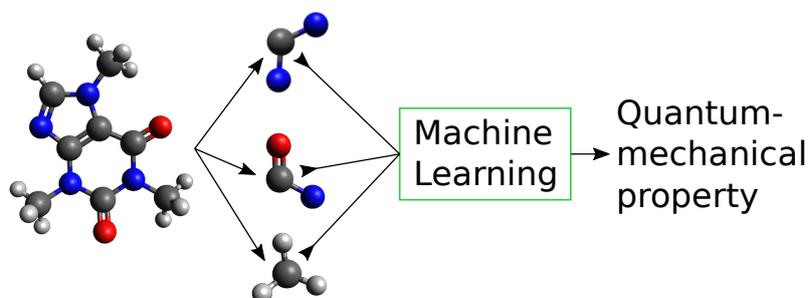


Towards more Efficient and Performant Computations in Quantum Chemistry with Machine Learning

Wiktor Pronobis

Supervisor: Prof. Klaus-Robert Müller

March 26, 2020



Abstract

Towards more Efficient and Performant Computations in Quantum Chemistry with Machine Learning Kernel methods allow an efficient solution of highly non-linear regression problems often encountered in quantum chemistry. Due to its flexibility it is unclear how to design a similarity matrix represented by the kernel which encodes a given learning problem in a compact and beneficial way. In this thesis, we propose novel kernels for quantum mechanical systems which are composed of two- and three-body interaction terms. Specifically, we develop descriptors of molecules which are of fixed size and invariant with respect to translation, rotation and atom indexing. For these representations, we demonstrate their ability to accurately predict quantum mechanical properties in combination with kernel ridge regression. A feature importance analysis reveals insights about the two- and three-body interactions in small organic molecules. Our descriptors are extended by novel decomposition kernels which encode the comparison of two- and three-body combinations of atoms directly into the similarity matrix. The special structure of these kernels is used to analyse interaction potentials for molecular dynamics data sets. These kernel methods are complemented by a new approximate matrix inversion scheme based on banded Toeplitz matrices. For all of these three methodologies, we demonstrate their efficiency and performance to solve quantum chemical problems.

Zusammenfassung

Richtung effizienteren und performanten Berechnungen in der Quantenchemie mit Maschinellern Lernen Kern-basierte Methoden erlauben es, effizientere Lösungen von hoch nichtlinearen Regressionsproblemen zu entwickeln, die häufig in der Quantenchemie auftreten. Wegen ihrer Flexibilität, ist es unklar, wie man eine kern-basierte Ähnlichkeitsmatrix aufstellt, die das Lernproblem in einer kompakten und nutzbaren Weise kodiert. In dieser Arbeit schlagen wir neue Kerne für quantenmechanische Systeme vor, die auf Zwei- und Dreikörperinteraktionstermen beruhen. Genauer entwickeln wir Deskriptoren von Molekülen, die konstante Größe haben und invariant sind gegenüber der Translation, Rotation und Atomindizierung. Für diese Repräsentationen demonstrieren wir die akkurate Vorhersage von quantenmechanischen Eigenschaften in Kombination mit Gaußprozess-Regression. Eine Wichtigkeitsanalyse der Features enthüllt Einsichten über die Zwei- und Dreikörperinteraktionen in kleinen organischen Molekülen. Unsere Deskriptoren werden weiterentwickelt zu zerlegbaren Kernen, die den Vergleich zwischen Zwei- und Dreierkörperkombinationen von Atomen direkt in der Ähnlichkeitsmatrix kodieren. Die spezielle Struktur dieser Kerne wird genutzt um die Interaktionspotentiale für moleküldynamik Datensätze zu untersuchen. Unsere Kernmodelle werden ergänzt durch ein approximiertes Matrix-Invertierungsverfahren für schmalbandige Toeplitzmatrizen. Für alle diese drei Ansätze zeigen wir deren Effizienz und Leistungsfähigkeit, quantenchemische Probleme zu lösen.

Acknowledgements

First and foremost, I thank Klaus-Robert Müller for his support and inspiration in the intriguing intersection field of machine learning and physics. Klaus encouraged me to investigate my methods in a most fundamental way with the aim of extracting new insights which is crucial for trying to better understand nature. Klaus allowed me the freedom to study my ideas in a variety of different topics, letting me widely explore my interests and introduced me to the signal processing field. Such a freedom is key to truly pioneering research.

I thank Alexandre Tkatchenko for the invaluable discussions and inspiration which helped directing my objectives in a most elementary and principled way.

I especially thank Kristof Schütt for the countless invaluable interesting debates over the years and the outstanding and fruitful collaboration.

I thank Stefan Chmiela for the discussions and priceless feedback.

I thank Shinichi Nakajima for the exceptional collaboration and ongoing guidance.

I thank all my co-authors, notably Danny Panknin, Johannes Kirschnick, Vignesh Srinivasan, Wojciech Samek, Volker Markl and Manohar Kaul for their support and amazing collaborating experience. Special thanks to Stefan Chmiela, Oliver Unke, Mikolaj Czuchaj and Tomasz Pronobis for the excellent proof-reading of this thesis.

I preeminently thank Nino Ushikishvili for her invaluable support, backing and accompaniment along my way through the last years of this work.

I thank all my colleagues of the ML group at TU Berlin, especially my office mate over the years Danny Panknin.

Finally, I thank parents, Maria and Ryszard Pronobis, for their infinite support and advice.

Contents

1	Introduction	1
1.1	Theoretical background	2
1.2	Integration of machine learning with quantum chemistry	4
1.3	Description of the chapters	9
1.4	Previously published work	10
2	Many-body descriptors	13
2.1	Representations of physical systems	13
2.2	Invariant molecular many-body descriptors	16
2.3	Tests on molecular data sets	19
2.4	Feature importance analysis	23
2.5	Summary and discussion	29
3	Capturing intensive and extensive molecular properties with machine learning	33
3.1	Methods	34
3.2	Experiments	35
3.3	Summary and discussion	37
4	Kernel representations of quantum mechanical systems	41
4.1	Local invariant kernels	42
4.2	Tests on molecular data sets	45
4.3	Summary and discussion	56
5	Approximate banded Toeplitz matrix inversion	67
5.1	Method	68
5.2	Main theorems	72
5.3	Proofs	74
5.4	Analytic solution of the tridiagonal case	78
5.5	Time complexity experiments	80
5.6	Constructing Green's functions	81

CONTENTS

5.7 Banded approximation of deconvolution operators	83
5.8 Many-body van der Waals interactions	85
5.9 Interpolation of potential energy surfaces	90
5.10 Summary and discussion	93
6 Conclusions	99
A Supplemental results	101
A.1 Image filtering experiments	101
References	105

Introduction

The intersection of quantum physics and chemistry defines the underlying principles for understanding the structure and properties of matter in everyones lives. A profound knowledge about the behaviour of organic substances would help in targeting the design of new compounds for applications in medicine, the creation of new renewable power sources and fighting hunger. As the number of potentially useful materials grows exponentially with increasing system size, it is important to develop both accurate and efficient methods to predict the desired properties of given substances. Although the mathematical description of these systems is well understood since the start of quantum physics more than 100 years ago, this turns out to be a surprisingly difficult task, in spite of the exponential growth of hardware resources.

The goal of this thesis is to develop alternative tools to aid in the understanding of matter. Our methods will be characterized by being highly efficient at the cost of a certain level of approximation while still being performant at a specific degree of accuracy. Following recent trends, we will develop methods for predicting the properties of quantum mechanical systems directly with machine learning. More precisely, we will rely on kernels for regression analysis which we find a suitable choice for the highly non-linear functions typically encountered in quantum chemistry. As a key underlying concept, we will use decompositions of the studied systems into smaller computational units, specifically two- and three-body combinations of atoms composing a given compound. One of the central properties of our methods will be the invariance with respect to the translation, rotation and atom indexing. While encoding relatively little chemical knowledge, we will investigate whether our models are in accordance with chemical intuition.

One of our approaches will allow to explicitly construct Green's functions of differential operators of arbitrary order. These Green's functions play a crucial role in the understanding of quantum mechanical systems. We will achieve this by introducing a novel matrix inversion scheme. This inversion scheme will be applied to solve quantum-mechanical problems, specifically to compute long-

range van der Waals interactions and for the interpolation of potential energy surfaces more efficiently. Before diving into the details, we provide the basic knowledge needed to understand the main motivation of our approaches and how we calculated some of the targeted quantum mechanical properties to be predicted.

1.1 Theoretical background

A quantum mechanical system is composed of a set of elementary units called atoms identified by the atomic number $Z \in \mathbb{N}$ and position in three-dimensional space $\mathbf{R} \in \mathbb{R}^3$ and can be represented by the set $S = \{(Z_i, \mathbf{R}_i)\}_{i=1}^N$ where N is the number of atoms of the system. Such a system is described by the nonrelativistic time-dependent Schrödinger equation

$$i\hbar \frac{\partial}{\partial t} \Psi(\mathbf{r}, t) = \hat{\mathcal{H}} \Psi(\mathbf{r}, t) \quad (1.1)$$

with the Hamilton operator

$$\hat{\mathcal{H}} = -\frac{\hbar}{2m} \nabla^2 + V(\mathbf{r}, t) \quad (1.2)$$

where the wave function $\Psi(\mathbf{r}, t)$ describes a single particle of mass m in an external potential $V(\mathbf{r}, t)$, i is the imaginary unit and \hbar is the reduced Planck constant, respectively. The time-dependent Schrödinger equation (1.1) can be solved in terms of stationary states of the form

$$\hat{\mathcal{H}} \Psi(\mathbf{r}) = E \Psi(\mathbf{r}) \quad (1.3)$$

where E is the energy of the state $\Psi(\mathbf{r})$. For the wave function, a common interpretation is $|\Psi(\mathbf{r})|^2$ as the probability of locating the particle at position \mathbf{r} . In the Born-Oppenheimer approximation, the positions of the nuclei of the atoms are fixed and the problem is formulated by the n electron coordinates, replacing $\Psi(\mathbf{r})$ with $\Psi(\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_n)$. A key property of the system S is its ground state energy which can be defined by the variational principle

$$E_{\text{ground}} = \min_{\Psi(\cdot)} \langle \Psi(\cdot) | \hat{\mathcal{H}} | \Psi(\cdot) \rangle \quad \text{with} \quad \langle \Psi(\cdot) | \Psi(\cdot) \rangle = 1 \quad (1.4)$$

where $\langle \cdot | \cdot \rangle$ denotes the expectation value. Solving the Schrödinger equation for systems containing a number of particles greater than one is generally infeasible due to the high-dimensional nature and non-separability of the problem. To circumvent this problem, various types of approximations have been developed to obtain an accurate estimate of the ground state energy of the system.

Density-functional theory

Perhaps the most popular approximation of the Schrödinger equation is provided by the density-functional theory where the key player is the electron

density defined by

$$n(\mathbf{r}) = n \int d^3r_2 \cdots \int d^3r_n \Psi^*(\mathbf{r}, \mathbf{r}_2, \cdots, \mathbf{r}_n) \Psi(\mathbf{r}, \mathbf{r}_2, \cdots, \mathbf{r}_n) \quad (1.5)$$

where $\Psi^*(\cdot)$ denotes the complex conjugate of $\Psi(\cdot)$. As it turns out, the energy minimization in Eq. (1.4) can be carried out in terms of this density $n(\mathbf{r})$, thereby reducing the number of free variables from $3n$ to 3. This allows to formulate the many-body interactions by a set of much simpler single-particle problems, the so-called Kohn-Sham equations which are given by

$$\left[-\frac{\hbar}{2m} \nabla^2 + V_s(\mathbf{r}) \right] \varphi_i(\mathbf{r}) = \varepsilon_i \varphi_i(\mathbf{r}) \quad \text{for } i = 1, \cdots, n \quad (1.6)$$

$$V_s(\mathbf{r}) = V(\mathbf{r}) + e^2 \int \frac{n(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} d^3r' + V_{\text{XC}}[n(\mathbf{r})] \quad (1.7)$$

with the elementary charge e , the exchange-correlation potential V_{XC} and

$$n(\mathbf{r}) = \sum_{i=1}^n |\varphi_i(\mathbf{r})|^2 \quad (1.8)$$

The Kohn-Sham equations are usually solved in an iterative way, where $V_s(\mathbf{r})$ is calculated given the electron density $n(\mathbf{r})$ in Eq. (1.7) and reversely the potential $V_s(\mathbf{r})$ defines the spectrum of energies and single-particle states ε_i and $\varphi_i(\mathbf{r})$ in Eq. (1.6). Most of the properties of quantum mechanical systems in this work have been calculated using density-functional theory.

Many-body expansion

The ground-state energy is a function of the coordinates of the nuclei composing a given physical system. This dependence can be modeled by a many-body expansion which is a decomposition of the ground-state energy into many-body terms

$$E_{\text{ground}} = \sum_{i=1}^N E_i + \sum_{i=1}^N \sum_{j=i+1}^N E_{ij} + \sum_{i=1}^N \sum_{j=i+1}^N \sum_{k=j+1}^N E_{ijk} + \cdots \quad (1.9)$$

where E_i depends only on the coordinates of atom i , E_{ij} depends on the coordinates of the atoms i and j and E_{ijk} depends on the coordinates of the atoms i , j and k , respectively. Such a decomposition is very flexible in the sense that the many-body terms can be defined for each combination of atom-types individually. Typically, the above expansion is applied for extensive properties which are characterized by growing in magnitude with increasing system size N , an example of which being the atomization energy. For the atomization energy, we will investigate if the inclusion of up to three-body interactions can accurately model a given system under investigation. This assumption will be verified for systems restricted to a specific domain, e.g. stable small organic

molecules and molecular dynamics data sets near equilibrium. Our methods are specifically designed to learn the two- and three-body interaction potentials occurring in Eq. (1.9) with machine learning. These interaction potentials will be analyzed across different models and molecules for their conformity with chemical intuition.

1.2 Integration of machine learning with quantum chemistry

Large parts of this section have appeared in previously published work [1].

Recently, machine learning has been increasingly applied in the quantum mechanical domain in a variety of ways [2, 3, 4, 5, 6, 7, 8]. The many-body expansion from above can be used as the starting point to develop quantum chemistry models based on machine learning. Although machine learning based approaches take input data obtained from highly accurate first principles methods like density-functional theory, the generalization property inherent of these approaches allows to reversely deduce conclusions for the data domain under study. In this work, we will design regression methods based on kernels which can be used to specifically analyse the two- and three-body interactions in chemical compounds. Kernel based methods allow an efficient convex solution of highly non-linear optimization problems often encountered in quantum chemistry. As typical settings for a chemist or physicist include a low number of data points paired with a highly non-linear learning problem, kernel based formulations are considered as suitable and powerful methods of choice. Motivated by the intrinsic efficiency of kernel methods, in this thesis we will develop new kernels encoding local physical environments based on the many-body expansion with the aim of designing accurate and performant models for predicting quantum-mechanical properties of molecules.

In view of these considerations it is important to understand the kernel properties relevant for an efficient solution in a possibly much higher-dimensional, sometimes even unknown feature space induced by the kernel [9]. This is especially true, as it is non-trivial how a kernelized formulation can circumvent the so-called curse of dimensionality [10, 11, 12]. In the next section, we provide insights on how the choice of the kernel helps solving these problems by introducing a function class of limited complexity from which the final model is chosen.

Implicit feature mapping – the kernel trick

Kernel ridge regression (KRR) is one of the most popular methods of nonlinear regression analysis in quantum chemistry. One of the main ingredients of KRR is the representation of the underlying physical system which mainly determines the performance of predicting quantum-mechanical properties based on KRR. Several such representations have been developed for both, solids and molecules; all of them with different advantages and limitations. In the

Chap. 2 of this thesis, we will propose and investigate the importance of invariant two- and three-body descriptors and use these representations to analyse two- and three-body interactions in molecules. These descriptors correspond to a similarity measure between two chemical compounds which is represented by the kernel. As recent approaches define the kernel directly from the underlying physical system, it is important to understand the properties of kernels and how these kernel properties can be used to improve the performance of machine learning models for quantum chemistry.

The second important ingredient of KRR (the first being the representation) is the kernel. But what is a kernel in general and how can it be useful? With a kernel, the data can be nonlinearly mapped onto a feature space, where the learning may become easier and where optimal generalization can be guaranteed. A key concept here is that this mapping can be done *implicitly* by the choice of the kernel. This implicit feature mapping to a possibly much higher dimensional space is very flexible. More intuitively, the kernel encodes a real valued similarity measure between two chemical compounds. This similarity measure is primarily encoded by the representation of the physical system which is then used in combination with standard nonlinear kernel functions like the Gaussian or Laplace kernel. Alternatively, the similarity measure can be encoded *directly* into the kernel, leading to a variety of kernels in the chemistry domain, e.g. for predicting the atomization energy with KRR. We will develop a set of local kernels in the Chap. 4 of this thesis which are designed to compare atomic environments across molecules with each other.

One way to better understand the role of the kernel is to apply existing learning methods in a projected space $\phi : \mathbb{R}^{n_i} \rightarrow \mathbb{R}^{n_o}$ with the input and feature dimension n_i and n_o , respectively. Specifically, it is required that a given algorithm (together with predictions based on this algorithm) works solely on scalar products of type $\mathbf{x}^\top \mathbf{y}$ which can then be translated into scalar products in feature space $\phi(\mathbf{x})^\top \phi(\mathbf{y})$. Then, it turns out that such scalar products in feature space can be done *implicitly*, replacing them with an evaluation of the kernel function $k(\mathbf{x}, \mathbf{y}) := \phi(\mathbf{x})^\top \phi(\mathbf{y})$ [13]. This is known as the kernel trick [14] and interestingly enough, many algorithms can be kernelized this way [15]. Using the kernel trick, one never has to explicitly perform the potentially computationally expensive transformation $\phi(\cdot)$.

The kernel function $k(\cdot)$ thus allows to reduce some of the intrinsic difficulties of the non-linear mapping $\phi(\cdot)$. The question remains, which kernel functions allow for such implicit feature mappings. Mercer’s theorem [16] guarantees that such a mapping exists, if for all elements f of the Hilbert space L^2 defined on a compact set $C \subset \mathbb{R}^{n_i}$

$$\int_C f(\mathbf{x})k(\mathbf{x}, \mathbf{y})f(\mathbf{y})d\mathbf{x}d\mathbf{y} > 0. \quad (1.10)$$

From the kernel and a set of input samples $\{\mathbf{x}_i\}_{i=1}^N$ we can construct a discrete

version of Mercer’s theorem by composing the matrix

$$K := \begin{pmatrix} k(\mathbf{x}_1, \mathbf{x}_1) & \cdots & k(\mathbf{x}_1, \mathbf{x}_N) \\ & \ddots & \\ k(\mathbf{x}_N, \mathbf{x}_1) & \cdots & k(\mathbf{x}_N, \mathbf{x}_N) \end{pmatrix}. \quad (1.11)$$

Mercer’s theorem now implies that the matrix K is a Gram matrix, i.e. positive-semidefinite for any set of inputs $\{\mathbf{x}_i\}_{i=1}^N$. Thus, practically if the matrix K would have negative eigenvalues then it will not fulfill Mercer’s theorem. Examples of popular kernels in the quantum chemistry domain are shown in Tab. 1.1.

Name	Kernel $k(\mathbf{x}, \mathbf{y})$
Gaussian	$\exp(-\ \mathbf{x} - \mathbf{y}\ _2^2 / (2\sigma^2))$
Laplace	$\exp(-\ \mathbf{x} - \mathbf{y}\ _1 / \sigma)$
Polynomial	$(\mathbf{x}^\top \cdot \mathbf{y} + c)^d$
Matérn	$\frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\sqrt{2\nu}}{l} \ \mathbf{x} - \mathbf{y}\ \right)^\nu K_\nu \left(\frac{\sqrt{2\nu}}{l} \ \mathbf{x} - \mathbf{y}\ \right)$

Table 1.1: List of Mercer kernels often used in the quantum chemistry domain. For the Matérn kernel, Γ denotes the Gamma function and K_ν is the modified Bessel function of the second kind, respectively. This table has appeared in previously published work [1].

For some kernels like the Gaussian kernel, the feature map $\phi(\cdot)$ can be infinite dimensional. Due to the *curse of dimensionality* it is then a question whether such a feature mapping to a much higher-dimensional space is a good idea at all, especially as the training set size increases (which corresponds to the dimension of the linear span of the projected input samples in feature space). As it turns out, one can still leverage the feature mapping if the learning algorithm is kept *simple* [15]. The intuitive complexity of the learning problem induced by the kernel, the data and the learning algorithm is a measure of how well a kernel matches the data. Note that translation invariant kernels have natural regularization properties which help reducing the complexity of a learning algorithm. The Gaussian kernel for example is smooth in all its derivatives [17].

While there is a wide variety of representations of physical systems, it is less obvious how to encode prior knowledge into the kernel (see Zien et al. [18] for the first kernels engineered to reflect prior knowledge). In the quantum chemistry domain this is typically done by limiting the similarity measure to *local* information [19, 4]. The definition of such locality depends on the chemical system as it limits correlations between such localized kernels and emphasizes local correlations. Due to the scalar product properties of the mapping $\phi(\cdot)$ in feature space, these local kernels can be combined by a sum to yield a new kernel function.

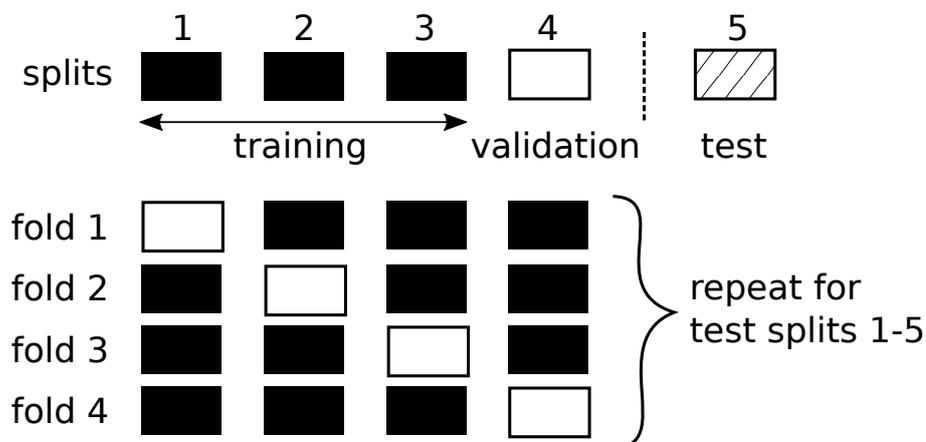


Figure 1.1: Schematic 4-fold cross-validation (inner loop) together with 5-fold nested cross-validation (outer loop), respectively. This image has appeared in previously published work [1].

To conclude this section, we will describe a method for choosing a good kernel among a set of candidate kernels for a given learning problem, a procedure that is commonly called model selection [20]. Typically, a class of kernels is defined by a set of hyperparameters which e.g. control the scaling of the kernel with respect to the data in the chosen distance metric. These hyperparameters have to be determined (i.e. a kernel is selected from a given class) in order to minimize the generalization error, a measure of how good unseen data can be predicted [20]. Note that minimizing a given criterion on the training data alone with respect to the hyperparameters usually results in poor generalization due to *overfitting*. The most common procedure to estimate the generalization error is cross-validation. In cross-validation, the data set is divided into k subsets of equal sample size. Then, the model is trained on the remaining $k - 1$ subsets and evaluated on the k -th subset, called the *validation set*. The average of the error over the k validation sets is a good estimate of the generalization error. After heuristically choosing a set of hyperparameters (kernels), this cross-validation scheme yields the best hyperparameters among the set which are then evaluated on an unseen test set. Repeating this procedure for different test splits is called *nested cross-validation*. Both, cross-validation and nested cross-validation is schematically shown in Fig. 1.1.

Kernel methods

After reviewing key concepts of kernels, we present two practical applications which have been extensively used in the quantum chemistry domain, one for supervised and one for unsupervised learning, respectively.

Kernel ridge regression

A typical setting in machine learning problems includes the prediction of response variables $\{y_i\}_{i=1}^N$ for a set of samples $\{\mathbf{x}_i\}_{i=1}^N$. The kernel trick introduced in the previous section can be applied to the linear ridge regression model. In ridge regression, a cost function typically given by

$$C(\mathbf{w}) := \frac{1}{N} \sum_{i=1}^N (y_i - \mathbf{w}^\top \mathbf{x}_i)^2 + \lambda \cdot \|\mathbf{w}\|^2 \quad (1.12)$$

is minimized with respect to the weight coefficients \mathbf{w} , where λ is a regularization parameter of the model which penalizes the norm of the weights. Given the regularization parameter λ , the weights which minimize Eq. (1.12) are given by

$$\mathbf{w}_{\text{ridge}} = (\lambda \cdot I + X^\top X)^{-1} X^\top \mathbf{y}, \quad (1.13)$$

with the design matrix X whose rows are composed of the inputs $\{\mathbf{x}_i\}_{i=1}^N$ and the identity matrix I , respectively. Increasing the complexity regularizer λ results in smoother functions, thereby avoiding purely interpolating the training data and reducing overfitting (see [20]). Due to its form, the ridge regression model often exhibits good stability in terms of generalization error. However, for most real-world problems the linear model is not powerful enough to accurately predict quantum mechanical properties as it is difficult to find features of the underlying system which linearly correlate with the response variables $\{y_i\}_{i=1}^N$. As the simplicity of the linear ridge regression model turns out to be the main limitation, there is a need for a non-linear variant.

This non-linear variant can be provided by kernelizing the ridge regression model. In kernel ridge regression, the parameters of the model $\boldsymbol{\alpha} := (\alpha_1, \dots, \alpha_N)$ are calculated by

$$(\lambda \cdot I + K) \cdot \boldsymbol{\alpha} = \mathbf{y}, \quad (1.14)$$

with the already introduced Gram matrix K and $\mathbf{y} := (y_1, \dots, y_N)$, respectively. From the parameters $\boldsymbol{\alpha}$, a new prediction for a sample \mathbf{x} is given by

$$y_{\text{est}} = \sum_{i=1}^N \alpha_i \cdot k(\mathbf{x}, \mathbf{x}_i) \quad (1.15)$$

Due to its nice practical and theoretical properties, kernel ridge regression has been extensively used in the quantum chemistry domain [21, 22, 23, 5, 24]. Note that the formally same solution of Eq. (1.14) is also obtained when training Gaussian processes and starting from the framework of Bayesian statistics [25].

Kernel principal component analysis

Kernel principal component analysis (kernel PCA) [26, 27, 28] is a kernelized extension to one of the most popular data dimensionality reduction techniques,

namely principal component analysis (PCA). To recall, PCA is an unsupervised method that uses an orthogonal transformation to project the high dimensional data onto a linearly uncorrelated set of low dimensional variables called the *principal components*. These principal components are defined in a compact way in the sense that a given component accounts for the highest variance under the constraint of being orthogonal to the preceding ones, the first principal component having the largest possible variance.

PCA can be kernelized by virtue of the kernel trick: the evaluation of the data on the m -th principal component equals the m -th eigenvector of the kernel matrix [29]. As PCA requires the data to be centered which is not guaranteed in feature space, one common preliminary step is to centralize the kernel beforehand by

$$K' := K - \mathbf{1}_N \cdot K - K \cdot \mathbf{1}_N + \mathbf{1}_N \cdot K \cdot \mathbf{1}_N, \quad (1.16)$$

where $\mathbf{1}_N$ is the $N \times N$ -matrix with entries $1/N$. From the normalized eigenvectors $\{\mathbf{u}_i\}_{i=1}^N$ of the centralized kernel K' , we compute the m -th principal component of a new sample \mathbf{x} by

$$p_m(\mathbf{x}) = \sum_{i=1}^N \mathbf{u}_{m,i} \cdot k(\mathbf{x}, \mathbf{x}_i), \quad (1.17)$$

where $\mathbf{u}_{m,i}$ is the i -th element of the eigenvector \mathbf{u}_m . Kernel PCA is often used in the quantum chemistry domain to display the data in its first *two* principal components [30, 31, 32, 33], along with the label information if present. Such a projection separates the structure of the data as induced by the kernel from the response variable, possibly learning something about the difficulties to predict a given response variable.

Kernel PCA can be used in a supervised fashion by projecting the label vector \mathbf{y} on the normalized eigenvectors of the centralized kernel matrix

$$z_i := \mathbf{u}_i^\top \mathbf{y} \quad i = 1, \dots, N \quad (1.18)$$

where we call the $\{z_i\}_{i=1}^N$ the *kernel PCA coefficients*. Analyzing the kernel PCA coefficients allows to gain additional information about the complexity of the learning problem at hand [34].

1.3 Description of the chapters

Chapter 2 (Many-body descriptors) We design descriptors of quantum mechanical systems which are invariant to translation, rotation and atom indexing. These representations are used to analyse the importance of two- and three-body interactions in stable small organic molecules.

Chapter 3 (Capturing intensive and extensive molecular properties with machine learning) We analyse the ability of machine learning models

based on our invariant molecular many-body descriptors and an artificial neural network for their ability to predict intensive and extensive quantum-mechanical properties.

Chapter 4 (Kernel representations of quantum mechanical systems)

We develop a similarity measure of quantum mechanical systems based on kernels. The decomposition property of these kernels is used to construct two- and three-body interaction potentials for molecular dynamics data sets.

Chapter 5 (Approximate banded Toeplitz matrix inversion)

We analyse the inverse of banded Toeplitz matrices. For a certain class of these matrices we prove the regularity and develop an efficient algorithm to construct an approximate inverse from the band. This algorithm is implemented reversely, where we estimate the band which best reconstructs a given non-banded Toeplitz matrix. We apply our methods to construct Green's functions of differential operators, to approximate deconvolution operators, to compute long-range van der Waals interactions and for the interpolation of potential energy surfaces, respectively.

The schematic relationship between the chapters of this thesis is depicted in Fig. 1.2. Based on the decomposition of the exemplary ethanol molecule into smaller components, we develop many-body descriptors which are invariant with respect to translation, rotation and atom indexing in Chap. 2. The ability of these representations to predict intensive and extensive molecular properties on equal footing will be subsequently analyzed in greater detail in Chap. 3. In Chap. 4, an alternative approach using the same molecular many-body decompositions as in Chap. 2 is developed, where we construct similarity measures for quantum mechanical systems directly by local kernels. In Chap. 5, we propose an approximate inversion scheme for banded Toeplitz matrices and apply this method to interpolate potential energy surfaces including the ethanol rotor predicted by the local kernels of Chap. 4 and examine whether we can compute another molecular property, namely the polarizabilities including self-consistent electrodynamic screening effects more efficiently. The methods introduced in Chap. 5 can be used for a wider range of applications, which we demonstrate by efficiently constructing Green's functions of differential operators.

1.4 Previously published work

Many results in this thesis have previously been published in journals and books. They are taken from the following articles:

- W. Pronobis, A. Tkatchenko, and K.-R. Müller. "Many-Body Descriptors for Predicting Molecular Properties with Machine Learning: Analysis of Pairwise and Three-Body Interactions in Molecules". *Journal of Chemical Theory and Computation* 14 (6), pp. 2991–3003, 2018

- W. Pronobis, K. T. Schütt, A. Tkatchenko, and K.-R. Müller. "Capturing intensive and extensive DFT/TDDFT molecular properties with machine learning". *The European Physical Journal B* 91 (8), p. 178, 2018
- W. Pronobis, D. Panknin, J. Kirschnick, V. Srinivasan, W. Samek, V. Markl, M. Kaul, K.-R. Müller, and S. Nakajima. "Sharing hash codes for multiple purposes". *Japanese Journal of Statistics and Data Science* 1 (1), pp. 215–246, 2018
- W. Pronobis, and K.-R. Müller. Kernel Methods for Quantum Chemistry. In: *Machine Learning for Quantum Simulations of Molecules and Materials*. Springer Nature, 2020, pp. 27–40
- K. Hansen, F. Biegler, R. Ramakrishnan, W. Pronobis, O. A. von Lilienfeld, K.-R. Müller, and A. Tkatchenko. "Machine Learning Predictions of Molecular Properties: Accurate Many-Body Potentials and Nonlocality in Chemical Space". *The journal of physical chemistry letters* 6 (12), pp. 2326–2331, 2015
- H. Marienwald, W. Pronobis, K.-R. Müller, and S. Nakajima. "Tight Bound of Incremental Cover Trees for Dynamic Diversification". *arXiv preprint arXiv: 1806.06126*, 2018

Figures and tables that are fully or partially taken from previously published work, cite the original source.

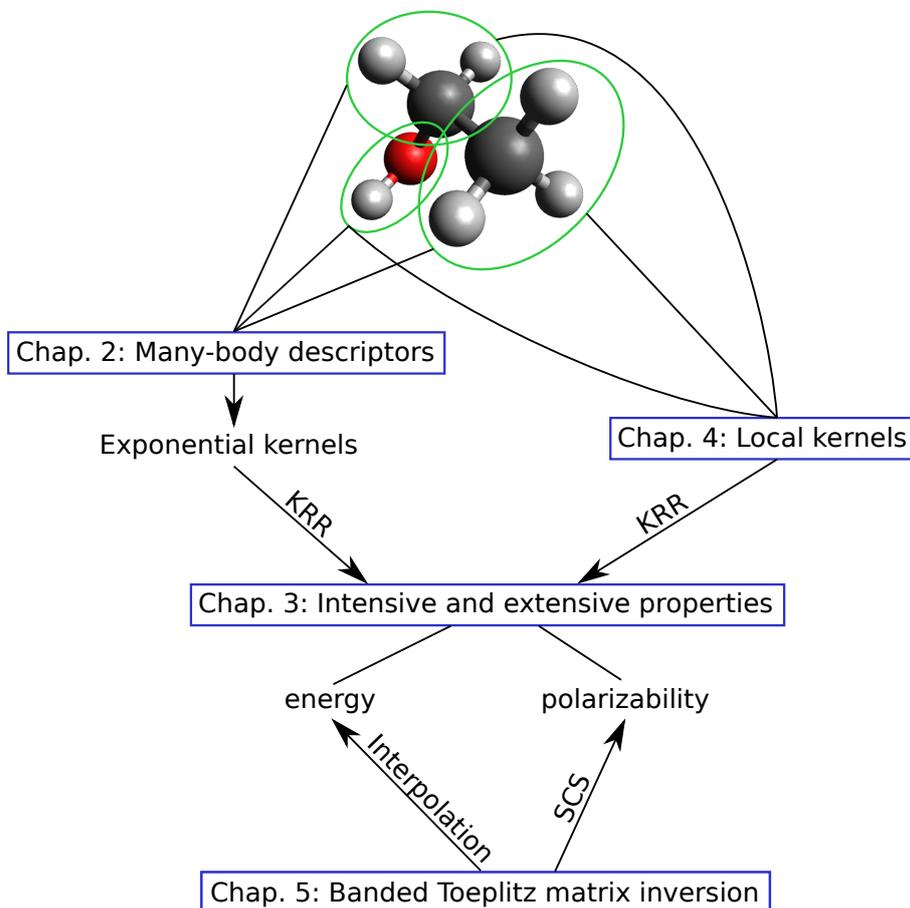


Figure 1.2: Schematic relationship between the chapters of this thesis. Based on the decomposition of the exemplary ethanol molecule (top) into smaller components (green ellipses), we will develop many-body descriptors which are invariant with respect to translation, rotation and atom indexing in Chap. 2. In Chap. 4, these decompositions are used to construct similarity measures for quantum mechanical systems represented directly by local kernels. A set of exponential kernels on top of the descriptors of Chap. 2 and the kernels of Chap. 4 are used to predict quantum-mechanical properties with KRR. The ability of these approaches to predict intensive and extensive molecular properties on equal footing will be analyzed in greater detail in Chap. 3. In Chap. 5, we propose an approximate inversion scheme for banded Toeplitz matrices which we apply to interpolate potential energy surfaces and to compute polarizabilities including self-consistent electrodynamic screening effects more efficiently.

Many-body descriptors

Large parts of this chapter have appeared in previously published work [1, 8].

Recently, machine learning has been ubiquitously used in the industry and sciences. The possibility of parallel implementations using GPU cards in addition to new deep learning architectures has enabled powerful learning machines which reach and even surpass human performance in a variety of applications. From imperfect information games like heads-up no-limit Texas hold'em poker [35] over real-time strategy games like StarCraft [36], the program AlphaGo Zero [37] has been trained without human knowledge and is arguably the strongest Go player in history. Machine learning approaches reach human performance in human interaction tasks like speech recognition [38], image recognition [39] and speech generation [40].

In this chapter, we follow one of the most intriguing applications of machine learning in sciences: the prediction of highly complex properties of quantum mechanical systems. Specifically, we are interested in the prediction of the properties of small sized molecules and the analysis of the pairwise and three-body interactions. Before proceeding, we put our work in the context of existing literature on machine learning of molecular properties and in particular on molecular representations.

This chapter is structured as follows. The next Section 2.1 reviews some existing representations of physical systems. This is followed by the definition of our invariant many-body descriptors for molecules in Sec. 2.2. Sec. 2.3 details the data sets as well as the learning task and the prediction of several properties of small organic molecules and contains an analysis of the importance of the presented two- and three-body molecular features. The chapter is summarized and discussed in Sec. 2.5.

2.1 Representations of physical systems

Recently, machine learning has been successfully used to predict the atomization energies of small molecules [41, 20, 3, 2] and molecular dynamics simula-

tions [5, 42, 43, 44] as well as for studying properties of quantum-mechanical densities [45, 46]. Descriptors of molecules are constructed to provide an invariant, unique and efficient representation as input to machine learning models [47, 48, 21, 49, 50, 44, 51]. Such representations encoding a physical system will be defined by a set of 4-dimensional points $\{(Z_i, \mathbf{r}_i)\}_{i=1}^N$, where Z_i is the atomic number and \mathbf{r}_i is the position of the atom i in three-dimensional space, respectively. While the system size N is well defined for molecules (by the total number of atoms of the molecule), one workaround for solids is to use atomic environment descriptors together with a cutoff distance to limit the number of neighboring atoms used to compute the atomic representation. Alternatively, any molecular descriptor can also be combined with a modified distance metric to account for the periodic boundary conditions [52]. A raw encoding of the physical system by the atomic positions is unsuited for use in combination with machine learning methods as it neglects invariance with respect to basic symmetry operations. Instead, a representation is defined

$$R : \{(Z_i, \mathbf{x}_i)\}_{i=1}^N \rightarrow \mathbb{R}^{N_F} \quad (2.1)$$

with the number of features N_F . Such a mapping should encode the underlying chemical system in a complete, unique and efficient way, including as much problem symmetries as possible. One way to incorporate translational and rotational invariance is to use pairwise atomic distances to construct the representation R . For molecules, a pioneering work which utilizes this observation is the Coulomb matrix (CM) [41] which is defined as

$$C_{ij} = \begin{cases} 0.5Z_i^{2.4} & , i = j \\ \frac{Z_i Z_j}{\|\mathbf{r}_i - \mathbf{r}_j\|} & , i \neq j \end{cases} \quad (2.2)$$

Being composed based on inverse pairwise distances, the off-diagonal elements of the CM account well for Coulomb interaction terms of the atomization energy. The diagonal elements of the CM correspond to a polynomial fit of atomic energies to nuclear charge [41]. From the set of all pairwise distances a given molecule can be uniquely reconstructed, which is not the case for the following representations of this section. For equilibrium molecules a variant of the CM has been proposed which sorts the row (or equivalently column) norms, and which better suits the feature comparison needed for applying kernel methods [20]. The CM is a *global* descriptor in the sense that it lacks a direct encoding of local atomic environment features. Due to its simplicity and predictive power, the CM provides the basis for various following molecular descriptors. Being composed of two-body terms, the three-body interactions of a given molecule are implicitly learned by the intrinsic feature mapping of the kernel (see Sec. 1.2). Although sorting of the rows solves some of its problems, one possible flaw of the CM is the comparison of different kinds of atom combinations within the distance metric which brings us to the next descriptor.

The bag-of-bonds (BOB) molecular representation is a development of the CM which rearranges the elements of the CM into bags defined by a given bond

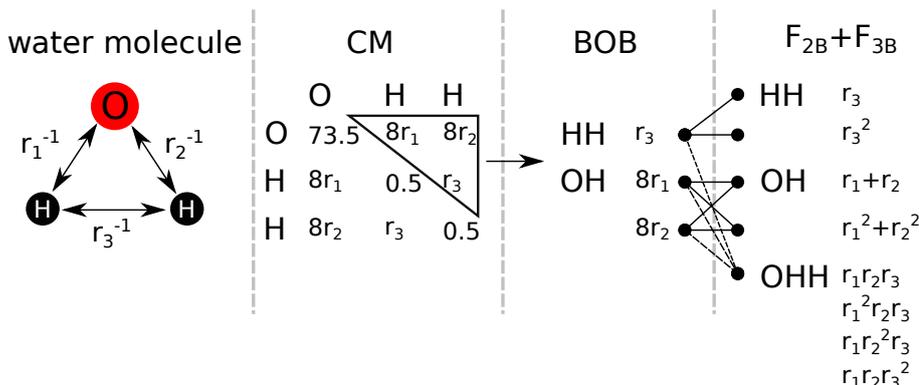


Figure 2.1: Molecular representations of a water molecule (left) defined by a set of three pairwise distances. From the Coulomb matrix (CM), the off-diagonal elements are reordered by the bag-of-bonds (BOB) descriptor. These two-body terms are then combined to atomic index invariant two-body and three-body features F_{2B} and F_{3B} , respectively. This image has appeared in previously published work [1].

type [53]. Within each feature group, the elements of BOB are sorted, thereby ensuring atomic permutation invariance. Due to this grouping, chemically more similar elements are compared with each other as compared to the CM. In addition, three-body interactions in molecules can possibly be better implicitly learned by the kernel. Similarly to the bag-of-words descriptor used in natural language processing and information retrieval applications, BOB encodes the frequencies of bonds present in a given molecule. As such, the BOB descriptor is inspired by interatomic potentials, which model a quantum-mechanical property as a sum over such potentials. In fact, a Taylor series expansion in combination with kernel ridge regression yields a low-order approximation of the BOB model by a sum over bonds and pairwise potentials [53]. This important finding indicates that the BOB model is able to *learn* optimal pairwise potentials better than the CM, which is beneficial for some extensive properties like the atomization energy and the polarizability, respectively [54]. For BOB, the Laplace kernel performs better than the Gaussian kernel, indicating that the Laplace kernel might be superior in utilizing non-local information in chemical compound space [53].

Shapeev et al. [55, 56] introduces systematically improvable interatomic potential descriptors based on invariant polynomials. These moment tensor potentials are invariant with respect to permutation, rotation and reflection and have the advantage that the computational complexity of computing these polynomials scales like $O(N)$, where N is the number of atoms. One possible limitation is that these potentials treat all atoms as chemically equivalent. Shapeev et al. suggest a future extension to alleviate this issue, namely to let the radial basis functions depend on the types of atoms.

Faber et al. [3] studied a representation using the histogram of distances, angles and dihedral angles with kernel ridge regression and achieves a mean absolute error (MAE) of 0.58 kcal/mol on the GDB-9 set, when trained on 118000 molecules. An angle representation based on molecular atomic radial angular distributions (MARAD) achieves a MAE of 1.2 kcal/mol with kernel ridge regression and 4.0 kcal/mol with the linear Bayesian ridge regression model when trained on 118000 molecules.

The recently introduced BAML (bonds angles machine learning) representation [57] can be viewed as many-body extension of BOB and constructs arbitrary distance functions between pairwise distances. BAML reaches a MAE of 1.15 kcal/mol on the GDB-7 set trained on 5000 molecules [49] and a MAE of 1.2 kcal/mol on the GDB-9 set when trained on 118000 molecules [3].

Huo et al. [49] introduces a many-body tensor representation which improves on the histogram descriptors of Faber et al. by “smearing” the histograms of given many-body features. For one of their best models, a MAE of 0.60 kcal/mol on GDB-7 using Gaussian kernel ridge regression and a MAE of 0.74 kcal/mol using a linear model (with many-body interactions) has been reported.

Most of the above approaches use explicit three-body (e.g. angle) or four-body (e.g. dihedral angle) features to construct the respective representation. In the next section, we propose novel translational, rotational and atom indexing invariant molecular descriptors which build on the success of inverse pairwise distances for predicting the atomization energy [20, 53, 41, 2, 21]. In particular, we construct many-body interaction features of *arbitrary order* from inverse pairwise distances which helps to alleviate sorting challenges encountered in e.g. CM. Similarly, our model learns e.g. a three-body interatomic potential, which is not necessarily a function of angle. Our novel descriptors allow to construct an invariant two-body and many-body interaction representation at *fixed* descriptor size. Note that fixed sized molecular descriptors are useful in practice as they can be easily used in combination with kernel ridge regression or deep neural networks or other models that expect fixed size input data. Also, such fixed size representations are generally extensible to large molecules and solids, while incorporating informative higher-order interaction terms. Furthermore when using these novel descriptors we observe that linear models perform only slightly worse than the non-linear methods. The latter is helpful in practice as linear models allow to easily analyse the importance of the proposed two-, three-body or many-body interaction features for predicting atomization energies of the molecules. This allows to extract insights from the learned model.

2.2 Invariant molecular many-body descriptors

We can represent a physical system with N atoms by the set $S = \{\mathbf{r}_i, Z_i\}_{i=1}^N$ where \mathbf{r}_i denotes the position of atom i in three-dimensional coordinate space and Z_i stands for the corresponding atomic number, respectively. A general

form of many-body interaction descriptors is defined using this set by

$$f_{\mathbf{Z},p}(S) := \sum_{(j_1, \dots, j_k) \in G(k, N)} \delta_{\mathbf{Z}}(\bar{\mathbf{Z}}) \cdot p(\mathbf{r}_{j_1}, Z_{j_1}, \dots, \mathbf{r}_{j_k}, Z_{j_k}) \quad (2.3)$$

where $\bar{\mathbf{Z}} := (Z_{j_1}, \dots, Z_{j_k})$, \mathbf{Z} is a given k -tuple of atomic numbers with $k \leq N$, p is a k -body interactions term, and the partial permutations set $G(k, N)$ consists of the sequences without repetition of k elements from the set $\{1, 2, \dots, N\}$. The cardinality of the k -permutation of N set $G(k, N)$ is $\frac{N!}{(N-k)!}$. For example if $k = 1$, the sum in Eq. (2.3) is taken over all atoms of the system S of a given type. The descriptors in Eq. (2.3) are intrinsically invariant to the indexing of the atoms comprising the system S , as the sum is formed over all elements of the set $G(k, N)$. If the k -body interactions term p satisfies invariance with respect to the translation and rotation of the atoms of S , this carries over to the descriptors $f_{\mathbf{Z},p}(S)$. For systems with a large number of atoms N , the sum can be limited to the largest interaction terms p contained in Eq. (2.3). In the following, we propose a set of translational and rotational invariant two-body and three-body interaction terms p , which will define our invariant many-body interaction descriptors.

Invariant two-body interaction descriptors

We define the set of translational and rotational invariant two-body interaction terms

$$p_m^{2B}(\mathbf{r}_1, Z_1, \mathbf{r}_2, Z_2) := \|\mathbf{r}_1 - \mathbf{r}_2\|^{-m} \quad (2.4)$$

where $m \in \mathbb{N}^+$. For a given set of n different atomic numbers $A_n := \{Z_i\}_{i=1}^n$ with $Z_i \neq Z_j \forall i, j \in \{1, \dots, n\}$, let S_{2B} denote the set of all tuples (Z_i, Z_j) with $Z_i \leq Z_j$ and $Z_i, Z_j \in A_n$. Let M_{2B} denote the set $\{1, 2, \dots, n_{2B}\}$ where n_{2B} is a parameter of the model and defines the largest occurring exponent. For a given physical system S , the two-body interaction descriptors F_{2B} are now given by

$$F_{2B} := \{f_{\mathbf{Z}, p_m^{2B}}(S)\}_{m \in M_{2B}, \mathbf{Z} \in S_{2B}} \quad (2.5)$$

Typically, the set A_n contains the atomic numbers present in the data set. The dimension of the two-body interaction descriptors is $n_{2B} \cdot n \cdot (n + 1)/2$. The pseudocode for computing these descriptors is shown in Alg. 1.

Invariant three-body interaction descriptors

We define the set of translational and rotational invariant three-body interaction terms

$$\bar{p}_{m_1, m_2, m_3}^{3B}(\mathbf{r}_1, Z_1, \mathbf{r}_2, Z_2, \mathbf{r}_3, Z_3) := \|\mathbf{r}_{12}\|^{-m_1} \cdot \|\mathbf{r}_{13}\|^{-m_2} \cdot \|\mathbf{r}_{23}\|^{-m_3} \quad (2.6)$$

Algorithm 1 Descriptors2B**Input:**

molecule $M = \{(Z_i, \mathbf{r}_i)\}_{i=1}^N$
maximal exponent n_{2B}

Output: feature vector $f(\cdot)$

```

1:  $f(\cdot) \leftarrow 0$ 
2: for  $i, j \leftarrow$  atoms of molecule do
3:    $\triangleright$  each atom is represented by position  $\mathbf{r}$  and atomic number  $Z$ 
4:    $r_{ij} \leftarrow \|\mathbf{r}_i - \mathbf{r}_j\|$ 
5:    $\mathbf{Z} \leftarrow$  sorted tuple  $(Z_i, Z_j)$ 
6:   for  $m = 1$  to  $n_{2B}$  do
7:      $f(\mathbf{Z}, m) += r_{ij}^{-m}$   $\triangleright$  feature entries are indexed by  $(\mathbf{Z}, m)$ 

```

and

$$p_{m_1, m_2, m_3}^{3B}(\mathbf{r}_1, Z_1, \mathbf{r}_2, Z_2, \mathbf{r}_3, Z_3) := \|\mathbf{r}_{12}\|^{-m_1} \cdot \|\mathbf{r}_{13}\|^{-m_2} \cdot \|\mathbf{r}_{23}\|^{-m_3} \quad (2.7)$$

$$\cdot \theta(Z_1, Z_2, Z_3, \|\mathbf{r}_{12}\|, \|\mathbf{r}_{13}\|, \|\mathbf{r}_{23}\|) \quad (2.8)$$

where $m_1, m_2, m_3 \in \mathbb{N}^+$, $\mathbf{r}_{ij} := \mathbf{r}_i - \mathbf{r}_j$ for $i, j = \{1, 2, 3\}$, and the bond angle indicator

$$\theta(\cdot) = \begin{cases} 1, & \|\mathbf{r}_{12}\| < B(Z_1, Z_2) \wedge \|\mathbf{r}_{13}\| < B(Z_1, Z_3) \\ 1, & \|\mathbf{r}_{13}\| < B(Z_1, Z_3) \wedge \|\mathbf{r}_{23}\| < B(Z_2, Z_3) \\ 1, & \|\mathbf{r}_{12}\| < B(Z_1, Z_2) \wedge \|\mathbf{r}_{23}\| < B(Z_2, Z_3) \\ 0, & \text{otherwise} \end{cases} \quad (2.9)$$

where $B(Z_i, Z_j) := 1.1 \cdot L(Z_i, Z_j)$ for $i, j = \{1, 2, 3\}$, and the values for the bond length function $L(\cdot)$ are given in Tab. 2.1. For a given set of n different atomic numbers $A_n := \{Z_i\}_{i=1}^n$ with $Z_i \neq Z_j \forall i, j \in \{1, \dots, n\}$, let S_{3B} denote the set of all 3-tuples (Z_i, Z_j, Z_k) with $Z_i < Z_j < Z_k$ and $Z_i, Z_j, Z_k \in A_n$. Let M_{3B} be the set of partial permutations $G(3, n_{3B})$ as defined above, where n_{3B} is a parameter of the model and defines the largest occurring exponent for the three-body terms. For a given physical system S , the three-body interaction descriptors F_{3B} and \bar{F}_{3B} are now given by

$$F_{3B} := \left\{ f_{\mathbf{Z}, p_{m_1, m_2, m_3}^{3B}}(S) \right\}_{(m_1, m_2, m_3) \in M_{3B}, \mathbf{Z} \in S_{3B}} \quad (2.10)$$

and

$$\bar{F}_{3B} := \left\{ f_{\mathbf{Z}, \bar{p}_{m_1, m_2, m_3}^{3B}}(S) \right\}_{(m_1, m_2, m_3) \in M_{3B}, \mathbf{Z} \in S_{3B}} \quad (2.11)$$

The dimension of the three-body interaction descriptors is $n^2 \cdot (n+1) / 2 \cdot \frac{n_{3B}!}{(n_{3B}-3)!}$. The pseudocode for computing these descriptors is shown in Alg. 2.

Fig. 2.1 schematically shows the representations CM, BOB and the descriptors F_{2B} and F_{3B} for the example of a water molecule.

Algorithm 2 Descriptors3B**Input:**

molecule $M = \{(Z_i, \mathbf{r}_i)\}_{i=1}^N$
 maximal exponent n_{3B}

Output: feature vector $f(\cdot)$

```

1:  $G(3, n_{3B}) \leftarrow$  sequences without repetition of 3 elements from the set
    $\{1, 2, \dots, n_{3B}\}$ 
2:  $B(\cdot) \leftarrow 1.1 \cdot L(\cdot)$ 
3:  $f(\cdot) \leftarrow 0$ 
4: for  $i, j, k \leftarrow$  atoms of molecule do
5:      $\triangleright$  each atom is represented by position  $\mathbf{r}$  and atomic number  $Z$ 
6:      $r_{ij} \leftarrow \|\mathbf{r}_i - \mathbf{r}_j\|$ 
7:      $r_{ik} \leftarrow \|\mathbf{r}_i - \mathbf{r}_k\|$ 
8:      $r_{jk} \leftarrow \|\mathbf{r}_j - \mathbf{r}_k\|$ 
9:     hasAngle  $\leftarrow$  False
10:    if  $r_{ij} < B(Z_i, Z_j)$  and  $r_{ik} < B(Z_i, Z_k)$  then
11:        hasAngle  $\leftarrow$  True
12:    else if  $r_{ik} < B(Z_i, Z_k)$  and  $r_{jk} < B(Z_j, Z_k)$  then
13:        hasAngle  $\leftarrow$  True
14:    else if  $r_{ij} < B(Z_i, Z_j)$  and  $r_{jk} < B(Z_j, Z_k)$  then
15:        hasAngle  $\leftarrow$  True
16:    if not hasAngle then
17:        continue
18:     $\mathbf{Z} \leftarrow$  sorted tuple  $(Z_i, Z_j, Z_k)$ 
19:    for  $m_1, m_2, m_3 \leftarrow G(3, n_{3B})$  do
20:         $f(\mathbf{Z}, m_1, m_2, m_3) += r_{ij}^{-m_1} \cdot r_{ik}^{-m_2} \cdot r_{jk}^{-m_3}$ 
21:         $\triangleright$  feature entries are indexed by  $(\mathbf{Z}, m_1, m_2, m_3)$ 

```

2.3 Tests on molecular data sets

We use the following two reference data sets for the evaluation of the predictive power of machine learning models with our proposed invariant many-body interaction descriptors.

GDB-7. The GDB-7 data set is a subset of the freely available small molecule database GDB-13 [58] with up to seven heavy atoms CNO. For this data set, electronic ground- and excited state properties have been calculated. Hybrid density functional theory with the Perdew-Burke-Ernzerhof hybrid functional approximation (PBE0) [59, 60] has been used to calculate the atomization energy of the molecules. The electron affinity, ionization potential, excitation energies and maximal absorption intensity have been obtained from ZINDO [61, 62, 63]. For the static polarizability, PBE0 and self-consistent screening (SCS) [64] have been used. The frontier orbital (HOMO and LUMO) eigenvalues have been calculated using PBE0, SCS and Hedin’s GW approximation [65]. The SCS, PBE0 and GW calculations have been performed using

bond-type	(Z_1, Z_2)	$L(Z_1, Z_2)$
H-H	(1, 1)	0.74
H-C	(1, 6)	1.08
H-O	(1, 8)	0.96
H-N	(1, 7)	1.01
C-C	(6, 6)	1.51
C-O	(6, 8)	1.43
C-N	(6, 7)	1.47
O-O	(8, 8)	1.48
O-N	(8, 7)	1.40
N-N	(7, 7)	1.45
F-H	(9, 1)	0.92
F-C	(9, 6)	1.35
F-O	(9, 8)	1.42
F-N	(9, 7)	1.36
F-F	(9, 9)	1.42

Table 2.1: Bond lengths in Ångström (right column) for all combinations of the elements H, C, N, O and F. Used for computing the three-body interaction descriptors F_{3B} . This table has appeared in previously published work [8].

FHI-AIMS [66] (tight settings/tier2 basis set), ZINDO/s calculations are based on the ORCA [67] code.

GDB-9. The GDB-9 data set is a subset of the chemical universe database GDB-17 [68] of 166 billion organic small molecules. The subset contains molecules with up to nine heavy atoms CNO with corresponding harmonic frequencies, dipole moments, polarizabilities, along with energies, enthalpies, and free energies of atomization, all calculated at the B3LYP/6-31G(2df,p) level of quantum chemistry [69].

We evaluate the performance of predicting the properties of the molecules of these two data sets by using our proposed invariant many-body interaction descriptors F_{2B} and $F_{2B} + F_{3B}$. Additionally, we computed the sorted Coulomb matrices (CM) [20] and the popular bag-of-bonds (BOB) [53] molecular representations. For the atomization energy, we use the models kernel-ridge-regression (KRR) [70, 71], ridge regression (RR) [72], k-nearest neighbors (KNN) [73] and the mean predictor (MEAN). For the other properties, we use kernel-ridge-regression with the Laplace kernel for CM and BOB which works better compared to the Gauss kernel for these descriptors [20], the Gauss kernel in combination with the F_{2B} and $F_{2B} + F_{3B}$ descriptors, respectively. To fit the model parameters (hyper parameters), we use 10-fold cross-validation [74], see [20] for details. Unless otherwise noted, the models are trained on 5000

Method	Features	MAE	RMSE	Max. dev.
mean	-	174	219	1166
RR	CM	25	33	134
RR	BOB	23	30	144
RR	F_{2B}	4.9	12	350
RR	$F_{2B} + F_{3B}$	1.0	8.3	327
RR	$F_{2B} + \bar{F}_{3B}$	1.0	8.1	301
KNN	CM	80	104	461
KNN	BOB	70	102	424
KNN	F_{2B}	49	73	230
KNN	$F_{2B} + F_{3B}$	10	28	306
KNN	$F_{2B} + \bar{F}_{3B}$	13	35	395
KRR (Gaussian)	CM	8.6	15	433
KRR (Laplace)	CM	3.7	5.8	89
KRR (Gaussian)	BOB	7.6	10	99
KRR (Laplace)	BOB	1.8	3.9	103
KRR (Gaussian)	F_{2B}	1.9	4.7	155
KRR (Laplace)	F_{2B}	4.2	6.1	62
KRR (Gaussian)	$F_{2B} + F_{3B}$	0.83	1.5	28
KRR (Laplace)	$F_{2B} + F_{3B}$	2.4	3.8	51
KRR (Gaussian)	$F_{2B} + \bar{F}_{3B}$	0.81	1.4	31
KRR (Laplace)	$F_{2B} + \bar{F}_{3B}$	2.7	4.1	67

Table 2.2: Prediction errors of the PBE0 atomization energy of the molecules of the set GDB-7 by various machine learning models with random 5k train molecules and the remaining 1868 molecules as test set. The errors are given in kcal/mol. The models used are ridge regression (RR), kernel ridge regression (KRR) and k-nearest neighbors (KNN). The results in this table have appeared in previously published work [8].

random molecules. The performance is evaluated on the remaining molecules of the respective set, by the mean absolute error (MAE), the root-mean-square error (RMSE) and the maximum deviation (Max. dev.), respectively.

For the atomization energy, the results of the machine learning models are given in Tab. 2.2 and 2.3. The results for predicting diverse quantum mechanical properties are given in Tab. 2.4 and 2.5, respectively. The MAE in dependence of the number of training samples is shown in Figs. 2.2 and 2.3, respectively.

The $F_{2B} + F_{3B}$ and $F_{2B} + \bar{F}_{3B}$ models outperforms the BOB descriptor in

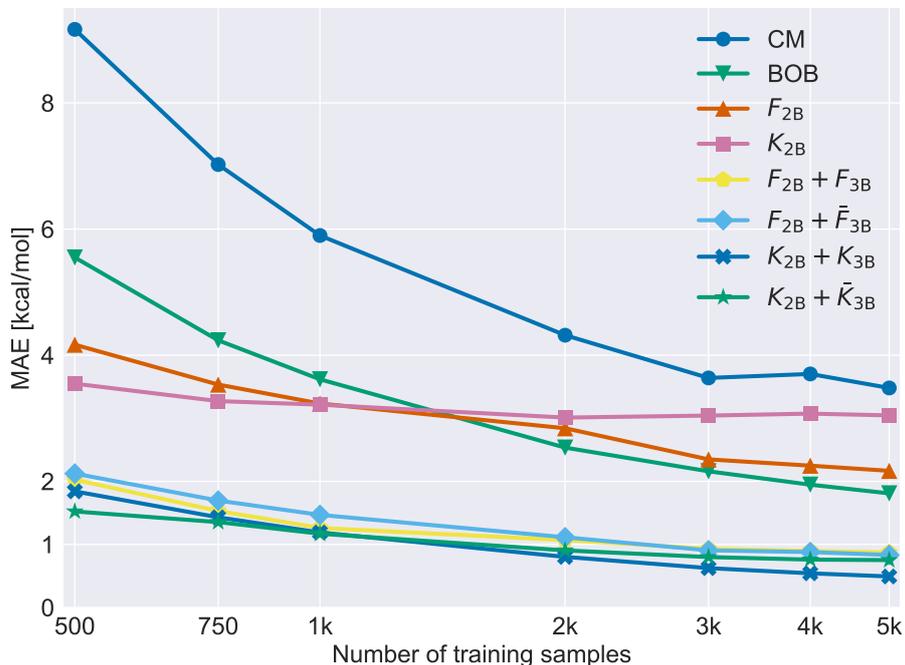


Figure 2.2: Mean absolute error of predicting the PBE0 atomization energy of the molecules of the set GDB-7 with KRR in dependence of the number of training samples. The errors are given in kcal/mol. For CM, BOB, F_{2B} , $F_{2B} + F_{3B}$ and $F_{2B} + \bar{F}_{3B}$, the Gaussian kernel has been used. The kernel parameters have been determined by 10-fold nested cross-validation. The kernel methods K_{2B} , $K_{2B} + K_{3B}$ and $K_{2B} + \bar{K}_{3B}$ will be presented in the Chap. 4.

the prediction of the static polarizability computed with self-consistent screening (20% improvement), the first excitation energy (20% improvement) and the atomization energy (50% improvement) of the molecules of the GDB-7 set. Additionally, the prediction errors of the electron affinity and the HOMO eigenvalues are improved by 5%. The largest correlation between prediction and reference is achieved for the static polarizability computed with SCS as well as the atomization energy. Noticeably, both models $F_{2B} + F_{3B}$ and $F_{2B} + \bar{F}_{3B}$ display similar prediction accuracies for the atomization energy, indicating that three-body interactions are local for molecules at equilibrium.

The $F_{2B} + F_{3B}$ and $F_{2B} + \bar{F}_{3B}$ models outperforms the BOB descriptor in the prediction of the heat capacity (40% improvement), the zero point vibrational energy (50% improvement), the isotropic polarizability (30% improvement) and the atomization energies (60% improvement) of the molecules of the GDB-9 set. Additionally, the prediction errors of the HOMO and LUMO eigenvalues as well as the gap are improved by 15%, 10% and 9%, respectively. The largest correlation between prediction and reference is achieved for the

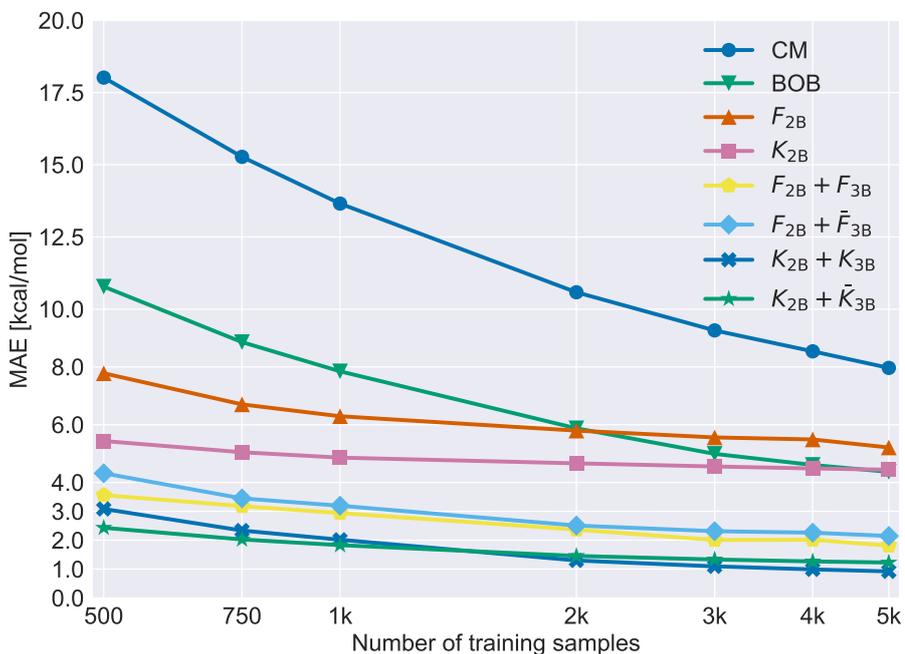


Figure 2.3: Mean absolute error of predicting the B3LYB/6-31G(2df,p) atomization energy of the molecules of the set GDB-9 with KRR in dependence of the number of training samples. The errors are given in kcal/mol. For CM, BOB, F_{2B} , $F_{2B} + F_{3B}$ and $F_{2B} + \bar{F}_{3B}$, the Gaussian kernel has been used. The kernel parameters have been determined by 10-fold nested cross-validation. The kernel methods K_{2B} , $K_{2B} + K_{3B}$ and $K_{2B} + \bar{K}_{3B}$ will be presented in the Chap. 4.

electronic spatial extent, zero point vibrational energy, the heat capacity, the isotropic polarizability and the atomization energies.

The prediction of the atomization energy by using the linear RR model is comparable to the KRR model. This makes the $F_{2B} + F_{3B}$ descriptors interesting candidates for alternative linear regression models such as Bayesian linear regression [75], partial least squares [76] or generalized least squares [77]. In this work, we will utilize this fact to compute a feature ranking measure in the next section.

2.4 Feature importance analysis

The inclusion of the three-body descriptors F_{3B} increases the predictive power of the KRR model by more than 50% over using the two-body descriptors F_{2B} for both data sets GDB-7 and GDB-9. Due to the non-linear kernels used, it is not obvious, how the three-body features improve the performance. The fre-

Method	Features	MAE	RMSE	Max. dev.
mean	-	185	235	1544
RR	CM	235	308	1289
RR	BOB	89	134	653
RR	\bar{F}_{2B}	6.8	10	462
RR	$F_{2B} + F_{3B}$	1.6	2.8	88
RR	$F_{2B} + \bar{F}_{3B}$	1.6	2.9	81
KNN	CM	239	279	898
KNN	BOB	231	272	758
KNN	\bar{F}_{2B}	151	177	556
KNN	$F_{2B} + F_{3B}$	25	42	358
KNN	$F_{2B} + \bar{F}_{3B}$	27	55	476
KRR (Gaussian)	CM	17	22	181
KRR (Laplace)	CM	7.9	10	129
KRR (Gaussian)	BOB	11	16	253
KRR (Laplace)	BOB	4.0	6.0	132
KRR (Gaussian)	F_{2B}	4.8	6.4	45
KRR (Laplace)	F_{2B}	8.2	11	190
KRR (Gaussian)	$F_{2B} + F_{3B}$	1.5	2.8	96
KRR (Laplace)	$F_{2B} + F_{3B}$	4.5	6.4	147
KRR (Gaussian)	$F_{2B} + \bar{F}_{3B}$	1.5	2.7	91
KRR (Laplace)	$F_{2B} + \bar{F}_{3B}$	4.7	7.5	170

Table 2.3: Prediction errors of the B3LYP/6-31G(2df,p) atomization energy of the molecules of the set GDB-9 by various machine learning models with random 5k train molecules and the remaining 126722 molecules as test set. The errors are given in kcal/mol. The models used are ridge regression (RR), kernel ridge regression (KRR) and k-nearest neighbors (KNN). The results in this table have appeared in previously published work [8].

frequencies of the bond types corresponding to three bonded atoms which have an angle (Fig. 2.5 and Fig. 2.6 top) suggest the top-three most important connections C–C–H, H–C–H and C–C–C, respectively. On the other hand, using the F_{2B} descriptors in combination with the H–C–H subset of F_{3B} features (Fig. 2.5 and Fig. 2.6 bottom) shows negligible decrease of the mean absolute error of the KRR model as compared to the inclusion of the C–C–H and C–C–C subsets.

There are a number of ways to define feature importance [78, 79, 80, 81] respectively to explain nonlinear models [82, 83, 84, 85, 86, 87, 88]. Here,

2.4. Feature importance analysis

Property	CM	BOB	F_{2B}	$F_{2B} + F_{3B}$	$F_{2B} + \bar{F}_{3B}$	Unit
ae-pbe0	3.7	1.8	1.9	0.83	0.81	kcal/mol
homo-gw	0.212	0.138	0.167	0.128	0.130	eV
lumo-gw	0.187	0.142	0.155	0.147	0.129	eV
homo-pbe0	0.202	0.130	0.156	0.120	0.118	eV
lumo-pbe0	0.174	0.108	0.133	0.108	0.092	eV
homo-zindo	0.279	0.144	0.173	0.132	0.132	eV
lumo-zindo	0.252	0.134	0.168	0.112	0.115	eV
p-pbe0	0.130	0.083	0.103	0.088	0.073	Ångström ³
p-scs	0.065	0.042	0.061	0.032	0.022	Ångström ³
e1-zindo	0.37	0.19	0.21	0.15	0.17	eV
ea-zindo	0.29	0.15	0.18	0.13	0.14	eV
imax-zindo	0.084	0.067	0.074	0.071	0.072	a.u.
emax-zindo	1.47	1.20	1.29	1.26	1.15	eV
ip-zindo	0.32	0.18	0.21	0.18	0.20	eV

Table 2.4: Mean absolute errors of predicting several ground- and excited state properties by kernel ridge regression trained on 5000 random molecules and tested on the remaining 1868 molecules of the GDB-7 data set. The best performing models are marked in bold. The results in this table have appeared in previously published work [8].

we use the feature importance ranking measure (FIRM) [83], which defines the feature importance according to the standard deviation of a conditional expected output of the learner. FIRM can be applied to a broad family of learning machines, the measure is robust with respect to perturbation of the problem and invariant with respect to irrelevant transformations. In general, the computation of the exact FIRM is infeasible. For the unregularized linear regression model and normally distributed input features, the FIRM of a feature f can be computed analytically [83] by

$$\text{FIRM}(f) := \frac{1}{n} \cdot \frac{1}{\sigma(f)} \cdot \text{cov}(f, y), \quad (2.12)$$

where n is the number of samples, $\sigma(\cdot)$ is the standard deviation, y denotes the labels and $\text{cov}(\cdot)$ the covariance. In the above formula, FIRM is computed for each feature independently. To capture the importance of the inclusion of the three-body descriptors F_{3B} , we propose to use FIRM on the signed deviation of labels and prediction of the KRR model with the two-body features F_{2B}

$$\text{FIRM}_{3B}(f) := \frac{1}{n} \cdot \frac{1}{\sigma(f)} \cdot \text{cov}(f, y - p_{2B}), \quad (2.13)$$

2. MANY-BODY DESCRIPTORS

Property	CM	BOB	F_{2B}	$F_{2B} + F_{3B}$	$F_{2B} + \bar{F}_{3B}$	Unit
U0	7.9	4.0	4.8	1.5	1.5	kcal/mol
U	7.9	4.0	4.8	1.5	1.5	kcal/mol
H	7.9	4.0	4.8	1.5	1.5	kcal/mol
G	7.9	4.0	4.8	1.5	1.5	kcal/mol
HOMO	5.8	4.3	4.7	3.6	3.9	kcal/mol
LUMO	8.9	5.7	6.0	5.1	5.4	kcal/mol
gap	11	6.8	7.9	6.2	6.6	kcal/mol
alpha	1.00	0.63	0.72	0.49	0.58	Bohr ³
mu	0.77	0.65	0.67	0.61	0.61	Debye
r2	16	8.5	7.3	9.0	11.5	Bohr ²
zpve	0.33	0.20	0.18	0.10	0.10	kcal/mol
A	0.42	0.37	0.40	0.42	0.45	GHz
B	0.12	0.10	0.12	0.13	0.11	GHz
C	0.052	0.045	0.046	0.050	0.042	GHz
cv	0.38	0.20	0.21	0.12	0.11	cal/(mol K)

Table 2.5: Mean absolute errors of predicting several properties calculated at the B3LYP/6-31G(2df,p) level of quantum chemistry and predicted by kernel ridge regression trained on 5000 random molecules and tested on the remaining 126722 molecules of the GDB-9 data set. The best performing descriptors are marked in bold. The results in this table have appeared in previously published work [8].

where p_{2B} is the prediction of the KRR model using the F_{2B} descriptors, see also [89]. Additionally, we compute the product of the above $FIRM_{3B}$ of the feature f with the frequency of its corresponding bond-type

$$FIRM_{\text{freq}}(f) := \text{freq}(f) \cdot \frac{1}{n} \cdot \frac{1}{\sigma(f)} \cdot \text{cov}(f, y - p_{2B}), \quad (2.14)$$

where $\text{freq}(f)$ is the frequency of the bond-type corresponding to the feature f . Fig. 2.5 and Fig. 2.6 show the $FIRM$, $FIRM_{3B}$ and $FIRM_{\text{freq}}$ for the three-body descriptors F_{3B} for both data sets GDB-7 and GDB-9. Additionally, we show the frequency of the bond-type corresponding to the feature f and the error improvement of using the KRR model with the F_{2B} features augmented with the corresponding subset of three-body features F_{3B} .

The $FIRM_{3B}$ indicates low importance of the H–C–H and increased importance of the C–C–C features, which correlates with the error improvement by using these features in combination with the F_{2B} descriptors. This indicates, that three-body interactions relevant for prediction improvement are more dominant for the C–C–C bond type as compared to the H–C–H bond

property	description
ae-pbe0	atomization energy (DFT/PBE0)
homo-gw	highest occupied molecular orbital (GW)
lumo-gw	highest unoccupied molecular orbital (GW)
homo-pbe0	highest occupied molecular orbital (DFT/PBE0)
lumo-pbe0	highest unoccupied molecular orbital (DFT/PBE0)
homo-zindo	highest occupied molecular orbital (ZINDO/s)
lumo-zindo	highest unoccupied molecular orbital (ZINDO/s)
p-pbe0	polirazability (DFT/PBE0)
p-scs	polarizability (self-consistent-screening)
e1-zindo	first excitation energy (ZINDO)
ea-zindo	electron affinity (ZINDO/s)
imax-zindo	excitation energy at maximal absorption (ZINDO)
emax-zindo	maximal absorption intensity (ZINDO)
ip-zindo	ionization potential (ZINDO/s)

Table 2.6: Description of the molecular properties contained in the data set GDB-7. This table has appeared in previously published work [8].

type, where the correlation with the atomization energy can be captured by using the corresponding two-body features F_{2B} .

The measure $FIRM_{3B}$ reduces the importance of the hydrogen type bonds in favour of the non-hydrogen features, as compared to $FIRM$. The correlation of a molecular descriptor with the target (atomization energy) is not necessarily a good predictor variable in presence of other features. In this sense, $FIRM_{3B}$ captures the importance of the three-body descriptors F_{3B} in the presence of the two-body interactions modelled by the two-body descriptors F_{2B} . For the non-hydrogen type three-body features, $FIRM$ indicates approximately equal importance of the C–C–C, C–C–N and N–C–O bonds, in contrast to $FIRM_{3B}$, which lifts the C–C–C importance. This shows, that for non-hydrogen bonds, our set of descriptors are better able to capture three-body interactions of the C–C–C type as compared to the other bond types. In spite of the 5 times lower frequency of the N–C–O bond compared to C–C–N, both, the error improvement and $FIRM_{3B}$ show approximately equal importance of these three-body interactions.

For the three-body features, we can use the parameters of the linear RR model to compute the energy of a given bond-type

$$E_{3B}(b) := \sum_{i=1}^N \delta_b(\text{bond}(i)) \cdot c_i \cdot f_i, \quad (2.15)$$

property	description
U0	internal energy at 0 K
U	internal energy at 298.15 K
H	enthalpy at 298.15 K
G	free energy at 298.15 K
HOMO	energy of highest occupied molecular orbital
LUMO	energy of lowest unoccupied molecular orbital
gap	gap, difference between LUMO and HOMO
alpha	isotropic polarizability
mu	dipole moment
r2	electronic spatial extent
zpve	zero point vibrational energy
A	rotational constant A
B	rotational constant B
C	rotational constant C
cv	heat capacity at 298.15 K

Table 2.7: Description of the molecular properties contained in the data set GDB-9. This table has appeared in previously published work [8].

where c_i are the coefficients of the trained RR model, f_i are the three-body features, N are the number of three-body features, b is the bond-type under examination and $\text{bond}(i)$ indicates the bond-type corresponding to the feature f_i . Fig. 2.4 shows E_{3B} in dependence of the bond angle exemplary for the C–C–C type bond type of the GDB-7 and GDB-9 set, respectively.

Physically, these results indicate, that for intermediate size molecules, the interaction of the hydrogen atom with all other atoms (of type C, N, O) can be captured effectively by pairwise interactions. In fact, if we use the F_{2B} features in combination with the non-hydrogen subset of F_{3B} , we get a mean absolute error of 0.9 kcal/mol for the GDB-7 set and 1.8 kcal/mol for the GDB-9 set on the rest of the molecules, respectively. In view of the fact, that the hydrogen atom constitutes by far the dominant atom type for both data sets, the errors degrade by 13% and 20% as compared to the full $F_{2B} + F_{3B}$ descriptors, respectively. This intriguing result lets us formulate to the following conjecture:

For the accurate prediction of the atomization energy of intermediate size molecules, the interaction potential of the hydrogen atom with all other atoms can be effectively approximated as a pairwise interaction potential.

The interatomic interaction between non-hydrogen atoms goes beyond pairwise interactions. Interestingly, for the C–C–C bond-type, the energy E_{3B}

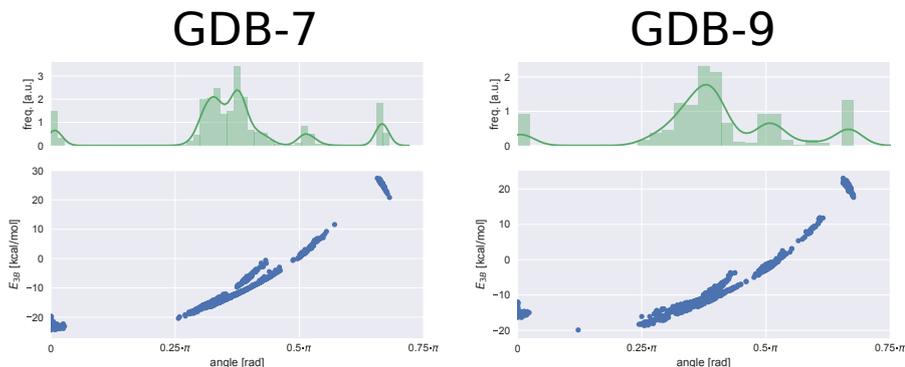


Figure 2.4: E_{3B} by Eq. (2.15) in dependence of the bond angle for the C–C–C bond type for the molecules of the set GDB-7 (left) and GDB-9 (right) along with the distribution of the angles (top). These images have appeared in previously published work [8].

shows a clear dependence of the bond angle, as compared to the other bond-types. This result indicates, that there is a simple relation between the angle at the C-atom of the C–C–C bond-type and the atomization energy. Between the angles $\pi/4$ and $\pi/2$, there exist two branches of the dependence of the atomization energy of the angle. This indicates, that for C–C–C, our model learns two angle-type functions, distinguishing single-double and single-single C–C–C bonds, see the C–C–C angle dependence of E_{3B} in Fig. 2.4.

2.5 Summary and discussion

In this chapter we have developed representations of physical systems composed of atoms which have a fixed size and are translation, rotation and atom indexing invariant. We have used these descriptors to predict quantum mechanical properties of a set of small organic molecules containing the heavy atoms CNO with kernel ridge regression in combination with the Gaussian kernel. On these data sets our best models outperform the CM and BOB, the improvement ratio being better for extensive than for intensive properties. A more detailed analysis in the next chapter, where we compare our models with an artificial neural network approach which learns atom-wise decompositions directly from first principles shows the difficulty in using our descriptors for predicting highly non-local properties like transition energies. Using a linear model which performs only slightly worse compared to kernel ridge regression with our descriptors, a feature importance analysis has indicated that for the accurate prediction of the atomization energy of small sized molecules, the interaction potential of the hydrogen atom with all other atoms can be effectively approximated as a pairwise interaction potential.

Although our proposed molecular descriptors display a superior perfor-

mance compared to CM and BOB for predicting the atomization energy of stable molecules, a possible difficulty is the number of features which grows exponentially for increasing the maximal exponent in the respective definitions. This can be a problem for molecular dynamics data sets where a larger exponent is expected to better model the highly non-linear energy surfaces. We tackle this problem in Chap. 4, where we not only solve this issue, but design an even more performant similarity measures for quantum mechanical systems based on kernels.

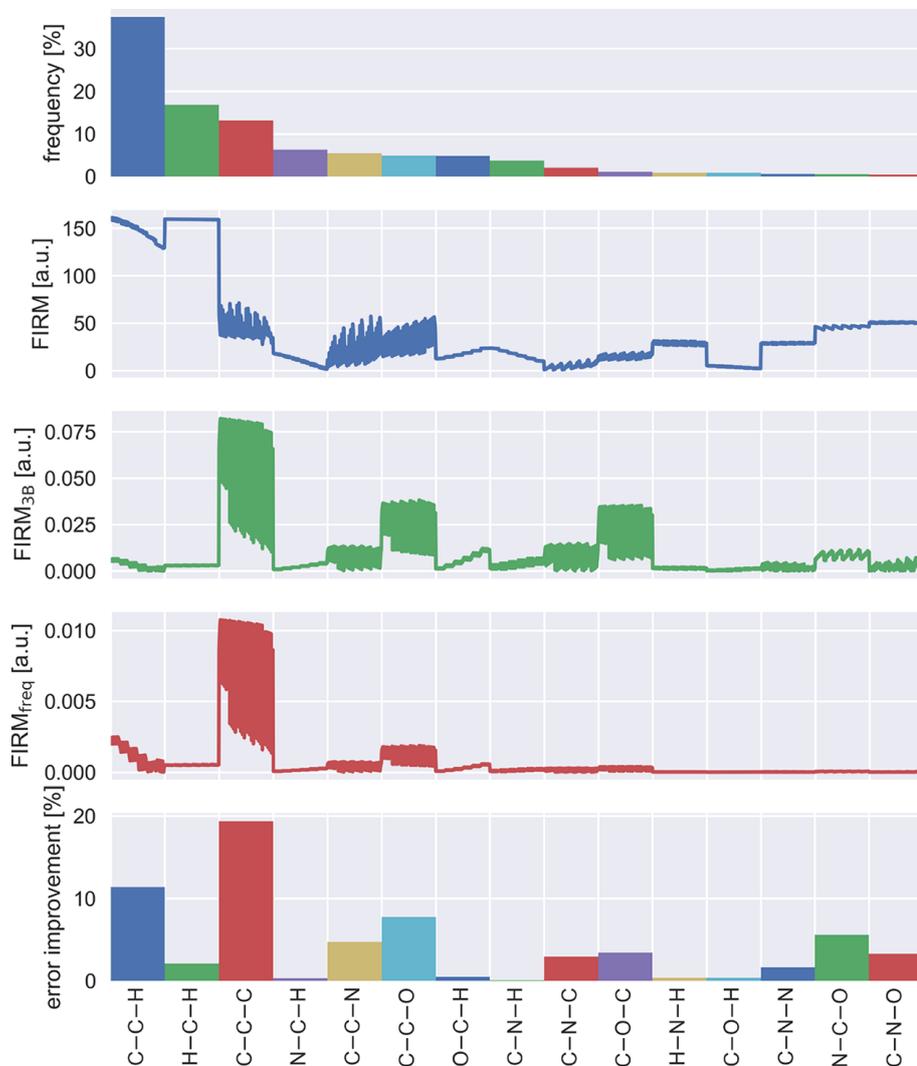


Figure 2.5: FIRM by Eq. (2.12), (second from top), FIRM_{3B} by Eq. (2.13) (third from top), and FIRM_{freq} by Eq. (2.14) (fourth from top) for the F_{3B} descriptors of the data set GDB-7. Additionally, the frequency of the corresponding bond-type (top) and the error improvement by using KRR with the F_{2B} features in combination with the bond-type subset of the F_{3B} descriptors (bottom) are shown. This image has appeared in previously published work [8].

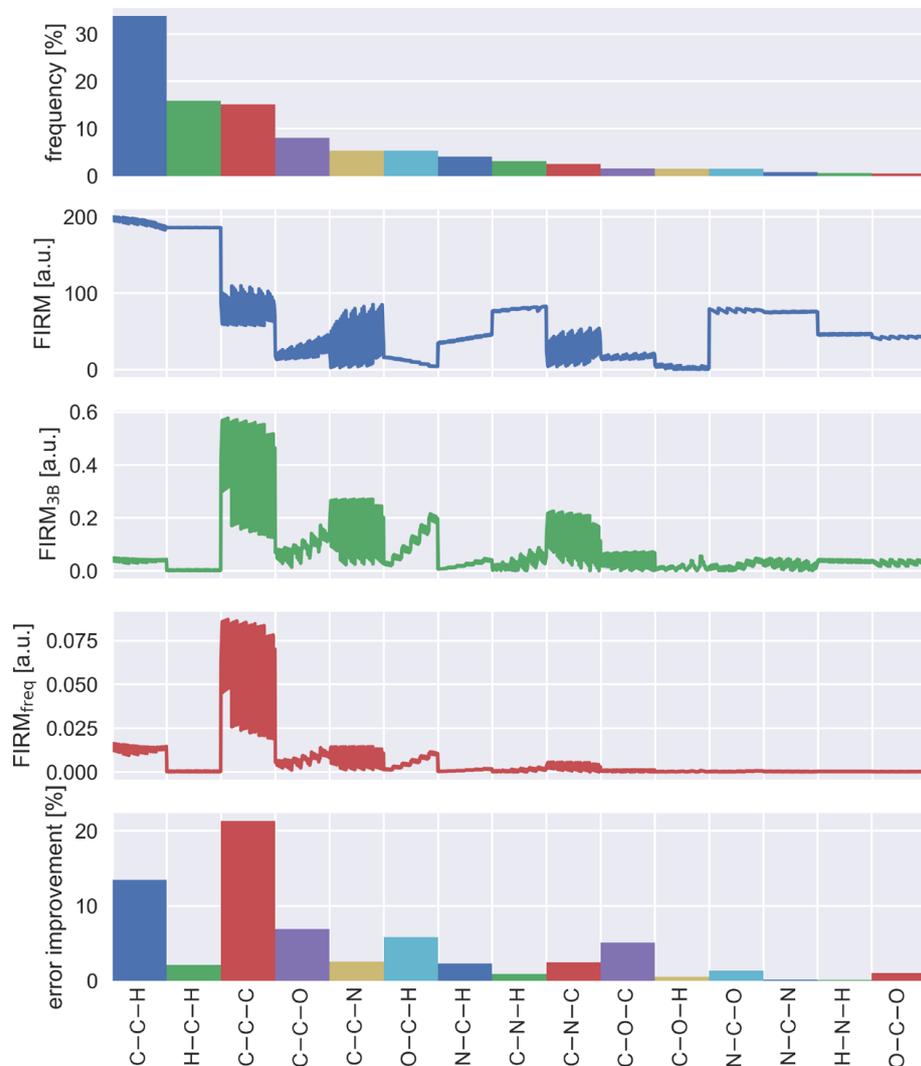


Figure 2.6: FIRM by Eq. (2.12), (second from top), FIRM_{3B} by Eq. (2.13) (third from top), and FIRM_{freq} by Eq. (2.14) (fourth from top) for the F_{3B} descriptors of the data set GDB-9. Additionally, the frequency of the corresponding bond-type (top) and the error improvement by using KRR with the F_{2B} features in combination with the bond-type subset of the F_{3B} descriptors (bottom) are shown. This image has appeared in previously published work [8].

Capturing intensive and extensive molecular properties with machine learning

Large parts of this chapter have appeared in previously published work [54].

Recently, machine learning has been successfully applied to the fast and accurate prediction of molecular properties across chemical compound space [41, 20, 90, 3, 2, 91] and molecular dynamics simulations [52, 42, 43, 92] as well as for studying properties of quantum-mechanical densities [45, 46]. An indispensable ingredient to most machine learning models are molecular descriptors, which are constructed to provide an invariant, unique and efficient representation as input to machine learning models [47, 48, 21, 49, 50, 44, 51]. A popular molecular descriptor is the bag-of-bonds (BOB) model [53], which is an extension of the Coulomb matrix (CM) approach [41] and groups the pairwise distances according to pairs of atom types. In the previous chapter, we have demonstrated how machine learning can be successfully applied to the prediction of chemical properties of small organic molecules such as energies or polarizabilities using a set of translation, rotation and atom indexing invariant two- and three-body descriptors.

Compared to these properties, the electronic excitation energies pose a much more challenging learning problem. Studying the valence electronic spectra of small molecules can yield insights into the properties and discovery of solar cell materials [93] and organic diodes [94]. Attractive candidates for computing such properties are time-dependent DFT or wavefunction-based methods. One popular method is to use linear response time-dependent density functional theory (LR-TDDFT) within the adiabatic approximation [95]. Although less computationally expensive than corresponding coupled-cluster approaches, computing the spectra via LR-TDDFT is still a demanding task, in particular across chemical compound space, where the properties of a diverse data set of compounds need to be obtained in a fast and reliable manner.

3. CAPTURING INTENSIVE AND EXTENSIVE MOLECULAR PROPERTIES WITH MACHINE LEARNING

In this chapter, we examine how recent machine learning approaches can be transferred to predicting intensive properties, in particular the singlet-singlet transition energies computed with TDDFT. Intensive properties are characterized by being independent of the system size, as opposed to extensive properties, which increase with increasing system size. To trace the source of possible difficulties back to intensiveness or descriptor, we choose a set of different types of properties to be predicted with machine learning. Specifically, we select the atomization energy and the isotropic polarizability as extensive property. In addition, we choose three intensive properties: the gap between the highest occupied and lowest unoccupied molecular orbital energies (HOMO-LUMO gap), together with the transition energy of the ground state (S_0) to the lowest two vertical electronic excited states (S_1 and S_2), E_1 and E_2 , respectively.

On these selected quantum mechanical properties, we perform experiments with various types of molecular descriptors. We examine the two- and three-body translational and rotationally invariant molecular descriptors used in the previous chapter which are especially suited for this study as they are invariant w.r.t. atom indexing and independent of the size of the molecule. Thus they can easily be used in combination with kernel ridge regression and artificial neural networks. Furthermore, the used molecular representation is extensible to large molecules and solids as well as to incorporate higher-order interaction terms. Additionally, we examine and compare the performance with the neural network SchNet [96], which learns a local representation of the property under investigation.

The chapter is structured as follows. In Sec. 3.1 we introduce the methods used to predict intensive and extensive quantum-mechanical properties. This is followed by the experiment Sec. 3.2. The summary and discussion in Sec. 3.3 concludes this chapter.

3.1 Methods

Invariant two-body interaction descriptors

We use the translation, rotation and atom indexing invariant two-body descriptors already encountered in the previous chapter. To recall, for two atoms of the molecule with the atomic numbers and coordinates (Z_1, \mathbf{r}_1) and (Z_2, \mathbf{r}_2) , the set of two-body descriptors is given by

$$F_{2B,Z_1,Z_2} := \left\{ \|\mathbf{r}_1 - \mathbf{r}_2\|^{-m} \right\}_{m=1,\dots,M} \quad (3.1)$$

where we choose $M = 15$ for this study. For the whole set of two-body descriptors F_{2B} , we concatenate the descriptors F_{2B,Z_1,Z_2} for the set of pairs of atomic numbers (Z_1, Z_2) present in the data set.

Invariant three-body interaction descriptors

We use a variant of the translation, rotation and atom indexing invariant three-body descriptors already encountered in the previous chapter. To recall, for

three atoms of the molecule with the atomic numbers and coordinates (Z_1, \mathbf{r}_1) , (Z_2, \mathbf{r}_2) and (Z_3, \mathbf{r}_3) , the set of three-body descriptors is given by

$$F_{3B, Z_1, Z_2, Z_3} := \left\{ \frac{1}{\|\mathbf{r}_{12}\|^{m_1} \|\mathbf{r}_{13}\|^{m_2} \|\mathbf{r}_{23}\|^{m_3}} \right\} \quad (3.2)$$

where $m_1, m_2, m_3 = 1, \dots, P$ and we choose $P = 7$ for this study. In Eq. (3.2), all combinations of three atoms of the molecule are taken into account. In this chapter, we use the local variant of F_{3B, Z_1, Z_2, Z_3} (see Sec. 2.2), where we select three-body interactions, which are formed by two sets of bonded atoms which have a common atom. We define two atoms to be bonded, if their euclidean distance is smaller than the threshold function $B(Z_1, Z_2) := 1.1 \cdot L(Z_1, Z_2)$ and the values for the bond length function L given in Tab. 2.1. For the whole set of three-body descriptors F_{3B} , we concatenate the descriptors F_{3B, Z_1, Z_2, Z_3} for the set of 3-tuples of atomic numbers (Z_1, Z_2, Z_3) present in the data set.

SchNet

The neural network SchNet [97] is a variant of the earlier proposed deep tensor neural networks [2] and is based on the principle of learning atom-wise representations directly from first-principles. Given the atoms of type Z_1, \dots, Z_N , initial atom embeddings $\mathbf{x}_{Z_i}^{(0)} \in \mathbb{R}^{n_F}$, where n_F is the dimension of the feature space, depend only on the atom type. Then, a series of pairwise interaction refinements

$$\mathbf{x}_i^{t+1} = \mathbf{x}_i^t + \sum_{j \neq i} V^t(\mathbf{x}_j^t, \|\mathbf{r}_{ij}\|)$$

introduces information about the chemical environment into the embeddings. In SchNet, this is modeled using continuous-filter convolutions with filter-generating networks [91]. Through multiple of these corrections, SchNet is able to include complex many-body terms in the representation. Finally, an output neural network O predicts atom-wise property contributions, such that the final prediction is $\hat{y} = \sum_{i=1}^N O(\mathbf{x}_i^{(n_T)})$ for extensive properties and $\hat{y} = \frac{1}{N} \sum_{i=1}^N O(\mathbf{x}_i^{(n_T)})$ for intensive properties. During training, the initial embedding vectors $\mathbf{x}_{Z_i}^{(0)}$ as well as the parameters of the interaction network V and the output network O are optimized. In this chapter, we use $n_T = 6$ interaction refinements and $n_F = 64$ feature dimensions.

3.2 Experiments

We use the 21786 molecules from the GDB-9 benchmark dataset with up to 8 heavy atoms of type CNOF. GDB-9 includes relaxed geometries and properties computed using DFT at the B3LYP/6-31G(2df,p) level of theory [69]. This data set was previously used to predict deviations from reference second-order approximate coupled-cluster (CC2) singles and doubles spectra from their TDDFT counterparts [22]. The singlet-singlet transition energies from the

3. CAPTURING INTENSIVE AND EXTENSIVE MOLECULAR PROPERTIES WITH MACHINE LEARNING

Method	U0	α	gap	E_1	E_2
mean pred.	185.0	6.27	25.4	22.4	18.0
CM	4.8	0.60	7.8	12.7	10.2
BOB	2.3	0.36	4.8	11.5	9.6
F_{2B}	2.9	0.45	5.8	11.6	9.5
F_{3B}	2.9	0.45	5.3	11.3	9.4
$F_{2B} + F_{3B}$	1.1	0.33	4.6	11.1	9.2
SchNet	1.0	0.22	3.4	11.4	10.0

Table 3.1: Mean absolute errors of predicting the atomization energy (U0), isotropic polarizability (α), difference between the HOMO and LUMO energies (gap) and the transition energy to the first (E_1) and second (E_2) electronic excited singlet state. The properties U0, α and gap were calculated with DFT at the B3LYP/6-31G(2df,p) level of theory, the transition energies were calculated with LR-TDDFT at the PBE0/def2TZVP level of theory. The energy units are kcal/mol, the polarizability is given in units Bohr³. Best results are marked bold. The results in this table have appeared in previously published work [54].

ground state to the first and second excited state were calculated at the LR-TDDFT [98] level employing the hybrid XC functional PBE0 [59, 99] with def2TZVP basis set [100]. Instead of applying the delta learning approach [22], we attempt to learn the transition energies *directly*. We additionally use the atomization energy U0, isotropic polarizability α and HOMO-LUMO gap from GDB-9 for evaluation.

For all models, we use 10k random molecules for training and the remaining unseen 11786 molecules for computing the prediction errors. The results are listed in Tab. 3.1. For the CM and BOB descriptors, the Laplace kernel has been used. For the two- and three-body descriptors F_{2B} and F_{3B} , the Gaussian kernel achieves smaller prediction errors.

The mean predictor (mean pred.) is given by the average value of the property to be predicted. In general, the mean predictor yields an upper bound of the mean absolute prediction error of the machine learning models under investigation. While the Coulomb matrix (CM) uniquely encodes the structure of a given molecule, it performs worst of the evaluated descriptors. A major reason for this is that it implies a similarity measure of atom types based on the Coulomb interaction of the atomic nuclei, which does not reflect chemistry well. The bag-of-bonds (BOB) model is an extension to the Coulomb matrix where atom types are sorted into bags, thereby avoiding an unsuited atom similarity. This significantly boosts the performance compared to the CM for the atomization energy, polarizability and gap. Still, the bags are not invariant to atom indexing, which allows for multiple possible descriptors of the same

molecule.

The F_{2B} descriptors solve some of the sorting problems encountered in the CM and BOB representation. The prediction error of the F_{2B} descriptors is significantly better than the CM result, while being slightly worse compared to the BOB model. Combining the local three-body descriptors F_{3B} with the two-body descriptors F_{2B} significantly increases the predictive performance of the atomization energy, polarizability and the HOMO-LUMO gap (62%, 27% and 21% improvement, respectively). For the transition energies E_1 and E_2 , only a minor performance gain is observed by including the local three-body descriptors F_{3B} . SchNet slightly improves upon the descriptors $F_{2B} + F_{3B}$ for the prediction of the atomization energy, polarizability and gap, indicating that these properties can be well-represented by atom-wise contributions.

For the transition energies, the two- and three-body descriptors do not improve upon the performance of the baseline methods BOB and CM as much as for the extensive properties. Moreover, SchNet only achieves a performance that lies on the level of CM and BOB. As SchNet is able to include complex many-body terms in the representation, the non-locality of the transition energies do not allow a decomposition into atomic contributions. This indicates the need for much more complex global many-body terms for predicting transition energies, possibly encoding higher order interactions with order larger than three.

As most descriptors are either size-dependent or encode a sum or average term over local many-body interactions, they are naturally better suited to predict extensive properties. Such descriptors are typically limited by the order of the explicitly included many-body interactions. This can be a problem for predicting more complex quantum mechanical properties as demonstrated by the HOMO-LUMO gap, where SchNet performs better than explicit pairwise and three-body interaction descriptors. For the transition energies, SchNet does not improve upon the $F_{2B} + F_{3B}$ result. Even though the three-body descriptors are only applied to local bond angles, they perform better than two-body descriptors. In light of the SchNet results, this indicates that explicit many-body terms are more suitable to model transition energies using machine learning. As SchNet is designed to include high-order *local* interactions, we speculate on the need to develop *global* descriptors for intensive properties. As such properties are in general more difficult to predict than their localized counterparts, we conjecture that such kind of descriptors will describe both extensive and intensive properties on equal footing. In addition, as seen by the learning curves in Fig. 3.1, more data may be exceedingly helpful for further improving the predictive performance of the intensive properties under investigation.

3.3 Summary and discussion

We have evaluated a variety of machine learning techniques for intensive and extensive properties. As expected, all of them perform better on extensive properties than on intensive quantities. For the gap, SchNet performs 25%

3. CAPTURING INTENSIVE AND EXTENSIVE MOLECULAR PROPERTIES WITH MACHINE LEARNING

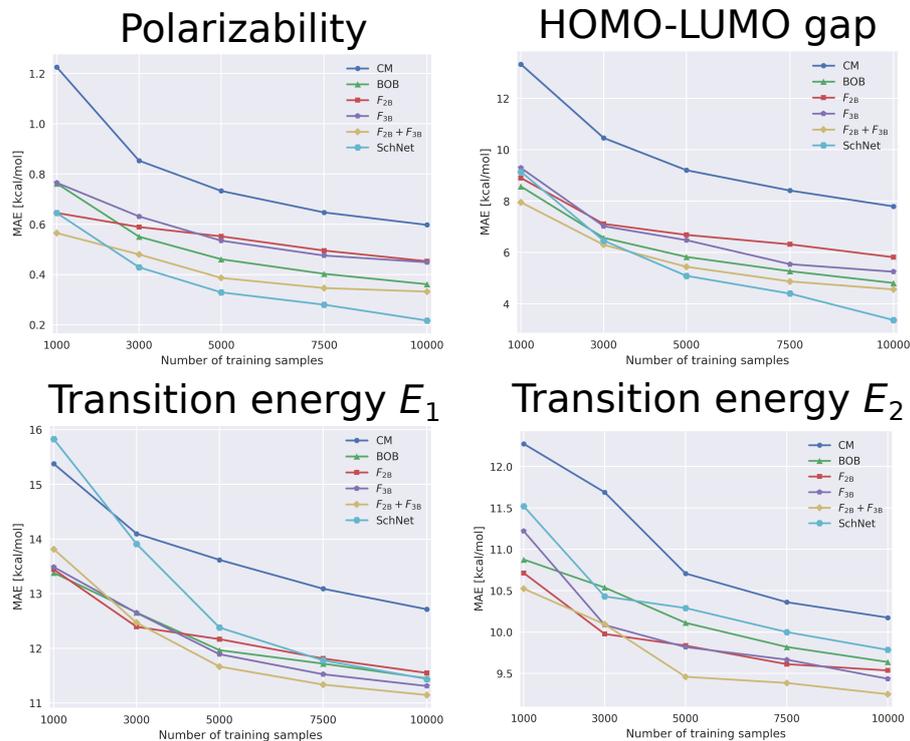


Figure 3.1: Mean absolute error of predicting the B3LYP/6-31G(2df,p) polarizability in Bohr³ (top left), the B3LYP/6-31G(2df,p) HOMO-LUMO gap in kcal/mol (top right), the singlet-singlet transition energy TDPBE0/def2TZVP- E_1 (bottom left) and TDPBE0/def2TZVP- E_2 (bottom right) in kcal/mol in dependence of the number of training samples. The model hyperparameters have been determined by 10-fold cross validation. These images have appeared in previously published work [54].

better than the explicit combination of pairwise and three-body descriptors. As SchNet is able to include complex many-body terms in principle, this result indicates the need for descriptors with many-body interactions with order larger than three for predicting the HOMO-LUMO gap. For the intensive properties E_1 and E_2 , the three-body descriptors work best, in particular combined with the two-body terms. In contrast, the decomposition into atom-wise contributions of SchNet, while working well for extensive properties, can be considered a drawback when attempting to predict transition energies by the averaging approach in the last output layer of SchNet.

Still, even with the best-performing descriptors the error of transition energy prediction may still be too high for any practical use. More advanced non-local descriptors will be necessary to predict transition energies more accurately, possibly encoding higher many-body terms or electronic state infor-

3.3. Summary and discussion

mation. In addition, as seen by the learning curves, more data may be exceedingly helpful for further improving the predictive performance of the intensive properties under investigation.

Kernel representations of quantum mechanical systems

Kernel based learning methods [101, 102, 15, 103, 104] allow an efficient convex solution of highly non-linear optimization problems often encountered in quantum chemistry. One possible advantage of using kernels is the relatively low number of training samples to achieve a certain accuracy as compared to highly parametric models like artificial neural networks. Additionally it might be easier for a chemist or physicist to incorporate domain-specific knowledge into the model, especially in view of the possibility of kernel compositions which can comply well with the many-body expansion of certain quantum-mechanical properties. A common task for the practitioner is to find a (kernel) representation of the problem at hand which encodes the distribution of the data in a complete, unique and efficient way [48], favorably taking into account the inherent symmetries of the system such as rotational, translational and atomic indexing invariance. As typical settings for a chemist or physicist include a low number of data points paired with a highly non-linear learning problem, kernel based formulations are considered as suitable and powerful methods of choice.

In the previous two chapters, we have developed a set of molecular descriptors which have been used in combination with exponential kernels like the Gaussian and Laplace kernel to predict various quantum-mechanical properties with kernel ridge regression (KRR). The invariant many-body descriptors presented in Chap. 2 are composed of two- and three-body combinations of atoms which are combined by taking the sum over their respective feature representation to ensure invariance with respect to atom indexing. Here, we investigate how these many-body decompositions can be used to construct a similarity measure represented by the kernel *directly*. The Figs. 4.1 and 4.2 schematically show these two conceptually different approaches. Such composite kernels have appeared previously in the literature [105, 44, 106, 4] and typically encode prior knowledge about the learning problem at hand.

In this chapter, we propose novel composite kernels which contain relatively

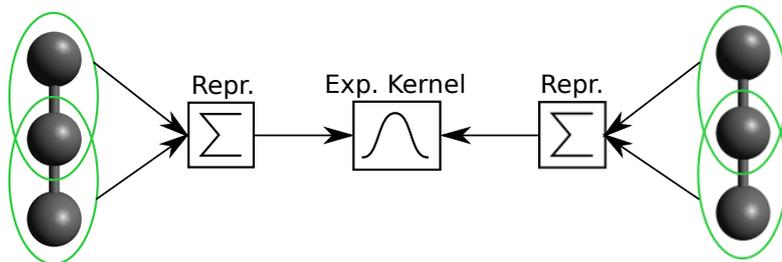


Figure 4.1: Schematic construction of an exponential kernel (Exp. Kernel) as similarity measure between two linear C-C-C carbon chains using many-body descriptors. From the decomposition of these exemplary molecules (green ellipses), local features are computed which are then combined by a sum to yield the final feature vector or representation (Repr.). The resulting feature vectors of both molecules are compared by an exponential kernel like the Gaussian or the Laplace kernel, respectively.

little prior chemical knowledge. Intuitively, our similarity measures will distinguish chemical environments by the set of atom types composing the two- and three-body terms which will be defined in the next section. By doing so, correlations between groups of atoms with different types are attenuated and the computation of our kernels is faster compared to the case including explicit correlation terms, as less many-body combinations have to be compared with each other. As we will see, in spite of encoding relatively little prior chemical knowledge, our models are able to learn interaction potentials which comply with chemical intuition.

This chapter is structured as follows. In Sec. 4.1, our local kernels are proposed. We extensively test kernel methods based on these local kernels in Sec. 4.2 on both stable molecules and molecular dynamics data sets as well as on more controlled experiments for an ethanol data set, where we analyse the interaction potentials learned by our models. The summary and discussion in Sec. 4.3 completes the chapter.

4.1 Local invariant kernels

We propose a set of kernels which are composed of many-body interaction terms and which are invariant in terms of the indexing of the atoms of both participating compounds. Similarly to the previous Chap. 2, we define a physical system by the set $S = \{\mathbf{r}_i, Z_i\}_{i=1}^N$ where N is the total number of atoms. Associated with this system, we denote the set of all combinations of k -tuples of atoms by

$$T_k(S) := \{((Z_{i_1}, \mathbf{r}_{i_1}), (Z_{i_2}, \mathbf{r}_{i_2}), \dots, (Z_{i_k}, \mathbf{r}_{i_k}))\}_{(i_1, i_2, \dots, i_k) \in G(k, N)} \quad (4.1)$$

where $G(k, N)$ is the k -permutation of N set as defined in Sec. 2.2. Along with an element of this set of tuples $t = ((Z_{i_1}, \mathbf{r}_{i_1}), (Z_{i_2}, \mathbf{r}_{i_2}), \dots, (Z_{i_k}, \mathbf{r}_{i_k})) \in T_k(S)$

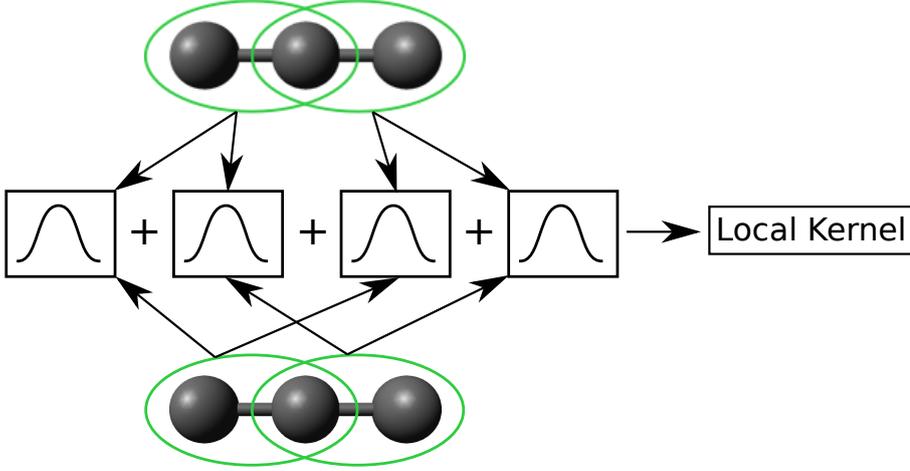


Figure 4.2: Schematic construction of a local kernel as similarity measure between two linear C-C-C carbon chains. From the decomposition of these exemplary molecules (green ellipses), local features are computed and all pairwise combinations of these features are compared with each other using an exponential kernel of choice like the Gaussian or the Laplace kernel, respectively. These local similarity measures are then combined by a sum to yield the final local kernel.

we define

$$\mathbf{R}(t) := \{\|\mathbf{r}_{i_p} - \mathbf{r}_{i_q}\|^{-1}\}_{p,q \in G(2,k)} \quad (4.2)$$

$$\mathbf{Z}(t) := (Z_{i_1}, Z_{i_2}, \dots, Z_{i_k})_{\text{sorted}} \quad (4.3)$$

where $Z_{i_1} \leq Z_{i_2} \leq \dots \leq Z_{i_k}$. The vector $\mathbf{R}(t)$ comprises the set of inverse pairwise distances of the given k -tuple of atoms t . For example if $k = 2$ we get $\mathbf{R}(t) = \{\|\mathbf{r}_{i_1} - \mathbf{r}_{i_2}\|^{-1}\}$ and in the case of $k = 3$ we have $\mathbf{R}(t) = \{\|\mathbf{r}_{i_1} - \mathbf{r}_{i_2}\|^{-1}, \|\mathbf{r}_{i_1} - \mathbf{r}_{i_3}\|^{-1}, \|\mathbf{r}_{i_2} - \mathbf{r}_{i_3}\|^{-1}\}$, respectively. The general form of our proposed local kernels is now given by

$$K(S_1, S_2) := \sum_{\substack{t_1 \in T_k(S_1) \\ t_2 \in T_k(S_2)}} \delta_{\mathbf{Z}(t_1), \mathbf{Z}(t_2)} \cdot D(\mathbf{R}(t_1)) \cdot D(\mathbf{R}(t_2)) \cdot K_{\text{exp}}(\|\mathbf{R}(t_1) - \mathbf{R}(t_2)\|) \quad (4.4)$$

with the kernel function $K_{\text{exp}}(\cdot)$ which can be an arbitrary exponential kernel and a damping function $D(\cdot)$ which can be viewed as a weighting term and will be defined in the next section.

Algorithm 3 Kernels2B**Input:**molecule $M_1 = \{(Z_i, \mathbf{r}_i)\}_{i=1}^{n_1}$ molecule $M_2 = \{(\bar{Z}_i, \bar{\mathbf{r}}_i)\}_{i=1}^{n_2}$ σ

▷ kernel parameter

 c_T

▷ cutoff distance

Output: $K_{2B}(M_1, M_2)$

```

1:  $K_{2B} \leftarrow 0$ 
2: for  $i, j \leftarrow G(2, n_1)$  do
3:    $R_{ij} \leftarrow \|\mathbf{r}_i - \mathbf{r}_j\|$ 
4:    $D_{ij} \leftarrow f_{c_T}(R_{ij})$ 
5: for  $p, q \leftarrow G(2, n_2)$  do
6:    $\bar{R}_{pq} \leftarrow \|\bar{\mathbf{r}}_p - \bar{\mathbf{r}}_q\|$ 
7:    $\bar{D}_{pq} \leftarrow f_{c_T}(\bar{R}_{pq})$ 
8: for  $i, j \leftarrow G(2, n_1)$  do
9:   for  $p, q \leftarrow G(2, n_2)$  do
10:     $\mathbf{Z} \leftarrow$  sorted tuple  $(Z_i, Z_j)$ 
11:     $\bar{\mathbf{Z}} \leftarrow$  sorted tuple  $(\bar{Z}_p, \bar{Z}_q)$ 
12:    if  $\mathbf{Z} == \bar{\mathbf{Z}}$  then
13:       $K_{2B} += D_{ij} \cdot \bar{D}_{pq} \cdot e^{-\|R_{ij} - \bar{R}_{pq}\|^2 / (2 \cdot \sigma^2)}$ 
14: return  $K_{2B}$ 

```

Two-body kernels

For the two-body kernels, our intuition is to model the interaction potential for a given pair of atom types as an univariate function of the corresponding pairwise distance. Our two-body kernels are defined by

$$K_{2B}(S_1, S_2) := \sum_{\substack{t_1 \in T_2(S_1) \\ t_2 \in T_2(S_2)}} \delta_{\mathbf{Z}(t_1), \mathbf{Z}(t_2)} \cdot D(\mathbf{R}(t_1)) \cdot D(\mathbf{R}(t_2)) \cdot e^{-\|\mathbf{R}(t_1) - \mathbf{R}(t_2)\|^2 / (2 \cdot \sigma^2)} \quad (4.5)$$

where for a given element $t = ((Z_{i_1}, \mathbf{r}_{i_1}), (Z_{i_2}, \mathbf{r}_{i_2})) \in T_2(S)$ we set

$$\mathbf{R}(t) = \|\mathbf{r}_{i_1} - \mathbf{r}_{i_2}\|^{-1} \quad (4.6)$$

$$D(\mathbf{R}(t)) := f_{c_T}(\|\mathbf{r}_{i_1} - \mathbf{r}_{i_2}\|) \quad (4.7)$$

$$f_{c_T}(x) := \begin{cases} \cos^2\left(\frac{x}{c_T} \cdot \frac{\pi}{2}\right) & x \leq c_T \\ 0 & \text{otherwise} \end{cases} \quad (4.8)$$

with the cutoff distance c_T and the Gaussian kernel parameter σ , respectively. The pseudocode for generating these kernels for molecules is given in Alg. 3.

Three-body kernels

Similarly to the two-body case, we model the interaction potential for a given triple of atom types as a function of the corresponding triple of pairwise distances. While the use of inverse pairwise distances naturally induces a damping of larger distances, we find it beneficial to explicitly include the damping function $D(\cdot)$, as the number of three-body combinations rapidly grows for increasing system size. The sorting of the triple of atomic numbers \mathbf{Z} is necessary to compare pairwise distances of according atom types with each other. In the three-body case, the similarity measure between two systems S_1 and S_2 is defined by

$$K_{3B}(S_1, S_2) := \sum_{\substack{t_1 \in T_3(S_1) \\ t_2 \in T_3(S_2)}} \delta_{\mathbf{Z}(t_1), \mathbf{Z}(t_2)} \cdot D(\mathbf{R}(t_1)) \cdot D(\mathbf{R}(t_2)) \cdot e^{-\|\mathbf{R}(t_1) - \mathbf{R}(t_2)\|^2 / (2 \cdot \sigma^2)} \quad (4.9)$$

with $t = ((Z_{i_1}, \mathbf{r}_{i_1}), (Z_{i_2}, \mathbf{r}_{i_2}), (Z_{i_3}, \mathbf{r}_{i_3})) \in T_3(S)$ and

$$D(\mathbf{R}(t)) := f_{c_T}(\|\mathbf{r}_{i_1} - \mathbf{r}_{i_2}\|) \cdot f_{c_T}(\|\mathbf{r}_{i_1} - \mathbf{r}_{i_3}\|) \cdot f_{c_T}(\|\mathbf{r}_{i_2} - \mathbf{r}_{i_3}\|) \quad (4.10)$$

with the cutoff distance c_T and the Gaussian kernel parameter σ , respectively. Being composed of a possibly large sum of three-body terms, we potentially smear useful structural information about the quantum mechanical system. However, the convex nature of the optimization problem defined by kernel ridge regression can be viewed as simultaneously optimizing the composite kernels. The pseudocode for generating the three-body kernels for molecules is presented in Alg. 4.

4.2 Tests on molecular data sets

In a first experiment, we test our kernels on the sets of stable organic molecules already used in the previous chapters GDB-7 and GDB-9, respectively. There, our invariant many-body descriptors have provided accurate models for the atomization energy by using a combination of two- and three-body features. Encoding two- and three-body interactions directly into the similarity measure makes our proposed kernels particularly suitable for these data sets. Specifically, we hope to circumvent the problem of generating a potentially large number of features while maintaining a good prediction accuracy using our kernels. More specifically, we investigate the conjecture of Sec. 2.4 which states that for the atomization energy, the interaction potential of the hydrogen atom with all other atoms can be effectively modeled as a pairwise potential. To this end, we define a local kernel variant by explicitly excluding interactions containing

Algorithm 4 Kernels3B**Input:**molecule $M_1 = \{(Z_i, \mathbf{r}_i)\}_{i=1}^{n_1}$ molecule $M_2 = \{(\bar{Z}_i, \bar{\mathbf{r}}_i)\}_{i=1}^{n_2}$ σ

▷ kernel parameter

 c_T

▷ cutoff distance

Output: $K_{3B}(M_1, M_2)$

```

1:  $K_{3B} \leftarrow 0$ 
2: for  $i, j \leftarrow G(2, n_1)$  do
3:    $R_{ij} \leftarrow \|\mathbf{r}_i - \mathbf{r}_j\|$ 
4:    $D_{ij} \leftarrow f_{c_T}(R_{ij})$ 
5: for  $p, q \leftarrow G(2, n_2)$  do
6:    $\bar{R}_{pq} \leftarrow \|\bar{\mathbf{r}}_p - \bar{\mathbf{r}}_q\|$ 
7:    $\bar{D}_{pq} \leftarrow f_{c_T}(\bar{R}_{pq})$ 
8: for  $i, j, k \leftarrow G(3, n_1)$  do
9:   for  $p, q, r \leftarrow G(3, n_2)$  do
10:     $\mathbf{Z} \leftarrow$  sorted tuple  $(Z_i, Z_j, Z_k)$ 
11:     $\bar{\mathbf{Z}} \leftarrow$  sorted tuple  $(\bar{Z}_p, \bar{Z}_q, \bar{Z}_r)$ 
12:    if  $\mathbf{Z} == \bar{\mathbf{Z}}$  then
13:       $D \leftarrow D_{ij} \cdot D_{ik} \cdot D_{jk} \cdot \bar{D}_{pq} \cdot \bar{D}_{pr} \cdot \bar{D}_{rq}$ 
14:       $K_{3B} += D \cdot e^{-((R_{ij}-\bar{R}_{pq})^2+(R_{ik}-\bar{R}_{pr})^2+(R_{jk}-\bar{R}_{qr})^2)/(2 \cdot \sigma^2)}$ 
15: return  $K_{3B}$ 

```

hydrogen in the three-body kernel

$$\begin{aligned} \bar{K}_{3B}(S_1, S_2) := & \sum_{\substack{t_1 \in T_3(S_1) \\ t_2 \in T_3(S_2)}} \theta(\mathbf{Z}(t_1)) \cdot \theta(\mathbf{Z}(t_2)) \cdot \delta_{\mathbf{Z}(t_1), \mathbf{Z}(t_2)} \\ & \cdot D(\mathbf{R}(t_1)) \cdot D(\mathbf{R}(t_2)) \cdot e^{-\|\mathbf{R}(t_1) - \mathbf{R}(t_2)\|^2 / (2 \cdot \sigma^2)} \end{aligned} \quad (4.11)$$

with the indicator function

$$\theta((Z_1, Z_2, Z_3)) := \begin{cases} 0 & Z_1 = 1 \vee Z_2 = 1 \vee Z_3 = 1 \\ 1 & \text{otherwise} \end{cases} \quad (4.12)$$

To predict molecular properties, we apply the KRR technique as described in Sec. 1.2 paired with our local kernels as defined in Sec. 4.1. For kernel ridge regression, a common preprocessing step is to subtract the mean of the predicted target variable which originates from the stochastic framework of Gaussian processes which assumes a prior zero-mean distribution of the estimated functions. However, using our local kernels we observe that this step can be omitted, indicating that these kernels intrinsically encode chemically suitable information about a given molecular property. We apply 10-fold nested cross validation as describes in Sec. 1.2 to select the parameters of our kernels and sample

Property	$F_{2B} + F_{3B}$	K_{2B}	$K_{2B} + K_{3B}$	$K_{2B} + \tilde{K}_{3B}$	Unit
ae-pbe0	0.8	2.9	0.50	0.75	kcal/mol
homo-gw	0.13	0.45	0.17	0.23	eV
lumo-gw	0.15	0.22	0.14	0.16	eV
homo-pbe0	0.12	0.36	0.15	0.19	eV
lumo-pbe0	0.11	0.18	0.086	0.11	eV
homo-zindo	0.13	0.50	0.21	0.28	eV
lumo-zindo	0.11	0.31	0.20	0.23	eV
p-pbe0	0.088	0.16	0.056	0.062	Ångström ³
p-scs	0.032	0.12	0.038	0.055	Ångström ³
e1-zindo	0.15	0.64	0.43	0.48	eV
ea-zindo	0.13	0.37	0.24	0.28	eV
imax-zindo	0.071	0.086	0.069	0.068	a.u.
emax-zindo	1.26	1.54	1.30	1.34	eV
ip-zindo	0.18	0.51	0.26	0.32	eV

Table 4.1: Mean absolute prediction errors of several ground- and excited state properties by kernel ridge regression trained on 5000 random molecules and tested on the remaining 1868 molecules that were not used for training or validation of the GDB-7 data set. The properties are described in Tab. 2.6. The best performing models are marked in bold.

the molecules for training randomly. The results of predicting the molecular properties of the data sets GDB-7 and GDB-9 are shown in Tabs. 4.1 and 4.2, respectively.

For the atomization energy, the local kernel $K_{2B} + K_{3B}$ outperforms the CM, BOB and our methods based on invariant two- and three-body descriptors $F_{2B} + F_{3B}$, respectively. Specifically, this model achieves a mean absolute error of 0.50 kcal/mol for the molecules of the GDB-7 set and 0.88 kcal/mol for the molecules of the GDB-9 set when trained on 5000 random molecules and tested on the remaining molecules which corresponds to an 60% improvement over the $F_{2B} + F_{3B}$ descriptors. These results suggest that local kernels are more efficient for predicting extensive properties compared to our invariant two- and three-body descriptors for equilibrium molecules. A similar performance gain is achieved for the zero point vibrational energy (zpve), where $K_{2B} + K_{3B}$ reaches 0.066 kcal/mol. On the other hand, for the properties based on the energies of the HOMO and the LUMO, the invariant descriptors $F_{2B} + F_{3B}$ outperform the local kernels $K_{2B} + K_{3B}$ with exception of the property lumopbe0 of the molecules of the set GDB-7. This indicates that the invariant descriptors $F_{2B} + F_{3B}$ have an improved ability to capture non-local information as compared to the kernel $K_{2B} + K_{3B}$. This effect is even more dominant

4. KERNEL REPRESENTATIONS OF QUANTUM MECHANICAL SYSTEMS

Property	$F_{2B} + F_{3B}$	K_{2B}	$K_{2B} + K_{3B}$	$K_{2B} + \bar{K}_{3B}$	Unit
U0	1.5	4.5	0.88	1.2	kcal/mol
U	1.5	4.5	0.89	1.2	kcal/mol
H	1.5	4.5	0.89	1.2	kcal/mol
G	1.5	4.4	0.90	1.2	kcal/mol
HOMO	3.6	7.6	4.2	5.3	kcal/mol
LUMO	5.1	7.9	6.0	6.2	kcal/mol
gap	6.2	11.4	7.5	8.4	kcal/mol
alpha	0.49	0.82	0.50	0.51	Bohr ³
mu	0.61	0.80	0.67	0.71	Debye
r2	9.0	83	46	70	Bohr ²
zpve	0.10	0.17	0.066	0.097	kcal/mol
A	0.42	0.62	0.54	0.57	GHz
B	0.13	0.26	0.17	0.22	GHz
C	0.050	0.17	0.097	0.13	GHz
cv	0.12	0.30	0.11	0.18	cal/(mol K)

Table 4.2: Mean absolute prediction errors of several properties calculated at the B3LYP/6-31G(2df,p) level of quantum chemistry and predicted by kernel ridge regression trained on 5000 random molecules and tested on the remaining 126722 molecules that were not used for training or validation of the GDB-9 data set. The properties are described in Tab. 2.7. The best performing descriptors are marked in bold.

for the properties el-zindo, ea-zindo and ip-zindo of the molecules of the set GDB-7 and the properties r2, A, B and C of the molecules of the GDB-9 set, respectively. The two models $F_{2B} + F_{3B}$ and $K_{2B} + K_{3B}$ perform with similar mean absolute error for the polarizabilities p-pbe0 and p-scs, imax-zindo and emax-zindo of the GDB-7 molecules and the properties mu, r2 and cv of the molecules of the GDB-9 set. We already showed the learning curves for all our models in the previous Chap. 2. Figs. 4.3 and 4.4 show the mean absolute error of predicting the atomization energy of the data sets GDB-7 and GDB-9 for the four combinations of two- and three-body models $F_{2B} + F_{3B}$, $F_{2B} + \bar{F}_{3B}$, $K_{2B} + K_{3B}$ and $K_{2B} + \bar{K}_{3B}$ in dependence of the number of training samples in greater detail.

The best overall performing model $K_{2B} + K_{3B}$ achieves the chemical accuracy of 1 kcal/mol at 1500 and 4000 number of training molecules for the GDB-7 and GDB-9 set, respectively. The proposed local kernel $K_{2B} + \bar{K}_{3B}$ which models the interactions of the hydrogen atom with all other atoms exclusively by pairwise potentials performs only slightly worse compared to the full kernel variant $K_{2B} + K_{3B}$. This substantiates our conjecture of Sec. 2.4 as

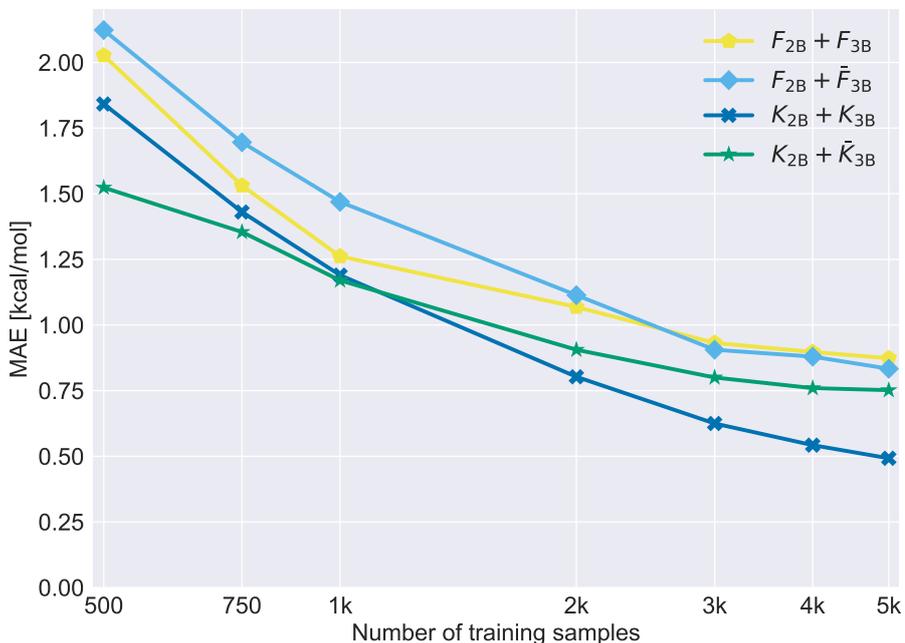


Figure 4.3: Mean absolute error of predicting the PBE0 atomization energy of the molecules of the set GDB-7 with KRR in dependence of the number of training samples. The errors are given in kcal/mol. For $F_{2B} + F_{3B}$ and $F_{2B} + \bar{F}_{3B}$, the Gaussian kernel has been used. The kernel parameters have been determined by 10-fold nested cross-validation.

this result indicates that it is possible to improve the performance even with such a considerable model restriction.

Equilibrium molecules form an important domain in chemical compound space as nature is dominated by stable molecules. However, for useful quantum chemistry models it is essential to additionally study non-equilibrium molecules which allows a wider range of applications and properties to be predicted. In a sense, the task of predicting equilibrium properties is of limited practical use except as a benchmark, as the data domain of such data sets lies outside the range of important quantum chemistry applications. Examples include the relaxation of geometries, generation of new molecules and the prediction of molecular forces. For the forces, recently more sophisticated models based on the Coulomb matrix have been designed [5, 6, 7]. We rely on the molecular dynamics (MD) data used in their work where ab initio molecular dynamics-quality thermodynamic observables using path-integral MD for organic molecules containing the four chemical elements C, N, O and H have been computed. The atomization energy for the alkanes in this work has been computed analogously.

As the indexing of the atoms of the molecules in the respective data set

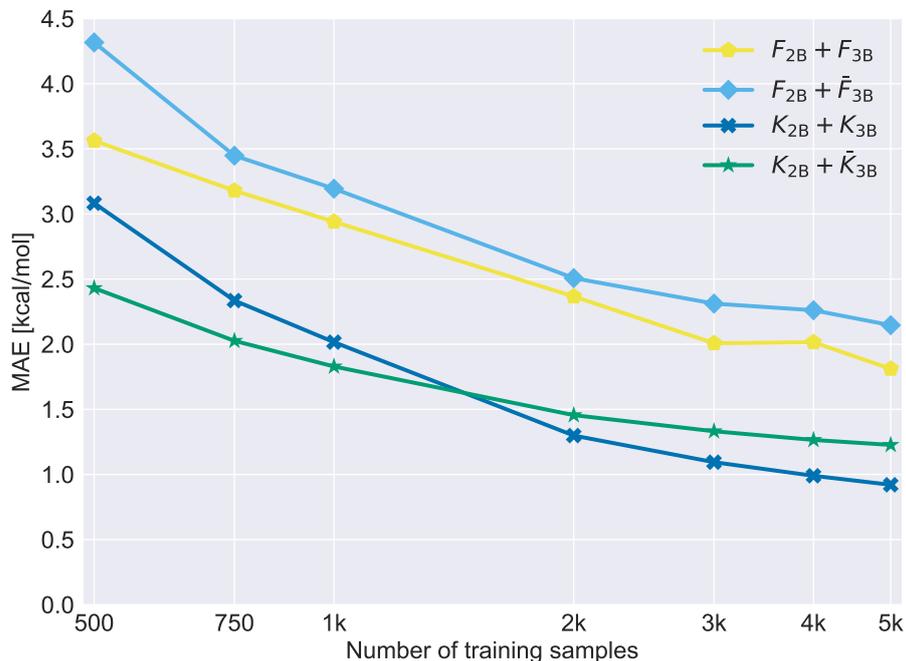


Figure 4.4: Mean absolute error of predicting the B3LYB/6-31G(2df,p) atomization energy of the molecules of the set GDB-9 with KRR in dependence of the number of training samples. The errors are given in kcal/mol. For $F_{2B} + F_{3B}$ and $F_{2B} + \bar{F}_{3B}$, the Gaussian kernel has been used. The kernel parameters have been determined by 10-fold nested cross-validation.

are fixed, we use a variant of the CM which is composed of inverse pairwise distances, thereby uniquely identifying the atomic positions of the molecule. The Gaussian kernel performs better than the Laplace kernel for CM and BOB which indicates a better conditioned learning problem than for the stable molecules sets due to the lack of somewhat artificially sorting the features entries in the descriptors. For our invariant two- and three-body descriptors we use the non-local variant $F_{2B} + F_{3B}$ due to the lack of the concept of bonding distances for the molecular dynamics data sets. We use a maximum exponent of 9 for both the F_{2B} and the F_{3B} descriptors, emphasizing the three-body interactions more than in the stable molecules sets GDB-7 and GDB-9. We adopt KRR to predict the atomization energy of the molecules along a molecular dynamics (MD) trajectory, where we apply 10-fold nested cross validation to select the parameters of our kernels and sample the molecules for training randomly. The results for predicting the atomization energy of the remaining molecules of the respective data set are shown in Tab. 4.3.

Our local kernel $K_{2B} + K_{3B}$ outperforms the other models for most of the molecules with the exceptions of the salicylic and uracil set, where the

4.2. Tests on molecular data sets

data set	std	CM	BOB	F_{2B}	$F_{2B} + F_{3B}$	K_{2B}	$K_{2B} + K_{3B}$
aspirin	6.1	3.7	3.1	3.3	1.2	3.0	1.0
benzene	5.5	0.17	0.60	0.69	0.22	0.66	0.18
azobenzene	6.5	3.2	1.7	2.0	0.77	1.9	0.60
ethanol	4.1	0.73	1.35	1.7	0.46	1.6	0.36
malonaldehyde	4.2	0.60	1.33	1.4	0.65	1.4	0.48
paracetamol	5.8	3.2	2.4	2.7	1.0	2.5	0.81
resorcinol	4.9	0.62	1.7	2.0	0.48	1.9	0.46
salicylic	5.5	0.48	1.8	2.1	0.66	1.8	0.56
naphthalene	5.5	0.52	1.1	1.4	0.51	1.2	0.37
toluene	4.9	0.64	1.1	1.4	0.50	1.3	0.40
uracil	4.9	0.30	1.4	1.4	0.33	1.3	0.44
methane	4.2	0.129	0.370	0.53	0.037	0.55	0.033
ethane	4.2	0.24	0.68	0.80	0.24	0.79	0.11
propane	4.7	0.78	0.94	1.1	0.39	1.1	0.25
butane	5.2	2.0	1.2	1.4	0.51	1.3	0.29
pentane	5.8	3.5	1.7	1.7	0.70	1.5	0.37
hexane	6.1	4.3	1.7	1.7	0.85	1.6	0.43
heptane	6.5	4.6	2.0	1.8	0.87	1.7	0.46
octane	6.8	5.1	2.0	2.0	0.90	1.8	0.49

Table 4.3: Mean absolute prediction errors of the atomization energy of the molecules along a MD trajectory by kernel ridge regression using the Gaussian kernel for the CM, BOB, F_{2B} and $F_{2B} + F_{3B}$ descriptors. The models have been trained on 1000 random molecules and tested on the remaining molecules of the respective data set. The first column `std` denotes the standard deviation of the target atomization energy. The best performing methods are marked in bold. Note that gradient-domain machine learning models including symmetries (sGDML) outperform our best models, as these models include forces in their learning procedure [6].

CM achieves a better performance and benzene, where both models CM and $K_{2B} + K_{3B}$ have similar mean absolute error when trained on 1000 random molecules. In contrast to the stable molecules sets, the BOB descriptor performs worse than the CM, indicating the disadvantageous feature sorting of BOB for the MD sets. Although the CM perform well for most of the molecules, it is challenged by larger molecules like aspirin, paracetamol and azobenzene, where the error is only slightly worse than the standard deviation of the target atomization energy. Some of the problems of the CM can be seen exemplary in the error dependence of the number of training samples for the benzene and

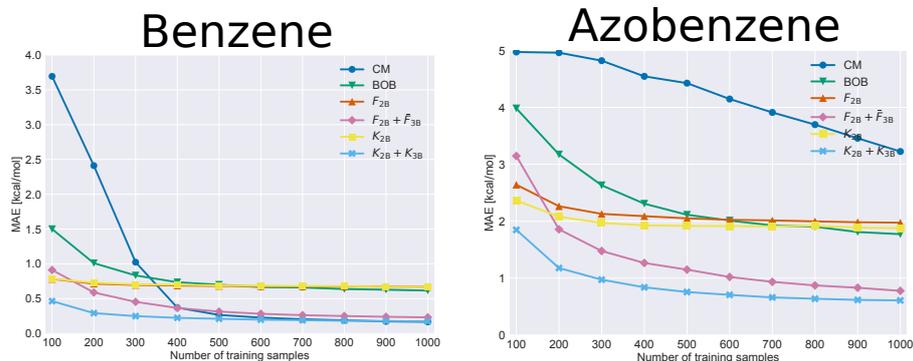


Figure 4.5: Mean absolute prediction errors of the atomization energy of the molecules of the MD benzene (left) and MD azobenzene set (right) with KRR in dependence of the number of training samples. The errors are given in kcal/mol. For the CM, BOB, F_{2B} and $F_{2B} + \bar{F}_{3B}$, the Gaussian kernel has been used. The kernel parameters have been determined by 10-fold nested cross-validation.

azobenzene molecule in Fig. 4.5. For benzene the learning curve is an order of magnitude steeper, which we attribute to the CM smearing the different oscillatory motions of azobenzene into the pairwise distances. Although the performance of the best model $K_{2B} + K_{3B}$ decreases accordingly, our local kernels have less problems in distinguishing these oscillations, as it is far easier to learn three-dimensional compared to $3n$ -dimensional energy surfaces, where n is the number of atoms of the molecule. By summing over sets of many-body interactions in the kernel definition, training the model can be viewed as simultaneously learning the corresponding interaction potentials. This is even more apparent for the alkane sets, where the performance of the CM quickly decreases for increasing the molecule size, indicating problems of the CM with linear chain molecules. The error of the local kernel $K_{2B} + K_{3B}$ grows much slower, showing superior learning efficiency for alkanes.

Our local kernels show good performance for the analyzed MD data sets in spite of encoding relatively little chemical knowledge. However, due to the decomposition into independent sums of kernels, our methods are more widely applicable. Specifically, we use our local kernels to analyse the two- and three-body interactions in chemical compounds in greater detail. To this end, we study the local energy surfaces learned by our models which are functions of the types of atoms defining the interaction given by the sorted tuple of atomic numbers \mathbf{Z} and their corresponding set of pairwise distances \mathbf{R} . The interaction

energies are defined by

$$E(\mathbf{R}, \mathbf{Z}) := \sum_{i=1}^N \alpha_i \cdot \sum_{t_1 \in T_k(S_i)} \delta_{\mathbf{Z}(t_1), \mathbf{Z}} \cdot D(\mathbf{R}(t_1)) \cdot D(\mathbf{R}) \cdot e^{-\|\mathbf{R}(t_1) - \mathbf{R}\|^2 / (2 \cdot \sigma^2)} \quad (4.13)$$

where the $\{\alpha_i\}_{i=1}^N$ are the model parameters trained on the full kernel and the $\{S_i\}_{i=1}^N$ comprise the training set of molecules, respectively. The total predicted energy can be expressed as the sum over these interaction energies. However, as sums of energies are somewhat difficult to interpret, we investigate these interaction energies for models trained with different combinations of kernels. Specifically, we analyse the two-body interactions for a model trained with the two-body and mixed kernel K_{2B} and $K_{2B} + K_{3B}$, respectively. Similarly, we investigate the three-body interactions for the K_{3B} and $K_{2B} + K_{3B}$ kernels.

To help with the interpretation of the interaction energies, we compute the density at a given point defined by \mathbf{Z} and \mathbf{R} by a common technique called kernel density estimation which is defined by

$$\text{KDE}_k(\mathbf{R}, \mathbf{Z}) := \frac{1}{N} \sum_{i=1}^N \sum_{t_1 \in T_k(S)} \frac{1}{N(\mathbf{Z}(t_1))} \cdot \delta_{\mathbf{Z}(t_1), \mathbf{Z}} \cdot e^{-\|\mathbf{R}(t_1) - \mathbf{R}\|^2 / (2 \cdot \sigma^2)} \quad (4.14)$$

where $k = 2$ for the two-body and $k = 3$ for the three-body density estimates and $N(\mathbf{Z}(t_1))$ is a normalization factor, respectively. In the following, we apply these methods for studying interaction potentials for a variety of molecules. To increase the reproducibility of the learned interaction energies, we perform an average over an ensemble of 30 models trained on 1000 random molecules from a total number of 30000 molecules in the data set.

First, we apply the above methods exemplary to some extent for the ethanol molecule for which the learning curve is shown in Fig. 4.6. The gradient of the learning curve for the K_{2B} and $K_{2B} + K_{3B}$ kernels is relatively flat at 1000 training points which indicates the models have converged at learning interaction potentials. These interaction potentials for ethanol are shown in Fig. 4.7, where we have selected the three-body interactions H-H-C, H-C-C and H-H-O for displaying purposes. In this notation, the two former atoms have the distance r_1 and r_2 to the latter atom and form a fixed angle of 120° unless mentioned otherwise. Noticeably, both the two- and three-body interaction energies decay to zero for large distances. This shows the ability of our models to learn chemically plausible energy surfaces, even at locations with a low number of data points.

Now, we analyse the two-body interactions in greater detail. For densities containing larger distances, the interaction energy shows lower variation, which indicates that our kernel models emphasize local interactions more. The bond length between two hydrogen atoms is approximately 0.74 Ångström. However, the H-H interaction energy of the two-body model K_{2B} displays a local minimum around 1.1 Ångström, which equals the C-H bonding distance. This is

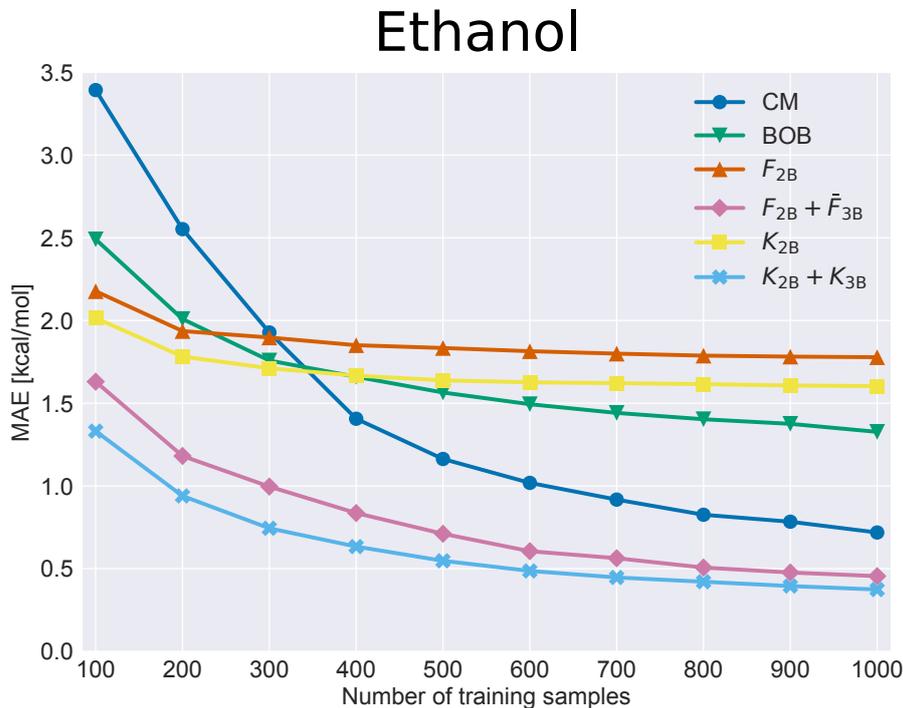


Figure 4.6: Mean absolute prediction errors of the atomization energy of the molecules of the ethanol data set with KRR in dependence of the number of training samples. The errors are given in kcal/mol. For the CM, BOB, $F_{2B} + F_{3B}$ and $F_{2B} + \bar{F}_{3B}$, the Gaussian kernel has been used. The kernel parameters have been determined by 10-fold nested cross-validation.

possibly due to the two-body model K_{2B} mapping a part of the C-H on the H-H interactions for the molecules of the ethanol data set. The local minimum at 1.1 Ångström for the H-H interaction vanishes if the two- and three-body kernel $K_{2B} + K_{3B}$ is used, indicating that the local minimum for the pairwise H-H interaction is a mapped three-body interaction containing carbon and hydrogen atoms, respectively. In contrast, the distance 1.1 Ångström remains a local minimum for the C-H interaction energy by using the combined kernel. For the C-C and C-O interactions, the two-body model learns a local minimum at 1.5 and 1.4 Ångström, which corresponds well with the true bonding distances in Tab. 2.1. However, these minima vanish as we include the three-body model, which we attribute to these interactions being better modeled by three-body terms. In contrast, the C-H and O-H interactions can be partially expressed by two-body terms. Interestingly, the two-body potentials display local minima at distance larger than 1.5 Ångström. We interpret these minima as the tendency of our models to pull the atoms together and thereby forming the molecule. This can be useful for applications in MD simulations as this property will keep

the models from diverging away from potentially non-interesting data regimes.

While the two-body interaction potentials change significantly by adding the three-body kernel, this effect is less dominant for the three-body potentials. Examples include the H-H-C and the H-H-O interactions at the selected angle of 120° in Fig. 4.7. When both distances are at 1.0 Ångström in the H-H-C energy surface (the bottom left corner of the heat map), the pure three-body kernel model K_{3B} is repulsive, indicating a mapping of two-body interactions. This effect vanishes by inclusion of the two-body kernel, enforcing the above statement, that the C-H interaction in ethanol can be modeled by a pairwise potential. The H-C-C interaction displays a structural change around the high-density region around the point with the coordinates at 1.5 and 1.0 Ångström which corresponds to the transformation of the two-body interaction C-H.

To better interpret the interaction potentials, we next examine molecules with a limited number of two- and three-body interactions, namely benzene and toluene together with the two alkanes propane and octane. Toluene is structurally similar to benzene, where one of the H atoms is replaced by a CH_3 group. Propane and octane are linear chain alkanes. The corresponding results are shown in Figs. 4.8, 4.9, 4.10 and 4.11, respectively.

For benzene and toluene, the local minima of the C-H and C-C interaction potentials lie at approximately 1.0 and 1.2 Ångström when trained with the two-body kernel K_{2B} . By using the mixed model $K_{2B} + K_{3B}$ the local C-H minimum becomes less dominant, indicating that for benzene and toluene the C-H interactions are better mapped by three-body terms. For propane and octane, this effect appears for both the C-H and C-C interaction potentials for which the minima lie at 1.0 and 1.4 Ångström when trained with the K_{2B} model. Both of these local minima vanish by using the combined kernel $K_{2B} + K_{3B}$, indicating that these interactions are better modeled by three-body terms in alkanes. Noticeably, the interaction potentials for benzene and toluene and accordingly for propane and octane are qualitatively similar. This demonstrates the ability our local kernel models to learn similar energy surfaces for chemically related molecules.

To gain even further understanding of our learned models, we perform more controlled experiments on the stable ethanol molecule where we selectively vary a single fixed structural degree of freedom. Specifically, we examine the following three cases

1. move a selected hydrogen atom along the corresponding C-H connection
2. rotate the dihedral angle of the CH_3 -group
3. vary the angle between two of the hydrogen atoms of the CH_3 -group

The corresponding results for the two- and three-body interaction potentials together with the total two- and three-body energies and kernel density estimates according to Eq. (4.14) are shown in Figs. 4.12, 4.13 and 4.14, respectively.

In the first experiment of varying the C-H distance in Fig. 4.12, both two- and three-body energies show a local minimum at 1.1 Ångström, approximately

the C-H bond length. However, at large distances the two-body model displays the lowest energy and therefore prefers the selected hydrogen atom to be removed from the ethanol molecule which is somewhat unphysical. This effect is corrected by adding three-body terms which lifts this asymptotic energy, resulting in a global minimum at the bond length of 1.1 Ångström. This further demonstrates, that our models learn chemically reasonable interaction potentials. If we look at the three-body interaction potentials, this global minimum is modeled dominantly by H-H-C interactions as perhaps chemically expected. Interestingly, the O-H interaction shows significantly less variation than the H-H and C-H potentials which is not encoded into the model directly. This indicates that in this experiment the energies are dominated by H-H and C-H interactions, respectively.

A similar phenomenon is observed in the varying H-C-H angle experiment in Fig. 4.14, where the two-body model learns the lowest total energy at larger angles and the combined two- and three-body kernel displays a global minimum around the equilibrium angle zero.

For the dihedral angle rotation experiment in Fig. 4.13, both local kernels show exact periodicity which is directly encoded into the kernels as the sum is taken over all permutations of the respective many-body interactions. By including the three-body terms, the average magnitude of the O-H energy gets lowered as compared to the H-H energy which we attribute to this interaction being partially modeled by the H-H-O term. Fig. 4.15 shows the comparison with the untrained model, where we set $\alpha_i = 1$ for $i = 1, \dots, N$ in Eq. (1.15). Although we encode relatively little chemical information into our kernels, even the untrained model shows the expected threefold periodicity. We attribute this to a chemically plausible smoothness for the interaction potentials induced by our kernels. Training the models shifts the local minima to the correct equilibrium position at $\varphi = 0$ and lowers the amplitude of the oscillations.

4.3 Summary and discussion

In this chapter, we have developed new kernels for quantum chemistry which are based on many-body interactions between the atoms forming a given compound. Except for the decomposition into the different types of two- and three-body interactions, these kernels encode no additional chemical knowledge, for example angles and bond lengths are not a priori included in the computation of the kernel. Nevertheless, our best model $K_{2B} + K_{3B}$ outperforms the CM, BOB and the invariant descriptors $F_{2B} + F_{3B}$ of the previous chapter for stable molecules and molecular dynamics data sets when predicting the atomization energy with kernel ridge regression. The superior prediction error is a consequence of the kernel decompositions into much lower dimensional interaction potentials, making the model learning more efficiently in practice than for example global descriptors like the CM, a phenomenon most dominant for a low number of training points.

We have analyzed the two- and three-body interaction potentials for a set

of non-equilibrium molecules with models trained by two-body kernels, three-body kernels and a combination of the two. This analysis has shown that our local kernels can learn models which are in agreement with chemical intuition. In a more controlled experiment where we varied a single degree of freedom of the ethanol molecule, we have enforced this tendency, as the model $K_{2B} + K_{3B}$ learns the global energy minimum near the true equilibrium position.

Paired with the high learning efficiency, our local kernels can potentially be used to study the relation of interaction potentials with the type, size and composition of molecules. In particular, we pose the question whether there are global interaction potentials for a set of structurally similar molecules and whether transfer learning between these molecules is feasible. In fact, our analysis indicates such possibilities for alkanes.

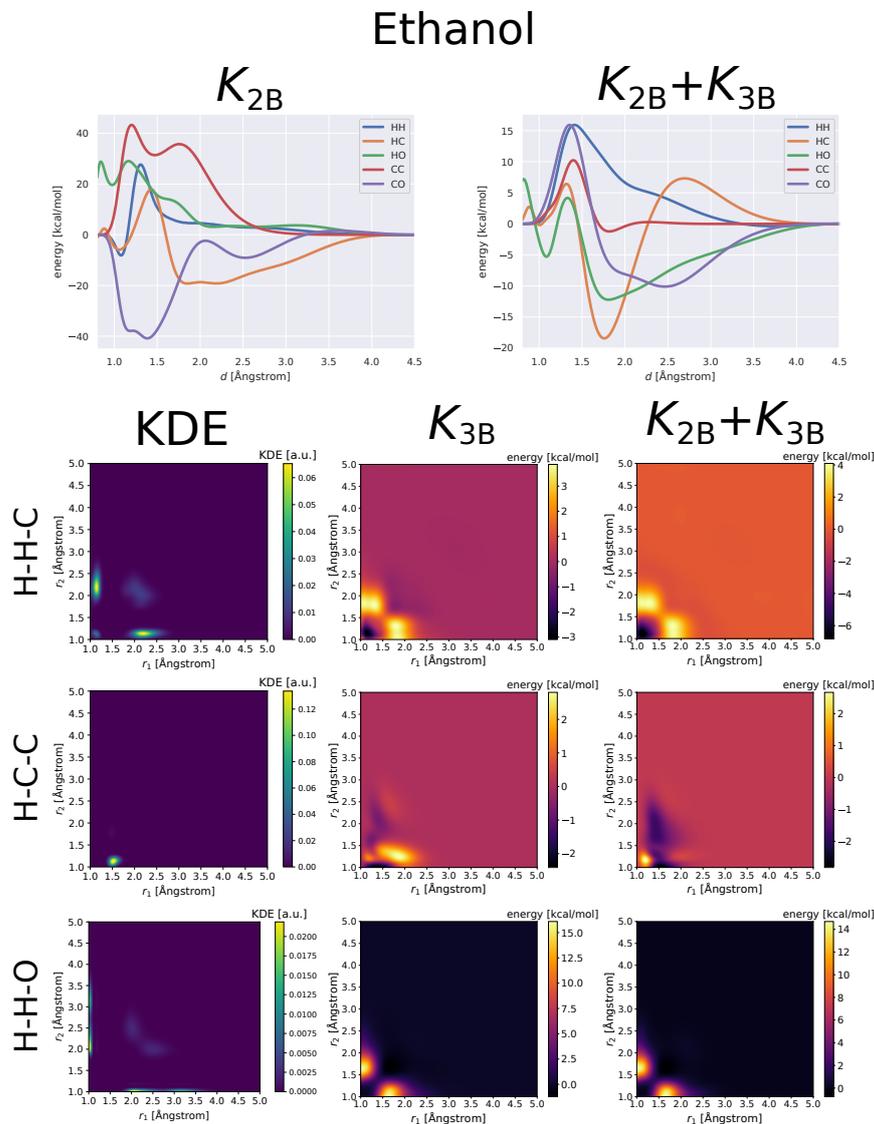


Figure 4.7: Interaction potentials of ethanol according to Eq. (4.13). Two-body potentials for a model trained with the two-body kernel K_{2B} (top left) and the mixed kernel $K_{2B} + K_{3B}$ (top right), respectively. The three-body potentials of the H-H-C (top row), H-C-C (middle row) and H-H-O (bottom row) potentials are shown for a model trained with the three-body kernel K_{3B} (middle column) and the mixed kernel $K_{2B} + K_{3B}$ (right column). The left column shows the kernel density estimate according to Eq. (4.14). All models have been trained on 1000 random molecules and the kernel parameters have been determined by 10-fold nested cross-validation. The interaction potentials have been averaged by an ensemble of 30 models to increase reproducibility.

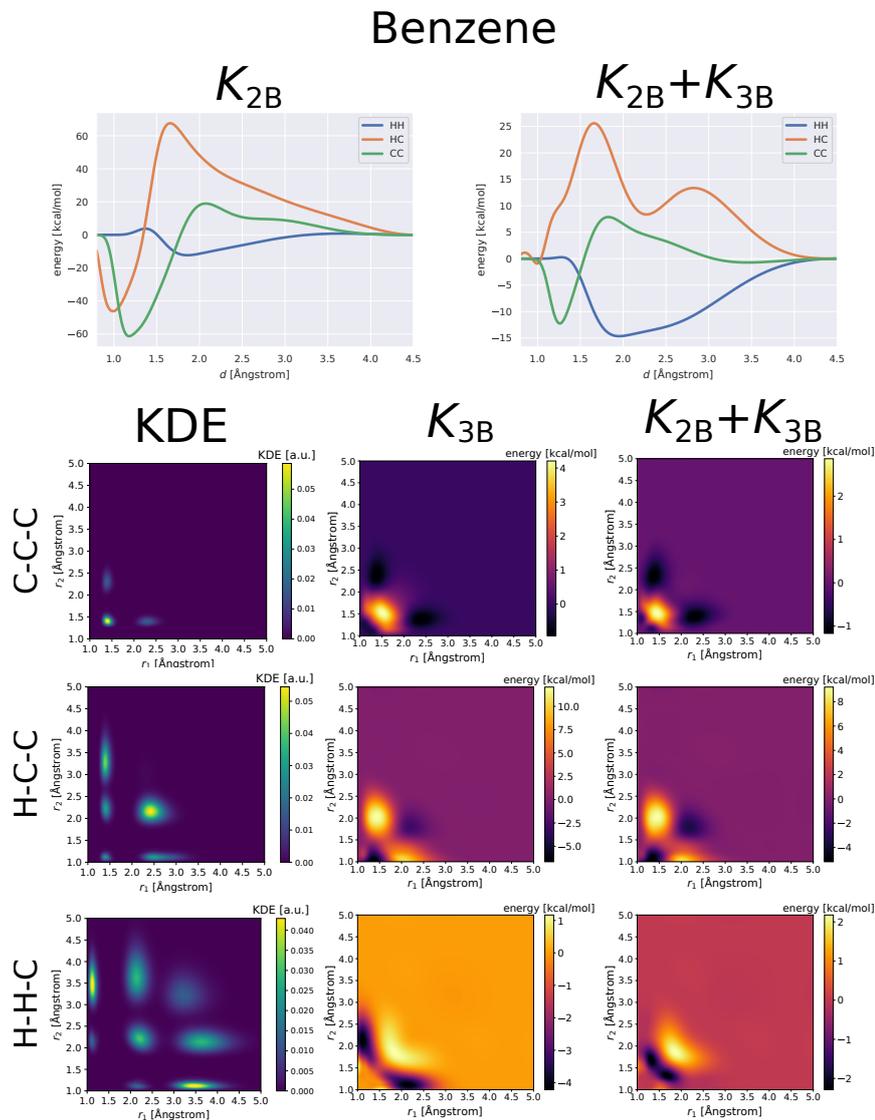


Figure 4.8: Interaction potentials of benzene according to Eq. (4.13). Two-body potentials for a model trained with the two-body kernel K_{2B} (top left) and the mixed kernel $K_{2B} + K_{3B}$ (top right), respectively. The three-body potentials of the C-C-C (top row), H-C-C (middle row) and H-H-C (bottom row) potentials are shown for a model trained with the three-body kernel K_{3B} (middle column) and the mixed kernel $K_{2B} + K_{3B}$ (right column). The left column shows the kernel density estimate according to Eq. (4.14). All models have been trained on 1000 random molecules and the kernel parameters have been determined by 10-fold nested cross-validation. The interaction potentials have been averaged by an ensemble of 30 models to increase reproducibility.

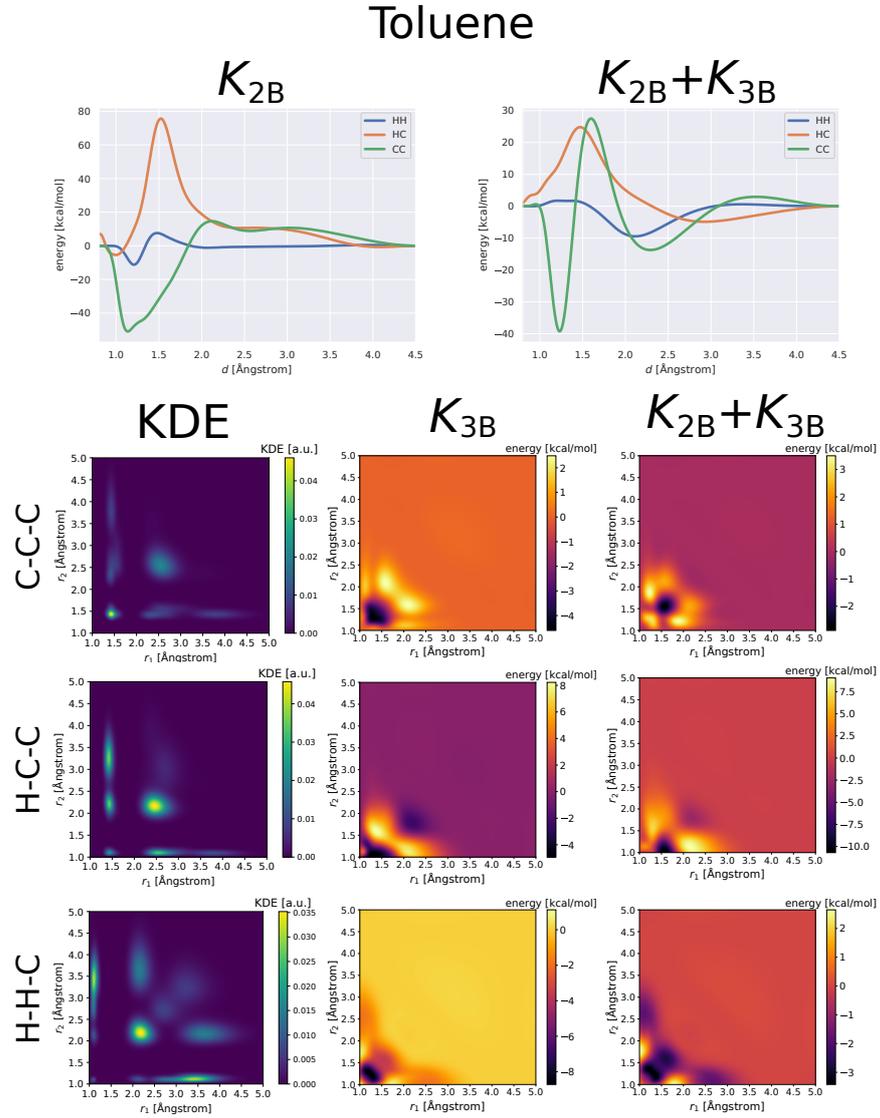


Figure 4.9: Interaction potentials of toluene according to Eq. (4.13). Two-body potentials for a model trained with the two-body kernel K_{2B} (top left) and the mixed kernel $K_{2B} + K_{3B}$ (top right), respectively. The three-body potentials of the C-C-C (top row), H-C-C (middle row) and H-H-C (bottom row) potentials are shown for a model trained with the three-body kernel K_{3B} (middle column) and the mixed kernel $K_{2B} + K_{3B}$ (right column). The left column shows the kernel density estimate according to Eq. (4.14). All models have been trained on 1000 random molecules and the kernel parameters have been determined by 10-fold nested cross-validation. The interaction potentials have been averaged by an ensemble of 30 models to increase reproducibility.

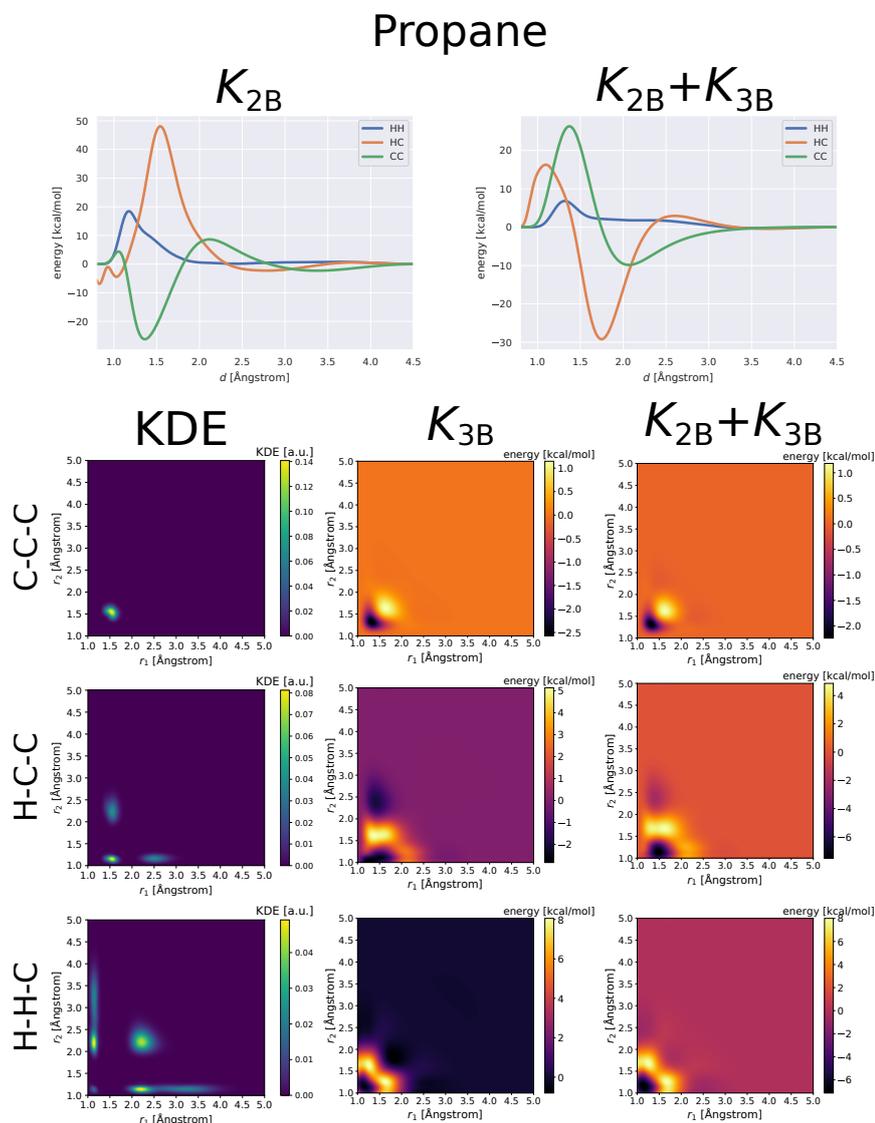


Figure 4.10: Interaction potentials of propane according to Eq. (4.13). Two-body potentials for a model trained with the two-body kernel K_{2B} (top left) and the mixed kernel $K_{2B} + K_{3B}$ (top right), respectively. The three-body potentials of the C-C-C (top row), H-C-C (middle row) and H-H-C (bottom row) potentials are shown for a model trained with the three-body kernel K_{3B} (middle column) and the mixed kernel $K_{2B} + K_{3B}$ (right column). The left column shows the kernel density estimate according to Eq. (4.14). All models have been trained on 1000 random molecules and the kernel parameters have been determined by 10-fold nested cross-validation. The interaction potentials have been averaged by an ensemble of 30 models to increase reproducibility.

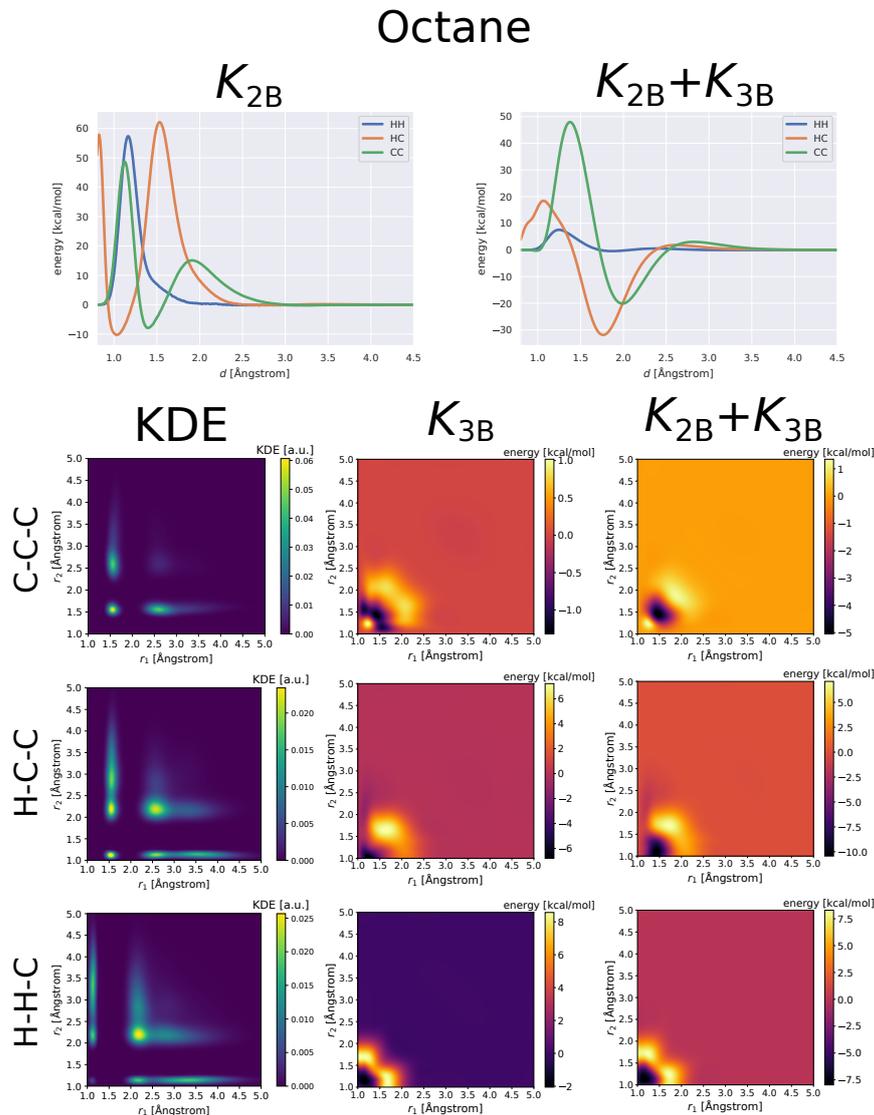


Figure 4.11: Interaction potentials of octane according to Eq. (4.13). Two-body potentials for a model trained with the two-body kernel K_{2B} (top left) and the mixed kernel $K_{2B} + K_{3B}$ (top right), respectively. The three-body potentials of the C-C-C (top row), H-C-C (middle row) and H-H-C (bottom row) potentials are shown for a model trained with the three-body kernel K_{3B} (middle column) and the mixed kernel $K_{2B} + K_{3B}$ (right column). The left column shows the kernel density estimate according to Eq. (4.14). All models have been trained on 1000 random molecules and the kernel parameters have been determined by 10-fold nested cross-validation. The interaction potentials have been averaged by an ensemble of 30 models to increase reproducibility.

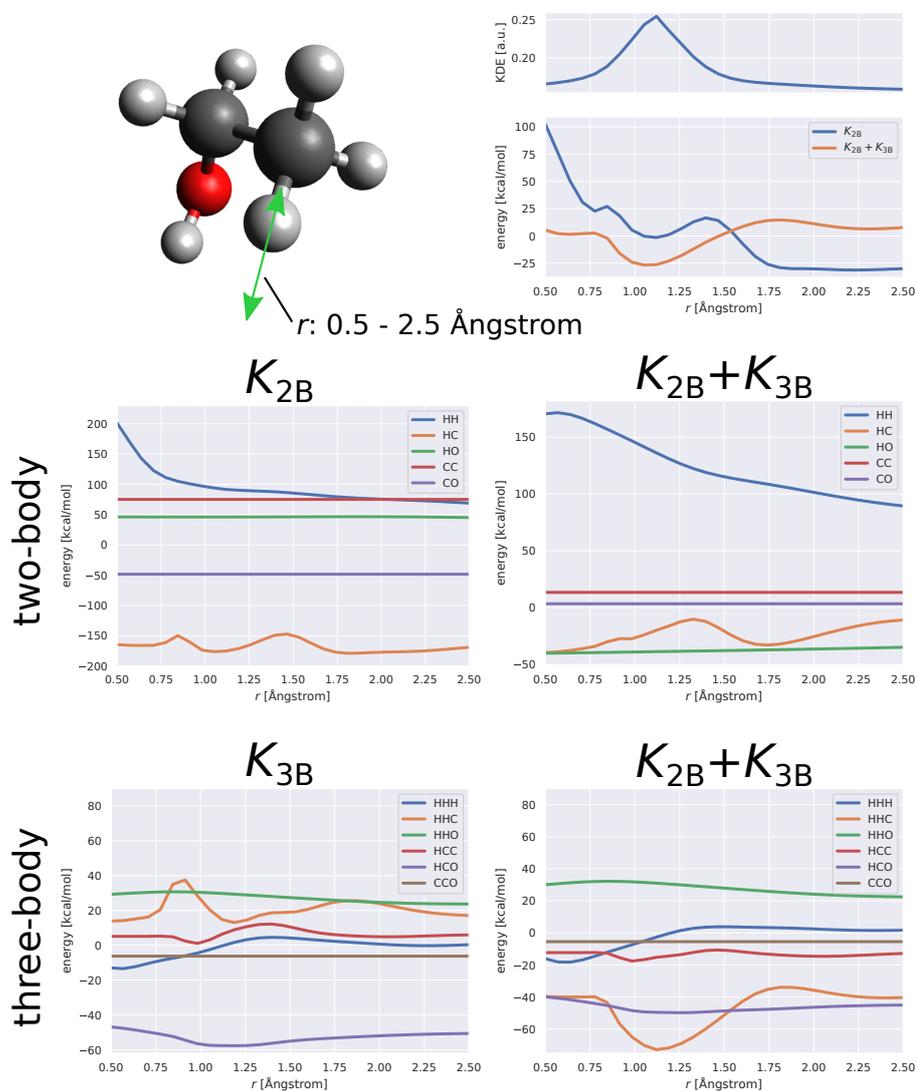


Figure 4.12: Interaction potentials for varying the C-H distance in ethanol as indicated by the green arrow (top left). For these configurations, the predicted total energies of the two- and three-body model is shown together with the kernel density estimate according to Eq. (4.14) (top right). The two-body (middle row) and three-body (bottom row) interaction potentials are shown for models trained with the two-body kernel K_{2B} (middle left), the three-body kernel K_{3B} (bottom left) and the mixed kernel $K_{2B} + K_{3B}$ (right column), respectively. All models have been trained on 1000 random molecules and the kernel parameters have been determined by 10-fold nested cross-validation. The interaction potentials have been averaged by an ensemble of 30 models to increase reproducibility.

4. KERNEL REPRESENTATIONS OF QUANTUM MECHANICAL SYSTEMS

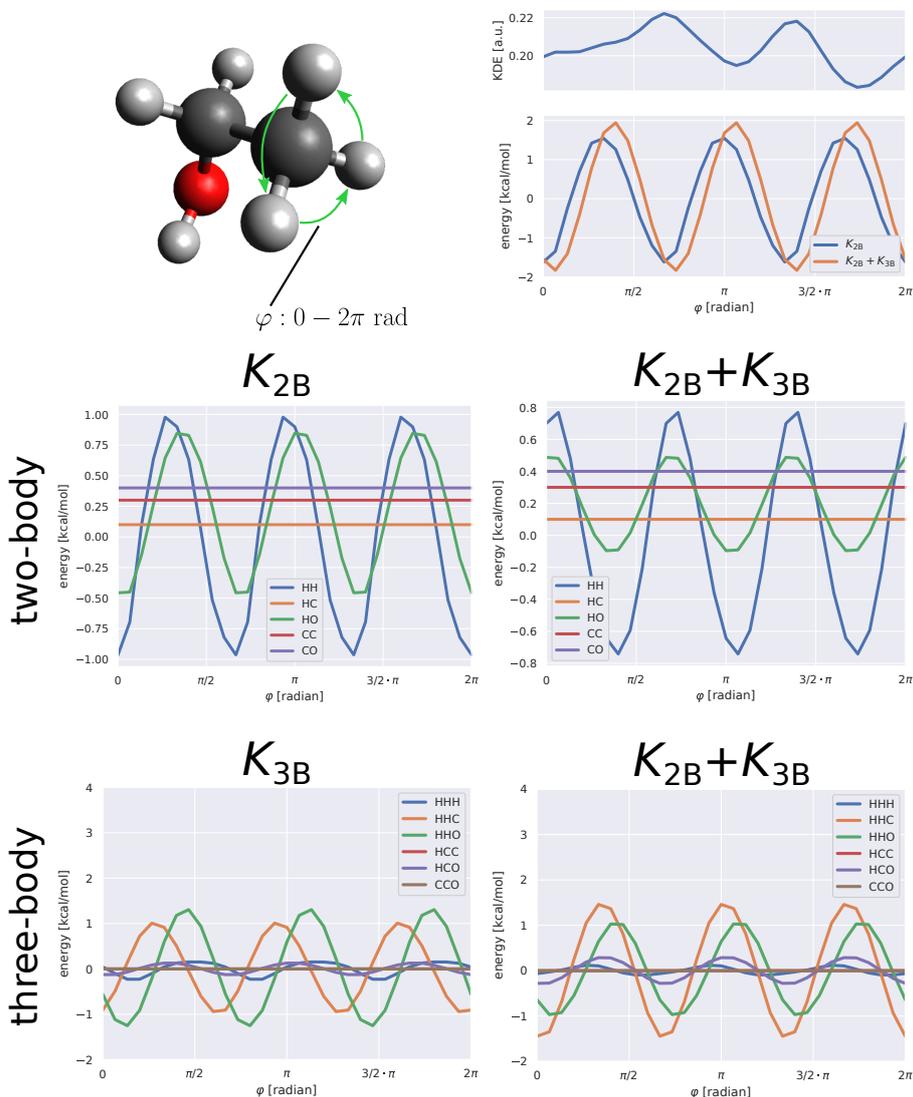


Figure 4.13: Interaction potentials for varying the dihedral angle of the CH_3 group in ethanol as indicated by the green arrow (top left). For these configurations, the predicted total energies of the two- and three-body model is shown together with the kernel density estimate according to Eq. (4.14) (top right). The two-body (middle row) and three-body (bottom row) interaction potentials are shown for models trained with the two-body kernel K_{2B} (middle left), the three-body kernel K_{3B} (bottom left) and the mixed kernel $K_{2B} + K_{3B}$ (right column), respectively. All models have been trained on 1000 random molecules and the kernel parameters have been determined by 10-fold nested cross-validation. The interaction potentials have been averaged by an ensemble of 30 models to increase reproducibility.

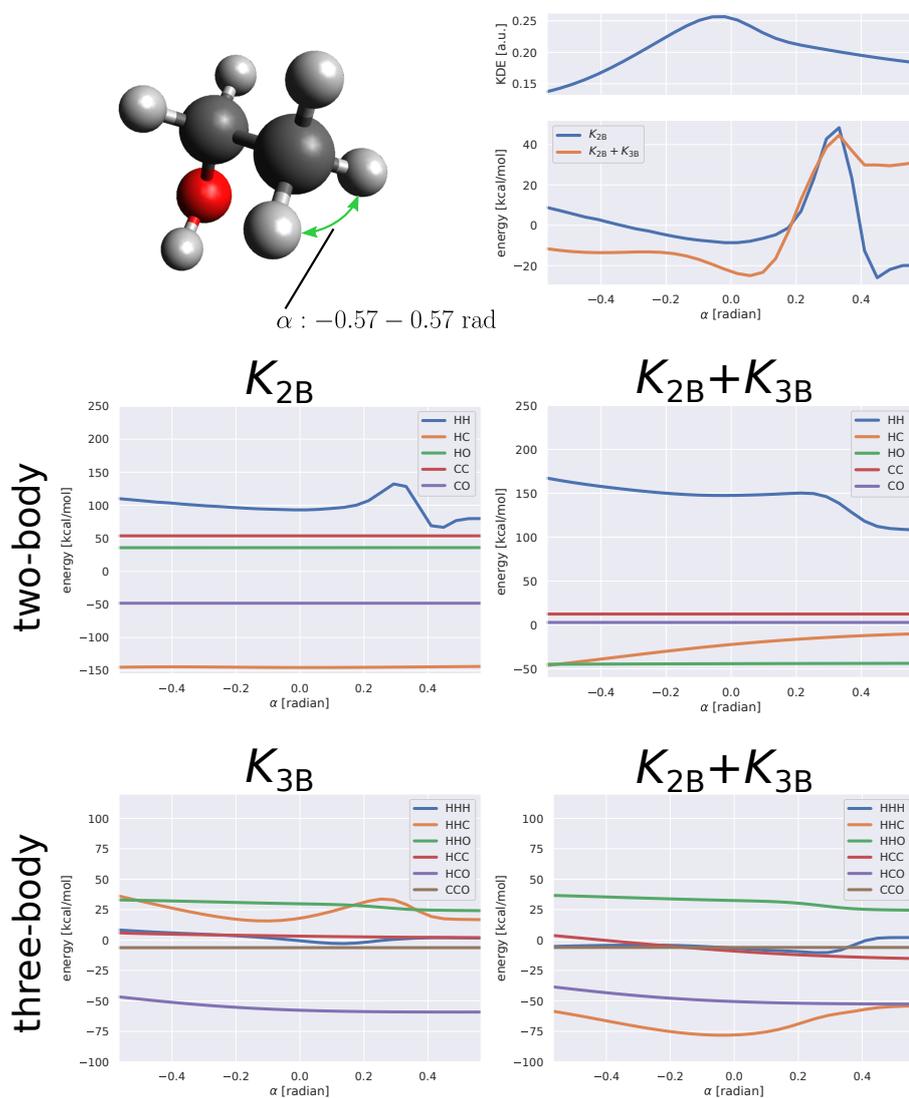


Figure 4.14: Interaction potentials for varying the H-C-H angle in ethanol as indicated by the green arrow (top left). For these configurations, the predicted total energies of the two- and three-body model is shown together with the kernel density estimate according to Eq. (4.14) (top right). The two-body (middle row) and three-body (bottom row) interaction potentials are shown for models trained with the two-body kernel K_{2B} (middle left), the three-body kernel K_{3B} (bottom left) and the mixed kernel $K_{2B} + K_{3B}$ (right column), respectively. All models have been trained on 1000 random molecules and the kernel parameters have been determined by 10-fold nested cross-validation. The interaction potentials have been averaged by an ensemble of 30 models to increase reproducibility.

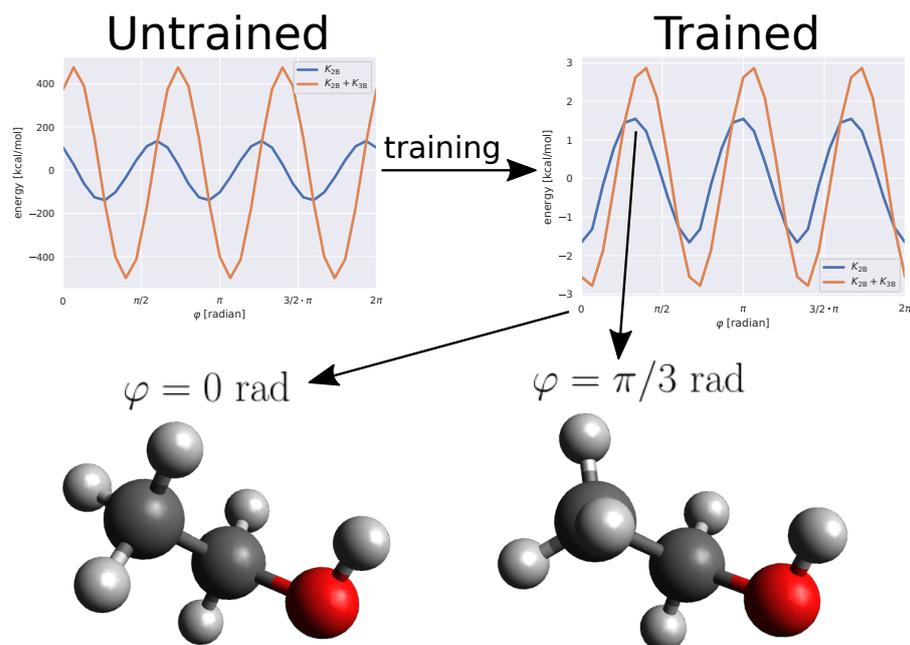


Figure 4.15: Atomization energies for varying the dihedral angle of the CH_3 group in ethanol predicted using the two-body kernel $K_{2\text{B}}$ (blue) and mixed two- and three-body kernel $K_{2\text{B}} + K_{3\text{B}}$ (orange) for the untrained (left) and trained model (right). For the untrained model, we set $\alpha_i = 1$ for $i = 1, \dots, N$ in Eq. (1.15). The trained model has been averaged by an ensemble of 30 models to increase reproducibility. In the bottom, the molecule for the dihedral angle $\varphi = 0$ rad (left) and $\varphi = \pi/3$ rad (right) is shown.

Approximate banded Toeplitz matrix inversion

Many real-world problems can be formulated in terms of a special structure called Toeplitz matrix. These matrices naturally arise from discretizing differential equations and convolution operators. Positive definite covariance matrices are typically represented by a Toeplitz matrix in stationary stochastic processes. A Toeplitz matrix is characterized by constant entries along the diagonals. For such matrices, more efficient algorithms as compared to the case of general matrices can be designed, for example to find the determinant [107, 108, 109], to solve matrix equations, i.e. Toeplitz systems [110, 111, 112, 113], to invert general Toeplitz matrices [114] and to compute matrix decompositions [115, 116, 117], respectively. A recent review on Toeplitz matrices is given by Gray [118]. So-called banded Toeplitz matrices allow even more efficient algorithms, the elements of these matrices are equal to zero if the absolute difference between row and column indices exceeds a certain positive number [119, 120, 121]. The problem of inverting banded Toeplitz matrices has been discussed by many authors [122, 123, 124, 125, 126]. Banded matrices which are not in Toeplitz form but instead have Toeplitz inverses have been analyzed in the literature [127, 128, 129].

Most of the above methods are exact in the sense that their solution depends on the matrix dimension. Intuitively, the matrix dimension should play a minor role for large matrices as the structure of these matrices is qualitatively similar. In this chapter, we propose to model the inverses of banded Toeplitz matrices by matrices which are themselves Toeplitz. Specifically, we develop methods to construct the unique Toeplitz matrices which approximate the inverses of a specific class of banded Toeplitz matrices as the matrix dimension tends to infinity. A characteristic of our method is that the computation time for a single element of the inverse is independent of the defining matrix dimension. Besides the proof of regularity of this class of banded Toeplitz matrices, we provide insights about the relationship between the matrices involved and their

corresponding inverses and show how they can be transformed into one another.

The chapter is structured as follows. We propose our main algorithms in the next Sec. 5.1. Our main theorems are provided in Sec. 5.2. This is followed by a more detailed list of proofs in Sec. 5.3. We apply our method to the special case of tridiagonal matrices, where we obtain analytical solutions in Sec. 5.4. In Sec. 5.5, we evaluate the time complexity by comparing to state-of-the-art methods for inverting banded Toeplitz matrices. Our approach can be applied to efficiently construct the Green's functions of one-dimensional linear differential operators with constant coefficients, which is demonstrated in Sec. 5.6. In another application, we show how to obtain a banded approximation of deconvolution operators, where we use one of our methods in *reverse* fashion in Sec. 5.7. We apply our method to compute van der Waals interactions including long-range electrodynamic response screening more efficiently in Sec. 5.8. This is followed by another application for quantum chemistry in Sec. 5.9, where we interpolate potential energy surfaces including the CH₃ ethanol rotor already encountered in the previous chapter. Finally, we conclude the chapter in Sec. 5.10.

5.1 Method

Let M_n be a $n \times n$ banded Toeplitz matrix

$$M_n := \begin{bmatrix} m_0 & m_1 & \cdots & m_s & 0 & \cdots & 0 \\ & \ddots & & & & & \vdots \\ m_{-1} & & & & & & 0 \\ \vdots & & & & & & \vdots \\ m_{-r} & & & & & & m_s \\ 0 & & & & & & \vdots \\ \vdots & & & & & & \vdots \\ 0 & \cdots & 0 & m_{-r} & \cdots & m_{-1} & m_0 \end{bmatrix}$$

defined by the band $m_i \in \mathbb{C}$ for $i = -r, \dots, 0, \dots, s$ with $m_{-r}, m_s \neq 0$ and $r, s \in \mathbb{N}^+$. The inverse M_n^{-1} of M_n cannot be Toeplitz (see Proofs section). However, for a specific class of matrices of form M_n which we will define later, the inverse M_n^{-1} exists and can be approximated by a Toeplitz matrix

$$M_n^{-1} \approx \bar{M}_n^{-1} := \begin{bmatrix} \phi_0 & \phi_1 & \cdots & \phi_{n-2} & \phi_{n-1} \\ \phi_{-1} & \phi_0 & \cdots & \cdots & \phi_{n-2} \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \phi_{-n+2} & \cdots & \ddots & \ddots & \phi_1 \\ \phi_{-n+1} & \phi_{-n+2} & \cdots & \phi_{-1} & \phi_0 \end{bmatrix}$$

as the matrix dimension tends to infinity in the following sense

$$\lim_{n \rightarrow \infty} (M_n^{-1} - \bar{M}_n^{-1})_{[a \cdot n]+i, [a \cdot n]+j} = 0 \quad (5.1)$$

for any given $i, j \in \mathbb{N}^+$ and $a \in]0, 1[$. In the next section, we propose an efficient algorithm to construct the Toeplitz matrix \bar{M}_n^{-1} from M_n . In the reverse application of this algorithm, we show how to estimate the parameters r and s and to construct the band $\{m_i\}_{i=-r}^s$ starting from the Toeplitz matrix \bar{M}_n^{-1} .

Constructing approximate inverses from banded Toeplitz matrices

The band $\{m_i\}_{i=-r}^s$ defines the polynomial

$$P(x) := \sum_{i=0}^{r+s} m_{i-r} \cdot x^i = m_s \cdot \prod_{i=1}^{r+s} (x - z_i) \quad (5.2)$$

where the $\{z_i\}_{i=1}^{r+s} \in \mathbb{C}$ denote the complex roots of $P(\cdot)$ and we assume

$$|z_1| \leq |z_2| \leq \dots \leq |z_{r+s}| \quad (5.3)$$

without loss of generality. From these roots we define the two polynomials

$$A(x) := \sum_{i=0}^r a_{r-i} \cdot x^i = \prod_{i=1}^r (x - z_i) \quad (5.4)$$

$$B(x) := \sum_{i=0}^s b_i \cdot x^i = m_s \cdot \prod_{i=1}^s (x - z_{r+i}) \quad (5.5)$$

with the coefficients $\{a_i\}_{i=0}^r$ and $\{b_i\}_{i=0}^s$ and $A(x) \cdot B(x) = P(x)$. The Toeplitz matrix $(\bar{M}_n^{-1})_{i,j} =: \phi_{i-j}$ is constructed by the set of two recurrence relations

$$\phi_k = -\frac{1}{a_0} \cdot \sum_{i=1}^r a_i \phi_{k-i} \quad s \leq k \leq n-1 \quad (5.6)$$

$$\phi_{-k} = -\frac{1}{b_0} \cdot \sum_{i=1}^s b_i \phi_{-k+i} \quad r \leq k \leq n-1 \quad (5.7)$$

with the initial values $(\phi_{-r+1}, \dots, \phi_0, \dots, \phi_{s-1})^\top$ given by solving the system of linear equations

$$\sum_{i=0}^r a_i \phi_{k-i} = 0 \quad 1 \leq k \leq s-1 \quad (5.8)$$

$$\sum_{i=0}^s b_i \phi_{-k+i} = 0 \quad 1 \leq k \leq r-1 \quad (5.9)$$

$$\sum_{i=0}^{r-1} b_0 a_i \phi_{-i} - \sum_{i=1}^s a_r b_i \phi_{-r+i} = 1 \quad (5.10)$$

Algorithm 5 Band2Toeplitz

Input:

$$\vec{m} = (m_{-r}, \dots, m_s)$$

Output:

$$\vec{\varphi} = (\phi_{-r+1}, \dots, \phi_{s-1})$$

- 1: $z_1, z_2, \dots, z_{r+s} \leftarrow$ ordered roots of the polynomial $\sum_{i=0}^{r+s} m_{-r+i} \cdot x^i$
 - 2: $a_0, a_1, \dots, a_r \leftarrow$ such that $\sum_{i=0}^r a_{r-i} \cdot x^i = \prod_{i=1}^r (x - z_i)$
 - 3: $b_0, b_1, \dots, b_s \leftarrow$ such that $\sum_{i=0}^s b_i \cdot x^i = m_s \cdot \prod_{i=r+1}^{r+s} (x - z_i)$
 - 4: $y_0 \leftarrow 1$
 - 5: **for** $i = 1$ to $r + s - 2$ **do**
 - 6: $y_i \leftarrow 0$
 - 7: **for** $i = 0$ to $r + s - 2$ **do**
 - 8: $M_{0,i} \leftarrow 0$ \triangleright The dimension of M is $(r + s - 1, r + s - 1)$
 - 9: **for** $i = 0$ to $r - 1$ **do**
 - 10: $M_{0,r-1-i} \leftarrow M_{0,r-1-i} + b_0 \cdot a_i$
 - 11: **for** $i = 1$ to s **do**
 - 12: $M_{0,i-1} \leftarrow M_{0,i-1} - a_s \cdot b_i$
 - 13: **for** $i = 1$ to $s - 1$ **do**
 - 14: **for** $j = 0$ to r **do**
 - 15: $M_{i,r-1+i-j} \leftarrow a_j$
 - 16: **for** $i = 1$ to $r - 1$ **do**
 - 17: **for** $j = 0$ to s **do**
 - 18: $M_{s-1+i,i-r+1+j} \leftarrow b_j$
 - 19: $\vec{\varphi} \leftarrow$ solve $M \cdot \vec{\varphi} = \vec{y}$
 - 20: **return** $\vec{\varphi}$
-

The pseudocode for obtaining the initial values $(\phi_{-r+1}, \dots, \phi_{s-1})^\top$ from the band $\{m_i\}_{i=-r}^s$ is given in Alg. 5. The total sequence $\phi = (\phi_{n-1}, \dots, \phi_{-n+1})^\top$ which defines the inverse \bar{M}_n^{-1} can be calculated by applying the recurrence relations (5.6) - (5.7).

Constructing approximate banded inverses from Toeplitz matrices

The parameters r and s can be estimated independently by using the functions

$$f(\bar{r}) := \min_{\mathbf{a} \in \mathbb{R}^{\bar{r}}} \sum_{k=\bar{r}}^{n-1} \left(\sum_{i=1}^{\bar{r}} a_i \cdot \phi_{k-i} - \phi_k \right)^2 \quad (5.11)$$

$$g(\bar{s}) := \min_{\mathbf{b} \in \mathbb{R}^{\bar{s}}} \sum_{k=\bar{s}}^{n-1} \left(\sum_{i=1}^{\bar{s}} b_i \cdot \phi_{-k+i} - \phi_{-k} \right)^2 \quad (5.12)$$

which for given \bar{r} and \bar{s} are minimized by a least squares solution

$$\mathbf{a} := (X_a^\top X_a)^{-1} X_a^\top \boldsymbol{\varphi}_{\bar{r}, n-\bar{r}} \quad (5.13)$$

$$\mathbf{b} := (X_b^\top X_b)^{-1} X_b^\top \boldsymbol{\varphi}_{\bar{s}, n-\bar{s}} \quad (5.14)$$

with

$$\boldsymbol{\varphi}_{k,l} := (\phi_k, \phi_{k+1}, \dots, \phi_{k+l})^\top \quad (5.15)$$

and

$$X_a := (\boldsymbol{\varphi}_{0,-\bar{r}}, \boldsymbol{\varphi}_{1,-\bar{r}}, \dots, \boldsymbol{\varphi}_{n-\bar{r},-\bar{r}})^\top \quad (5.16)$$

$$X_b := (\boldsymbol{\varphi}_{0,\bar{s}}, \boldsymbol{\varphi}_{1,\bar{s}}, \dots, \boldsymbol{\varphi}_{n-\bar{s},\bar{s}})^\top \quad (5.17)$$

The parameters r and s can be determined by

$$s := \arg \min_{\bar{s}} f(\bar{s}) \quad (5.18)$$

$$r := \arg \min_{\bar{r}} g(\bar{r}) \quad (5.19)$$

$$(5.20)$$

In practice, we suggest to select the values for r and s at which the functions $f(\cdot)$ and $g(\cdot)$ exceed a given small value, for example a numerical precision threshold of 10^{-14} . Using these parameters r and s , solve the system of linear equations given the sequence $(\phi_{-r}, \dots, \phi_s)^\top$ in terms of the variables $\{\bar{a}_i\}_{i=0}^r$ and $\{\bar{b}_i\}_{i=0}^s$

$$\sum_{p=0}^r \bar{a}_p \phi_{k-p} = 0 \quad 1 \leq k \leq s \quad (5.21)$$

$$\sum_{p=0}^s \bar{b}_p \phi_{-k+p} = 0 \quad 1 \leq k \leq r \quad (5.22)$$

$$\sum_{p=0}^{r-1} \bar{a}_p \phi_{-p} - \sum_{p=1}^s \bar{b}_p \phi_{r+p} = 1 \quad (5.23)$$

$$\bar{a}_r = \bar{b}_0 \quad (5.24)$$

Using the variable transformation $a_i = \bar{a}_i/\bar{a}_0$ for $i = 0, \dots, r$ and $b_i = \bar{b}_i \cdot \bar{a}_0/\bar{b}_0$ for $i = 0, 1, \dots, s$ the band is now obtained by

$$m_i = \begin{cases} \sum_{k=0}^{\min(r+i,s)} a_{k-i} \cdot b_k & -r \leq i \leq 0 \\ \sum_{k=0}^{\min(r,s-i)} a_k \cdot b_{k+i} & 0 \leq i \leq s \end{cases} \quad (5.25)$$

The pseudocode of obtaining the band $\{m_i\}_{i=-r}^s$ from the sequence $(\phi_{-r}, \dots, \phi_s)^\top$ is given in Alg. 6.

5.2 Main theorems

The roots of the polynomial $P(\cdot)$ play a crucial role in understanding the properties of the inverse M_n^{-1} . Based on these roots, we define a sufficient condition for the applicability of our method for constructing approximate Toeplitz inverses from banded Toeplitz matrices.

Theorem 1. *If*

$$|z_1| \leq |z_2| \leq \dots \leq |z_r| < 1 < |z_{r+1}| \leq \dots \leq |z_{r+s}| \quad (5.26)$$

then there exists a unique Toeplitz matrix \bar{M}_n^{-1} that approximates M_n^{-1} in the sense given by (5.1).

The construction of the matrix \bar{M}_n^{-1} has been shown in the previous section. At first glance, the condition (5.26) looks very specific. In practice however, the inverses of banded Toeplitz matrices often decay to zero, as one moves away from the main diagonal. In fact, the practical usefulness of our method is provided by the following

Theorem 2. *If the inverse M_n^{-1} of a banded Toeplitz matrix of form M_n decays to zero as one moves away from the main diagonal in the limit*

$$\lim_{|k| \rightarrow \infty} \lim_{n \rightarrow \infty} (M_n^{-1})_{[n/2], [n/2]+k} = 0 \quad (5.27)$$

then the condition (5.26) holds for the band of M_n .

The relation of the four matrices $M_n, M_n^{-1}, \bar{M}_n, \bar{M}_n^{-1}$ is schematically depicted in Fig. 5.1.

Solving linear matrix equations

In practice, one often solves matrix equations of the form $M_n \cdot \mathbf{x} = \mathbf{y}$ with given M_n and \mathbf{y} . Using our approximate Toeplitz inverse \bar{M}_n^{-1} , we propose the approximate solution

$$\bar{\mathbf{x}} := \bar{M}_n^{-1} \cdot \mathbf{y} = \phi * \mathbf{y} \quad (5.28)$$

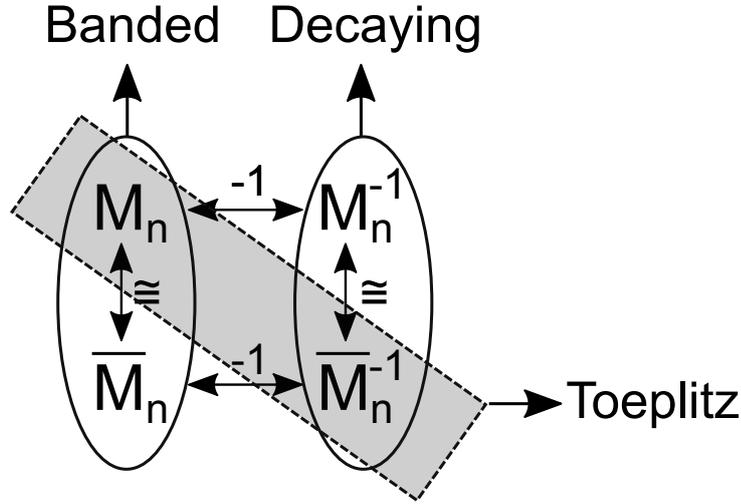


Figure 5.1: Relationship of the four matrices $M_n, M_n^{-1}, \bar{M}_n, \bar{M}_n^{-1}$. The matrices M_n and \bar{M}_n are banded. Their corresponding inverses M_n^{-1} and \bar{M}_n^{-1} decay crosswise to the main diagonal in the sense given by Eq. (5.27). The matrices in the bottom row \bar{M}_n and \bar{M}_n^{-1} are approximations of the respective matrices M_n and M_n^{-1} in the top row and *vice versa*. The matrices on the diagonal of the figure M_n and \bar{M}_n^{-1} are Toeplitz and are the main focus of this chapter.

where $*$ denotes the convolution operator. More specifically, our approximate solution is given by

$$(\bar{\mathbf{x}})_i = \sum_{k=0}^{n-1} \phi_{k-i} \cdot (\mathbf{y})_k \quad (5.29)$$

Computational complexity analysis

The computational complexity of our method is dominated by the parameters r and s . The roots $\{z_i\}_{i=1}^{r+s}$ of the polynomial $P(\cdot)$ in Eq. (5.2) can be determined in $\mathcal{O}((r+s)^3)$, for example by using the companion matrix [130]. Similarly, computing the coefficients $\{a_i\}_{i=0}^r$ and $\{b_i\}_{i=0}^s$ and the initial values $(\phi_{-r+1}, \dots, \phi_{s-1})^\top$ in Alg. 5 takes $\mathcal{O}((r+s)^3)$. The construction of the total sequence $\boldsymbol{\phi} = \{\phi_k\}_{k=-n+1}^{n-1}$ can be done in $\mathcal{O}((r+s)^3 + (r+s) \cdot n)$. We emphasize that the complexity of our method is independent of the matrix order n in the sense that a single element ϕ_k can be calculated in $\mathcal{O}((r+s)^3)$ which can be seen by Eqs. (5.49) and (5.50), respectively. Computing the approximate solution of the linear matrix equation $\bar{\mathbf{x}}$ can be done in $\mathcal{O}((r+s)^3 + n^2)$. As the elements of $\boldsymbol{\phi}$ decay to zero in the limit $\lim_{|k| \rightarrow \infty} \phi_k = 0$ if the condition (5.26) holds, one can perform the convolution step in (5.29) more efficiently in prac-

tice. Estimating the parameters r and s can be done in $\mathcal{O}(n \cdot (r^3 + s^3))$. To obtain the band from a Toeplitz matrix takes $\mathcal{O}((r + s)^3)$ for solving the linear matrix equations (5.21) - (5.24) and $\mathcal{O}((r + s)^2)$ to compute the band in Eq. (5.25), respectively.

5.3 Proofs

First, we summarize the results from a previous work by the following

Theorem 3 (Greville et al. [127]). *The matrix \bar{M}_n is banded and the inverse of a Toeplitz matrix if and only if the matrix-vector product $\bar{M}_n \cdot \mathbf{x}$ with $\mathbf{x} := (x^0, x^1, \dots, x^{n-1})^\top$ can be represented by*

$$(\bar{M}_n \cdot \mathbf{x})_i := \begin{cases} A(x) \sum_{j=0}^i b_{i-j} x^j & 0 \leq i \leq s-1 \\ x^{i-s} A(x) B(x) & s \leq i \leq n-r-1 \\ x^{i-s} B(x) \sum_{j=0}^{n-1-i} a_j x^j & n-r \leq i \leq n-1 \end{cases} \quad (5.30)$$

with $a_r b_0 \neq 0$ and the polynomials $A(x) = \sum_{k=0}^r a_{r-k} x^k$ and $B(x) = \sum_{k=0}^s b_k x^k$ are relatively prime.

We rely on their approach to efficiently construct the matrix \bar{M}_n^{-1} from the coefficients $\{a_i\}_{i=0}^r$ and $\{b_i\}_{i=0}^s$ in Sec. 5.1. From now on, we move to our results for analyzing the properties of banded Toeplitz matrices. If we choose the coefficients defined by Eqs. (5.4) and (5.5) for constructing the matrix \bar{M}_n , the resulting matrix equals the matrix M_n with exception of the upper left $s \times s$ and bottom right $r \times r$ corner block matrices. We define C and \bar{C} as the upper left $s \times s$ block matrix of M_n and \bar{M}_n , respectively. Let $D := C - \bar{C}$ and D_p be the difference of the corresponding bottom-right $r \times r$ block matrices of M_n and \bar{M}_n , respectively. Let $U^\top := (I_{s \times s}, 0_{s \times (n-s)})$ and $V^\top := (0_{r \times (n-r)}, I_{r \times r})$ where $I_{s \times s}$ and $I_{r \times r}$ is the $s \times s$ and $r \times r$ identity matrix, $0_{s \times (n-s)}$ is the $s \times (n-s)$ zero matrix and $0_{r \times (n-r)}$ is the $r \times (n-r)$ zero matrix, respectively. Using the Woodbury matrix identity [131] the relation between M_n^{-1} and \bar{M}_n^{-1} can be written as

$$M_n^{-1} = (\bar{M}_n + UDU^\top + VD_pV^\top)^{-1} \quad (5.31)$$

$$\begin{aligned} &= (\bar{M}_n + UDU^\top)^{-1} - (\bar{M}_n + UDU^\top)^{-1}V \\ &\quad \cdot (D_p^{-1} + V^\top(\bar{M}_n + UDU^\top)^{-1}V)^{-1}V^\top(\bar{M}_n + UDU^\top)^{-1} \end{aligned} \quad (5.32)$$

$$\begin{aligned} &= \bar{M}_n^{-1} - X_D - (\bar{M}_n^{-1} - X_D)V \\ &\quad \cdot (D_p^{-1} + V^\top(\bar{M}_n^{-1} - X_D)V)^{-1}V^\top(\bar{M}_n^{-1} - X_D) \end{aligned} \quad (5.33)$$

$$= \bar{M}_n^{-1} - X_D - \bar{M}_n^{-1}V(D_p^{-1} + V^\top\bar{M}_n^{-1}V)^{-1}V^\top\bar{M}_n^{-1} \quad (5.34)$$

$$= \bar{M}_n^{-1} - X_D - X_{D_p} \quad (5.35)$$

with

$$X_D := \bar{M}_n^{-1}U(D^{-1} + U^\top \bar{M}_n^{-1}U)^{-1}U^\top \bar{M}_n^{-1} \quad (5.36)$$

$$X_{D_p} := \bar{M}_n^{-1}V(D_p^{-1} + V^\top \bar{M}_n^{-1}V)^{-1}V^\top \bar{M}_n^{-1} \quad (5.37)$$

and we utilized $U^\top \bar{M}_n^{-1}V = V^\top \bar{M}_n^{-1}U = 0$ in the above steps. The matrix M_n^{-1} exists if the matrices D , $(D^{-1} + U^\top \bar{M}_n^{-1}U)$, D_p and $(D_p^{-1} + V^\top \bar{M}_n^{-1}V)$ are invertible. The regularity of the matrices D and $(D^{-1} + U^\top \bar{M}_n^{-1}U)$ is shown in Lemmas 1 and 2, respectively. The proofs for the regularity of the matrices D_p and $(D_p^{-1} + V^\top \bar{M}_n^{-1}V)$ are analogous and therefore omitted. Note that while the matrices M_n^{-1} and \bar{M}_n^{-1} depend on n , the elements of the two matrices D and $(D^{-1} + U^\top \bar{M}_n^{-1}U)$ are independent of n . We will utilize this fact in Lemma 3.

Lemma 1. *The matrix $D = U^\top (M_n - \bar{M}_n)U$ is regular.*

Proof. As $A(x) \cdot B(x) = P(x)$, we can express the element of M_n at the i -th row and j -th column in terms of the coefficients $\{a_i\}_{i=0}^r$ and $\{b_i\}_{i=0}^s$ by

$$(M_n)_{i,j} = m_{j-i} = \begin{cases} \sum_{k=0}^{\min(r+j-i,s)} a_{k+i-j} \cdot b_k & j \leq i \\ \sum_{k=0}^{\min(r,s-j+i)} a_k \cdot b_{k+j-i} & j > i \end{cases} \quad (5.38)$$

where we set $\sum_{i=i_1}^{i_2} = 0$ for $i_1 > i_2$. Similarly, we can express the element of \bar{M}_n at the i -th row and j -th column in terms of these coefficients by

$$(\bar{M}_n)_{i,j} = \begin{cases} \sum_{k=0}^j a_{k+i-j} \cdot b_k & j \leq i \\ \sum_{k=0}^i a_k \cdot b_{k+j-i} & j > i \end{cases} \quad (5.39)$$

For simplicity of the proof, we assume $r = s$. Then, we can express D in terms of the coefficients $\{a_i, b_i\}_{i=0}^s$ from Eqs. (5.38) and (5.39) by

$$(D)_{i,j} = \begin{cases} \sum_{k=1}^{s-i} a_{k+i} \cdot b_{k+j} & j \leq i \\ \sum_{k=1}^{s-j} a_{k+i} \cdot b_{k+j} & j > i \end{cases} \quad (5.40)$$

The matrix D can we written as

$$D = \sum_{m=1}^s T_m \quad (5.41)$$

where

$$(T_m)_{i,j} := \begin{cases} a_{m+i} \cdot b_{m+j} & i, j \leq s - m \\ 0 & \text{otherwise} \end{cases} \quad (5.42)$$

From the linear dependence of the rows of D follows $T_s = 0 \Rightarrow a_s \cdot b_s = 0$. This is in contradiction to our assumption $m_{-r}, m_s \neq 0$, as $m_{-r} = a_r \cdot b_0$ and $m_s = a_0 \cdot b_s$. This completes the proof of Lemma 1. \square

Lemma 2. *There exists a $\bar{n} \geq r + s$ such that the matrix $(D^{-1} + U^\top \bar{M}_{\bar{n}}^{-1} U)$ is regular.*

Proof. In the framework of infinite Toeplitz matrices [132], it can be shown that

$$\lim_{n \rightarrow \infty} \det M_n = \prod_{i=1}^r \prod_{j=1}^s \left(1 - \frac{|z_i|}{|z_{s+j}|}\right)^{-1} \neq 0 \quad (5.43)$$

This means there exists a $\bar{n} \geq 2s$ for which $M_{\bar{n}}$ is regular. For this \bar{n} we have

$$X_D = \bar{M}_{\bar{n}}^{-1} - M_{\bar{n}}^{-1} - X_{D_p} \quad (5.44)$$

$$U(D^{-1} + U^\top \bar{M}_{\bar{n}}^{-1} U)^{-1} U^\top = \bar{M}_{\bar{n}}(\bar{M}_{\bar{n}}^{-1} - M_{\bar{n}}^{-1})\bar{M}_{\bar{n}} - V(D_p^{-1} + V^\top \bar{M}_{\bar{n}}^{-1} V)^{-1} V^\top \quad (5.45)$$

$$(D^{-1} + U^\top \bar{M}_{\bar{n}}^{-1} U)^{-1} = U^\top \bar{M}_{\bar{n}} U - U^\top \bar{M}_{\bar{n}} M_{\bar{n}}^{-1} \bar{M}_{\bar{n}} U \quad (5.46)$$

where we utilized $U^\top U = I$ and $U^\top V = V^\top U = 0$. This completes the proof of Lemma 2. \square

Now we are ready to formalize the existence of the matrix M_n^{-1} is the following

Lemma 3. *If the condition (5.26) holds for M_n , then the matrix M_n is regular for all $n \geq r + s$.*

Proof. Both the matrices D and $(D^{-1} + U^\top \bar{M}_n^{-1} U)^{-1}$ are independent of n for $n \geq r + s$ which can be seen from the construction of the matrix \bar{M}_n . Their existence has been shown for a given $\bar{n} \geq r + s$ in Lemma 1 and 2. This completes the proof of Lemma 3. \square

After we have shown that our main matrix of interest M_n^{-1} exists, we move to the main

Proof of Theorem 1. The characteristic polynomials of the recurrence relations (5.6) - (5.7) are

$$A_c(x) = x^r + 1/a_0 \cdot \sum_{p=1}^r a_p x^{r-p} \quad (5.47)$$

$$B_c(x) = x^s + 1/b_0 \cdot \sum_{p=1}^s b_p x^{s-p} \quad (5.48)$$

which have the roots $\{z_i\}_{i=1}^r$ and $\{1/z_i\}_{i=r+1}^{r+s}$, respectively. The elements of \bar{M}_n^{-1} can be expressed as a linear combination of these roots by

$$\begin{aligned} \phi_k = & (c_0 z_1^k + c_1 k z_1^k + \dots + c_{w_1} k^{w_1-1} z_1^k) + \dots \\ & + (c_{r-w_r+1} z_a^k + c_{r-1} k^{w_r-1} z_a^k) \end{aligned} \quad (5.49)$$

$$\begin{aligned} \phi_{-k} = & (c_0^* z_{r+1}^{-k} + c_1^* k z_{r+1}^{-k} + \dots + c_{w_{r+1}}^* k^{w_{r+1}-1} z_{r+1}^{-k}) + \dots \\ & + (c_{s-w_{r+s}+1}^* z_{r+b}^{-k} + c_{s-1}^* k^{w_{r+s}-1} z_{r+b}^{-k}) \end{aligned} \quad (5.50)$$

for $k = 0, 1, \dots, n-1$ and where a and b denote the number of distinct roots of $A(\cdot)$ and $B(\cdot)$ and w_i stands for the multiplicity of the root z_i , respectively. The coefficients $\{c_i\}_{i=0}^{r-1}$ and $\{c_i^*\}_{i=0}^{s-1}$ can be determined from the initial values $(\phi_{-r+1}, \dots, \phi_{s-1})^\top$ with the additional requirement $\phi_0 = \phi_{-0}$ in Eqs. (5.49) and (5.50), respectively. From Eqs. (5.49) and (5.50) follows $\phi_k = \mathcal{O}(e^{-\alpha|k|})$ for some positive α which implies $\lim_{|k| \rightarrow \infty} \phi_k = 0$, where we utilized Eq. (5.26).

With Eq. (5.36), this translates into

$$\exists \alpha > 0 : \forall n \geq r+s, \forall i, j = 0, \dots, n-1 : (X_D)_{i,j} = \mathcal{O}(e^{-\alpha \cdot (i+j)}) \quad (5.51)$$

and similarly

$$\exists \beta > 0 : \forall n \geq r+s, \forall i, j = 0, \dots, n-1 : (X_{D_p})_{i,j} = \mathcal{O}(e^{-\beta \cdot (n-i+n-j)}) \quad (5.52)$$

Now, for any given $i, j \in \mathbb{N}^+$ and $a \in]0, 1[$ it holds

$$\begin{aligned} \lim_{n \rightarrow \infty} (M_n^{-1} - \bar{M}_n^{-1})_{[a \cdot n] + i, [a \cdot n] + j} &= \lim_{n \rightarrow \infty} (\mathcal{O}(e^{-\alpha \cdot (a \cdot n + i + a \cdot n + j)}) \\ &+ \mathcal{O}(e^{-\beta \cdot (n - a \cdot n - i + n - a \cdot n - j)})) \end{aligned} \quad (5.53)$$

$$\begin{aligned} &= \lim_{n \rightarrow \infty} (\mathcal{O}(e^{-\alpha \cdot (a \cdot n + i + a \cdot n + j)}) \\ &+ \mathcal{O}(e^{-\beta \cdot ((1-a) \cdot n - i + (1-a) \cdot n - j)})) \end{aligned} \quad (5.54)$$

$$= 0 \quad (5.55)$$

This completes the proof of Theorem 1. \square

The approximation behaviour for increasing matrix dimension is schematically depicted in Fig. 5.2. As a useful side-product, we show the following

Corollary 1. *The inverse of the banded Toeplitz matrix M_n cannot be Toeplitz for $n \geq r+s$.*

Proof. As the matrix M_n has Toeplitz inverse, it is of form \bar{M}_n with the coefficients $\{a_i\}_{i=0}^r$ and $\{b_i\}_{i=0}^s$ defined from the band of M_n . For $n \geq r+s$ this implies $D = 0$ which is a contradiction to Lemma 1. This completes the proof of Corollary 1. \square

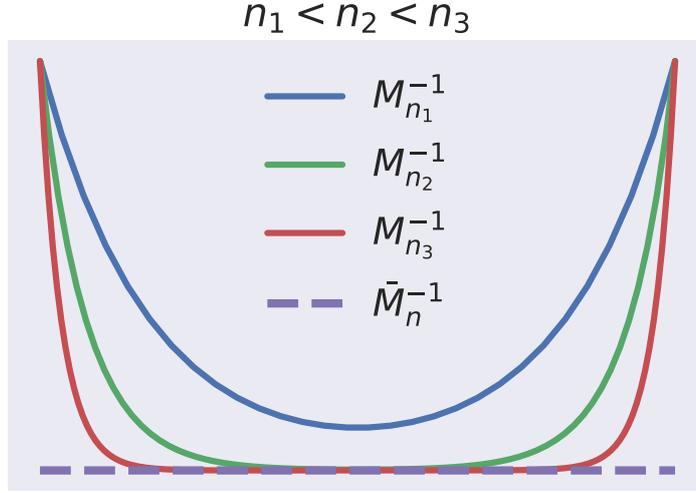


Figure 5.2: Schematic main diagonal of the matrices $M_{n_1}^{-1}$, $M_{n_2}^{-1}$, $M_{n_3}^{-1}$ with $n_3 > n_2 > n_1$ and of the approximate inverse \bar{M}_n^{-1} . The diagonals of the inverse M_n^{-1} converge to a constant value which corresponds to the main diagonal of \bar{M}_n^{-1} for $\lim_{n \rightarrow \infty}$ with exception of the borders.

Proof of Theorem 2. The upper left $s \times s$ and bottom right $r \times r$ corners of the matrix M_n can be modified based on two polynomials $A(x)$ and $B(x)$ of order r and s such that the inverse of the resulting matrix \bar{M}_n is Toeplitz. Let $|z_1|, |z_2|, \dots, |z_r|$ denote the roots of the polynomial $A(x)$ and $|z_{r+1}|, \dots, |z_{r+s}|$ the roots of the polynomial $B(x)$, respectively. The condition (5.27) implies

$$\lim_{|k| \rightarrow \infty} \lim_{n \rightarrow \infty} (\bar{M}_n^{-1})_{\lfloor n/2 \rfloor, \lfloor n/2 \rfloor + k} = 0 \quad (5.56)$$

as can be seen by Eq. (5.35). Then, the decay $\lim_{|k| \rightarrow \infty} \phi_k = 0$ implies

$|z_1|, |z_2|, \dots, |z_r| < 1$ and $1/|z_{r+1}|, \dots, 1/|z_{r+s}| < 1$, respectively. This completes the proof of Theorem 2. \square

5.4 Analytic solution of the tridiagonal case

In this section we assume M_n to be tridiagonal

$$M_n = \begin{pmatrix} m_1 & m_2 & 0 & \cdots & 0 \\ m_0 & m_1 & m_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & m_0 & m_1 & m_2 \\ 0 & \cdots & 0 & m_0 & m_1 \end{pmatrix}$$

In this case, the approximate Toeplitz inverse $(\bar{M}_n)_{i,j} = \phi_{i-j}$ has the property

$$\phi_k = K_1 \cdot \phi_{k-1} \quad 1 \leq k \leq n-1 \quad (5.57)$$

$$\phi_{-k} = K_2 \cdot \phi_{-k+1} \quad 1 \leq k \leq n-1 \quad (5.58)$$

and we can express K_1, K_2 and ϕ_0 in terms of the coefficients m_0, m_1 and m_2 , respectively. The polynomial $P(x) = m_0 + m_1 \cdot x + m_2 \cdot x^2$ has the two roots

$$p_{1,2} = -\frac{1}{2} \cdot \frac{m_1}{m_2} \pm \sqrt{\left(\frac{1}{2} \cdot \frac{m_1}{m_2}\right)^2 - \frac{m_0}{m_2}} \quad (5.59)$$

The condition (5.26) implies both roots to be real which means

$$m_1^2 > 4 \cdot m_0 \cdot m_2 \quad (5.60)$$

Let z_1, z_2 be the small and large root in absolute value of $P(\cdot)$, respectively. These roots can be expressed in terms of the sign of $m_1 \cdot m_2$

$$z_1 = -\frac{1}{2} \cdot \frac{m_1}{m_2} + \text{sign}(m_1 \cdot m_2) \cdot \sqrt{\left(\frac{1}{2} \cdot \frac{m_1}{m_2}\right)^2 - \frac{m_0}{m_2}} \quad (5.61)$$

$$z_2 = -\frac{1}{2} \cdot \frac{m_1}{m_2} - \text{sign}(m_1 \cdot m_2) \cdot \sqrt{\left(\frac{1}{2} \cdot \frac{m_1}{m_2}\right)^2 - \frac{m_0}{m_2}} \quad (5.62)$$

With

$$A(x) = a_1 + a_0 \cdot x = x - z_1 \quad (5.63)$$

$$B(x) = b_0 + b_1 \cdot x = m_2 \cdot (x - z_2) \quad (5.64)$$

and with the recurrence relations (5.6) - (5.7) follows

$$K_1 = -a_1/a_0 = z_1 \quad (5.65)$$

$$K_2 = -b_1/b_0 = 1/z_2 \quad (5.66)$$

$$(5.67)$$

and we can simplify the expression for K_2 by

$$1/z_2 = \frac{z_1}{z_2 \cdot z_1} \quad (5.68)$$

$$= \frac{m_2}{m_0} \cdot z_1 \quad (5.69)$$

$$= -\frac{1}{2} \cdot \frac{m_1}{m_0} + \text{sign}(m_1 \cdot m_2) \cdot \sqrt{\left(\frac{1}{2} \cdot \frac{m_1}{m_0}\right)^2 - \frac{m_2}{m_0}} \quad (5.70)$$

Finally, for ϕ_0 we have due to Eq. (5.9)

$$\phi_0 = 1/(a_0 b_0 - a_1 b_1) \quad (5.71)$$

$$= \frac{1}{m_2} \cdot \frac{1}{z_1 - z_2} \quad (5.72)$$

$$= \frac{1}{2} \cdot \text{sign}(m_1 \cdot m_2) \cdot (m_1^2/4 - m_0 \cdot m_2)^{-1/2} \quad (5.73)$$

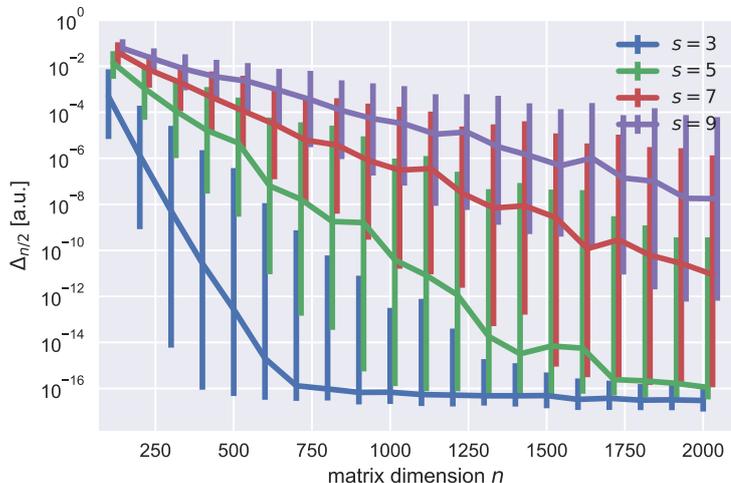


Figure 5.3: Second quartile (median) of the error $\Delta_{n/2}$ in dependence of the bandwidth s and matrix dimension n . The bottom and top caps indicate the first and third quartile of the error $\Delta_{n/2}$.

5.5 Time complexity experiments

In this section, we evaluate the accuracy of our approximate matrix inversion scheme in dependence of the bandwidth parameter $s = r$ and matrix dimension n . To this end, we generate a set of random banded Toeplitz matrices with m_i i.i.d. from the uniform distribution $U(0, 1)$ for $i = -s, \dots, s$. We select the subset of these matrices which meet the criterion (5.26). For these matrices, we compute the exact inverse with a state-of-the-art method for inverting banded Toeplitz matrices by Trench et al. [125]. We evaluate the accuracy of our approximate inverse by the mean absolute error of the row $n/2$ of the difference to the true inverse

$$\Delta_{n/2} := \frac{1}{n} \cdot \sum_{i=0}^{n-1} |(M_n^{-1})_{n/2,i} - (\bar{M}_n^{-1})_{n/2,i}| \quad (5.74)$$

For the experiment, we have chosen $s = r$ and the bandwidths $s = 3, 5, 7, 9$ and a set of matrix dimensions ranging from $n = 100$ to $n = 2000$, respectively. To get a more stable estimate of the error $\Delta_{n/2}$, we repeat the experiment 100 times for each combination of s and n . The result of the first three quartiles of the error $\Delta_{n/2}$ is shown in Fig. 5.3. As expected, the error $\Delta_{n/2}$ decreases with increasing matrix dimension. However, the convergence rate is slower for larger bandwidths. We attribute this to a larger probability of a root of the polynomial $P(\cdot)$ in Eq. (5.2) being closer to one for larger s for our input matrix distribution. The complexity of constructing the row $n/2$ of our approximate

inverse is $\mathcal{O}(s^3 + s \cdot n)$. In contrast, the implemented exact method by Trench et al. [125] computes s quotients of two $s \times s$ determinants for each element of the inverse and therefore takes $\mathcal{O}(s^4 \cdot n)$. In a second experiment, we evaluate the time complexity for the set of bandwidths $s = \{2, 5, 10, 15, 20, 25, 30\}$ and a fixed matrix dimension of $n = 100000$. We have repeated the experiment 100 times to get a more robust time estimate. The comparison of the time complexity is shown in Fig. 5.4.

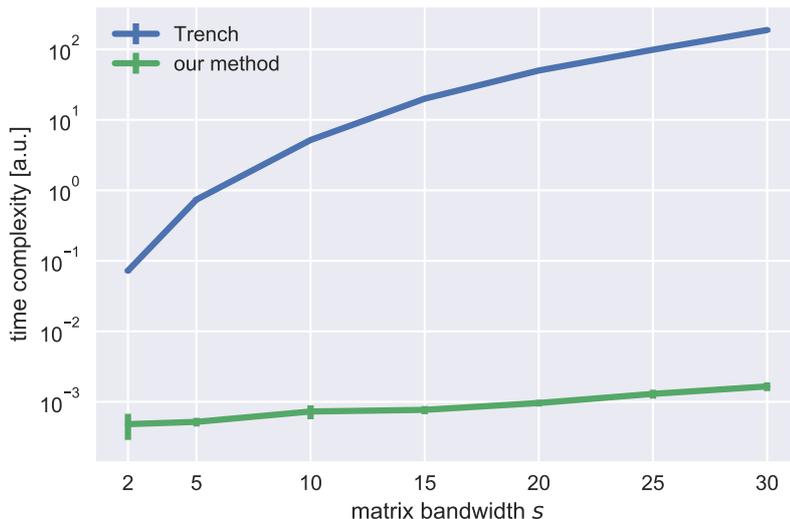


Figure 5.4: Time complexity of an exact banded Toeplitz matrix inversion scheme by Trench et al. (blue) and our approximate inversion scheme (green) in dependence of the matrix bandwidth s for a fixed matrix dimension of $n = 100000$.

In addition to having a larger slope, the method by Trench et al. is at least one order of magnitude slower compared to our method for small bandwidths s . The error of our method in this experiment is close to numerical precision which is attributed to the large matrix dimension of $n = 100000$. This experiment confirms the preferable theoretical time complexity of our method.

5.6 Constructing Green's functions

Typically, finding a Green's function $G(x, s)$ of a given differential operator $L(x)$ is a difficult task and can be solved only for special cases. The Green's function is defined by any solution of

$$L(x)G(x, s) = \delta(x - s) \quad (5.75)$$

with the Dirac delta function $\delta(\cdot)$. If $L(x)$ is translation invariant, i.e. it has constant coefficients with respect to x , the corresponding Green's function

$G(x, s)$ can be interpreted as a convolution operator $G(x, s) = G(x - s)$. Using finite-difference methods, the discretized operator of L results in a banded Toeplitz matrix form as the corresponding differential operator has finite order. Let $\{m_i\}_{i=-s}^s$ denote the band of the discretized operator of L with order s after applying a central difference scheme of desired accuracy. If the criterion (5.26) is conformed for this band, we propose to obtain the discrete Green's function in the interval $[x_l, x_r]$ by the Alg. 7.

Typically, the Green's function is not unique. Our proposed approximate Green's function \mathbf{g} meets the boundary condition $\lim_{|k| \rightarrow \infty} \mathbf{g}_k = 0$. Any other Green's function of the operator L can be constructed using the general homogeneous solution of $L\mathbf{x} = 0$. A particular solution of $L\mathbf{x} = \mathbf{y}$ is readily obtained by the convolution operation $\mathbf{x} = \mathbf{g} * \mathbf{y}$ (see Eq. (5.28)).

Analytic derivation of Green's functions

The method in Alg. 7 allows to construct Green's functions given a stepsize h which is used to obtain the discretized band of the differential operator L . This is useful in practice as numerical methods typically intrinsically incorporate such a stepsize parameter. However, we can go one step further and derive the theoretical Green's functions with our approach directly.

The principal idea is to perform the limit $h \rightarrow 0$. We exemplify the derivation for the differential operator of order one $L(x) = \frac{d}{dx} + \gamma$ with the Green's function $G(x) = \Theta(x)e^{-\gamma x}$. The band of the discretized operator is given by $\vec{m} = (-1/(2h), \gamma, 1/(2h))^T$ and the corresponding matrix is tridiagonal. For tridiagonal matrices our methods have been analytically solved in the previous Sec. 5.4. With $h = x/n$ we can derive the theoretical Green's function with our method by

$$G(x) = \lim_{n \rightarrow \infty} \left(\frac{1}{h} \cdot \phi_0 \cdot z_1^n \right) \quad (5.76)$$

$$= \lim_{n \rightarrow \infty} \left(\frac{1}{2} \cdot \frac{1}{\sqrt{\frac{\gamma^2 h^2}{4} + \frac{1}{4}}} \cdot z_1^n \right) \quad (5.77)$$

$$= \lim_{n \rightarrow \infty} z_1^n \quad (5.78)$$

$$= \lim_{n \rightarrow \infty} \left(-\frac{\gamma x}{n} + \sqrt{\left(\frac{\gamma x}{n}\right)^2 + 1} \right)^n \quad (5.79)$$

$$= \lim_{n \rightarrow \infty} \left(-\frac{\gamma x}{n} + (1 + O(1/n^2)) \right)^n \quad (5.80)$$

$$= \lim_{n \rightarrow \infty} \left(1 - \frac{\gamma x}{n} \right)^n \quad (5.81)$$

$$= e^{-\gamma x} \quad (5.82)$$

which corresponds to the theoretical result. The $\frac{1}{h}$ term in Eq. (5.76) is due to the discrete representation of the Dirac delta function $\delta(\cdot)$. The step from

5.7. Banded approximation of deconvolution operators

Name	Operator L	Green's function G	Parameters
Critically damped harmonic oscillator	$\partial_t^2 + 2\gamma\partial_t + \gamma^2$	$\Theta(t)te^{-\gamma t}$	$\gamma = 5$
Overdamped harmonic oscillator	$\partial_t^2 + 2\gamma\partial_t + \omega_0$	$\Theta(t)e^{-\gamma t} \sin(\omega t)/\omega$	$\omega_0 = 15, \gamma = 5,$ $\omega := \sqrt{\omega_0^2 - \gamma^2}$
Screened Poisson equation	$\partial_t^2 - k^2$	$-\frac{1}{\sqrt{2\pi}} \frac{t}{k} K_{-1/2}(kt)$	$k = 2.5$

Table 5.1: Differential operators used in the experiment with their corresponding Green's functions and parameters.

Eq. (5.80) to Eq. (5.81) can be shown by Bernoulli's inequality. The exact derivation of the Green's functions can be performed for second order differential operators as there is a closed form solution of the roots of quartic functions (polynomials of degree four). We leave the derivation for differential operators of order higher than two for future work.

Experiments

We apply the Alg. 7 to three differential operators with well known non-trivial Green's functions. The operators and their corresponding Green's functions and parameters used are listed in Tab. 5.1, where $\Theta(\cdot)$ is the Heaviside step function and $K_\nu(\cdot)$ is a modified Bessel function of the second kind, respectively. We measure the accuracy by the integrated error

$$\Delta_{\text{Green}} = h \cdot \sum_{i=0}^{\lfloor x_r/h \rfloor} (G(h \cdot i) - (\mathbf{g})_i)^2 \quad (5.83)$$

where we used the list of stepsizes $h = \{0.1, 0.05, 0.01, 0.005, 0.001\}$ and the interval $[0, x_r]$ with $x_r = 10$ in the Alg. 7 to obtain \mathbf{g} , respectively. The results in dependence of the stepsize h are shown in Fig. 5.5. It can be seen, that our approximate Green's functions converge to the theoretical solution as the stepsize h decreases.

A more complex experiment is shown in Sec. A.1, where we apply our approximate deconvolution operations on blurred images in two dimensions which demonstrates the application of our methods on real-world image data.

5.7 Banded approximation of deconvolution operators

Toeplitz matrices can be viewed as one-dimensional convolution operators. The functions with which we perform convolutions often decay to zero in practice, as often there is a natural correlation between the geometric distance of sampled points and the covariance of the corresponding random variables. In principal, this suggests to approximate such correlation functions or kernels by a banded

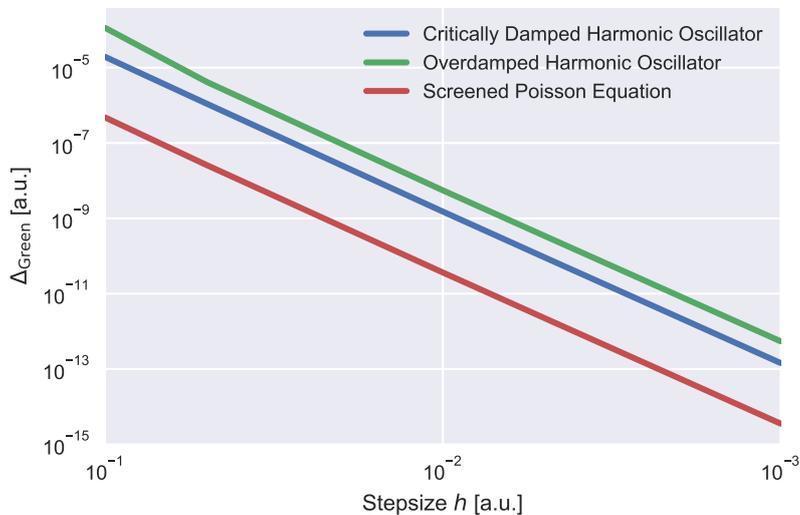


Figure 5.5: Error of approximate Green's functions in dependence of the stepsize h . The errors are relative to the theoretical Green's function.

Toeplitz matrix. However, this distorts the convolution function at the cutoff which introduces building errors in the recurrence relations. Instead, we propose to model the inverse procedure, the deconvolution operation, by a banded Toeplitz matrix. In this way, we construct an approximate deconvolution operator which best reconstructs the original local kernel around the position zero, thereby circumventing a signal distortion. This can be done with the Alg. 6.

Experiments

We use the base convolution functions (or kernels) listed in Tab. 5.2 to construct the corresponding deconvolution operators as the inverse of $K + \lambda \cdot I$ with the regularization parameter λ . We measure the error of our approximation by the relative error

$$\Delta_{\text{band}} = \frac{\sqrt{\frac{1}{2s+1} \sum_{i=-s}^s (d_{n/2+i} - m_i)^2}}{\sqrt{\frac{1}{2s+1} \sum_{i=-s}^s (d_{n/2+i})^2}} \quad (5.84)$$

where $d_i := ((K + \lambda \cdot I)^{-1})_{n/2,i}$ and we used $n = 5000$ in our experiment. The results for a set of regularization parameters λ in dependence of the bandwidth s expressed in terms of the standard deviation of the convolution operator $x_{\text{cutoff}} \cdot \sigma := s/n$ are shown in Figs. 5.6, 5.7 and 5.8, respectively.

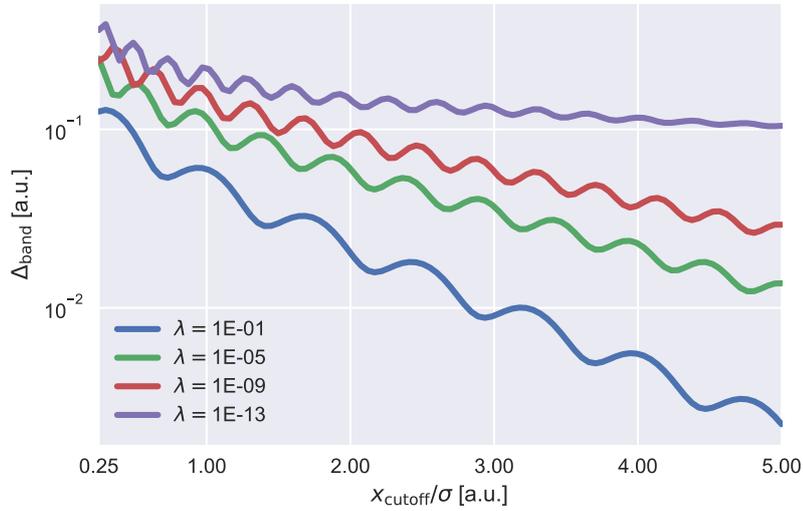


Figure 5.6: Error of Gaussian deconvolutions for different regularization parameters in dependence of the assumed bandwidth of the deconvolution operator expressed in terms of the scale of the kernel.

Name	Kernel $K(x)$	Parameters
Gauss	$e^{-\frac{x^2}{2\sigma^2}}$	$\sigma = 0.1$
Laplace	$e^{-\frac{ x }{\sigma}}$	$\sigma = 0.1$
Matérn	$(1 + \frac{\sqrt{5} x }{\sigma} + \frac{5x^2}{3\sigma^2}) \cdot e^{-\frac{\sqrt{5} x }{\sigma}}$	$\sigma = 0.1$

Table 5.2: Base kernels used in the experiment to compute the respective deconvolution operators.

For the Gaussian kernel, the error increases with decreasing the regularizer λ . The error is worse compared to the Laplace and Matérn case, as the Gaussian deconvolution operators are not well approximated by a banded matrix. For the Laplace and Matérn deconvolutions, the error rapidly decreases as one both, decreases the regularization or increases the bandwidth, respectively. Intriguingly, we only need the knowledge of a small portion of the Laplace and Matérn convolution operator around zero to compute its inverse operator.

5.8 Many-body van der Waals interactions

In this section, we apply our matrix inversion scheme to the computation of long-range dispersion or van der Waals interactions in molecules and condensed matter. Van der Waals interactions play a crucial role in a variety of fields

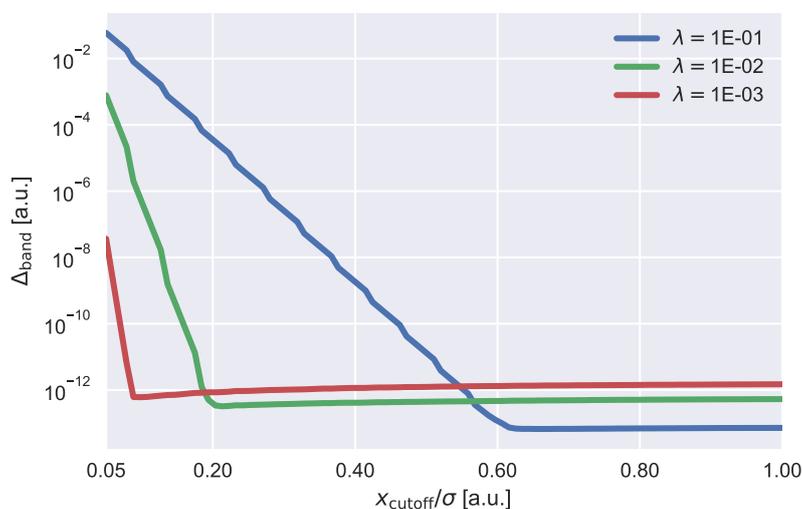


Figure 5.7: Error of Laplace deconvolutions for different regularization parameters in dependence of the assumed bandwidth of the deconvolution operator expressed in terms of the scale of the kernel.

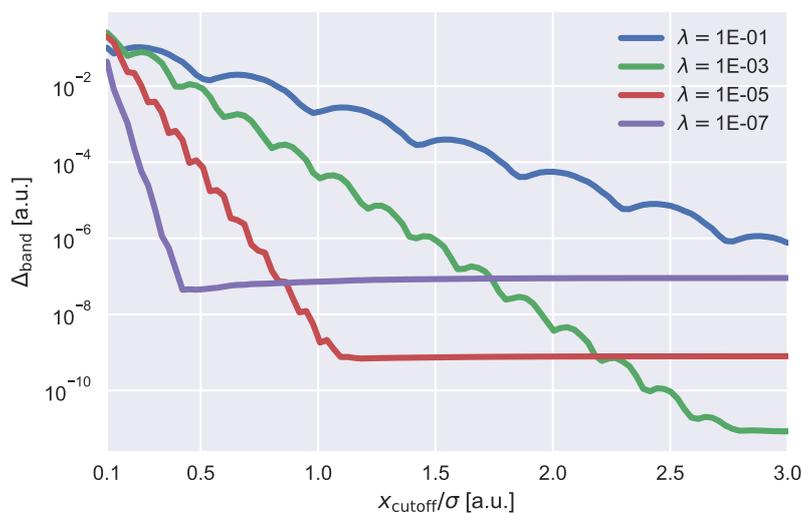


Figure 5.8: Error of Matérn deconvolutions for different regularization parameters in dependence of the assumed bandwidth of the deconvolution operator expressed in terms of the scale of the kernel.

such as supramolecular chemistry, structural biology, nanotechnology, surface science and condensed matter physics. The nature of the van der Waals force is

attractive for distances between particles larger than typical covalent bonding distances and becomes repulsive for smaller distances. Originating from correlations of fluctuating instantaneous polarizations, these many-body dispersion effects can only be described by quantum dynamics.

The frequency-dependent polarizability in molecules and materials with a finite electronic gap can be accurately computed by a many-body dispersion method which uses a system of coupled quantum harmonic oscillators (QHOs) within the random-phase approximation in combination with density functional theory [133], where each of the QHO represents an atom in a molecular system of interest. This method starts from the ground state and the corresponding charge density of a QHO which is a spherical Gaussian function, from which the Coulomb interaction between two QHOs is derived as

$$\nu_{pq} = \frac{\text{erf}(R_{pq}/\sigma_{pq})}{R_{pq}} \quad (5.85)$$

with the Gauss error function $\text{erf}(\cdot)$, the distance between the QHOs R_{pq} and the effective width $\sigma_{pq} = \sqrt{\sigma_p^2 + \sigma_q^2}$ obtained from the widths of the QHOs σ_p and σ_q , respectively. These widths are related to the polarizabilities by $\sigma_p = (\sqrt{2/\pi} \cdot \alpha_p/3)^{1/3}$ in a classical electrodynamics treatment. The dipole-dipole interaction tensor can be now derived as

$$\begin{aligned} T_{pq}^{ab} = & -\frac{3R_a R_b - R_{pq}^2 \delta_{ab}}{R_{pq}^5} \times \left(\text{erf}(R_{pq}/\sigma_{pq}) - \frac{2}{\sqrt{\pi}} \frac{R_{pq}}{\sigma_{pq}} \exp(-(R_{pq}/\sigma_{pq})^2) \right) \\ & + \frac{4}{\sqrt{\pi}} \frac{R_a R_b}{\sigma_{pq}^3 R_{pq}^2} \exp(-(R_{pq}/\sigma_{pq})^2) \end{aligned} \quad (5.86)$$

where a and b represent the Cartesian coordinates $\{x, y, z\}$ and R_a, R_b are the respective components of the QHO distance R_{pq} . In the Tkatchenko-Scheffler scheme, the local chemical environment is accounted for by modeling the frequency-dependent polarizability by

$$\alpha_p(i\omega) = \frac{\alpha_p^0}{1 + (\omega/\omega_p)^2} \quad (5.87)$$

where the static dipole polarizability $\alpha_p^0[n(\mathbf{r})]$ and the effective excitation frequency $\omega_p[n(\mathbf{r})]$ are functionals of the ground-state electron density $n(\mathbf{r})$ of the system. This model for the frequency-dependent polarizability lacks to capture long-range electrodynamic response screening and anisotropy effects which can be included by self-consistently solving the system of linear equations for a given frequency ω

$$\bar{\alpha}_p(i\omega) = \alpha_p(i\omega) - \alpha_p(i\omega) \sum_{q \neq p}^N T_{pq} \bar{\alpha}_q(i\omega) \quad (5.88)$$

for $p = 1, 2, \dots, N$ where N denotes the number of QHOs and the $\bar{\alpha}_p(i\omega)$ are the polarizabilities of the system that account for both short-range and

long-range electrodynamic response screening effects. The Eqs. (5.88) can be solved by constructing the Hermitian matrix A which contains the inverse of the frequency-dependent polarizability tensors $\alpha_p^{-1}(i\omega)$ on the diagonal 3×3 subblocks and the dipole-dipole interaction tensor T_{pq} on the non-diagonal 3×3 subblocks. By inverting the matrix A , the screened set of polarizability tensors are obtained by

$$\bar{\alpha}_p(i\omega) = \sum_q^N B_{pq} \quad (5.89)$$

We now proceed to solve the Eqs. (5.88) for a model system of QHOs by using our matrix inversion scheme introduced in Sec. 5.1.

Linear chain of QHOs

We compute the set of screened polarizabilities for a system which is composed of a linear chain of atoms where neighboring atoms are separated by a distance d . Due to the equivalence of the atoms of the chain, we model both the frequency-dependent polarizability and its screened counterpart as independent of the QHO index $a(i\omega) := a_p(i\omega)$ and $\bar{a}(i\omega) := \bar{a}_p(i\omega)$ for all $p = 1, 2, \dots, N$. For simplicity we analyse a single component of the polarizabilities by replacing the 3×3 tensor by real numbers $a(i\omega), \bar{a}_p(i\omega) \in \mathbb{R}$ for a given frequency ω . The self-consistency cycle is now given by

1. select a fixed frequency and corresponding polarizability $a(i\omega)$ and choose the starting $\bar{a}(i\omega)$ randomly
2. construct the matrix A from the polarizability $a(i\omega)$ and the dipole-dipole interaction tensor T_{pq} which is a function of $\bar{a}(i\omega)$
3. compute $B = A^{-1}$ and the screened polarizability $\bar{a}(i\omega) = \sum_q^N B_{pq}$ for $p = N/2$
4. iterate steps (2.) - (3.) until convergence of $\bar{a}(i\omega)$

To apply our matrix inversion scheme, we introduce a banded variant of the matrix A by

$$(\bar{A}_{n_B})_{ij} := \begin{cases} (A)_{ij} & |i - j| \leq n_B \\ 0 & \text{otherwise} \end{cases} \quad (5.90)$$

which corresponds to a cutoff distance $d \cdot n_B$ for which the dipole-dipole interaction tensor T_{pq} becomes zero, if the distance between the oscillators p and q exceeds that distance. The number $2 \cdot n_B$ corresponds to the maximal number of interacting oscillators.

We apply the self-consistency cycle above to solve the system of linear Eqs. (5.88) for $N = 150$ using the full matrix A and the banded variants \bar{A}_{n_B}

for $n_B = 5, 10, 15$. We use our approximate matrix inversion scheme of Sec. 5.1 to compute the inverses $\bar{B}_{n_B} = \bar{A}_{n_B}^{-1}$ in the self-consistency cycle. The result for the screened polarizability $\bar{a}(i\omega)$ in dependence of QHO distance d for a polarizability of $a(i\omega) = 12 \text{ Bohr}^3$ is shown in Fig. 5.9.

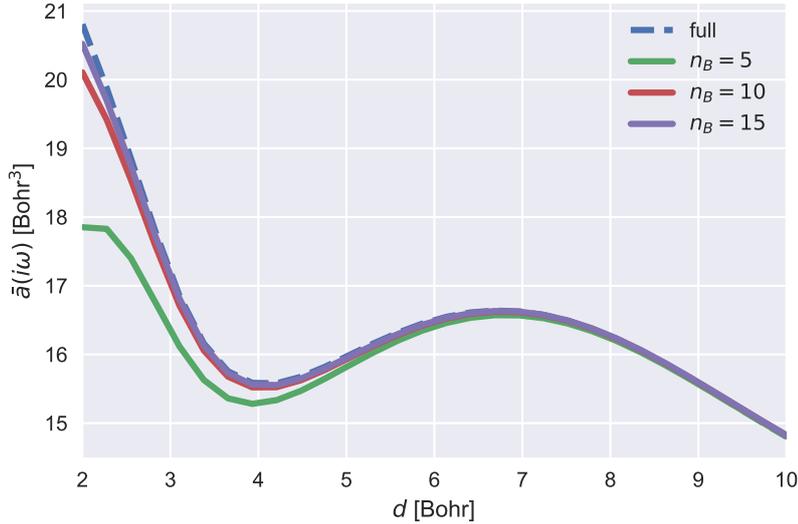


Figure 5.9: Screened polarizability in dependence of the QHO distance d of the full model (blue) and approximate matrix inversion schemes for $n_B = 5, 10, 15$ at the frequency-dependent polarizability of $a(i\omega) = 12 \text{ Bohr}^3$.

The screened polarizability is in good agreement with the full model by taking a small number of interacting oscillators into account $n_B \ll N$ and the local minimum around 4 Bohr is qualitatively reconstructed by all models used. For larger number of interacting oscillators n_B , the screened polarizability converges to the full model of using the matrix A to solve the Eqs. (5.88). The convergence rate is better for larger distances d , indicating the importance of many-body effects for small distances between the particles of the system to accurately compute dispersion effects.

Fig. 5.10 shows the screened polarizability $\bar{a}(i\omega)$ in dependence of the frequency or equivalently the polarizability $a(i\omega)$ at a QHO distance of $d = 4$ Bohr. The convergence rate is better for larger frequencies (or equivalently smaller polarizabilities $a(i\omega)$), indicating the importance of many-body effects for small frequencies to accurately compute dispersion effects.

The complexity of our approximate matrix inversion scheme is $\mathcal{O}(n_B^3)$ which drastically improves upon the complexity of the full model $\mathcal{O}(N^3)$ for $n_B \ll N$. However, as the system of QHOs is qualitatively similar by decreasing the total number of oscillators N , there is the question how well an exact but smaller model $N_{\text{small}} < N$ can approximate the full problem at $N = 150$. Especially, we

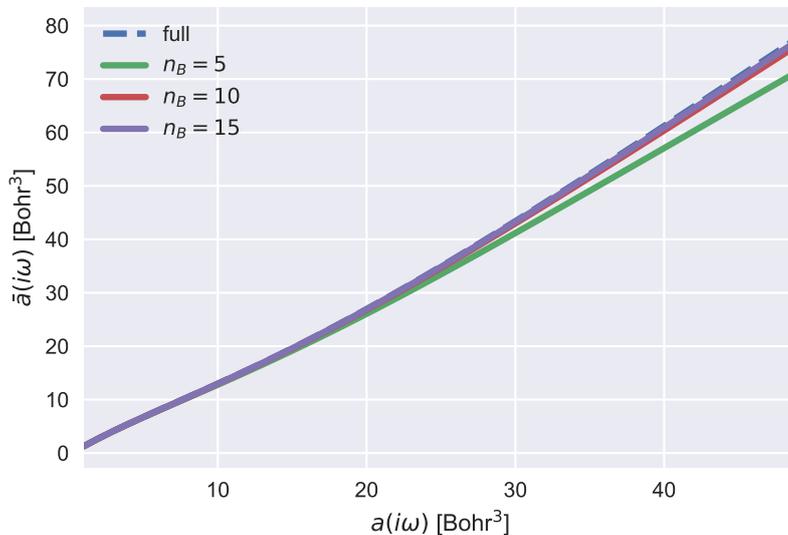


Figure 5.10: Screened polarizability in dependence of the frequency-dependent polarizability of the full model (blue) and approximate matrix inversion schemes for $n_B = 5, 10, 15$ at a QHO distance of $d = 4$ Bohr.

compare the setting where both, the smaller exact model and our approximate scheme share the same time complexity $N_{\text{small}} = n_B$. Fig. 5.11 shows the accuracy of a set of smaller exact and approximate models for computing the full screened polarizability of our initial setting at $N = 150$ in dependence of the number of interacting oscillators used (N_{small} for the small exact and n_B for the approximate model, respectively) which dictate the time complexity. The result shows the true benefit of our method as the accuracy is one order of magnitude better compared to the exact smaller model at the same time complexity.

5.9 Interpolation of potential energy surfaces

In this section, we apply our approximate matrix inversion scheme to the interpolation of potential energy surfaces. In polynomial interpolation, the values of a function of interest at a set of positions is predicted based on a polynomial which passes through a set of training points. The interpolation technique used depends on the assumptions which are made about the defining polynomials and can incorporate prior knowledge about the data. Common applications include the interpolation of potential energy surfaces like molecular potential energies [134, 135, 136], reaction surfaces [137, 138, 139] and energy-minimization, where spline interpolation [140, 141] and variants like discrete

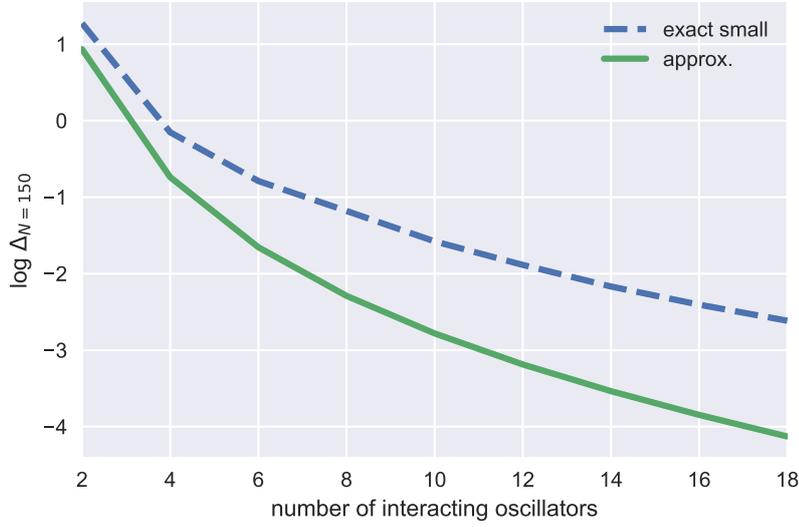


Figure 5.11: Accuracy of the small exact (blue) and approximate (green) model in dependence of the time complexity which is dictated by the number of interacting oscillators (N_{small} for the full and n_B for the approximate model, respectively).

splines [142] has been used in the literature.

In cubic spline interpolation, a set of points $\{(x_i, y_i)\}$ for $i = 0, 1, \dots, n$ is interpolated by a set of piecewise polynomials of order three called splines. These splines are constructed to be twice continuously differentiable at the interpolation points $\{x_i\}_{i=0}^n$. For simplicity, we assume equidistant interpolation points with separation distance h for which the splines are constructed by

$$s_i(x) = \frac{1}{6h} \cdot ((x_{i+1} - x)^3 \cdot M_i + (x - x_i)^3 \cdot M_{i+1}) + c_i \cdot (x - x_i) + d_i \quad (5.91)$$

for $i = 0, 1, \dots, n - 1$ and $x \in [x_i, x_{i+1}]$ with the moments $\{M_i\}_{i=0}^n$ and

$$d_i := y_i - \frac{h^2}{6} \cdot M_i \quad (5.92)$$

$$c_i := \frac{y_{i+1} - y_i}{h} - \frac{h}{6} \cdot (M_{i+1} - M_i) \quad (5.93)$$

The moments $\vec{M} := (M_0, M_1, \dots, M_n)^\top$ can be obtained by solving the system of linear equations

$$T \cdot \vec{M} = \vec{Y} \quad (5.94)$$

where T is the tridiagonal Toeplitz matrix with the value 4 on the main diagonal and the value 1 on the first left and right sub-diagonal and $\vec{Y} :=$

$(Y_0, Y_1, \dots, Y_n)^\top$ is defined by

$$Y_i = \begin{cases} \frac{6}{h^2} \cdot (y_{i+1} - 2 \cdot y_i + y_{i-1}) & i = 1, 2, \dots, n-1 \\ \frac{6}{h^2} \cdot (y_1 - y_0) & i = 0 \\ \frac{6}{h^2} \cdot (y_{n-1} - y_n) & i = n \end{cases} \quad (5.95)$$

Note that for the tridiagonal matrix T the corresponding linear matrix equations can be efficiently solved by the Thomas algorithm. However, for computing the interpolated function value at a given point x , the exact Thomas algorithm has the time complexity $\mathcal{O}(n)$ which depends on the number of interpolation or training points n . Intuitively, the function values at the position x are dominated by its nearest neighbors. Our matrix inversion scheme implements this intuition by computing the inverse of T which is Toeplitz efficiently around the main diagonal. As the matrix T is strongly diagonally dominant, i.e. the diagonal entry is larger in magnitude than the sum of the non-diagonal entries, the inverse T^{-1} decays quickly which suggests an accurate approximation of our method. We propose approximate cubic splines by introducing a cutoff number n_B of the nearest neighbors taken into account for computing the convolution operation in Eq. (5.29) by

$$\bar{M}_i := \sum_{k=0}^{n_B-1} \phi_{k-i} \cdot Y_k \quad (5.96)$$

for $i = 0, 1, \dots, n$ which defines the approximate moments $\{\bar{M}_i\}_{i=0}^n$ of the cubic splines.

We apply our approximate spline interpolation on the potential energy surfaces of the ethanol molecule, where we rotate the CH_3 -group (methyl group) and the OH -group (hydroxy group) along an angle between $[0, 2\pi]$ from the equilibrium position, where the C-C axis is fixed for the methyl group and the C-O axis is fixed for the hydroxy group, respectively. Each point represents the energy of the molecule for a fixed angle of the respective functional group while all other degrees of freedom have been relaxed. The energies have been computed using all-electron coupled cluster with single, double, and perturbative triple excitations (CCSD(T)) with the Dunning's correlation-consistent basis set cc-pVTZ.

Fig. 5.12 shows the potential energy surfaces (left) along with the interpolation points and the approximate spline interpolation of the remaining points together with the mean absolute prediction error (right) by increasing the number of nearest neighbors n_B , respectively. The mean absolute prediction error rapidly decreases and saturates at $n_B = 2$ at the prediction error of the exact spline interpolation.

In a second experiment, we apply our approximate spline interpolation on minimum energy paths of reactions of the kind $\text{X}^- + \text{H}_3\text{C}-\text{Y} \rightarrow \text{X}-\text{CH}_3 + \text{Y}^-$ for all combinations of $\text{X}, \text{Y} \in \{\text{F}, \text{Cl}, \text{Br}, \text{I}\}$. The reaction coordinate of this reaction is defined as $r_{\text{CY}} - r_{\text{CX}}$. Figs. 5.13 and 5.14 show minimum energy paths (left) along with the interpolation points and the approximate spline interpolation of

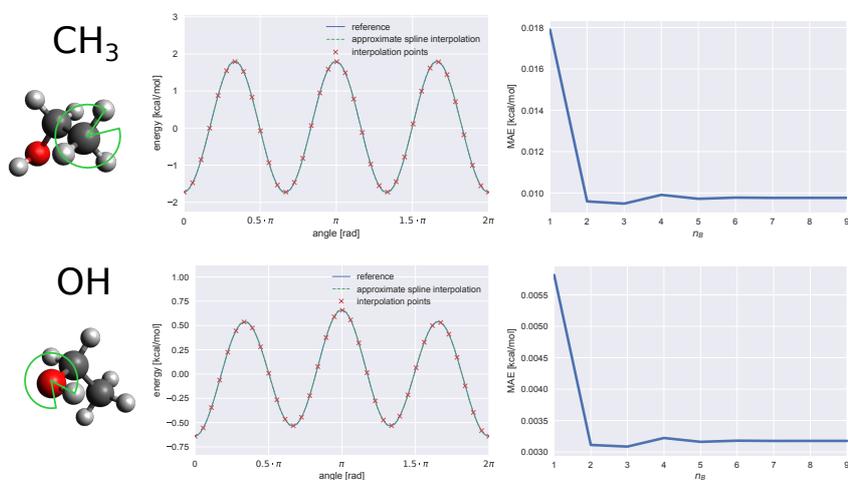


Figure 5.12: Approximate spline interpolation (middle) for $n_B = 3$ of the potential energy surfaces obtained by rotating the CH₃-group (top) and the OH-group (bottom) of ethanol from equilibrium position. The reference CCSDFT energy in kcal/mol (blue solid line) is shown along with the interpolation points (red crosses) and the approximate spline interpolation of the remaining points (green dashed line). The mean absolute error of approximate spline interpolation is given in kcal/mol in dependence of the number of interpolation points n_B (right).

the remaining points together with the mean absolute prediction error (right) by increasing the number of nearest neighbors n_B , respectively.

The mean absolute prediction error rapidly decreases and saturates at $n_B = 3$ at the prediction error of the exact spline interpolation. A larger number of neighbors n_B have to be taken into account compared to the ethanol rotor energy surfaces which we attribute to more complex functions at equal sampling density.

The results of both experiments indicate that our approximate cubic spline interpolation can be used to accurately predict potential energy surfaces by taking into account a low number of training points. This can potentially be useful for energy-minimization applications which typically involve a low number of sample points at which the energy is evaluated.

5.10 Summary and discussion

In this chapter, we have developed algorithms to efficiently approximate the inverses of a special class of banded Toeplitz matrices in a numerically stable way. Specifically, we have approximated these inverses by matrices which are themselves Toeplitz. We have provided sufficient and necessary conditions

when this is possible in terms of the roots of a polynomial which is given by the band defining the original matrix. This criterion can be efficiently computed without having to construct the involved matrices which potentially saves storage space and computational cost in practice.

Our approach can be used to construct the Green's functions of one-dimensional operators with constant coefficients. To the knowledge of the author, this is a new approach for computing the Green's functions of differential operators of arbitrary order in a direct and explicit way. In another application, we have demonstrated the applicability of our method to approximate the inverses of convolution kernel functions while retaining a certain reconstruction accuracy. We have applied our method to compute the polarizabilities including long-range electrodynamic response screening effects more efficiently. Finally, we have proposed approximate cubic spline interpolation using our matrix inversion scheme to accurately predict potential energy surfaces using a small number of training points, an approach which can potentially be useful for energy minimization strategies.

Algorithm 6 Toeplitz2Band**Input:**

$$\vec{\varphi} = (\phi_{-r}, \dots, \phi_s)$$

Output:

$$\vec{m} = (m_{-r}, \dots, m_s)$$

```

1:  $y_0 \leftarrow 1$ 
2: for  $i = 1$  to  $r + s + 1$  do
3:    $y_i \leftarrow 0$ 
4: for  $i = 0$  to  $s - 1$  do
5:    $M_{0,i} \leftarrow \varphi_{s+1+i}$  ▷ The dimension of  $M$  is  $(r + s + 2, r + s + 2)$ 
6:    $M_{0,2s+1-i} \leftarrow -\varphi_{s+1+i}$ 
7:  $M_{1,r} \leftarrow 1$ 
8:  $M_{1,r+1} \leftarrow -1$ 
9: for  $i = 1$  to  $s$  do
10:  for  $j = 0$  to  $r$  do
11:     $M_{1+i,j} \leftarrow \varphi_{r+i-j}$ 
12: for  $i = 1$  to  $r$  do
13:  for  $j = 0$  to  $s$  do
14:     $M_{1+s+i,r+1+j} \leftarrow \varphi_{-r+i+j}$ 
15:  $\vec{x} \leftarrow \text{solve } M \cdot \vec{x} = \vec{y}$ 
16: for  $i = 0$  to  $r$  do
17:    $a_i \leftarrow x_i/x_0$ 
18: for  $i = 0$  to  $s$  do
19:    $b_i \leftarrow x_{r+1+i} \cdot x_0/x_r$ 
20: for  $i = 0$  to  $r + s$  do
21:    $m_i \leftarrow 0$ 
22: for  $i = 0$  to  $s$  do
23:    $m_s \leftarrow m_s + a_i \cdot b_i$ 
24: for  $i = 1$  to  $r$  do
25:    $l \leftarrow \min(r - i, s)$ 
26:   for  $j = 0$  to  $l$  do
27:      $m_{i-1} \leftarrow m_{i-1} + a_{i+j} \cdot b_j$ 
28: for  $i = 1$  to  $s$  do
29:    $l \leftarrow \min(r, s - i)$ 
30:   for  $j = 0$  to  $l$  do
31:      $m_{r+i-1} \leftarrow m_{r+i-1} + b_{i+j} \cdot a_j$ 
32: return  $\vec{m}$ 

```

Algorithm 7 PronobisGreens

Input:

$$\vec{m} = (m_{-s}, \dots, m_s)$$

$$h$$

 \triangleright stepsize parameter

$$x_l, x_r$$

 \triangleright lower and upper bounds**Output:**

$$\vec{g}$$

 \triangleright Green's function in the interval $[x_l, x_r]$ 1: $\vec{\varphi} \leftarrow \text{Band2Toeplitz}(\vec{m})$ 2: $z_1, z_2, \dots, z_{2s} \leftarrow$ ordered roots of the polynomial $\sum_{i=0}^{2s} m_{-s+i} \cdot x^i$ 3: $a_0, a_1, \dots, a_s \leftarrow$ such that $\sum_{i=0}^s a_{s-i} \cdot x^i = \prod_{i=1}^s (x - z_i)$ 4: $b_0, b_1, \dots, b_s \leftarrow$ such that $\sum_{i=0}^s b_i \cdot x^i = m_s \cdot \prod_{i=s+1}^{2s} (x - z_i)$ 5: $n_l \leftarrow x_l/h$ 6: $n_r \leftarrow x_r/h$ 7: **for** $i = 1$ to $s - 1$ **do**8: $a_i \leftarrow -a_i/a_0$ 9: $b_i \leftarrow -b_i/b_0$ 10: $g_{nl} \leftarrow \varphi_s$ 11: **for** $i = 1$ to $s - 1$ **do**12: $g_{n_l+i} \leftarrow \varphi_{s+i}/h$ 13: $g_{n_l-i} \leftarrow \varphi_{s-i}/h$ 14: **for** $i = s$ to $n_r - 1$ **do**15: **for** $j = 1$ to s **do**16: $g_{n_l+i} \leftarrow g_{n_l+i} + a_j \cdot g_{n_l+i-j}$ 17: **for** $i = s$ to $n_l - 1$ **do**18: **for** $j = 1$ to s **do**19: $g_{n_l-i} \leftarrow g_{n_l-i} + b_j \cdot g_{n_l-i-j}$ 20: **return** \vec{g}

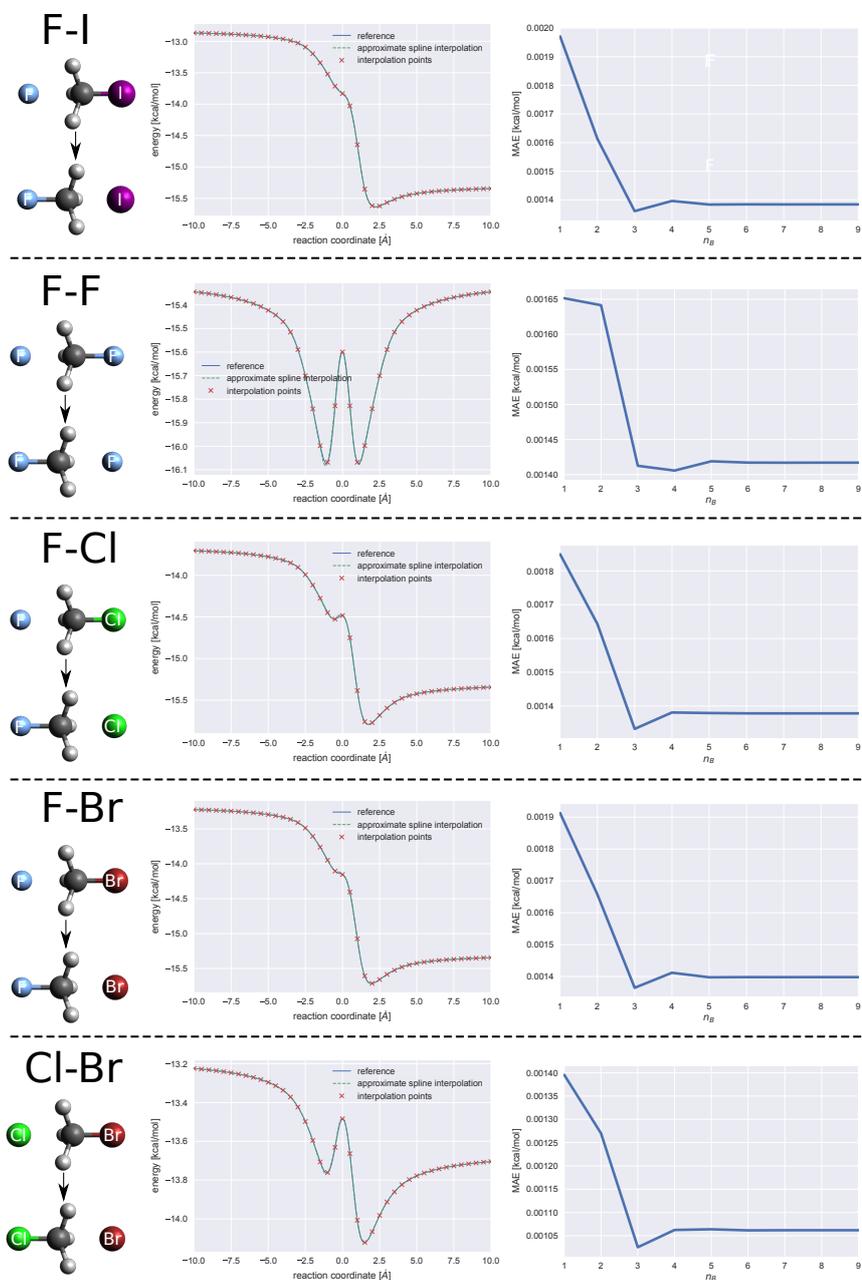


Figure 5.13: Approximate spline interpolation (middle) for $n_B = 3$ of the minimum energy paths of the reaction $X^- + H_3C-Y \rightarrow X-CH_3 + Y^-$ (left) for $(X,Y) = \{ (F-I), (F-F), (F-Cl), (F-Br), (Cl-Br) \}$. The reference energy in kcal/mol (blue solid line) has been computed at the DSD-BLYP-D3(BJ)/def2-TZVP level of theory [139] and is shown along with the interpolation points (red crosses) and the approximate spline interpolation of the remaining points (green dashed line). The mean absolute error of approximate spline interpolation is given in kcal/mol in dependence of the number of interpolation points n_B (right).

5. APPROXIMATE BANDED TOEPLITZ MATRIX INVERSION

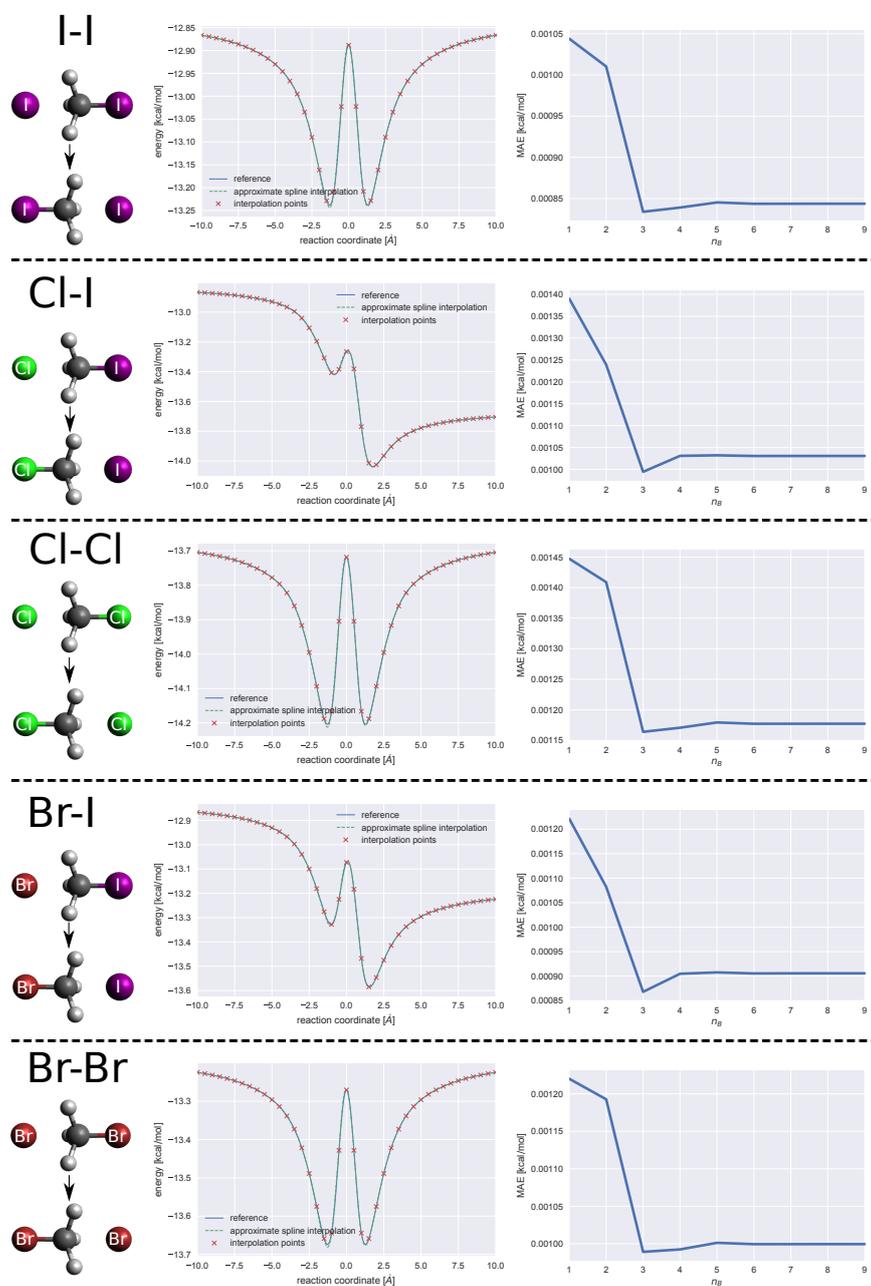


Figure 5.14: Approximate spline interpolation (middle) for $n_B = 3$ of the minimum energy paths of the reaction $X^- + H_3C-Y \rightarrow X-CH_3 + Y^-$ (left) for $(X,Y) = \{ (I-I), (Cl-I), (Cl-Cl), (Br-I), (Br-Br) \}$. The reference energy in kcal/mol (blue solid line) has been computed at the DSD-BLYP-D3(BJ)/def2-TZVP level of theory [139] and is shown along with the interpolation points (red crosses) and the approximate spline interpolation of the remaining points (green dashed line). The mean absolute error of approximate spline interpolation is given in kcal/mol in dependence of the number of interpolation points n_B (right).

Conclusions

The goal of this thesis has been to develop alternative efficient and performant tools for application in quantum chemistry. We have achieved this goal using three distinct methodologies: our set of invariant molecular many-body descriptors, our decomposition kernels and our approximate banded Toeplitz matrix inversion scheme.

In our first approach, we have developed representations of quantum mechanical systems which are invariant with respect to translation, rotation and atom indexing. This invariance has been achieved by summing over the permutations of two- and three-body combinations of atoms composing a physical system, potentially saving storage space in practice if the number of features is low for a given learning. For these descriptors, a linear ridge regression model has performed only slightly worse compared to the non-linear kernelized variant for stable small organic molecules and molecular dynamics data sets. This characteristic of our molecular representations potentially allows a practitioner to apply linear analysis tools for exploration of important features. In fact, our feature importance analysis has indicated that the interactions of hydrogen with all other atoms can be effectively modeled as a pairwise potential for stable small organic molecules. One limitation of these descriptors is the exponential growth of the number of features for increasing the highest exponent in their respective definition.

This problem has been circumvented by our second methodology, where we have proposed a similarity measure of physical systems encoding two- and three-body environments directly into the kernel, which can save even more storage space in practice, as the set of local many-body descriptors do not need to be stored explicitly. We have demonstrated that these decomposition kernels perform even better than our invariant descriptors for predicting the atomization energy for both stable small organic molecules and molecular dynamics simulation sets. The decomposition property of these kernels not only allows more efficient learning as demonstrated by the learning curves, it also allows us to model two- and three-body interaction potentials for which we

have shown to agree with chemical intuition. In future projects, we plan to use these kernels to investigate the possibility of transfer learning of interaction potentials across chemical compound space. We speculate on the need of interaction terms which model the interaction between specific two- and three-body atom combinations for this purpose.

In our third methodology, we have developed an approximate inversion algorithm for banded Toeplitz matrices which is designed to compute the inverses in a computationally efficient and robust way. Specifically, our matrix inversion scheme allows to uniquely approximate the inverse of a certain class of banded Toeplitz matrices. This class has been defined by a criterion which is independent of the matrix dimension and can therefore be readily verified. The proposed algorithms have been theoretically underpinned by a proof of regularity of the matrices of this class. We have demonstrated the applicability of our matrix inversion methods for quantum chemistry, where we have computed the polarizabilities including long-range electrodynamic response screening effects more efficiently. Finally, we have proposed approximate cubic spline interpolation using our matrix inversion scheme to accurately predict potential energy surfaces using a small number of training points, an approach which can potentially be useful for energy minimization procedures.

A limitation which prohibits the wider applicability of our matrix inversion method is the constant entries along the diagonals of the Toeplitz matrices. In the future, we plan to lift this assumption and investigate other extensions, e.g. the two- and three-dimensional case and sub-bandwidth time complexity, respectively.

Supplemental results

A.1 Image filtering experiments

In this experiment, we demonstrate the applicability of our approximate matrix inversion scheme of Chap. 5 outside the quantum chemistry domain on real-world image data. Specifically, we compute two-dimensional image convolutions where the filter kernel can be separated into a product of two one-dimensional functions using the kernel functions encountered in Sec 5.7. The two-dimensional convolutions are defined by

$$I_{\text{blurred}}(i, j) := \sum_{k_1=-n}^n \sum_{k_2=-n}^n I_{\text{in}}(i - k_1, j - k_2) \cdot K(h \cdot k_1) \cdot K(h \cdot k_2) \quad (\text{A.1})$$

where I_{in} is a single-channel input image, I_{blurred} is the resulting blurred filtered image, $K(\cdot)$ is a kernel with size $2n + 1$ and h is a stepsize parameter, respectively. The time complexity for the convolution step in Eq. (A.1) depends quadratically on the kernel size given by n . For increasing image resolutions, the parameter n has to increase accordingly to achieve the same filtering effect. To circumvent this problem, we apply our approximation scheme to perform the inverse operation

$$I_{\text{recon}}(i, j) := \sum_{k_1=-s}^s \sum_{k_2=-s}^s I_{\text{blurred}}(i - k_1, j - k_2) \cdot \bar{K}_s(k_1) \cdot \bar{K}_s(k_2) \quad (\text{A.2})$$

with inverse kernel sizes given by $s \ll n$. The functions $\bar{K}_s(\cdot)$ are obtained from Alg. 6 applied on the original kernels $K(\cdot)$ in Tab. 5.2 with kernel parameters listed in Tab. A.1. The results with different inverse kernel sizes s are shown in Fig. A.1.

Our method can be used to efficiently revert the filtering operation with one order of magnitude smaller kernel sizes $s \ll n$. The inverse Gaussian kernels have problems to fully reconstruct the original image which is in accordance

A. SUPPLEMENTAL RESULTS

Kernel	n	h	σ	λ
Gauss	301	10.0	10.0	0.1
Laplace	301	0.033	2.0	0.0
Matérn	301	0.033	0.1	0.0

Table A.1: Base kernels used in the experiment to compute the respective deconvolution operators.

with the results of the one-dimensional deconvolution experiments in Sec. 5.7. Interestingly, the inverse Laplace kernel has the smallest possible kernel size $s = 1$ for the most drastically blurred image which potentially can be used in combination with image compression algorithms.

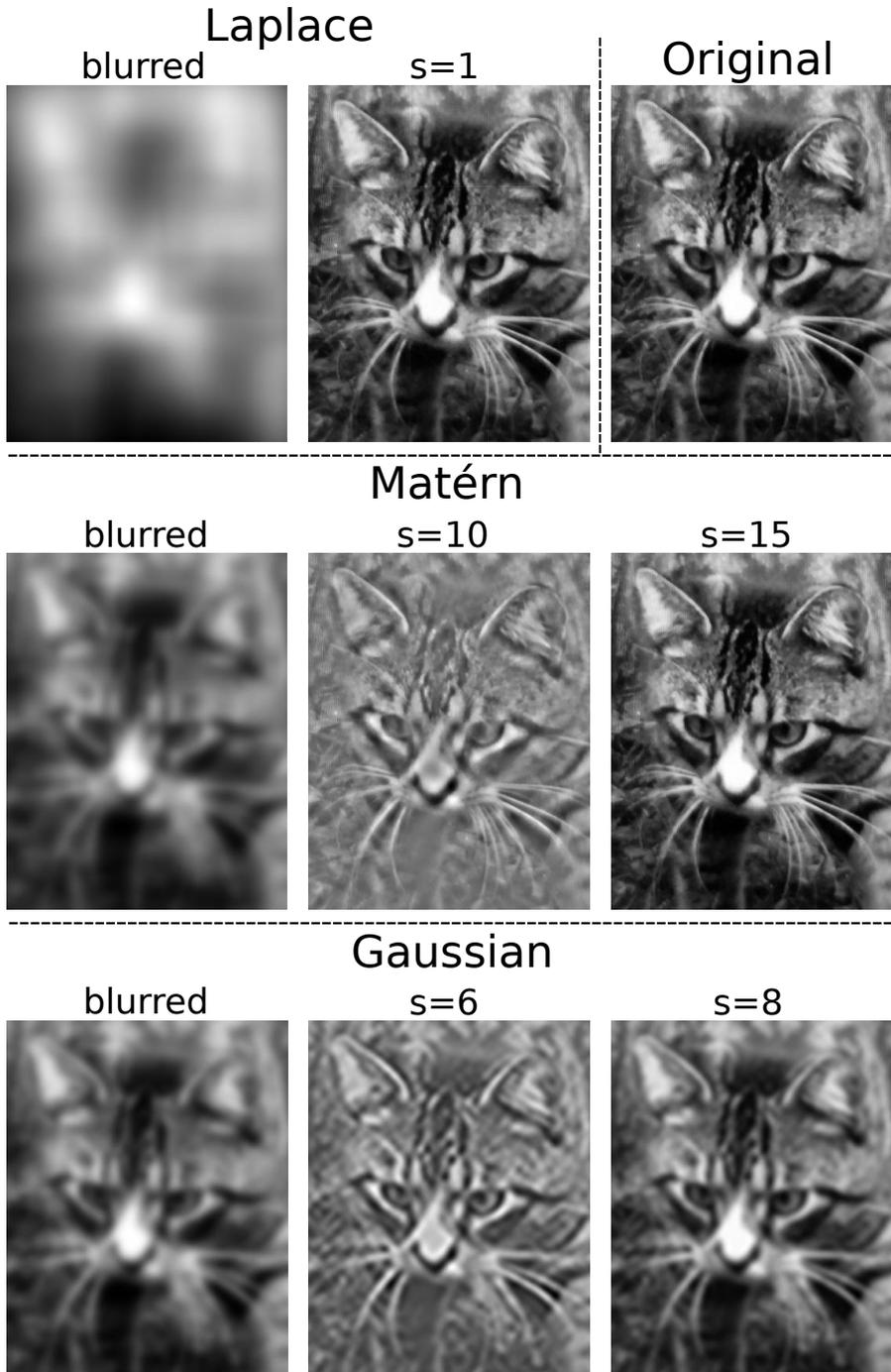


Figure A.1: Image convolutions (left) column from the original image (top right) according to Eq. (A.1) using the Laplace kernel (top row), Matérn kernel (middle row) and Gaussian kernel (bottom row) with the parameters listed in Tab. 5.2. The deconvolution is shown according to 103 Eq. (A.2) for kernel widths $s = 1$ for the Laplace kernel, $s = 10, 15$ for the Matérn kernel and $s = 6, 8$ for the Gaussian kernel, respectively.

References

- [1] W. Pronobis and K.-R. Müller. “Kernel Methods for Quantum Chemistry”. In: *Machine Learning for Quantum Simulations of Molecules and Materials*. Springer Nature, 2020, pp. 27–40.
- [2] K. T. Schütt, P.-J. Kindermans, H. E. S. Felix, S. Chmiela, A. Tkatchenko, and K.-R. Müller. “Schnet: A continuous-filter convolutional neural network for modeling quantum interactions”. In: *Advances in Neural Information Processing Systems*. 2017, pp. 991–1001.
- [3] F. A. Faber, L. Hutchison, B. Huang, J. Gilmer, S. S. Schoenholz, G. E. Dahl, O. Vinyals, S. Kearnes, P. F. Riley, and O. A. von Lilienfeld. “Prediction Errors of Molecular Machine Learning Models Lower than Hybrid DFT Error”. *Journal of Chemical Theory and Computation* 13 (11), pp. 5255–5264, 2017.
- [4] F. A. Faber, A. S. Christensen, B. Huang, and O. A. von Lilienfeld. “Alchemical and structural distribution based representation for universal quantum machine learning”. *The Journal of Chemical Physics* 148 (24), p. 241717, 2018.
- [5] S. Chmiela, A. Tkatchenko, H. E. Sauceda, I. Poltavsky, K. T. Schütt, and K.-R. Müller. “Machine Learning of Accurate Energy-conserving Molecular Force Fields”. *Science Advances* 3 (5), e1603015, 2017.
- [6] S. Chmiela, H. E. Sauceda, K.-R. Müller, and A. Tkatchenko. “Towards exact molecular dynamics simulations with machine-learned force fields”. *Nature Communications* 9 (1), p. 3887, 2018.
- [7] S. Chmiela, H. E. Sauceda, I. Poltavsky, K.-R. Müller, and A. Tkatchenko. “sGDML: Constructing accurate and data efficient molecular force fields using machine learning”. *Computer Physics Communications* 240, pp. 38–45, 2019.
- [8] W. Pronobis, A. Tkatchenko, and K.-R. Müller. “Many-Body Descriptors for Predicting Molecular Properties with Machine Learning: Analysis of Pairwise and Three-Body Interactions in Molecules”. *Journal of Chemical Theory and Computation* 14 (6), pp. 2991–3003, 2018.

- [9] B. Schölkopf, S. Mika, C. J. Burges, P. Knirsch, K.-R. Müller, G. Rätsch, and A. J. Smola. “Input space versus feature space in kernel-based methods”. *IEEE transactions on neural networks* 10 (5), pp. 1000–1017, 1999.
- [10] P. Indyk and R. Motwani. “Approximate nearest neighbors: towards removing the curse of dimensionality”. In: *Proceedings of the thirtieth annual ACM symposium on Theory of computing*. ACM. 1998, pp. 604–613.
- [11] J. H. Friedman. “On bias, variance, 0/1—loss, and the curse-of-dimensionality”. *Data mining and knowledge discovery* 1 (1), pp. 55–77, 1997.
- [12] J. Rust. “Using randomization to break the curse of dimensionality”. *Econometrica: Journal of the Econometric Society*, pp. 487–516, 1997.
- [13] B. E. Boser, I. M. Guyon, and V. N. Vapnik. “A training algorithm for optimal margin classifiers”. In: *Proceedings of the fifth annual workshop on Computational learning theory*. ACM. 1992, pp. 144–152.
- [14] K.-R. Müller, A. Smola, G. Rätsch, B. Schölkopf, J. Kohlmorgen, and V. Vapnik. “Using support vector machines for time series prediction”. *Advances in kernel methods—support vector learning*, pp. 243–254, 1999.
- [15] K.-R. Müller, S. Mika, G. Rätsch, K. Tsuda, and B. Schölkopf. “An introduction to kernel-based learning algorithms”. *IEEE Transactions on Neural Networks* 12 (2), pp. 181–201, 2001.
- [16] M. James and F. A. Russell. “XVI. Functions of positive and negative type, and their connection the theory of integral equations”. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character* 209 (441-458), pp. 415–446, 1909.
- [17] A. J. Smola, B. Schölkopf, and K.-R. Müller. “The connection between regularization operators and support vector kernels”. *Neural Networks* 11 (4), pp. 637–649, 1998.
- [18] A. Zien, G. Rätsch, S. Mika, B. Schölkopf, T. Lengauer, and K.-R. Müller. “Engineering support vector machine kernels that recognize translation initiation sites”. *Bioinformatics* 16 (9), pp. 799–807, 2000.
- [19] A. P. Bartók, R. Kondor, and G. Csányi. “On representing chemical environments”. *Phys. Rev. B* 87 (18), p. 184115, 2013.
- [20] K. Hansen, G. Montavon, F. Biegler, S. Fazli, M. Rupp, M. Scheffler, O. A. von Lilienfeld, A. Tkatchenko, and K.-R. Müller. “Assessment and Validation of Machine Learning Methods for Predicting Molecular Atomization Energies”. *Journal of Chemical Theory and Computation* 9 (8), pp. 3404–3419, 2013.
- [21] G. Montavon, K. Hansen, S. Fazli, M. Rupp, F. Biegler, A. Ziehe, A. Tkatchenko, A. V. Lilienfeld, and K.-R. Müller. “Learning invariant representations of molecules for atomization energy prediction”. In: *Advances in Neural Information Processing Systems*. 2012, pp. 440–448.

-
- [22] R. Ramakrishnan and O. A. von Lilienfeld. “Many molecular properties from one kernel in chemical space”. *CHIMIA International Journal for Chemistry* 69 (4), pp. 182–186, 2015.
- [23] G. Ferré, T. Haut, and K. Barros. “Learning molecular energies using localized graph kernels”. *The Journal of chemical physics* 146 (11), p. 114107, 2017.
- [24] D. Hu, Y. Xie, X. Li, L. Li, and Z. Lan. “Inclusion of Machine Learning Kernel Ridge Regression Potential Energy Surfaces in On-the-Fly Nonadiabatic Molecular Dynamics Simulation”. *The Journal of Physical Chemistry Letters* 9 (11), pp. 2725–2732, 2018.
- [25] C. K. Williams and C. E. Rasmussen. *Gaussian processes for machine learning*. Vol. 2. 3. MIT press Cambridge, MA, 2006.
- [26] B. Schölkopf, A. Smola, and K.-R. Müller. “Kernel principal component analysis”. In: *International conference on artificial neural networks*. Springer, 1997, pp. 583–588.
- [27] Z. Liu, D. Chen, and H. Bensmail. “Gene expression data classification with kernel principal component analysis”. *BioMed Research International* 2005 (2), pp. 155–159, 2005.
- [28] D. Antoniou and S. D. Schwartz. “Toward identification of the reaction coordinate directly from the transition state ensemble using the kernel PCA method”. *The Journal of Physical Chemistry B* 115 (10), pp. 2465–2469, 2011.
- [29] B. Schölkopf, A. Smola, and K. Müller. “Nonlinear Component Analysis as a Kernel Eigenvalue Problem”. *Neural Computation* 10 (5), pp. 1299–1319, 1998.
- [30] Y. M. Koyama, T. J. Kobayashi, S. Tomoda, and H. R. Ueda. “Perturbational formulation of principal component analysis in molecular dynamics simulation”. *Physical Review E* 78 (4), p. 046702, 2008.
- [31] X. Han. “Nonnegative principal component analysis for cancer molecular pattern discovery”. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)* 7 (3), pp. 537–549, 2010.
- [32] A. Varnek and I. I. Baskin. “Chemoinformatics as a theoretical chemistry discipline”. *Molecular Informatics* 30 (1), pp. 20–32, 2011.
- [33] X. Deng, X. Tian, and S. Chen. “Modified kernel principal component analysis based on local structure analysis and its application to nonlinear process fault diagnosis”. *Chemometrics and Intelligent Laboratory Systems* 127, pp. 195–209, 2013.
- [34] M. L. Braun, J. M. Buhmann, and K.-R. Müller. “On Relevant Dimensions in Kernel Feature Spaces”. *Journal of Machine Learning Research* 9 (Aug), pp. 1875–1908, 2008.

- [35] M. Moravčík, M. Schmid, N. Burch, V. Lisý, D. Morrill, N. Bard, T. Davis, K. Waugh, M. Johanson, and M. Bowling. “DeepStack: Expert-level artificial intelligence in heads-up no-limit poker”. *Science* 356 (6337), pp. 508–513, 2017.
- [36] S. Ontanon, G. Synnaeve, A. Uriarte, F. Richoux, D. Churchill, and M. Preuss. “A Survey of Real-Time Strategy Game AI Research and Competition in StarCraft”. *IEEE Transactions on Computational Intelligence and AI in Games* 5 (4), pp. 293–311, 2013.
- [37] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton, Y. Chen, T. Lillicrap, F. Hui, L. Sifre, G. van den Driessche, T. Graepel, and D. Hassabis. “Mastering the game of Go without human knowledge”. *Nature* 550 (7676), pp. 354–359, 2017.
- [38] G. Hinton, L. Deng, D. Yu, G. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury. “Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups”. *IEEE Signal Processing Magazine* 29 (6), pp. 82–97, 2012.
- [39] K. He, X. Zhang, S. Ren, and J. Sun. “Deep Residual Learning for Image Recognition”. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2016.
- [40] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. W. Senior, and K. Kavukcuoglu. “WaveNet: A Generative Model for Raw Audio”. *CoRR* abs/1609.03499, 2016.
- [41] M. Rupp, A. Tkatchenko, K.-R. Müller, and O. A. von Lilienfeld. “Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning”. *Phys. Rev. Lett.* 108, p. 058301, 2012.
- [42] F. Noé and C. Clementi. “Kinetic Distance and Kinetic Maps from Molecular Dynamics Simulation”. *Journal of Chemical Theory and Computation* 11 (10), pp. 5002–5011, 2015.
- [43] M. Gastegger, J. Behler, and P. Marquetand. “Machine learning molecular dynamics for the simulation of infrared spectra”. *Chemical Science* 8 (10), pp. 6924–6935, 2017.
- [44] A. P. Bartók, S. De, C. Poelking, N. Bernstein, J. R. Kermode, G. Csányi, and M. Ceriotti. “Machine learning unifies the modeling of materials and molecules”. *Science Advances* 3 (12: e1701816), 2017.
- [45] J. C. Snyder, M. Rupp, K. Hansen, K.-R. Müller, and K. Burke. “Finding Density Functionals with Machine Learning”. *Phys. Rev. Lett.* 108, p. 253002, 2012.
- [46] F. Brockherde, L. Vogt, L. Li, M. E. Tuckerman, K. Burke, and K.-R. Müller. “Bypassing the Kohn-Sham equations with machine learning”. *Nature Communications* 8 (1), p. 872, 2017.

-
- [47] C. R. Collins, G. J. Gordon, O. A. von Lilienfeld, and D. J. Yaron. “Constant size descriptors for accurate machine learning models of molecular properties”. *The Journal of Chemical Physics* 148 (24), p. 241718, 2018.
- [48] B. Huang and O. A. von Lilienfeld. “Communication: Understanding molecular representations in machine learning; The role of uniqueness and target similarity”. *The Journal of Chemical Physics* 145 (16), p. 161102, 2016.
- [49] H. Huo and M. Rupp. “Unified representation for machine learning of molecules and crystals”. *arXiv preprint arXiv:1704.06439*, pp. 13754–13769, 2017.
- [50] J. Behler and M. Parrinello. “Generalized Neural-Network Representation of High-Dimensional Potential-Energy Surfaces”. *Phys. Rev. Lett.* 98, p. 146401, 2007.
- [51] A. P. Bartók, M. C. Payne, R. Kondor, and G. Csányi. “Gaussian Approximation Potentials: The Accuracy of Quantum Mechanics, without the Electrons”. *Phys. Rev. Lett.* 104, p. 136403, 2010.
- [52] S. Chmiela. “Towards exact molecular dynamics simulations with invariant machine-learned models”. dissertation. Technische Universität Berlin, 2019.
- [53] K. Hansen, F. Biegler, R. Ramakrishnan, W. Pronobis, O. A. von Lilienfeld, K.-R. Müller, and A. Tkatchenko. “Machine Learning Predictions of Molecular Properties: Accurate Many-Body Potentials and Nonlocality in Chemical Space”. *The Journal of Physical Chemistry Letters* 6 (12), pp. 2326–2331, 2015.
- [54] W. Pronobis, K. T. Schütt, A. Tkatchenko, and K.-R. Müller. “Capturing intensive and extensive DFT/TDDFT molecular properties with machine learning”. *The European Physical Journal B* 91 (8), p. 178, 2018.
- [55] A. V. Shapeev. “Moment Tensor Potentials: A Class of Systematically Improvable Interatomic Potentials”. *Multiscale Modeling & Simulation* 14 (3), pp. 1153–1173, 2016.
- [56] E. V. Podryabinkin and A. V. Shapeev. “Active learning of linearly parametrized interatomic potentials”. *Computational Materials Science* 140, pp. 171–180, 2017.
- [57] C. R. Collins, G. J. Gordon, O. Anatole von Lilienfeld, and D. J. Yaron. “Constant Size Molecular Descriptors For Use With Machine Learning”. *ArXiv e-prints*, 2017.
- [58] L. C. Blum and J.-L. Reymond. “970 Million Druglike Small Molecules for Virtual Screening in the Chemical Universe Database GDB-13”. *Journal of the American Chemical Society* 131 (25), pp. 8732–8733, 2009.

- [59] J. P. Perdew, M. Ernzerhof, and K. Burke. “Rationale for mixing exact exchange with density functional approximations”. *The Journal of Chemical Physics* 105 (22), pp. 9982–9985, 1996.
- [60] M. Ernzerhof and G. E. Scuseria. “Assessment of the Perdew–Burke–Ernzerhof exchange–correlation functional”. *The Journal of Chemical Physics* 110 (11), pp. 5029–5036, 1999.
- [61] J. Ridley and M. Zerner. “An intermediate neglect of differential overlap technique for spectroscopy: Pyrrole and the azines”. *Theoretica chimica acta* 32 (2), pp. 111–134, 1973.
- [62] A. D. Bacon and M. C. Zerner. “An intermediate neglect of differential overlap theory for transition metal complexes: Fe, Co and Cu chlorides”. *Theoretica chimica acta* 53 (1), pp. 21–54, 1979.
- [63] M. C. Zerner. “Semiempirical molecular orbital methods”. *Reviews in computational chemistry* 2, pp. 313–365, 1991.
- [64] A. Tkatchenko, R. A. DiStasio, R. Car, and M. Scheffler. “Accurate and Efficient Method for Many-Body van der Waals Interactions”. *Phys. Rev. Lett.* 108, p. 236402, 2012.
- [65] L. Hedin. “New Method for Calculating the One-Particle Green’s Function with Application to the Electron-Gas Problem”. *Phys. Rev.* 139, A796–A823, 1965.
- [66] V. Blum, R. Gehrke, F. Hanke, P. Havu, V. Havu, X. Ren, K. Reuter, and M. Scheffler. “Ab initio molecular simulations with numeric atom-centered orbitals”. *Computer Physics Communications* 180 (11), pp. 2175–2196, 2009.
- [67] F. Neese. *ORCA—an Ab Initio, Density Functional and Semiempirical Program Package*. Vol. 2. 2012.
- [68] L. Ruddigkeit, R. van Deursen, L. C. Blum, and J.-L. Reymond. “Enumeration of 166 Billion Organic Small Molecules in the Chemical Universe Database GDB-17”. *Journal of Chemical Information and Modeling* 52 (11), pp. 2864–2875, 2012.
- [69] R. Ramakrishnan, P. O. Dral, M. Rupp, and O. A. von Lilienfeld. “Quantum chemistry structures and properties of 134 kilo molecules”. *Scientific Data* 1, 2014.
- [70] J. Vert, K. Tsuda, and B. Schölkopf. “A Primer on Kernel Methods”. In: *Kernel Methods in Computational Biology*. Cambridge, MA, USA: MIT Press, 2004, pp. 35–70.
- [71] V. Vovk. “Kernel Ridge Regression”. In: *Empirical Inference*. Springer Berlin Heidelberg, 2013, pp. 105–116.
- [72] A. E. Hoerl and R. W. Kennard. “Ridge Regression: Applications to Nonorthogonal Problems”. *Technometrics* 12 (1), pp. 69–82, 1970.
- [73] “k-Nearest Neighbor Algorithm”. In: *Discovering Knowledge in Data*. John Wiley & Sons, Inc., 2014, pp. 149–164.

-
- [74] R. Kohavi et al. “A study of cross-validation and bootstrap for accuracy estimation and model selection”. 14 (2), pp. 1137–1145, 1995.
- [75] B. M. Hill. “Bayesian Inference in Statistical Analysis”. *Technometrics* 16 (3), pp. 478–479, 1974.
- [76] S. Wold, M. Sjöström, and L. Eriksson. “PLS-regression: a basic tool of chemometrics”. *Chemometrics and Intelligent Laboratory Systems* 58 (2), pp. 109–130, 2001.
- [77] A. Aitken. “On least squares and linear combination of observations”. 55, pp. 42–48, 1934.
- [78] R. Tibshirani. “Regression Shrinkage and Selection Via the Lasso”. *Journal of the Royal Statistical Society, Series B* 58, pp. 267–288, 1994.
- [79] K. P. Bennett and O. L. Mangasarian. “Robust linear programming discrimination of two linearly inseparable sets”. *Optimization Methods and Software* 1 (1), pp. 23–34, 1992.
- [80] J. H. Friedman. “Greedy function approximation: a gradient boosting machine”. *The Annals of Statistics* 29 (5), pp. 1189–1232, 2001.
- [81] S. Sonnenburg, A. Zien, P. Philips, and G. Rätsch. “POIMs: positional oligomer importance matrices—understanding support vector machine-based signal detectors”. *Bioinformatics* 24 (13), pp. i6–i14, 2008.
- [82] D. Baehrens, T. Schroeter, S. Harmeling, M. Kawanabe, K. Hansen, and K.-R. Müller. “How to explain individual classification decisions”. *Journal of Machine Learning Research* 11 (Jun), pp. 1803–1831, 2010.
- [83] A. Zien, N. Krämer, S. Sonnenburg, and G. Rätsch. “The Feature Importance Ranking Measure”. In: *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2009, Bled, Slovenia, September 7-11, 2009, Proceedings, Part II*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 694–709.
- [84] G. Montavon, S. Lapuschkin, A. Binder, W. Samek, and K.-R. Müller. “Explaining nonlinear classification decisions with deep taylor decomposition”. *Pattern Recognition* 65, pp. 211–222, 2017.
- [85] G. Montavon, W. Samek, and K.-R. Müller. “Methods for interpreting and understanding deep neural networks”. *Digital Signal Processing* 73, pp. 1–15, 2018.
- [86] M. D. Zeiler and R. Fergus. “Visualizing and Understanding Convolutional Networks”. In: *Computer Vision –ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I*. Ed. by D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars. Cham: Springer International Publishing, 2014, pp. 818–833.
- [87] K. Simonyan, A. Vedaldi, and A. Zisserman. “Deep inside convolutional networks: Visualising image classification models and saliency maps”. *arXiv preprint arXiv:1312.6034*, 2013.

- [88] K. Lenc and A. Vedaldi. *Understanding image representations by measuring their equivariance and equivalence*. 2014.
- [89] S. Haufe, F. Meinecke, K. Görgen, S. Dähne, J.-D. Haynes, B. Blankertz, and F. Bießmann. “On the interpretation of weight vectors of linear models in multivariate neuroimaging”. *Neuroimage* 87, pp. 96–110, 2014.
- [90] G. Montavon, M. Rupp, V. Gobre, A. Vazquez-Mayagoitia, K. Hansen, A. Tkatchenko, K.-R. Müller, and O. A. von Lilienfeld. “Machine learning of molecular electronic properties in chemical compound space”. *New Journal of Physics* 15 (9), p. 095003, 2013.
- [91] K. T. Schütt, F. Arbabzadah, S. Chmiela, K. R. Müller, and A. Tkatchenko. “Quantum-chemical insights from deep tensor neural networks”. *Nature communications* 8, p. 13890, 2017.
- [92] A. Mardt, L. Pasquali, H. Wu, and F. Noé. “VAMPnets for deep learning of molecular kinetics”. *Nature communications* 9 (1), p. 5, 2018.
- [93] M. Grätzel. “Photoelectrochemical cells”. *Nature* 414 (6861), pp. 338–344, 2001.
- [94] M. Gross, D. C. Müller, H.-G. Nothofer, U. Scherf, D. Neher, C. Bräuchle, and K. Meerholz. *Nature* 405, pp. 661–665, 2000.
- [95] E. Runge and E. K. U. Gross. “Density-Functional Theory for Time-Dependent Systems”. *Phys. Rev. Lett.* 52, pp. 997–1000, 1984.
- [96] K. T. Schütt, P.-J. Kindermans, H. E. Saucedo Felix, S. Chmiela, A. Tkatchenko, and K.-R. Müller. “SchNet: A continuous-filter convolutional neural network for modeling quantum interactions”. In: *Advances in Neural Information Processing Systems 30*. Ed. by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett. Curran Associates, Inc., 2017, pp. 992–1002.
- [97] K. T. Schütt, H. E. Saucedo, P.-J. Kindermans, A. Tkatchenko, and K.-R. Müller. “SchNet—A deep learning architecture for molecules and materials”. *The Journal of Chemical Physics* 148 (24), p. 241722, 2018.
- [98] F. Furche and R. Ahlrichs. “Adiabatic time-dependent density functional methods for excited state properties”. *The Journal of Chemical Physics* 117 (16), pp. 7433–7447, 2002.
- [99] C. Adamo and V. Barone. “Toward reliable density functional methods without adjustable parameters: The PBE0 model”. *The Journal of Chemical Physics* 110 (13), pp. 6158–6170, 1999.
- [100] F. Weigend and R. Ahlrichs. “Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for H to Rn: Design and assessment of accuracy”. *Physical Chemistry Chemical Physics* 7 (18), p. 3297, 2005.
- [101] C. Cortes and V. Vapnik. “Support-vector networks”. *Machine learning* 20 (3), pp. 273–297, 1995.

-
- [102] V. Vapnik, S. E. Golowich, and A. J. Smola. “Support vector method for function approximation, regression estimation and signal processing”. In: *Advances in neural information processing systems*. 1997, pp. 281–287.
- [103] B. Schölkopf and A. J. Smola. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2001.
- [104] B. Schölkopf, A. J. Smola, F. Bach, et al. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2002.
- [105] S. De, A. P. Bartók, G. Csányi, and M. Ceriotti. “Comparing molecules and solids across structural and alchemical space”. *Physical Chemistry Chemical Physics* 18 (20), pp. 13754–13769, 2016.
- [106] T. Olsen and K. S. Thygesen. “Accurate ground-state energies of solids and molecules from time-dependent density-functional theory”. *Physical Review Letters* 112 (20), p. 203001, 2014.
- [107] Z. Cinkir. “A fast elementary algorithm for computing the determinant of Toeplitz matrices”. *Journal of Computational and Applied Mathematics* 255, pp. 353–361, 2014.
- [108] H.-C. Li. “On calculating the determinants of Toeplitz matrices”. *Journal of Applied Mathematics and Bioinformatics* 1 (1), p. 55, 2011.
- [109] J. F. Monahan. *Numerical methods of statistics*. Cambridge University Press, 2011.
- [110] N. Levinson. “The Wiener (Root Mean Square) Error Criterion in Filter Design and Prediction”. *Journal of Mathematics and Physics* 25 (1-4), pp. 261–278, 1946.
- [111] J. Durbin. “The fitting of time-series models”. *Revue de l’Institut International de Statistique* 28 (3), pp. 233–244, 1960.
- [112] P. Delsarte and Y. Genin. “The split Levinson algorithm”. *IEEE transactions on acoustics, speech, and signal processing* 34 (3), pp. 470–478, 1986.
- [113] F. de Hoog. “A new algorithm for solving Toeplitz systems of equations”. *Linear Algebra and its Applications* 88, pp. 123–138, 1987.
- [114] W. F. Trench. “An algorithm for the inversion of finite Toeplitz matrices”. *Journal of the Society for Industrial and Applied Mathematics* 12 (3), pp. 515–522, 1964.
- [115] E. H. Bareiss. “Numerical solution of linear equations with Toeplitz and vector Toeplitz matrices”. *Numerische Mathematik* 13 (5), pp. 404–424, 1969.
- [116] A. Bojanczyk, R. Brent, and F. De Hoog. “QR factorization of Toeplitz matrices”. *Numerische Mathematik* 49 (1), pp. 81–94, 1986.

- [117] A. W. Bojanczyk, R. P. Brent, F. R. De Hoog, and D. R. Sweet. “On the stability of the Bareiss and related Toeplitz factorization algorithms”. *SIAM Journal on Matrix Analysis and Applications* 16 (1), pp. 40–57, 1995.
- [118] R. M. Gray. “Toeplitz and Circulant Matrices: A Review”. *Foundations and Trends® in Communications and Information Theory* 2 (3), pp. 155–239, 2006.
- [119] W. W. Barrett and P. J. Feinsilver. “Inverses of banded matrices”. *Linear Algebra and its Applications* 41, pp. 111–130, 1981.
- [120] P. Amodio and L. Brugnano. “The conditioning of Toeplitz band matrices”. *Mathematical and Computer Modelling* 23 (10), pp. 29–42, 1996.
- [121] A. Kavcic and J. M. Moura. “Matrices with banded inverses: Inversion algorithms and factorization of Gauss-Markov processes”. *IEEE transactions on Information Theory* 46 (4), pp. 1495–1509, 2000.
- [122] D. Meek. “The inverses of Toeplitz band matrices”. *Linear Algebra and its Applications* 49, pp. 117–129, 1983.
- [123] W. F. Trench. “Inversion of Toeplitz Band Matrices”. *Mathematics of Computation* 28 (128), pp. 1089–1095, 1974.
- [124] A. Jain. “Fast inversion of banded Toeplitz matrices by circular decompositions”. *IEEE Transactions on Acoustics, Speech, and Signal Processing* 26 (2), pp. 121–126, 1978.
- [125] W. F. Trench. “Explicit Inversion Formulas for Toeplitz Band Matrices”. *SIAM Journal on Algebraic Discrete Methods* 6 (4), pp. 546–554, 1985.
- [126] M. Elouafi. “Explicit inversion of Band Toeplitz matrices by discrete Fourier transform”. *Linear and Multilinear Algebra* 66 (9), pp. 1767–1782, 2018.
- [127] T. N. E. Greville and W. F. Trench. “Band matrices with Toeplitz inverses”. *Linear Algebra and its Applications* 27, pp. 199–209, 1979.
- [128] M. Hanke and J. G. Nagy. “Toeplitz approximate inverse preconditioner for banded Toeplitz matrices”. *Numerical Algorithms* 7 (2), pp. 183–199, 1994.
- [129] B. Romain, P. Antonio, and G. Rainer. “An L-Banded Approximation to the Inverse of Symmetric Toeplitz Matrices”. *Stochastics and Quality Control* 25 (1), pp. 13–30, 2010.
- [130] A. Edelman and H. Murakami. “Polynomial roots from companion matrix eigenvalues”. *Mathematics of Computation* 64 (210), pp. 763–776, 1995.
- [131] N. J. Higham. *Accuracy and stability of numerical algorithms*. Vol. 80. Siam, 2002.
- [132] A. Böttcher and S. Grudsky. *Spectral Properties of Banded Toeplitz Matrices*. Society for Industrial and Applied Mathematics, 2005.

-
- [133] R. A. DiStasio Jr, V. V. Gobre, and A. Tkatchenko. “Many-body van der Waals interactions in molecules and condensed matter”. *Journal of Physics: Condensed Matter* 26 (21), p. 213202, 2014.
- [134] J. Ischtwan and M. A. Collins. “Molecular potential energy surfaces by interpolation”. *The Journal of chemical physics* 100 (11), pp. 8080–8088, 1994.
- [135] K. C. Thompson, M. J. Jordan, and M. A. Collins. “Polyatomic molecular potential energy surfaces by interpolation in local internal coordinates”. *The Journal of chemical physics* 108 (20), pp. 8302–8316, 1998.
- [136] R. P. Bettens and M. A. Collins. “Learning to interpolate molecular potential energy surfaces with confidence: A Bayesian approach”. *The Journal of chemical physics* 111 (3), pp. 816–826, 1999.
- [137] C. Crespos, M. A. Collins, E. Pijper, and G.-J. Kroes. “Application of the modified Shepard interpolation method to the determination of the potential energy surface for a molecule–surface reaction: $\text{H}_2 + \text{Pt}(111)$ ”. *The Journal of chemical physics* 120 (5), pp. 2392–2404, 2004.
- [138] M. J. Jordan, K. C. Thompson, and M. A. Collins. “Convergence of molecular potential energy surfaces by interpolation: Application to the $\text{OH} + \text{H}_2 \rightarrow \text{H}_2\text{O} + \text{H}$ reaction”. *The Journal of chemical physics* 102 (14), pp. 5647–5657, 1995.
- [139] O. T. Unke and M. Meuwly. “PhysNet: a neural network for predicting energies, forces, dipole moments, and partial charges”. *Journal of chemical theory and computation* 15 (6), pp. 3678–3693, 2019.
- [140] G. Wolberg and I. Alfy. “An energy-minimization framework for monotonic cubic spline interpolation”. *Journal of Computational and Applied Mathematics* 143 (2), pp. 145–188, 2002.
- [141] T. I. Vassilev. “Fair interpolation and approximation of B-splines by energy minimization and points insertion”. *Computer-Aided Design* 28 (9), pp. 753–760, 1996.
- [142] O. L. Mangasarian and L. L. Schumaker. “Best summation formulae and discrete splines”. *SIAM Journal on Numerical Analysis* 10 (3), pp. 448–459, 1973.

Vorveröffentlichungen und Eigenanteile

Publikation:

W. Pronobis, A. Tkatchenko, and K.-R. Müller. "Many-Body Descriptors for Predicting Molecular Properties with Machine Learning: Analysis of Pairwise and Three-Body Interactions in Molecules". *Journal of Chemical Theory and Computation* 14 (6), pp. 2991–3003, 2018

Die Hauptbeiträge stammen von mir. Ich habe die Deskriptoren entwickelt, die in diesem Artikel eingeführt wurden. Die Methoden wurden von mir implementiert, evaluiert und ausgewertet. Alle Bilder in diesem Artikel stammen von mir. Alle Autoren haben die Ergebnisse diskutiert und zum finalen Text beigetragen.

Publikation:

W. Pronobis, K. T. Schütt, A. Tkatchenko, and K.-R. Müller. "Capturing intensive and extensive DFT/TDDFT molecular properties with machine learning". *The European Physical Journal B* 91 (8), p. 178, 2018

Ich habe die Deskriptoren entwickelt, die in diesen Artikel auftauchen und die entsprechenden Experimente durchgeführt. Alle Bilder in diesem Artikel stammen von mir. Die Hauptbeiträge stammen zu gleichen Teilen von K. T. Schütt, der die Experimente basierend auf seinem neuronalen Netz SchNet durchgeführt hat, und mir. Alle Autoren haben die Ergebnisse diskutiert und zum finalen Text beigetragen.

Publikation:

W. Pronobis, D. Panknin, J. Kirschnick, V. Srinivasan, W. Samek, V. Markl, M. Kaul, K.-R. Müller, and S. Nakajima. "Sharing hash codes for multiple purposes". *Japanese Journal of Statistics and Data Science* 1 (1), pp. 215–246, 2018

Die Hashing Methoden in diesem Artikel wurden zu gleich Teilen von D. Panknin und mir entwickelt. Die entwickelten Hashing Verfahren wurden von

mir implementiert und anhand von den Datensätzen MovieLens10M und Netflix evaluiert und ausgewertet. Alle Bilder basierend auf diesen Datensätzen und den theoretischen Gütefunktionen stammen von mir. Alle Autoren haben die Ergebnisse diskutiert und zum finalen Text beigetragen.

Buchkapitel:

W. Pronobis, and K.-R. Müller. Kernel Methods for Quantum Chemistry. In: *Machine Learning for Quantum Simulations of Molecules and Materials*. Springer Nature, 2020, pp. 27–40

Die Hauptbeiträge stammen von mir. Alle Bilder in diesem Artikel stammen von mir. S. Chmiela und K.-R. Müller haben das Kapitel diskutiert und zum finalen Text beigetragen.

Publikation:

K. Hansen, F. Biegler, R. Ramakrishnan, W. Pronobis, O. A. von Lilienfeld, K.-R. Müller, and A. Tkatchenko. "Machine Learning Predictions of Molecular Properties: Accurate Many-Body Potentials and Nonlocality in Chemical Space". *The journal of physical chemistry letters* 6 (12), pp. 2326–2331, 2015

Ich habe die Deskriptoren, die in diesem Artikel eingeführt wurden, implementiert und evaluiert.

Publikation:

H. Marienwald, W. Pronobis, K.-R. Müller, and S. Nakajima. "Tight Bound of Incremental Cover Trees for Dynamic Diversification". *arXiv preprint arXiv:1806.06126*, 2018

Ich habe die Ergebnisse der Nearest-Neighbor-Methode, die in diesem Artikel eingeführt wurde, diskutiert.