



HAL
open science

Chemically accurate simulations by machine learning correlated approximations

Basile Herzog

► **To cite this version:**

Basile Herzog. Chemically accurate simulations by machine learning correlated approximations. Condensed Matter [cond-mat]. Université de Lorraine, 2023. English. NNT : 2023LORR0126 . tel-04473287

HAL Id: tel-04473287

<https://hal.univ-lorraine.fr/tel-04473287>

Submitted on 22 Feb 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



**UNIVERSITÉ
DE LORRAINE**

**BIBLIOTHÈQUES
UNIVERSITAIRES**

AVERTISSEMENT

Ce document est le fruit d'un long travail approuvé par le jury de soutenance et mis à disposition de l'ensemble de la communauté universitaire élargie.

Il est soumis à la propriété intellectuelle de l'auteur. Ceci implique une obligation de citation et de référencement lors de l'utilisation de ce document.

D'autre part, toute contrefaçon, plagiat, reproduction illicite encourt une poursuite pénale.

Contact bibliothèque : ddoc-theses-contact@univ-lorraine.fr
(Cette adresse ne permet pas de contacter les auteurs)

LIENS

Code de la Propriété Intellectuelle. articles L 122. 4

Code de la Propriété Intellectuelle. articles L 335.2- L 335.10

http://www.cfcopies.com/V2/leg/leg_droi.php

<http://www.culture.gouv.fr/culture/infos-pratiques/droits/protection.htm>



UNIVERSITÉ
DE LORRAINE



Laboratoire
 $\phi\alpha$
Physique et Chimie
Théoriques



C2MP

THÈSE

réalisée au Laboratoire de Physique et Chimie Théoriques
présentée et soutenue publiquement le 30 août 2023
pour l'obtention du titre de

DOCTEUR EN PHYSIQUE

par

BASILE HERZOG

CHEMICALLY ACCURATE SIMULATIONS
BY MACHINE LEARNING CORRELATED
APPROXIMATIONS

Jury composé de :

Émilie Gaudry	Professeur, IJL, Université de Lorraine	Présidente du jury
Sébastien Lebègue	Directeur de recherche, LPCT, Université de Lorraine	Directeur de thèse
Julien Toulouse	Maître de conférences, LCT, Sorbonne Université	Rapporteur
Fabio Pietrucci	Maître de conférences, IMPMC, Sorbonne Université	Rapporteur
Emmanuel Fromager	Professeur, LCQ, Université de Strasbourg	Examineur

Thèse réalisée par Basile Herzog entre le 1er octobre 2020 et le 1er octobre 2023
au Laboratoire de Physique et Chimie Théorique de Nancy
sous la supervision de Dario Rocca et Sébastien Lebègue

ABSTRACT

The ability to systematically compute physical properties at chemical accuracy would be of tremendous help for efficient *in silico* material discovery, drug design or chemical reaction pathways study. While density functional theory has become the workhorse of materials simulations, the quality of results unfortunately often varies depending on the specific choice of the exchange–correlation functional, and this significantly limits the predictive power of this approach. More correlated approaches, such as a coupled cluster theory, the random phase approximation, or configuration interaction, provide more reliable results in a systematically improvable way. Their computational cost however render them inaccessible in a lot of applications, and this is particularly true for finite temperature properties, where many samples calculations are necessary to reproduce ensemble averages. Machine learning (ML) on the other, has proven its ability to efficiently reduce computational costs by using knowledge from previous calculations. The aim of this thesis is to contribute to the on-going development of ML assisted methods to obtain chemically accurate calculations. Machine learning perturbation theory (MLPT) is a recent promising approach capable of obtaining chemically accurate finite temperature properties by producing ensemble property estimates for highly expensive computational methods. This is achieved by learning the energy difference between two methods, using few samples from a reference statistics produced from a computationally feasible level of theory, and reweighting the later statistics to the computationally inaccessible target method, with thermodynamic perturbation theory (TPT). In this thesis, the MLPT method is tested against possible limitations of TPT, when a poor overlap between reference and target configurational space produce biased target estimates. A diagnostic test for this problem is proposed, together with a solution in the form of Monte Carlo resampling using the same ML model. MLPT is further applied to compute a coupled cluster estimate of the adsorption enthalpy of carbon dioxide inside a periodic zeolite. Finally, on another topic, a new method is presented, in the framework of configuration interaction (CI), to efficiently iteratively sample new Slater determinants using a generative ML model. This algorithm, called CIGen, is shown to be competitive or outperform other Monte Carlo and ML approaches to the CI problem.

REMERCIEMENTS

Je souhaite ici remercier différentes personnes qui m'ont soutenu, aidé, et appris durant ces dernières années. Dario Rocca tout d'abord qui m'a accueilli sur ce sujet de thèse. Grazie mille per il vostro sostegno, la vostra pazienza, il vostro incoraggiamento e la libertà che mi avete dato. È stato un piacere averla come superviseur. Je remercie Sébastien Lebègue d'avoir accepté de me superviser après le départ de Dario. Merci infiniment pour ta confiance et pour ton aide. À Bastien, Philippe et Aurélia, merci pour tous ces bons moments ici, je suis très heureux de vous avoir rencontré et de pouvoir vous appeler mes amis ; tout comme Maurício muito obrigado por sua ajuda e conhecimento (não sei se posso me considerar um programador, mas agora consigo compilar algumas coisas) e pelos ótimos momentos que passei com você. Elisabeth merci de m'avoir écouté et encouragé, Guillaume pour tes appels et tes conseils. Je tiens à remercier également ma famille (et Léo) pour leur soutien. Léa, mille mercis, tu as tout rendu plus facile.

CONTENTS

1	INTRODUCTION	1
1.1	Chemical accuracy	1
1.2	Prior work	2
1.3	Outline	8
2	ELECTRONIC STRUCTURE METHODS	9
2.1	Introduction	9
2.2	Wave function-based methods	10
2.2.1	The many electron problem	10
2.2.2	Hartee-Fock and the correlation energy	14
2.2.3	Configuration Interaction	18
2.2.4	Perturbation theory expansion of the correlation energy	19
2.2.5	Møller-Plesset Perturbation Theory	20
2.2.6	Coupled Cluster	22
2.2.7	Basis sets	25
2.2.8	Methods comparison	26
2.3	Density functional theory	27
2.3.1	Hohenberg-Kohn theorems	27
2.3.2	Kohn-Sham Density functional theory	29
2.3.3	Exchange correlation	33
2.3.4	Approximations to the exchange correlation energy	36
2.3.5	Random Phase Approximation	38
2.3.6	Methods Comparison continued	39
2.4	Finite temperature properties	41
2.4.1	<i>Ab initio</i> Molecular Dynamics	41
2.4.2	Monte Carlo sampling	43
2.4.3	Thermodynamic perturbation theory	43
3	MACHINE LEARNING METHODS	46
3.1	Introduction	46
3.2	Physical descriptors	47
3.3	Kernel Ridge Regression	48
3.4	Dimensionality reduction	51
3.5	Generative models: Restricted Boltzmann Machine	52
4	ASSESSING THE ACCURACY OF MACHINE LEARNING THERMODYNAMIC PERTURBATION THEORY: DENSITY FUNCTIONAL THEORY AND BEYOND	56
4.1	Introduction	56
4.2	Methodology	60
4.3	Results and Discussions	64

4.3.1	Assessing the accuracy of MLPT for different density functional approximations	64
4.3.2	Machine learning Monte Carlo resampling	74
4.3.3	Assessing the accuracy of MLPT for the random phase approximation	82
4.4	Conclusions	84
5	COUPLED CLUSTER FINITE TEMPERATURE SIMULATIONS OF PERIODIC MATERIALS VIA MACHINE LEARNING	85
5.1	Introduction	85
5.2	Methods	87
5.2.1	Periodic boundary conditions coupled cluster	87
5.2.2	Machine Learning Thermodynamic Perturbation Theory and Monte Carlo resampling	88
5.3	Results and discussion	90
5.4	Conclusions	94
6	GENERATIVE MACHINE LEARNING TO SOLVE THE SCHRÖDINGER EQUATION IN THE CONFIGURATION INTERACTION SPACE	95
6.1	Introduction	95
6.2	Methodological approach	100
6.2.1	Configuration interaction approach	100
6.2.2	Generative model	100
6.2.3	Direct generation of single and double excitations	103
6.2.4	Computational complexity	106
6.2.5	Computational details	108
6.3	Numerical Results	108
6.4	Conclusions	111
7	CONCLUSION AND PERSPECTIVE	113
	BIBLIOGRAPHY	116
	RÉSUMÉ	139

LIST OF FIGURES

Figure 1	Jacob’s ladder of density functional theory.	2
Figure 2	Feed forward NN model.	3
Figure 3	Δ – ML: PBE, RPA electronic energies, and their difference for zeolite configurations.	5
Figure 4	Restricted Boltzmann Machine.	7
Figure 5	Self Consistent construction of the Hartree-Fock wave function.	16
Figure 6	Dissociation curve of the BH molecule using various wave function methods	26
Figure 7	Self consistent field procedure to find the ground state eigenvectors of the Kohn-Sham Hamiltonian.	32
Figure 8	Illustration of the pseudo-potential method.	34
Figure 9	Dissociation curve of the BH molecule, using various wave function and DFT methods.	40
Figure 10	Pictorial representation of thermodynamic perturbation theory	45
Figure 11	Smooth Overlap of Atomic Positions descriptor	49
Figure 12	Kernel Ridge Regression of the sinus function, using a gaussian kernel	51
Figure 13	Adsorbed systems studied	60
Figure 14	t-SNE representation of the configurational spaces spanned by the different functionals.	70
Figure 15	Distributions of deviations of SCAN energy from the average value of the corresponding reference calculation	76
Figure 16	Radial distribution functions calculated using the SCAN functional for the first and second coordination sphere of O atoms around Si.	77
Figure 17	t-SNE representation of the configurational spaces spanned by the different functionals for the CH ₄ @HChab.	78
Figure 18	The unit cell of the system studied in this work, CO ₂ in protonated chabazite.	90
Figure 19	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.	93

Figure 20	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.	93
Figure 21	Restricted Boltzmann machine (RBM) as a generative model and its application to the solution of the Schrödinger equation.	98
Figure 22	Pictorial representation of the tower sampling algorithm.	101
Figure 23	Effect of the temperature in the generative procedure.	104
Figure 24	Total energy convergence as a function of the number of iterations.	109
Figure 25	CIgen dissociation curve of N ₂	111

LIST OF TABLES

Table 1	Experimental and reference theoretical enthalpies of adsorption	65
Table 2	Enthalpies of adsorption estimates of CH ₄ @HChab and CO ₂ @SiChab	66
Table 3	Deviations of the MLPT estimates	68
Table 4	Values of the I _w index corresponding to each individual MLPT estimate of the ensemble internal energy.	73
Table 5	Deviations of the SCAN and SCAN+rVV10 internal energies predicted using MLPT and MLMC.	80
Table 6	Experimental enthalpies of adsorption of CH ₄ @HChab and CO ₂ @SiChab and their calculated reference MLPT, and MLMC values.	81
Table 7	Values of the I _w index corresponding to the MLPT predictions of RPA energies from PBE+D2 trajectories.	83
Table 8	Correlation energy contributions to the adsorption energy (kcal/mol) computed using different levels of theory and basis sets.	88
Table 9	Enthalpy of adsorption of CO ₂ in protonated chabazite (kcal/-mol) computed using different target and sampling methods.	91
Table 10	Computational complexity of the different methods discussed in this work	107
Table 11	Total energies (Ha) for the different systems/basis sets . . .	110

ACRONYMS

ML	Machine Learning
MLPT	Machine Learning Perturbation Theory
MLMC	Machine Learning Monte Carlo
TPT	Thermodynamic Perturbation Theory
DFT	Density Functional Theory
LDA	Local Density Approximation
GGA	Generalized Gradient Approximation
PBE	Perdew–Burke–Ernzerhof
HF	Hartree-Fock
CI	Configuration Interaction
MP	Møller-Plesset
CC	Coupled Cluster
RPA	Random Phase Approximation
MD	Molecular Dynamics
RBM	Restricted Boltzmann Machine
KRR	Kernel Ridge Regression

INTRODUCTION

1.1	Chemical accuracy	1
1.2	Prior work	2
1.3	Outline	8

1.1 CHEMICAL ACCURACY

“A target accuracy must be selected. A model is not likely to be of much value unless it is able to provide clear distinction between possible different modes of molecular behavior. As the model becomes quantitative, the target should be that data is reproduced and predicted within experimental accuracy. For energies, such as heats of formation or ionization potentials, a global accuracy of 1 kcal/mole would be appropriate.” This quote from the 1998 Nobel lecture of John Pople [154] defined the chemical accuracy as an energy resolution goal for computational methods in material science. For example, the ratio between reactants and products of a chemical reaction at equilibrium can be obtained by exponentiation of the ratio of the free energy difference between both states and the thermal energy [21] ($K_{R \rightarrow P} = \exp(-\Delta A_{R \rightarrow P}/k_B T)$), and a variation of 1.4 kcal/mol leads to a variation of ~ 10.7 in the equilibrium constant $K_{R \rightarrow P}$.

While the simplest approximations in quantum chemistry (Hartree-Fock) and in solid state physics (local and semi-local Density Functional Theories (DFT)) are affordable for moderately big systems (up to few thousands atoms, with a numerical cost roughly scaling as $\mathcal{O}(K^4)$ and $\mathcal{O}(K^3)$ respectively with basis size [55]), they cannot be improved in a systematic way, and lack chemical accuracy in a lot of applications [37]. One needs to resort to more involve approximations – post Hartree-Fock methods such as Configuration Interaction, Møller-Plesset Perturbation Theory or Coupled Cluster, or beyond-DFT models such as the Random Phase Approximation – but the later are computationally more demanding and are restricted to smaller systems [80]. In particular, while periodic implementations have been recently proposed for condensed-matter applications for second order Møller-Plesset and Coupled Cluster [20, 45, 49, 127, 151], finite temperature results (which require few hundreds of thousands of single point calculations) are still unobtainable for condensed matter systems using those theories. This was conceptualised in 2000 for DFT by Perdew and Schmidt in a famous figure (see the

adaptation in Figure 1): the Jacob's ladder of DFT approximations for exchange-correlation energy [145].

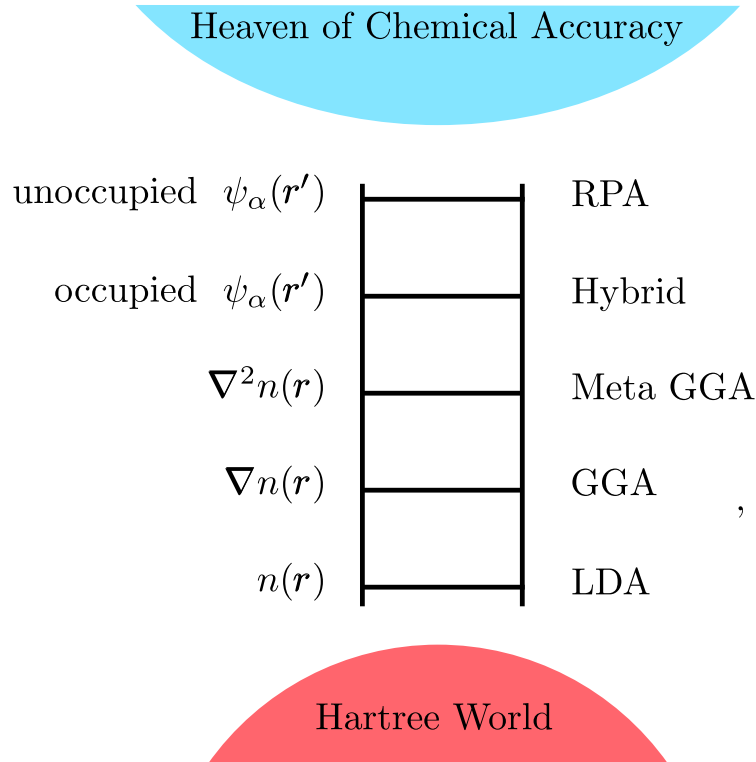


Figure 1: Jacob's ladder of density functional theory, proposed by Perdew and Schmidt. The ladder allows to climb from the mean field description of the Hartree(-Fock) product to chemical accuracy by adding more and more non-locality in the description of the exchange-correlation energy [145].

In this context Machine Learning methods have seen in the last decade an intensive rise in order to alleviate the computational burden of quantum mechanical methods and make them more affordable [24].

1.2 PRIOR WORK

In 1995 the work of Blank *et al.* introduced the use of a feed-forward neural network (NN) to learn the potential energy surface (PES) of a small molecule (CO) adsorbed on a Nickel surface using data from an empirical force fields, and from density functional theory (DFT) calculations of H₂ desorption from a silicon surface [17]. A feed-forward NN, pictured in Figure 2, is a graph model with one input layer that encode the data points, and one or more hidden layers that encode high-order correlations between input points. Between two successive layers all vertices are connected. After the hidden layers, the target property is written

as a non-linear function of a linear combination of the layers, and is learned using variational optimization algorithm such as gradient-descent [125].

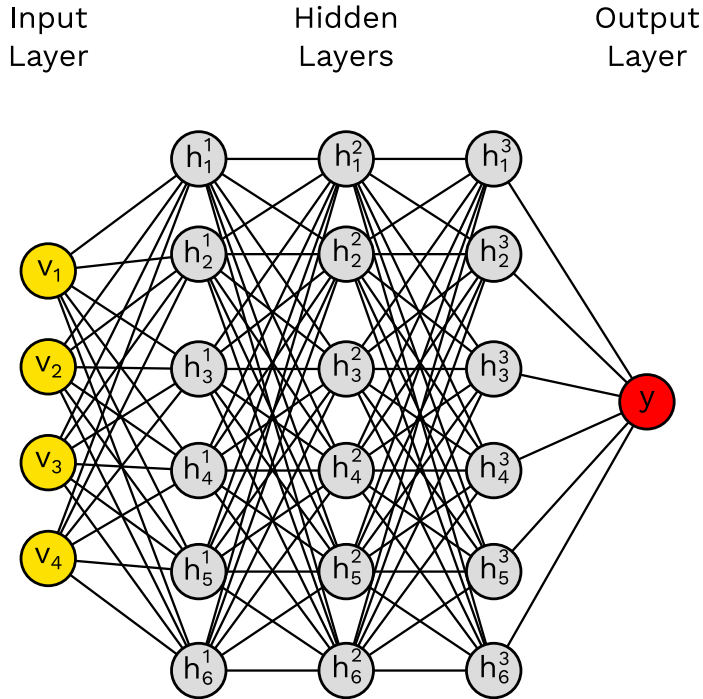


Figure 2: Feed forward NN model consisting on four input neurons, two hidden layers with six neurons each, and a scalar target y . The hidden neurons take value $h_i^j = f_i^j(b_i^j + \sum_{k=1}^6 a_{k,i}^{j-1,j} h_k^{j-1})$, with biases parameters b_i^j and weights $a_{k,i}^{j-1,j}$. The same definition holds for the first hidden layer (fed by the visible input) and the target y . The functions f_i^j are usually non-linear such as the sigmoid function.

If reactions coordinate were used as input for the NN, Lorenz *et al.* performed a similar work in 2004 to study the interaction of H_2 with a Palladium surface, using instead symmetry adapted functions to feed the NN, reducing the need for the model to learn the PES symmetries, and thus reducing the computational cost. With an error comparable to direct DFT computations for a fraction of the computational cost, those models were however reduced to low dimensional systems: the number of weight parameters augment too rapidly with the system size, so as the number of training points required for a proper fit, and the number of iterations necessary in the optimization procedure. Behler and Parinello introduced High-Dimensional Neural Network Potential in a seminal paper in 2007 [15]. They pursued the aforementioned ideas by defining atomic environments: instead of a single high-dimensional feed-forward NN, they used a architecture where each atom is represented by a single atomic feed-forward NN, and the total energy is given by a sum of all atomic contributions. Thus, the size of the NN model

scales linearly with the number of atoms, and the model does not depend on the number and type of atoms, giving it a high flexibility. Each atomic NN describes the configuration of all the neighboring atoms inside a cutoff radius. Symmetry adapted functions representing radial and angular distributions are used in those models. They demonstrated their model by performing molecular dynamics (MD) of periodic Silicon, with 64 atoms in the cell. About ten thousand configurations computed in the local density approximation of DFT were used to train the model, for a negligible validation error (~ 0.1 kcal/mol). While this approach is able to give large scale molecular dynamics the accuracy of DFT with several orders of magnitude of numerical cost avoided, the number of necessary training points remain unattainable for the most accurate quantum chemical methods.

Kernel methods, another paradigm of machine learning algorithms, have seen an important parallel development in quantum chemistry and in the solid state. Bartók *et al.* used Gaussian Processes [160] to generate inter-atomic potential and learn the PES of bulk Carbon, Silicon and Germanium [11]. Rupp *et al.* predicted atomization energies of organic molecules with Kernel Ridge Regression [169].

Databases generation have also been an important step in the process of testing and validating different approaches to learn and predict quantum chemical or solid properties with *e.g.* the QM9 dataset [155] which contain 134 thousand organic molecules with different properties such as atomization energies, dipole moments, polarizability or harmonic frequencies.

Whichever paradigm is used, suitable and efficient representation of atomic configurations are needed. While Behler and coworkers use in their aforementioned model atom-centered symmetry functions [13], which are sum of radial and angular symmetry functions describing the environment inside a cutoff radius for each atom, other atomic environment descriptions have been proposed. For example Bartók *et al.* created the Smooth Overlap of Atomic Positions in 2013 [9], which is a decomposition into spherical harmonics of sum of gaussian centered on atoms inside a cutoff radius. Differently, global representations encode the whole atomic configuration: for instance, the Coulomb matrix depends on pair distances and chemical species [169], the Many Body Tensor Representation contain different weighted broadened distance, angle, and dihedral distribution terms [93].

In 2015, Ramakrishnan *et al.* introduced the idea of Δ -Machine Learning: instead of directly fitting the energy or other property of interest to the configurational space, it can be easier to learn the target value difference as given by two quantum mechanical approximations [156]. Considering two different DFT exchange-correlation functional for example, both approximation include the same kinetic term and differ in their treatment of the correlation energy. Learning the difference between the two approximations is easier than directly learn the whole

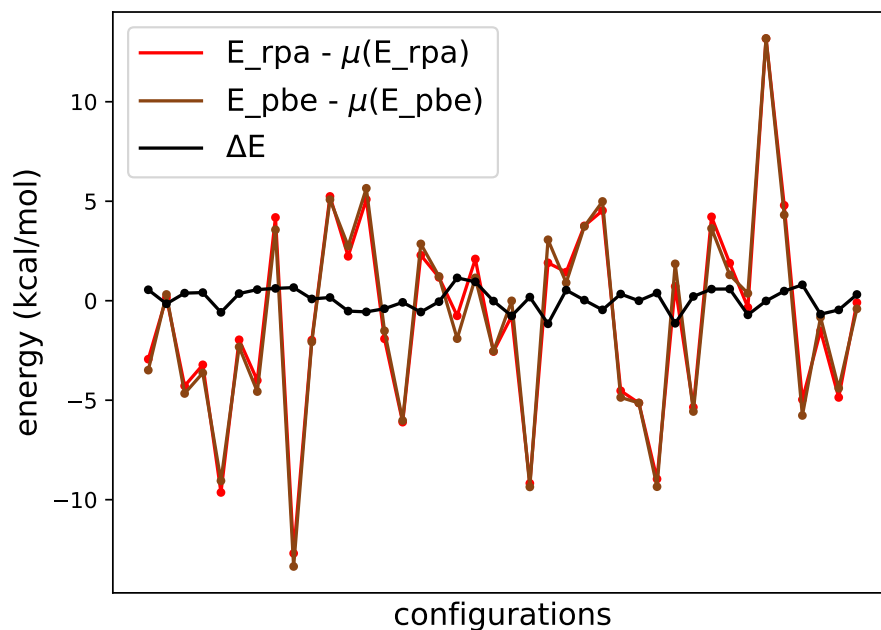


Figure 3: Total electronic energies, centered around their mean, of forty protonated chabazite configurations, within the PBE DFT functional and the random phase approximation. The black curve shows their difference.

physical description of the target quantum mechanical approximation. In Figure 3 is shown as an example the energies and their difference for several zeolite configurations using the Random Phase Approximation (RPA) and the PBE DFT functional. It is seen how the ΔE curve is smoother and possesses a smaller variation. If several thousands of configurations are necessary to learn the PES of a complex system, this number can be reduced to few tens or one hundred using this technique. Dorner *et al.* used in 2018 thermodynamics perturbation theory to reweight an *ab initio* molecular dynamics performed using semi-local DFT to predict the melting temperature of Silicon at RPA level of theory [52]. In a similar work, Rocca *et al.*, in 2019, used thermodynamics perturbation theory to reweight *ab initio* molecular dynamics performed using semi-local DFT onto RPA averages to compute the adsorption enthalpy of carbon dioxide and methane into a zeolite [164]. In thermodynamic perturbation theory, one can express the canonical ensemble average of one Hamiltonian as a function of averages in another Hamiltonian, if the difference between both Hamiltonian is known [31]. Given the computational cost of RPA calculations, they extracted few tens of configurations on which to compute RPA energy and perform the perturbation theory, and were able to recover chemical accuracy with respect to the experiment. The same year, this

work was continued by Chehaibou *et al.*: a delta machine learning model was devised to learn, from those selected configurations, the energy difference between the DFT and RPA energy using a Kernel Ridge Regression model, thus allowing to perform the thermodynamic perturbation theory on the full trajectory and reducing the error bar [28]. This method was coined Machine Learning Perturbation Theory (MLPT), and has been applied to compute enthalpy of adsorption of small molecules into chabazite [28], and free energies of activation of the proton exchange reaction in protonated chabazite [21].

The works mentioned until now assume that quantum mechanical calculations are possible prior to create a model to train on those. It could be advantageous to use ML directly during the process of solving the Schrödinger equation. In 2017, Carleo and Troyer [25] introduced the idea of Neural Quantum States (NQS). Trial functions have been commonly used in Variational Quantum Monte Carlo (QMC), since their introduction by Jastrow in 1955 [99]: one chooses, in front of a mean-field solution such as Hartree-Fock, a variational ansatz (Jastrow factor) to express the correlation of a quantum mechanical system. The ansatz is subsequently optimized in a variational way, notably using Monte Carlo sampling [187]. The Jastrow factor is typically a non-linear function of the pair distances in the problem. While these scheme allows for very accurate calculations (often used as benchmark [70]), there is a certain degree of arbitrary in the choice of the Jastrow function, and a lack of flexibility in the ansatz. In this context, Carleo and Troyer suggested to use the representative power of neural networks [90, 167] instead to parametrize the wave function to be optimized. More specifically, they chose a Restricted Boltzmann Machine (RBM) as illustrated in Figure 4.

The associated wave function amplitude for a configuration \mathbf{v} of binary input $\{v_i\}$ is written as

$$\Psi(\mathbf{v}) = \sum_{\{h_i\}} e^{\mathbf{a}^\top \mathbf{v} + \mathbf{b}^\top \mathbf{h} + \mathbf{h}^\top \mathbf{W} \mathbf{v}} \quad (1)$$

and the optimization is performed using a Variational QMC method, the Stochastic Reconfiguration algorithm [180]. Introduced for spins models such as the Antiferromagnetic Heisenberg model were a better convergence compared to Jastrow functions was shown, this work was further brought to electronic structure problems by Choo *et al.* [35]: by mapping the electronic structure Hamiltonian into a spin Hamiltonian (using among other the Jordan-Wigner transformation [101]), recovering the Configuration Interaction (CI) framework [184]. They studied total electronic energy calculations of small molecules (up to fourteen electrons) in the STO-3G and 6-31G basis sets. While able to converge most of the calculations to chemical accuracy in the STO-3G basis, this was not feasible in the 6-31G basis set. It appears that it is yet difficult to elaborate efficient schemes to optimize a neural network - moreover with complex weights - on the CI problem. The sampling is

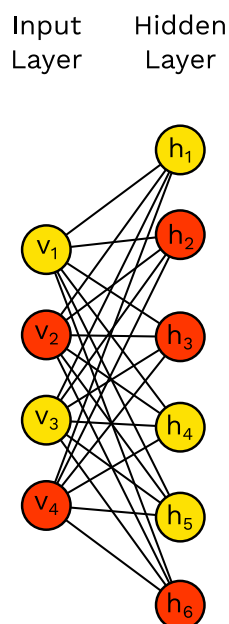


Figure 4: Restricted Boltzmann Machine: this neural network model is composed of one input (or visible) layer (the v_i), and one unique hidden layer (h_i). The model parameter is made of the weight W_{ij} between both layers, plus two additional bias vectors composed of a_i in the visible layer and b_i in the hidden one.

complicated by a predominant weight of the most important configurations in the wave function, resulting in a problematic slow-down of the algorithm for the non-minimal basis set 6-31G. Barrett *et al.* addressed that issue using an autoregressive neural network and could obtain chemically accurate calculations for basis set up to twenty-eight spin orbitals [6].

Another difficulty for the RBM or other simple models is the highly complex nature of the correlations between occupation of orbitals in the configurations. By simple model is understood a model with few parameters, *e.g.* a similar number of hidden and input variables, as the representative power of neural networks grows with the number of those hidden variables [167]. One solution is to use then deep neural networks, by, for example, stacking RBM together, one hidden layer being an input for the next hidden layer (deep Restricted Boltzmann Machine [119]). Real space wave functions have been encoded as NQS in such deep learning architectures in the last four years [76, 82, 105, 146], outperforming previous state of the art Variational QMC ansatzes for systems up to thirty electrons. Truncated CI wave functions remain however a highly valuable method to achieve systematic accuracy and calibrate other methods [42]. Since the CI space grows combinatorially with the system size, prior CI algorithms usually select the most important configurations using perturbation theory [95] or random sampling [67]

before diagonalization of the resulting Hilbert subspace in an iterative way. Machine learning has been proposed as a mean to replace perturbation theory as selection criterion of the important configurations by Coe *et al.* in 2018 [36], using a regressive neural network to learn the important configurations.

1.3 OUTLINE

The aim of this thesis is to contribute to the development of novel machine learning methodologies in order to bring down the computational cost needed in obtaining chemically accurate properties. Machine Learning Perturbation theory is further developed and tested, using various DFT functionals, the Random Phase Approximation, Møller-Plesset and Coupled Cluster theories. On another topic, a generative machine learning model is developed to sample new Slater determinants in the Configuration Interaction framework. The next two chapters are dedicated to understand those diverse quantum mechanical and machine learning methods. The second chapter is a review of the quantum many-electron problem, with an introduction to wavefunction methods and density functional theory. It also introduces *ab initio* Molecular Dynamics and Thermodynamic perturbation Theory. The third chapter explains the theoretical basics of the machine learning models used in this thesis. The results of this thesis can be found in the fourth, fifth and sixth chapters. In Chapter four, the Machine Learning Perturbation Theory is continued to address the possible configurational mismatch inherent to thermodynamic perturbation theory, by studying total electronic energies and adsorption enthalpies with various DFT functionals. A diagnosis test is presented to detect the existence of an overlap problem, in which case a solution is proposed to overcome it, in the form of a Monte Carlo resampling of the biased trajectory, using the original machine learning model. Chapter five carries on with this work by a calculation of the enthalpy of adsorption of carbon dioxide in a zeolite at the coupled cluster single double and perturbative triple (CCSD(T)) level of theory, bringing the first finite temperature simulation of a periodic material at the Coupled Cluster level of theory to our knowledge. Chapter 6 shows how a generative machine learning model can be used to efficiently sample new Slater determinants in an iterative configuration interaction algorithm, improving upon random sampling. The seventh and last chapter discuss the results of this thesis and conclude.

ELECTRONIC STRUCTURE METHODS

2.1	Introduction	9
2.2	Wave function-based methods	10
2.2.1	The many electron problem	10
2.2.2	Hartree-Fock and the correlation energy	14
2.2.3	Configuration Interaction	18
2.2.4	Perturbation theory expansion of the correlation energy	19
2.2.5	Møller-Plesset Perturbation Theory	20
2.2.6	Coupled Cluster	22
2.2.7	Basis sets	25
2.2.8	Methods comparison	26
2.3	Density functional theory	27
2.3.1	Hohenberg-Kohn theorems	27
2.3.2	Kohn-Sham Density functional theory	29
2.3.3	Exchange correlation	33
2.3.4	Approximations to the exchange correlation energy	36
2.3.5	Random Phase Approximation	38
2.3.6	Methods Comparison continued	39
2.4	Finite temperature properties	41
2.4.1	<i>Ab initio</i> Molecular Dynamics	41
2.4.2	Monte Carlo sampling	43
2.4.3	Thermodynamic perturbation theory	43

2.1 INTRODUCTION

The aim of this chapter is to introduce the various quantum mechanical methods considered in this thesis. It is organized as follows: the first section is dedicated to wave function-based methods. It begins by an exposition of the many electron problem, and shows its general solution as a linear combination of Slater determinants. The mean field Hartree-Fock Slater determinant is presented as a first level of approximation, together with a definition of the correlation energy. The Configuration Interaction framework is then presented as a straightforward way to

recover correlation. Møller-Plesset Perturbation Theory and Coupled Cluster theory are introduced as alternative methodologies, without some of the shortcomings of the Configuration Interaction method. A brief introduction to basis sets in quantum chemistry is given before a comparison of those different methods with a dissociation curve of the Hydrogen-Bore molecule. In the second section, the basis of density functional theory is introduced, with the Hohenberg and Kohn theorems, and the Kohn-Sham self consistent scheme. An exact formulation of the exchange-correlation energy is given, with a presentation of various levels of approximations to the later, ending with the Random Phase Approximation. The previous Hydrogen-Bore dissociation curve is continued with those methods. Finally in the third and last section, finite temperature properties are considered, with an introduction to the adiabatic approximation and the Born-Oppenheimer molecular dynamics, and Monte Carlo sampling. That section ends with a presentation of thermodynamic perturbation theory.

2.2 WAVE FUNCTION-BASED METHODS

2.2.1 *The many electron problem*

2.2.1.1 *The complete Hamiltonian*

Let us consider a non-relativistic system of N electrons of mass m and N_α nuclei of masses $\{M_1, \dots, M_{N_\alpha}\}$ and atomic number $\{Z_1, \dots, Z_{N_\alpha}\}$ interacting via the Coulomb potential. The corresponding Hamiltonian for this system is

$$\hat{H} = \hat{T}_e + \hat{T}_n + \hat{V}_{ee} + \hat{V}_{nn} + \hat{V}_{en} \quad (2)$$

with

- $\hat{T}_e = \sum_{i=1}^N \frac{-\hbar^2 \hat{\nabla}_{\mathbf{r}_i}^2}{2m}$
- $\hat{T}_n = \sum_{i=1}^{N_\alpha} \frac{-\hbar^2 \hat{\nabla}_{\mathbf{R}_i}^2}{2M_i}$
- $\hat{V}_{ee} = \sum_{i=1}^N \sum_{j>i}^N \frac{e^2}{4\pi\epsilon_0 |\mathbf{r}_i - \mathbf{r}_j|}$
- $\hat{V}_{nn} = \sum_{i=1}^{N_\alpha} \sum_{j>i}^{N_\alpha} \frac{Z_i Z_j e^2}{4\pi\epsilon_0 |\mathbf{R}_i - \mathbf{R}_j|}$
- $\hat{V}_{en} = \sum_{i=1}^N \sum_{j=1}^{N_\alpha} \frac{-Z_j e^2}{4\pi\epsilon_0 |\mathbf{r}_i - \mathbf{R}_j|}$

where the $\{\mathbf{r}_i\}$, $\{\mathbf{R}_i\}$ are the positions of the electrons and nuclei. The wave function $\Psi(\{\mathbf{r}_i\}, \{\mathbf{R}_i\})$ of this system is in position basis a function of all electrons and nuclei positions, and a solution of the time-independent Schrödinger equation

$$\hat{H}|\Psi\rangle = E|\Psi\rangle \quad (3)$$

Additionally, our system is made of half-integer spin particles (fermions), and the wave function must be antisymmetric under the exchange of any two indistinguishable particles.

2.2.1.2 Electronic Schrödinger equation

In the many electron problem we are interested in the eigenvectors $|\psi_i^e\rangle$ and eigenvalues E_i^e of the electronic Hamiltonian.

$$\hat{H}_e = \hat{T}_e + \hat{V}_{ee} + \hat{V}_{ext} \quad (4)$$

with $\hat{V}_{ext} = \hat{V}_{en}$. When solving the electronic Schrödinger equation, the nuclei positions in \hat{V}_{ext} are fixed parameters and the set of eigenvalues $\{E_i^e(\mathbf{R}_j)\}$ where \mathbf{R}_j are the nuclei positions is called the potential energy surface of the problem. This is the **Born-Oppenheimer approximation**, that will be studied more in depth in Section 2.4.1.

2.2.1.3 General form of the total wave function: Slater determinants

Consider a single electron system, with wave function $|\psi\rangle = \sum_i c_i |\phi_i\rangle$, where the set of $\{|\phi_i\rangle\}$ is a complete orthonormal basis, and i completely specify both space and spin. If we now add another electron, we can describe the new wave function as $|\psi(1,2)\rangle = \sum_i \sum_j a_{ij} |\phi_i\rangle \otimes |\phi_j\rangle = \sum_i \sum_j a_{ij} |\phi_i, \phi_j\rangle = \sum_i \sum_j a_{ij} |\phi_i(1)\phi_j(2)\rangle$. The wave function of fermions being antisymmetric under the exchange of two particles, this translate into $|\psi(2,1)\rangle = \sum_i \sum_j a_{ij} |\phi_j, \phi_i\rangle = \sum_i \sum_j a_{ij} |\phi_i(2)\phi_j(1)\rangle = -|\psi(1,2)\rangle$, so that we must impose $a_{ij} = -a_{ji}$ for all i, j . This yields

$$|\psi(1,2)\rangle = \sum_i \sum_{j>i} a_{ij} (|\phi_i(1)\phi_j(2)\rangle - |\phi_i(2), \phi_j(1)\rangle) \quad (5)$$

If we add another electron, the total wave function is

$$|\psi(1,2,3)\rangle = \sum_i \sum_j \sum_k a_{ijk} |\phi_i(1)\phi_j(2)\phi_k(3)\rangle. \quad (6)$$

The antisymmetry of $|\psi(1,2,3)\rangle$ impose that a_{ijk} change sign under any transposition of two indexes among i, j, k since for example $|\psi(1,2,3)\rangle = -|\psi(2,1,3)\rangle = +|\psi(2,3,1)\rangle$, we will have then

$$a_{\sigma(ijk)} = \begin{cases} a_{ijk} & \text{if } \sigma(ijk) \text{ is an even permutation of } i, j, k \\ -a_{ijk} & \text{if } \sigma(ijk) \text{ is an odd permutation of } i, j, k \\ 0 & \text{otherwise.} \end{cases} \quad (7)$$

or, using the Levi-Civita symbol,

$$a_{\sigma(ijk)} = a_{ijk} \epsilon_{\sigma(ijk)} \quad (8)$$

and

$$|\psi(1,2,3)\rangle = \sum_i \sum_{j>i} \sum_{k>j} a_{ijk} \sum_{\sigma(ijk)} \epsilon_{\sigma(ijk)} |\phi_{\sigma_i}(1)\phi_{\sigma_j}(2)\phi_{\sigma_k}(3)\rangle \quad (9)$$

where σ_i gives the corresponding value of i in the permutation $\sigma(ijk)$. For example, if $\sigma(ijk) = jik$, then $\sigma_i = j$.

For a squared 3×3 matrix M with elements $M_{\alpha\beta}$, the determinant can be defined as $\det(M) = \sum_{\sigma(123)} \epsilon_{\sigma(123)} M_{1\sigma_1} M_{2\sigma_2} M_{3\sigma_3}$. We see that we can use the notation of determinants to define

$$\begin{aligned} |ij \cdots n\rangle &= \sum_{\sigma(ij \cdots n)} \epsilon_{\sigma(ij \cdots n)} |\phi_{\sigma_i}(1)\phi_{\sigma_j}(2) \cdots \phi_{\sigma_n}(N)\rangle \\ &\equiv \begin{vmatrix} \phi_i(1) & \phi_j(1) & \cdots & \phi_n(1) \\ \phi_i(2) & \phi_j(2) & \cdots & \phi_n(2) \\ \cdots & \cdots & \cdots & \cdots \\ \phi_i(N) & \phi_j(N) & \cdots & \phi_n(N) \end{vmatrix} \end{aligned} \quad (10)$$

For N electrons, we will obtain the total wave function as

$$|\Psi(1,2, \cdots, N)\rangle = \sum_i \sum_{j>i} \cdots \sum_{n>m} a_{ijk \cdots lmn} |ijk \cdots lmn\rangle \quad (11)$$

The kets $|ijk \cdots lmn\rangle$ are called Slater determinants [130, 184] and the total wave function is written as a linear combinations of such Slater determinants. They form a convenient basis of the problem as they directly encode the anti-symmetry requirement.

2.2.1.4 Second quantization

It is sometime convenient to work in the so-called second quantization formalism where the Hamiltonian reads

$$\hat{H}_e = \sum_{ij} t_{ij} \hat{c}_i^\dagger \hat{c}_j + \sum_{ij} v_{ij}^{\text{ext}} \hat{c}_i^\dagger \hat{c}_j + \sum_{ijkl} v_{ijkl} \hat{c}_j^\dagger \hat{c}_i^\dagger \hat{c}_k \hat{c}_l \quad (12)$$

with

$$t_{ij} = \langle \phi_i | \hat{T}_e | \phi_j \rangle \quad (13)$$

$$v_{ij}^{ext} = \langle \phi_i | \hat{V}_{ext} | \phi_j \rangle \quad (14)$$

$$v_{ijkl} = \langle \phi_i \phi_j | \hat{V}_{ee} | \phi_k \phi_l \rangle \quad (15)$$

where \hat{T}_e , \hat{V}_{ext} and \hat{V}_{ee} are the one-particle kinetic, external and the two-particles Coulomb operators respectively (without sum as in eq. (2)). $\hat{c}_i, \hat{c}_i^\dagger$ are the annihilation and creation operators, obeying (for fermions) the anti-commutation rules

$$\{\hat{c}_i, \hat{c}_j^\dagger\} = \delta_{ij} \quad (16)$$

$$\{\hat{c}_i, \hat{c}_j\} = 0 \quad (17)$$

$$\{\hat{c}_i^\dagger, \hat{c}_j^\dagger\} = 0 \quad (18)$$

Those operators act on Slater determinants in the occupation basis

$$\hat{c}_i^\dagger |n_1, \dots, n_i, \dots\rangle = (-1)^{\sum_{k=n_1}^{n_i-1} n_k} (1 - n_i) |n_1, \dots, n_i + 1, \dots\rangle \quad (19)$$

$$\hat{c}_i |n_1, \dots, n_i, \dots\rangle = (-1)^{\sum_{k=n_1}^{n_i-1} n_k} n_i |n_1, \dots, n_i - 1, \dots\rangle \quad (20)$$

where $|n_1, n_2, \dots\rangle = |1, 0, 0, 1, 1, 0, \dots\rangle$ corresponds to the 3-electron state $|1, 4, 5\rangle$ in eq. (10). We can see then that $\hat{c}_1 |1, 4, 5\rangle = |4, 5\rangle$, $\hat{c}_1 |4, 5\rangle = 0$, $\hat{c}_1^\dagger |1, 4, 5\rangle = 0$ and $\hat{c}_1^\dagger |4, 5\rangle = |1, 4, 5\rangle$.

It is possible to change the one-electron basis $|\phi_k\rangle$ by $|x\rangle = \sum_k |\phi_k\rangle \langle \phi_k | x \rangle = \sum_k \phi_k^*(x) |\phi_k\rangle = \sum_k \phi_k^*(x) \hat{c}_k^\dagger |0\rangle$, where $|x\rangle$ specifies a new basis, e.g. in position basis $|x\rangle = |\mathbf{r}, \sigma\rangle$ with $\sigma = \pm \frac{1}{2}$ the spin projection. This defines the fields creation operators

$$\hat{\psi}^\dagger(x) = \sum_k \phi_k^*(x) \hat{c}_k^\dagger. \quad (21)$$

The Hamiltonian can again be rewritten in this basis [56, 130] with

$$\hat{H}_e = \int dx \hat{\psi}^\dagger(x) (\hat{T}_e(x) + \hat{V}_{ext}(x)) \hat{\psi}(x) + \frac{1}{2} \int dx dx' \hat{\psi}^\dagger(x) \hat{\psi}^\dagger(x') \hat{V}_{ee}(x, x') \hat{\psi}(x) \hat{\psi}(x') \quad (22)$$

2.2.1.5 Slater-Condon rules

Using this formalism, we can more easily evaluate the Hamiltonian matrix elements $\langle ijk \dots lmn | \hat{H}_e | ijk \dots lmn \rangle$. The one-body operators \hat{T}_e and \hat{V}_{ext} involve matrix elements $\langle n_1, n_2, \dots | \hat{c}_i^\dagger \hat{c}_j | n_1, n_2, \dots \rangle$ which are only non-zero if $i = j$. The two-body operator \hat{V}_{ee} involve matrix elements $\langle n_1, n_2, \dots | \hat{c}_j^\dagger \hat{c}_i^\dagger \hat{c}_k \hat{c}_l | n_1, n_2, \dots \rangle$ which are only non-zero if $(i = k, j = l)$ or $(i = l, j = k)$. This means that *Slater determinants differing by more than two orbitals have a corresponding Hamiltonian matrix element equal to zero*. Let $|A\rangle, |A_a^r\rangle, |A_{ab}^{rs}\rangle$ be Slater determinants where the orbital

a in $|A\rangle$ becomes r in $|A_{ab}^{rs}\rangle$ and the orbitals a, b in $|A\rangle$ become r, s in $|A_{ab}^{rs}\rangle$. The following **Slater-Condon rules** hold [184]:

$$\langle A|\hat{H}_e|A\rangle = \sum_{i=1}^N (t_{ii} + v_{ii}^{\text{ext}}) + \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N (v_{ijij} - v_{ijji}) \quad (23)$$

$$\langle A|\hat{H}_e|A_a^r\rangle = (t_{ar} + v_{ar}^{\text{ext}}) + \sum_{i=1}^N (v_{airi} - v_{aiir}) \quad (24)$$

$$\langle A|\hat{H}_e|A_{ab}^{rs}\rangle = v_{abrs} - v_{absr} \quad (25)$$

To summarize, the many-problem consist in solving the eigenvalue problem defined by the time-independent Schrödinger equation, for which there exist no analytical solution. The exponential growth of the Hilbert space size with the number of electrons prevents a direct numerical resolution and we need to resort to approximations. The first of such consist in reducing the one-electron basis size by any kind of numerical cut-off. For a real-space representation ($\phi_i(\mathbf{r}) = \delta(\mathbf{r} - \mathbf{r}_i)\sigma_i$), this means introducing a minimal numerical space resolution. Quantum chemistry use atomic orbitals based on the solution of the Schrödinger equation for the Hydrogen atom, and the cut-off here corresponds to the size (the number and type) of the basis set used. An additional type of approximation is to neglect terms in the expansion of the full wave function, retaining only the most important. We will now describe the Hartree-Fock procedure, that express the wave function as a single of such terms.

2.2.2 Hartree-Fock and the correlation energy

2.2.2.1 Hartree-Fock method

Consider the first element $|123 \cdots N\rangle$ of the expansion in eq. (11). Remember that we had defined the single electron basis from $|\psi\rangle = \sum_i c_i |\phi_i\rangle$. This basis is not unique and we can obtain all other basis through unitary transformations of the coefficients c_i . By doing so, the Hamiltonian matrix element $\langle 123 \cdots N|\hat{H}_e|123 \cdots N\rangle$ will vary. By minimizing this energy with respect to the c_i , we will obtain the Hartree-Fock Slater determinant. It is the best single Slater determinant approximation to the true ground state N electrons wave function.¹

Let construct the following Lagrangian functional with the basis orthonormality constraint:

$$\mathcal{L}[\{\phi_i\}] = E[\{\phi_i\}] - \sum_{ij} \epsilon_{ij} (\langle \phi_j|\phi_i\rangle - \delta_{ij}). \quad (26)$$

¹ In particular, if there are no interaction between electrons, and if the $|\phi_i\rangle$ are eigenvectors of the single electron problem, with sorted eigenvalues ϵ_i , then the Slater determinant $|12 \cdots N\rangle$ is the ground-state to the N electron problem (up to a normalizing constant $\frac{1}{\sqrt{N!}}$), with energy $E = \sum_{i=1}^N \epsilon_i$.

where $E[\{\phi_i\}]$ is given by eq. (23). By cancelling the functional derivatives $\partial\mathcal{L}/\partial\phi_i$, one obtains the following eigenvalue problem

$$F\phi_i(x) = \left(h(x) + \sum_{j=1}^N (J_j(x) - K_j(x)) \right) \phi_i(x) = \epsilon_i \phi_i(x) \quad (27)$$

with the one-electron Coulomb and exchange operators acting on a state $\phi_i(x)$ as

$$J_j(x)\phi_i(x) = \int dx' \frac{|\phi_j(x')|^2}{r_{ij}} \phi_i(x) \quad (28)$$

$$K_j(x)\phi_i(x) = \int dx' \frac{\phi_j^*(x')\phi_i(x')}{r_{ij}} \phi_j(x) \quad (29)$$

and the one-electron operator $\hat{h} = \hat{T}_e + \hat{V}_{\text{ext}}$.² The total operator \hat{F} is called the **Fock operator**. This procedure yields a new one-electron basis written as a linear combination of the old one. For molecular systems (by opposition to periodic systems), the one-electron orbitals $\phi_i(x)$ are called **molecular orbitals**.

The one-electron Coulomb operator acts on each point by averaging the Coulomb repulsion due to the other orbitals. Note the non-locality of the exchange operator whose action on ϕ_i at point x depends on its value at all points.

Since both contain the orbitals in their definition, one needs to resort to a self consistent scheme in order to find the solutions of this problem, as illustrated in Figure 5.

The one-electrons energies ϵ_i are eigenvalues of the Fock operator:

$$\epsilon_i \equiv \langle \phi_i | \hat{F} | \phi_i \rangle = t_{ii} + v_{ii}^{\text{ext}} + \sum_{j=1}^N (v_{ijij} - v_{ijji}). \quad (30)$$

and the Hartree-Fock state $|\Phi\rangle = |12 \dots N\rangle$ has the energy

$$E_{\text{HF}} = \sum_{i=1}^N \epsilon_i - \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N (v_{ijij} - v_{ijji}) \quad (31)$$

The Fock operator can be rewritten as

$$\hat{F} = \hat{T}_e + \hat{V}_{\text{ext}} + \hat{V}_{\text{HF}} \quad (32)$$

where \hat{V}_{HF} is to be seen as an effective one-particle potential, averaging on the effect of the occupation of the N first orbitals.

In this thesis, we will always consider closed-shell systems, *i.e.* systems with an equal number of spin up and down electrons (as opposed to open-shell systems).

² It is possible to alternatively enforce the wavefunction normalization and keep the freedom of non-orthonormal basis. In that case a general eigenvalue problem $F\phi = ES\phi$ is found where $S_{ij} = \langle \phi_i | \phi_j \rangle$ is the overlap matrix [184].

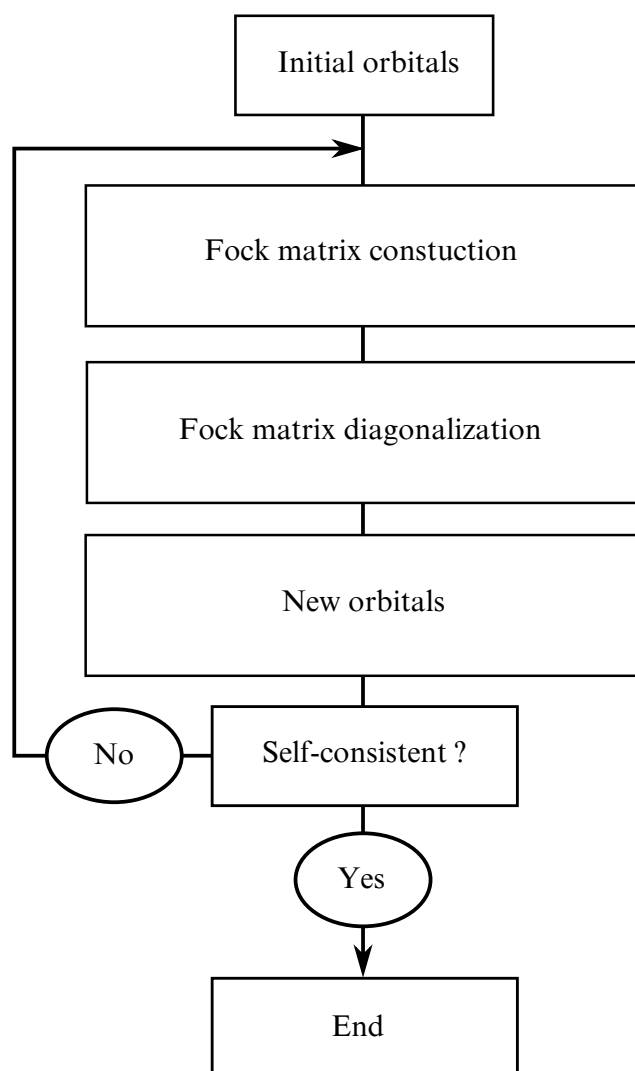


Figure 5: Self Consistent construction of the Hartree-Fock wave function: After construction and diagonalization of the Fock matrix, starting from a guess of the orbitals (possibly random), a new set of orbitals ϕ_i and eigenvalues ϵ_i are found. This is repeated until some convergence criterion is met on the later.

For such systems, it is possible to consider only spatial atomic orbitals in the minimization yielding the Hartree-Fock state: the same spatial orbital hold two electrons with opposite spins. This procedure is called **Restricted Hartree-Fock (RHF)**. The **Unrestricted Hartree-Fock (UHF)** method consider spin-orbitals with a different spatial part. Finally, let us note that the molecular spin-orbitals are in practice written as linear combination of atomic spin-orbitals in the form $\phi_i(x) = \sum_{\alpha} c_{\alpha i} f_{\alpha}(x)$, as mentioned earlier. The most computationally intensive step of the Hartree-Fock method is to compute all v_{ijkl} integral terms of eq. (15), making it

quartic with the atomic basis set size. Those atomic basis sets will be introduced in subsection 2.2.7.

2.2.2.2 Correlation energy

The **correlation energy** E_c is defined as the difference between the ground state energy E and the Hartree-Fock energy E_{HF} :

$$E_c = E - E_{\text{HF}} \quad (33)$$

Note that in practice, as we are using finite basis sets, the ground state energy E must be understood with respect to that basis set. The correlation energy is commonly decomposed into **dynamic correlation** near the equilibrium geometry and **static correlation** near dissociated geometry. The term “dynamic” refers to the lack of relaxation in the Hartree-Fock approximation where the electrons only feel an average potential. The non-dynamic correlation, hence “static”, refers to the multi-configurational character of the wave function near dissociated geometries. Considering for example the hydrogen molecule H_2 with only two basis functions, one $1s$ function attached on each hydrogen. The wave function can be written as a sum of the Hartree-Fock state and a doubly excited determinant. At equilibrium geometry, the Hartree-Fock state represent the largest contribution to the wave function among the two, while in the dissociation limit, the Hartree Fock determinant is degenerated with the doubly excited determinant [184].

2.2.2.3 Size-consistency and size-extensivity

A method is said to be **size-consistent** if, when considering two sub-systems A and B without interaction between each other, the energy of the total system is given by the sum of the energy of the sub-systems, as it should. Moreover, the total wave function should be the antisymmetrized product of the wave functions of A and B . Typically A and B could be the atoms forming a molecule, which become non-interacting in the *dissociation limit* where the pair distances between atoms goes to infinity.

The Fock matrix is constituted of elements $F_{ij} = \langle \phi_i | \hat{F} | \phi_j \rangle$. In the case where A and B are two closed-shell systems, the Fock matrix elements F_{ij} where i (j) are localized on A (B) will be zeros in the dissociation limit of the RHF procedure. Thus the Fock matrix is block diagonal already from the first iteration, and the Hartree-Fock determinant is the product of the two subsystems Hartree-Fock determinants, hence its energy is size-consistent. If A and B are open-shell however, while their sum is not, only the UHF method can recover this block diagonal form. The RHF method is then non size-consistent in general, while the UHF method is.

Size-extensivity is a property obeyed by physical systems such that the energy of M monomers must scale linearly with M when $M \rightarrow \infty$ [184]. (This does not mean that the energy is M times the energy of a single monomer). As such, it is necessary for periodic systems to be described by an approximation that respects this constraint. Obviously a non size-consistent method is not size-extensive and truncated CI methods, which lack size-consistency as discussed in the next section, cannot be used for periodic systems.

In the so-called **Post Hartree-Fock methods**, after the Hartree-Fock Slater determinant has been constructed, we try to recover (part or all) of the correlation energy.

2.2.3 Configuration Interaction

The Configuration Interaction is the most straightforward method to do so, given our previous definitions. Indeed, we can rewrite eq. (11) as

$$|\Psi\rangle = |\Phi\rangle + \sum_{\substack{a \\ r}} c_a^r |\Phi_a^r\rangle + \sum_{\substack{a < b \\ r < s}} c_{ab}^{rs} |\Phi_{ab}^{rs}\rangle + \sum_{\substack{a < b < c \\ r < s < t}} c_{abc}^{rst} |\Phi_{abc}^{rst}\rangle + \dots \quad (34)$$

where the reference state $|\Phi\rangle$ is chosen to be the Hartree-Fock Slater determinant, and the wave function possesses the intermediate normalization condition (both $|\Psi\rangle$ and $|\Phi\rangle$ have unit norm). In that formalism, we can restrict the wave function to a given number of orbital substitutions with respect to the Hartree-Fock state and solve the Schrödinger equation in that particular truncated Hilbert space. Denoting by S,D,T,Q all the singly, doubly, triply and quadruply substituted states, there exist different approximated wave functions such as CIS, CISD, CID, CIS-DTQ. Considering all possible substitutions (and thus recovering the total wave function form) is called **Full Configuration Interaction (FCI)** and yields the exact (in the basis set) energy E_0 .

For a RHF solution with $2K$ molecular spin-orbitals, $N_\alpha = N_\beta$ spin up and spin down electrons, there are $\binom{K}{N_\alpha}^2$ such substitutions. Thus, FCI scales combinatorially with the basis set size, and only small molecules and small basis sets are attainable. Let consider the CID wave function, made of the Hartree-Fock state plus all doubly substituted states. If we attempt to describe two non-interacting subsystems A and B as defined earlier, we cannot write the CID wave function of the total system as the product of CID wave functions of each subsystem: the former will contain products of singly substituted determinants, while the later will hold quadruply substituted determinants (product of doubly substituted determinants). This is a general feature of the truncated CI method to lack size-consistency. Of course the FCI wave function, being exact, is size-consistent.³

³ The product of the two non-interacting FCI wave functions will correctly yield the FCI wave function of the total system.

2.2.4 Perturbation theory expansion of the correlation energy

We will first describe a formal expansion of the ground state energy in terms of *non-degenerate* Rayleigh-Schrödinger perturbation theory [184] (RSPT). In this framework, one expresses the Hamiltonian of interest as

$$\hat{\mathcal{H}} = \hat{\mathcal{H}}_0 + \lambda \hat{V} \quad (35)$$

$$\hat{\mathcal{H}} |\Psi_i\rangle = \mathcal{E}_i |\Psi_i\rangle \quad (36)$$

where the chosen unperturbed Hamiltonian $\hat{\mathcal{H}}_0$ has a known non-degenerate eigenspace:

$$\hat{\mathcal{H}}_0 |\Phi_i^{(0)}\rangle = E_i^{(0)} |\Phi_i^{(0)}\rangle \quad (37)$$

and the perturbation \hat{V} is turned on with λ where $0 \leq \lambda \leq 1$. The eigenstates and eigenvalues are expanded as

$$|\Psi_i\rangle = \sum_{k=0}^{\infty} \lambda^k |\Phi_i^{(k)}\rangle \quad (38)$$

$$\mathcal{E}_i = \sum_{k=0}^{\infty} \lambda^k E_i^{(k)} \quad (39)$$

Using intermediate normalisation ($\langle \Phi_i^{(0)} | \Psi_i \rangle = 1$), insertion of eq. (38) and eq. (39) in eq. (36), identifying power of λ , and insertion of the $\hat{\mathcal{H}}_0$ eigenstates, one finds for the three first orders

$$E_i^{(1)} = \langle \Phi_i^{(0)} | \hat{V} | \Phi_i^{(0)} \rangle \quad (40)$$

$$E_i^{(2)} = \sum_{k \neq i} \frac{|\langle \Phi_i^{(0)} | \hat{V} | \Phi_k^{(0)} \rangle|^2}{E_i^{(0)} - E_k^{(0)}} \quad (41)$$

$$E_i^{(3)} = \sum_{k \neq i} \sum_{l \neq i} \frac{\langle \Phi_i^{(0)} | \hat{V} | \Phi_k^{(0)} \rangle \langle \Phi_k^{(0)} | \hat{V} | \Phi_l^{(0)} \rangle \langle \Phi_l^{(0)} | \hat{V} | \Phi_i^{(0)} \rangle}{(E_i^{(0)} - E_k^{(0)})(E_i^{(0)} - E_l^{(0)})} \quad (42)$$

$$- E_i^{(1)} \sum_{k \neq 0} \frac{|\langle \Phi_i^{(0)} | \hat{V} | \Phi_k^{(0)} \rangle|^2}{(E_i^{(0)} - E_k^{(0)})^2}$$

If the first order term $E_i^{(1)}$ is size-consistent, we might suspect the third order term not to be: as $E_i^{(1)} = e(A) + e(B)$ for two independent systems A and B, we will find product of A term and B terms in the second sum of $E_i^{(3)}$.

However it turns out that those size inconsistent terms are cancelled by some terms in the first sum. This is a general property and all terms in the perturbative expansion are size-consistent [80, 184]. Nonetheless, it is impractical that those size inconsistent terms remains in the formal expansion, if their cancellation in numerical schemes is not totally recovered. An alternative and equivalent derivation

of the energy expansion can be found using time-dependent perturbation theory [56, 102]. Using Gell-Mann and Low theorem, the correlation energy is rewritten as [56, 61]

$$\varepsilon - E^{(0)} = \frac{\langle \Phi^{(0)} | \hat{V}(0) \hat{U}(0, -\infty) | \Phi^{(0)} \rangle}{\langle \Phi^{(0)} | \hat{U}(0, -\infty) | \Phi^{(0)} \rangle} \quad (43)$$

where $U_I(t, t')$ is the evolution operator and interaction picture is used:

$$U_I(t, t') = T \exp \left(-i \int_{t'}^t \hat{V}_I(t) dt \right) \quad (44)$$

T designates the time ordering operator, that reorders terms upon which it act such that the latest is to the left, with a global sign given by the parity of the number of permutations needed. The terms arising from the exponential serie are further decomposed through Wick theorem and finally can be drawn as Goldstone diagrams (time-ordered Feynman diagrams) [56, 130]. The numerator is then seen to factorize into a product of two sums, one of which cancels the denominator, with only *linked* diagrams surviving (this is one formulation of the linked-cluster theorem), resulting upon further time-integration in the following formula for the correlation energy, due to Goldstone [63]

$$\varepsilon - E^{(0)} = \langle \Phi^{(0)} | \hat{V} \sum_{n=0}^{\infty} \left(\frac{1}{E^{(0)} - \mathcal{H}_0} \hat{V} \right)^n | \Phi^{(0)} \rangle_{\text{connected}} \quad (45)$$

The subscript *connected* means that the sum only involves diagrams that are linked, that is, topologically connected diagrams, which correspond to terms that are size-consistent. the diagrams are constructed by successively (n times at order n) applying the perturbation, which creates couples of particle-holes pairs for the Coulomb interaction, before the last \hat{V} returns the pairs to the ground state [56].

2.2.5 Møller-Plesset Perturbation Theory

Møller-Plesset Perturbation Theory (MPPT) is a RSPT (or equivalently the Goldston expansion of eq. (45)) where \mathcal{H}_0 is chosen as the Hartree-Fock operator \hat{F} (which is the sum over electrons of one-particle Fock operators), and the perturbation is then $\hat{H}_e - \hat{F} = \hat{V}_{ee} - \hat{V}_{HF}$. Thus the corresponding eigenstates are Slater determinants. The zeroth and first orders of the ground state energy are

$$E^{(0)} = \langle \Phi | \hat{F} | \Phi \rangle = \sum_{i=1}^N \varepsilon_i \quad (46)$$

$$E^{(1)} = \langle \Phi | \hat{V}_{ee} - \hat{V}_{HF} | \Phi \rangle = -\frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N (v_{ijij} - v_{ijji}) \quad (47)$$

yielding together the Hartree-Fock energy. In second order, only the doubly substituted Slater determinants contribute due to Brillouin theorem.⁴ If a, b label the occupied molecular orbitals in the Hartree-Fock state, and r, s the unoccupied ones, we obtain

$$E^{(2)} = \frac{1}{4} \sum_{a,b,r,s} \frac{|v_{abrs} - v_{absr}|^2}{\epsilon_a + \epsilon_b - \epsilon_r - \epsilon_s} \quad (48)$$

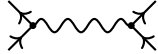
or

$$E^{(2)} = \frac{1}{2} \sum_{a,b,r,s} \frac{|v_{abrs}|^2}{\epsilon_a + \epsilon_b - \epsilon_r - \epsilon_s} - \frac{1}{2} \sum_{a,b,r,s} \frac{v_{abrs}^* v_{absr}}{\epsilon_a + \epsilon_b - \epsilon_r - \epsilon_s} \quad (49)$$

where the correction has been separated into *direct* and *exchange* contributions, and the second order MPPT energy, or **MP2 energy**, reads

$$E_{\text{MP2}} = E_{\text{HF}} + E^{(2)} \quad (50)$$

In higher order, it is easier to use the diagrammatic expansion of eq. (45). The successive diagrams of the expansion can be constructed using the following rules [184]:

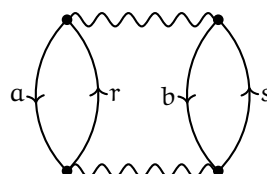
1. The Coulomb interaction is represented by an interacting wiggly line between two vertices, with a pair of directed lines (one in and one out) at each vertex 
2. At order n , stack n such pieces vertically and connect them with the straight line in all possible ways upon deformation of the lines and horizontal reflection(s) of the interaction(s), allowing only connected diagrams, and forbidding self interacting vertices (Brillouin theorem).
3. Label each directed line with an index

The associated value of the diagram is computed by

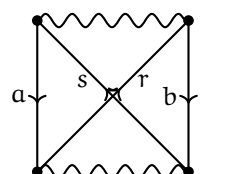
- A factor $v_{klmn} = \langle \text{left in}; \text{right in} | \hat{V}_{ee} | \text{left out}; \text{right out} \rangle$ per interaction line,
- Count the lines up (particle) and down (hole) between each interaction line and associate a factor $(\sum_h \epsilon_h - \sum_p \epsilon_p)^{-1}$
- Count the number of holes lines h and the number of closed loops l and associate a factor $(-1)^{h+p}$
- A factor $\frac{1}{2}$ for diagrams with horizontal symmetry
- Sum the resulting term over each index

⁴ $\langle \Phi | H | \Phi_a^r \rangle = h_{ar} + \sum_i (v_{airi} - v_{aiir})$ constitutes an off-diagonal element of the Fock matrix.

For example the two second order diagrams (direct and exchange terms above) are

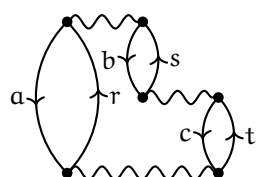


$$= (-1)^{2+2} \frac{1}{2} \sum_{a,b,r,s} \frac{v_{abrs}^* v_{abrs}}{\epsilon_a + \epsilon_b - \epsilon_r - \epsilon_s} \quad (51)$$



$$= (-1)^{2+1} \frac{1}{2} \sum_{a,b,r,s} \frac{v_{absr}^* v_{abrs}}{\epsilon_a + \epsilon_b - \epsilon_r - \epsilon_s} \quad (52)$$

And one of the twelve third order diagrams:



$$= (-1)^{3+3} \sum_{a,b,c,r,s,t} \frac{v_{acrt} v_{btsc} v_{rsab}}{(\epsilon_a + \epsilon_c - \epsilon_r - \epsilon_t)(\epsilon_a + \epsilon_b - \epsilon_r - \epsilon_s)} \quad (53)$$

MPPT is thus size-consistent at all order for closed-shell fragment systems if RHF is used as unperturbed wave function. Convergence of the serie should be considered carefully: first of all the difference between the highest occupied molecular orbital (HOMO) and the lowest unoccupied one (LUMO) should be big enough to avoid the divergence of all terms (thus homonuclear molecules will have a divergent MPPT energy near dissociation, see also ref. [74] for the MP2 divergence in metals). Moreover, the radius of convergence of the serie with λ can be inferior to one. The convergence is usually assured near equilibrium positions but not near dissociation. The MP2 method represent a net improvement over Hartree-Fock near equilibrium, for a cost scaling in M^5 with the number of basis functions [39]. The higher orders are less used as the ratio of correlation energy recovered over computational cost is less appealing [80]. As a final note, let us mention that since perturbation theory is not variational, the MPPT energy will oscillate with the order of perturbation and can occasionnaly be lower than the FCI energy.

2.2.6 Coupled Cluster

We will finish our presentation of wave function methods with **Coupled Cluster (CC)** theory. The starting point is the CC exponential ansatz for the ground state wave function:

$$|\Psi_{CC}\rangle = e^{\hat{T}} |\Phi\rangle \quad (54)$$

with $|\Phi\rangle$ the Hartree-Fock wave function and the CC operator defined by

$$\hat{T} = \sum_n \hat{T}_n, \quad (55)$$

the \hat{T}_n being n -substitutions operators. For example we have

$$\hat{T}_1 = \sum_a \sum_r t_a^r \hat{c}_r^\dagger \hat{c}_a \quad (56)$$

$$\hat{T}_2 = \frac{1}{4} \sum_{ab} \sum_{rs} t_{ab}^{rs} \hat{c}_r^\dagger \hat{c}_s^\dagger \hat{c}_b \hat{c}_a \quad (57)$$

that generate respectively all singly and doubly excited Slater determinant from the reference $|\Phi\rangle$, t_a^r and t_{ab}^{rs} being the single and double CC amplitudes. Expansion of the exponential in eq. (54) yields:

$$e^{\hat{T}} = (1 + \hat{T} + \frac{1}{2}\hat{T}^2 + \hat{T}^3 + \dots) = (1 + \hat{T}_1 + \hat{T}_2 + \dots + \frac{1}{2}\hat{T}_1^2 + \frac{1}{2}\hat{T}_1\hat{T}_2 + \frac{1}{2}\hat{T}_2\hat{T}_1 + \dots) \quad (58)$$

that can be collected into n -fold excitation operators \hat{C}_n :

$$\hat{C}_1 = \hat{T}_1 \quad (59)$$

$$\hat{C}_2 = \frac{1}{2}\hat{T}_1^2 + \hat{T}_2 \quad (60)$$

$$\hat{C}_3 = \frac{1}{3!}\hat{T}_1^3 + \frac{1}{2}\hat{T}_1\hat{T}_2 + \hat{T}_3 \quad (61)$$

and (54) can be rewritten as

$$|\Psi_{CC}\rangle = |\Phi\rangle + \sum_r c_a^r |\Phi_a^r\rangle + \sum_{\substack{a<b \\ r<s}} c_{ab}^{rs} |\Phi_{ab}^{rs}\rangle + \sum_{\substack{a<b<c \\ r<s<t}} c_{abc}^{rst} |\Phi_{abc}^{rst}\rangle + \dots \quad (62)$$

which is exactly the FCI wave function (34). The coefficients are here given by

$$c_a^r = t_a^r \quad (63)$$

$$c_{ab}^{rs} = t_{ab}^{rs} + t_a^r * t_b^s \quad (64)$$

$$c_{abc}^{rst} = t_{abc}^{rst} + t_a^r * t_{bc}^{st} + t_a^r * t_b^s * t_c^t \quad (65)$$

where $*$ is an antisymmetric product with respect to exchange of occupied orbitals and exchange of unoccupied orbitals. (e.g. $t_a^r * t_b^s = t_a^r t_b^s - t_b^r t_a^s$). Since the exponential ansatz gives back the FCI wave function with the complication of an additional non-linear parametrization of the amplitudes, it has little interest as it is. Truncation of the CC ansatz is however an improvement over truncation of the CI wave function [80]. The truncation is done on the CC operator, and for example the Coupled Cluster Single Double (CCSD) ansatz is given by retaining only \hat{T}_1 and \hat{T}_2 in \hat{T} . Contrary to the CISD case, triple and higher excitations are still present through product of the single and double substitution operators, while

the coefficients of those are expressed as products of the single and double CC amplitudes t_a^r, t_{ab}^{st} only. The correlation energy is readily found by projecting the Schrödinger equation for $|\Psi_{CC}\rangle$ on $|\Phi_0\rangle$:

$$E\langle\Phi_0|\Psi_{CC}\rangle = \langle\Phi_0|H|\Psi_{CC}\rangle \quad (66)$$

$$E = \langle\Phi_0|H|\Phi_0\rangle + \sum_{\substack{a<b \\ r<s}} c_{ab}^{rs} \langle\Phi_0|H|\Phi_{ab}^{rs}\rangle \quad (67)$$

so that

$$E_c = \frac{1}{4} \sum_{a,b,r,s} (t_{ab}^{rs} + 2t_a^r t_b^s) (v_{abrs} - v_{absr}). \quad (68)$$

Note that this equation is valid for any truncation order of the CC ansatz, the correlation energy is expressed only as function of the single and double CC amplitudes. Those are obtained by projecting the CC Schrödinger equation on single, double, triple, ..., n-substituted Slater determinants for a n truncated CC ansatz. Note that due to the exponential nature of the ansatz, the excitations degrees that were lost in the size-consistency check of the CID wave function are recovered in the CCD one, with additional terms involving single excitations of each fragment A and B. Those last terms however cancels when computing the amplitudes, and the CC method is size-consistent when constructed from a size-consistent reference.

The CC ansatz can also be treated by means of Rayleigh-Schrödinger perturbation theory. The wave function is expanded as in eq. (38) with

$$|\Phi_i^{(k)}\rangle = \sum_{\mu} c_{\mu}^{(k)} \hat{C}_{\mu}^{(k)} |\Phi_i^{(0)}\rangle \quad (69)$$

where c_{μ} design the coefficient associated to the excitation operator \hat{C}_{μ} . To first order in the ground state we have

$$E^{(1)} |\Phi\rangle = (\hat{F} - E^{(0)}) |\Phi_0^{(1)}\rangle + (\hat{V}_{ee} - \hat{V}_{KS}) |\Phi\rangle \quad (70)$$

so that by projecting onto a doubly excited determinant we find

$$\langle\Phi_{ab}^{rs} | (\hat{F} - E^{(0)}) |\Phi_0^{(1)}\rangle = -\langle\Phi_{ab}^{rs} | (\hat{V}_{ee} - \hat{V}_{KS}) |\Phi\rangle \quad (71)$$

$$\langle\Phi_{ab}^{rs} | \hat{F} c_2^{(1)} \hat{C}_2^{(1)} |\Phi\rangle - E^{(0)} \langle\Phi_{ab}^{rs} | c_2^{(1)} \hat{C}_2^{(1)} |\Phi\rangle = -(v_{abrs} - v_{absr})^* \quad (72)$$

$$c_{abrs}^{(1)} = \frac{(v_{abrs} - v_{absr})^*}{\epsilon_a + \epsilon_a - \epsilon_r - \epsilon_s} \quad (73)$$

If we consider the Coupled Cluster Double (CCD) ansatz we recover by inserting (73) into (68) the MP2 correlation energy. The **CCSD(T)** method consist in solving for the CC amplitudes up to second order, then computing the perturbative triple amplitude to second order by using the CCSD amplitudes (rather than the perturbations to second order), and finally correcting the CCSD energy by the third order perturbative energy term.

2.2.7 Basis sets

For electronic molecular calculations, it is customary to use atomic orbitals basis sets. Pople basis sets were constructed to approximate Slater-type orbitals (STO) [80, 184]. Those are nodeless hydrogen-like orbitals that possess an exponential decrease with the distance from the nuclei and thus satisfy the nuclear cusp condition.⁵ To facilitate the numerical evaluation of the electrons integrals in (23-25), these orbitals can be approximated by linear combinations of Gaussian functions, which possess a $e^{-\alpha r^2}$ radial dependence. The **STO-LG** basis set are **minimal basis sets**, meaning they use only one STO for each atomic orbital, up to the valence shell. The STO are formed by linear combinations of L Gaussian functions. For example, the STO-3G basis for the Lithium is made of five atomic orbitals ($\phi_{1s}, \phi_{2s}, \phi_{2p_x}, \phi_{2p_y}, \phi_{2p_z}$, each of which is a sum of three Gaussian functions. **Double zeta** basis sets are constructed by adding one STO to each orbital, usually only to the valence ones (**split-valence basis sets**). The 3-21G basis for the Lithium is made of nine orbitals: one ϕ_{1s} core orbital made of three gaussian (3-21G), and four couple of valence orbitals $\phi_{2s}^1, \phi_{2s}^2, \phi_{2^1 p_x}, \phi_{2^2 p_x}$, dots, where the ϕ_k^1 are made of two Gaussian functions, (3-21G), and the ϕ_k^2 are made of one Gaussian function (3-21G). To have more flexibility in describing the effect on atoms of the non-uniform electric field due to the others atoms, **polarizations functions** can be added. Those are orbitals with higher angular momentum, p-like for Hydrogen, d-like for Lithium, etc. The 3-21G* add thus five d_k -like orbitals made of one Gaussian function each, for a total of fourteen orbitals. The 3-21G** add again p-like orbitals to the core shell.

In order to have more reliable post Hartree-Fock calculations, it is necessary to augment the number of virtual orbitals, while accounting for correlation [80, 184]. A first treatment of the sort it to diagonalize the one-particle reduced density matrix $\rho(\mathbf{x}, \mathbf{x}') = \langle \hat{\psi}^\dagger(\mathbf{x}) \hat{\psi}(\mathbf{x}') \rangle$. The obtained eigenvectors are called **natural orbitals** (NO), and the associated eigenvalues are occupation numbers. They permit a faster convergence of the post-Hartree Fock methods [80, 184]: one perform for example a truncated CI calculation on top of the Hartree-Fock canonical orbitals, after what the natural orbitals are computed. Only the ones with the highest occupation numbers are kept, and the procedure is repeated. Similarly, **Atomic Natural Orbitals** (ANO) can be computed from post-Hartree Fock methods on atoms alone, and subsequent molecular calculations are performed while retaining only a subset of such ANO.⁶

⁵ The divergent Coulomb interaction at coinciding positions of particles in the wave function must be compensated by a non-differentiability of the wave function, that will in turn produce an opposite divergent kinetic energy term [80].

⁶ Note that since the reduced density matrix has positive or zeros eigenvalues [80], its diagonalization is equivalent to a Singular Value Decomposition [128], and keeping only the highest eigenvalues is a dimensionality reduction technique (it is a Principal Component Analysis, see Section 3.4).

The **correlation-consistent basis sets** were constructed by Dunning [53] in a similar way than that of ANO, but the orbitals retained in each stage of their hierarchy are those with similar contributions to the correlation energy. The *correlation-consistent polarized valence basis sets* are denoted by cc-pVXZ, where "cc" stands for correlation consistent, "p" for polarization, "V" for valence (only valence electron correlation has been considered in their design), and "XZ" corresponds to double zeta (DZ), triple zeta (TZ), etc.

2.2.8 Methods comparison

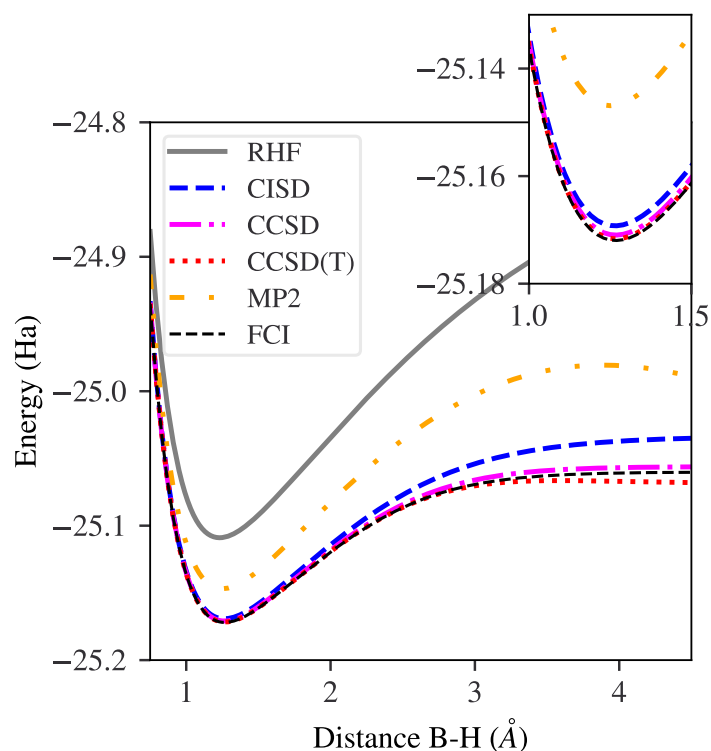


Figure 6: Dissociation curve of the BH molecule using various wave function methods presented in this chapter. Calculations were realized with Nwchem [5] in the 6-31G basis set.

In Figure 6 is shown a dissociation curve of the closed-shell Hydrogen Bore molecule, using different methods presented *supra*. The FCI curve, is the exact one in the 6-31G basis set considered here. The RHF solution is qualitatively correct near equilibrium but obviously lack size-consistency in the dissociation limit as both Bore and Hydrogen are open-shell atoms. The CISD wave function, is a net improvement toward the FCI reference, while it overestimates the atomization energy. The MP2 energy is also a good description. The atomization energy

overestimation is bigger in MP2, together with a too short equilibrium bond, as it is constructed on Hartree Fock orbitals only, while the orbitals are relaxed in CISD. The coupled cluster energies are the best approximation on this small system, with CCSD(T) giving the most precise atomization energy. Note that both MP2 and CCSD(T) are not qualitatively well behaved near the dissociation limit: neither is variational, the Møller–Plesset perturbative potential (also present in the triple (T) perturbation of CCSD(T) cannot be considered a small perturbation near the dissociation.

Note that we only considered single-reference methods in this thesis. Multi-reference methods are extension of the former where additional determinants serve as reference upon which the correlation is restored. This is particularly efficient for static correlation and the correct description of bond breaking.

2.3 DENSITY FUNCTIONAL THEORY

The previous methods are commonly called wave-function methods. They rely on constructing the missing correlation energy from the Hartree-Fock single Slater determinant. While they can produce very accurate results, they still possess an unfavorable numerical scaling with the number of electrons. This renders them impracticable to use for moderate size systems (more than hundreds of electrons). In Density functional theory (DFT), the interest is shifted from the wave function to the *density*: the many electron problem is reformulated as a variational problem over the density of electrons. The latter being a function of only three spatial coordinates, while the wave function is a N-fold tensor product of one-electron functions of three spatial coordinates (without considering spin), this is an enormous simplification. In practice however, the exact formulation of DFT involves an unknown universal functional of the density. A self-consistent method has been devised to obtain quantitative results from this theoretical framework, in which the many electron problem is mapped onto a non interacting one with the same density. This non interacting system must be solved exactly, leading to a cubic scaling with the basis set size, still more efficient than the quartic scaling of Hartree-Fock.⁷ In this context, DFT has become the very basis of material simulations and systems with thousands of atoms can be studied routinely [131].

2.3.1 Hohenberg-Kohn theorems

Let us rewrite here the electronic Hamiltonian of eq. (4) by dropping the e subscript:

$$\hat{H} = \hat{T} + \hat{V}_{ee} + \hat{V}_{\text{ext}} \quad (74)$$

⁷ In this chapter and the last one, we do not discuss recent linear scaling algorithms, see *e.g.* ref. [129] for DFT, and ref. [138] for CCSD(T).

and let $\hat{V}_{\text{ext}} = \sum_{i=1}^N v(\mathbf{r}_i)$ be a local (*i.e.* multiplicative) potential. Since we are not interested in spin polarized systems or non-collinear spin systems, we will drop the spin variable σ_i to alleviate the notation, but we shall keep in mind a missing sum over this variable. Moreover, the $4\pi\epsilon_0$ factor in the Coulomb potential will be dropped from now on. For any N -electron wave function $\Psi(\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_N)$ the electron density is given by⁸

$$n(\mathbf{r}) = N \int d\mathbf{r}_2 \cdots d\mathbf{r}_N |\Psi(\mathbf{r}, \mathbf{r}_2, \dots, \mathbf{r}_N)|^2 \quad (75)$$

2.3.1.1 First Hohenberg-Kohn theorem

The first of Hohenberg-Kohn theorems reads [89]:

Theorem 1 *There exist a one-to-one mapping between the external potential \hat{V}_{ext} , the non-degenerate ground state, and its corresponding density.*

The proof proceeds by assuming that two external potentials \hat{V}_{ext}^1 and \hat{V}_{ext}^2 that differ by more than a constant shift lead to the same density $n_1(\mathbf{r}) = n_2(\mathbf{r}) = n(\mathbf{r})$. Two different external potentials corresponds to two different Hamiltonians \hat{H}_1 and \hat{H}_2 , which themselves correspond to two distinct ground states $|\Psi_1\rangle$ and $|\Psi_2\rangle$.⁹ We have then

$$\begin{aligned} \langle \Psi_1 | \hat{H}_1 | \Psi_1 \rangle &< \langle \Psi_2 | \hat{H}_1 | \Psi_2 \rangle \\ E_1 &< \langle \Psi_2 | \hat{H}_2 + \hat{V}_{\text{ext}}^1 - \hat{V}_{\text{ext}}^2 | \Psi_2 \rangle \\ E_1 &< E_2 + \int n(\mathbf{r})(v_1(\mathbf{r}) - v_2(\mathbf{r})) \end{aligned}$$

and reciprocally

$$E_2 < E_1 + \int n(\mathbf{r})(v_2(\mathbf{r}) - v_1(\mathbf{r}))$$

so that

$$E_1 + E_2 < E_1 + E_2.$$

Since this is impossible, we must have $n_1(\mathbf{r}) \neq n_2(\mathbf{r})$.

This unique correspondence between the Hamiltonian (through the external potential) and the ground state density allows one to write either the ground state or the ground state energy as a functional of the ground state density, *i.e.*

$$|\Psi_0\rangle = |\Psi[n_0]\rangle \quad (76)$$

$$E_0 = E[n_0] \quad (77)$$

⁸ or alternatively by $n(\mathbf{r}) = \langle \hat{n}(\mathbf{r}) \rangle = \langle \Psi | \sum_i \delta(\mathbf{r} - \mathbf{r}_i) | \Psi \rangle$

⁹ If $|\Psi_1\rangle = |\Psi_2\rangle$, then $(\hat{V}_{\text{ext}}^2 - \hat{V}_{\text{ext}}^1) |\Psi_1\rangle = (\hat{T} + \hat{V}_{ee} + \hat{V}_{\text{ext}}^2) |\Psi_1\rangle - (\hat{T} + \hat{V}_{ee} + \hat{V}_{\text{ext}}^1) |\Psi_1\rangle = (E_2 - E_1) |\Psi_1\rangle$ so that both potential only differ by a constant in contradiction with the assumption.

It is important to further note that this correspondence also says that *in principle* all properties of the system are determined by the ground state density [129].

2.3.1.2 Second Hohenberg-Kohn theorem

Theorem 2 *The ground state energy associated to a given external potential is the minimum value of the energy functional of the density at the point where the latter coincides with the ground state density.*

Let us first define the functional

$$F[n(\mathbf{r})] = \langle \Psi[n] | \hat{T} + \hat{V}_{ee} | \Psi[n] \rangle. \quad (78)$$

Note that this functional is *universal*, as it does only depend on the electron number, all the specifics of the system under consideration being in the external potential. The energy functional $E[n(\mathbf{r})]$ is now defined as

$$E[n] = F[n] + \int d\mathbf{r} v(\mathbf{r}) n(\mathbf{r}) \quad (79)$$

and we have that

$$\begin{aligned} \langle \Psi[n] | \hat{H}[v] | \Psi[n] \rangle &\geq E_0 \\ \langle \Psi[n] | \hat{T} + \hat{V}_{ee} | \Psi[n] \rangle + \langle \Psi[n] | \hat{V}_{ext} | \Psi[n] \rangle &\geq E_0 \\ \langle \Psi[n] | \hat{T} + \hat{V}_{ee} | \Psi[n] \rangle + \int d\mathbf{r} v(\mathbf{r}) n(\mathbf{r}) &\geq E_0 \\ F[n] + \int d\mathbf{r} v(\mathbf{r}) n(\mathbf{r}) &\geq E_0 \\ E[n] &\geq E_0 \end{aligned}$$

which gives the second theorem according to which

$$\min_n E[n] = E[n_0] = E_0 \quad (80)$$

Thus, in principle, to solve the Schrödinger equation for the electronic ground state, one simply needs to minimize the functional $E[n]$ with respect to the density $n(\mathbf{r})$. The functional $F[n]$ is however not known analytically and this remains an open problem [129]. To face this problem, a practical scheme was devised by Kohn and Sham in 1965 [109], one year after Hohenberg and Kohn's paper.

2.3.2 Kohn-Sham Density functional theory

The Kohn-Sham scheme for DFT relies on the following assumption:

It is possible to associate to the exact ground state density a non-interacting auxiliary system with the same ground state density

Although this statement remains unproven for general purposes [129] we will assume its validity.

The HK theorems also apply to non-interacting system. Let $|\Phi\rangle$ belong to the space of ground-states of such non-interacting systems in correspondence with external potentials.

Since there are no electrons-electrons interactions, those kets $|\Phi\rangle$ are single Slater determinants in some basis $\{\phi_i\}$, in accordance with Section 2.2:

$$|\Phi\rangle = \frac{1}{\sqrt{N!}} \hat{A} |\phi_1 \phi_2 \cdots \phi_N\rangle. \quad (81)$$

Their density reads

$$n(\mathbf{r}) = \sum_i |\phi_i(\mathbf{r})|^2 \quad (82)$$

One can now redefine the universal functional $F[n]$ as

$$F[n] = T_s[n] + E_{Hxc}[n] \quad (83)$$

where $T_s[n]$ is the kinetic energy functional of the auxiliary system:

$$T_s[n] = \left\langle \Phi[n] \left| \frac{-\hbar^2 \nabla_{\mathbf{r}}^2}{2m} \right| \Phi[n] \right\rangle \quad (84)$$

E_{Hxc} is further decomposed as

$$E_{Hxc}[n] = E_H[n] + E_{xc}[n] \quad (85)$$

with

$$E_H[n] = \frac{1}{2} \int d\mathbf{r} d\mathbf{r}' \frac{n(\mathbf{r}')n(\mathbf{r})}{|\mathbf{r} - \mathbf{r}'|}. \quad (86)$$

$E_H[n]$ is the so-called Hartree functional and corresponds to the mean-field electrostatic repulsion energy. In the Hartree-Fock formalism we had $E_{HF} = \langle \hat{T} \rangle_{HF} + \langle \hat{V}_{ee} \rangle_{HF}$ where $\langle \hat{V}_{ee} \rangle_{HF}$ decompose into the subtraction of a direct and exchange term. $E_H[n]$ reproduce the direct term but includes a spurious self-energy term by missing the exchange term. $E_{xc}[n]$ is called the *exchange correlation functional* and reads explicitly

$$E_{xc}[n] = T[n] - T_s[n] + V_{ee}[n] - E_H[n]. \quad (87)$$

It contains the difference between the kinetic energy of the interacting and non-interacting system, as well as the difference between the Coulomb interaction and the mean-field Hartree term. As such, all the intricacies of the interacting many-body problem are contained in this functional.

In order to minimize $E[n]$ with the universal functional of eq. (83), let $\{\phi_i\}$ be a set of orbitals subject to the constraint $\langle \phi_j | \phi_i \rangle = \delta_{ij}$. One can construct the following Lagrangian functional:

$$\mathcal{L}[\{\phi_i\}] = E[\{\phi_i\}] - \sum_{ij} \epsilon_{ij} (\langle \phi_j | \phi_i \rangle - \delta_{ij}). \quad (88)$$

By cancelling the functional derivatives $\partial \mathcal{L} / \partial \phi_i$, one obtains the Kohn-Sham equations:

$$\left(-\frac{\hbar^2 \nabla^2}{2m} + v_{\text{KS}}(\mathbf{r}) \right) \phi_i = \epsilon_i \phi_i \quad (89)$$

with the Kohn-Sham potential $v_{\text{KS}}(\mathbf{r})$ being defined as

$$v_{\text{KS}}(\mathbf{r}) = v_{\text{ext}}(\mathbf{r}) + \frac{\partial E_{\text{Hxc}}[n]}{\partial n(\mathbf{r})} \quad (90)$$

$$= v_{\text{ext}}(\mathbf{r}) + \int d\mathbf{r}' \frac{n(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} + \frac{\partial E_{\text{xc}}[n]}{\partial n(\mathbf{r})} \quad (91)$$

$$= v_{\text{ext}}(\mathbf{r}) + v_{\text{H}}(\mathbf{r}) + v_{\text{xc}}(\mathbf{r}) \quad (92)$$

Thus according to the Kohn-Sham assumption, there exist a correspondence between the exact interacting many-electron problem and the Kohn-Sham auxiliary non interacting system with Hamiltonian

$$\hat{H}_{\text{KS}} = -\frac{\hbar^2 \hat{\nabla}^2}{2m} + \hat{V}_{\text{KS}} \quad (93)$$

If the exchange-correlation functional was known, it would suffice to solve the later to obtain the exact ground-state density of the interacting problem. Unfortunately the exchange-correlation functional, so as the universal $F[n]$ functional, remains unknown.

In the next section we will see some approximations to $E_{\text{xc}}[n]$, but before to go on, let mention a few practical points:

2.3.2.1 Self consistent field procedure

In the Kohn-Sham Hamiltonian, the potential depends upon the basis functions ϕ_i , so that in practice, once an approximation has been chosen for $E_{\text{xc}}[n]$, the Kohn-Sham equations needs to be solved in a self-consistent fashion, as illustrated in Figure 7

2.3.2.2 Basis sets

Until now we have not discussed periodic potentials that appears in solids. We will restrict ourselves to one dimension for simplicity. To describe such systems, the crystal is approximated as the periodic repetition of N unit cells of length a ,

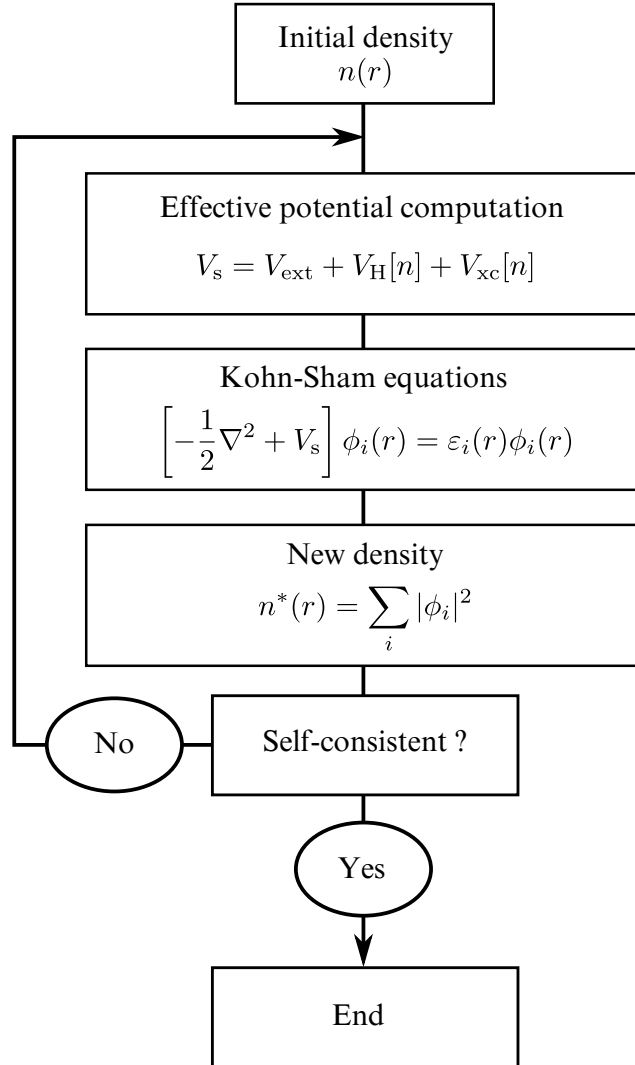


Figure 7: Self consistent field procedure to find the ground state eigenvectors of the Kohn-Sham Hamiltonian associated to a given external potential V_{ext} in a given one-electron basis set $\{\phi_i\}$.

with periodic boundary conditions. The one-electron wave function is thus such that $\psi(x + Na) = \psi(x)$. The translation operator is $T(na) = e^{na \frac{d}{dx}} = e^{-ina\hat{p}/\hbar}$ with eigenvalues e^{ikna} . It commutes with the Hamiltonian so that it exist a simultaneous basis $\phi(x)$ with $H_e T(na)\phi(x) = E\phi(x+a) = Ee^{ikna}\phi(x)$. Moreover the boundary condition imposes $k = \frac{2\pi m}{Na}$ where m is an integer. The function $u_k(x) = e^{-ikx}\phi(x)$ is such that $T(na)u_k(x) = e^{-ik(x+a)}\phi(x+a) = u_k(x)$. The one-particle solutions of the Schrödinger equation in a periodic potential can thus be written as (**Bloch theorem**):

$$\phi_k(x) = e^{ikx}u_k(x) \quad (94)$$

where $u_k(x)$ possesses the periodicity of the potential. k can be restricted to the values $\frac{2\pi m}{Na}$, $-\frac{N}{2} < m < \frac{N}{2}$, if we add another quantum number n to specify the congruence. The reduced wavenumber values defines the Brillouin zone, and the n number defines the band index.

Since the function $u_{nk}(x)$ is periodic, it can be further decomposed into plane waves through Fourier decomposition. Close to the nuclei, a high number of plane-waves is necessary to reproduce the correct one-electron wave functions, and pseudo-potentials were devised to alleviate the associated computational burden, thus allowing accurate calculations with lower energy cutoff.

2.3.2.3 Pseudo-potentials

The idea of pseudo-potential (PP) is to use an effective potential for the nuclei and the core electrons, only to consider the valence electrons left, interacting with the effective core potential (see Figure 8). They are constructed by solving an approximate all-electron problem for the atoms alone, and come in different flavors: **norm-conserving** PP impose the equality of the charge integral inside the core radius R_c of the all-electron valence wave functions and the corresponding wave functions obtained with the PP (pseudo-wave function) [129]. **Ultrasoft** PP release this constraint and report it into the Kohn-Sham eigenvalue problem [129]. **Projector augmented waves** (PAW), employ a similar strategy than the ultrasoft PP but reproduce the true oscillatory behavior of the valence all-electron wave function. They are implemented in the code VASP [113] which was used for the results of chapters 4 and 5

2.3.3 Exchange correlation

While the true exchange functional is not known, we can obtain for it a formal expression that will be convenient to discuss beyond-DFT approaches. We will use the **adiabatic connection** concept [79]. The idea is to define a Hamiltonian parametrized by λ , decomposed into a non-interacting (mean-field) one-body Hamiltonian \hat{H}_0 and a many-body Hamiltonian $\hat{H}_1(\lambda)$. By slowly (adiabatically) varying λ , one can interpolate between the mean-field and the fully interacting problem. It is moreover assumed that the λ -parametrized external potential is such that the **density is kept constant** during the adiabatic switch [55].

$$\hat{H}(\lambda) = \hat{H}_0 + \lambda \hat{H}_1(\lambda) = \hat{T} + \hat{V}_\lambda^{\text{ext}} + \lambda \hat{V}_{ee}, \quad (95)$$

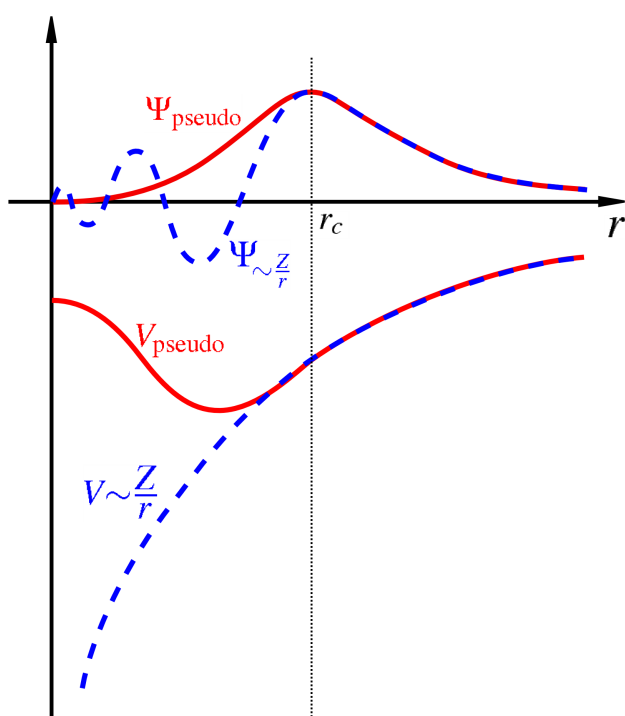


Figure 8: Illustration of the pseudo-potential method: the real diverging Coulomb potential (dashed blue line, bottom), is replaced by an effective potential (red line, bottom) inside some cutoff radius around the nuclei. The highly oscillatory behavior of the wave function (dashed blue line, above) is thus suppressed in this region (red line, above), and a smaller number of plane waves are necessary.

with

$$\hat{H}_0 = \hat{T} + \hat{V}_{\lambda=0}^{\text{ext}} \quad (96)$$

$$\hat{H}_1 = \hat{V}_{ee} + \frac{1}{\lambda} [\hat{V}_{\lambda}^{\text{ext}} - \hat{V}_{\lambda=0}^{\text{ext}}] \quad (97)$$

$$\hat{V}_{\lambda=0}^{\text{ext}} = \mathbf{v}^{\text{ext}} + \hat{V}^{\text{MF}} \quad (98)$$

$$\hat{V}_{\lambda=1}^{\text{ext}} = \mathbf{v}^{\text{ext}} \quad (99)$$

$$(100)$$

From the Hellmann-Feynman theorem we have that

$$\frac{dE_{\lambda}}{d\lambda} = \langle \psi_{\lambda} | \frac{d\hat{H}_{\lambda}}{d\lambda} | \psi_{\lambda} \rangle \quad (101)$$

so that

$$E(\lambda = 1) = \langle \psi_0 | H_0 | \psi_0 \rangle + \int_0^1 d\lambda \langle \psi_{\lambda} | \left[\hat{H}_1(\lambda) + \lambda \frac{d\hat{H}_1(\lambda)}{d\lambda} \right] | \psi_{\lambda} \rangle \quad (102)$$

Hence the true energy $E(\lambda = 1)$ is recovered from the mean-field ground state $|\psi_0\rangle$. This is closely related to the Gell-Mann and Low theorem mentioned in subsection 2.2.4. As in the later, level-crossing (from symmetry breaking for example) must be absent on the adiabatic path for this scheme to work (in particular, one must have $\langle \psi_0 | \psi_1 \rangle \neq 0$ [130]).

Going on, we have that

$$E = E_0 + \int_0^1 d\lambda \langle \psi_{\lambda} | \hat{V}_{ee} | \psi_{\lambda} \rangle + \langle \psi_{\lambda} | \frac{d\hat{V}_{\lambda}^{\text{ext}}}{d\lambda} | \psi_{\lambda} \rangle \quad (103)$$

$$= E_0 + \int_0^1 d\lambda \langle \psi_{\lambda} | \hat{V}_{ee} | \psi_{\lambda} \rangle + \int d\mathbf{r} n(\mathbf{r}) (v_1^{\text{ext}} - v_0^{\text{ext}}) \quad (104)$$

where we used the fact that the density is kept constant when going from the first to the second line. From second quantization we can rewrite $\langle \hat{V}_{ee} \rangle$ as [56]

$$\begin{aligned} \langle \hat{V}_{ee} \rangle &= \frac{1}{2} \int d\mathbf{r} d\mathbf{r}' \frac{1}{|\mathbf{r} - \mathbf{r}'|} (\langle \hat{n}(\mathbf{r}) \hat{n}(\mathbf{r}') \rangle - \langle \hat{n}(\mathbf{r}) \rangle \langle \hat{n}(\mathbf{r}') \rangle) \\ &= \frac{1}{2} \int d\mathbf{r} d\mathbf{r}' \frac{1}{|\mathbf{r} - \mathbf{r}'|} [\langle \delta \hat{n}(\mathbf{r}) \delta \hat{n}(\mathbf{r}') \rangle + \langle \hat{n}(\mathbf{r}) \rangle \langle \hat{n}(\mathbf{r}') \rangle - \langle \hat{n}(\mathbf{r}) \rangle \langle \hat{n}(\mathbf{r}') \rangle] \end{aligned} \quad (105)$$

with the density fluctuation

$$\delta \hat{n}(\mathbf{r}) = \hat{n}(\mathbf{r}) - \langle \hat{n}(\mathbf{r}) \rangle. \quad (106)$$

From Kohn-Sham DFT we have

$$E[n] = T_s[n] + \int d\mathbf{r} n(\mathbf{r}) v^{\text{ext}}(\mathbf{r}) + E_H[n] + E_{xc}[n] \quad (107)$$

As the density is kept constant during the integration, we can identify the middle term in eq. (105) with $E_H[n]$, so that

$$E_{xc}[n] = \frac{1}{2} \int_0^1 d\lambda \int d\mathbf{r} d\mathbf{r}' \frac{\langle \delta \hat{n}(\mathbf{r}) \delta \hat{n}(\mathbf{r}') \rangle_\lambda - n(\mathbf{r}) \delta(\mathbf{r} - \mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} \quad (108)$$

Let us now separate the exchange correlation energy in an exchange and a correlation contribution :

$$E_{xc} = E_x + E_c \quad (109)$$

The total exchange energy is defined as [55, 129]

$$E_x = \langle \psi_0 | \hat{V}_{ee} | \psi_0 \rangle - E_H[n] = \langle V_{ee} \rangle_0 - E_H[n] \quad (110)$$

where $|\psi_0\rangle$ is a the Kohn-Sham Slater determinant. This definition allows to write the correlation energy as

$$E_c = E - \langle \psi_0 | \hat{H} | \psi_0 \rangle \quad (111)$$

in complete analogy with the definition of eq. (33).

Using eq. (105), we can rewrite eq. (110) as

$$E_x = \frac{1}{2} \int d\mathbf{r} d\mathbf{r}' \frac{1}{|\mathbf{r} - \mathbf{r}'|} [\langle \delta \hat{n}(\mathbf{r}) \delta \hat{n}(\mathbf{r}') \rangle_0 - n(\mathbf{r}) \delta(\mathbf{r} - \mathbf{r}')] \quad (112)$$

and finally

$$E_c = \frac{1}{2} \int_0^1 d\lambda \int d\mathbf{r} d\mathbf{r}' \frac{1}{|\mathbf{r} - \mathbf{r}'|} [\langle \delta \hat{n}(\mathbf{r}) \delta \hat{n}(\mathbf{r}') \rangle_\lambda - \langle \delta \hat{n}(\mathbf{r}) \delta \hat{n}(\mathbf{r}') \rangle_0] \quad (113)$$

The correlation energy is thus expressed as the integral of the difference between the density-density correlation function of the interacting and non-interacting system times the Coulomb kernel while the interaction is gradually turned on.

2.3.4 Approximations to the exchange correlation energy

One way to approximate the exchange correlation is to replace the true system by an idealization in eq. (108). The homogeneous electron gas (HEG or "jellium") is one such approximate system. It consists of N interacting electrons, with a positive uniform background charge to model the nuclei. The HEG energy is function of the density, or equivalently of the average dimensionless distance r_s between electrons (or Wigner-Seitz radius). The later is defined by the relation

$$\frac{4}{3} \pi (r_s a_0)^3 = \frac{1}{n} \quad (114)$$

with $a_0 = \frac{\hbar^2}{m_e e^2}$ the Bohr radius. While being an approximate model of reality, the HEG is not analytically tractable, but numerical results of its correlation energy

$e_c^{\text{HEG}}(r_s)$ have been computed and are tabulated [55, 129]. As an example, in the high-density limit, one can find the following energy density [130]

$$e_c^{\text{HEG}}(r_s \ll 1) \sim \frac{e^2 n}{a_0} (0.0622 \ln r_s - 0.096). \quad (115)$$

The exchange energy density is however known and reads [55]

$$e_x^{\text{HEG}}(n) = -\frac{3(3\pi^2)^{1/3} e^2}{4\pi} \int d\mathbf{r} n(\mathbf{r})^{4/3}. \quad (116)$$

The **local density approximation (LDA)** consists in using those values of the exchange and correlation energy of the HEG in $E_{xc}[n(\mathbf{r})]$.

$$E_{xc}^{\text{LDA}}[n] = \int d\mathbf{r} [e_x^{\text{HEG}}(n) + e_c^{\text{HEG}}(n)] n(\mathbf{r}) \quad (117)$$

That is, for each point \mathbf{r} , the XC energy is given by the HEG XC energy for the corresponding density $n(\mathbf{r})$, hence the "local" in the name. This scheme is expected to perform well for approximately free electrons systems such as conduction electrons in metals and poorly for localized electronic system such as atoms [129].

The next class of approximations are coined **generalized gradient approximations (GGA)**, and consist in adding a dependence on the density gradient through some function f :

$$E_{xc}^{\text{GGA}}[n] = \int d\mathbf{r} n(\mathbf{r}) e_{xc}^{\text{HEG}}(n) f(n, \nabla n) \quad (118)$$

Different such functionals have been proposed, one of the most widely used [55, 129] being the GGA functional of Perdew, Burke and Ernzerhof (PBE) [143], which will be used in this thesis.

This gradient dependence allows a better treatment of inhomogeneous systems and improve upon the local aspect of the LDA (GGA functional are usually referred as semi-local functionals [55]). There exist also **Meta-GGA** functionals which add a dependence on the density laplacian $\nabla^2 n$ and on the non-interacting kinetic energy density $\tau(\mathbf{r}) \propto \sum_i |\nabla \phi_i(\mathbf{r})|^2$.

Finally let us mention the **hybrid functionals** that combine one functional (either LDA/GGA/Meta-GGA) with the so-called exact Hartree-Fock exchange energy¹⁰

$$E_x^{\text{HF}} = -\frac{1}{2} \sum_{i,j} \sum_{\sigma} \int d\mathbf{r} d\mathbf{r}' \frac{\phi_{i\sigma}^*(\mathbf{r}) \phi_{j\sigma}^*(\mathbf{r}') \phi_{j\sigma}(\mathbf{r}) \phi_{i\sigma}(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|}. \quad (119)$$

The idea behind this is that $E_{xc} = \int_0^1 d\lambda \langle \hat{V}_{ee} \rangle_{\lambda} - E_{\text{H}}$ from eq. (108) so that by turning on the interaction one interpolates between $E^{\text{HF}} - E_{\text{H}} = E_x^{\text{HF}}$ and the fully interacting E_{xc} [129]. The PBE0 functional for example is given by

$$E_{xc}^{\text{PBE0}} = \frac{1}{4} E_x^{\text{HF}} + \frac{3}{4} E_x^{\text{PBE}} + E_c^{\text{PBE}} \quad (120)$$

¹⁰ Or sometime simply named exact exchange. It is exact in the sense that it is the same expression as for the Hartree-Fock exchange but the two do not coincide as the orbitals differ. Note that spin has been explicitly re-introduced in this expression.

Note that GGA, Meta-GGA or hybrid functionals all incorporate the LDA approximation in their core.

2.3.5 Random Phase Approximation

An alternative route to the LDA and subsequent approximations is to perform an approximation on the exact form of the exchange-correlation of eq. (113). Let us define the time-ordered density correlation function

$$\chi(\mathbf{r}, \mathbf{t}, \mathbf{r}', \mathbf{t}') = -i\langle\Psi|T\delta\hat{n}(\mathbf{r}, \mathbf{t})\delta\hat{n}(\mathbf{r}', \mathbf{t}')|\Psi\rangle \quad (121)$$

which is related to the retarded density correlation function

$$\chi_R(\mathbf{r}, \mathbf{t}, \mathbf{r}', \mathbf{t}') = -i\Theta(\mathbf{t} - \mathbf{t}')\langle\Psi|[\delta\hat{n}(\mathbf{r}, \mathbf{t}), \delta\hat{n}(\mathbf{r}', \mathbf{t}')]\Psi\rangle \quad (122)$$

defined in linear response theory as the kernel of the density response to an external perturbation ($\langle\delta\hat{n}\rangle = \int\chi_R\hat{V}$) [55, 130]. The time-ordered density correlation is subject to the following Dyson equation written (in Fourier space, for the **homogeneous** problem)

$$\chi(\mathbf{q}, \omega) = \Pi(\mathbf{q}, \omega) + \Pi(\mathbf{q}, \omega)V_{ee}(\mathbf{q})\chi(\mathbf{q}, \omega) \quad (123)$$

with $\Pi(\mathbf{q}, \omega)$ the irreducible polarization, sum of all processes beginning by the creation of a particle-hole pair and ending by its annihilation, such that their corresponding Feynman diagrams are not disconnected when cutting one interaction line. The polarization equally defines the effective, dressed or **screened interaction** V_{eff} by

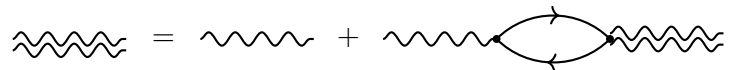
$$V_{eff}(\mathbf{q}) = V_{ee}(\mathbf{q}) + V_{ee}(\mathbf{q})\Pi(\mathbf{q}, \omega)V_{eff}(\mathbf{q}). \quad (124)$$

In this context, the random phase approximation (RPA) consist in keeping only the zeroth order term in the interaction for the polarization, which correspond to the time-ordered correlation function of the unperturbed system:

$$\Pi^{RPA} \equiv \chi_0 \equiv \text{---} \text{---} \text{---} \quad (125)$$


And the unperturbed response function χ_0 is given as function of the unperturbed one-electron orbitals. Eqs. (123) and (124) in the random phase approximation are diagrammatically depicted as [55, 56, 130]

$$\text{---} \text{---} \text{---}^{RPA} = \text{---} \text{---} \text{---} + \text{---} \text{---} \text{---} \text{---} \text{---} \text{---}^{RPA} \quad (126)$$


$$\text{---} \text{---} \text{---} = \text{---} \text{---} \text{---} + \text{---} \text{---} \text{---} \text{---} \text{---} \text{---} \quad (127)$$


After inserting χ^{RPA} and χ_0 in eq. (113), one finds after performing the integration over λ , taking the time Fourier transform of the response function, and restricting the frequency integral to positive value using the response function symmetry [55, 161]:

$$E_c^{\text{RPA}} = \frac{1}{2\pi} \int_0^\infty d\omega \int d\mathbf{r} d\mathbf{r}' \left[\ln \left(1 - \chi_0(\mathbf{r}, \mathbf{r}', \omega) \frac{1}{|\mathbf{r}' - \mathbf{r}|} \right) + \chi_0(\mathbf{r}, \mathbf{r}', \omega) \frac{1}{|\mathbf{r}' - \mathbf{r}|} \right] \quad (128)$$

The exchange energy E_x in eq. (112) is similarly written using χ_0 . By explicitly expanding χ^{RPA} in eq. (121) or the logarithm in eq. (128), we obtain the correlation energy as a sum involving terms $\sim (V_{ee}\chi_0)^n$, $n \geq 2$. Those correspond to the so-called **ring diagrams**, the first of them ($n = 2$) being the MP2 direct diagram, and the next one was given in eq. (53), evaluated with the Kohn-Sham orbitals and eigenvalues [55, 56, 130, 161].

In the high density regime of the HEG the RPA correlation energy yields the exact correlation energy [56, 130] (eq. (115)). For inhomogeneous systems it is still a valuable approximation, that systematically accounts for correlation at all orders in perturbation, and stays size-consistent. The RPA is particularly suited to study systems where long range weak interaction, such as van der Waals, are important [161]. By taking the exact exchange and computing E_c^{RPA} as in eq. (128) we are considering the so-called **direct RPA** method. Beyond RPA methods can include higher order exchange terms [161]. The numerical cost of direct RPA algorithms typically range between $O(N^4)$ and $O(N^6)$ [161]. This method is also closely related to the so-called GW approximation to the self-energy, the later being approximated as the product of the interacting one-particle green function and the RPA screened interaction [64, 161], yielding accurate quasiparticle energies¹¹ for solids and thus furnish reliable band gaps values.

2.3.6 Methods Comparison continued

In Figure 9 is shown again the dissociation curve of the Hydrogen Bore molecule in 6-31G basis with various DFT approximations. The RPA curve was computed using PBE Kohn-Sham orbitals as input. Since DFT methods are not constructed on top of a Hartree-Fock reference, all curves has been vertically shifted so that they share the same minimum energy (as one can do: only energy differences are physically meaningful). The LDA present a better behavior than the Hartree-Fock curve but overestimate considerably the atomization energy. PBE is an improvement over LDA, but the atomization energy is again noticeably overestimated. The RPA behavior is much closer to the CCSD(T) one but the RPA equilibrium bond length is overestimated.

¹¹ which can be interpreted as electron addition and removal energies

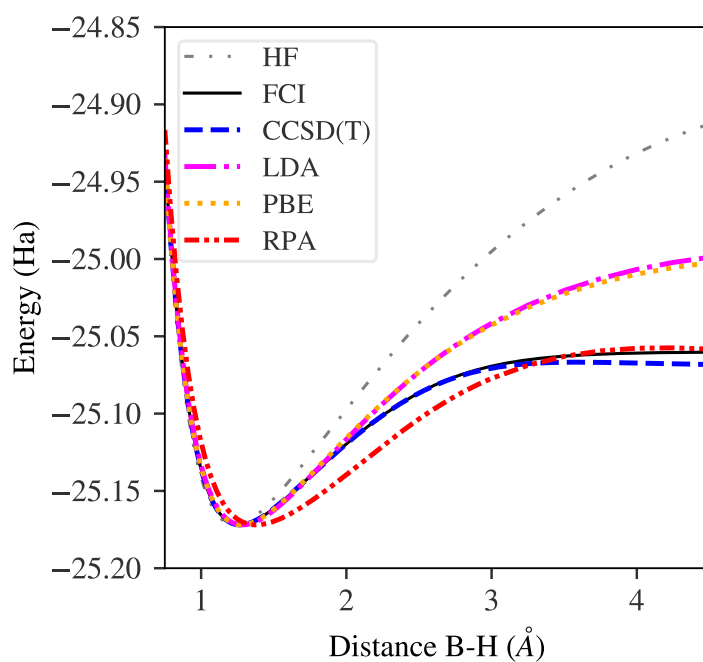


Figure 9: Dissociation curve of the BH molecule using various wave functions methods, the LDA and PBE DFT functionals, and the RPA method. Calculations were realized with Nwchem [5] in the 6-31G basis set. The RPA correlation energy was computed using PBE orbitals

2.4 FINITE TEMPERATURE PROPERTIES

We have seen various approximations to solve the electronic time independent Schrödinger equation. We would like now to be able to obtain finite temperatures properties: that is, we are interested in computing averages of some observable O in a given thermodynamical ensemble to which a density operator ρ is associated:

$$\langle \hat{O} \rangle = \frac{\text{tr} \hat{\rho} \hat{O}}{\text{tr} \hat{\rho}} \quad (129)$$

In the canonical ensemble the density matrix is given by

$$\hat{\rho} = e^{-\beta \hat{H}} \quad (130)$$

with the inverse temperature $\beta = (k_B T)^{-1}$, and the trace runs over all accessible configurations, that is, all positions of the electrons and nuclei. By assuming the equality of the time and ensemble averages (ergodic hypothesis [2]), we can obtain the former by studying the dynamics of the system of interest, whether it is a molecule or a crystal. In order to do so, we need to come back to the full Hamiltonian (2) $\hat{H} = \hat{H}_e + \hat{T}_n + \hat{V}_{nn}$. We will describe an approximation in which the electronic state is assumed to remain the lowest one (adiabatic approximation), while the nuclear and electronic degrees of freedom are separated (Born-Oppenheimer approximation). The nucleus dynamics is further solved by solving classical equations of motion using electronic forces.

2.4.1 Ab initio Molecular Dynamics

2.4.1.1 The adiabatic approximation

Say $\{\psi_i(\mathbf{r}|\mathbf{R})\}$ are the eigenstates of the electronic Hamiltonian \hat{H}_e with eigenvalues $\{\epsilon_i\}$, where \mathbf{r} (\mathbf{R}) is a shorthand for all the electrons (nuclei) positions. Let us call **Potential Energy Surfaces** (PES) the energy surfaces $\epsilon_i(\mathbf{R})$.

It is possible to decompose a wave function belonging to the full Hilbert space of \hat{H} as the following product of wave functions belonging to the Hilbert spaces of the electronic and nuclear Hamiltonian¹²:

$$\Psi(\mathbf{r}, \mathbf{R}) = \sum_{i=0}^{\infty} \psi_i(\mathbf{r}; \mathbf{R}) \chi_i(\mathbf{R}). \quad (131)$$

Thus, the $\{\chi_i(\mathbf{R})\}$ are wave functions of the nuclei positions. Applying \hat{H} and taking the product by $\langle \psi_j |$ yields :

¹² This factorization, known as Schmidt decomposition, is a direct application of the singular value decomposition [193], that will be described in Section 3.4

$$(\epsilon_j(\mathbf{R}) + \hat{V}_{nn} + \hat{T}_n - E) \chi_j(\mathbf{R}) = \sum_{i=0}^{\infty} \Lambda_{ji} \chi_i(\mathbf{R}) \quad (132)$$

with

$$\Lambda_{ji} = \sum_{\alpha} \frac{-\hbar^2}{2M_{\alpha}} (\langle \psi_j | \nabla_{\mathbf{R}_{\alpha}}^2 | \psi_i \rangle + 2 \langle \psi_j | \nabla_{\mathbf{R}_{\alpha}} | \psi_i \rangle \nabla_{\mathbf{R}_{\alpha}}) \quad (133)$$

The Λ term contains all couplings between the nuclear and electronic degrees of freedom. By setting all the off diagonal elements of Λ to zero, we put ourselves in the **adiabatic approximation**. Indeed, the electronic Hamiltonian contains all its time dependence inside the nuclei position. Expressing the time derivative as $\partial_t = \sum_{\alpha} \nabla_{\mathbf{R}_{\alpha}} \dot{\mathbf{R}}_{\alpha}$, and taking the dot product by $\langle \psi_j |$ we get for our general electronic wave function $\psi(\mathbf{r}; \mathbf{R}) = \sum_i a_i \psi_i(\mathbf{r}; \mathbf{R})$ that

$$\dot{a}_j(t) = -\frac{i}{\hbar} \epsilon_j(t) a_j(t) - \sum_i \sum_{\alpha} \langle \psi_j | \nabla_{\mathbf{R}_{\alpha}} | \psi_i \rangle \dot{\mathbf{R}}_{\alpha} a_i(t) \quad (134)$$

By neglecting the coupling terms between different instantaneous electronic eigenstates, if the system is at some time t_0 on a given PES $\epsilon_k(\mathbf{R})$, it will remain in the corresponding instantaneous eigenstate $\psi_k(\mathbf{r}; \mathbf{R})$ for all future times $t > t_0$. The adiabatic theorem states that this is accurate as long as the PES are well separated and the electronic Hamiltonian is not varying too quickly [170] :

$$\frac{\langle \psi_j, t | \dot{H}_e | \psi_i, t \rangle}{\epsilon_j(t) - \epsilon_i(t)} \ll \frac{\epsilon_j(t)}{\hbar} \quad (135)$$

2.4.1.2 The Born-Oppenheimer approximation

Moreover, by also equating to zero the diagonal terms in eq. (133), we are now performing the **Born-Oppenheimer (BO) approximation**. In that case we have to solve

$$(\epsilon_j(\mathbf{R}) + \hat{V}_{nn} + \hat{T}_n) \chi_j(\mathbf{R}) = E \chi_j(\mathbf{R}) \quad (136)$$

in which the PES $\epsilon_j(\mathbf{R})$ is a local external potential felt by the nuclei (hence the name *potential energy surface*). The nuclei and electronic degree of freedom are now totally separated, resulting in a huge simplification of the problem to solve.

This is a good approximation near equilibrium points of the nuclei (where $\nabla_{\mathbf{R}}$ is close to zero). This however breaks down in the vicinity of **avoided crossing** points, where the instantaneous electronic eigenstates abruptly change from one diabatic state to the other, or near **conical intersections**, where the denominator of eq. (135) diverge due to the degeneracy. It should moreover not be trusted in metals, for which groundstate and first excited states are not well separated [150].

2.4.1.3 Classical motion of the nuclei

A last approximation which is made in what is generally called *ab initio* Molecular Dynamics (AIMD) is that the nuclei are moved by solving **classical** equations of motion. That is, once the electronic ground state density has been found for a given set of nuclei positions, one compute the forces acting on the nuclei by using the Hellmann-Feynman theorem while varying the nuclei positions, and update the later. This classical approximation neglects the so-called **nuclear quantum effects**. Those can be important for light elements, or in order to account for nuclei delocalization. The later for example is responsible for a bandgap reduction (zero point renormalization) of ~ 0.4 eV in the diamond at zero temperature [104]. Finally, in order to account for the temperature, one must choose a statistical ensemble within which to work and fix the temperature. For example to maintain temperature in the canonical (N,V,T) ensemble, one can choose the *Andersen thermostat* [4] procedure, in which new nuclei velocities will be drawn from a Maxwell-Boltzmann statistics at random times. To summarize, what we will call AIMD in the next parts of this thesis consist in the Born-Oppenheimer separation of nuclear and electronic degrees of freedom, the later being solved using DFT, before classical equations of motions are applied to the nuclei using forces derived from the DFT.

2.4.2 Monte Carlo sampling

Instead of computing time averages by performing Molecular Dynamics, one can directly compute ensemble averages by sampling configurations in order to reproduce the target probability density given by eq. (130). The Metropolis algorithm is a popular and simple method to achieve this: from any starting configuration \mathbf{R} with energy ϵ , a new configuration \mathbf{R}' with energy ϵ' is proposed with proposal probability density $g(\mathbf{R}'|\mathbf{R})$ and is accepted with probability $\min(1, e^{-\beta(\epsilon' - \epsilon)})$. The sequence of configurations sampled in this way is ensured to converge to the canonical Boltzmann probability [125, 135]. The choice of the proposal distribution is arbitrary and should be tuned in order to obtain a reasonable acceptance ratio. For example a Maxwell-Boltzmann distribution can be used to obtain random displacements of the atoms going from \mathbf{R} to \mathbf{R}' , as done in Section 4.3.2. Since it only demands energy calculations, this methodology can be a cheaper alternative to methods where the forces are computationally demanding.

2.4.3 Thermodynamic perturbation theory

Consider two electronic structure methods \mathcal{X} and \mathcal{Y} with corresponding Hamiltonians $H^{\mathcal{X}}$ and $H^{\mathcal{Y}} = H^{\mathcal{X}} - \Delta V$. Let $\langle O \rangle_{\mathcal{X}}$ and $\langle O \rangle_{\mathcal{Y}}$ be canonical ensemble averages

of the observable O evaluated with \mathcal{X} and \mathcal{Y} Hamiltonians. Thermodynamic perturbation theory relates both averages [31, 152]:

$$\langle O \rangle_{\mathcal{Y}} = \frac{\int O e^{-\beta H^{\mathcal{Y}}} d\Gamma_{\mathcal{Y}}}{\int e^{-\beta H^{\mathcal{Y}}} d\Gamma_{\mathcal{Y}}} = \frac{\int O e^{\beta \Delta V} e^{-\beta H^{\mathcal{X}}} d\Gamma_{\mathcal{X}}}{\int e^{\beta \Delta V} e^{-\beta H^{\mathcal{X}}} d\Gamma_{\mathcal{X}}} \quad (137)$$

so that

$$\langle O \rangle_{\mathcal{Y}} = \frac{\langle O e^{\beta \Delta V} \rangle_{\mathcal{X}}}{\langle e^{\beta \Delta V} \rangle_{\mathcal{X}}}. \quad (138)$$

Thus the average in the \mathcal{Y} ensemble can be expressed as a reweighting of the average in the \mathcal{X} ensemble. The corresponding weights are the Boltzmann factors of the energy difference associated with the configurations upon which the average is performed.

If one has performed an AIMD simulation using some DFT functional (\mathcal{X} method) to obtain some average property O , it is then possible to obtain the corresponding ensemble property within another method \mathcal{Y} , by computing the energies using said \mathcal{Y} method for each configuration of the AIMD.

Note that we used the fact that the configurational spaces are identical when going from one integral to the other. In practice the integrals are replaced by summation over finite sets of configurations around equilibrium configurations. In that case, for formula (138) to be a valid, the probability densities of both methods must sufficiently overlap on the space sampled using the \mathcal{X} method (see Figure 10).

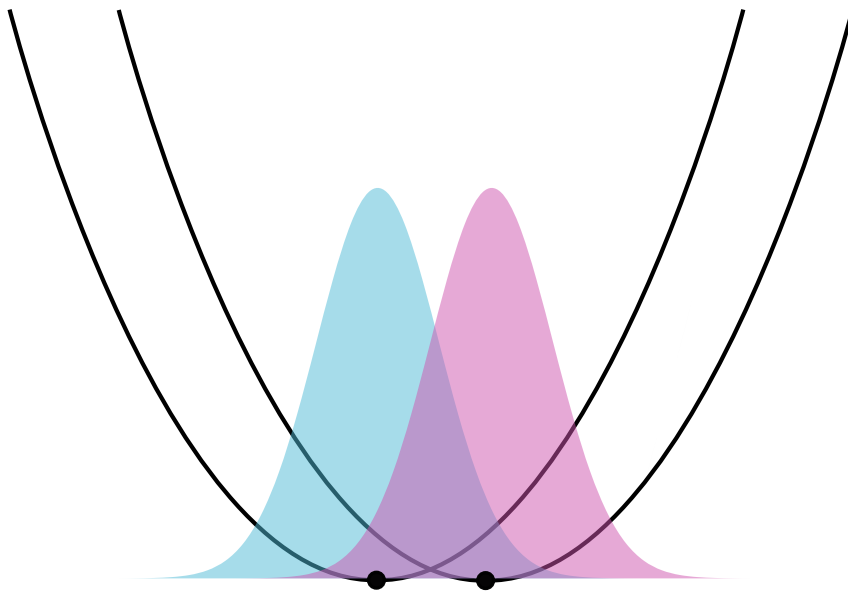


Figure 10: Pictorial representation of thermodynamic perturbation theory: the black curves are the potential energies of two harmonic Hamiltonian with different equilibrium positions (black dots). The bell curves are the associated density probabilities distributions along the configurational space.

MACHINE LEARNING METHODS

3.1	Introduction	46
3.2	Physical descriptors	47
3.3	Kernel Ridge Regression	48
3.4	Dimensionality reduction	51
3.5	Generative models: Restricted Boltzmann Machine	52

3.1 INTRODUCTION

In machine learning (ML), one starts with a set of observations - the training dataset - and a model. The model is a function or a set of functions of the training points and of some parameters. One then use an algorithm (machine) to adjust the parameters of the model (learning) in order to optimize some objective. A lot of different tasks can be tackled by ML, such as regression, classification, clustering, sampling, pattern recognition, etc. ML is generally divided into three classes: supervised learning, unsupervised learning, and reinforcement learning. In **supervised learning**, the training dataset is composed of the input data points and the targets. The inputs characterize the system studied (for example the geometry and chemical species of a molecule) while the target is the observable one would like to predict, given new unseen input. It can be a continuous variable (regression) such as the atomization energy, or a discrete variable (classification) such as the stability of the molecule. In **unsupervised learning**, there are only input, without target. The objective is to uncover patterns in the training dataset in order to generalize. For example clustering consists in dividing the data into different clusters (the unsupervised pendant of classification), dimensionality reduction aims at extracting the most important features from a set of datapoints, while generative models learn the probability distribution of the dataset, from which one can sample new input points. Finally in reinforcement learning, the ML model learn to interact with an environment in order to maximize its reward.

In the next sections we will first focus on the regression task of fitting a set of atomic configurations to their total energy. To this purpose we will discuss physical descriptors, which are optimized representations of molecular and periodic physical systems in order to produce useful inputs. Then we will describe how to go from linear regression to kernel ridge regression. Dimensionality reduction will

be discussed afterward, with the principal component analysis and t-stochastic neighbor embedding algorithms. Finally to prepare the last chapter of this thesis we will introduce generative models and a description of the restricted Boltzmann machine model.

3.2 PHYSICAL DESCRIPTORS

Suppose we would like to find a ML model that learn the relationship between a set of atomic configurations (input) and their total energy (target), thus avoiding explicitly solving the Schrödinger equation. The later being invariant under global translations and rotations of the chemical configuration, it is desirable that the input representation, or **descriptor**, also possesses these symmetries, to circumvent the need for the model to learn the later. This disqualifies the traditional Cartesian coordinates or Z-matrix. Additionally, a good representation should also be invariant to permutations of chemical elements. The last condition is less stringent as one can impose a unique way of labelling the different atoms or chemical species. Different solutions have been proposed [13, 77, 93]. For example the Coulomb matrix is defined by [169]

$$M_{ij} = \begin{cases} \frac{Z_i^{2.4}}{2} & \text{if } i = j \\ \frac{Z_i Z_j}{R_{ij}} & \text{if } i \neq j \end{cases} \quad (139)$$

where Z_i are the nuclear charges and R_{ij} the distances between atoms. A more involved description consist of considering local environments: a cutoff radius r_c is defined, and for each atom, a description of each atoms inside the sphere of radius r_c centered at the given atom will be given. This gives much more flexibility to the representation and allow the final model to better generalize to unseen configurations. Considering the size extensivity of the energy, this amounts to decompose the target property into atomic contributions. The Smooth Overlap of Atomic Positions [43] is one of such descriptors, that will be used in the next two chapters. Let us consider a chemical structure χ made of N_χ atoms. First, gaussian atomic densities

$$\rho(\mathbf{r}, \chi_a) = \sum_{i \in \chi_a} \exp(-|\mathbf{x}_i - \mathbf{r}|/2\sigma^2) \quad (140)$$

are constructed, centered at the atom considered, where the sum runs over atoms within r_c around the atom a . Those densities are then decomposed into spherical harmonics, before to take their rotationally invariant power spectrum $\mathbf{p}(\chi_a)$. The later vector is the SOAP descriptor associated to the atomic environment χ_a . Once a descriptor has been chosen, the ML model should be able to assess the degree of similarity between different configuration in order to predict the energy of an

unseen configuration. This can be achieved in kernel methods through the use of kernel functions. The later take as input two vectors and return a number, which is in correspondence to some inner product in another space, and will be defined in the next section. In case of local descriptors such as SOAP, the global similarity between two structures is decomposed in local similarities between atomic environments. The local SOAP kernel $k(\chi_a, \chi'_b)$ which measures the similarity between atomic region a of structure χ and atomic region b of structure χ' is defined as

$$k(\chi_a, \chi'_b) = \mathbf{p}(\chi_a) \cdot \mathbf{p}(\chi'_b) \quad (141)$$

This is in fact equivalent [43] to the overlap integration of the two atomic environment densities upon rotation:

$$k(\chi_a, \chi'_b) = \int d\hat{R} \left| \int d\mathbf{r} \rho(\hat{R}\mathbf{r}, \chi_a) \rho(\mathbf{r}, \chi'_b) \right|^2 \quad (142)$$

where \hat{R} represent the rotation operator. See also Figure 11.

Finally, in order to globally compare two structures, different approaches exist [43], one of such is to define the following global kernel:

$$\mathcal{K}(\chi, \chi') = \sum_{a=1}^{N_x} \sum_{b=1}^{N_{x'}} \frac{k(\chi_a, \chi'_b)}{\sqrt{k(\chi_a, \chi_a) k(\chi'_b, \chi'_b)}}. \quad (143)$$

This kernel defines the similarity between the structures χ and χ' by summing the local similarities between each atomic environment of the two. This is divided by a normalization factor that ensure that the local similarity is at most one for identical atomic environments.

3.3 KERNEL RIDGE REGRESSION

In regression, the training dataset is composed of a set of inputs $\mathbf{x}_i \in \mathbb{R}^d$ and targets $y_i \in \mathbb{R}$. The objective is to optimize a model to reproduce the target given the input, such that the model is able to generalize to unseen input $\tilde{\mathbf{x}}$ and predict the target \tilde{y} for those.

In linear regression (LR), we wish to find the hyperplane defined by $\boldsymbol{\beta} \cdot \mathbf{x} = y$ - where we assumed that our plane goes through zero¹ - from which the sum of the distances of the training points is minimized. This corresponds to the model

$$f(\tilde{\mathbf{x}}) = \boldsymbol{\beta} \cdot \tilde{\mathbf{x}}, \quad (144)$$

We find the parameters β_i of the model by minimizing the squared error on the training set:

$$L = \sum_{i=1}^n (f(\mathbf{x}_i) - y_i)^2 = \|\mathbf{X}\boldsymbol{\beta} - \mathbf{y}\|^2 \quad (145)$$

¹ we can always center our dataset, by subtracting the mean of the x_i and y_i

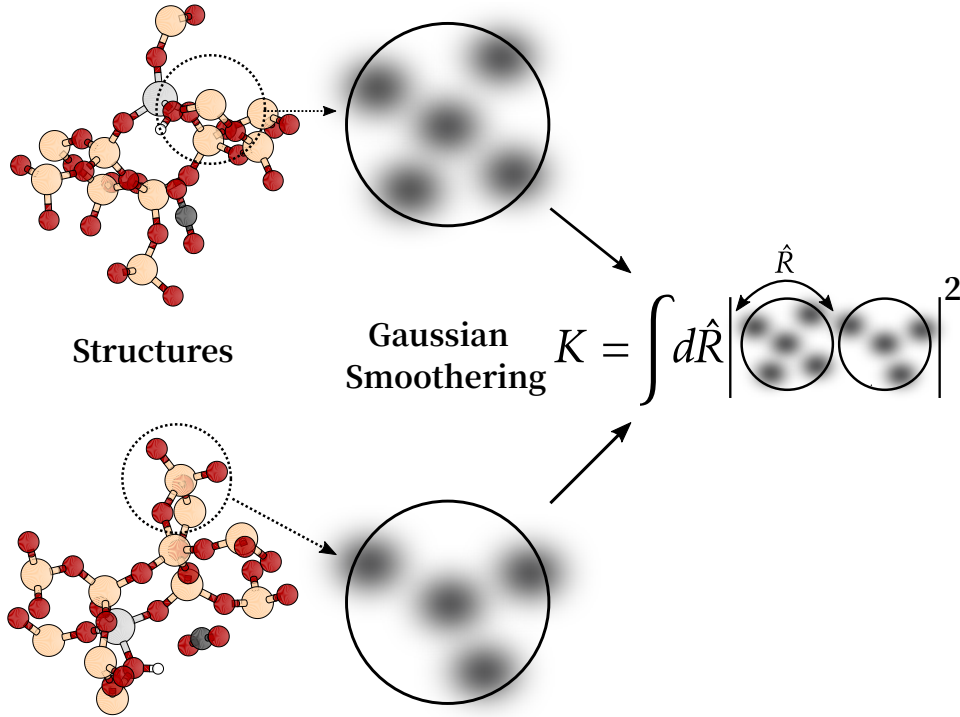


Figure 11: Smooth Overlap of Atomic Positions descriptor: two local atomic environments are described as gaussian centered densities inside some cutoff. The similarity between both environments is assessed by the amplitude of the overlap of those densities, upon rotation to preserve rotational invariance of the similarity measure.

where \mathbf{X} is a matrix whose i^{th} row is the training input \mathbf{x}_i . By cancelling $\partial L / \partial \boldsymbol{\beta}$, we find² the notorious normal equations

$$\boldsymbol{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} \quad (146)$$

We can always write β_i as a linear function of the x_i , such that $\boldsymbol{\beta} = \mathbf{X}\boldsymbol{\alpha}$, and our linear model is rewritten as

$$\begin{aligned} f(\tilde{\mathbf{x}}) &= \boldsymbol{\beta} \cdot \tilde{\mathbf{x}} = \sum_{i=1}^d \beta_i \tilde{x}_i \\ &= \sum_{i=1}^d \left(\sum_{j=1}^n \alpha_j x_j \right)_i \tilde{x}_i \\ &= \sum_{j=1}^n \alpha_j \mathbf{x}_j \cdot \tilde{\mathbf{x}} \end{aligned}$$

Solving for $\boldsymbol{\alpha}$ gives

$$\boldsymbol{\alpha} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{y}. \quad (147)$$

² using $\frac{\partial \mathbf{u}^T \mathbf{v}}{\partial \mathbf{x}} = \mathbf{u}^T \frac{\partial \mathbf{v}}{\partial \mathbf{x}} + \mathbf{v}^T \frac{\partial \mathbf{u}}{\partial \mathbf{x}}$

That is, linear regression only depends of inner products (hence it only depends on distances and angles between input points), and its solution can be expressed through the Gram matrix $\mathbf{X}^T\mathbf{X}$, or equivalently the covariance matrix, which measure the similarity between inputs.

Say now that we are interested in a phenomenon where the relation between input points is non-linear, and we want to perform a polynomial fit instead of a simple linear fit, with some polynomial **feature map** $\phi(\mathbf{x}) \in \mathbb{R}^D$, $D \geq d$. We can still use our linear regression model to fit an hyperplane defined in the **feature space** \mathbb{R}^{D^3}

$$f(\tilde{\mathbf{x}}) = \boldsymbol{\beta} \cdot \phi(\tilde{\mathbf{x}}) = \sum_{j=1}^n \alpha_j \phi(\mathbf{x}_j) \cdot \phi(\tilde{\mathbf{x}}) = \sum_{j=1}^n \alpha_j k(\mathbf{x}_j, \tilde{\mathbf{x}}). \quad (148)$$

In the last line we have defined the **kernel function** k . A kernel function should be positive definite, symmetric, and continuous [168].

The simplest kernel is the linear kernel $k(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i \cdot \mathbf{x}_j$, which corresponds to linear regression.

Let us now consider the feature map $\phi(x) = e^{-\frac{x^2}{2\sigma^2}} (1, \frac{x}{\sigma}, \sqrt{\frac{1}{2!} \frac{x^2}{\sigma^2}}, \sqrt{\frac{1}{3!} \frac{x^3}{\sigma^3}}, \dots)$. The inner product $\phi(x) \cdot \phi(y)$ is equal to the gaussian kernel $k(x, y) = e^{-\frac{(x-y)^2}{2\sigma^2}}$. We can perform a linear regression in the infinite dimensional space defined by ϕ (or equivalently, an infinite dimension polynomial fit) by computing directly the inner products through k . This in essence in the **kernel trick** [160].

The coefficients of $\boldsymbol{\alpha}$, solution of the **kernel regression**, are now given by

$$\boldsymbol{\alpha} = \mathbf{K}^{-1} \mathbf{y} \quad (149)$$

where $K_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$. The parameter σ in the gaussian kernel is called an hyperparameter and should be optimized separately. Other popular kernels are the laplacian kernel or the polynomial kernel [168].

If the fit is “too good” on the training points, the model can fail to generalize to unseen data, a phenomenon know as **overfitting**. This can be checked by splitting the data into a training set and a test set. The model is trained on the training set while the error estimate is computed on the test set. The test error should not be much higher than the training error. To prevent overfitting, one can add a $\lambda \|\mathbf{f}\|^2 = \lambda \boldsymbol{\alpha}^T \mathbf{K} \boldsymbol{\alpha}$ penalty term to the squared error, λ being a hyperparameter to optimize. This is known as Kernel Ridge Regression (KRR), with solution

$$\boldsymbol{\alpha} = (\mathbf{K} + \lambda \mathbf{I})^{-1} \mathbf{y} \quad (150)$$

KRR has been used to learn the atomization energy of molecules as well as other properties (HOMO/LUMO levels, polarizability, etc.) inside a dataset. In those works, chemical accuracy was achievable with training sets whose size were of the order of 10^3 to 10^4 [8, 77, 93].

3 Consider a model where $g(x) = \alpha x^2$, considering the map $\phi(x) = x^2$, we have a linear model in the space defined by ϕ , $g(x) = \alpha \cdot \phi(x)$.

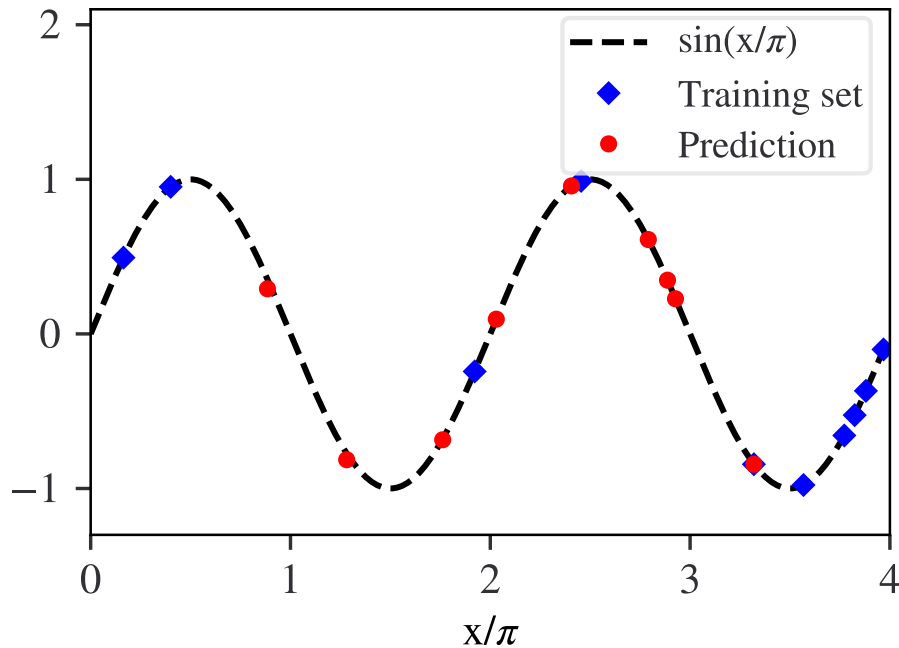


Figure 12: Kernel Ridge Regression of the sinus function, using a gaussian kernel.

3.4 DIMENSIONALITY REDUCTION

Dimensionality reduction is a method to reduce the number of variables in a dataset (thus lowering the dimension d of the input space) while retaining as much of the original information as possible. This can be helpful for a variety of reasons, such as improving computation efficiency, reducing noise, clustering and simplifying the visualization of data, the two later being used in the next two chapters.

Principal component analysis (PCA) is one of the most commonly used techniques in machine learning. PCA works by identifying the directions of maximum variance in the data and projecting the data onto a lower-dimensional subspace that captures the most important features. In practice, PCA is a low-rank approximation of our input matrix \mathbf{X} through singular value decomposition (SVD): any $n \times d$ real matrix \mathbf{X} can be decomposed as

$$\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T \quad (151)$$

where $\mathbf{\Sigma}$ is a $n \times d$ diagonal matrix whose diagonal entries are named singular values, and \mathbf{U}, \mathbf{V} are $n \times n$ and $d \times d$ orthogonal matrices. Geometrically we can see that most of the information on the application \mathbf{X} is contained in those singular values that have the biggest absolute value. Similarly, SVD amounts to diagonalize

the covariance matrix $\mathbf{C} = \mathbf{X}\mathbf{X}^T$ or the Gram matrix $\tilde{\mathbf{C}} = \mathbf{X}^T\mathbf{X}$.⁴ We wish to approximate the $n \times d$ matrix \mathbf{X} with a $k \times d$ matrix \mathbf{X}_k , k being lower than d .⁵ One possibility is to minimize the Frobenius norm of the distance between \mathbf{X} and \mathbf{X}_k : $\|\mathbf{X} - \mathbf{X}_k\|_F = \sqrt{\text{tr}[(\mathbf{X} - \mathbf{X}_k)^T(\mathbf{X} - \mathbf{X}_k)]}$. Now the Young-Eckart Theorem states [54] that this is achieved by $\mathbf{X}_k = \sum_{i=1}^k \sigma_i \mathbf{u}_i \mathbf{v}_i^T$, where σ_i , \mathbf{u}_i and \mathbf{v}_i are the i^{th} singular value, in decreasing order of magnitude, and the i^{th} columns of \mathbf{U} and \mathbf{V} .

Another route for visualizing high-dimensional data consist in mapping the data onto a low-dimensional space while preserving the pairwise distances between the data points. For example one may want to find a linear projection: $\mathbf{x}' = \frac{1}{\sqrt{d}}\mathbf{A}\mathbf{x}$ where \mathbf{A} is a $k \times d$ matrix such that $\|\mathbf{x}'_i - \mathbf{x}'_j\|^2 \approx \|\mathbf{x}_i - \mathbf{x}_j\|^2$. This can be easily achieved with $\mathbf{A} \sim \mathcal{N}(0, 1)$ according to Johnson-Lindenstrauss lemma [41]. Another popular technique is t-distributed stochastic neighbor embedding (t-SNE). t-SNE uses a gaussian kernel instead of the cartesian norm to measure the distance between two points (hence is able to capture nonlinear relationships). It defines a similarity measure by constructing a gaussian conditional probability of the distance between two points, both in the original d -dimensional space (p_{ij}), and in the reduced k -dimensional space (q_{ij}). The mapping is found by minimizing the sum of the Kullback-Leibler divergence between both probabilities:

$$\mathcal{C} = \sum_i \sum_j p_{ij} \log\left(\frac{p_{ij}}{q_{ij}}\right) = \sum_i \text{KL}(P_i \| Q_i) \quad (152)$$

The Kullback-Leibler divergence $\text{KL}(P \| Q) = \sum_i P(i) \log\left(\frac{P_i}{Q_i}\right)$, or relative entropy, between two discrete probability distributions P and Q is commonly used as a measure of the similarity between both [125]. Note that the distance itself between points which appears in the Gaussian probabilities is not necessarily Cartesian, and one can use the SOAP kernel to define a distance between two chemical structures χ_a and χ_b :

$$D(\chi_a, \chi_b) = (1 - \mathcal{K}(\chi_a, \chi_b))^{1/2}. \quad (153)$$

3.5 GENERATIVE MODELS: RESTRICTED BOLTZMANN MACHINE

Generative models are a class of unsupervised ML models that can learn to generate new data that is similar to the training data. Among those, energy based models define an *energy*, function of the model parameters and input points, and a probability distribution over the inputs and parameters. In correspondence with statistical physics, the most probable inputs across the training dataset are then to be associated with the lowest energy and reciprocally during the learning stage. The Boltzmann Machine model is a complete weighted graph with the inputs \mathbf{v}_i

⁴ $\mathbf{C} = \mathbf{U}\Sigma\mathbf{V}^T(\mathbf{U}\Sigma\mathbf{V}^T)^T = \mathbf{U}\Sigma\mathbf{V}^T\mathbf{V}\Sigma^T\mathbf{U}^T = \mathbf{U}\Sigma\Sigma^T\mathbf{U}^T = \mathbf{U}\mathbf{D}\mathbf{U}^T$.

⁵ Actually \mathbf{X}_k has the same dimensions of \mathbf{X} but is of *rank* k , we can still pad it with zeros.

living on the vertices, and the parameters J_{ij} are the link weights. The weight matrix is symmetric, and there is no self interaction ($J_{ii} = 0 \forall i$). An energy function $E(\mathbf{v}, \mathbf{J}) = -\mathbf{v}^T \mathbf{J} \mathbf{v}$ is defined and associated to a Boltzmann probability $P(\mathbf{v} | \mathbf{J}) = \exp(-E(\mathbf{v}, \mathbf{J})) / \sum_{\{\mathbf{v}\}} \exp(-E(\mathbf{v}, \mathbf{J}))$. For binary input values, this is obviously a classical spin system with arbitrary long-range interactions. One can add a bias vector into the energy (an external field with contribution $-\mathbf{a}^T \mathbf{v}$). This model however, cannot account for correlations higher than of order two as seen from its energy function. A solution is to introduce a **hidden layer** of vertices h_i , that will model higher order statistics between visible input through their interaction with the hidden layer [125]. For the simple Boltzmann machine, the idea is to perform a Hubbard-Stratonovich transformation of the model to mediate the input interactions through a new degree of freedom \mathbf{h} [132].⁶

The restricted Boltzmann machine (RBM), pictorially represented in Figure 4, is a neural network model, consisting of one input layer of D visible binary units $v_i \in \{0, 1\}$, one layer of P binary hidden units $h_j \in \{0, 1\}$, and $D \times P$ weights W_{ij} between both layers. Two bias vectors of components a_i and b_j are added to the visible and hidden layer, respectively.

By introducing an energy function of $\Lambda = \{\mathbf{a}, \mathbf{b}, \mathbf{W}\}$ for a given configuration $\{\mathbf{v}, \mathbf{h}\}$ as

$$E(\mathbf{v}, \mathbf{h}, \Lambda) = -(\mathbf{a}^T \mathbf{v} + \mathbf{b}^T \mathbf{h} + \mathbf{v}^T \mathbf{W} \mathbf{h}), \quad (154)$$

and an inverse temperature $\beta = 1/T$, one can define the probability distribution associated with the RBM over the input configurations \mathbf{v} to be

$$P(\mathbf{v}) = \frac{\sum_{\{\mathbf{h}\}} e^{-\beta E(\mathbf{v}, \mathbf{h})}}{\sum_{\{\mathbf{h}, \mathbf{v}\}} e^{-\beta E(\mathbf{v}, \mathbf{h})}} = \frac{\sum_{\{\mathbf{h}\}} e^{-\beta E(\mathbf{v}, \mathbf{h})}}{Z}, \quad (155)$$

where the partition function Z was defined. After training on a set of vectors \mathbf{v} distributed according to some (usually unknown) probability distribution, the RBM can be used to generate new vectors \mathbf{v} according to the $P(\mathbf{v})$ distribution learnt by the model. To this purpose the Gibbs sampling can be used [125]: Starting from an initial random trial vector as input, one can obtain a hidden layer vector \mathbf{h} using the conditional probability $p(h_i | v_i)$ and then obtain a new input-layer vector \mathbf{v} using the conditional probability $p(v_i | h_i)$; the repetition of this operation a certain number of times forms a Markov chain that generates a vector \mathbf{v} according to the probability $P(\mathbf{v})$.

The Gibbs sampling algorithm is important for both the training and generative step of restricted Boltzmann machines (RBMs) [125]. It is based on the following steps:

⁶ The Hubbard-Stratonovich transformation use the simple Gaussian integral identity $\left(\frac{K}{2\pi}\right)^{\frac{1}{2}} \int dh \exp\left(-\frac{1}{2}Kh^2 + Ks^2h\right) = \exp\left(\frac{1}{2}Ks^2\right)$. If J_{ij} is positive semi-definite then $\exists \mathbf{W} \mid \mathbf{J} = \mathbf{W}^T \mathbf{W}$ so that $E(\mathbf{v}) = -\mathbf{v}^T \mathbf{W}^T \mathbf{W} \mathbf{v}$ and by taking $K = 1$ and $s = (\mathbf{W} \mathbf{v})^T$ one obtains $p(\mathbf{v}) = Z^{-1} \exp(-E(\mathbf{v})) = Z^{-1} \exp\left(\frac{1}{2}(\mathbf{W} \mathbf{v})^T (\mathbf{W} \mathbf{v})\right) = Z^{-1} \int dh \exp\left(-\frac{1}{2}h^T h + (\mathbf{W} \mathbf{v})^T h\right) = Z^{-1} \int dh \exp(-E(\mathbf{v}, \mathbf{h}))$, see [132].

1. The components of the initial input vector $\mathbf{v}^{(0)}$ are drawn from a uniform probability distribution $\mathcal{U}(0, 1)$. Namely, each $v_i^{(0)}$ with $i = 1, \dots, D$ contains a random number in the interval $[0, 1[$.
2. The values $h_j^{(0)}$ with $j = 1, \dots, P$ are initialized to 0. For each $h_j^{(0)}$ a number d_j is drawn from $\mathcal{U}(0, 1)$ and the value of $h_j^{(0)}$ is then updated to 1 if $d_j < p(h_j = 1 | \mathbf{v}^{(0)})$.
3. In order to update the vector \mathbf{v} , the new components $v_i^{(1)}$ are first set to 0. For each $v_i^{(1)}$ a number d_i is drawn from $\mathcal{U}(0, 1)$ and the value of $v_i^{(1)}$ is updated to 1 if $d_i < p(v_i = 1 | \mathbf{h}^{(0)})$.
4. The steps 2 and 3 are repeated L times.

Restricted Boltzmann machines are usually trained using the contrastive divergence learning procedure [87]. The contrastive divergence approach maximizes the log-likelihood (divided by M) of this dataset, which is given by

$$\mathcal{L} = \frac{1}{M} \sum_{i=1}^M \log P(\mathbf{v}^{(i)}). \quad (156)$$

By using eqs. (154-155) this becomes explicitly

$$\begin{aligned} \mathcal{L} &= \frac{1}{M} \sum_{i=1}^M \log \sum_{\{\mathbf{h}\}} \exp(\mathbf{a}^T \mathbf{v}^{(i)} + \mathbf{b}^T \mathbf{h} + \mathbf{v}^{(i)T} \mathbf{W} \mathbf{h}) - \log Z \\ &= \frac{1}{M} \sum_{i=1}^M \mathbf{a}^T \mathbf{v}^{(i)} + \frac{1}{M} \sum_{i=1}^M \log \prod_j \sum_{h_j \in \{0,1\}} \exp\left(h_j \left(\sum_k W_{kj} v_k^{(i)} + b_j\right)\right) - \log Z \\ &= \frac{1}{M} \sum_{i=1}^M \mathbf{a}^T \mathbf{v}^{(i)} + \frac{1}{M} \sum_{i=1}^M \sum_j \log \left(1 + \exp\left(\sum_k W_{kj} v_k^{(i)} + b_j\right)\right) - \log Z. \end{aligned} \quad (157)$$

The derivatives of \mathcal{L} with respect to the parameters \mathbf{a}_k , \mathbf{b}_k , W_{jk} are

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \mathbf{a}_k} &= \frac{1}{M} \sum_{i=1}^M v_k^{(i)} - \frac{1}{Z} \sum_{\{\mathbf{h}, \mathbf{v}\}} v_k e^{-E(\mathbf{v}, \mathbf{h})} \\ &= \langle v_k \rangle_{\text{data}} - \langle v_k \rangle_{\text{model}}, \end{aligned} \quad (158)$$

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \mathbf{b}_k} &= \frac{1}{M} \sum_{i=1}^M \frac{\exp\left(\sum_j W_{jk} v_j^{(i)} + h_k\right)}{1 + \exp\left(\sum_j W_{jk} v_j^{(i)} + h_k\right)} - \langle h_k \rangle_{\text{model}} \\ &= \frac{1}{M} \sum_{i=1}^M p(h_k = 1 | \mathbf{v}^{(i)}) - \langle h_k \rangle_{\text{model}} \\ &= \langle h_k \rangle_{\text{data}} - \langle h_k \rangle_{\text{model}} \end{aligned} \quad (159)$$

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial W_{jk}} &= \frac{1}{M} \sum_{i=1}^M v_j^{(i)} p(h_k = 1 | \mathbf{v}^{(i)}) - \langle v_j h_k \rangle_{\text{model}} \\ &= \langle v_j h_k \rangle_{\text{data}} - \langle v_j h_k \rangle_{\text{model}}. \end{aligned} \tag{160}$$

Given these equations the learning algorithm is based on the following iterative steps: (1) A minibatch of N_b training elements $\mathbf{v}^{(i)}$ is obtained from the dataset; (2) The conditional probabilities $p(h_k = 1 | \mathbf{v}^{(i)}) = \frac{1}{1 + \exp(-\beta(\mathbf{v}^{(i)\top} \mathbf{W})_k - b_k)}$ are computed and used to evaluate the averages $\langle \dots \rangle_{\text{data}}$; (3) N_b pairs $\{\mathbf{v}^{(L)}, \mathbf{h}^{(L)}\}$ are generated by a L step Gibbs sampling and used to evaluate the averages $\langle \dots \rangle_{\text{model}}$; (4) the parameters of the RBM model are updated with a learning rate ϵ :

$$\mathbf{a}_k = \mathbf{a}_k + \epsilon (\langle v_k \rangle_{\text{data}} - \langle v_k \rangle_{\text{model}}) \tag{161}$$

$$\mathbf{b}_k = \mathbf{b}_k + \epsilon (\langle h_k \rangle_{\text{data}} - \langle h_k \rangle_{\text{model}}) \tag{162}$$

$$W_{jk} = W_{jk} + \epsilon (\langle v_j h_k \rangle_{\text{data}} - \langle v_j h_k \rangle_{\text{model}}). \tag{163}$$

With the simpler Boltzmann machine, the weights would have been updated by $\langle v_j v_k \rangle_{\text{data}} - \langle v_j v_k \rangle_{\text{model}}$, thus modelling only the second order statistics of the inputs. To see how the RBM is instead able to go beyond second order, we can use the marginal probability $P(\mathbf{v})$ to write

$$P(\mathbf{v}) = \frac{1}{Z'} e^{-E(\mathbf{v})} = \frac{1}{Z} \sum_{\{\mathbf{h}\}} e^{-E(\mathbf{v}, \mathbf{h})} = \frac{e^{\mathbf{a}^\top \mathbf{v}}}{Z} \sum_{\{\mathbf{h}\}} e^{\mathbf{b}^\top \mathbf{h} + \mathbf{v}^\top \mathbf{W} \mathbf{h}} \tag{164}$$

so that

$$E(\mathbf{v}) = -\mathbf{a}^\top \mathbf{v} - \log \sum_{\mathbf{h}} e^{\mathbf{b}^\top \mathbf{h} + \mathbf{v}^\top \mathbf{W} \mathbf{h}} \tag{165}$$

Let us now consider the probability distribution $q_a(h_a) = Z^{-1} e^{b_a h_a}$ and its cumulant expansion $K_a(t) = \log \mathbb{E}[e^{t h_a}] = \log \sum q_a(h_a) e^{t h_a} = \sum_n \kappa_a^{(n)} \frac{t^n}{n!}$ where the n -th cumulant is given by $\kappa_a^{(n)} = \frac{d^n K_a(0)}{dt^n}$. For $t = (\mathbf{v}^\top \mathbf{W})_a$ we obtain

$$E(\mathbf{v}) = -\mathbf{a}^\top \mathbf{v} - \sum_{a,n} \kappa_a^{(n)} \frac{((\mathbf{v}^\top \mathbf{W})_a)^n}{n!} \tag{166}$$

Thus, by integrating out the hidden variable degree of freedom, one can recover in the energy of eq. (166) all orders of the input correlations.⁷ This explains the representative power of the RBM model, and of hidden variables in general [125, 132].

⁷ Note that the Boltzmann machine model is recovered by taking a normal probability for $q_a(h_a)$, whose only non-zero cumulant is the second one.

ASSESSING THE ACCURACY OF MACHINE LEARNING THERMODYNAMIC PERTURBATION THEORY: DENSITY FUNCTIONAL THEORY AND BEYOND

4.1	Introduction	56
4.2	Methodology	60
4.3	Results and Discussions	64
4.3.1	Assessing the accuracy of MLPT for different density functional approximations	64
4.3.2	Machine learning Monte Carlo resampling	74
4.3.3	Assessing the accuracy of MLPT for the random phase approximation	82
4.4	Conclusions	84

The content of this chapter is nearly identical to the following published work: [84]

“Assessing the Accuracy of Machine Learning Thermodynamic Perturbation Theory: Density Functional Theory and Beyond”, by Basile Herzog, Maurício Chagas da Silva, Bastien Casier, Michael Badawi, Fabien Pascale, Tomas Bucko, Sébastien Lebègue and Dario Rocca, published in the *Journal of Chemical Theory and Computation* 18.3 (2022), pp. 1382–1394.

4.1 INTRODUCTION

Correlated quantum chemical methods could provide an alternative to density functional theory (DFT) in (periodic) materials simulations possibly reaching the threshold of chemical accuracy (1 kcal/mol) with respect to experimental data. Traditional quantum chemical methods, such as Møller-Plesset perturbation theory to second order (MP2) [137] or coupled-cluster theory [7], have been recently implemented for condensed phase materials applications [20, 45, 49, 127, 151]. An alternative particularly suitable for condensed matter applications is represented by the random phase approximation (RPA) [19, 51, 58, 60, 78, 118, 123] and its variants which include higher order corrections [12, 38, 48, 50, 73, 81, 139]. Due to the significantly high computational cost of these approaches, their use is limited, especially in (finite-temperature) molecular dynamics (MD) simulations.

The use of machine learning (ML) techniques could be highly beneficial for MD simulations and, since the seminal work of Behler and Parrinello (BP) in 2007 [15], has seen an increasing popularity in this field [14, 33, 34, 194]. While keeping a level of accuracy comparable to *ab initio* calculations, ML approaches can be used to replace most of the expensive quantum mechanical calculations with numerically cheap predictions. This allows for an increase in the system size and timescale that would be normally accessible by traditional *ab initio* MD simulations. However, ML models typically employed in MD simulations require a significant amount of data to be trained and this represents an issue for the most expensive quantum mechanical approximations.

In a previous work was applied a scheme that couples machine learning techniques with thermodynamic perturbation theory (MLPT) to compute enthalpies of adsorption of molecules in the zeolite chabazite at the RPA level of theory [28] (the same methodology can be used in the context of free energies of activation [21, 62]). Achieving these results with a brute force molecular dynamics simulation would be completely out of reach. Indeed, by considering only the cost of the RPA energy calculations for the zeolite, completing a 200000 steps MD simulation would require 65 million CPU hours and more than 110 years on 64 cores at 2.6 GHz (this estimate does not include the significant additional cost of computing RPA forces [159]). With our MLPT technique, as few as 10 single point RPA energy calculations were sufficient to train a machine learning model that was then used to predict the RPA energy for several other configurations [28].

In this previous work the van der Waals (vdW) corrected PBE+D2 functional [22, 68] was used to generate a production MD trajectory (in general, a numerically inexpensive approximation should be chosen as production method). Based on the Δ -ML method [157], a model was thereafter trained to predict the difference between RPA and PBE+D2 energies. Following the thermodynamic perturbation approach [32, 153, 165], the energy differences were subsequently used to reweight the statistical weights of the PBE+D2 configurations in order to obtain the RPA canonical distribution and compute the RPA ensemble energies and enthalpies (in this context the RPA is considered the “target” level of theory). Thanks to the inexpensive predictions of the ML model, this technique involves a computational cost that is several orders of magnitude smaller than a full RPA MD simulation.

While the error involved in the ML procedure is rather well controlled [21, 28], the application of thermodynamic perturbation theory (TPT) might be at the origin of a bias in the results estimated at the target level of theory. Indeed, as explained in subsection 2.4.3 the configurational space sampled by the production MD might have suboptimal superposition with the target configurational space [32, 153]. In

certain cases, this might lead to a strong loss of statistical significance when the contributions of the production configurations are reweighted to obtain the target distribution (namely only few configurations could contribute to the whole target level statistics).

In this work, the accuracy of the MLPT approach is assessed using the adsorption of molecules in zeolites as a test case. Indeed, many separation and adsorption processes require the improvement of the dedicated materials, and DFT calculations often provide valuable suggestions of optimized formulations in this regard [3, 30, 85, 106]. However, additional efforts have to be done to find a better compromise between accuracy of prediction and calculation cost. By considering five different DFT functionals, including generalized gradient approximation (GGA), meta-GGA, and non-local vdW corrected functionals, full MD simulations are performed to provide reference values for the ensemble energies and enthalpies of adsorption. Each one of these five functionals is then used as a MLPT production method to obtain the energy/enthalpy estimates for the remaining four functionals, which are considered as target approximations.

By comparing MLPT estimates with reference results the level of accuracy of this approach is established. Some anomalous cases are found, for example when GGA functionals are used as starting point for meta-GGA functionals and vice versa (in certain cases the MLPT estimate of the ensemble energy can deviate by more than 8 kcal/mol). By using machine learning dimensionality reduction algorithms, a qualitative visualization of the relative distributions of the configurations from the different MD simulations is shown. This analysis confirms that the most problematic cases can be ascribed to a poor superposition of the configurational spaces. In order to detect possible failure of MLPT a diagnostic test is proposed, the I_w index, which is evaluated from ML energy predictions and does not rely on the knowledge of any reference results.

We then propose a scheme to significantly improve results even when the I_w coefficient is close to 0 (lowest production-target superposition). This approach, denoted as MLMC, is based on a Monte Carlo resampling of the target configurational space that reuses the machine learning model already trained for MLPT. Without any additional target level calculations, this scheme decreases even the largest deviations within the threshold of chemical accuracy.

Finally, the tools developed in this work are applied to analyze previous results based on the RPA [28], where reference values cannot be produced. The relatively large values of the I_w index hints that the PBE+D2 functional provides a reliable starting point for RPA target properties. The results are stable even if a full MLMC

resampling of the configurational space is performed. This shows that, with a proper choice of the production approximation, MLPT allows for a quick and accurate estimate of target level properties. The MLPT approach opens the way to a more systematic application of accurate but expensive DFT and quantum chemical approximations in finite temperature simulations of materials.

In the following Section 4.2, the methodological part of this work and computational details are discussed. In Section 4.3, numerical results are presented and analyzed. Section 4.4 contains the conclusions.

4.2 METHODOLOGY

The enthalpies of adsorption ($\Delta_{\text{ads}}H$) of the CH_4 molecule in protonated chabazite (HChab) and CO_2 molecule in siliceous chabazite (SiChab) were investigated by employing *ab initio* molecular dynamics (AIMD) and machine learning thermodynamic perturbation theory [21, 28, 62] (MLPT). The models of the adsorbed molecules are shown in Figure 13. The enthalpy of adsorption is defined by $\Delta_{\text{ads}}H = \Delta_{\text{ads}}U + \Delta_{\text{ads}}(pV)$. In the low coverage limit (few molecules are adsorbed), the change of pV is negligible between the adsorbed system and the substrate ($pV(\text{M@Z}) \approx pV(\text{Z})$), so that, assuming an ideal gas behaviour for the adsorbate, $\Delta_{\text{ads}}(pV) = -k_B T$. The enthalpy of adsorption is then given by

$$\Delta_{\text{ads}}H(\text{M@Z}) = \langle E(\text{M@Z}) \rangle - (\langle E(\text{M}) \rangle + \langle E(\text{Z}) \rangle) - k_B T \quad (167)$$

where $\langle E \rangle$ is the internal energy computed as ensemble average of potential energy via the AIMD simulations or MLPT (M denotes the molecules, Z the zeolites, and M@Z the adsorbed system), k_B is the Boltzmann constant, and T is the system temperature (equal to 300 K in all our simulations).

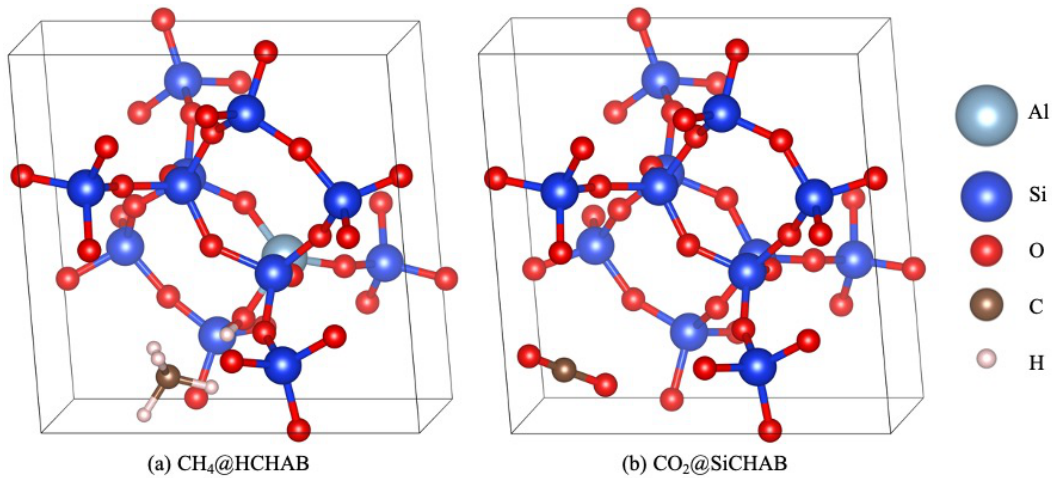


Figure 13: Adsorbed systems studied. (a) CH_4 in protonated Chabazite (HChab) and (b) CO_2 in siliceous chabazite (SiChab).

The goal of this work is to benchmark the accuracy of MLPT by comparing its estimates with reference values obtained from full MD simulations. We chose five DFT functionals in order to cover a range of different characteristics but keeping also into account their numerical cost, as long and stable MD trajectories have to be produced for reference. Specifically, was selected: a generalized gradient approximation (GGA) functional, PBE [144] and its version with corrections for van der Waals interactions, PBE+D2 [22, 68]; a meta-GGA functional, SCAN [182, 183];

a non-local vdW functional, vdW-DF2 [47, 166]; a version of SCAN with non-local vdW corrections, SCAN+rVV10 [142]. We will also discuss results based on the random phase approximation (RPA), as implemented in the VASP code [112, 113, 115]; due to the high numerical cost a direct MD simulation at this level of theory is completely out of reach and the MLPT method becomes instrumental to estimate finite temperature properties.

All AIMD simulations were performed within the NVT ensemble, whereby temperature was maintained by Andersen [4] thermostat with collision probability of 0.05. An integration time step of 0.5 fs was considered and the total simulation time of each MD run was 100 ps, i.e. $\sim 200 \cdot 10^3$ configurations were generated with each method for each system investigated in this work. The initial part of 10 ps of each trajectory was considered as an equilibration period and the corresponding data were discarded. The remaining part of the trajectory was tested for the absence of any drift via Mann-Kendall tests [175]. All the standard errors on the ensemble energies of these trajectories are below 0.2 kcal/mol.

In the AIMD simulations a cell fixed at the lattice parameters of the chabazite optimized at the PBE level is used ($a = b = c = 9.34 \text{ \AA}$ and $\alpha = \beta = \gamma = 95^\circ$). All AIMD simulations as well as single point calculations were carried out with the VASP electronic structure package [78, 112, 113, 115]. PAW pseudopotentials [117] for all atoms with the default kinetic energy cutoffs were used to expand the wavefunctions. The “accurate” precision setting was used and the hydrogen atomic mass was increased to 3.0 a.u. The Γ point approximation was applied in all calculations.

The results produced by AIMD are used as a reference for our MLPT method. Within this approach an MD simulation is first performed at a certain production level of theory and then “corrected” to obtain results at one or multiple target levels of theory. One of the advantages of MLPT is that an estimate of a finite-temperature property for the target method can be obtained with only a few tens or few hundreds of single point calculations [28]. Out of about $180 \cdot 10^3$ configurations generated in production runs, a reduced sample of about $19 \cdot 10^3$ evenly separated configurations (X_i) and corresponding potential energies obtained with the production method (E_i) are selected to be used for the predictions of the energies at the new target functional level. We note that this configuration reduction is not strictly necessary but, since MD configurations close in time are correlated, this procedure speeds up the predictions of the ML model without compromising the accuracy of prediction.

In order to compute target level ensemble energies $\langle E' \rangle$, MLPT applies TPT to reweight the production statistical distribution [153]

$$\langle E' \rangle = \frac{\sum_i^M E'_i \exp(-\beta \Delta E_i)}{\sum_i^M \exp(-\beta \Delta E_i)} \quad (168)$$

where E'_i denotes the target level energies of the i th configuration, $\Delta E_i = E'_i - E_i$ is the difference between target and production energies, $\beta = 1/k_B T$, and M is the number of configurations, in this case $19 \cdot 10^3$.

The application of eq. (168) still requires a large number (M) of calculations at the target level of theory. In MLPT only a very small number N_{train} of these calculations is actually performed and then used to train an ML model that predicts all the remaining $M - N_{\text{train}}$ energies. In order to decrease the number of target calculations required to train the ML algorithm instead of considering directly the target energy E'_i , we build a ML regression model using ΔE_i . This approach, which is just one of the many possible applications of the Δ -ML idea [46, 157], takes advantage of the much smoother dependence of ΔE_i on E_i in comparison with E'_i and, accordingly, is easier to predict. To represent the configurations within our ML model we use the *Smooth Overlap of Atomic Positions* (SOAP) [9] descriptor, as implemented in the Dscribe library [86]. While several other descriptors for periodic materials have been proposed in the literature [13, 26, 94], the choice of the SOAP descriptor provides already a satisfactory level of accuracy for MLPT applications [21, 28, 62].

The SOAP approach leads to a natural definition of a kernel for local atomic environments that can be subsequently used to define the global rematch kernel [181]. This kernel, which can be intuitively seen as a measure of similarity between configurations, is then used in the framework of the kernel ridge regression (KRR) [181] ML algorithm; the KRR implementation in the Scikit-learning package is used in this work [141]. To train the ML model, 200 evenly distributed configurations are chosen from the $19 \cdot 10^3$ structures of the production trajectory. An additional independent set of 25 configurations is also selected to test the accuracy of the predictions of the ML model; for all the applications considered in this work the root mean square error (RMSE) in the prediction of the energy of single configurations is at most 0.3 kcal/mol. Single point calculations at the target level are performed exclusively for these 225 configurations. The 18775 (i.e., 19000-225) remaining values of ΔE_i (and $E'_i = E_i + \Delta E_i$) necessary to evaluate eq. (168) are inexpensively predicted by the ML model. This shows the clear advantage of using MLPT, in particular if it is necessary to consider multiple target level theories or highly expensive approximations (e.g., the RPA).

As discussed in the next sections, MLPT does not provide satisfactory results in certain cases. This happens for the estimate of SCAN and SCAN+rVV10 adsorption enthalpies from PBE production calculations. This issue is related to the unsatisfactory superposition of the target and production configurational spaces. To improve the results in these cases we have reused the ML model from the MLPT application to resample the target level configurational space using a Monte Carlo (MC) algorithm [16, 111, 133, 140]. We will denote this procedure as MLMC.

In our MC procedure, multiple types of trial moves were used. First type of the move applied to all systems was based on a random displacement of atomic positions (\mathbf{x}) along the velocity vector (\mathbf{v}) drawn at random from the Maxwell-Boltzmann distribution corresponding to our target T of 300 K realized according to the formula $\mathbf{x}_{new} = \mathbf{x}_{old} + \mathbf{v}\Delta t$, with Δt set to 0.5 fs. In the second type of move, used only for the adsorbate+substrate systems ($\text{CO}_2@SiChab$ and $\text{CH}_4@HChab$), an extra random translation (maximal magnitude of 0.5 Å) and rotation (by up to 35.0 deg.) of the molecular adsorbate was attempted. Further improvements in efficiency could probably be achieved by implementing more sophisticated sampling techniques. This is, however, beyond the scope of the present work, which primarily focuses on design of strategy to overcome the overlap problem occurring in the MLPT simulations.

The MC approach used to sample the canonical ensemble is equivalent to MD but is simpler to implement and does not require the calculation of forces (for certain methods forces are not available in most solid state physics software implementations). By using our in-house software, the MLMC approach recursively sample new configurations following these steps: (1) A new geometric configuration is randomly generated; (2) A production level *ab initio* calculation is performed to obtain E_i for this new configuration; (3) the ML learning model is used to predict ΔE_i and E'_i ; (4) the new configuration is accepted or rejected according to the Metropolis criterion [16, 111, 133, 140]. The advantage of MLMC is that with respect to MLPT no additional calculations are required at the target level of theory and this is particularly important when computationally expensive quantum chemical methods are used. The trade-off is that each step of the MC procedure requires a new production level calculation. In the applications of next section we have sampled $400 \cdot 10^3$ steps for each MLMC trajectory to get converged ensemble statistics (for all the systems considered in this work the standard error on the MC ensemble energies is at most 0.3 kcal/mol).

4.3 RESULTS AND DISCUSSIONS

4.3.1 *Assessing the accuracy of MLPT for different density functional approximations*

The experimentally measured enthalpies of adsorption of the investigated systems are rather small, -4.06 kcal/mol and -5.02 kcal/mol for $\text{CH}_4 @\text{HChab}$ and $\text{CO}_2 @\text{SiChab}$, respectively. Indeed, for both systems, the adsorption process is dominated by weak long-range van der Waals (vdW) interactions, which are, due to the small size of both adsorbate molecules, only modest. Ideally, very accurate methods such as RPA, MP2 or CCSD(T) would be necessary to model with systematic improvements in accuracy such processes in which long-range interactions play an important role. While the use of such methods is prohibitive due to their enormous computational cost, machine learning techniques could play an important role in extending their applicability to realistic systems [28]. This requires ML techniques that reasonably preserve the accuracy of these high-level quantum chemical methods and have sufficient predictive power for configurations beyond the training set. In this section we will discuss the accuracy of MLPT for the case of computationally affordable DFT functionals, for which a reference can be easily obtained. At the end of this section we will consider the RPA enthalpies of adsorption.

In Table 1, we report the values of the enthalpies of adsorption (eq. (167)) for CH_4 and CO_2 in zeolite as obtained from the full MD simulations. These results will be used as reference values for assessing the quality of our MLPT methodology. Comparing the calculated and the experimental values for the enthalpies of adsorption in Table 1, it can be noticed that only functionals with some vdW corrections can achieve a reasonable agreement with the experimental values, although the accuracy improvement is not systematic. For the $\text{CH}_4 @\text{HChab}$ system, SCAN+rVV10 was identified as the best performing functional, with a deviation of -0.71 kcal/mol (17% of deviation), whereas for the $\text{CO}_2 @\text{SiChab}$ the best functional was PBE+D2 with a deviation of -0.46 kcal/mol (9% of deviation). On the other hand, the uncorrected PBE and SCAN functionals show a sizable underestimation in the calculated enthalpies of adsorption for the CH_4 and CO_2 molecules in the zeolites, with relative errors with respect to experiment higher than 35%. However, the main purpose of this work is not to compare the accuracy of different functionals compared to experiment but to assess the performance and limitations of MLPT to reproduce reference AIMD results obtained using different quantum mechanical approximations.

Starting from each of the 5 trajectories obtained using PBE, PBE+D2, SCAN, vdW-DF2, and SCAN+rVV10 production methods, the MLPT approach is used

Table 1: Experimental and reference theoretical enthalpies of adsorption for the CH₄ and CO₂ molecules in zeolites obtained using different functionals. All values are expressed in kcal/mol.

System	PBE	PBE+D2	SCAN	vdW-DF2	SCAN+rVV10	Exp.
CH ₄ @HChab	-1.13	-6.08	-2.60	-5.64	-4.77	-4.06 [148]
CO ₂ @SiChab	-1.18	-5.48	-2.95	-6.93	-7.11	-5.02 [126]

to predict the results for all remaining approximations. The results are shown in Table 2. In this table, columns and lines represent the production and target methods, respectively. If, for example, we consider the PBE column (below Production - MD simulations), and the SCAN line, we find an estimate of the SCAN enthalpy of adsorption that is obtained by applying MLPT to the PBE production trajectory; this estimate avoids completely the generation of a new SCAN molecular dynamics trajectory and requires only 200 additional SCAN calculations that are used to train the ML model. For the sake of completeness, we also report values on the diagonal of this table, namely MLPT values based on the same production and target levels of theory. This shows that some small fluctuations can be introduced by the ML model even in this particularly simple case, additionally to those regular/general fluctuations due to the AIMD convergence and statistical uncertainty.

The accuracy of the MLPT estimates can be established by comparing the MLPT results with reference values of Table 1. The corresponding deviations are reported in parenthesis in Table 2. Altogether, it can be noticed that the predictions within the PBE/PBE+D2/vdW-DF2 group and the SCAN/SCAN+rVV10 are reasonable with deviations typically within few tenths of kcal/mol and only in one case slightly beyond chemical accuracy (a deviation of 1.01 kcal/mol for the prediction of SCAN+rVV10 from SCAN for CO₂ @SiChab). When the production and target methods belong to two different groups of functional, the MLPT is significantly less reliable and significantly large deviations appear. The most problematic cases involve the prediction of CH₄ @HChab enthalpy of adsorption at the SCAN and SCAN+rVV10 levels of theory from PBE, where the deviation is larger than 8 kcal/mol.

Table 2: Enthalpies of adsorption estimates of CH₄ @HChab and CO₂ @SiChab computed directly from straightforward MD (Ref. column) as well as using the MLPT method, whereby all functionals have been used as production as well as target methods. Deviations of MLPT results from the reference values (Ref. column) are given in parenthesis. All values are expressed in kcal/mol. The results with deviations above 1.0 kcal/mol are in bold.

System	Target - MLPT	Ref.	Production - MD simulations				
			PBE	PBE+D2	SCAN	vdW-DF2	SCAN+rVV10
CH ₄ @HChab Exp. -4.06 [148]	PBE	-1.13	-1.18 ^(-0.05)	-1.05 ^(0.07)	1.25^(2.38)	-0.66 ^(0.47)	0.23^(1.36)
	PBE+D2	-6.08	-5.99 ^(0.09)	-6.08 ^(0.00)	-2.64^(3.44)	-5.10 ^(0.98)	-5.65 ^(0.43)
	SCAN	-2.60	-11.13^(-8.54)	-4.78^(-2.18)	-2.65 ^(-0.05)	-1.84 ^(0.75)	-2.37 ^(0.23)
	vdW-DF2	-5.64	-6.24 ^(-0.59)	-5.82 ^(-0.18)	-0.92^(4.73)	-5.53 ^(0.11)	-5.75 ^(-0.11)
	SCAN+rVV10	-4.77	-13.63^(-8.86)	-6.71^(-1.94)	-5.03 ^(-0.26)	-4.13 ^(0.64)	-4.74 ^(0.03)
CO ₂ @SiChab Exp. -5.02 [126]	PBE	-1.18	-1.15 ^(0.02)	-1.47 ^(-0.29)	-1.46 ^(-0.29)	-1.24 ^(-0.06)	-4.68^(-3.50)
	PBE+D2	-5.48	-5.61 ^(-0.14)	-5.44 ^(0.04)	-5.40 ^(0.08)	-5.60 ^(-0.12)	-7.56^(-2.08)
	SCAN	-2.95	-1.42^(1.54)	-2.74 ^(0.21)	-2.96 ^(-0.01)	-4.30^(-1.34)	-3.72 ^(-0.77)
	vdW-DF2	-6.93	-7.47 ^(-0.54)	-7.57 ^(-0.64)	-8.11^(-1.17)	-6.94 ^(-0.00)	-8.90^(-1.96)
	SCAN+rVV10	-7.11	-3.97^(3.14)	-5.52^(1.59)	-6.01^(1.10)	-7.17 ^(-0.06)	-7.09 ^(0.02)

As the enthalpies of adsorption are computed from differences of internal energies of interacting and non-interacting systems, fortuitous error cancellations might arise that could mask the real accuracy of MLPT predictions. To shed some light onto this problem, we report in Table 3 MLPT total ensemble energies for the individual components involved in the adsorption process (adsorbed system, zeolite alone, molecule alone). These results confirm the conclusions previously drawn from Table 2, with the predictions made for the cases where the production and target methods are both from the same group of functionals (i.e., either PBE/PBE+D2/vdW-DF2 or SCAN/SCAN+rVV10) being well within the chemical accuracy.

MLPT becomes unreliable when predictions mix these two groups, with deviations that often reach several kcal/mol (these observations do not hold for the standalone CH₄ and CO₂ molecules, whose deviations are always below 0.1 kcal/mol). The worst performance is confirmed to correspond to the prediction of SCAN and SCAN+rVV10 from the PBE production trajectory of CH₄@HChab. For these cases, we notice that the largest deviations in the enthalpies of adsorption (>8 kcal/mol) are not only due to the low accuracy of the MLPT ensemble energy estimates for CH₄@HChab and HChab, but also to the opposite sign in deviations of predictions made for these two systems, which leads to an error accumulation when evaluating eq. (167).

While MLPT allows, in principle, for a quick evaluation of finite-temperature properties at one or more target levels of theory from a single MD production run (typically based on the most computationally inexpensive approximation), the results in Tables 2-3 show that such a strategy does not always allow one to achieve the required level of accuracy. Accordingly, special care should be taken in choosing a production method suitable for the target approximation(s) of interest. The failure of MLPT in certain cases can be explained by the limitations of the thermodynamic perturbation theory (TPT) itself, rather than by an inaccurate ML model. Indeed, if the configurational spaces visited with high likelihood by the production and target approximations do not overlap sufficiently, the TPT has low predictive power [153]. In the most anomalous cases it can happen that this overlap is so poor that only one or few individual configurations effectively contribute to the ensemble average reweighted according to eq. (168).

Table 3: Deviations of the MLPT estimates of the target internal energies of individual systems from reference values. All values are in kcal/mol and all values larger than 1.0 kcal/mol in absolute value are in bold.

System	Target - MLPT	Production - MD simulations				
		PBE	PBE+D2	SCAN	vdW-DF2	SCAN+rVV10
CH ₄ @HChab	PBE	-0.007	0.053	-1.348	-0.237	-1.582
	PBE+D2	0.197	0.001	-0.547	0.405	-2.987
	SCAN	-4.380	-2.137	-0.029	1.497	0.204
	vdW-DF2	-0.378	0.206	0.285	0.060	-3.799
	SCAN+rVV10	-5.026	-2.036	-0.221	1.012	-0.001
CO ₂ @SiChab	PBE	0.014	-0.693	-2.946	-0.434	-3.689
	PBE+D2	0.245	0.041	-2.267	0.015	-1.665
	SCAN	2.249	1.244	-0.013	2.448	-0.594
	vdW-DF2	-0.053	-0.622	-4.138	-0.002	-2.122
	SCAN+rVV10	3.395	2.289	0.899	3.251	0.025
HChab	PBE	0.035	0.025	-3.706	-0.712	-2.953
	PBE+D2	0.054	0.000	-4.007	-0.626	-3.479
	SCAN	4.142	0.084	0.025	0.718	-0.034
	vdW-DF2	0.230	0.453	-4.407	-0.039	-3.678
	SCAN+rVV10	3.830	-0.045	0.051	0.359	-0.027
SiChab	PBE	-0.012	-0.380	-2.634	-0.348	-0.230
	PBE+D2	0.360	0.001	-2.337	0.145	0.353
	SCAN	0.659	1.050	0.002	3.809	0.139
	vdW-DF2	0.453	0.008	-2.972	0.001	-0.228
	SCAN+rVV10	0.246	0.745	-0.158	3.358	0.000
CH ₄	PBE	0.006	-0.047	-0.024	0.009	0.012
	PBE+D2	0.053	0.000	0.021	0.053	0.062
	SCAN	0.013	-0.039	-0.001	0.027	0.012
	vdW-DF2	-0.014	-0.067	-0.033	-0.007	-0.014
	SCAN+rVV10	0.000	-0.052	-0.013	0.014	-0.000
CO ₂	PBE	0.003	-0.018	-0.026	-0.025	0.041
	PBE+D2	0.021	-0.000	-0.007	-0.007	0.060
	SCAN	0.054	-0.018	-0.006	-0.018	0.037
	vdW-DF2	0.030	0.007	0.004	-0.001	0.070
	SCAN+rVV10	0.013	-0.049	-0.043	-0.051	0.001

In order to analyze qualitatively this behavior we consider all the production MD simulations based on the 5 different functionals. By using the t-distributed stochastic neighbor embedding (t-SNE) algorithm [43, 190] we present in Figure 14 two-dimensional visualizations of the high-dimensional configurational space spanned by these trajectories (to improve the readability, figures were created using only 500 uncorrelated structures selected from each trajectory). To be consistent with the ML learning algorithm used in this study for the regression, the t-SNE approach was applied using a definition of distance D between two configurations χ^A and χ^B based on the normalized SOAP kernel K :

$$D(\chi^A, \chi^B) = (1 - K(\chi^A, \chi^B))^{\frac{1}{2}}; \quad (169)$$

this definition also inherits some of the properties of the SOAP kernel, such as the rotational and translational invariance required in materials and molecular modelling.

It can be noticed in Figure 14 that for $\text{CH}_4 @ \text{HChab}$, $\text{CO}_2 @ \text{SiChab}$, HChab , and SiChab the trajectories generated by the two groups of functionals (PBE/PBE+D2/vdW-DF2 and SCAN/SCAN+rVV10) form two clusters with very limited overlap. This shows that the configurational spaces spanned by the two groups of approximations are to a large extent different, which is at the origin of the poor performance of MLPT in the cases discussed above. For the gas phase molecules CH_4 and CO_2 , the configurational space has a much simpler structure and the five different approximations produce trajectories which always largely overlap; this is also consistent with the results in Table 3, where MLPT is highly accurate for all molecular energies.

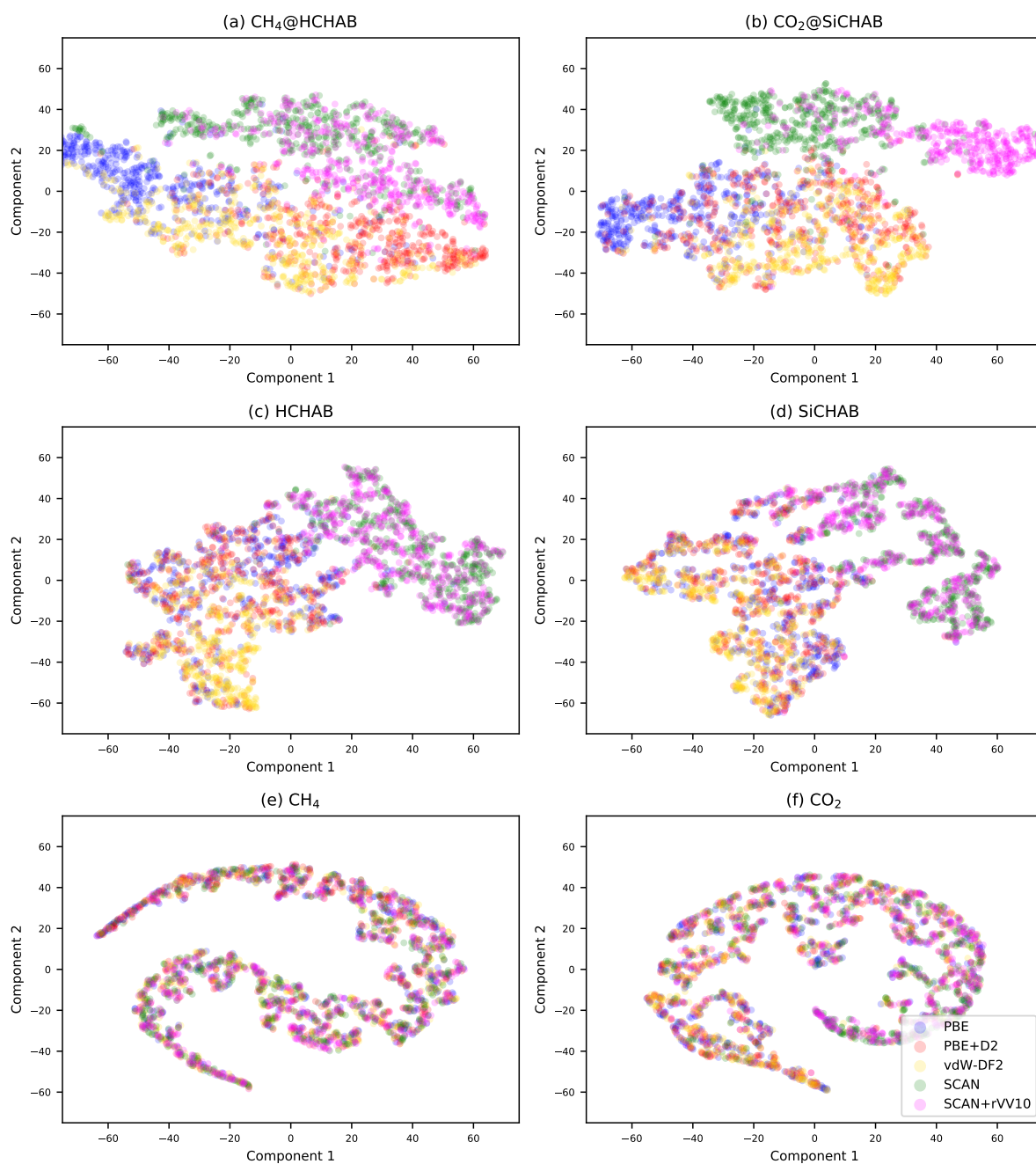


Figure 14: t-SNE representation of the configurational spaces spanned by the different functionals. This visualization is based on 500 selected configurations from each molecular dynamics trajectory. principal component of the SOAP descriptor for each 500 configurations of each functionals.

The t-SNE visualization provides useful insights into the potential failures of TPT and, consequently, MLPT. However, a diagnostic approach that requires the full MD trajectories is highly impractical in reality, since the goal of MLPT is indeed to completely avoid the MD simulation at the target level of theory. In order to overcome this difficulty and introduce a tool to analyze and predict the failures of MLPT, we follow the approach introduced in ref. [21] and introduce the I_w index. By considering the weight factor (w_i) in eq. (168)

$$w_i = \exp(-\beta\Delta E_i) \quad (170)$$

we define $I_w = \frac{(M-N)}{M}$ where M is the total number of configurations sampled in the production run and N is the lowest integer which satisfies the condition

$$\frac{\sum_i^N w_i}{\sum_j^M w_j} \geq 0.5, \quad (171)$$

where the weights w_i have been sorted in ascending order. In practice, the I_w index gives the fraction of configurations that contribute, after the reweighting by eq. (168), one half of the total statistical weight.

In the case of an extremely small overlap between configurational spaces sampled by the production and target approximations, we would have $N \approx M$ (since most weights w_i would be close to 0, N has to approach M to satisfy the condition in eq. (171)) and $I_w \approx 0$. This corresponds to a significant loss in the statistical significance since the large number of configurations sampled in the production run is reduced to a few individual structures. In the opposite case of excellent overlap we would have $N \approx M/2$ and $I_w \approx 0.5$; if for example the production and target approximations provides energies that differ only by a constant, we would obtain $I_w = 0.5$.

The I_w values for all systems investigated in this work are presented in Table 4; in order to simplify the analysis, we are also reporting in parenthesis the deviations of the MLPT ensemble energies with respect to the MD references (same as in Table 3). As expected the I_w index is equal to 0.5 when the production and target methods are the same. The values of I_w allow for an interpretation of the deviations previously reported in Table 3. All the largest errors of MLPT (often significantly above chemical accuracy) correspond to $I_w = 0.00$, namely to a small correlation/overlap of production and target methods. In few cases, small deviations are found for systems with $I_w = 0.00$ (for example the MLPT evaluation of the SCAN ensemble energy from PBE+D2 for HChab) but these results must be considered as coincidental, whereby the reweighting of a very small number of configurations led, by chance, to a satisfactory reconstruction of the target statisti-

cal distribution.

As discussed above, the MLPT estimates become reliable for the production and target method combinations from within the two groups of functionals PBE/PBE+D2/vdW-DF2 and SCAN/SCAN+rVV10. As shown in Table 4 the I_w values within this two groups are greater than 0.03. For the isolated gas phase molecules CH₄ and CO₂ we always find large values of I_w (>0.1) and consequently the errors on the energy predictions are small (within 0.18 kcal/mol). These remarks point to the fact that the I_w can be used for a qualitative *a posteriori* estimate of the reliability of MLPT. In the case of an I_w close to 0, the MLPT estimates are unreliable and often deviate by several kcal/mol from the reference values. Non-vanishing values of I_w , ideally as close as possible to 0.5, correspond to accurate MLPT predictions. In the specific case of the systems considered here, an I_w of about 0.03 already provides ensemble energies well within chemical accuracy.

Table 4: Values of the I_w index corresponding to each individual MLPT estimate of the ensemble internal energy. The values in parentheses represent the deviations in kcal/mol from the reference values obtained from straightforward MD. In bold, all deviations large than 1.0 kcal/mol.

System	Target - MLPT	Production - MD simulations				
		PBE	PBE+D2	SCAN	vdW-DF2	SCAN+rVV10
CH ₄ @HChab	PBE	0.50 ^(-0.01)	0.06 ^(0.05)	0.00 ^(-1.35)	0.03 ^(-0.24)	0.00 ^(-1.58)
	PBE+D2	0.03 ^(0.20)	0.50 ^(0.00)	0.00 ^(-0.55)	0.06 ^(0.41)	0.00 ^(-2.99)
	SCAN	0.00 ^(-4.38)	0.00 ^(-2.14)	0.50 ^(-0.03)	0.00 ^(1.50)	0.35 ^(0.20)
	vdW-DF2	0.03 ^(-0.38)	0.05 ^(0.21)	0.00 ^(0.28)	0.50 ^(0.06)	0.00 ^(-3.80)
	SCAN+rVV10	0.00 ^(-5.03)	0.00 ^(-2.04)	0.34 ^(-0.22)	0.00 ^(1.01)	0.50 ^(-0.00)
CO ₂ @SiChab	PBE	0.50 ^(0.01)	0.10 ^(-0.69)	0.00 ^(-2.95)	0.04 ^(-0.43)	0.00 ^(-3.69)
	PBE+D2	0.08 ^(0.24)	0.50 ^(0.04)	0.00 ^(-2.27)	0.12 ^(0.02)	0.00 ^(-1.67)
	SCAN	0.00 ^(2.25)	0.00 ^(1.24)	0.50 ^(-0.01)	0.00 ^(2.45)	0.35 ^(-0.59)
	vdW-DF2	0.03 ^(-0.05)	0.13 ^(-0.62)	0.00 ^(-4.14)	0.50 ^(-0.00)	0.00 ^(-2.12)
	SCAN+rVV10	0.00 ^(3.39)	0.00 ^(2.29)	0.32 ^(0.90)	0.00 ^(3.25)	0.50 ^(0.02)
HChab	PBE	0.50 ^(0.03)	0.31 ^(0.03)	0.00 ^(-3.71)	0.09 ^(-0.71)	0.00 ^(-2.95)
	PBE+D2	0.32 ^(0.05)	0.50 ^(0.00)	0.00 ^(-4.01)	0.11 ^(-0.63)	0.00 ^(-3.48)
	SCAN	0.00 ^(4.14)	0.00 ^(0.08)	0.50 ^(0.02)	0.00 ^(0.72)	0.39 ^(-0.03)
	vdW-DF2	0.08 ^(0.23)	0.11 ^(0.45)	0.00 ^(-4.41)	0.50 ^(-0.04)	0.00 ^(-3.68)
	SCAN+rVV10	0.00 ^(3.83)	0.00 ^(-0.05)	0.39 ^(0.05)	0.00 ^(0.36)	0.50 ^(-0.03)
SiChab	PBE	0.50 ^(-0.01)	0.36 ^(-0.38)	0.00 ^(-2.63)	0.14 ^(-0.35)	0.00 ^(-0.23)
	PBE+D2	0.35 ^(0.36)	0.50 ^(0.00)	0.00 ^(-2.34)	0.14 ^(0.14)	0.00 ^(0.35)
	SCAN	0.00 ^(0.66)	0.00 ^(1.05)	0.50 ^(0.00)	0.00 ^(3.81)	0.40 ^(0.14)
	vdW-DF2	0.13 ^(0.45)	0.15 ^(0.01)	0.00 ^(-2.97)	0.50 ^(0.00)	0.00 ^(-0.23)
	SCAN+rVV10	0.00 ^(0.25)	0.00 ^(0.74)	0.40 ^(-0.16)	0.00 ^(3.36)	0.50 ^(0.00)
CH ₄	PBE	0.50 ^(0.01)	0.46 ^(-0.05)	0.29 ^(-0.02)	0.33 ^(0.01)	0.28 ^(0.01)
	PBE+D2	0.46 ^(0.05)	0.50 ^(0.00)	0.26 ^(0.02)	0.29 ^(0.05)	0.25 ^(0.06)
	SCAN	0.29 ^(0.01)	0.26 ^(-0.04)	0.50 ^(-0.00)	0.45 ^(0.03)	0.49 ^(0.01)
	vdW-DF2	0.33 ^(-0.01)	0.30 ^(-0.07)	0.45 ^(-0.03)	0.50 ^(-0.01)	0.44 ^(-0.01)
	SCAN+rVV10	0.29 ^(0.00)	0.25 ^(-0.05)	0.49 ^(-0.01)	0.44 ^(0.01)	0.50 ^(-0.00)
CO ₂	PBE	0.50 ^(0.00)	0.50 ^(-0.02)	0.12 ^(-0.03)	0.48 ^(-0.02)	0.14 ^(0.04)
	PBE+D2	0.50 ^(0.02)	0.50 ^(-0.00)	0.11 ^(-0.01)	0.48 ^(-0.01)	0.14 ^(0.06)
	SCAN	0.11 ^(0.05)	0.12 ^(-0.02)	0.50 ^(-0.01)	0.11 ^(-0.02)	0.45 ^(0.04)
	vdW-DF2	0.48 ^(0.03)	0.48 ^(0.01)	0.11 ^(0.00)	0.50 ^(-0.00)	0.13 ^(0.07)
	SCAN+rVV10	0.14 ^(0.01)	0.15 ^(-0.05)	0.45 ^(-0.04)	0.14 ^(-0.05)	0.50 ^(0.00)

4.3.2 Machine learning Monte Carlo resampling

Up to this point we have discussed the origin of MLPT inaccuracies occurring in certain cases. Ideally, the production trajectory should be chosen to suit the target approximations of interest. However, this is not always possible and it is of interest to improve MLPT results also in the most anomalous cases with $I_w \approx 0$. Recently, Rizzi *et al.* [163] have discussed a ML-based approach to overcome the limitations of TPT for free energy calculations in cases where there is none or very limited overlap of the configurational spaces of production and target methodologies. Following the previous approach presented by Wirnsberger *et al.* [192], Rizzi *et al.* [163] reported the use of a configurational space transformation built on a neural network model.

While this methodology is appealing, in this work we decided to pursue an approach based on the Monte Carlo resampling of the trajectory. This machine learning Monte Carlo (MLMC) method has the following advantages: The implementation is straightforward and does not require the calculation of forces, which often are not implemented in solid state physics codes for methods such as RPA, MP2, or CC; the same ML model of MLPT is reused and no additional calculations at the target level of theory are required; the MC resampling can be used also to verify results when a good or reasonable value of I_w is found. Finally, let us remark that the kernel methods used in our MLPT and MLMC methods typically require much smaller number of training configurations than neural networks [97] used in the method of Wirnsberger *et al.* [192].

To test the MLMC approach, we consider the prediction of SCAN and SCAN+rVV10 enthalpies of adsorption from PBE production calculations which showed the largest deviations between MLPT estimates and reference values. In Figure 15 we show the following probability distributions of the SCAN energy determined for each zeolite and adsorbed system: the reference from the SCAN molecular dynamics (in blue labeled as reference); the distribution of the ML predicted SCAN energies for the configurations sampled by the PBE AIMD (denoted as ML in green); the distribution of SCAN energies from the MLMC resampling (in red). The reference and ML distributions show an unsatisfactory overlap, further confirming the unsatisfactory sampling of PBE to predict SCAN properties. Since no reweighting of energies (such as in eq. (168)) has been performed to generate the histograms shown in Figure 15, the ML distributions are strongly biased, and therefore deviate significantly from the reference MD results. The MLMC approach largely overcomes this issue, as it can be noticed from the good overlap of the blue and red curves. Computing the partial radial distribution function for the Si-O pairs, Figure 16, it is also evident that the MLMC generates geometries which are in a

better agreement with those obtained from straightforward MD simulations with the target functional SCAN/SCAN+rVV₁₀.

The improvement in the geometric sampling can be also demonstrated by including the MC data in the t-SNE analysis. Figure 17 shows indeed that the clusters of the SCAN and SCAN+rVV₁₀ configurations sampled by the Monte Carlo have an excellent overlap with those obtained from the reference MD simulation performed at the SCAN level. In this context it is important to emphasize again that the MLMC results are obtained without any additional explicit target level calculation.

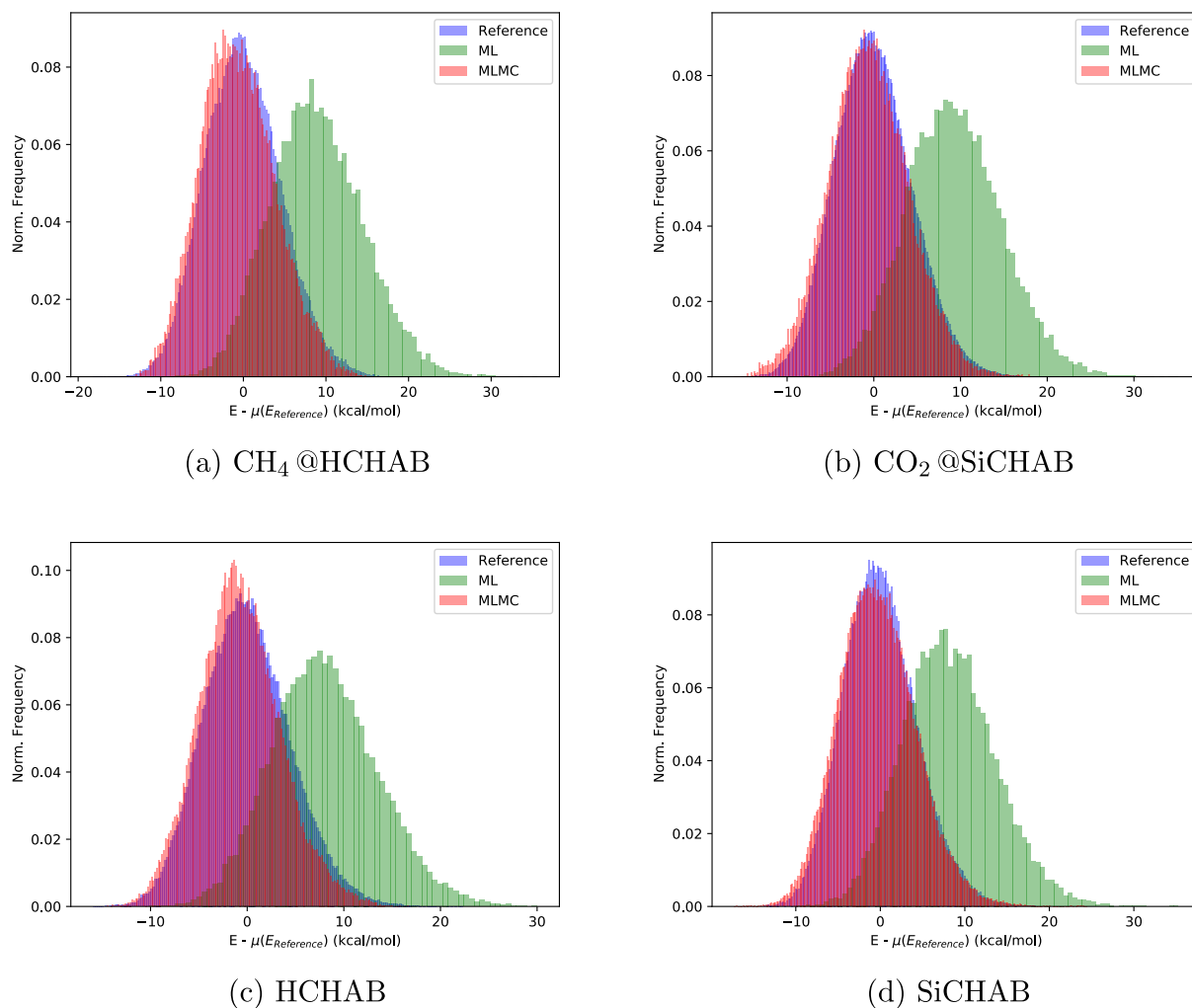


Figure 15: Distributions of deviations of SCAN energy from the average value of the corresponding reference calculation ($\mu(E_{\text{Reference}})$) obtained in three different simulations: straightforward MD with the SCAN functional (blue), ML (green) and MLMC (red) with PBE production method.

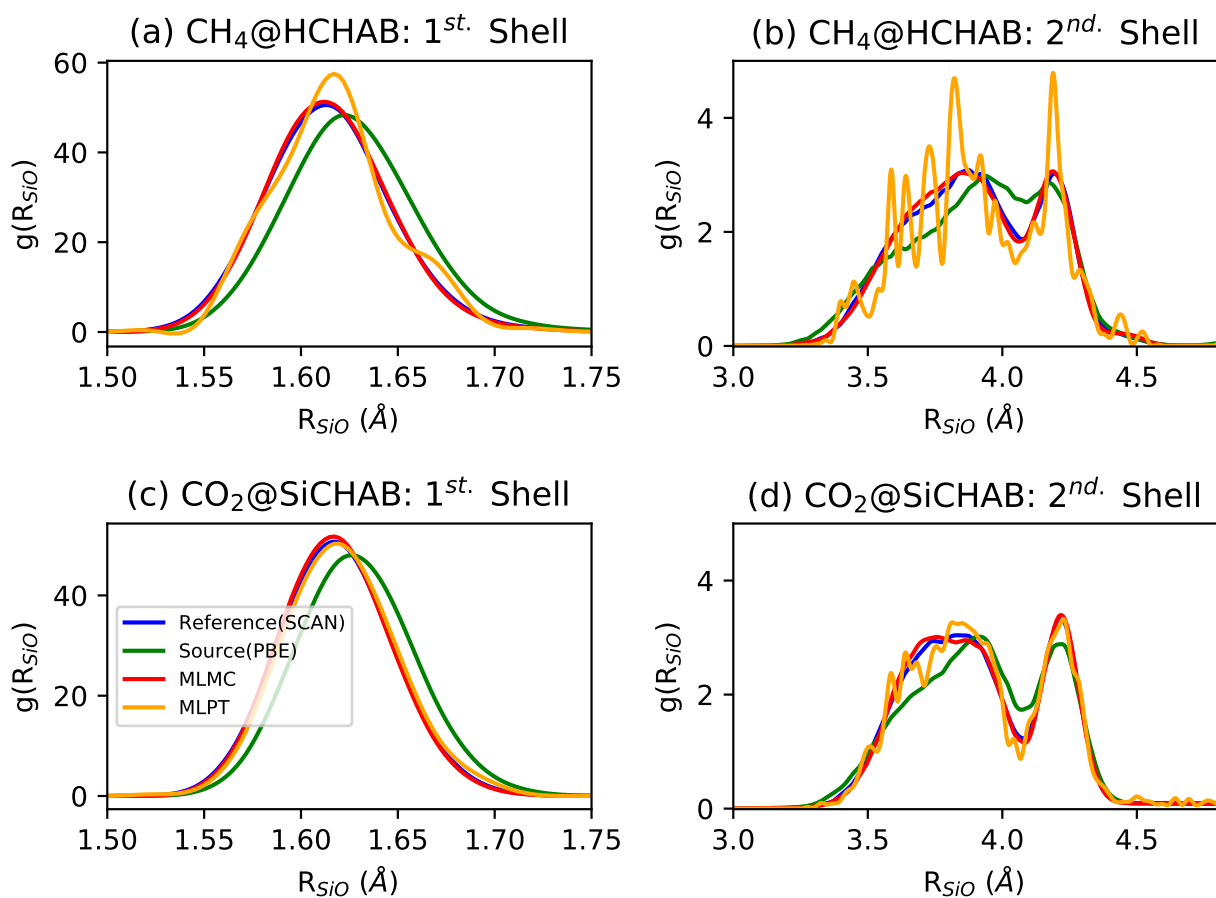


Figure 16: Radial distribution functions calculated using the SCAN functional for the first and second coordination sphere of O atoms around Si. The calculations were performed using a straightforward MD simulation performed at the SCAN level (Reference), and via the MLPT and MLMC methods based on the PBE production method.

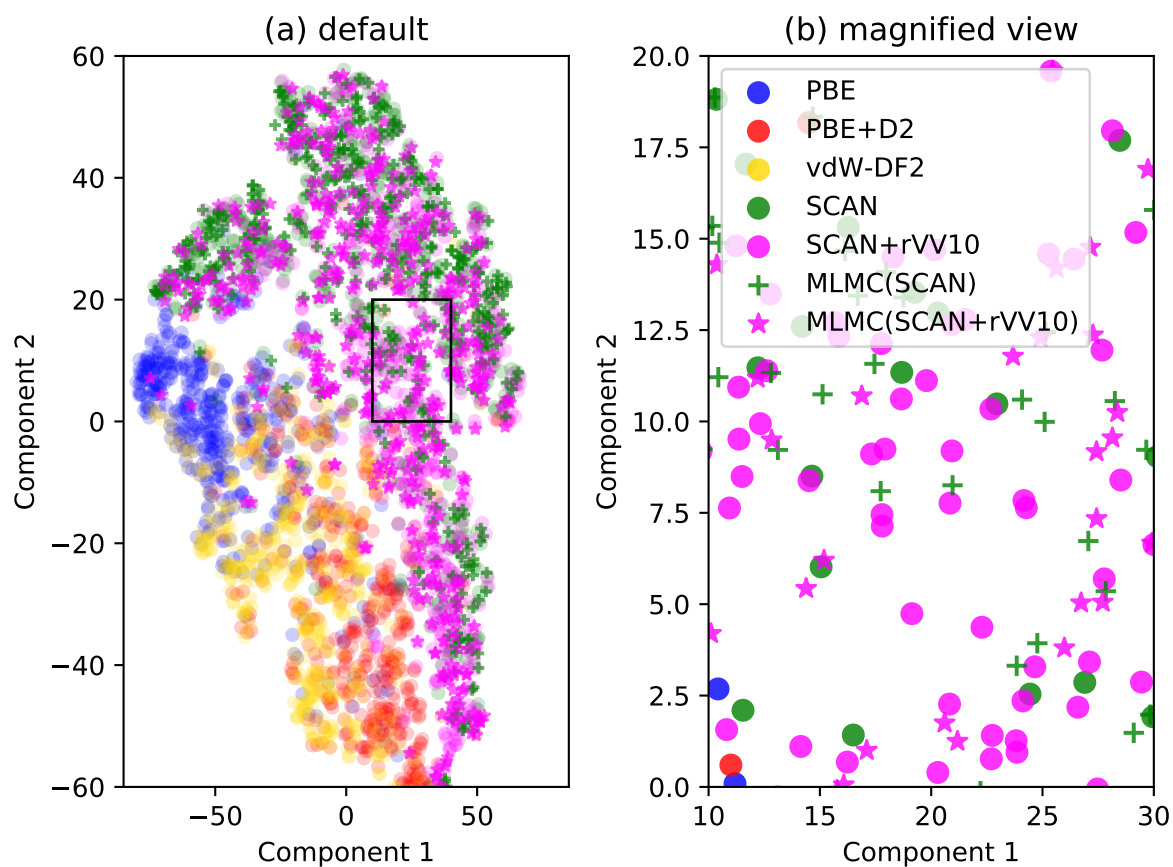


Figure 17: t-SNE representation of the configurational spaces spanned by the different functionals for the $\text{CH}_4 @ \text{HChab}$. This visualization is based on 500 selected configurations from each molecular dynamics trajectory and from MLMC trajectories of the SCAN (MLMC(SCAN)) and SCAN+rVV10 (MLMC(SCAN+rVV10)) functionals using PBE as production method.

We have shown insofar that the MLMC can significantly improve the sampling of the geometries at the target level. The MLMC also clearly reduces error in computed internal energies of individual systems, as shown in Table 5. For instance, the deviation in the internal energy of the $\text{CH}_4 @ \text{HChab}$ at the SCAN level was reduced from -4.38 kcal/mol from MLPT with PBE production method to about -0.64 kcal/mol. Similarly, the deviations were also reduced for the predicted SCAN and SCAN+rVV10 enthalpies of adsorptions, see Table 6. For $\text{CH}_4 @ \text{HChab}$ enthalpy, for instance, the MLPT presented a deviation of -8.58 kcal/mol which reduced to only 0.12 kcal/mol when the MLMC resampling was employed.

Once again, the greatest improvements have been achieved for the adsorbate with substrate systems, followed by the clean zeolites, while the effect of MLPT or MLMC is negligible for the isolated molecular systems for which most of the observed deviations are close to zero and probably are related to numerical errors. In Tables 5-6, all the predictions of SCAN/SCAN+rVV10 energies from PBE were significantly improved by the MLMC method. The general improvement reduced the relative internal energy deviations below 1.0 kcal/mol.

Table 5: Deviations of the SCAN and SCAN+rVV10 internal energies predicted using MLPT and MLMC from their respective reference values obtained in straightforward MD simulations (Table 1). All values are expressed in kcal/mol.

System	Prediction	MLPT	MLMC
CH ₄ @HChab	SCAN	-4.380	-0.639
	SCAN+rVV10	-5.026	-0.858
CO ₂ @SiChab	SCAN	2.249	-0.564
	SCAN+rVV10	3.395	0.961
HChab	SCAN	4.142	-0.749
	SCAN+rVV10	3.830	-0.230
SiChab	SCAN	0.659	-0.345
	SCAN+rVV10	0.246	-0.267
CH ₄	SCAN	0.013	-0.011
	SCAN+rVV10	0.000	-0.014
CO ₂	SCAN	0.054	-0.005
	SCAN+rVV10	0.013	-0.058

Table 6: Experimental enthalpies of adsorption of CH₄@HChab and CO₂@SiChab and their calculated reference MLPT, and MLMC values. The deviations between the predictions and the reference MD values are given in parentheses. All values are expressed in kcal/mol.

System	Prediction	Ref.	MLPT	MLMC
CH ₄ @HChab	SCAN	-2.60	-11.13 ^(-8.54)	-2.48 ^(0.12)
Exp. -4.06 [148]	SCAN+rVV10	-4.77	-13.63 ^(-8.86)	-5.38 ^(-0.61)
CO ₂ @SiChab	SCAN	-2.95	-1.42 ^(1.54)	-3.17 ^(-0.21)
Exp. -5.02 [126]	SCAN+rVV10	-7.11	-3.97 ^(3.14)	-5.82 ^(1.29)

4.3.3 *Assessing the accuracy of MLPT for the random phase approximation*

In the previous work referenced earlier, (ref. [28]), MLPT has been applied to compute enthalpies of adsorption at the RPA level of theory. In addition to the two systems studied in this work, also $\text{CH}_4 @ \text{CHAB}$ and $\text{CO}_2 @ \text{HChab}$ were considered in this previous paper. The RPA enthalpies of adsorption showed an excellent agreement with experiment and for the molecules adsorbed in HChab zeolite a significant improvement with respect to the production method (PBE+D2) was found. Because of the high computational cost involved in the RPA calculations, in this case it is not possible to produce reference trajectories and use them to repeat some of the previous analysis (indeed, MLPT is meant to make finite-temperature calculations involving computationally expensive approximations possible).

However, the I_w index can be straightforwardly obtained and we present its values in Table 7. This type of analysis, which we did not develop in the original paper, ref. [28], shows that a good performance of MLPT should be expected, as the I_w coefficient takes values above 0.1 for all systems. While not strictly necessary, the accuracy can be further confirmed by applying MLMC, which requires several additional production level calculations but avoids completely the highly expensive RPA calculations. To this purpose, we consider all the systems but $\text{CO}_2 @ \text{SiChab}$, which is the adsorbed system with the highest I_w coefficient. (the two adsorbed systems with the lowest I_w and the two zeolites). With respect to the MLPT values reported in ref. [28], the MLMC results present only small deviations: from -0.05 kcal/mol to 0.66 kcal/mol for the studied systems (see Table 7). Keeping in mind that also statistical errors should be considered (the standard error in the MC sampling is about 0.3 kcal/mol), we conclude that the agreement between MLPT and MLMC results is very good from which we deduce that PBE+D2 is a suitable production method for the RPA target level calculations and that the Monte Carlo resampling of the MLPT results is not strictly necessary.

Table 7: Values of the I_w index corresponding to the MLPT predictions of RPA energies from PBE+D2 trajectories for the different systems considered in ref. [28], together with the MLPT and MLMC total electronic energies, and their difference. Energies are in kcal/mol.

System	I_w	MLPT	MLMC	Difference
CH ₄ @HChab	0.13	-21398.09	-21398.65	-0.57
CO ₂ @HChab	0.21	-21945.80	-21945.97	-0.17
CH ₄ @SiChab	0.26	-21371.82	-21371.74	-0.08
CO ₂ @SiChab	0.27	-21916.60	-	-
HChab	0.28	-20214.40	-20215.06	-0.66
SiChab	0.32	-20188.00	-20188.05	-0.05

4.4 CONCLUSIONS

We assessed the accuracy of the machine learning thermodynamic perturbation theory approach by computing enthalpies of adsorption of molecules in zeolites. By considering a set of five DFT functionals with different characteristics, MLPT estimates are compared to reference results produced from full MD simulations. The largest deviations are found when the configurational space accessible to the target level has a small overlap with that sampled by the production method. Even when a reference result is not available, the most problematic cases can be detected by using a diagnostic tool, the I_w index. In these cases the results can be significantly improved by coupling the Monte Carlo approach to the MLPT model to resample the configurational space at the target level of theory which is, however, performed at the cost of the production method.

Finally, the analysis is extended to some recently published MLPT results that used the RPA as target approximation [28]; in this case the high computational cost of the RPA makes the generation of reference ensemble energies and enthalpies completely unpractical. The application of the I_w diagnostic test and of the MC resampling highlights the full reliability of those previous results, which were based on PBE+D2 production MD.

The main advantage of the MLPT approach stands in the possibility of evaluating finite temperature properties at a certain target level of theory by performing only a minimal number of single-point calculations. Additionally, MLPT is also suitable for approximations where the computation of gradients is not available or adds a significant computational cost [159]. This approach opens the possibility of systematically applying high-accuracy/high-cost *ab initio* methodologies to achieve a new level of predictive power in materials simulations.

COUPLED CLUSTER FINITE TEMPERATURE SIMULATIONS
OF PERIODIC MATERIALS VIA MACHINE LEARNING

5.1	Introduction	85
5.2	Methods	87
5.2.1	Periodic boundary conditions coupled cluster	87
5.2.2	Machine Learning Thermodynamic Perturbation Theory and Monte Carlo resampling	88
5.3	Results and discussion	90
5.4	Conclusions	94

The content of this chapter is nearly identical to the following Chemrxiv preprint: “Coupled cluster finite temperature simulations of periodic materials via machine learning”, by Basile Herzog, Alejandro Gallo, Felix Hummel, Michael Badawi, Tomas Bucko, Sébastien Lebègue, Andreas Grüneis and Dario Rocca, to be published.

5.1 INTRODUCTION

The last chapter was devoted to a methodological assessment of the Machine Learning Perturbation Theory. In this chapter, we will be interested in applying those methods to provide correlated quantum chemical wave function finite temperature results. The later, based on post-Hartree-Fock (post-HF) approximations are indeed systematically improvable and could potentially overcome some of the limitations of DFT for materials simulations. Among those, second-order Møller-Plesset perturbation theory (MP2) [137] and coupled cluster theory [7], have been recently implemented for periodic materials [20, 45, 49, 127, 151]. 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.

While machine learning (ML) is nowadays a well established tool in the context of MD simulations, using notably Neural Network Potentials [15, 33, 34, 194], the ML-accelerated MD typically requires large amounts of data and becomes rapidly

challenging for the more expensive approximations. In this chapter, we show how finite temperature observables for periodic materials can be evaluated using the ‘gold standard’ coupled cluster ansatz, including single, double, and perturbative triple particle-hole excitation operators (CCSD(T)) in combination with Machine Learning Perturbation Theory and Monte Carlo sampling.

For molecular systems, the application of ML techniques has already been proven to be effective in enhancing the efficiency of CCSD(T) MD simulations [18, 34, 172, 176, 179]. Very recently, applications to molecular condensed phase systems, specifically to liquid water, have also been considered. In ref. [40] the ML model for periodic water was trained with data produced for finite water clusters using near-linear scaling coupled cluster theory. In ref. [29], CCSD(T) calculations were restricted to very small periodic models based on a box of 16 H₂O 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, this work represents the first application of CCSD(T) to finite temperature simulations of periodic solid materials. This 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 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. (138)).

This work is based on the combination of the Machine Learning Perturbation Theory (MLPT), together with an efficient periodic coupled cluster theory implementation. This implementation is based on a plane-wave basis set and novel finite size and basis set correction techniques that accelerate the convergence to the complete-basis-set limit and thermodynamic limit significantly [71, 96]. 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 [120, 121, 173, 188].

As a specific application for 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 [66, 98, 191]. In this field, more quantitative and systematically improvable theoretical predictions are instrumental to interpret experimental findings and predict new materials. Although the calculations presented in this work are still significantly more expensive than those based on standard density functional theory, this proof-of-principle work paves the way to a more systematic use of highly accurate post-HF methods in materials simulations.

5.2 METHODS

5.2.1 *Periodic boundary conditions coupled cluster*

The coupled cluster theory calculations are performed using the Cc4s code [27], which is interfaced to the Vienna *ab initio* simulation package (VASP) [114, 116]. The calculations are performed in several steps involving Hartree–Fock and MP2 theory to obtain corresponding energies and optimized approximate natural orbitals [72]. Once the natural orbitals have been computed, the Cc4s interface to VASP is used to compute intermediate quantities [92] that are needed for the subsequent coupled cluster energy calculations including the corresponding finite size [71] and basis set corrections [96]. In ref. [173], 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. 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.

The convergence of the CCSD and (T) correlation energy contributions to the molecular adsorption energy was tested on a single configuration. In particular, the following difference has been considered: $\Delta E^{\text{corr.}} = E_{\text{CO}_2\text{-HChab}}^{\text{corr.}} - E_{\text{CO}_2}^{\text{corr.}} - E_{\text{HChab}}^{\text{corr.}}$, where $E_{\text{CO}_2\text{-HChab}}^{\text{corr.}}$, $E_{\text{CO}_2}^{\text{corr.}}$, and $E_{\text{HChab}}^{\text{corr.}}$ are, respectively, the correlation energies of the chabazite containing the CO₂ molecule, the isolated CO₂ molecule and the pristine chabazite. Table 8 lists the $\Delta E^{\text{corr.}}$ values obtained for different basis set sizes and methods. In particular, the CCSD results with (CCSD-FP) and without (CCSD) a focal point finite basis set error correction and the perturbative triples contributions with (T*) and without (T) an approximate finite basis set error correction have been compared. We stress that CCSD-FP and (T*) have been thoroughly tested in ref. [96]. As discussed in ref. [96], 10 virtual natural orbitals per occupied state suffice in combination with the focal point basis set correction to achieve chemical accuracy for a large number of reaction energies on the level of CCSD and (T). Although we can not fully converge our binding energy estimates for the present system, we find satisfactory convergence of the correlation energy contributions with respect to the number of virtual orbitals. From the second-largest to the largest basis set size the CCSD-FP and (T*) estimates change by 0.5 kcal/mol and 0.4 kcal/mol only. From this we conclude that the remaining basis set error when using $N_v/N_o = 10$ for CCSD-FP and $N_v/N_o = 5$ for (T*) will not exceed 1 kcal/mol.

N_v/N_o	CCSD	CCSD-FP	(T)	(T*)
5	2.4	-6.9	-0.5	-1.0
10	-4.9	-7.9	-1.1	-1.4
15	-7.0	-8.4	N/A	N/A

Table 8: Correlation energy contributions to the adsorption energy (kcal/mol) computed using different levels of theory and basis sets. N_v and N_o correspond to the number of virtual and occupied orbitals, respectively. CCSD refers to the CCSD correlation energy, whereas CCSD-FP includes a focal point finite basis set error correction, as thoroughly tested in ref. [96]. (T) and (T*) correlation energy estimates are without and with an approximate finite basis set error correction, respectively.

5.2.2 Machine Learning Thermodynamic Perturbation Theory and Monte Carlo resampling

In this work we consider the calculation of the enthalpy of adsorption of carbon dioxide in a porous zeolitic material (see Section 5.3). In practice this quantity is computed as

$$\begin{aligned}\Delta_{\text{ads}}H(\text{M@Z}) &= \Delta_{\text{ads}}U(\text{M@Z}) + \Delta_{\text{ads}}(pV)(\text{M@Z}) \\ &= \langle E(\text{M@Z}) \rangle - (\langle E(\text{M}) \rangle + \langle E(\text{Z}) \rangle) - k_B T,\end{aligned}\quad (172)$$

where $\Delta_{\text{ads}}U$ is the internal energy of adsorption, $\langle E(i) \rangle$ denotes the ensemble average of 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 $\Delta_{\text{ads}}(pV)(\text{M@Z}) = -k_B T$ is obtained by assuming an ideal gas behavior of M and a negligible change of pV of the zeolite due to adsorption. The 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 unpractical at these levels of theory. Starting from an AIMD trajectory obtained using numerically affordable semi-local DFT with empirical van der Waals corrections (PBE+D2) [23, 69], the post-HF ensemble energies are estimated using the MLPT approach trained on a small number of single point calculations. This approach was described in Chapter 4 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 can be obtained from thermodynamic perturbation theory by reweighting:

$$\langle E_1 \rangle_1 = \frac{\sum_{i=1}^M E_1(\mathbf{R}_i) \exp(-\beta \Delta E(\mathbf{R}_i))}{\sum_{i=1}^M \exp(-\beta \Delta E(\mathbf{R}_i))}, \quad (173)$$

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 production trajectory.

2. While the application of eq. (173) 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 [10, 44] and Δ -ML [158]. $E_0(\mathbf{R})$ is known and the evaluation of eq. (173) requires only the energy difference $\Delta E(\mathbf{R})$. 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.

In order to correct for the basis set incompleteness, the (T^*) correlation energy was computed for 25 configurations of each system. The standard deviation of the difference between the (T^*) and (T) correlation energies across the 25 configurations of the M@Z system was found to be very small (within 0.19 kcal/mol). Hence, to a very good accuracy, (T^*) can be obtained from (T) by a uniform shift of contribution of each configuration by the average difference between the (T^*) and (T) energies obtained using 25 configurations. We applied this procedure to obtain our final CCSD(T) MLPT result. To validate this approach, we computed the MLPT enthalpy result using only those 25 configurations as training set, both with the (T^*) energy and with the (T) energy plus the uniform shift, and found the two results to agree within 0.02 kcal/mol.

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 Chapter 4, uses Metropolis MC [136] 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 subsequently evaluated at the post-HF level 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.

5.3 RESULTS AND DISCUSSION

In this section we 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 [178, 185]. The primitive cell of the model considered here is shown in Figure 18.

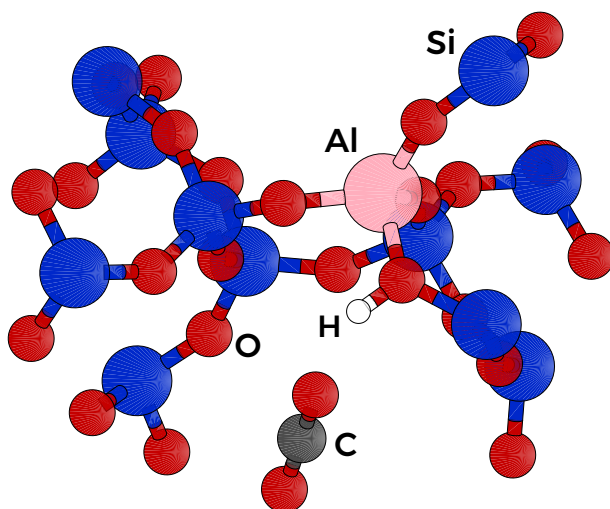


Figure 18: The unit cell of the system studied in this work, CO_2 in protonated chabazite.

The experimental value of the enthalpy of adsorption of CO_2 in HChab, -8.41 kcal/mol [147], 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. [147] and we cannot exactly quantify the uncertainty in the experimental reference.

The computed results are presented in Table 9, where the error bars related to the finite sampling and the ML model are also indicated [28]. The molecular dynamics at PBE+D2 level leads to an estimate for the adsorption energy which is more than 1 kcal/mol 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 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, it was demonstrated that also the RPA provides accurate enthalpies of adsorption of molecules in zeolites [28]. Specifically, the value for CO₂ in protonated chabazite is -8.01 kcal/mol. Although the RPA has a diagrammatic structure it is not as straightforward to systematically improve its accuracy as for post-HF methods [12, 48, 57, 73, 81, 91, 139, 162]. In practice, the RPA often provides more realistic results starting from a DFT approximation rather than from HF [162], and this starting point dependence makes this approximation less reliable as a general predictive method.

Target Method	Sampling Method	Enthalpy (kcal/mol)
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 [147]	adsorption isotherms	-8.41

Table 9: Enthalpy of adsorption of CO₂ in protonated chabazite (kcal/mol) computed using different target and sampling methods.

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 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 Chapter 4 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. (173)); in case of a poor overlap only 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 Chapter 4. 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. The reweighting of the trajectories at the MP2 level provides large values for I_w (> 0.15) and the corresponding enthalpies in Table 9 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 with 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, 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 [15, 33, 34, 194], this MLMC approach avoids training on atomic forces, which are not readily available in the current periodic CCSD(T) implementation and that 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 9, differ by only 0.2 kcal/mol from the MLPT value. In the MLMC case the error bar is however sizeably larger because of the long auto-correlation 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 [124]. As shown in Figure 19, 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 Chapter 4, 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, Figure 20 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.

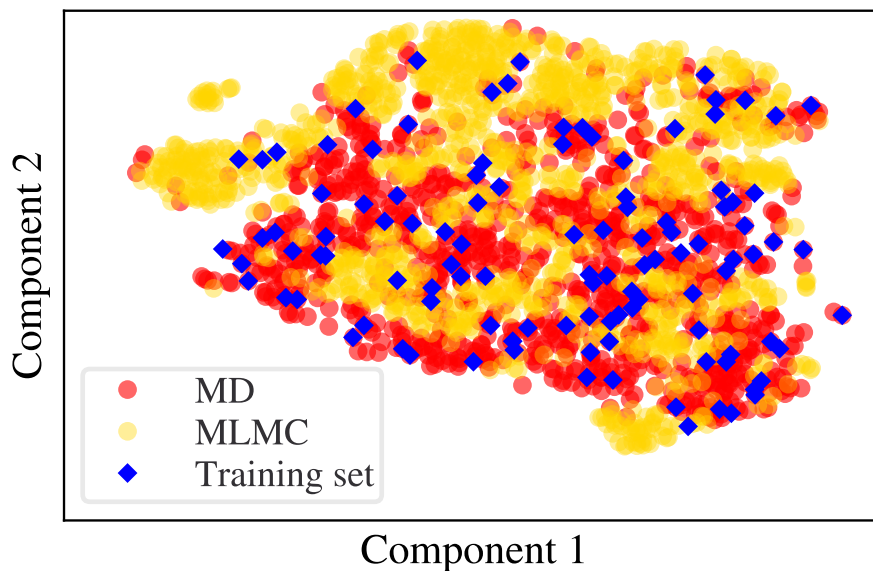


Figure 19: t-SNE representation of the configurational spaces spanned by the PBE+D2 molecular dynamics (MD) trajectory and the CCSD(T) machine learning Monte Carlo (MLMC) trajectory. The configurations included in the training set are also shown to demonstrate that they cover essentially whole relevant part of the configurational space sampled at the CCSD(T) target level. The axes represent the two components of the t-SNE projection.

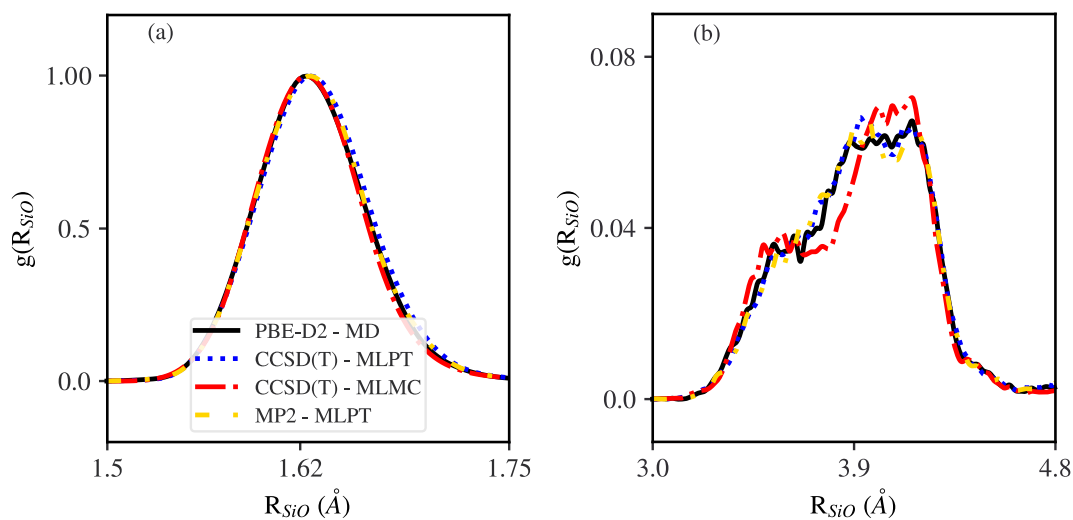


Figure 20: First (a) and second (b) series of peaks of the partial radial distribution function for the Si-O pairs determined at different levels of theory.

5.4 CONCLUSIONS

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 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 modelling of catalytic reactions.

GENERATIVE MACHINE LEARNING TO SOLVE THE SCHRÖDINGER EQUATION IN THE CONFIGURATION INTERACTION SPACE

6.1	Introduction	95
6.2	Methodological approach	100
6.2.1	Configuration interaction approach	100
6.2.2	Generative model	100
6.2.3	Direct generation of single and double excitations	103
6.2.4	Computational complexity	106
6.2.5	Computational details	108
6.3	Numerical Results	108
6.4	Conclusions	111

The content of this chapter is nearly identical to the following published work: [83]
 “Solving the Schrödinger Equation in the Configuration Space with Generative Machine Learning”, by Basile Herzog, Bastien Casier, Sébastien Lebègue, and Dario Rocca, published in the Journal of Chemical Theory and Computation 2023 19 (9), 2484-2490

6.1 INTRODUCTION

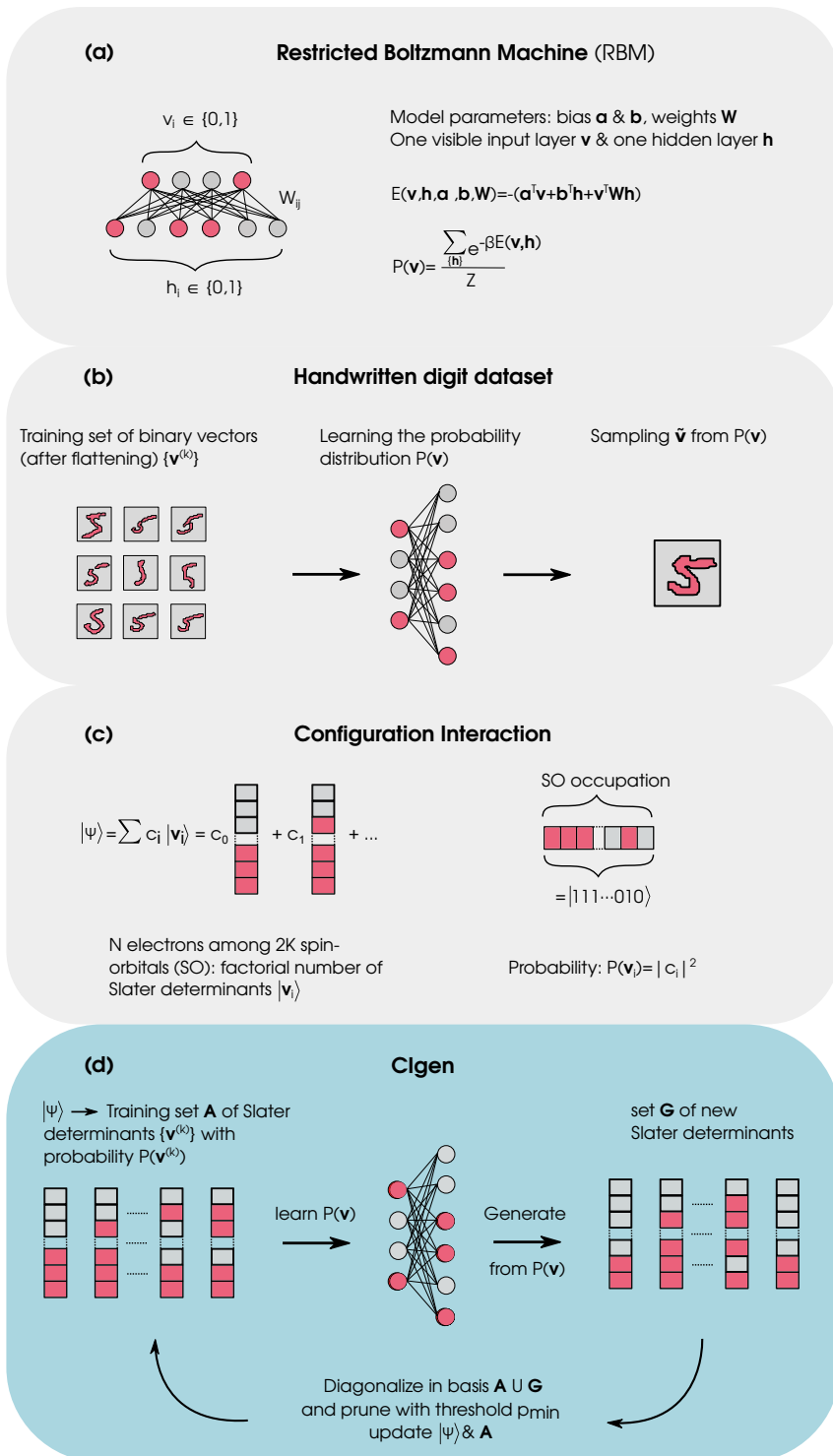
The application of machine learning (ML) to quantum chemistry and computational materials science has experienced an impressive growth in the past few years. However, methods that apply ML to molecular property predictions [122, 171] or molecular dynamics [15, 189] usually imply the availability of a certain amount of data previously produced by approximating the solution of the Schrödinger equation, as seen for instance in the two previous chapters. By considering the exponential numerical complexity involved in the exact solution of this equation it would be highly desirable to take advantage of ML techniques also in this context. For realistic molecules and materials this is a widely open field of research and the full potential of ML has yet to be widely exploited in its full potential.

Following earlier work on quantum neural networks [177], in 2017 Carleo and Troyer showed that neural networks can effectively represent the quantum states of spin models, thus reducing the exponential complexity of the many-body problem

[25]. Specifically, they showed that a restricted Boltzmann machine (RBM) used as an ansatz within variational Monte Carlo can achieve variational energies lower than those obtained with traditional approaches. This wavefunction ansatz and similar variants based on different architectures are generally referred to as neural-network quantum states (NQS). From a theoretical point of view this success relies on universal approximation theorems [90, 119], which imply that neural networks can approximate complex (but “reasonably smooth”) high-dimensional functions, including quantum states. This method was demonstrated numerically considering the one- and two-dimensional Ising and Heisenberg models. More recently Carleo and coworkers have applied a similar approach to realistic Hamiltonians of small molecules [35]. This was achieved by mapping the electronic structure Hamiltonian into a spin-like Hamiltonian by using quantum information encodings. The most accurate results were obtained with the Jordan-Wigner mapping [101], which leads to an approach equivalent to the configuration interaction (CI) [184]. Within the CI method the fully interacting wavefunction is expressed as a linear combination of excited Slater determinants (the “configurations”); while the coefficients of this expansion are typically computed as solution of an eigenvalue problem, in ref. [35] they were learnt by the RBM in an unsupervised way using a Monte Carlo sampling and the variational principle. While this approach was achieving chemical accuracy for small basis sets, the Monte Carlo sampling was repeatedly drawing the most dominant state (i.e. the Hartree-Fock determinant) and this was at the origin of a significant slow down of convergence for basis sets beyond the minimal STO-3G. It was recently shown that this issue can be alleviated by using autoregressive neural networks, that allowed calculations with up to 30 spin orbitals [6]. Our numerical results presented below show that a simple RBM architecture used as a generative model can easily double this number, performing calculations with up to 56 spin-orbitals. More in general it should be noticed that the representation of the wavefunction in the configuration space is strongly non-smooth (determinants with similar occupations can provide significantly different contributions to the wavefunction) and this could be challenging for ML regression approaches. To overcome this issue alternative methods use instead the high expressive power of deep neural networks to represent electronic structure wavefunctions in real-space [76, 82, 105, 146]; as an advantage some of these techniques can achieve the complete basis set limit in a rather straightforward way but on the other side they require deep architectures involving the optimization of a large number of parameters and special care to keep into account the antisymmetry of the electronic wavefunction. In a recent work this type of approaches has also been extended to efficiently describe the wavefunction at different molecular geometries: Most of the weights in the neural network are shared among different different structures and only a small percentage of them is actually reoptimized [174].

The antisymmetry is instead naturally included in the CI space which, however, grows unfavorably with the system size. This has led to the development of a series of methods that select the excited determinants that contribute the most to the wavefunction, either based on perturbation theory [95] or Monte Carlo sampling [67]. More recently machine learning techniques have been coupled to these methods [36, 100, 149]. In this context Coe was the first to propose a machine learning approach that explores the configuration space but, instead of explicitly computing the coefficient of each configuration, this is predicted by a regression neural network [36]. While Carleo and coworkers apply a neural network approach to “exactly” fit the wavefunction in the configuration space using the variational principle [35], the alternative approaches of refs. [36, 100, 149] are more qualitative and use supervised learning techniques.

Figure 21: Restricted Boltzmann machine (RBM) as a generative model and its application to the solution of the Schrödinger equation.



(a) RBM architecture consisting of one visible input layer and one hidden layer of binary values (differently from standard neural networks a proper output layer is not present); for a given configuration (\mathbf{v}, \mathbf{h}) the parameters $(\mathbf{a}, \mathbf{b}, \mathbf{W})$ are used to define an energy function E and an associated Boltzmann-like probability density P . (b) As an example, the RBM can be trained on a set of handwritten digits and afterward used to generate new realistic ones; to this purpose the digit's images are flattened to become unidimensional binary vectors $\mathbf{v}^{(k)}$ where 1 and 0 correspond to the digit and background pixels, respectively. (c) The configuration interaction (CI) approach expands the wavefunction of a molecule as a linear combination of excited Slater determinants, which can be represented as a sort of unidimensional binary image. (d) The Clgen algorithm presented in this work iteratively trains an RBM on the distribution of determinants in the current approximation of the wavefunction and subsequently uses it to expand it by generating new important contributions.

In this work we propose an alternative approach named CIgen that uses generative machine learning to directly generate the configurations that contribute the most to the electronic wavefunction. This is done with the general aim of avoiding the exploration of the overwhelmingly large configuration space to search for the most important determinants. In this respect our approach lies between the selective CI and NQS approaches: Similarly to selective CI, the determinants with the most important contributions are selected and used in an exact diagonalization procedure; this selection procedure is based on a generative NQS.

Notable examples of generative machine learning algorithms include the restricted Boltzmann machines [88, 134], the variational autoencoders [107], and the generative adversarial networks (GANs) [65]. CIgen is based on RBMs, a type of generative neural network whose architecture is shown in Figure 21(a). As for other generative models the RBM can learn in an unsupervised way the statistical distribution behind a series of input objects and then be used to generate new ones. A qualitative example is shown in Figure 21(b): An RBM can be trained with a series of pictures, e.g. of handwritten digits, in practice represented as a flattened matrix of pixels and afterward used to generate new realistic images of the same digit.

For a given molecule or material the excited Slater determinants are associated to an underlying probability distribution that could only be exactly evaluated by solving the Schrödinger equation to obtain the wavefunction. By representing the space of excited determinants simply as binary vectors (a sort of onedimensional binary image), the RBM is here iteratively trained using data from the current approximation of the wavefunction and subsequently used to generate new important determinants to improve this approximation (see Figure 21(c-d)).

Within the previous approach of Carleo and coworkers [25, 35] the RBM was not used as a generative model but rather as a regression model for the wavefunction (to this purpose the RBM architecture was generalized to include complex weights); here we use instead the RBM as a model to qualitatively represent the probability distribution associated with the wavefunction, which is afterward used to generate the most likely configurations. In this respect the CIgen approach has some similarities with the use of RBMs to learn the statistical distribution of data from experimental measurements of quantum states and to subsequently generate new configurations to compute quantum averages [134, 186]. Differently from selective CI or Monte Carlo CI our approach does not “explore” the huge determinant space to find significant contributions but rather directly generate them with the current approximation of the probability distribution associated with the RBM. Comparison with previous approaches will be further discussed below.

6.2 METHODOLOGICAL APPROACH

6.2.1 Configuration interaction approach

Considering the non-relativistic N electron problem in the Born-Oppenheimer approximation, the configuration interaction method solves the Schrödinger equation for a fixed basis set by expanding the wavefunction in the following way [184]:

$$|\Psi\rangle = c_0|\Phi_0\rangle + \sum_{ra} c_a^r|\Phi_a^r\rangle + \sum_{\substack{a<b \\ r<s}} c_{ab}^{rs}|\Phi_{ab}^{rs}\rangle + \sum_{\substack{a<b<c \\ r<s<t}} c_{abc}^{rst}|\Phi_{abc}^{rst}\rangle + \dots \quad (174)$$

where $|\Phi_0\rangle$ is the Hartree-Fock ground state, $|\Phi_a^r\rangle$ is a singly excited Slater determinant from occupied spin orbital a to unoccupied spin orbital r , and all the other terms correspond to multiple excitations; while $|\Psi\rangle$ could denote any excited state of a given system here the discussion will be focused on the ground-state. When all the possible excited determinants are included in eq. (174) the approach is called full configuration interaction (FCI) and the solution of the Schrödinger equation becomes exact for a given basis set. Within the “brute-force” FCI approach the coefficients of the expansion c_0, c_a^r, \dots are determined by computing the expectation value of the Hamiltonian $\langle\Psi|\hat{H}|\Psi\rangle$ and applying the variational principle. Since the total number of determinants grows as $\binom{2K}{N}$, where $2K$ is the number of spin-orbitals, the FCI approach becomes quickly unpractical for most applications. It is however well known that often the wavefunction can be accurately represented by a limited number of excited determinants and here a ML model is proposed to directly generate the important contributions.

6.2.2 Generative model

In this work, it is shown that restricted Boltzmann machines (RBM) can be used to efficiently generate the excited determinants that contribute the most to the wavefunction in eq. (174). The RBM model was defined in Chapter 3 and we refer the reader to this section for more details. For the purpose of this work the input vectors \mathbf{v} of the visible layer represent the Slater determinants and their associated probability $P(\mathbf{v})$ should ideally be proportional to their contribution to the wavefunction (the square of the c coefficients in eq. (174)). The determinants are represented as binary vectors, where 1 denotes occupied and 0 unoccupied states, with size equal to the number of spin-orbitals (see Figure 21). The architecture of the RBM used in this work is rather simple, with $2K$ (the number of spin orbitals) neurons in the visible and in the hidden layers. Given a certain fixed subspace of determinants, the Hamiltonian is diagonalized to obtain the wavefunction coeffi-

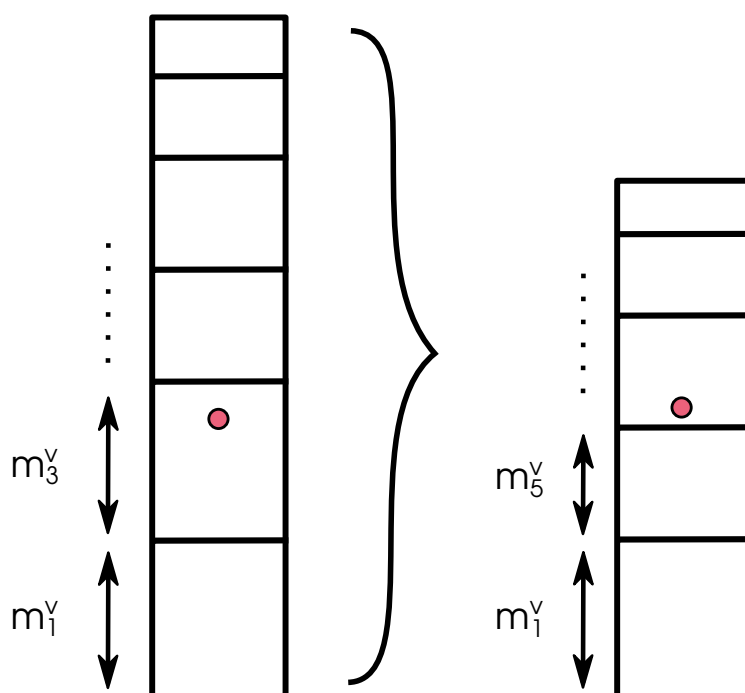


Figure 22: Pictorial representation of the tower sampling algorithm, in which the probability of sampling the i -th orbital is proportional to the height m_i^v of each rectangle. Left: Considering at first the spin down states (odd indexes), a random number in the interval between 0 and $\sum_{\text{odd } i} m_i^v$ is drawn (depicted as a red dot) resulting in the population of the corresponding spin-orbital. Right: In a second iteration the occupation probability m_3^v of the third state is now set to zero to prevent a new sampling of this already occupied state and the procedure is repeated, resulting in the population of a new spin-orbital. The procedure is repeated until $N/2$ spin-orbitals for each spin are occupied.

cients; the model is then trained with random configurations generated according to the corresponding probability distribution by updating the RBM parameters with the contrastive divergence algorithm [87]. The computational parameters are fixed to $M = 50000$ for the training set size, $N_b = 10$ for the minibatch size, $L = 20$ for the number of Gibbs sampling iterations, and $\epsilon = 0.01$ as learning rate (see Chapter 3). In Chapter 3 was given the Gibbs sampling algorithm as applied to the RBM. For our specific application we have a constraint for the values of v_i , whose sum has to be equal to the number of electrons $\sum_i v_i = N$. Since we exclusively consider closed-shell systems, we have the additional constraint that the number of spin up and down electrons must be equal. While we could generate vectors without restrictions and reject them *a posteriori*, it is significantly more efficient to intrinsically include this constraint in the procedure. This is achieved by replacing the step 3 of the Gibbs sampling with an approach based on the tower sampling

algorithm [110]. Within this approach the N states are randomly populated with a probability proportional to $m_i^y = p(v_i = 1 | \mathbf{h})$ (this quantity can indeed be interpreted as the occupation probability of each visible neuron). The basic idea of the application of the tower sampling algorithm to our problem is depicted in Figure 22. It is important to mention that in our approach we represent Slater determinants with $2K$ component binary vectors whose *even* and *odd* indexes label spin *up* and *down*, respectively. The Gibbs sampling approach is used both in the training of the model and in the generative phase. While $L = 20$ iterations are used in the training part, the generative process, which expands the configuration space, uses only one step $L = 1$. This (underconverged) choice introduces some “noise” in the model that allows to more easily sample determinants that are not already in the training set. As also shown by numerical experiments that we performed, a similar behavior can be obtained by lowering the number of iterations in the training or by increasing the temperature of the RBM during the generative procedure. However, in the former case it is harder to determine the reliability of the model and in the latter the generative procedure loses efficiency as larger values of L are used. The generation of new determinants and the retraining of the model are performed iteratively. Specifically, the CIgen algorithm performs the following steps:

1. Start: The configuration interaction singles and doubles (CISD) is used to generate an initial guess wavefunction.
2. Pruning: The determinants whose squared wavefunction coefficients (eq. (174)) are below a certain threshold p_{\min} are pruned.
3. Training: The RBM model is trained only on the non pruned determinants but the Hartree-Fock (HF) determinant, which is typically associated to a very high probability, is not included in the training set (otherwise the ML would mainly generate the Hartree-Fock state in the following step). While discarded at the training stage, the HF determinant is included by default in the diagonalization step below.
4. Generation: The trained RBM is used to generate a set of new determinants, whose number is proportional to the number of determinants already included in the current approximation of the wavefunction. During this procedure determinants are automatically discarded if they are already included in the wavefunction, if they do not couple through single or double substitutions with the current determinant set, and if they do not have correct spin and point group symmetries.
5. Diagonalization: The Hamiltonian is diagonalized in the subspace that includes the newly generated determinants and new coefficients are obtained for eq. (174)

6. Iterate: The procedure is repeated from step 2 until convergence of the energy is achieved.

When sampling from the RBM model, the conditional probabilities $p(v_i = 1|\mathbf{h})$ and $p(h_i = 1|\mathbf{v})$ are sigmoid functions, as one can deduce from eqs. (154-155) For example, for the visible cells we have

$$p(v_k|\mathbf{h}) = \frac{\exp\left(\beta v_k(\sum_j W_{kj}h_j + a_k)\right)}{\sum_{v_k} \exp\left(\beta v_k(\sum_j W_{kj}h_j + a_k)\right)} = \frac{\exp\left(\beta v_k(\sum_j W_{kj}h_j + a_k)\right)}{1 + \exp\left(\beta v_k(\sum_j W_{kj}h_j + a_k)\right)} \quad (175)$$

and

$$p(v_i = 1|\mathbf{h}) = \frac{1}{1 + \exp(-\beta((\mathbf{W}\mathbf{h})_k - a_k))}. \quad (176)$$

One can easily see that rising the temperature will increase noise in the sampling and for $T \rightarrow \infty$ all probabilities of occupation will be equal to one half. In this limit the specific parameters obtained training the model become irrelevant and the RBM generates new configurations completely randomly (with uniform probability). This trend as a function of the temperature can be seen in Figure 23, that shows the performance of CIgen to predict the total energy of H₂O for three different temperatures. In our problem, once the energy is close to convergence and there are few determinants left to find, we will repeatedly sample the same determinants that are already in the wavefunction, which slows down the procedure. To reduce this computational burden and to obtain more out of equilibrium samples, we add a temperature increment in the generative procedure in the following form: at a given iteration, with N_d the number of determinants after pruning, we try to generate $f_{gen} \cdot N_d$ new determinants. Let k count each attempt to add a new determinant. The temperature in the sampling is fixed to be $T = T + \lfloor k/(50 \cdot f_{gen} \cdot N_d) \rfloor$. This number was fixed by numerical experiments.

Differently from previous approaches based on RBMs for quantum states [25, 35], CIgen does not require an RBM architecture generalized to include complex weights and, while the model still learns in an unsupervised way, the procedure is not based on the variational principle. It can be further noted that if traditional NQS aim at fitting the wave function $|\Psi\rangle$, our approach is instead a representation of the squared amplitude $|\Psi|^2$. While the training of a ML model based on the variational principle is more elegant and physically motivated [25, 35], the CIgen approach requires a less strict training as it does not aim to exactly fit the wavefunction.

6.2.3 Direct generation of single and double excitations

The CIgen algorithm is based on a rather straightforward way of using a generative neural network model to enlarge the configuration subspace of interest.

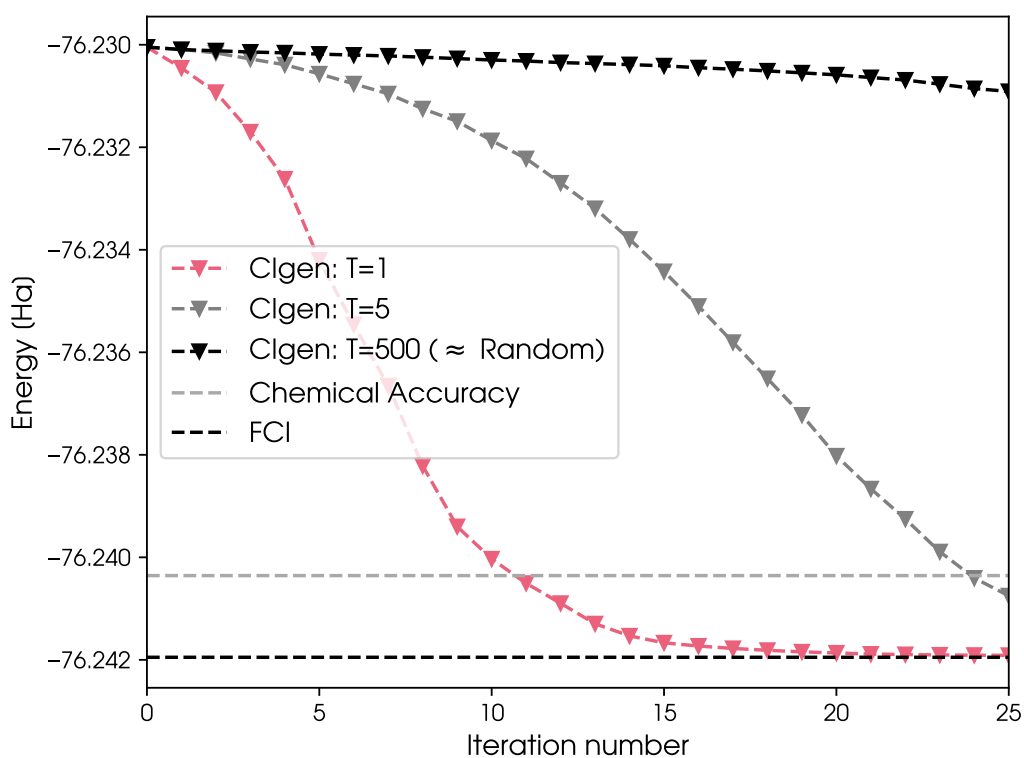


Figure 23: Effect of the temperature in the generative procedure: energy convergence for water in the cc-pVDZ basis with one frozen orbital against the number of iteration of the Clgen method, using temperatures $T = 1, 5,$ and 500 . The $T = 500$ curve corresponds to a fully uniform sampling of the connected excited determinants. The chemical accuracy limit and Full CI values are depicted as horizontal lines.

However, the generative step (step 4 of the Cigen algorithm) often samples determinants that clearly cannot contribute to the wavefunction and this makes the overall procedure less efficient. For example, it is well known that a fixed set of determinants can only be coupled through the Hamiltonian to new determinants that differ by single and double substitutions. This observation, which has some subtle implications, is exploited by the Monte Carlo Configuration Interaction (MCCI) algorithm proposed by Greer *et al.* [67]. In the MCCI algorithm, new determinants are sampled by randomly proposing single and double excitations of the determinants already included in the wavefunction set. More specifically, a determinant inside the current wave function is randomly chosen, the decision to perform a single or double substitution is chosen at random, and the substituted orbital(s) is (are) randomly chosen. At first sight this could seem a rather inefficient random procedure, equivalent to a uniform sampling among the space of connected determinants, but in practice the important (highly contributing) determinants have often multiple connections to other important determinants through single and double substitutions. Within this sampling method, among all the single and double excited determinants to a given wave function, those that are more connected through single or double excitation are more likely to be sampled, compared to a uniform sampling among them. Consider a wave function made of two determinants a and b , which is connected to four determinants d_1, d_2, d_3, d_4 . Assume that d_1, d_2, d_3 are connected to a through single or double excitation, while d_2, d_3, d_4 are connected to b . A uniform sampling will choose among the four with probability one fourth. If instead we first choose a or b with probability one half, before to randomly choose one of their excited determinants, d_1 and d_4 will be sampled with probability one sixth, and d_2, d_3 with probability one third. After sampling this way, the resulting Hamiltonian to be diagonalized will be less sparse, resulting in more precise estimations of the FCI coefficients. Hence there is less chance to discard important determinants. To take advantage of this feature of the configuration space we developed a variant of our approach named Cigen-SD, using the following algorithm to generate one determinant:

1. A determinant \bar{v} is randomly selected (with uniform probability) among those in the current wavefunction.
2. Gibbs sampling generates the probabilities of occupation of each visible neuron $m_i^v = p(v_i = 1|\mathbf{h})$. In this case, instead of directly applying the tower sampling algorithm, these probabilities are used to build a transition matrix whose elements are defined as $S_{ij} = (1 - m_i^v) \cdot m_j^v$, where i is an occupied state and j an unoccupied state of \bar{v} .
3. It is randomly decided if a single or a double substitution to \bar{v} is constructed (the two options having the same probability).

4. The specific substitution is sampled using the tower sampling algorithm (Figure 22) but, in this case, the transition probability matrix elements S_{ij} replace m_i^v .

The CIgen, CIgen-SD, and MCCI will be compared in the Numerical Results subsection.

6.2.4 Computational complexity

This section will discuss the numerical complexity of CIgen(-SD) and compare it with the Monte Carlo configuration interaction and the NQS method of Choo *et al.* [35]. We start the discussion by evaluating the cost of a single iteration. Let N_{det} be the number of retained determinants (namely whose squared coefficients are larger than p_{min}) at a given iteration of the CIgen(-SD)/MCCI methods. Similarly, in the NQS method of Choo *et al.*, N_{det} corresponds to the number of unique determinants inside the sample of configurations. In each of those methods, the cost of a single iteration scales as $O(N_{\text{det}}^2 K^4)$, where the dependence on K^4 can be improved when the sparsity of the Hamiltonian matrix elements can be exploited. Indeed, selective CI methods such as CIgen(-SD) and MCCI involve a diagonalization procedure that seeks the lowest eigenstate; by using iterative algorithms this can be achieved by a repeated product of the Hamiltonian matrix and a guess vector, which costs $O(N_{\text{det}}^2 K^4)$ operations. Similarly, the optimization procedure in the NQS-based variational quantum Monte Carlo implies the computation of the expectation value of the Hamiltonian for a given sample of configurations. This is again a $O(N_{\text{det}}^2 K^4)$ procedure (more precisely, the calculation of the local energy of each sampled configuration requires the evaluation of expectation values with all the other configurations already sampled). It should be noticed that the sampling of new determinants scales better than the diagonalization/optimization step. All of the methods considered here propose new determinants at a cost that is $O(K)$ or $O(K^2)$. Since $K \ll N_{\text{det}}$ in practice the most expensive task in this phase is to verify if the proposed determinants are already included in the current wavefunction set, that leads to a cost of $O(N_{\text{det}} \log N_{\text{det}})$. This can be explained considering that determinants can be stored as a sorted list of decimal integers corresponding to the binary occupation arrays. Using the binary search algorithm the identification of duplicates scales as $O(\log N_{\text{det}})$ and this task has to be repeated for all the newly generated determinants, whose number is itself proportional to N_{det} . Differently from the other methods, CIgen also requires a check to determine if the proposed determinants are among the single-double substituted determinants included in the current wavefunction. This task does not take advantage of the sorting of the list and, accordingly, requires a brute-force search through the list of the N_{det} determinants in the wavefunction; repeating this task for each one of the

newly proposed determinants yields a $O(N_{\text{det}}^2)$ complexity. For the sake of clarity the computational complexity involved in the sampling of new determinants are summarized in Table 10.

As shown by the numerical applications in Sec. 6.3 the MCCI sampling requires a higher number of iterations to converge than both CIgen and CIgen-SD. While we do not report NQS results in this paper in the original reference of Choo *et al.* it was shown that achieving chemical accuracy for the energy of H₂O in the 6-31G basis set could be already problematic and a number of required iterations larger than CIgen(-SD) is expected (our results for this system are discussed in Sec. 6.3). These comparisons show that CIgen(-SD) is competitive with similar methods based on the configuration space sampling and, by decreasing the number of required iterations, avoids several computationally expensive diagonalization/optimization steps, which scale as $O(N_{\text{det}}^2 K^4)$ and dominate the computational cost per iteration.

A direct comparison of the numerical complexity of CIgen(-SD) with deep neural network methods such as FermiNet [146] or PauliNet [82] is not straightforward. As a main difference those deep learning methods use a continuous space representation rather than the configuration space of CIgen. While the Monte Carlo wavefunction ansatzes of FermiNet and PauliNet still involve a linear combination of Slater determinants, their number is limited and the electronic correlation is to a large extent included by optimized orbitals and the Jastrow factor. These methodologies achieve the complete basis limit in a seamless way but require a number of parameters that is of magnitude larger than the RBM model used by CIgen.

Method	Sampling complexity	Diagonalization/optimization complexity
CIgen	$O(N_{\text{det}}^2)$	$O(N_{\text{det}}^2 K^4)$
CIgen-SD	$O(N_{\text{det}} \log N_{\text{det}})$	$O(N_{\text{det}}^2 K^4)$
MCCI	$O(N_{\text{det}} \log N_{\text{det}})$	$O(N_{\text{det}}^2 K^4)$
NQS	$O(N_{\text{det}} \log N_{\text{det}})$	$O(N_{\text{det}}^2 K^4)$

Table 10: Computational complexity of the different methods discussed in this work. This summary shows that the cost involved in the sampling of new determinants is smaller than the diagonalization/optimization part, which dominates the complexity of all these methods. Accordingly, the CIgen methods, by requiring less iterations to converge, can decrease the number of required diagonalizations and the overall computational cost (see text).

6.2.5 Computational details

The equilibrium geometries of the molecules considered in this work are optimized at the CCSD(T) level of theory in the corresponding basis set (6-31G or cc-pVDZ) and can be obtained from the Computational Chemistry Comparison and Benchmark DataBase [1]. The reference results for these molecules at the CCSD(T) and FCI level of theory (Table 11 and Figs. 24-25) have been obtained from the Molpro code [75, 108].

The electronic structure part of our implementation (specifically the diagonalization of the Hamiltonian in the space of the generated determinants) is based on the Quantum Package code [59]. For the CIgen, CIgen-SD, and MCCI methods the selection of the determinants to be included in the wavefunction in eq. (174) is based on an acceptance threshold of 10^{-12} on the squared coefficients.

6.3 NUMERICAL RESULTS

We now discuss the efficiency and accuracy of the CIgen approach by considering applications to molecular systems. To this purpose total energy calculations for the C_2 , N_2 , and H_2O molecules are performed for the 6-31G and cc-pVDZ basis sets considering the active spaces indicated in the third column of Table 11. For the largest system the number of spin orbitals is about double than what previously achieved with neural network quantum states [6].

As a preliminary step it is important to discuss the effect of the temperature on the generative power of the RBM (see eq. (155)). Figure 23 shows the convergence of the total energy of H_2O in the cc-pVDZ for three values of the temperature. The convergence of the CIgen approach is optimal at $T=1$ (chemical accuracy is achieved within 11 iterations) but sizeably slows down when T is increased to 5. In the limit of a very large temperature ($T=500$) the method performs poorly. Indeed, as explained earlier, the $T \rightarrow \infty$ limit leads to a completely random generation of determinants with uniform probability. This shows that for reasonable choices of the temperature the CIgen method actually learns the probability associated with the wavefunction and significantly speeds up the generation of the excited determinants with respect to random sampling.

We now consider the convergence of the total energy in the cc-pVDZ basis for the three molecules C_2 , N_2 , and H_2O in the cc-pVDZ basis and compare the performance of the CIgen and CIgen-SD algorithms with the MCCI algorithm. In all three cases the two CIgen variants considered here outperform the MCCI method. This is particularly true for the CIgen-SD approach that achieves chemical accuracy with a significantly smaller number of iterations.

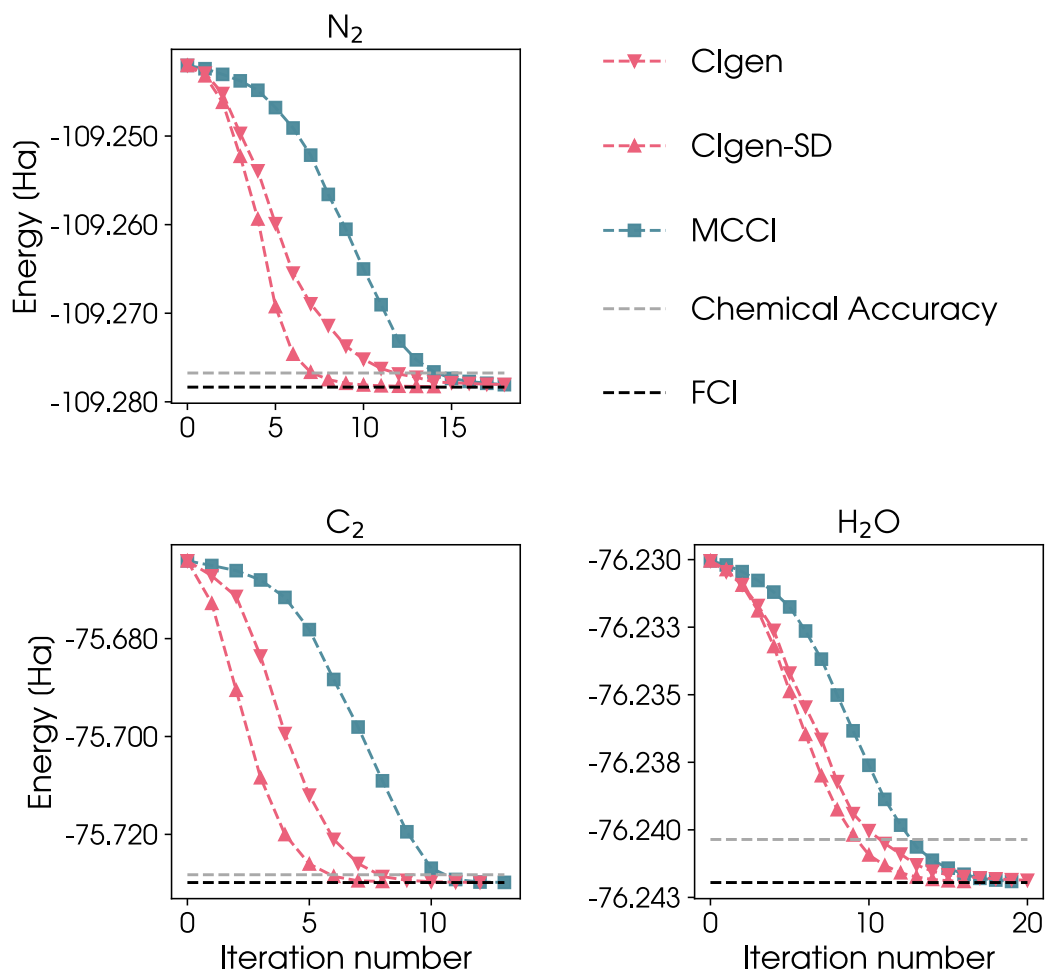


Figure 24: Total energy convergence as a function of the number of iterations: The three molecules N_2 , C_2 , and H_2O in the cc-pVDZ basis set are considered and the plot compares the CIgen, CIgen-SD and MCCI algorithms. The full configuration interaction (FCI) reference values and the chemical accuracy threshold are represented by horizontal lines.

The detailed results for the total energies of the molecules considered here are presented in Table 11, where CIgen-SD values are compared to coupled-cluster with singles, doubles and perturbative triples CCSD(T) and to FCI reference values. By considering a limited number of iterations (from a minimum of 9 for C₂ in 6-31G to a maximum of 17 for water in cc-pVDZ), CIgen-SD converges to FCI values within chemical accuracy. This is achieved by generating an excited determinant subspace that is significantly smaller than the full space. For example, for the molecule N₂ in the cc-pVDZ basis set only 18 millions determinants are generated and included in the wavefunction out of the about 540 millions that would be allowed by spin and space symmetries. This significant reduction in the number of determinants is not achieved by a search in the configuration space but rather by an iterative generation (and model training) of the determinants that contribute the most to the total energy. If for certain applications a lower level of accuracy can be tolerated the number of selected determinants can be further decreased by increasing the value of p_{\min} . Using a threshold of 10^{-8} instead of 10^{-12} , we obtain a total energy of -109.27216 Ha, corresponding to 98.120% of the correlation energy instead of 99.979 % and with only about 156 thousand determinants included in the wavefunction instead of almost 19 millions.

System	Basis	(N,K)	$E_{\text{CIgen-SD}}$	$E_{\text{CCSD(T)}}$	E_{FCI}	N_{conv}
C ₂	6-31G	(12,18)	-75.64416	-75.64415	-75.64418	12
N ₂	6-31G	(14,18)	-109.10824	-109.10635	-109.10842	12
H ₂ O	6-31G	(10,13)	-76.12220	-76.12182	-76.12237	10
C ₂	cc-pVDZ	(8,26)	-75.72982	-75.72781	-75.72984	9
N ₂	cc-pVDZ	(10,26)	-109.27827	-109.27829	-109.27834	13
H ₂ O	cc-pVDZ	(8,23)	-76.24192	-76.24131	-76.24195	17

Table 11: Total energies (Ha) for the different systems/basis sets: The CIgen-SD energies are compared to reference FCI and CCSD(T) values. The (N,K) column indicate the size of the active space, where N is the number of correlated electrons and K is the number of active molecular orbitals, and the N_{conv} column indicates the number of iterations to achieve convergence (i.e when the energy in two successive iterations differs by less than 10^{-5} Hartree).

As a final result in Figure 25 we present a full binding curve for the N₂ molecule. This represent a much more challenging example since for large interatomic distances the binding is characterized by strong static correlation whose description could become problematic for several traditional quantum chemical approaches. This is the case of CCSD(T) that starting at around 1.8 Å produces a binding curve

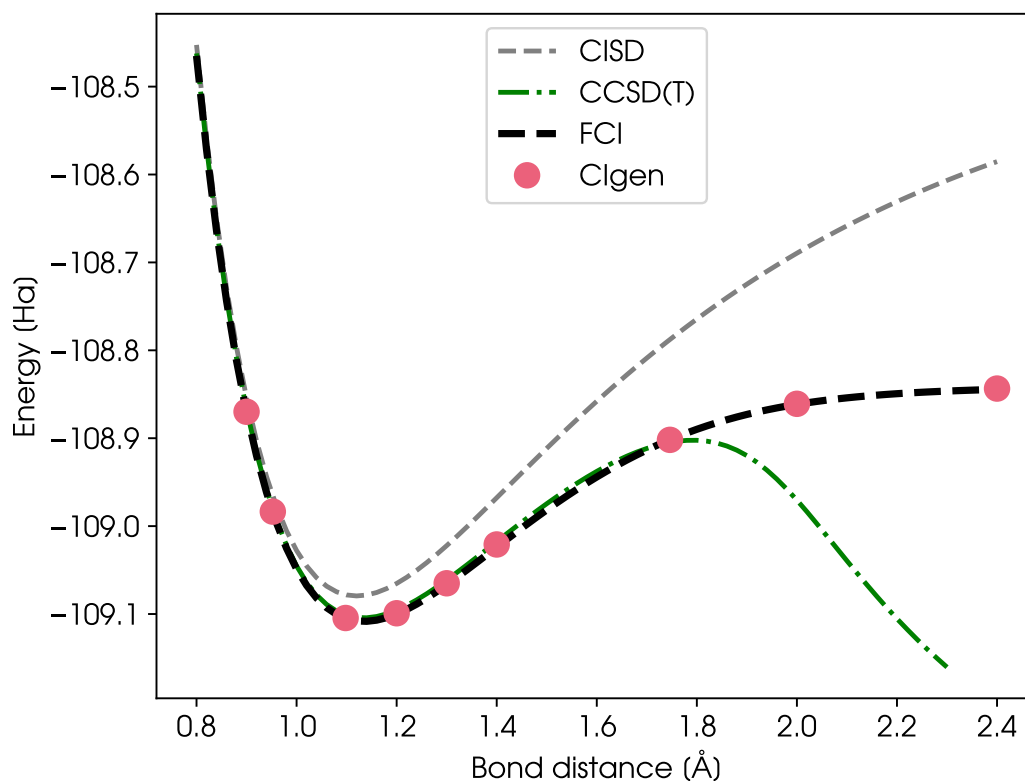


Figure 25: CIgen dissociation curve of N_2 : The results are obtained using the 6-31G basis set and the CCSD(T), FCI, and CISD curves are provided for comparison purposes.

with an unphysical behavior. The CIgen approach well describes the N_2 binding curve at every distance and produces a curve in excellent agreement with FCI. The comparison with the CISD curve, which provides the starting training data for the generative model, shows clearly that CIgen recovers an increasing amount of electronic correlation energy as a function of the interatomic distance. The increase in the electronic correlation manifests itself in the number of determinants that have to be included in the wavefunction to achieve convergence: At equilibrium distance the CIgen method selects a few hundreds of thousands of determinants while closer to dissociation (2.4 Å), around three millions determinants are necessary. While the number of the required determinants changes at different distances it still remains significantly smaller than the 127 millions of determinants in the full CI solution.

6.4 CONCLUSIONS

In summary we have shown how a generative model can be used to solve the Schrödinger equation by sampling the excited Slater determinants that contribute

the most to the wavefunction. Numerical applications show that this approach is already competitive with previous approaches based on Monte Carlo sampling of the excited determinants [67] or on machine learning ansatzes to represent the wavefunction in the configuration space [6, 35]. An improvement that should be addressed in future work involves the development of a generative model that intrinsically takes into account the symmetry of the determinants and, more in general, the properties of the determinants contributing to the wavefunction. It has already been shown that a sizeable improvement in the convergence speed can be achieved by directly generating the determinants that couple with double and single substitutions to the determinants already included in the wavefunction. In the current implementation, however, spin and spatial symmetries of the generated determinants are verified and enforced only *a posteriori*. The development of a generative machine learning model that keeps into account these symmetries would certainly increase the numerical efficiency of CIgen to possibly perform calculations with significantly larger numbers of spin orbitals. This will be considered in future work by developing new symmetry compliant architectures or by directly excluding symmetry forbidden substitutions in the transition probability matrix.

NOTE

The programs needed to perform CIgen and CIgen-SD calculations are available at <https://github.com/bslhrzg/cigen>.

CONCLUSION AND PERSPECTIVE

This thesis has brought contributions to the computational physics emerging field of machine learning (ML) aided *ab initio* calculations. More specifically, it has been seen how machine learning can significantly lower the computational cost of correlated quantum methods to ease the reach of chemical accuracy.

In Chapter 4, an assessment of the accuracy and of the limitations of Machine Learning Perturbation Theory (MLPT) was provided. MLPT uses a ML model to learn the energy difference between two quantum mechanical approximations, and use that model to perform thermodynamic perturbation theory (TPT) to a fraction of its cost, allowing to obtain finite temperature estimates in theories that would be inaccessible from a straightforward molecular dynamics. Using different exchange-correlation functionals to study the enthalpy of adsorption of small molecules inside zeolites, it was shown how TPT can be sensible to a potential mismatch between reference and target configurational spaces. This was made apparent from a quantitative deviation of the MLPT estimates with respect to the reference values of total electronic energies and enthalpies of adsorption, and from a clustering of the non overlapping configurational spaces using dimensionality reduction techniques. A diagnostic test, in the form of a number named I_w , was proposed to detect those biased estimates. This number measures the fraction of configurations necessary to recover half of the sorted statistical weights after the TPT reweighting. A half unit for I_w corresponds to a uniform shift of all energies, while a I_w close to zero indicates that a negligible fraction of the configurations holds most of the statistical weights, resulting in a biased ensemble average. A solution to this problem was devised in the form of a Monte Carlo resampling of the statistics, in combination with the same ML model (MLMC). The MLMC method is a straightforward Metropolis algorithm, where the energy evaluation is performed in two steps: the energy of the configuration is first computed using the reference theory, and the previous MLPT Δ -ML model is used in a second step to infer the target value of the energy. The correct configurational space can be recovered in this way, at the price of a new *ab initio* simulation, but without any new target level calculation. MLMC can also be used to further confirm some MLPT estimate with a good *a priori* overlap between reference and target levels of theory, as was done to study previous results in the Random Phase Approximation.

Those results were used in Chapter 5, to provide an estimate of the ambient temperature enthalpy of adsorption of carbon dioxide in a protonated chabazite with Coupled Cluster Single Double and perturbative Triple (CCSD(T)). Using a periodic implementation of the Coupled Cluster method, in a localized basis set formed by the MP2 natural orbitals, one hundred and ten CCSD(T) calculations were performed for each (molecule, zeolite and adsorbed zeolite) system. The further MLPT estimate was found to be in excellent agreement with the experimental reference, with an error below 0.1 kcal/mol. Because of a rather small value of I_w in this work, even if reliable from the results of Chapter 4, a MLMC estimate was computed, confirming the MLPT result, and providing what is to our knowledge the first finite temperature CCSD(T) estimate for a periodic material.

There is however still many roads to further develop and assess MLPT and MLMC. In the work presented here, the training set was chosen from evenly separated points in the reference molecular dynamics. A more informed training set choice could allow to reduce even more the size of the training set or allow for a better learning. Other level of reference theories could be considered, such as semi empirical tight binding models, for example associated to density functional theory (DFT) to provide fast estimates with DFT accuracy. More complex systems could be studied, chemical reactions for instance, to have more insight into the capabilities of MLPT. A perfectionned MLMC algorithm, using, *e.g.*, parallel tempering, could be equally profitable, in order to reduce the autocorrelation length of the samples, which induce a higher error bar compared to the MD.

Finally, in Chapter 6, it was shown, by computing total energies and dissociation curve of N_2 , C_2 and H_2O in the 6-31G and cc-PVDZ basis sets, that a generative ML model can be used to efficiently generate new Slater determinants in an iterative selective Configuration Interaction (CI) algorithm. The ML model is a restricted Boltzmann machine (RBM): a two-layers energy based generative neural network, that models a probability distribution, from which one can sample. Contrary to neural quantum states where a ML model is used to model the wave function, this is the wave function squared amplitude that is fitted here. From a starting wave function, the RBM is trained to learn the Born probabilities associated to each Slater determinant in the wave function. In a second step new determinants are sampled from the RBM, providing a bigger Hilbert space to be diagonalized for the ground state. The new ground state coefficients provide new probabilities to be learned, and this is repeated iteratively until convergence. Two versions of this method were proposed: the first one sample new determinants in a direct manner, by populating orbitals according to their RBM occupation probability. The second version exploit the properties of the connections among determinants in the CI space, by directly proposing single and double substitutions to the current wave

function. This is achieved by constructing a transition probability matrix from the RBM occupation probabilities. This methodology, coined CIgen, was seen to outperform the uniform sampling of the determinants, and to be already competitive with previous Monte Carlo approaches or ML quantum Monte Carlo Configuration Interaction ansatzes. By reducing the number of iterations necessary to converge, several of the most costly iteration (the obtention of the ground state) are avoided. One future improvement could be to directly learn the symmetries of the system, instead of enforcing them *a posteriori*. Other ML models on which exact sampling is possible could be tested, as well as other physical systems. Encouraging results were found using the CIgen code to solve spin systems such as the antiferromagnetic J_1 - J_2 model in one and two dimensions. From a more general point of view, the CIgen model is essentially about replacing a Monte Carlo proposal distribution by a more informed one, in a Bayesian inference framework. This idea has a lot of potential to bring more efficient exploration of the configurational space, allowing in the case of CIgen for a faster convergence of the CI energy minimization. It can also be used in a more traditional Monte Carlo setup to reduce the autocorrelation length of the sample, as was done in the context of lattice field theory with normalizing flow models [103] for example. This could be brought to continuous-time quantum monte carlo as well, or in finite temperature simulation, such as the MLMC algorithm, by learning the configurational space produced by the reference dynamics, and propose Monte Carlo updates in an active or reinforcement learning framework.

BIBLIOGRAPHY

- [1] NIST Computational Chemistry Comparison and Benchmark Database, NIST Standard Reference Database Number 101 Release 22, May 2022, Editor: Russell D. Johnson III <http://cccbdb.nist.gov/>.
- [2] Landau Lev Davidovitch 1908-1968. *Physique théorique. Tome 5. Physique statistique*. fre. [2e édition]. Moscou: Mir, 1967.
- [3] Karim Adil, Youssef Belmabkhout, Renjith S. Pillai, Amandine Cadiau, Prashant M. Bhatt, Ayalew H. Assen, Guillaume Maurin, and Mohamed Eddaoudi. "Gas/vapour separation using ultra-microporous metal-organic frameworks: insights into the structure/separation relationship." In: *Chem. Soc. Rev* 46.11 (2017), pp. 3402–3430. ISSN: 0306-0012. DOI: [10.1039/C7CS00153C](https://doi.org/10.1039/C7CS00153C). URL: <https://pubs.rsc.org/en/content/articlehtml/2017/cs/c7cs00153c><https://pubs.rsc.org/en/content/articlelanding/2017/cs/c7cs00153c><http://xlink.rsc.org/?DOI=C7CS00153C>.
- [4] Hans C. Andersen. "Molecular dynamics simulations at constant pressure and/or temperature." In: *J. Chem. Phys.* 72.4 (1980), pp. 2384–2393. ISSN: 0021-9606. DOI: [10.1063/1.439486](https://doi.org/10.1063/1.439486). URL: <http://aip.scitation.org/doi/10.1063/1.439486>.
- [5] E. Aprà et al. "NWChem: Past, present, and future." In: *The Journal of Chemical Physics* 152.18 (2020), p. 184102. DOI: [10.1063/5.0004997](https://doi.org/10.1063/5.0004997).
- [6] Thomas D Barrett, Aleksei Malyshev, and AI Lvovsky. "Autoregressive neural-network wavefunctions for ab initio quantum chemistry." In: *Nature Machine Intelligence* 4.4 (2022), pp. 351–358.
- [7] Rodney J. Bartlett and Monika Musiał. "Coupled-cluster theory in quantum chemistry." In: *Rev. Mod. Phys.* 79.1 (2007), pp. 291–352. ISSN: 0034-6861. DOI: [10.1103/RevModPhys.79.291](https://doi.org/10.1103/RevModPhys.79.291). URL: <https://journals.aps.org/rmp/abstract/10.1103/RevModPhys.79.291><https://link.aps.org/doi/10.1103/RevModPhys.79.291>.
- [8] Albert P. Bartok, Sandip De, Carl Poelking, Noam Bernstein, James Kermode, Gabor Csanyi, and Michele Ceriotti. "Machine Learning Unifies the Modelling of Materials and Molecules." en. In: *Science Advances* 3.12 (Dec. 2017). DOI: [10.1126/sciadv.1701816](https://doi.org/10.1126/sciadv.1701816).

- [9] Albert P. Bartók, Risi Kondor, and Gábor Csányi. "On representing chemical environments." In: *Phys. Rev. B* 87.18 (2013), p. 184115. ISSN: 1098-0121. DOI: [10.1103/PhysRevB.87.184115](https://doi.org/10.1103/PhysRevB.87.184115). arXiv: [1209.3140](https://arxiv.org/abs/1209.3140). URL: <https://journals.aps.org/prb/abstract/10.1103/PhysRevB.87.184115><https://link.aps.org/doi/10.1103/PhysRevB.87.184115>.
- [10] Albert P. Bartók, Risi Kondor, and Gábor Csányi. "On representing chemical environments." In: *Phys. Rev. B* 87.18 (2013), p. 184115.
- [11] Albert P. Bartók, Mike C. Payne, Risi Kondor, and Gábor Csányi. "Gaussian Approximation Potentials: The Accuracy of Quantum Mechanics, without the Electrons." In: *Phys. Rev. Lett.* 104 (13 2010), p. 136403. DOI: [10.1103/PhysRevLett.104.136403](https://doi.org/10.1103/PhysRevLett.104.136403). URL: <https://link.aps.org/doi/10.1103/PhysRevLett.104.136403>.
- [12] Jefferson E. Bates and Filipp Furche. "Communication: Random phase approximation renormalized many-body perturbation theory." In: *J. Chem. Phys.* 139.17 (2013), p. 171103. ISSN: 0021-9606. DOI: [10.1063/1.4827254](https://doi.org/10.1063/1.4827254). URL: <https://aip.scitation.org/doi/abs/10.1063/1.4827254><http://aip.scitation.org/doi/10.1063/1.4827254>.
- [13] Jörg Behler. "Atom-centered symmetry functions for constructing high-dimensional neural network potentials." In: *J. Chem. Phys.* 134.7 (2011), p. 074106. ISSN: 0021-9606. DOI: [10.1063/1.3553717](https://doi.org/10.1063/1.3553717). URL: <https://aip.scitation.org/doi/abs/10.1063/1.3553717><http://aip.scitation.org/doi/10.1063/1.3553717>.
- [14] Jörg Behler. "Perspective: Machine learning potentials for atomistic simulations." In: *J. Chem. Phys.* 145.17 (2016), p. 170901. ISSN: 0021-9606. DOI: [10.1063/1.4966192](https://doi.org/10.1063/1.4966192). URL: <https://aip.scitation.org/doi/abs/10.1063/1.4966192><http://aip.scitation.org/doi/10.1063/1.4966192>.
- [15] Jörg Behler and Michele Parrinello. "Generalized Neural-Network Representation of High-Dimensional Potential-Energy Surfaces." In: *Phys. Rev. Lett.* 98.14 (2007), p. 146401. ISSN: 0031-9007. DOI: [10.1103/PhysRevLett.98.146401](https://doi.org/10.1103/PhysRevLett.98.146401). URL: <https://journals.aps.org/prl/abstract/10.1103/PhysRevLett.98.146401><https://link.aps.org/doi/10.1103/PhysRevLett.98.146401>.
- [16] Michael Betancourt. "The Convergence of Markov Chain Monte Carlo Methods: From the Metropolis Method to Hamiltonian Monte Carlo." In: *Ann. Phys.* 531.3 (2019), p. 1700214. ISSN: 00033804. DOI: [10.1002/andp.201700214](https://doi.org/10.1002/andp.201700214). URL: <https://onlinelibrary.wiley.com/doi/full/10.1002/andp.201700214><https://onlinelibrary.wiley.com/doi/abs/10.1002/andp.201700214><https://onlinelibrary.wiley.com/doi/10.1002/andp.201700214>.

- [17] Thomas B. Blank, Steven D. Brown, August W. Calhoun, and Douglas J. Doren. "Neural network models of potential energy surfaces." In: *The Journal of Chemical Physics* 103.10 (Sept. 1995), pp. 4129–4137. ISSN: 0021-9606. DOI: [10.1063/1.469597](https://doi.org/10.1063/1.469597). eprint: https://pubs.aip.org/aip/jcp/article-pdf/103/10/4129/10777334/4129_1_online.pdf. URL: <https://doi.org/10.1063/1.469597>.
- [18] Mihail Bogojeski, Leslie Vogt-Maranto, Mark E Tuckerman, Klaus-Robert Müller, and Kieron Burke. "Quantum chemical accuracy from density functional approximations via machine learning." In: *Nature communications* 11.1 (2020), p. 5223.
- [19] David Bohm and David Pines. "A Collective Description of Electron Interactions: III. Coulomb Interactions in a Degenerate Electron Gas." In: *Phys. Rev.* 92.3 (1953), pp. 609–625. ISSN: 0031-899X. DOI: [10.1103/PhysRev.92.609](https://doi.org/10.1103/PhysRev.92.609). URL: <https://journals.aps.org/pr/abstract/10.1103/PhysRev.92.609><https://link.aps.org/doi/10.1103/PhysRev.92.609>.
- [20] George H. Booth, Andreas Grüneis, Georg Kresse, and Ali Alavi. "Towards an exact description of electronic wavefunctions in real solids." In: *Nature* 493.7432 (2013), pp. 365–370. ISSN: 0028-0836. DOI: [10.1038/nature11770](https://doi.org/10.1038/nature11770). URL: <https://www.nature.com/articles/nature11770><http://www.nature.com/articles/nature11770>.
- [21] Tomas Bucko, Monika Gesvandtnerova, and Dario Rocca. "Ab Initio Calculations of Free Energy of Activation at Multiple Electronic Structure Levels Made Affordable: An Effective Combination of Perturbation Theory and Machine Learning." In: *J. Chem. Theory Comput.* 16.10 (2020), pp. 6049–6060. ISSN: 1549-9618. DOI: [10.1021/acs.jctc.0c00486](https://doi.org/10.1021/acs.jctc.0c00486). URL: <https://pubs.acs.org/doi/abs/10.1021/acs.jctc.0c00486><https://pubs.acs.org/doi/10.1021/acs.jctc.0c00486>.
- [22] Tomas Bucko, Jürgen Hafner, Sébastien Lebègue, and Janos G. Angyan. "Improved Description of the Structure of Molecular and Layered Crystals: Ab Initio DFT Calculations with van der Waals Corrections." In: *J. Phys. Chem. A* 114.43 (2010), pp. 11814–11824. ISSN: 1089-5639. DOI: [10.1021/jp106469x](https://doi.org/10.1021/jp106469x). URL: <https://pubs.acs.org/doi/abs/10.1021/jp106469x><https://pubs.acs.org/doi/10.1021/jp106469x>.
- [23] Tomas Bucko, Sebastien Lebegue, Tim Gould, and Janos G. Angyan. "Many-body dispersion corrections for periodic systems: an efficient reciprocal space implementation." In: *J. Phys. - Condens. Matter* 28 (2016), p. 045201.
- [24] Giuseppe Carleo, Ignacio Cirac, Kyle Cranmer, Laurent Daudet, Maria Schuld, Naftali Tishby, Leslie Vogt-Maranto, and Lenka Zdeborová. "Machine learning and the physical sciences." In: *Rev. Mod. Phys.* 91 (4 2019), p. 045002.

- DOI: [10.1103/RevModPhys.91.045002](https://doi.org/10.1103/RevModPhys.91.045002). URL: <https://link.aps.org/doi/10.1103/RevModPhys.91.045002>.
- [25] Giuseppe Carleo and Matthias Troyer. "Solving the quantum many-body problem with artificial neural networks." In: *Science* 355.6325 (2017), pp. 602–606.
- [26] Bastien Casier, Mauricio Chagas da Silva, Michael Badawi, Fabien Pascale, Tomas Bucko, Sébastien Lebègue, and Dario Rocca. "Hybrid localized graph kernel for machine learning energy-related properties of molecules and solids." In: *J. Comp. Chem.* 42.20 (2021), pp. 1390–1401. ISSN: 0192-8651. DOI: [10.1002/jcc.26550](https://doi.org/10.1002/jcc.26550). URL: <https://onlinelibrary.wiley.com/doi/full/10.1002/jcc.26550><https://onlinelibrary.wiley.com/doi/abs/10.1002/jcc.26550><https://onlinelibrary.wiley.com/doi/10.1002/jcc.26550>.
- [27] *Cc4s User Documentation*. URL: <https://manuals.cc4s.org/user-manual/>.
- [28] Bilal Chehaibou, Michael Badawi, Tomas Bucko, Timur Bazhurov, and Dario Rocca. "Computing RPA Adsorption Enthalpies by Machine Learning Thermodynamic Perturbation Theory." In: *Journal of Chemical Theory and Computation* 15.11 (Nov. 2019), pp. 6333–6342. DOI: [10.1021/acs.jctc.9b00782](https://doi.org/10.1021/acs.jctc.9b00782).
- [29] Michael S. Chen, Joonho Lee, Hong-Zhou Ye, Timothy C. Berkelbach, David R. Reichman, and Thomas E. Markland. "Data-Efficient Machine Learning Potentials from Transfer Learning of Periodic Correlated Electronic Structure Methods: Liquid Water at AFQMC, CCSD, and CCSD(T) Accuracy." In: *Journal of Chemical Theory and Computation* 0.0 (0). PMID: 36730728, null. DOI: [10.1021/acs.jctc.2c01203](https://doi.org/10.1021/acs.jctc.2c01203). eprint: <https://doi.org/10.1021/acs.jctc.2c01203>. URL: <https://doi.org/10.1021/acs.jctc.2c01203>.
- [30] Siwar Chibani, Mouheb Chebbi, Sébastien Lebègue, Tomas Bucko, and Michael Badawi. "A DFT investigation of the adsorption of iodine compounds and water in H-, Na-, Ag-, and Cu- mordenite." In: *J. Chem. Phys.* 144.24 (2016), p. 244705. ISSN: 0021-9606. DOI: [10.1063/1.4954659](https://doi.org/10.1063/1.4954659). URL: <https://aip.scitation.org/doi/abs/10.1063/1.4954659><http://aip.scitation.org/doi/10.1063/1.4954659>.
- [31] Ch Chipot and A. Pohorille. "Free energy calculations: theory and applications in chemistry and biology." In: *Springer* 86 (2007). DOI: [10.1007/978-3-540-38448-9](https://doi.org/10.1007/978-3-540-38448-9).
- [32] Christophe Chipot and Andrew Pohorille. *Free energy calculations: Theory and Applications in Chemistry and Biology*. Springer, 2016.

- [33] Stefan Chmiela, Huziel E. Sauceda, Klaus-Robert Müller, and Alexandre Tkatchenko. "Towards exact molecular dynamics simulations with machine-learned force fields." In: *Nat. Commun.* 9.1 (2018), p. 3887. ISSN: 2041-1723. DOI: [10.1038/s41467-018-06169-2](https://doi.org/10.1038/s41467-018-06169-2). URL: <https://www.nature.com/articles/s41467-018-06169-2>
- [34] Stefan Chmiela, Alexandre Tkatchenko, Huziel E. Sauceda, Igor Poltavsky, Kristof T. Schütt, and Klaus-Robert Müller. "Machine learning of accurate energy-conserving molecular force fields." In: *Sci. Adv.* 3.5 (2017). ISSN: 2375-2548. DOI: [10.1126/sciadv.1603015](https://doi.org/10.1126/sciadv.1603015). URL: <https://www.science.org/doi/abs/10.1126/sciadv.1603015>
- [35] Kenny Choo, Antonio Mezzacapo, and Giuseppe Carleo. "Fermionic neural-network states for ab-initio electronic structure." In: *Nature communications* 11.1 (2020), pp. 1–7.
- [36] Jeremy P Coe. "Machine learning configuration interaction." In: *Journal of Chemical Theory and Computation* 14.11 (2018), pp. 5739–5749.
- [37] Aron J. Cohen, Paula Mori-Sánchez, and Weitao Yang. "Challenges for Density Functional Theory." In: *Chemical Reviews* 112.1 (2012). PMID: 22191548, pp. 289–320. DOI: [10.1021/cr200107z](https://doi.org/10.1021/cr200107z). eprint: <https://doi.org/10.1021/cr200107z>. URL: <https://doi.org/10.1021/cr200107z>.
- [38] Nicola Colonna, Maria Hellgren, and Stefano de Gironcoli. "Correlation energy within exact-exchange adiabatic connection fluctuation-dissipation theory: Systematic development and simple approximations." In: *Phys. Rev. B* 90.12 (2014), p. 125150. ISSN: 1098-0121. DOI: [10.1103/PhysRevB.90.125150](https://doi.org/10.1103/PhysRevB.90.125150). URL: <https://journals.aps.org/prb/abstract/10.1103/PhysRevB.90.125150>
- [39] Dieter Cremer. "Møller–Plesset perturbation theory: from small molecule methods to methods for thousands of atoms." In: *WIREs Computational Molecular Science* 1.4 (2011), pp. 509–530. DOI: <https://doi.org/10.1002/wcms.58>. eprint: <https://wires.onlinelibrary.wiley.com/doi/pdf/10.1002/wcms.58>. URL: <https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/wcms.58>.
- [40] János Daru, Harald Forbert, Jörg Behler, and Dominik Marx. "Coupled Cluster Molecular Dynamics of Condensed Phase Systems Enabled by Machine Learning Potentials: Liquid Water Benchmark." In: *Physical Review Letters* 129.22 (2022), p. 226001.

- [41] Sanjoy Dasgupta and Anupam Gupta. “An elementary proof of a theorem of Johnson and Lindenstrauss.” In: *Random Structures & Algorithms* 22.1 (2003), pp. 60–65. DOI: <https://doi.org/10.1002/rsa.10073>. eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/rsa.10073>. URL: <https://onlinelibrary.wiley.com/doi/abs/10.1002/rsa.10073>.
- [42] C. David Sherrill and Henry F. Schaefer. “The Configuration Interaction Method: Advances in Highly Correlated Approaches.” In: ed. by Per-Olov Löwdin, John R. Sabin, Michael C. Zerner, and Erkki Brändas. Vol. 34. *Advances in Quantum Chemistry*. Academic Press, 1999, pp. 143–269. DOI: [https://doi.org/10.1016/S0065-3276\(08\)60532-8](https://doi.org/10.1016/S0065-3276(08)60532-8). URL: <https://www.sciencedirect.com/science/article/pii/S0065327608605328>.
- [43] Sandip De, Albert P. Bartók, Gábor Csányi, and Michele Ceriotti. “Comparing molecules and solids across structural and alchemical space.” In: *Phys. Chem. Chem. Phys.* 18.20 (2016), pp. 13754–13769. ISSN: 1463-9076. DOI: [10.1039/C6CP00415F](https://doi.org/10.1039/C6CP00415F). arXiv: [1601.04077](https://arxiv.org/abs/1601.04077). URL: <https://pubs.rsc.org/en/content/articlehtml/2016/cp/c6cp00415f><https://pubs.rsc.org/en/content/articlelanding/2016/cp/c6cp00415f><http://xlink.rsc.org/?DOI=C6CP00415F>.
- [44] Sandip De, Albert P. Bartók, Gábor Csányi, and Michele Ceriotti. “Comparing molecules and solids across structural and alchemical space.” In: *Phys. Chem. Chem. Phys.* 18.20 (2016), pp. 13754–13769.
- [45] Mauro Del Ben, Jürg Hutter, and Joost VandeVondele. “Second-Order Møller–Plesset Perturbation Theory in the Condensed Phase: An Efficient and Massively Parallel Gaussian and Plane Waves Approach.” In: *J. Chem. Theory Comput.* 8.11 (2012), pp. 4177–4188. ISSN: 1549-9618. DOI: [10.1021/ct300531w](https://doi.org/10.1021/ct300531w). URL: <https://pubs.acs.org/doi/abs/10.1021/ct300531w><https://pubs.acs.org/doi/10.1021/ct300531w>.
- [46] Volker L. Deringer, Noam Bernstein, Gábor Csányi, Chiheb Ben Mahmoud, Michele Ceriotti, Mark Wilson, David A. Drabold, and Stephen R. Elliott. “Origins of structural and electronic transitions in disordered silicon.” In: *Nature* 589.7840 (2021), pp. 59–64. ISSN: 0028-0836. DOI: [10.1038/s41586-020-03072-z](https://doi.org/10.1038/s41586-020-03072-z). URL: <http://www.nature.com/articles/s41586-020-03072-z>.
- [47] M. Dion, H. Rydberg, E. Schröder, D. C. Langreth, and B. I. Lundqvist. “Van der Waals Density Functional for General Geometries.” In: *Phys. Rev. Lett.* 92.24 (2004), p. 246401. ISSN: 0031-9007. DOI: [10.1103/PhysRevLett.92.246401](https://doi.org/10.1103/PhysRevLett.92.246401). arXiv: [0402105](https://arxiv.org/abs/0402105) [cond-mat]. URL: <https://journals.aps.org/prl/abstract/10.1103/PhysRevLett.92.246401><https://link.aps.org/doi/10.1103/PhysRevLett.92.246401>.

- [48] Anant Dixit, János G. Ángyán, and Dario Rocca. “Improving the accuracy of ground-state correlation energies within a plane-wave basis set: The electron-hole exchange kernel.” In: *J. Chem. Phys.* 145.10 (2016), p. 104105. ISSN: 0021-9606. DOI: [10.1063/1.4962352](https://doi.org/10.1063/1.4962352). URL: <https://aip.scitation.org/doi/abs/10.1063/1.4962352><http://aip.scitation.org/doi/10.1063/1.4962352>.
- [49] Anant Dixit, Julien Claudot, Sébastien Lebègue, and Dario Rocca. “Communication: A novel implementation to compute MP2 correlation energies without basis set superposition errors and complete basis set extrapolation.” In: *J. Chem. Phys.* 146.21 (2017), p. 211102. ISSN: 0021-9606. DOI: [10.1063/1.4985096](https://doi.org/10.1063/1.4985096). URL: <https://aip.scitation.org/doi/abs/10.1063/1.4985096><http://aip.scitation.org/doi/10.1063/1.4985096>.
- [50] Anant Dixit, Julien Claudot, Sébastien Lebègue, and Dario Rocca. “Improving the Efficiency of Beyond-RPA Methods within the Dielectric Matrix Formulation: Algorithms and Applications to the A24 and S22 Test Sets.” In: *J. Chem. Theory Comput.* 13.11 (2017), pp. 5432–5442. ISSN: 1549-9618. DOI: [10.1021/acs.jctc.7b00837](https://doi.org/10.1021/acs.jctc.7b00837). URL: <https://pubs.acs.org/doi/abs/10.1021/acs.jctc.7b00837><https://pubs.acs.org/doi/10.1021/acs.jctc.7b00837>.
- [51] John F. Dobson and Jun Wang. “Successful Test of a Seamless van der Waals Density Functional.” In: *Phys. Rev. Lett.* 82.10 (1999), pp. 2123–2126. ISSN: 0031-9007. DOI: [10.1103/PhysRevLett.82.2123](https://doi.org/10.1103/PhysRevLett.82.2123). URL: <https://journals.aps.org/prl/abstract/10.1103/PhysRevLett.82.2123><https://link.aps.org/doi/10.1103/PhysRevLett.82.2123>.
- [52] Florian Dorner, Zoran Sukurma, Christoph Dellago, and Georg Kresse. “Melting Si: Beyond Density Functional Theory.” In: *Phys. Rev. Lett.* 121.19 (2018), p. 195701. ISSN: 0031-9007. DOI: [10.1103/PhysRevLett.121.195701](https://doi.org/10.1103/PhysRevLett.121.195701). arXiv: [1808.01826](https://arxiv.org/abs/1808.01826). URL: <https://journals.aps.org/prl/abstract/10.1103/PhysRevLett.121.195701><https://link.aps.org/doi/10.1103/PhysRevLett.121.195701>.
- [53] Thom H. Dunning. “Gaussian basis sets for use in correlated molecular calculations. I. The atoms boron through neon and hydrogen.” In: *The Journal of Chemical Physics* 90.2 (1989), pp. 1007–1023. DOI: [10.1063/1.456153](https://doi.org/10.1063/1.456153). eprint: <https://doi.org/10.1063/1.456153>. URL: <https://doi.org/10.1063/1.456153>.
- [54] C. Eckart and G. Young. “The approximation of one matrix by another of lower rank.” In: *Psychometrika* 1.3 (1936), pp. 211–218. DOI: [10.1007/BF02288367](https://doi.org/10.1007/BF02288367).

- [55] Eberhard Engel and Reiner Dreizler. *Density Functional Theory: An Advanced Course*. Vol. -1. Jan. 2010. ISBN: 978-3-642-14089-1. DOI: [10.1007/978-3-642-14090-7](https://doi.org/10.1007/978-3-642-14090-7).
- [56] A.L. Fetter and J.D. Walecka. *Quantum Theory of Many-particle Systems*. Dover Books on Physics. Dover Publications, 2003. ISBN: 9780486428277. URL: <https://books.google.fr/books?id=0wekf1s83b0C>.
- [57] F. Furche and T. van Voorhis. "Fluctuation-dissipation theorem density-functional theory." In: *J. Chem. Phys.* 122.16 (Apr. 2005), p. 164106.
- [58] Philipp Furche. "Molecular tests of the random phase approximation to the exchange-correlation energy functional." In: *Phys. Rev. B* 64.19 (2001), p. 195120. ISSN: 0163-1829. DOI: [10.1103/PhysRevB.64.195120](https://doi.org/10.1103/PhysRevB.64.195120). URL: <https://journals.aps.org/prb/abstract/10.1103/PhysRevB.64.195120><https://link.aps.org/doi/10.1103/PhysRevB.64.195120>
- [59] Yann Garniron et al. "Quantum Package 2.0: An Open-Source Determinant-Driven Suite of Programs." In: *Journal of Chemical Theory and Computation* 15.6 (2019), pp. 3591–3609.
- [60] Murray Gell-Mann and Keith A. Brueckner. "Correlation Energy of an Electron Gas at High Density." In: *Phys. Rev.* 106.2 (1957), pp. 364–368. ISSN: 0031-899X. DOI: [10.1103/PhysRev.106.364](https://doi.org/10.1103/PhysRev.106.364). URL: <https://journals.aps.org/pr/abstract/10.1103/PhysRev.106.364><https://link.aps.org/doi/10.1103/PhysRev.106.364>.
- [61] Murray Gell-Mann and Francis Low. "Bound States in Quantum Field Theory." In: *Phys. Rev.* 84 (2 1951), pp. 350–354. DOI: [10.1103/PhysRev.84.350](https://doi.org/10.1103/PhysRev.84.350). URL: <https://link.aps.org/doi/10.1103/PhysRev.84.350>.
- [62] Monika Gesvandtnerova, Dario Rocca, and Tomas Bucko. "Methanol carbonylation over acid mordenite: Insights from ab initio molecular dynamics and machine learning thermodynamic perturbation theory." In: *J. Catal.* 396 (2021), pp. 166–178. ISSN: 00219517. DOI: [10.1016/j.jcat.2021.02.011](https://doi.org/10.1016/j.jcat.2021.02.011). URL: <https://linkinghub.elsevier.com/retrieve/pii/S0021951721000695>.
- [63] Jeffrey Goldstone and Nevill Francis Mott. "Derivation of the Brueckner many-body theory." In: *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences* 239.1217 (1957), pp. 267–279. DOI: [10.1098/rspa.1957.0037](https://doi.org/10.1098/rspa.1957.0037). eprint: <https://royalsocietypublishing.org/doi/pdf/10.1098/rspa.1957.0037>. URL: <https://royalsocietypublishing.org/doi/abs/10.1098/rspa.1957.0037>.
- [64] Dorothea Golze, Marc Dvorak, and Patrick Rinke. "The GW Compendium: A Practical Guide to Theoretical Photoemission Spectroscopy." In: *Frontiers in Chemistry* 7 (2019). DOI: [10.3389/fchem.2019.00377](https://doi.org/10.3389/fchem.2019.00377). URL: <https://doi.org/10.3389/fchem.2019.00377>.

- [65] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. "Generative adversarial nets." In: *Advances in neural information processing systems* 27 (2014).
- [66] Lukas Grajciar, Christopher J Heard, Anton A Bondarenko, Mikhail V Polynski, Jittima Meeprasert, Evgeny A Pidko, and Petr Nachtigall. "Towards operando computational modeling in heterogeneous catalysis." In: *Chemical Society Reviews* 47.22 (2018), pp. 8307–8348.
- [67] JC Greer. "Monte Carlo configuration interaction." In: *Journal of Computational Physics* 146.1 (1998), pp. 181–202.
- [68] Stefan Grimme. "Semiempirical GGA-type density functional constructed with a long-range dispersion correction." In: *J. Comp. Chem.* 27.15 (2006), pp. 1787–1799. ISSN: 0192-8651. DOI: [10.1002/jcc.20495](https://doi.org/10.1002/jcc.20495). URL: <http://doi.wiley.com/10.1002/jcc.20495>.
- [69] Stefan Grimme. "Semiempirical GGA-type density functional constructed with a long-range dispersion correction." In: *J. Comput. Chem.* 27.15 (2006), pp. 1787–1799.
- [70] Jeffrey C. Grossman. "Benchmark quantum Monte Carlo calculations." In: *The Journal of Chemical Physics* 117.4 (July 2002), pp. 1434–1440. ISSN: 0021-9606. DOI: [10.1063/1.1487829](https://doi.org/10.1063/1.1487829). eprint: https://pubs.aip.org/aip/jcp/article-pdf/117/4/1434/10844793/1434_1_online.pdf. URL: <https://doi.org/10.1063/1.1487829>.
- [71] Thomas Gruber, Ke Liao, Theodoros Tsatsoulis, Felix Hummel, and Andreas Grüneis. "Applying the coupled-cluster ansatz to solids and surfaces in the thermodynamic limit." In: *Physical Review X* 8.2 (2018), p. 021043.
- [72] Andreas Grüneis, George H. Booth, Martijn Marsman, James Spencer, Ali Alavi, and Georg Kresse. "Natural orbitals for wave function based correlated calculations using a plane wave basis set." In: *J. Chem. Theory Comput.* 7.9 (2011), pp. 2780–2785. ISSN: 15499618. DOI: [10.1021/ct200263g](https://doi.org/10.1021/ct200263g).
- [73] Andreas Grüneis, Martijn Marsman, Judith Harl, Laurids Schimka, and Georg Kresse. "Making the random phase approximation to electronic correlation accurate." In: *J. Chem. Phys.* 131.15 (2009), p. 154115. ISSN: 0021-9606. DOI: [10.1063/1.3250347](https://doi.org/10.1063/1.3250347). URL: <https://aip.scitation.org/doi/abs/10.1063/1.3250347><http://aip.scitation.org/doi/10.1063/1.3250347>.
- [74] Andreas Grüneis, Martijn Marsman, and Georg Kresse. "Second-order Møller–Plesset perturbation theory applied to extended systems. II. Structural and energetic properties." In: *The Journal of Chemical Physics* 133.7 (2010), p. 074107. DOI: [10.1063/1.3466765](https://doi.org/10.1063/1.3466765). eprint: <https://doi.org/10.1063/1.3466765>. URL: <https://doi.org/10.1063/1.3466765>.

- [75] P. J. Knowles H.-J. Werner, G. Knizia, F. R. Manby, and M. Schütz. *MOL-PRO, a package of ab initio programs*. see <https://www.molpro.net>. Stuttgart, Germany.
- [76] Jiequn Han, Linfeng Zhang, and E Weinan. "Solving many-electron Schrödinger equation using deep neural networks." In: *Journal of Computational Physics* 399 (2019), p. 108929.
- [77] Katja Hansen, Franziska Biegler, Raghunathan Ramakrishnan, Wiktor Pronobis, O. Anatole von Lilienfeld, Klaus-Robert Müller, and Alexandre Tkatchenko. "Machine Learning Predictions of Molecular Properties: Accurate Many-Body Potentials and Nonlocality in Chemical Space." In: *The Journal of Physical Chemistry Letters* 6.12 (June 2015), pp. 2326–2331. DOI: [10.1021/acs.jpcllett.5b00831](https://doi.org/10.1021/acs.jpcllett.5b00831).
- [78] Judith Harl and Georg Kresse. "Cohesive energy curves for noble gas solids calculated by adiabatic connection fluctuation-dissipation theory." In: *Phys. Rev. B* 77.4 (2008), p. 045136. ISSN: 1098-0121. DOI: [10.1103/PhysRevB.77.045136](https://doi.org/10.1103/PhysRevB.77.045136). URL: <https://journals.aps.org/prb/abstract/10.1103/PhysRevB.77.045136><https://link.aps.org/doi/10.1103/PhysRevB.77.045136>.
- [79] J. Harris. "Adiabatic-connection approach to Kohn-Sham theory." In: *Phys. Rev. A* 29 (4 1984), pp. 1648–1659. DOI: [10.1103/PhysRevA.29.1648](https://doi.org/10.1103/PhysRevA.29.1648). URL: <https://link.aps.org/doi/10.1103/PhysRevA.29.1648>.
- [80] T Helgaker, P Jørgensen, and J Olsen. *Molecular Electronic Structure Theory*. Chichester: John Wiley & Sons, LTD, 2000.
- [81] Maria Hellgren, Nicola Colonna, and Stefano de Gironcoli. "Beyond the random phase approximation with a local exchange vertex." In: *Phys. Rev. B* 98.4 (2018), p. 045117. ISSN: 2469-9950. DOI: [10.1103/PhysRevB.98.045117](https://doi.org/10.1103/PhysRevB.98.045117). URL: <https://journals.aps.org/prb/abstract/10.1103/PhysRevB.98.045117><https://link.aps.org/doi/10.1103/PhysRevB.98.045117>.
- [82] Jan Hermann, Zeno Schätzle, and Frank Noé. "Deep-neural-network solution of the electronic Schrödinger equation." In: *Nature Chemistry* 12.10 (2020), pp. 891–897.
- [83] Basile Herzog, Bastien Casier, Sébastien Lebègue, and Dario Rocca. "Solving the Schrödinger Equation in the Configuration Space with Generative Machine Learning." In: *Journal of Chemical Theory and Computation* 19.9 (2023). PMID: 37043718, pp. 2484–2490. DOI: [10.1021/acs.jctc.2c01216](https://doi.org/10.1021/acs.jctc.2c01216). eprint: <https://doi.org/10.1021/acs.jctc.2c01216>. URL: <https://doi.org/10.1021/acs.jctc.2c01216>.

- [84] Basile Herzog, Maurício Chagas da Silva, Bastien Casier, Michael Badawi, Fabien Pascale, Tomas Bucko, Sébastien Lebègue, and Dario Rocca. "Assessing the Accuracy of Machine Learning Thermodynamic Perturbation Theory: Density Functional Theory and Beyond." In: *Journal of Chemical Theory and Computation* 18.3 (2022), pp. 1382–1394.
- [85] Etienne Paul Hessou, Hicham Jabraoui, Ibrahim Khalil, Marie-Antoinette Dziurla, and Michael Badawi. "Ab initio screening of zeolite Y formulations for efficient adsorption of thiophene in presence of benzene." In: *Appl. Surf. Sci.* 541 (2021), p. 148515. ISSN: 01694332. DOI: [10.1016/j.apsusc.2020.148515](https://doi.org/10.1016/j.apsusc.2020.148515). URL: <https://linkinghub.elsevier.com/retrieve/pii/S0169433220332736>.
- [86] Lauri Himanen, Marc O.J. Jäger, Eiaki V. Morooka, Filippo Federici Canova, Yashasvi S. Ranawat, David Z. Gao, Patrick Rinke, and Adam S. Foster. "DScribe: Library of descriptors for machine learning in materials science." In: *Comput. Phys. Commun.* 247 (2020), p. 106949. ISSN: 00104655. DOI: [10.1016/j.cpc.2019.106949](https://doi.org/10.1016/j.cpc.2019.106949). arXiv: [1904.08875](https://arxiv.org/abs/1904.08875). URL: <https://linkinghub.elsevier.com/retrieve/pii/S0010465519303042>.
- [87] Geoffrey E Hinton. "A practical guide to training restricted Boltzmann machines." In: *Neural networks: Tricks of the trade*. Springer, 2012, pp. 599–619.
- [88] Geoffrey E Hinton and Ruslan R Salakhutdinov. "Reducing the dimensionality of data with neural networks." In: *Science* 313.5786 (2006), pp. 504–507.
- [89] P. Hohenberg and W. Kohn. "Inhomogeneous Electron Gas." In: *Phys. Rev.* 136 (3B 1964), B864–B871. DOI: [10.1103/PhysRev.136.B864](https://doi.org/10.1103/PhysRev.136.B864). URL: <https://link.aps.org/doi/10.1103/PhysRev.136.B864>.
- [90] Kurt Hornik. "Approximation capabilities of multilayer feedforward networks." In: *Neural networks* 4.2 (1991), pp. 251–257.
- [91] Felix Hummel, Andreas Grüneis, Georg Kresse, and Paul Ziesche. "Screened Exchange Corrections to the Random Phase Approximation from Many-Body Perturbation Theory." en. In: *Journal of Chemical Theory and Computation* 15 (5 May 2019), pp. 3223–3236. DOI: [10.1021/acs.jctc.8b01247](https://doi.org/10.1021/acs.jctc.8b01247). URL: <http://dx.doi.org/10.1021/acs.jctc.8b01247>.
- [92] Felix Hummel, Theodoros Tsatsoulis, and Andreas Grüneis. "Low rank factorization of the Coulomb integrals for periodic coupled cluster theory." In: *The Journal of Chemical Physics* 146.12 (2017), p. 124105. DOI: [10.1063/1.4977994](https://doi.org/10.1063/1.4977994). eprint: <https://doi.org/10.1063/1.4977994>. URL: <https://doi.org/10.1063/1.4977994>.
- [93] Haoyan Huo and Matthias Rupp. "Unified Representation of Molecules and Crystals for Machine Learning." In: *arXiv:1704.06439* (Jan. 2018).

- [94] Haoyan Huo and Matthias Rupp. “Unified Representation of Molecules and Crystals for Machine Learning.” In: *arXiv:1704.06439 [physics.chem-ph]* (2018). arXiv: 1704.06439 [physics.chem-ph]. URL: <https://arxiv.org/abs/1704.06439><http://arxiv.org/abs/1704.06439>.
- [95] Bernard Huron, JP Malrieu, and P Rancurel. “Iterative perturbation calculations of ground and excited state energies from multiconfigurational zeroth-order wavefunctions.” In: *The Journal of Chemical Physics* 58.12 (1973), pp. 5745–5759.
- [96] Andreas Irmeler, Alejandro Gallo, and Andreas Grüneis. “Focal-point approach with pair-specific cusp correction for coupled-cluster theory.” In: *The Journal of Chemical Physics* 154.23 (2021), p. 234103.
- [97] Kevin Maik Jablonka, Daniele Ongari, Seyed Mohamad Moosavi, and Berend Smit. “Big-Data Science in Porous Materials: Materials Genomics and Machine Learning.” In: *Chem. Rev.* 120.16 (2020), pp. 8066–8129. ISSN: 0009-2665. DOI: [10.1021/acs.chemrev.0c00004](https://doi.org/10.1021/acs.chemrev.0c00004). URL: <https://pubs.acs.org/doi/abs/10.1021/acs.chemrev.0c00004><https://pubs.acs.org/doi/10.1021/acs.chemrev.0c00004>.
- [98] Kevin Maik Jablonka, Daniele Ongari, Seyed Mohamad Moosavi, and Berend Smit. “Big-data science in porous materials: materials genomics and machine learning.” In: *Chemical reviews* 120.16 (2020), pp. 8066–8129.
- [99] Robert Jastrow. “Many-Body Problem with Strong Forces.” In: *Phys. Rev.* 98 (5 1955), pp. 1479–1484. DOI: [10.1103/PhysRev.98.1479](https://doi.org/10.1103/PhysRev.98.1479). URL: <https://link.aps.org/doi/10.1103/PhysRev.98.1479>.
- [100] WooSeok Jeong, Carlo Alberto Gaggioli, and Laura Gagliardi. “Active Learning Configuration Interaction for Excited-State Calculations of Polycyclic Aromatic Hydrocarbons.” In: *Journal of chemical theory and computation* 17.12 (2021), pp. 7518–7530.
- [101] Pascual Jordan and Eugene Paul Wigner. “Über das paulische äquivalenzverbot.” In: *Z. Phys.* 47 (1928), p. 631.
- [102] P. Jørgensen and J. Simons. *Second Quantization-based Methods in Quantum Chemistry*. Academic Press, 1981. ISBN: 9780123902207. URL: <https://books.google.bi/books?id=sgCGAAAIAAJ>.
- [103] Gurtej Kanwar. *Machine Learning and Variational Algorithms for Lattice Field Theory*. 2021. arXiv: [2106.01975](https://arxiv.org/abs/2106.01975) [hep-lat].
- [104] Ferenc Karsai, Manuel Engel, Espen Flage-Larsen, and Georg Kresse. “Electron-phonon coupling in semiconductors within the GW approximation.” In: *New Journal of Physics* 20 (Dec. 2018). DOI: [10.1088/1367-2630/aaf53f](https://doi.org/10.1088/1367-2630/aaf53f).

- [105] Jan Kessler, Francesco Calcavecchia, and Thomas D Kühne. “Artificial neural networks as trial wave functions for quantum monte carlo.” In: *Advanced Theory and Simulations* 4.4 (2021), p. 2000269.
- [106] Ibrahim Khalil, Hicham Jabraoui, Sébastien Lebègue, Won June Kim, Luis-Jacobo Aguilera, Karine Thomas, Françoise Maugé, and Michael Badawi. “Biofuel purification: Coupling experimental and theoretical investigations for efficient separation of phenol from aromatics by zeolites.” In: *Chem. Eng. J.* 402 (2020), p. 126264. ISSN: 13858947. DOI: [10.1016/j.cej.2020.126264](https://doi.org/10.1016/j.cej.2020.126264). URL: <https://linkinghub.elsevier.com/retrieve/pii/S1385894720323925>.
- [107] Diederik P Kingma and Max Welling. “Auto-encoding variational bayes.” In: *arXiv preprint arXiv:1312.6114* (2013).
- [108] Peter J Knowles and Nicholas C Handy. “A determinant based full configuration interaction program.” In: *Computer physics communications* 54.1 (1989), pp. 75–83.
- [109] W. Kohn and L. J. Sham. “Self-Consistent Equations Including Exchange and Correlation Effects.” In: *Phys. Rev.* 140 (4A 1965), A1133–A1138. DOI: [10.1103/PhysRev.140.A1133](https://doi.org/10.1103/PhysRev.140.A1133). URL: <https://link.aps.org/doi/10.1103/PhysRev.140.A1133>.
- [110] Werner Krauth. *Statistical mechanics: algorithms and computations*. Vol. 13. OUP Oxford, 2006, pp. 33–34.
- [111] Kurt Kremer and Gary S Grest. “Monte Carlo and molecular dynamics simulations of polymers.” In: *Phys. Scr.* T35.T35 (1991), pp. 61–65. ISSN: 0031-8949. DOI: [10.1088/0031-8949/1991/T35/013](https://doi.org/10.1088/0031-8949/1991/T35/013). URL: <https://iopscience.iop.org/article/10.1088/0031-8949/1991/T35/013><https://iopscience.iop.org/article/10.1088/0031-8949/1991/T35/013/meta>.
- [112] G. Kresse and J. Furthmüller. “Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set.” In: *Comp. Mater. Sci.* 6.1 (1996), pp. 15–50. ISSN: 09270256. DOI: [10.1016/0927-0256\(96\)00008-0](https://doi.org/10.1016/0927-0256(96)00008-0). URL: <https://linkinghub.elsevier.com/retrieve/pii/0927025696000080>.
- [113] G. Kresse and J. Furthmüller. “Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set.” In: *Phys. Rev. B* 54.16 (1996), pp. 11169–11186. ISSN: 0163-1829. DOI: [10.1103/PhysRevB.54.11169](https://doi.org/10.1103/PhysRevB.54.11169). URL: <https://journals.aps.org/prb/abstract/10.1103/PhysRevB.54.11169><https://link.aps.org/doi/10.1103/PhysRevB.54.11169>.
- [114] G. Kresse and J. Furthmüller. “Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set.” In: *Physical Review B* 54.16 (Oct. 1996), pp. 11169–11186. DOI: [10.1103/PhysRevB.54.11169](https://doi.org/10.1103/PhysRevB.54.11169). URL:

- <http://link.aps.org/doi/10.1103/PhysRevB.54.11169> (visited on 09/26/2015).
- [115] G Kresse and J Hafner. “Ab initio molecular dynamics for liquid metals.” In: *Phys. Rev. B* 47.1 (1993), pp. 558–561. ISSN: 0163-1829. DOI: [10.1103/PhysRevB.47.558](https://doi.org/10.1103/PhysRevB.47.558). URL: <https://link.aps.org/doi/10.1103/PhysRevB.47.558>.
- [116] G. Kresse and J. Hafner. “Norm-conserving and ultrasoft pseudopotentials for first-row and transition elements.” en. In: *Journal of Physics: Condensed Matter* 6.40 (1994), p. 8245. ISSN: 0953-8984. DOI: [10.1088/0953-8984/6/40/015](https://doi.org/10.1088/0953-8984/6/40/015). URL: <http://stacks.iop.org/0953-8984/6/i=40/a=015> (visited on 10/17/2016).
- [117] G. Kresse and D. Joubert. “From ultrasoft pseudopotentials to the projector augmented-wave method.” In: *Phys. Rev. B* 59.3 (1999), pp. 1758–1775. ISSN: 0163-1829. DOI: [10.1103/PhysRevB.59.1758](https://doi.org/10.1103/PhysRevB.59.1758). URL: <https://journals.aps.org/prb/abstract/10.1103/PhysRevB.59.1758><https://link.aps.org/doi/10.1103/PhysRevB.59.1758>.
- [118] D.C. Langreth and J.P. Perdew. “The exchange-correlation energy of a metallic surface.” In: *Solid State Commun.* 17.11 (1975), pp. 1425–1429. ISSN: 00381098. DOI: [10.1016/0038-1098\(75\)90618-3](https://doi.org/10.1016/0038-1098(75)90618-3). URL: <https://linkinghub.elsevier.com/retrieve/pii/0038109875906183>.
- [119] Nicolas Le Roux and Yoshua Bengio. “Representational power of restricted Boltzmann machines and deep belief networks.” In: *Neural computation* 20.6 (2008), pp. 1631–1649.
- [120] Ke Liao, Xin-Zheng Li, Ali Alavi, and Andreas Grüneis. “A comparative study using state-of-the-art electronic structure theories on solid hydrogen phases under high pressures.” In: *npj Computational Materials* 5.1 (2019), pp. 1–6.
- [121] Ke Liao, Tong Shen, Xin-Zheng Li, Ali Alavi, and Andreas Grüneis. “Structural and electronic properties of solid molecular hydrogen from many-electron theories.” In: *Physical Review B* 103.5 (2021), p. 054111.
- [122] O Anatole von Lilienfeld, Klaus-Robert Müller, and Alexandre Tkatchenko. “Exploring chemical compound space with quantum-based machine learning.” In: *Nature Reviews Chemistry* 4.7 (2020), pp. 347–358.
- [123] Deyu Lu, Yan Li, Dario Rocca, and Giulia Galli. “Ab initio Calculation of van der Waals Bonded Molecular Crystals.” In: *Phys. Rev. Lett.* 102.20 (2009), p. 206411. ISSN: 0031-9007. DOI: [10.1103/PhysRevLett.102.206411](https://doi.org/10.1103/PhysRevLett.102.206411). URL: <https://journals.aps.org/prl/abstract/10.1103/PhysRevLett.102.206411><https://link.aps.org/doi/10.1103/PhysRevLett.102.206411>.

- [124] Laurens van der Maaten and Geoffrey Hinton. "Visualizing data using t-SNE." In: *J. Mach. Learn. Res.* 9.Nov (2008), pp. 2579–2605.
- [125] David J. C. MacKay. *Information Theory, Inference, and Learning Algorithms*. Copyright Cambridge University Press, 2003.
- [126] Hafez Maghsoudi, Mohammad Soltanieh, Hamidreza Bozorgzadeh, and Ali Mohamadizadeh. "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." In: *Adsorption* 19.5 (2013), pp. 1045–1053. ISSN: 0929-5607. DOI: [10.1007/s10450-013-9528-1](https://doi.org/10.1007/s10450-013-9528-1). URL: <https://link.springer.com/article/10.1007/s10450-013-9528-1><http://link.springer.com/10.1007/s10450-013-9528-1>.
- [127] M. Marsman, A. Grüneis, J. Paier, and G. Kresse. "Second-order Møller–Plesset perturbation theory applied to extended systems. I. Within the projector-augmented-wave formalism using a plane wave basis set." In: *J. Chem. Phys.* 130.18 (2009), p. 184103. ISSN: 00219606. DOI: [10.1063/1.3126249](https://doi.org/10.1063/1.3126249). URL: <https://aip.scitation.org/doi/abs/10.1063/1.3126249><http://scitation.aip.org/content/aip/journal/jcp/130/18/10.1063/1.3126249>.
- [128] Carla D. Martin and Mason A. Porter. "The Extraordinary SVD." In: *The American Mathematical Monthly* 119.10 (2012), pp. 838–851. ISSN: 00029890, 19300972. URL: <https://www.jstor.org/stable/10.4169/amer.math.monthly.119.10.838> (visited on 03/08/2023).
- [129] Richard M. Martin. *Electronic Structure: Basic Theory and Practical Methods*. Cambridge University Press, 2004. DOI: [10.1017/CB09780511805769](https://doi.org/10.1017/CB09780511805769).
- [130] R.D. Mattuck. *A Guide to Feynman Diagrams in the Many-body Problem*. Dover Books on Physics Series. Dover Publications, 1992. ISBN: 9780486670478. URL: <https://books.google.fr/books?id=pe-v8zfxE68C>.
- [131] Reinhard J. Maurer, Christoph Freysoldt, Anthony M. Reilly, Jan Gerit Brandenburg, Oliver T. Hofmann, Torbjörn Björkman, Sébastien Lebègue, and Alexandre Tkatchenko. "Advances in Density-Functional Calculations for Materials Modeling." In: *Annual Review of Materials Research* 49.1 (2019), pp. 1–30. DOI: [10.1146/annurev-matsci-070218-010143](https://doi.org/10.1146/annurev-matsci-070218-010143). eprint: <https://doi.org/10.1146/annurev-matsci-070218-010143>. URL: <https://doi.org/10.1146/annurev-matsci-070218-010143>.
- [132] Pankaj Mehta, Marin Bukov, Ching-Hao Wang, Alexandre G.R. Day, Clint Richardson, Charles K. Fisher, and David J. Schwab. "A high-bias, low-variance introduction to Machine Learning for physicists." In: *Physics Reports* 810 (2019). A high-bias, low-variance introduction to Machine Learn-

- ing for physicists, pp. 1–124. ISSN: 0370-1573. DOI: <https://doi.org/10.1016/j.physrep.2019.03.001>. URL: <https://www.sciencedirect.com/science/article/pii/S0370157319300766>.
- [133] Dimitrios Meimaroglou and Costas Kiparissides. “Review of Monte Carlo Methods for the Prediction of Distributed Molecular and Morphological Polymer Properties.” In: *Ind. Eng. Chem. Res.* 53.22 (2014), pp. 8963–8979. ISSN: 0888-5885. DOI: [10.1021/ie4033044](https://doi.org/10.1021/ie4033044). URL: <https://pubs.acs.org/doi/abs/10.1021/ie4033044><https://pubs.acs.org/doi/10.1021/ie4033044>.
- [134] Roger G Melko, Giuseppe Carleo, Juan Carrasquilla, and J Ignacio Cirac. “Restricted Boltzmann machines in quantum physics.” In: *Nature Physics* 15.9 (2019), pp. 887–892.
- [135] Nicholas Metropolis, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Edward Teller. “Equation of State Calculations by Fast Computing Machines.” In: *The Journal of Chemical Physics* 21.6 (1953), pp. 1087–1092. DOI: [10.1063/1.1699114](https://doi.org/10.1063/1.1699114).
- [136] Nicholas Metropolis, Arianna W Rosenbluth, Marshall N Rosenbluth, Augusta H Teller, and Edward Teller. “Equation of state calculations by fast computing machines.” In: *The journal of chemical physics* 21.6 (1953), pp. 1087–1092.
- [137] Chr. Møller and M. S. Plesset. “Note on an Approximation Treatment for Many-Electron Systems.” In: *Phys. Rev.* 46.7 (1934), pp. 618–622. ISSN: 0031-899X. DOI: [10.1103/PhysRev.46.618](https://doi.org/10.1103/PhysRev.46.618). URL: <https://journals.aps.org/pr/abstract/10.1103/PhysRev.46.618><https://link.aps.org/doi/10.1103/PhysRev.46.618>.
- [138] Péter R. Nagy, Gyula Samu, and Mihály Kállay. “Optimization of the Linear-Scaling Local Natural Orbital CCSD(T) Method: Improved Algorithm and Benchmark Applications.” In: *Journal of Chemical Theory and Computation* 14.8 (2018). PMID: 29965753, pp. 4193–4215. DOI: [10.1021/acs.jctc.8b00442](https://doi.org/10.1021/acs.jctc.8b00442). eprint: <https://doi.org/10.1021/acs.jctc.8b00442>. URL: <https://doi.org/10.1021/acs.jctc.8b00442>.
- [139] Thomas Olsen and Kristian S. Thygesen. “Extending the random-phase approximation for electronic correlation energies: The renormalized adiabatic local density approximation.” In: *Phys. Rev. B* 86.8 (2012), p. 081103. ISSN: 1098-0121. DOI: [10.1103/PhysRevB.86.081103](https://doi.org/10.1103/PhysRevB.86.081103). URL: <https://journals.aps.org/prb/abstract/10.1103/PhysRevB.86.081103><https://link.aps.org/doi/10.1103/PhysRevB.86.081103>.
- [140] Eric Paquet and Herna L. Viktor. “Molecular Dynamics, Monte Carlo Simulations, and Langevin Dynamics: A Computational Review.” In: *Biomed Res.*

- Int.* 2015 (2015), pp. 1–18. ISSN: 2314-6133. DOI: [10.1155/2015/183918](https://doi.org/10.1155/2015/183918). URL: <http://www.hindawi.com/journals/bmri/2015/183918/>.
- [141] Fabian Pedregosa et al. “Scikit-learn: Machine Learning in Python.” In: *J. Mach. Learn. Res.* 12.85 (2011), pp. 2825–2830. ISSN: 1533-7928. URL: <http://scikit-learn.sourceforge.net..>
- [142] Haowei Peng, Zeng-Hui Yang, John P. Perdew, and Jianwei Sun. “Versatile van der Waals Density Functional Based on a Meta-Generalized Gradient Approximation.” In: *Phys. Rev. X* 6.4 (2016), p. 041005. ISSN: 2160-3308. DOI: [10.1103/PhysRevX.6.041005](https://doi.org/10.1103/PhysRevX.6.041005). URL: <https://journals.aps.org/prx/abstract/10.1103/PhysRevX.6.041005><https://link.aps.org/doi/10.1103/PhysRevX.6.041005>.
- [143] John P. Perdew, Kieron Burke, and Matthias Ernzerhof. “Generalized Gradient Approximation Made Simple.” In: *Phys. Rev. Lett.* 77 (18 1996), pp. 3865–3868. DOI: [10.1103/PhysRevLett.77.3865](https://doi.org/10.1103/PhysRevLett.77.3865). URL: <https://link.aps.org/doi/10.1103/PhysRevLett.77.3865>.
- [144] John P. Perdew, Kieron Burke, and Matthias Ernzerhof. “Generalized Gradient Approximation Made Simple.” In: *Phys. Rev. Lett.* 77.18 (1996), pp. 3865–3868. ISSN: 0031-9007. DOI: [10.1103/PhysRevLett.77.3865](https://doi.org/10.1103/PhysRevLett.77.3865). URL: <https://journals.aps.org/prl/abstract/10.1103/PhysRevLett.77.3865><https://link.aps.org/doi/10.1103/PhysRevLett.77.3865>.
- [145] John P. Perdew and Karla Schmidt. “Jacob’s ladder of density functional approximations for the exchange-correlation energy.” In: *AIP Conference Proceedings* 577.1 (July 2001), pp. 1–20. ISSN: 0094-243X. DOI: [10.1063/1.1390175](https://doi.org/10.1063/1.1390175). eprint: https://pubs.aip.org/aip/acp/article-pdf/577/1/1/12108089/1_1_online.pdf. URL: <https://doi.org/10.1063/1.1390175>.
- [146] David Pfau, James S Spencer, Alexander GDG Matthews, and W Matthew C Foulkes. “Ab initio solution of the many-electron Schrödinger equation with deep neural networks.” In: *Physical Review Research* 2.3 (2020), p. 033429.
- [147] Trong D Pham, Qingling Liu, and Raul F Lobo. “Carbon dioxide and nitrogen adsorption on cation-exchanged SSZ-13 zeolites.” In: *Langmuir* 29.2 (2013), pp. 832–839.
- [148] Giovannimaria Piccini, Maristella Alessio, Joachim Sauer, Yuchun Zhi, Yuanshuai Liu, Robin Kolvenbach, Andreas Jentys, and Johannes A. Lercher. “Accurate Adsorption Thermodynamics of Small Alkanes in Zeolites. Ab initio Theory and Experiment for H-Chabazite.” In: *J. Phys. Chem. C* 119.11 (2015), pp. 6128–6137. ISSN: 1932-7447. DOI: [10.1021/acs.jpcc.5b01739](https://doi.org/10.1021/acs.jpcc.5b01739). URL: <https://pubs.acs.org/sharingguidelines><https://pubs.acs.org/doi/10.1021/acs.jpcc.5b01739>.

- [149] Sergio D Pineda Flores. "Chembot: A Machine Learning Approach to Selective Configuration Interaction." In: *Journal of Chemical Theory and Computation* 17.7 (2021), pp. 4028–4038.
- [150] Simone Pisana, Michele Lazzeri, Cinzia Casiraghi, Kostya S. Novoselov, A. K. Geim, Andrea C. Ferrari, and Francesco Mauri. "Breakdown of the adiabatic Born–Oppenheimer approximation in graphene." In: *Nature Materials* 6.3 (2007), pp. 198–201. ISSN: 1476-4660. DOI: [10.1038/nmat1846](https://doi.org/10.1038/nmat1846). URL: <https://doi.org/10.1038/nmat1846>.
- [151] Cesare Pisani, Lorenzo Maschio, Silvia Casassa, Migen Halo, Martin Schütz, and Denis Usvyat. "Periodic local MP2 method for the study of electronic correlation in crystals: Theory and preliminary applications." In: *J. Comp. Chem.* 29.13 (2008), pp. 2113–2124. ISSN: 01928651. DOI: [10.1002/jcc.20975](https://onlinelibrary.wiley.com/doi/full/10.1002/jcc.20975). URL: <https://onlinelibrary.wiley.com/doi/full/10.1002/jcc.20975><https://onlinelibrary.wiley.com/doi/abs/10.1002/jcc.20975><https://onlinelibrary.wiley.com/doi/10.1002/jcc.20975>.
- [152] Andrew Pohorille, Christopher Jarzynski, and Christophe Chipot. "Good Practices in Free-Energy Calculations." In: *The Journal of Physical Chemistry B* 114.32 (Aug. 2010). DOI: [10.1021/jp102971x](https://pubs.acs.org/doi/abs/10.1021/jp102971x).
- [153] Andrew Pohorille, Christopher Jarzynski, and Christophe Chipot. "Good Practices in Free-Energy Calculations." In: *J. Phys. Chem. B* 114.32 (2010), pp. 10235–10253. ISSN: 1520-6106. DOI: [10.1021/jp102971x](https://pubs.acs.org/doi/abs/10.1021/jp102971x). URL: <https://pubs.acs.org/doi/abs/10.1021/jp102971x><https://pubs.acs.org/doi/10.1021/jp102971x>.
- [154] John A. Pople. "Nobel Lecture: Quantum chemical models." In: *Rev. Mod. Phys.* 71 (5 1999), pp. 1267–1274. DOI: [10.1103/RevModPhys.71.1267](https://link.aps.org/doi/10.1103/RevModPhys.71.1267). URL: <https://link.aps.org/doi/10.1103/RevModPhys.71.1267>.
- [155] Raghunathan Ramakrishnan, Pavlo O Dral, Matthias Rupp, and O Anatole von Lilienfeld. "Quantum chemistry structures and properties of 134 kilo molecules." In: *Scientific Data* 1 (2014).
- [156] Raghunathan Ramakrishnan, Pavlo O. Dral, Matthias Rupp, and O. Anatole von Lilienfeld. "Big Data Meets Quantum Chemistry Approximations: The Δ -Machine Learning Approach." In: *Journal of Chemical Theory and Computation* 11.5 (May 2015), pp. 2087–2096. DOI: [10.1021/acs.jctc.5b00099](https://pubs.acs.org/doi/abs/10.1021/acs.jctc.5b00099).
- [157] Raghunathan Ramakrishnan, Pavlo O. Dral, Matthias Rupp, and O. Anatole von Lilienfeld. "Big Data Meets Quantum Chemistry Approximations: The Δ -Machine Learning Approach." In: *J. Chem. Theory Comput.* 11.5 (2015), pp. 2087–2096. ISSN: 1549-9618. DOI: [10.1021/acs.jctc.5b00099](https://pubs.acs.org/doi/pdf/10.1021/acs.jctc.5b00099). URL: <https://pubs.acs.org/doi/pdf/10.1021/acs.jctc.5b00099><https://pubs.acs.org/doi/10.1021/acs.jctc.5b00099>.

- [158] Raghunathan Ramakrishnan, Pavlo O Dral, Matthias Rupp, and O Anatole von Lilienfeld. "Big data meets quantum chemistry approximations: The Δ -machine learning approach." In: *J. Chem. Theory Comput.* 11.5 (2015), pp. 2087–2096.
- [159] Benjamin Ramberger, Tobias Schäfer, and Georg Kresse. "Analytic interatomic forces in the random phase approximation." In: *Phys. Rev. Lett.* 118.10 (2017), p. 106403. DOI: [10.1103/PhysRevLett.118.106403](https://doi.org/10.1103/PhysRevLett.118.106403). URL: <https://journals.aps.org/prl/abstract/10.1103/PhysRevLett.118.106403><http://journals.aps.org/prl/abstract/10.1103/PhysRevLett.118.106403>.
- [160] Carl Edward Rasmussen and Christopher K. I. Williams. "Gaussian processes for machine learning." In: *MIT Press* (2006). DOI: [10.1007/978-3-540-28650-9_4](https://doi.org/10.1007/978-3-540-28650-9_4).
- [161] Xinguo Ren, Patrick Rinke, Christian Joas, and Matthias Scheffler. "Random-phase approximation and its applications in computational chemistry and materials science." In: *Journal of Materials Science* 47.21 (2012), pp. 7447–7471. DOI: [10.1007/s10853-012-6570-4](https://doi.org/10.1007/s10853-012-6570-4). URL: <https://doi.org/10.1007/s10853-012-6570-4>.
- [162] Xinguo Ren, Alexandre Tkatchenko, Patrick Rinke, and Matthias Scheffler. "Beyond the Random-Phase Approximation for the Electron Correlation Energy: The Importance of Single Excitations." In: *Phys. Rev. Lett.* 106 (15 2011), p. 153003.
- [163] Andrea Rizzi, Paolo Carloni, and Michele Parrinello. "Targeted Free Energy Perturbation Revisited: Accurate Free Energies from Mapped Reference Potentials." In: *J. Phys. Chem. Lett.* (2021), pp. 9449–9454. ISSN: 1948-7185. DOI: [10.1021/acs.jpcllett.1c02135](https://doi.org/10.1021/acs.jpcllett.1c02135). URL: <https://pubs.acs.org/doi/abs/10.1021/acs.jpcllett.1c02135><https://pubs.acs.org/doi/10.1021/acs.jpcllett.1c02135>.
- [164] Dario Rocca, Anant Dixit, Michael Badawi, Sébastien Lebègue, Tim Gould, and Tomas Bucko. "Bridging molecular dynamics and correlated wavefunction methods for accurate finite-temperature properties." In: *Phys. Rev. Materials* 3.4 (2019), p. 040801.
- [165] Dario Rocca, Anant Dixit, Michael Badawi, Sébastien Lebègue, Tim Gould, and Tomas Bucko. "Bridging molecular dynamics and correlated wavefunction methods for accurate finite-temperature properties." In: *Phys. Rev. Mater.* 3.4 (2019), p. 040801. ISSN: 2475-9953. DOI: [10.1103/PhysRevMaterials.3.040801](https://doi.org/10.1103/PhysRevMaterials.3.040801). arXiv: [1904.05605](https://arxiv.org/abs/1904.05605). URL: <https://journals.aps.org/prmaterials/abstract/10.1103/PhysRevMaterials.3.040801><https://link.aps.org/doi/10.1103/PhysRevMaterials.3.040801>.

- [166] Guillermo Román-Pérez and José M. Soler. “Efficient Implementation of a van der Waals Density Functional: Application to Double-Wall Carbon Nanotubes.” In: *Phys. Rev. Lett.* 103.9 (2009), p. 096102. ISSN: 0031-9007. DOI: [10.1103/PhysRevLett.103.096102](https://doi.org/10.1103/PhysRevLett.103.096102). arXiv: [0812.0244](https://arxiv.org/abs/0812.0244). URL: <https://journals.aps.org/prl/abstract/10.1103/PhysRevLett.103.096102><https://link.aps.org/doi/10.1103/PhysRevLett.103.096102>.
- [167] Nicolas Roux and Y. Bengio. “1 Representational Power of Restricted Boltzmann Machines and Deep Belief Networks.” In: *Neural computation* 20 (July 2008), pp. 1631–49. DOI: [10.1162/neco.2008.04-07-510](https://doi.org/10.1162/neco.2008.04-07-510).
- [168] Matthias Rupp. “Machine learning for quantum mechanics in a nutshell.” en. In: *International Journal of Quantum Chemistry* 115 (Aug. 2015). DOI: [10.1002/qua.24954](https://doi.org/10.1002/qua.24954).
- [169] Matthias Rupp, Alexandre Tkatchenko, Klaus-Robert Müller, and O. Anatole von Lilienfeld. “Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning.” In: *Phys. Rev. Lett.* 108 (5 2012), p. 058301. DOI: [10.1103/PhysRevLett.108.058301](https://doi.org/10.1103/PhysRevLett.108.058301). URL: <https://link.aps.org/doi/10.1103/PhysRevLett.108.058301>.
- [170] J. J. Sakurai and Jim Napolitano. *Modern Quantum Mechanics*. 2nd ed. Cambridge University Press, 2017. DOI: [10.1017/9781108499996](https://doi.org/10.1017/9781108499996).
- [171] Benjamin Sanchez-Lengeling and Alán Aspuru-Guzik. “Inverse molecular design using machine learning: Generative models for matter engineering.” In: *Science* 361.6400 (2018), pp. 360–365.
- [172] Huziel E Saucedo, Stefan Chmiela, Igor Poltavsky, Klaus-Robert Müller, and Alexandre Tkatchenko. “Molecular force fields with gradient-domain machine learning: Construction and application to dynamics of small molecules with coupled cluster forces.” In: *The Journal of chemical physics* 150.11 (2019), p. 114102.
- [173] Tobias Schäfer, Alejandro Gallo, Andreas Irmler, Felix Hummel, and Andreas Grüneis. “Surface science using coupled cluster theory via local Wannier functions and in-RPA-embedding: the case of water on graphitic carbon nitride.” In: *The Journal of Chemical Physics* 155.24 (2021), p. 244103.
- [174] Michael Scherbela, Rafael Reisenhofer, Leon Gerard, Philipp Marquetand, and Philipp Grohs. “Solving the electronic Schrödinger equation for multiple nuclear geometries with weight-sharing deep neural networks.” In: *Nature Computational Science* 2.5 (2022), pp. 331–341.
- [175] Sheila K. Schiferl and Duane C. Wallace. “Statistical errors in molecular dynamics averages.” In: *J. Chem. Phys.* 83.10 (1985), pp. 5203–5209. ISSN: 0021-9606. DOI: [10.1063/1.449733](https://doi.org/10.1063/1.449733). URL: <http://aip.scitation.org/doi/10.1063/1.449733>.

- [176] Christoph Schran, Fabien Briec, and Dominik Marx. "Converged colored noise path integral molecular dynamics study of the zundel cation down to ultralow temperatures at coupled cluster accuracy." In: *Journal of chemical theory and computation* 14.10 (2018), pp. 5068–5078.
- [177] Maria Schuld, Ilya Sinayskiy, and Francesco Petruccione. "The quest for a quantum neural network." In: *Quantum Information Processing* 13 (2014), pp. 2567–2586.
- [178] Mutasem Omar Sinnokrot, Edward F Valeev, and C David Sherrill. "Estimates of the ab initio limit for π - π interactions: The benzene dimer." In: *Journal of the American Chemical Society* 124.36 (2002), pp. 10887–10893.
- [179] Justin S Smith, Roman Zubatyuk, Benjamin Nebgen, Nicholas Lubbers, Kipton Barros, Adrian E Roitberg, Olexandr Isayev, and Sergei Tretiak. "The ANI-1ccx and ANI-1x data sets, coupled-cluster and density functional theory properties for molecules." In: *Scientific data* 7.1 (2020), p. 134.
- [180] S. Sorella, M. Casula, and D. Rocca. "Weak binding between two aromatic rings: Feeling the van der Waals attraction by quantum Monte Carlo methods." In: *J. Chem. Phys.* 127.1 (July 2007), p. 014105.
- [181] Annika Stuke, Milica Todorović, Matthias Rupp, Christian Kunkel, Kunal Ghosh, Lauri Himanen, and Patrick Rinke. "Chemical diversity in molecular orbital energy predictions with kernel ridge regression." In: *J. Chem. Phys.* 150.20 (2019), p. 204121. ISSN: 00219606. DOI: [10.1063/1.5086105](https://doi.org/10.1063/1.5086105). arXiv: [1812.08576](https://arxiv.org/abs/1812.08576). URL: <http://aip.scitation.org/doi/10.1063/1.5086105>.
- [182] Jianwei Sun, Adrienn Ruzsinszky, and Johnp Perdew. "Strongly Constrained and Appropriately Normed Semilocal Density Functional." In: *Phys. Rev. Lett.* 115.3 (2015), p. 036402. ISSN: 0031-9007. DOI: [10.1103/PhysRevLett.115.036402](https://doi.org/10.1103/PhysRevLett.115.036402). URL: <https://journals.aps.org/prl/abstract/10.1103/PhysRevLett.115.036402><https://link.aps.org/doi/10.1103/PhysRevLett.115.036402>.
- [183] Jianwei Sun et al. "Accurate first-principles structures and energies of diversely bonded systems from an efficient density functional." In: *Nat. Chem.* 8.9 (2016), pp. 831–836. ISSN: 1755-4330. DOI: [10.1038/nchem.2535](https://doi.org/10.1038/nchem.2535). URL: <https://www.nature.com/articles/nchem.2535><http://www.nature.com/articles/nchem.2535>.
- [184] A. Szabo and N.S. Ostlund. *Modern Quantum Chemistry: Introduction to Advanced Electronic Structure Theory*. Dover Books on Chemistry. Dover Publications, 1996. ISBN: 9780486691862. URL: <https://books.google.fr/books?id=k-DcCgAAQBAJ>.

- [185] Tait Takatani, Edward G Hohenstein, Massimo Malagoli, Michael S Marshall, and C David Sherrill. "Basis set consistent revision of the S22 test set of noncovalent interaction energies." In: *The Journal of chemical physics* 132.14 (2010), p. 144104.
- [186] Giacomo Torlai, Guglielmo Mazzola, Juan Carrasquilla, Matthias Troyer, Roger Melko, and Giuseppe Carleo. "Neural-network quantum state tomography." In: *Nature Physics* 14.5 (2018), pp. 447–450.
- [187] Julien Toulouse, Roland Assaraf, and Cyrus J. Umrigar. "Chapter Fifteen - Introduction to the Variational and Diffusion Monte Carlo Methods." In: *Electron Correlation in Molecules – ab initio Beyond Gaussian Quantum Chemistry*. Ed. by Philip E. Hoggan and Telhat Ozdogan. Vol. 73. Advances in Quantum Chemistry. Academic Press, 2016, pp. 285–314. DOI: <https://doi.org/10.1016/bs.aiq.2015.07.003>. URL: <https://www.sciencedirect.com/science/article/pii/S0065327615000386>.
- [188] Theodoros Tsatsoulis, Sung Sakong, Axel Groß, and Andreas Grüneis. "Reaction energetics of hydrogen on Si (100) surface: A periodic many-electron theory study." In: *The Journal of Chemical Physics* 149.24 (2018), p. 244105.
- [189] Oliver T Unke, Stefan Chmiela, Huziel E Saucedo, Michael Gastegger, Igor Poltavsky, Kristof T Schutt, Alexandre Tkatchenko, and Klaus-Robert Müller. "Machine learning force fields." In: *Chemical Reviews* 121.16 (2021), pp. 10142–10186.
- [190] Laurens Van Der Maaten and Geoffrey Hinton. *Visualizing Data using t-SNE*. Tech. rep. 2008, pp. 2579–2605.
- [191] Veronique Van Speybroeck, Karen Hemelsoet, Lennart Joos, Michel Waroquier, Robert G Bell, and C Richard A Catlow. "Advances in theory and their application within the field of zeolite chemistry." In: *Chemical Society Reviews* 44.20 (2015), pp. 7044–7111.
- [192] Peter Wirnsberger, Andrew J. Ballard, George Papamakarios, Stuart Abercrombie, Sébastien Racanière, Alexander Pritzel, Danilo Jimenez Rezende, and Charles Blundell. "Targeted free energy estimation via learned mappings." In: *J. Chem. Phys.* 153.14 (2020), p. 144112. ISSN: 0021-9606. DOI: [10.1063/5.0018903](https://aip.scitation.org/doi/abs/10.1063/5.0018903). URL: <https://aip.scitation.org/doi/abs/10.1063/5.0018903><https://aip.scitation.org/doi/10.1063/5.0018903>.
- [193] Sebastian Wouters, Carlos A. Jiménez-Hoyos, Qiming Sun, and Garnet K.-L. Chan. "A Practical Guide to Density Matrix Embedding Theory in Quantum Chemistry." In: *Journal of Chemical Theory and Computation* 12.6 (2016). PMID: 27159268, pp. 2706–2719. DOI: [10.1021/acs.jctc.6b00316](https://doi.org/10.1021/acs.jctc.6b00316). eprint: <https://doi.org/10.1021/acs.jctc.6b00316>. URL: <https://doi.org/10.1021/acs.jctc.6b00316>.

- [194] Linfeng Zhang, Jiequn Han, Han Wang, Roberto Car, and Weinan E. “Deep Potential Molecular Dynamics: A Scalable Model with the Accuracy of Quantum Mechanics.” In: *Phys. Rev. Lett.* 120.14 (2018), p. 143001. ISSN: 0031-9007. DOI: [10.1103/PhysRevLett.120.143001](https://doi.org/10.1103/PhysRevLett.120.143001). URL: <https://journals.aps.org/prl/abstract/10.1103/PhysRevLett.120.143001><https://link.aps.org/doi/10.1103/PhysRevLett.120.143001>.

RÉSUMÉ

La précision chimique, définie comme une kilocalorie par mole, est un objectif de résolution énergétique pour les méthodes computationnelles en science des matériaux. Alors que les approximations les plus simples en chimie quantique (Hartree-Fock) et en physique du solide (théories fonctionnelles de la densité locales et semi-locales (DFT)) sont numériquement abordables pour des systèmes modérément grands (jusqu'à quelques milliers d'atomes, avec un coût numérique typiquement en $\mathcal{O}(K^4)$ et $\mathcal{O}(K^3)$ respectivement avec la taille de la base), elles ne peuvent pas être améliorés de manière systématique et ne parviennent pas à la précision chimique dans beaucoup d'applications. Il faut alors recourir à des approximations plus précises – méthodes post-Hartree-Fock telles que Configuration Interaction, la théorie des perturbations Mollër-Plesset, Coupled Cluster, ou des modèles au-delà de la DFT tels que la Random Phase Approximation (RPA) – mais ces dernières sont numériquement plus coûteuses et donc limitées à de plus petits systèmes. En particulier, alors que des implémentations périodiques de Mollër-Plesset ou Coupled Cluster ont été récemment proposées, les résultats à température finie (qui nécessitent quelques centaines de milliers de configurations calculées) sont toujours inatteignables pour les systèmes de matière condensée utilisant ces théories. Dans ce contexte, les méthodes d'apprentissage automatique ont connu au cours de la dernière décennie un important développement afin de réduire le coût numérique des méthodes de mécanique quantique. Parmi ces méthodes de machine learning (ML), nous nous intéresseront tout particulièrement dans cette thèse à la Machine Learning Perturbation Theory (MLPT). Cet algorithme repose sur deux ingrédients : le Δ -Machine Learning et la théorie de la perturbation thermodynamique. L'idée du Δ -Machine Learning est la suivante : plutôt que d'apprendre directement l'énergie (ou une autre propriété) en fonction de la configuration atomique, il est plus facile d'apprendre la différence d'énergie telle que donnée par deux approximations quantiques différentes. En considérant deux fonctionnelles d'échange et corrélation DFT différentes par exemple, les deux approximations incluent le même terme cinétique et diffèrent dans leur traitement de l'énergie de corrélation. Apprendre la différence entre les deux approximations est plus facile que d'apprendre directement toute la description physique contenue dans l'une des deux. Ainsi, si plusieurs milliers de configurations sont nécessaires pour apprendre la surface d'énergie potentielle d'un système complexe, ce nombre peut être réduit à quelques dizaines ou centaines grâce à cette technique. Dans la théorie des perturbations thermodynamiques maintenant, on peut exprimer la moyenne d'ensemble canonique d'un hamiltonien en fonction de moyennes d'un

autre hamiltonien, si la différence entre les deux hamiltoniens est connue. Plus explicitement, étant donnée une dynamique moléculaire réalisée avec une approximation X, on peut écrire la moyenne telle qu'elle aurait été obtenue avec une autre approximation Y si, pour chaque configuration de la dynamique moléculaire, la différence d'énergie entre les approximations X et Y est connue.

La technique MLPT, introduite en 2019 par Chehaibou *et al.*, utilise la théorie de la perturbation thermodynamique pour repondérer une dynamique moléculaire *ab initio*, réalisée avec une référence DFT semi locale, pour obtenir des moyennes d'ensemble RPA, et ainsi pouvoir calculer l'enthalpie d'adsorption du dioxyde de carbone et du méthane dans une zéolithe. Compte tenu du coût de calcul RPA, quelques dizaines de configurations ont été extraites des dynamiques moléculaire pour entraîner un modèle de Δ -Machine Learning et ainsi apprendre la différence d'énergie entre l'énergie DFT et RPA. Un travail ultérieur similaire a permis de calculer des énergies libres d'activation pour une réaction d'échange de proton dans une chabazite protonée, également au niveau RPA. Le travail présenté dans cette thèse se propose de continuer le développement de cette méthode MLPT. D'une part, la théorie de la perturbation thermodynamique repose sur l'égalité des espaces configurationnels accessibles aux théories de référence X et cible Y. En pratique, les dynamiques moléculaires numériques produisent des échantillons discrets, proches des positions d'équilibre associées au hamiltonien utilisé. Si ces configurations d'équilibre sont trop distantes entre les théories de référence et cible, l'estimation issue de la repondération risque d'être biaisée. Ce phénomène est étudié dans cette thèse : un test de diagnostic est proposé pour détecter ces cas problématiques, ainsi qu'une solution sous la forme d'un ré-échantillonnage Monte Carlo. Ce dernier utilise l'algorithme de Metropolis, mais le modèle de Δ -Machine Learning est réutilisé ici pour évaluer l'énergie au niveau cible en la calculant d'abord au niveau de référence. Ainsi on s'assure de procéder à un échantillonnage de l'espace configurationnel de la théorie cible et l'on peut retrouver des estimations non biaisées. Pour procéder à cette analyse, cinq fonctionnelles DFT différentes sont utilisées, pour lesquelles on dispose à chaque fois d'une dynamique de référence. La méthode MLPT est appliquée entre chaque couple parmi les cinq approximations. Le test de diagnostic permet bien de détecter les cas pathologiques, et le ré-échantillonnage Monte Carlo redonne des valeurs correctes dans ces cas là, à la précision chimique. Ce ré-échantillonnage est également réalisé pour confirmer de précédents calculs RPA, et le test de diagnostic semble indiquer un recouvrement configurationnel satisfaisant entre la fonctionnelle semi locale de Perdew-Burke-Ernzerhof plus une correction semi-empirique pour les interactions van der Waals (PBE+D2), et la RPA.

Dans un second temps, on utilisera cette méthodologie pour calculer l'enthalpie d'adsorption du dioxyde de carbone dans une zeolite, au niveau de la théorie du cluster couplé, avec excitations simples, doubles, et triples perturbatives (CCSD(T)).

Cette dernière est généralement considérée comme la méthode de référence en chimie quantique, permettant d'obtenir la précision chimique dans la plupart des applications. Son coût numérique est cependant bien trop élevé pour des applications à température finie en matière condensée, et l'on présente ici la première estimation CCSD(T) d'une propriété à température finie dans un matériau périodique.

Les travaux mentionnés jusqu'à présent nécessitent dans le cadre de l'entraînement d'un modèle d'apprentissage automatique qu'un certain nombre de calculs coûteux aient été réalisés au préalable. Il pourrait être avantageux d'utiliser le machine learning directement durant le processus de résolution de l'équation de Schrödinger. En 2017, Carleo et Troyer ont introduit l'idée des états quantiques neuronaux (NQS). Dans le cadre du Monte Carlo variationnel quantique, on utilise des ansatzes de fonction d'onde qui sont optimisées variationnellement. Dans ce contexte, Carleo et Troyer ont suggéré d'utiliser plutôt le pouvoir représentatif des réseaux de neurones pour paramétrer la fonction d'onde à optimiser. Plus précisément, ils choisissent une machine de Boltzmann restreinte (RBM) et optimisent ses paramètres au moyen d'un algorithme provenant des méthodes de Monte Carlo variationnel quantique. Cet ansatz neuronal est appliqué à des modèles de spins mais a ensuite été étendu aux problèmes de structure électronique par Choo *et al.* : en reformulant l'hamiltonien de structure électronique en un hamiltonien de spin (en utilisant la transformation de Jordan-Wigner par exemple), on peut retrouver la théorie des interactions de configuration (CI). Ils ont effectué des calculs d'énergie électronique sur de petites molécules (jusqu'à quatorze électrons) dans les bases STO-3G et 6-31G. Bien que la convergence des calculs à la précision chimique était observée dans la base STO-3G, ça n'était pas le cas pour la base 6-31G. Il apparaît encore difficile d'élaborer des algorithmes efficaces pour optimiser un réseau de neurones - de surcroît avec des poids complexes - sur le problème CI. L'échantillonnage est compliqué par un poids prédominant des configurations les plus importantes dans la fonction d'onde, entraînant un ralentissement problématique de l'algorithme pour l'ensemble de base non minimal 6-31G. Une autre difficulté pour le RBM ou d'autres modèles simples est la nature très complexe des corrélations entre l'occupation des orbitales dans les configurations. Par modèle simple, on entend un modèle avec peu de paramètres, *e.g.* un nombre similaire de variables cachées et d'entrée, car le pouvoir représentatif des réseaux de neurones augmente avec le nombre de ces variables cachées. Une solution consiste à utiliser des réseaux de neurones profonds, par exemple en empilant des RBM, une couche cachée étant une entrée pour la couche cachée suivante (machine de Boltzmann restreinte profonde). Des fonctions d'ondes spatiales ont été codées en tant que NQS dans de telles architectures d'apprentissage profond au cours des quatre dernières années, surpassant l'état de l'art précédent des ansatzes varia-

tionnels QMC, avec des systèmes jusqu'à trente électrons. Les fonctions d'onde CI tronquées restent cependant une méthode précieuse pour obtenir une précision systématique et calibrer d'autres méthodes. Étant donné que l'espace CI croît de manière combinatoire avec la taille du système, les algorithmes CI antérieurs sélectionnent généralement les configurations les plus importantes en utilisant la théorie des perturbations ou un échantillonnage aléatoire, avant une diagonalisation du sous-espace de Hilbert, dans un processus itératif. L'apprentissage automatique a été proposé comme moyen de remplacer la théorie des perturbations comme critère de sélection des configurations importantes par Coe *et al.* en 2018, en utilisant un réseau de neurones régressif pour apprendre les configurations importantes. Dans cette thèse, on propose une voie hybride. On se place dans le même type d'algorithmes itératifs, avec un échantillonnage aléatoire. Mais on remplace la distribution de probabilité uniforme pour la proposition de nouveaux déterminants par une distribution contenant de l'information quant aux déterminants actuellement dans la fonction d'onde. Ceci est réalisé en entraînant une machine de Boltzmann restreinte à reproduire les probabilités de Born associées aux coefficients des déterminants de la fonctions d'onde CI. On échantillonne ensuite de nouveaux déterminants à partir de ce modèle, avant diagonalisation. Contrairement aux états quantiques neuronaux, le réseau de neurone ne sert pas d'ansatz à la fonction d'onde, mais au carré de son amplitude, qui est donc une amplitude de probabilité. Le modèle RBM permettant l'échantillonnage, on peut donc proposer de nouvelles configurations plus probables qu'avec un échantillonnage non informé. Ce modèle est testé sur de petites molécules (jusqu'à quatorze électrons) dans les bases 6-31G et cc-PVDZ, et l'on montre que la convergence à la précision chimique est plus rapide qu'avec un Monte Carlo traditionnel.

Cette thèse est donc une contribution au développement de nouvelles méthodologies d'apprentissage automatique afin de réduire le coût de calcul nécessaire à l'obtention de propriétés chimiquement précises. Après une introduction aux problématiques considérées ici dans le premier chapitre, le chapitre deux procède à une revue du problème à N électrons et des méthodes considérées dans cette thèse, avec une première approximation appelée Hartree-Fock. Les méthodes dites post-Hartree Fock sont étudiées, notamment les interactions de configuration, la théorie du cluster couplé, et la perturbation Møller-Plesset. La théorie de la fonctionnelle de la densité est alors introduite, ainsi que la RPA. Dans le chapitre trois on s'intéressera aux différents concepts et méthodes d'apprentissage automatique utilisées dans le reste de ce manuscrit. Les chapitres quatre, cinq et six constituent les contributions de cette thèse. Le chapitre quatre est une étude de la méthode MLPT et du problème de recouvrement configurationnel mentionné plus haut. On y présente le ré-échantillonnage Monte Carlo comme solution. Le chapitre cinq applique ces méthodes (MLPT et ré-échantillonnage Monte Carlo) au calcul d'une

estimation de l'enthalpie d'adsorption du dioxyde de carbone dans une zeolite au niveau CCSD(T). Le chapitre six est consacré à un modèle génératif pour proposer de nouvelles configurations dans le cadre de la méthode CI. Enfin le septième et dernier chapitre discute les résultats de cette thèse et conclut.