67} or atom-centered symmetry functions (ACSF).\textsuperscript{68} While a systematic comparison of different ML approaches is beyond the purpose of this paper, other descriptors and/or ML algorithms might also perform well for the problem we are considering.

Below we summarize the main ideas of the SOAP and KRR approaches; a more detailed presentation can be found in Refs. 63–65. The numerical implementation of the ML models considered in this work is based on the DScribe libraries.\textsuperscript{69,70}

For a structure $A$ of a given material let us define an environment $X^A_i$ that includes the atoms surrounding a specific atom $i$ within a certain radius. The density of the atoms $\rho$ in the environment $X^A_i$ is defined as a sum of Gaussians

$$\rho_{X^A_i}(r) = \sum_{j \in X^A_i} \exp \left\{ -\frac{(r - x_j)^2}{2\sigma^2} \right\},$$

where the Gaussians of variance $\sigma^2$ are centered at the positions $x_j$ of all the atoms in the environment $X^A_i$ (including the central one). For systems containing different atomic species, as those considered in this work, a different environment has to be defined for each species. The SOAP kernel is then defined as an overlap integral of the densities of different environments, possibly from different structures $A$ and $B$:

$$\tilde{k}(X^A_i, X^B_j) = \int d\hat{R} \left( \int \rho_{X^A_i}(r) \rho_{X^B_j}(\hat{R}r) dr \right)^2,$$

where the first integral is performed over the three-dimensional rotations, as required in order to have a rotationally invariant representation. For practical purposes a normalised version of the kernel in Eq. 9 is typically used:

$$k(X^A_i, X^B_j) = \frac{\tilde{k}(X^A_i, X^B_j)}{\sqrt{\tilde{k}(X^A_i, X^A_i) \tilde{k}(X^B_j, X^B_j)}}.$$ 

This normalized kernel provides a measure of similarity since the maximum value (one) is
attained when the two environments $\mathcal{X}_i^A$ and $\mathcal{X}_j^B$ are identical (independently of rotation and translations of the environment). Because of the specific form of the density of atoms, expressed as an expansion of Gaussians, the integral over the rotations in Eq. 9 can be carried out analytically. This is achieved by using a basis of spherical harmonics and the corresponding formalism is discussed in details in Refs. 63,64.

The kernels introduced in Eqs. 9-10 can be used to compare environments surrounding specific atoms and, in this respect, describe only a local similarity. For the purpose of this work it is necessary to determine the similarity among different structures (all with the same chemical composition) generated by a MD simulation. This global information is included in the full covariance matrix $C(A,B)$, whose elements

$$C_{ij}(A,B) = k(\mathcal{X}_i^A, \mathcal{X}_j^B)$$

contain all the possible pairings of environments belonging to structure $A$ and to structure $B$. Based on the matrix in Eq. 11 different global kernels can be defined. For the purpose of this work the regularized entropy match (REMatch) kernel will be used, that is typically coupled with SOAP descriptors.63,64 The REMatch kernel is defined as

$$K^\gamma(A,B) = \text{Tr} P^\gamma C(A,B),$$

where $P$ is a doubly stochastic matrix that satisfies the condition

$$P^\gamma = \text{arg min}_P \, \sum_{i,j} P_{ij}(1 - C_{ij} + \gamma \ln P_{ij}),$$

and $\gamma$ is a hyperparameter that controls the entropic contribution.

The REMatch kernel in Eq. 12 can then be coupled with the kernel ridge regression ML algorithm. Since our purpose is to predict RPA total energies for the configurations generated by a PBE+D2 MD, a certain number of configurations $N_{\text{train}}$ is selected to train
the ML model (the specific selection criterion used will be discussed later). From these \( N_{\text{train}} \) configurations the REMatch kernel \( K_{\text{train}}^\gamma \) is built and used to train the ML model:

\[
w = (K_{\text{train}}^\gamma + \lambda 1)^{-1} y_{\text{train}} \tag{14}
\]

where \( y_{\text{train}} \) in this case contains the RPA energies corresponding to the configurations in the training set and \( \lambda \) is a regularization hyperparameter used to prevent overfitting. The weights \( w \) determined by Eq. 14 are then used to predict the RPA energies for all the remaining \( N_{\text{predict}} = N_{\text{tot}} - N_{\text{train}} \) configurations in the MD:

\[
y_{\text{predict}} = K_{\text{predict}}^\gamma w, \tag{15}
\]

where \( y_{\text{predict}} \) contains the predicted energies and \( K_{\text{predict}}^\gamma \) is the \( N_{\text{predict}} \times N_{\text{train}} \) REMatch kernel.

As RPA calculations are highly expensive it is crucial to develop a ML model that can be trained with a number of configurations which is as small as possible. Indeed, \( N_{\text{train}} \) explicit RPA calculations are required while the remaining \( N_{\text{predict}} \) are inexpensively predicted by Eq. 15. To this purpose it is useful to use the energy difference \( \Delta V_i = V'_i - V_i \) as a variable for training rather than the absolute energy \( V'_i \). Indeed, the energy difference \( \Delta V_i \) has a much smoother behavior than \( V'_i \) and the error on the prediction can be lowered by one order of magnitude with respect to a ML model that uses total energies. This idea is based on the \( \Delta \)-machine learning approach.\textsuperscript{48} For example, for CH\(_4\) adsorbed in protonated chabazite (see Sec. 3), \( \Delta V_i \) (and, correspondingly, \( V'_i \)) can be predicted with a root mean square error of about 0.4 kcal/mol by using only \( N_{\text{train}} = 10 \) and with a systematic decrease of the error as a function of \( N_{\text{train}} \).
2.3 Statistical error evaluation

The accuracy of a ML model is characterized by a statistical error that can be possibly decreased by increasing the quantity of data used for training. In our specific case the energies \( V'_i \) or, equivalently, the energy differences \( \Delta V_i \) are predicted with a statistical error that has implications on the estimate of the internal energy (Eq. 7) and enthalpy (Eq. 6).

As a starting point for our discussion it is necessary to determine the error involved in the prediction of \( \Delta V_i \) and \( V'_i \). To this purpose the root mean square error (RMSE) of the predicted energies with respect to the exact RPA results is computed for a small test set. This set includes configurations that are reasonably independent and uncorrelated with respect to configurations used for training the ML model (see Sec. 3 for additional details). Since the prediction is not biased (mean error is always well below 0.1 kcal/mol), the RMSE can be interpreted as the standard deviation of the distribution of the residuals in the regression model and will be denoted by \( s \). The RMSE in the prediction of the RPA energies of individual configurations leads to both a systematic and random error in the internal energy expressed by Eq. 7.

To discuss the systematic error we will split the ensemble energy in Eq. 7 into two contributions

\[
\langle V' \rangle_H = \frac{\sum_{i=1}^{N_{tot}} V_i \exp \{-\beta \Delta V_i\}}{\sum_{i=1}^{N_{tot}} \exp \{-\beta \Delta V_i\}} + \frac{\sum_{i=1}^{N_{tot}} \Delta V_i \exp \{-\beta \Delta V_i\}}{\sum_{i=1}^{N_{tot}} \exp \{-\beta \Delta V_i\}} = f_1 + f_2
\]

and we will focus on the last term \( f_2 \) that depends only on \( \Delta V_i \). Because of the statistical error associated with the machine learning model, we suppose that the exact \( \Delta V_i \) is replaced by values \( E \) normally distributed with variance \( s^2 \):

\[
P(E) = K \exp \left\{ -\frac{(E - \Delta V_i)^2}{2s^2} \right\}, \tag{17}
\]

where \( K \) is a normalizing factor. The estimate of \( f_2 \) that keeps into account this distribution
can be written in the following way:

\[
\tilde{f}_2 = \frac{\sum_i \int dE \ E \ exp \{-\beta E\} \times \ exp \left\{-\frac{(E-\Delta V_i)^2}{2s^2}\right\}}{\sum_i \int dE \ exp \{-\beta E\} \times \ exp \left\{-\frac{(E-\Delta V_i)^2}{2s^2}\right\}}
\]

\[
= \frac{\sum_i \int dE \ exp \{-\beta \Delta V_i\} \times \ exp \left\{-\frac{(E-\Delta V_i-\beta s)^2}{2s^2}\right\}}{\sum_i \int dE \ exp \{-\beta \Delta V_i\} \times \ exp \left\{-\frac{(E-\Delta V_i)^2}{2s^2}\right\}}
\]

\[
= \frac{\sum_i \int dE' \ exp \{-\beta \Delta V_i\} \times \ exp \left\{-\frac{(E'-\Delta V_i)^2}{2s^2}\right\}}{\sum_i \int dE' \ exp \{-\beta \Delta V_i\} \times \ exp \left\{-\frac{(E'-\Delta V_i)^2}{2s^2}\right\}}
\]

\[
= f_2 + \beta s^2,
\]  

(18)

where the sum over \(i\) is implicitly intended between 1 and \(N_{tot}\). Here \(f_2\) is the reweighted average obtained using the exact energy differences \(\Delta V_i\) while \(\tilde{f}_2\) is based on the values of \(\Delta V_i\) predicted by a ML algorithm that introduces a random statistical error. Accordingly, when ML and TPT are coupled, a systematic error of \(\beta s^2\) in the prediction of ensemble averages has to be expected. In the assumption that the residuals are normally distributed and that the variance \(s^2\) is known, this bias can be easily evaluated using Eq. 18. The results that will be discussed below have all been corrected for this systematic error. The value of \(s\) and, consequently, of \(\beta s^2\) decreases by increasing the number of training examples \(N_{\text{train}}\). For example, if we consider \(\text{CH}_4\) adsorbed in protonated chabazite (see Sec. 3 and in particular Table 1) the \(\beta s^2\) correction is 0.33 kcal/mol for \(N_{\text{train}}=10\) and 0.07 kcal/mol for \(N_{\text{train}}=200\). By the same logic, it is easy to see that \(\tilde{f}_1 = f_1\) and hence the first term on the right hand side of Eq. 16 is unaffected by the systematic error arising from \(s\).

We now consider the evaluation of the random error in the ensemble averages associated with the ML model. To this purpose it is important to notice that Eq. 7 is a non linear function of \(V'_i\) and, accordingly, an error propagation formula based on first derivatives will
be used here. We have that the standard error $\delta_{ML}$ in the ensemble average is

$$\delta_{ML} = \sqrt{\sum_{i=1}^{N_{tot}} \left( \frac{\partial \langle V' \rangle_{H'}^i}{\partial V'_i} \right)^2 s_{V'_i}^2} = s \sqrt{\sum_{i=1}^{N_{tot}} \left( \frac{\partial \langle V' \rangle_{H'}^i}{\partial V'_i} \right)^2}$$

(19)

where $s_{V'_i} = s$ is the standard deviation of the $V'_i$'s distribution and the derivative is explicitly given by

$$\frac{\partial \langle V' \rangle_{H'}^i}{\partial V'_i} = w_i (1 - \beta V'_i + \beta \langle V' \rangle_{H'}^i),$$

(20)

where we defined the weight

$$w_i = \frac{\exp \{-\beta (V'_i - V_i)\}}{\sum_{i=1}^{N_{tot}} \exp \{-\beta (V'_i - V_i)\}}.$$

(21)

In all calculations discussed further, $\delta_{ML}$ is typically much smaller than $s$.

Finally, we discuss the source of random error that is associated with the finite sample size. This error is not related to the ML approach but is typical of MD simulations, which span a finite time interval. As the term $\langle V' \rangle_{H'}$ in Eq. 7 can be considered as a weighted average of the form

$$\langle V' \rangle_{H'} = \sum_{i=1}^{N_{tot}} V'_i w_i$$

(22)

with the weight $w_i$ defined in Eq. 21, the corresponding statistical error can be estimated using the standard formula

$$\delta_s = \sqrt{\frac{N_s}{(N_s - 1)} \sum_{i=1}^{N_s} w_i^2 (V'_i - \langle V' \rangle_{H'}^i)^2},$$

(23)

where $N_s \ll N_{tot}$ is the number of uncorrelated samples. In practice, the latter can be identified e.g. by the block method of Flyvbjerg and Petersen. The value of $\delta_s$ depends on the length and the quality of the MD trajectory and it can, in principle, be made arbitrarily small by performing a sufficiently long simulation.
In our numerical examples discussed below, the error bars are evaluated as

\[ \delta = \sqrt{\delta_{ML}^2 + \delta_s^2} \]

where \( \delta_{ML} \) is defined in Eq. 19 and \( \delta_s \) in Eq. 23. The values of \( \delta_s \) are in all cases discussed in Sec. 3 significantly larger than those of \( \delta_{ML} \) and hence the former term represents a dominant contribution to \( \delta \).

### 3 Numerical results

The performance of the MLPT approach will be demonstrated on four specific applications involving CH\(_4\) and CO\(_2\) in protonated chabazite (denoted as HCHAB) and siliceous chabazite (CHAB). The four corresponding adsorbed systems will be indicated as HCHAB+CH\(_4\), HCHAB+CO\(_2\), CHAB+CH\(_4\), and CHAB+CO\(_2\). The HCHAB+CH\(_4\) example will be discussed in more detail to describe the numerical properties of the MLPT approach. The corresponding periodic model is shown in Fig. 1. While the paper is mainly focused on RPA properties, in Appendix A we consider an additional numerical application of MLPT to compute the PBE internal energy of HCHAB+CH\(_4\) starting from a PBE+D2 MD. For the PBE functional\(^{60}\) an exact value of the internal energy can be inexpensively obtained from a brute force application of TPT (Eq. 7) on all the configurations \( N_{tot} \). This exact result is used as a reference to better understand the convergence properties of MLPT and the role played by the systematic error correction in Eq. 18.

Following Ref. 46 we used MD trajectories based on the vdW corrected PBE+D2 functional\(^{59,60}\) as a starting point for the MLPT method. The VASP code\(^{74}\) was used to perform MD simulations in NVT ensemble. Temperature was set to 300 K and the length of trajectories was fixed at 100 ps with an integration step of 0.5 fs. In a production period succeeding the equilibration length of 5 ps, a total of 190,000 configurations were generated. RPA calculations are based on the implementation in the VASP code.\(^{20}\) All remaining computational
The first step of the MLPT is the selection of a proper training set with $N_{\text{train}}$ elements. Since configurations along a MD trajectory are time correlated, in this work we select configurations that are as distanced as possible in time. For example, $N_{\text{train}}=10$ indicates that out of the 190,000 configurations one configuration every 19,000 is selected to obtain a total of 10 training structures. Five training sets are considered here with $N_{\text{train}}=10, 25, 50, 100$, and 200. Additionally, a test set of 20 configurations is selected to estimate the RMSE in the prediction of the machine learning model. Those configurations are chosen randomly but with the constraint of being at a distance of at least 350 time steps from the geometries contained in the largest training set ($N_{\text{train}}=200$). In this way the test and training sets are reasonably independent. For the HCHAB+CH$_4$ system the RMSE error (also denoted with $s$) in the prediction of the RPA energies of configurations is shown in Table 1. The RMSE clearly decreases by increasing the number of training examples $N_{\text{train}}$.

Since configurations are selected using time separation as a criterion it is also important to understand to which extent they span the geometric space. This is particularly important for the adsorbed systems, where the molecule can move rather freely in the cage of the zeolite.
Table 1: RMSE (kcal/mol) in the prediction of RPA energies of the configurations of protonated chabazite (HCHAB), CH₄, and adsorbed system (HCHAB+CH₄) as a function of the number of training structures N_{train}. The RMSE is estimated on a test set of 20 configurations (see text).

<table>
<thead>
<tr>
<th>N_{train}</th>
<th>HCHAB+CH₄</th>
<th>HCHAB</th>
<th>CH₄</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>0.45</td>
<td>0.34</td>
<td>0.03</td>
</tr>
<tr>
<td>25</td>
<td>0.36</td>
<td>0.30</td>
<td>0.01</td>
</tr>
<tr>
<td>50</td>
<td>0.29</td>
<td>0.26</td>
<td>0.01</td>
</tr>
<tr>
<td>100</td>
<td>0.24</td>
<td>0.23</td>
<td>0.01</td>
</tr>
<tr>
<td>200</td>
<td>0.21</td>
<td>0.20</td>
<td>0.01</td>
</tr>
</tbody>
</table>

If certain parts of the configurational space do not have representative structures included in the training set, the ML learning model might have limited predictive power in these regions. Fig. 2 helps visualizing the selected configurations for HCHAB+CH₄ using the t-Distributed Stochastic Neighbor Embedding algorithm (t-SNE).

75 t-SNE is an unsupervised learning method that allows for a dimensionality reduction to visualize high-dimensional data: Points that are closer in the high-dimensional space have higher probability to be closer in the reduced two-dimensional space used for the plot. In order to generate Fig. 2, t-SNE is used to visualize the SOAP descriptors generated by considering only the carbon environment of CH₄ in HCHAB+CH₄. In this way the analysis is focused on the diffusion of the molecule in the zeolite rather than on the full system, that contains many more environments and degrees of freedom. Fig. 2(b) shows the configurations in the N_{train} = 200 training set with respect to the total number of configurations. The sampling is uniform and all the regions of the configuration space are fairly well represented by the training set. For N_{train} = 25 (Fig. 2(a)) the sampling is less satisfactory but, as shown below, the model can still be predictive. This is likely to be related to the use of the ∆-ML approach. Indeed, the model is trained on energy differences ∆V and a large contribution to the RPA total energy is already provided by the PBE+D2 total energy, whose exact value is available from the initial MD trajectory.

Detailed MLPT numerical results for CH₄ adsorbed in protonated chabazite are presented in Table 2, where values for the RPA ensemble averages of each system (CH₄, protonated
Figure 2: For HCHAB+CH$_4$ the dimensionality reduction t-SNE algorithm is used to visualize the configurations selected to train the machine learning model (in red) with respect to the total set of configurations (in blue). (a) $N_{\text{train}} = 25$; (b) $N_{\text{train}} = 200$. 
chabazite, and adsorbed system) and enthalpies of adsorption (in kcal/mol) are reported. Results are rather stable with respect to the $N_{\text{train}}$ variation and all the values of internal energies are well within a 0.2 kcal/mol interval. As few as 10 configurations can already provide highly accurate estimates of the RPA enthalphy of adsorption. As explained in Sec. 2, these values have been corrected for the systematic error $\beta s^2$ (see Eq. 18). This correction ranges from 0.33 kcal/mol ($N_{\text{train}}=10$) to 0.07 kcal/mol ($N_{\text{train}} = 200$). The table reports also error bars $\pm \delta$ associated with random errors Eq. 24. The values of $\delta$ are rather insensitive to the change of $N_{\text{train}}$; this is mostly due to the fact that the finite sampling component $\delta_s$, which is independent of $N_{\text{train}}$, provides a larger contribution than $\delta_{\text{ML}}$.

Table 2: RPA internal energies and enthalpy of adsorption (kcal/mol) of protonated chabazite (HCHAB), CH$_4$, and adsorbed system (HCHAB+CH$_4$) as estimated by MLPT for different values of $N_{\text{train}}$.

<table>
<thead>
<tr>
<th>$N_{\text{train}}$</th>
<th>HCHAB+CH$_4$</th>
<th>HCHAB</th>
<th>CH$_4$</th>
<th>$\Delta_{\text{ads}} H$</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>-21398.01 ± 0.19</td>
<td>-20213.82 ± 0.13</td>
<td>-1181.05 ± 0.02</td>
<td>-3.74 ± 0.23</td>
</tr>
<tr>
<td>25</td>
<td>-21398.13 ± 0.21</td>
<td>-20213.94 ± 0.13</td>
<td>-1181.07 ± 0.02</td>
<td>-3.72 ± 0.25</td>
</tr>
<tr>
<td>50</td>
<td>-21398.17 ± 0.21</td>
<td>-20213.99 ± 0.13</td>
<td>-1181.06 ± 0.02</td>
<td>-3.72 ± 0.25</td>
</tr>
<tr>
<td>100</td>
<td>-21398.13 ± 0.21</td>
<td>-20213.93 ± 0.12</td>
<td>-1181.06 ± 0.02</td>
<td>-3.74 ± 0.24</td>
</tr>
<tr>
<td>200</td>
<td>-21398.15 ± 0.20</td>
<td>-20213.95 ± 0.12</td>
<td>-1181.06 ± 0.02</td>
<td>-3.74 ± 0.23</td>
</tr>
</tbody>
</table>

Table 3: RPA internal energies and enthalpy of adsorption (kcal/mol) of protonated chabazite (HCHAB), CH$_4$, and adsorbed system (HCHAB+CH$_4$) as estimated by a brute force application of thermodynamic perturbation theory for a small number of configurations $N_{\text{conf}}$. Differently from Table 2, the ML approach is not applied in this case, causing thus a dramatic deterioration of convergence.

<table>
<thead>
<tr>
<th>$N_{\text{conf}}$</th>
<th>HCHAB+CH$_4$</th>
<th>HCHAB</th>
<th>CH$_4$</th>
<th>$\Delta_{\text{ads}} H$</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>-21401.99</td>
<td>-20213.75</td>
<td>-1180.97</td>
<td>-7.88</td>
</tr>
<tr>
<td>25</td>
<td>-21398.81</td>
<td>-20214.22</td>
<td>-1181.27</td>
<td>-3.91</td>
</tr>
<tr>
<td>50</td>
<td>-21399.19</td>
<td>-20214.20</td>
<td>-1181.24</td>
<td>-4.34</td>
</tr>
<tr>
<td>100</td>
<td>-21398.06</td>
<td>-20214.56</td>
<td>-1181.32</td>
<td>-2.78</td>
</tr>
<tr>
<td>200</td>
<td>-21398.41</td>
<td>-20214.56</td>
<td>-1181.16</td>
<td>-3.28</td>
</tr>
</tbody>
</table>

In order to highlight the importance of the ML approach in providing converged values for RPA internal energies and enthalpies, in Table 3 we present the results that would be obtained by applying thermodynamic perturbation theory only on the small set of training
configurations. Namely, in this case no ML model is used and, for example, $N_{\text{conf}} = 100$ means that TPT includes only the 100 configurations computed explicitly. Since configurations along a MD trajectory are time correlated, it is reasonable to expect convergence with a relatively small subset of configurations, as assumed in Ref. 47. However, as shown in Table 3 this is not the case for the sets considered in this work. For example, even the largest set $N_{\text{conf}} = 200$ still predicts results for the enthalpy of adsorption that deviate by about 0.5 kcal/mol with respect to the most converged MLPT values in Table 2. Importantly, the enthalpies $\Delta_{\text{ads}}H$ in Table 3 strongly oscillate and no convergence can be inferred from these results.

Finally, in Table 4 we summarize the RPA enthalpies of adsorption (in kcal/mol) as functions of $N_{\text{train}}$ for the all four systems considered in this work. The RPA, which is often considered a reference method for periodic condensed matter applications, provides systematically accurate estimates within chemical accuracy in comparison to experiments. The tendency to slightly underestimate is a well known behavior of the RPA.\textsuperscript{31} Enthalpies obtained from the PBE+D2 MD are instead less consistent, with large deviations for the adsorption in HCHAB and more reliable results for CHAB. Importantly, RPA results with $N_{\text{train}} = 10$ can be obtained with a CPU time comparable to or even shorter than the underlying PBE+D2 MD but can provide a significantly higher level of accuracy and more reliable results for different types of systems.

The HCHAB+CH$_4$ and HCHAB+CO$_2$ systems were already considered in our previous work on the selPT method.\textsuperscript{46} The best estimates of the RPA total energies computed using both methods are similar, the differences being typically within the indicated error bars. For the adsorbed system HCHAB+CH$_4$ the best selPT estimate of the internal energy (180 configurations) is $-21398.50 \pm 0.37$ kcal/mol, to be compared with $-21398.15 \pm 0.20$ kcal/mol of MLPT ($N_{\text{train}} = 200$). For HCHAB+CO$_2$ the agreement is slightly worse, with a selPT value of $-21946.49 \pm 0.35$ kcal/mol and a MLPT value of $-21945.92 \pm 0.32$ kcal/mol; however, since the enthalpy of adsorption for this system is larger, a larger deviation can also be
tolerated. For clean protonated chabazite and the gas phase molecules selPT and MLPT provide even closer agreement in internal energies (within 0.23 kcal/mol).

By design, the MLPT method should be superior in performance to selPT. selPT exploits only the information contained in the energy probability distribution to select a small set of configurations and to use them to compute RPA ensemble averages. MLPT takes advantage of the information contained in both the geometric structure of the configurations (through the SOAP descriptor and the REMatch kernel) and the energy (by predicting energy differences through the $\Delta - ML$ method). Importantly, MLPT shows a very fast convergence of internal energies with the size of the training set and even very small values of $N_{\text{train}}$ (10 or 25) yield very accurate predictions.

Table 4: PBE+D2, RPA, and experimental enthalpies of adsorption (kcal/mol) of CH$_4$ and CO$_2$ adsorbed in protonated (HCHAB) and siliceous (CHAB) chabazite. RPA results are provided as a function of $N_{\text{train}}$.

<table>
<thead>
<tr>
<th>$\Delta_{\text{ads}} H$</th>
<th>HCHAB+CH$_4$</th>
<th>HCHAB+CO$_2$</th>
<th>CHAB+CH$_4$</th>
<th>CHAB+CO$_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>PBE+D2</td>
<td>-6.09</td>
<td>-9.72</td>
<td>-4.64</td>
<td>-5.14</td>
</tr>
<tr>
<td>RPA ($N_{\text{train}} = 10$)</td>
<td>-3.74 ± 0.23</td>
<td>-8.06 ± 0.44</td>
<td>-3.22 ± 0.19</td>
<td>-4.57 ± 0.25</td>
</tr>
<tr>
<td>RPA ($N_{\text{train}} = 25$)</td>
<td>-3.72 ± 0.25</td>
<td>-7.85 ± 0.37</td>
<td>-3.19 ± 0.19</td>
<td>-4.48 ± 0.22</td>
</tr>
<tr>
<td>RPA ($N_{\text{train}} = 50$)</td>
<td>-3.72 ± 0.25</td>
<td>-7.93 ± 0.38</td>
<td>-3.32 ± 0.19</td>
<td>-4.66 ± 0.21</td>
</tr>
<tr>
<td>RPA ($N_{\text{train}} = 100$)</td>
<td>-3.74 ± 0.24</td>
<td>-8.08 ± 0.34</td>
<td>-3.41 ± 0.19</td>
<td>-4.73 ± 0.21</td>
</tr>
<tr>
<td>RPA ($N_{\text{train}} = 200$)</td>
<td>-3.74 ± 0.23</td>
<td>-8.01 ± 0.32</td>
<td>-3.36 ± 0.19</td>
<td>-4.65 ± 0.21</td>
</tr>
<tr>
<td>Exp.</td>
<td>-4.06$^a$</td>
<td>-8.41$^b$</td>
<td>-4.09$^c$</td>
<td>-5.02$^c$</td>
</tr>
</tbody>
</table>

$^a$ Refs. 76  
$^b$ Ref. 53  
$^c$ Ref. 54

4 Conclusions

We proposed the MLPT method that couples machine learning techniques with thermodynamic perturbation theory to efficiently evaluate finite-temperature properties using the RPA or other correlated quantum-chemical methods. By computing enthalpies of adsorption of small molecules in zeolites, we showed that the RPA provides consistently accurate
results compared to experiments. Importantly, as few as 10 RPA energies used to train a ML model provide a satisfactory level of numerical convergence. Future work will be aimed at further enhancing accuracy and efficiency of MLPT. For example, this will involve: (1) The development of algorithms for a more uniform sampling of the configuration space to avoid redundant information in the training set (i.e. configurations that are too close in the descriptor space). (2) Tuning the parameters of the DFT functional used for the initial MD simulation to obtain an improved starting point for perturbation theory; these parameters could be the C_6 coefficients for PBE+D2 or the percentage of exact exchange for a hybrid functional. (3) Using beyond-RPA methods or other more traditional quantum chemistry approaches to further improve the accuracy of thermodynamic calculations. The development of MLPT opens the path to the broader use of high-accuracy/high-cost ab initio methodologies to achieve a new level of predictive ability in materials simulations.

5 Acknowledgements

The authors thank Lauri Himanen and Chris Chipot for helpful discussions. This work was supported by Agence Nationale de la Recherche under grant number ANR-15-CE29-0003-01. B.C., M.B., and D.R. acknowledge financial support through the COMETE project (COncention 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”. T. Bučko acknowledges support from Slovak Research and Development Agency under the Contracts No. APVV-15-0105 and No. VEGA-1/0777/19. Part of the results of this research have been achieved using GENCI-CCRT/CINES computational resources under grant No. A0050805106 and No. A0060910433, and the Computing Center of the Slovak Academy of Sciences acquired in projects ITMS 26230120002 and 26210120002 supported by the Research and Development Operational Program funded by the ERDF. The Exabyte.io platform and parts of its infrastructure hosted on the Google Cloud Platform were used to enable RPA
calculations on up to several hundreds of structural configurations in parallel.

A  PBE internal energy from a PBE+D2 molecular dynamics

In this appendix we consider the application of MLPT to compute the PBE internal energy of HCHAB+CH$_4$ starting from a PBE+D2 molecular dynamics. This discussion is important to further understand the numerical properties of MLPT and validate this method against an exact reference result. Since PBE energies can be obtained as a byproduct of a PBE+D2 MD simulation, a brute force application of TPT (Eq. 7) including all the 190000 configurations is trivial and comes without any additional numerical cost (for the RPA several tens of millions of CPU hours would be required). This exact reference value is highlighted in bold in Table 5. In the same TPT column it is shown that the inclusion of a small number of configurations $N_{conf}$ in the perturbative approach leads to errors of several kcal/mol with respect to the reference internal energy. In contrast, when the $\beta s^2$ correction for the systematic error of Eq. 18 is included (second column) the MLPT approach shows an impressive level of accuracy, which does not seem to be influenced by the value of $N_{train}$ (at least for $N_{train} \geq 10$). This observation might seem counterintuitive since a ML approach is expected to systematically improve by increasing the number of training examples. However, the $\beta s^2$ term compensates almost exactly for the lower accuracy of the ML approach when small values of $N_{train}$ are employed. Indeed, if the systematic error is not corrected for (third column of Table 5) the MLPT results converge to the exact reference value only by increasing $N_{train}$, as it would be normally expected for a ML approach. Importantly, since $\beta s^2$ decreases as a function of $N_{train}$, MLPT with or without systematic error correction converges to the same result (for $N_{train} = 50$ this difference is already as small as 0.03 kcal/mol). All the RPA values presented in the main text include the $\beta s^2$ correction. In Table 5 it can be noticed that by decreasing $N_{train}$ from 10 to 5 the accuracy of the MLPT (no correction)
approach deteriorates sharply: The deviation with respect to the exact reference increases from 0.37 kcal/mol to 1.4 kcal/mol. For $N_{\text{train}}=5$ the number of configurations is clearly insufficient to train the ML model and also the $\beta s^2$ correction is of limited use. We can conclude that at least 10 training configurations should be included in order to achieve a reasonable convergence of the internal energies of the systems considered in this work.

Table 5: HCHAB+CH$_4$ internal energy (in kcal/mol) at the PBE level of theory as computed using MLPT with the $\beta s^2$ correction (Eq. 18), MLPT without this correction, and from the brute-force application of thermodynamic perturbation theory (TPT). $N_{\text{train}}$ refers to MLPT and $N_{\text{conf}}$ to TPT. The value in bold is the exact reference obtained by applying TPT on all the configurations generated by the initial molecular dynamics.

<table>
<thead>
<tr>
<th>$N_{\text{train}}/N_{\text{conf}}$</th>
<th>MLPT $\beta s^2$ correction</th>
<th>MLPT no correction</th>
<th>TPT no ML model</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>-7129.92</td>
<td>-7127.64</td>
<td>-7130.68</td>
</tr>
<tr>
<td>10</td>
<td>-7128.98</td>
<td>-7128.68</td>
<td>-7133.69</td>
</tr>
<tr>
<td>25</td>
<td>-7128.85</td>
<td>-7128.73</td>
<td>-7122.89</td>
</tr>
<tr>
<td>50</td>
<td>-7129.02</td>
<td>-7128.99</td>
<td>-7126.23</td>
</tr>
<tr>
<td>100</td>
<td>-7129.04</td>
<td>-7129.03</td>
<td>-7126.54</td>
</tr>
<tr>
<td>200</td>
<td>-7129.02</td>
<td>-7129.02</td>
<td>-7127.84</td>
</tr>
<tr>
<td>190000</td>
<td>/</td>
<td>/</td>
<td>-7129.05</td>
</tr>
</tbody>
</table>

References


(40) Bokdam, M.; Lahnsteiner, J.; Ramberger, B.; Schäfer, T.; Kresse, G. Assessing density


(49) Speybroeck, V. V.; Hemelsoet, K.; Joos, L.; Waroquier, M.; Bell, R. G.; Catlow, C.
R. A. Advances in theory and their application within the field of zeolite chemistry.  


(51) Bučko, T.; Chibani, S.; Paul, J.-F.; Cantrel, L.; Badawi, M. Dissociative iodomethane adsorption on Ag-MOR and the formation of AgI clusters: an ab initio molecular dynamics study.  


(54) Maghsoudi, H.; Soltanieh, M.; Bozorgzadeh, H.; Mohamadalizadeh, A. Adsorption isotherms and ideal selectivities of hydrogen sulfide and carbon dioxide over methane for the Si-CHA zeolite: comparison of carbon dioxide and methane adsorption with the all-silica DD3R zeolite.  

(55) Khalil, I.; Jabraoui, H.; Maurin, G.; Lebègue, S.; Badawi, M.; Thomas, K.; Maugé, F. Selective capture of phenol from biofuel using protonated faujasite zeolites with different Si/Al ratios.  


(70) The DScribe libraries are available at https://github.com/SINGROUP/dscribe.


(77) Bazhirov, T. Data-centric online ecosystem for digital materials science. *arxiv.org/abs/1902.10838* 2019,
Graphical TOC Entry

Random phase approximation
Thermodynamic perturbation theory
Machine learning
MLPT
Chemically accurate enthalpies of adsorption