MASTERARBEIT / MASTER'S THESIS

Titel der Masterarbeit / Title of the Master's Thesis

„Bayesian Optimization for the investigation of non collinear magnetism within density functional theory“

verfasst von / submitted by

Jakob Baumsteiger BSc

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of

Master of Science (MSc)

Wien, 2023 / Vienna, 2023

Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet:	UA 066 876
Studienrichtung lt. Studienblatt / degree programme as it appears on the student record sheet:	Physics UG2002
Betreut von / Supervisor:	Univ.-Prof. Dr.techn. Cesare Franchini

Contents

1 Introduction

Magnetism as a phenomenon is known to humans since at least 600 BC [1,2]. At these times ancient Greeks like Thales wondered about the peculiar behavior of certain stones, the lodestones. They tried to build theories for how these materials could attract or repel each other without physical contact. Not long after this, magnetism was also known in China [2]. There, further investigations and curiosities eventually lead to the invention of the first compass. In the following centuries, most laws of magnetism were discovered by experiments and research of numerous philosophers and scientists. However, a coherent explanation for the origin of magnetism in materials was first possible with the invention of quantum mechanics in the beginning of the 20th century. In 1911 and 1921 respectively, Niels Bohr [3] and Hendrika Johanna van Leeuwen [4] independently from each other proved that magnetism is inherently quantum mechanical (Bohr-Van Leeuwen theorem). Nowadays, quantum magnetism, the study of magnetic materials at the quantum level, is a thriving field of research that does not only provide new insights into fundamental interactions between quantum particles, but also offers interesting applications, for example in the fields of data storage, sensors and communication technologies. In today's research on magnetic materials it is often necessary to build models that describe the total energy as a function of the configurations of the electron spins and magnetic moments. This information is often crucial for testing new theories on the interactions that take place within the material or to extract important information like the magnetic anisotropy energy. Although the quantum mechanical framework for describing magnetic systems is known since the beginning of last century, actual calculations remained unfeasible due to the large number of freedom of real materials. However, the development of Density Functional Theory (DFT) [5,6] and supercomputers produced a paradigm shift that revolutionized the field [7]. Nevertheless, magnetic DFT calculations still require many computational resources. This is not only a problem considering the electricity consumption and the financial costs of the computers but also significantly limits the possibilities of today's research on quantum magnetism.

Obviously, the described problem makes it necessary to use the available computational resources as efficient as possible. Since it is hard to significantly reduce the computational costs of individual DFT calculations, the more promising approach is to find a way to build reliable models, that capture interesting magnetic properties of the material, based on only a small number of DFT calculations. Fortunately, there is an algorithm that was designed to do just that: Bayesian optimization [8,9]. Bayesian optimization is an active machine learning approach that calculates model functions using the mathematical concept of Gaussian processes and automatically acquires more data to efficiently improve the models. Bayesian optimization has already successfully been used in combination with DFT calculations [10–12]. However, until now Bayesian optimization has not been used for the investigation of magnetic properties of specific materials based on DFT calculations.

The goal of this thesis is to study weather Bayesian optimization can reduce the computational costs for the investigation of complex magnetic materials while preserving a high degree of accuracy and predictive power. In order to do this we combine the python package Bayesian Optimization Structure Search (BOSS) [13] with the Vienna Ab initio Simulation Package (VASP) [14–16], which we use for DFT calculations. The performance of the approach is tested by reproducing existing results on three known complex noncollinear magnets, Ba_2NaOsO_6, UO_2 and Sr_2IrO_4. Furthermore it will be tested if Bayesian Optimization can be used to simultaneously investigate multiple magnetic degrees of freedom. Such degrees of freedom can for example be the direction of the magnetic moment vector of different atoms. Thereby more and more complex spin configurations can be investigated.

In the sections 2 and 3 of this thesis we discuss the theoretical background of the investigation of magnetic materials with DFT and the theoretical background of Bayesian Optimization, respectively. In section 4 we explain basic concepts of magnetism in materials and introduce the three benchmark materials Ba_2NaOsO_6, UO_2 and Sr_2IrO_4. In section 5 the results and the performance of the approach are discussed. A detailed summary on important results also be found in the appendix in section A. Lastly we discuss the overall results of the thesis in section 6.

2 Quantum materials modeling

For the modeling of quantum materials, *ab initio* methods are typically used. Ab initio methods do not depend on phenomenological parameters such as measurement results. In the context of atomistic modeling they are solely based on fundamental quantum mechanical laws like the Schrödinger equation. There are several ab initio methods, that are frequently used for quantum materials modeling (*e.g.*, Hartree-Fock [17, 18], dynamical mean-field theory [19], coupled cluster methods [20] and DFT). Every method has its advantages and disadvantages. In this work we use the DFT method. All calculations were performed using the Vienna Ab initio Simulation Package (VASP) [14–16].

2.1 Quantum many-body Hamiltonian

Magnetic materials such as Ba_2NaOsO_6, UO_2 and Sr_2IrO_4 are quantum many-body systems consisting of large numbers of nuclei and electrons. A Hamiltonian that describes such a system can be written as

$$\hat{H} = T_e + T_n + U_{ee} + U_{nn} + U_{en}. \tag{1}$$

when relativistic effects are neglected. The individual terms have the following specific form and function:

$$\begin{aligned}
T_e &= -\frac{\hbar^2}{2m}\sum_{i_1}^{N}\nabla_i^2 && \text{Kinetic energy of electrons} \\
T_n &= -\sum_{k=1}^{M}\frac{\hbar^2}{2M_k}\nabla_k^2 && \text{Kinetic energy of nuclei} \\
U_{ee} &= \frac{1}{4\pi\epsilon_0}\frac{1}{2}\sum_{i,j=1}^{N}\frac{e^2}{|\boldsymbol{r}_i - \boldsymbol{r}_j|} && \text{Coulomb electron-electron interaction} \\
U_{nn} &= \frac{1}{4\pi\epsilon_0}\frac{1}{2}\sum_{\substack{k,l=1 \\ k\neq l}}^{M}\frac{Z_k Z_l e^2}{|\boldsymbol{R}_k - \boldsymbol{R}_l|} && \text{Coulomb nucleus-nucleus interaction} \\
U_{en} &= -\frac{1}{4\pi\epsilon_0}\sum_{k}^{M}\sum_{i}^{N}\frac{Z_k e^2}{|\boldsymbol{r}_i - \boldsymbol{R}_k|} && \text{Coulomb electron-nucleus interaction}
\end{aligned} \tag{2}$$

This Hamiltonian describes a system containing N electrons at the positions $\boldsymbol{r}_i$ with mass m and M nuclei at the positions $\boldsymbol{R}_k$. Here, M_k and Z_k are the mass and the atomic number of nucleus k, respectively.

The operators T_e and T_n are the kinetic energy operators of the electrons and nuclei, respectively. U_{ee}, U_{nn} and U_{en} are the Potential operators for the Coulomb interaction between electrons and electrons, nuclei and nuclei and electrons and nuclei, respectively.

It is important to note that finding the ground state of the quantum many-body Hamiltonian is not trivial. The interdependent interactions between the involved particles make the described systems very complex. In order to model such material systems it is thus necessary to make a series of approximations.

2.2 Born-Oppenheimer approximation

Electrons have much less mass than nuclei and therefore rearrange themselves on a timescale that is faster than the timescale of nuclear motion [21]. This difference is so remarkably large that the nuclei can be viewed as almost static compared to the electrons. In 1927, J. Robert Oppenheimer and Max Born argued that this instance can be used to simplify the molecular Schrödinger equation [22]. This argument is known as the *Born-Oppenheimer approximation.* Following the argument, Born and Oppenheimer argued that it is a reasonable approximation to write the time-independent wavefunction $\Psi(\boldsymbol{r}, \boldsymbol{R})$ as the product of an electronic wavefunction $\Psi_{el}(\boldsymbol{r}, \boldsymbol{R})$ and a nuclear wavefunction $\Psi_N(\boldsymbol{R})$:

$$\Psi(\boldsymbol{r}, \boldsymbol{R}) = \Psi_{el}(\boldsymbol{r}, \boldsymbol{R}) \cdot \Psi_N(\boldsymbol{R}) \tag{3}$$

With this approach it is possible to solve the electronic and nuclear problems separately. This simplifies the quantum many-body problem, since the Hilbert space of the electronic wavefunctions has a considerably lower number of dimensions than the original Hilbert space. The electronic wavefunction is determined by the electronic Schrödinger equation

$$\hat{H}_{el}\Psi_{el}(\boldsymbol{r}, \boldsymbol{R}) = E_{el}\Psi_{el}(\boldsymbol{r}, \boldsymbol{R}), \tag{4}$$

where the purely electronic Hamiltonian operator $\hat{H}_{el}$ defined as

$$\hat{H}_{el} = T_e + U_{ee} + U_{en}. \tag{5}$$

The purely electronic energy E_{el} depends on the nuclear coordinates $\boldsymbol{R}$, but only parametrically.

2.3 Density Functional Theory

Applying the Born-Oppenheimer approximation simplifies the problem of the many body Hamiltonian considerably. However, it is still unfeasible to solve the Schrödinger equation with this approximation for many materials. The problem is the term U_{ee} that describes the Coulomb interactions between the electrons. Because of this term the problem is still a many body problem, where all electrons need to be considered simultaneously.

Density Functional theory simplifies this problem. Instead of dealing with the complex many-electron wavefunction as in traditional quantum mechanical methods, DFT focuses on the electron density $\rho(\boldsymbol{r})$ as the central variable. This reduces the computational complexity and allows efficient calculations even for large systems. The approach of using the electron density as a key quantity was first used by Thomas and Fermi in 1927 [23], although with ambiguous success. Nowadays DFT is one of the major players in computational quantum chemistry [24]. Walter Kohn, probably the most important developer of modern DFT has been awarded the Nobel prize in 1998 for his work. The persistent success of DFT comes from its good balance between accuracy and computational cost.

2.3.1 Hohenberg-Kohn theorems

Density functional theory is based on the Hohenberg-Kohn theorems [5], formulated by Pierre Hohenberg and Walter Kohn in 1964. To understand them, let us again consider the purely electronic Hamiltonian (equation (5)). In atomic units it can be expressed as

$$\begin{aligned} \hat{H} &= -\frac{1}{2}\sum_{i=1}^{N}\nabla_i^2 + \sum_{i=1}^{N} v(\boldsymbol{r}_i) + \sum_{j}\sum_{i>j}\frac{1}{r_{ij}} \\ \text{with} \quad v(\boldsymbol{r}_i) &= -\sum_k \frac{Z_k}{r_{ik}}. \end{aligned} \tag{6}$$

The expressions r_{ij} and r_{ik} are shorthand notations for $|\boldsymbol{r}_i - \boldsymbol{r}_j|$ and $|\boldsymbol{r}_i - \boldsymbol{R}_k|$, respectively. In equation (6) we introduced the external potential $v(\boldsymbol{r})$, that represents the potential energy interactions between the electrons and the nuclei. It is called *external*, because it is caused by charges that are external to the systems of electrons. The first Hohenberg—Kohn theorem states that $v(\boldsymbol{r})$ is determined within a trivial additive constant by the knowledge of the electron density $\rho(\boldsymbol{r})$ [25]. Hence, $\rho(\boldsymbol{r})$ determines the electronic Hamiltonian and therefore also the corresponding wavefunction and total energy. Consequentially the ground state electronic energy E_0 can be expressed using the purely electronic Hamiltonian (5) as a functional of the ground state electron density ρ_0

$$E_0 = \bar{T}_e[\rho_0] + \bar{U}_{ee}[\rho_0] + \bar{U}_{en}[\rho_0], \tag{7}$$

The overbars in equation (7) indicate that the individual terms are not operators, but rather expectation values determined by the ground state electronic wavefunction. The wavefunctions, on the other hand, are determined by the ground state electron density. Therefore, each term of equation (7) is a functional of ρ_0. From equation (2) and the definition of $v(\boldsymbol{r}_i)$ in equation (6) we get

$$U_{en} = \sum_{i=1}^{N} v(\mathbf{r}_i), \tag{8}$$

Furthermore, calculating the expectation value of equation (8) we obtain

$$\bar{U}_{en}[\rho] = \left\langle \Psi \left| \sum_{i=1}^{N} v(\mathbf{r}_i) \right| \Psi \right\rangle = \int \rho(\mathbf{r}) v(\mathbf{r}) d\mathbf{r}. \tag{9}$$

This means that $\bar{U}_{en}$ is known and depends on the specific system. We can now rewrite (7) as

$$E_0 = \int \rho_0(\mathbf{r})v(\mathbf{r})d\mathbf{r} + F[\rho_0], \tag{10}$$

where $F[\rho_0] = \bar{T}_e[\rho_0] + \bar{U}_{ee}[\rho_0]$ is the universal part of the energy functional, that does not depend on the external potential $v(\mathbf{r})$ and is thus independent from the specific atomic system. $F[\rho_0]$ is not known but it can be obtained in a variational procedure. In order to see why this is possible we need the second Hohenberg-Kohn theorem.

The second Hohenberg-Kohn theorem states that the electronic energy of any trial ground state electron density ρ' can not be larger than the electronic energy of the ground state electron density:

$$\bar{T}_e[\rho'] + \bar{U}_{ee}[\rho'] + \int \rho' v(\mathbf{r})d\mathbf{r} \geq E_v(\rho_0). \tag{11}$$

Here, $E_v[\rho]$ is the functional that describes the electronic energy with external potential $v(\mathbf{r})$. This means that the ground state electron density ρ_0 minimizes the energy functional $E_v[\rho']$.

2.3.2 Kohn-Sham equation

In 1965, Walter Kohn and Lu Jeu Sham introduced a procedure to calculate ρ_0 and ultimately E_0 [6]. Their approach can principally yield correct results. However, it requires to calculate the wanted quantities by successive approximation. This is necessary, because the Kohn-Sham formalism includes an unknown functional, the exchange-correlation functional.

Kohn and Sham considered a fictitious reference system with non-interacting electrons, the so-called noninteracting system. Quantities that are defined with respect to this noninteracting system will from now on be marked with the index s. Let us now introduce two new expressions, $\Delta\bar{T}_e$ and $\Delta\bar{U}_{ee}$, dropping the 0 index of the ground state electron densities. The first one is

$$\Delta\bar{T}_e[\rho] := \bar{T}_e[\rho] - \bar{T}_{e,s}[\rho]. \tag{12}$$

This is the difference between the kinetic energy in the physical system and the kinetic energy in the noninteracting system with the same electron density. The second new expression is

$$\Delta\bar{U}_{ee}[\rho] := \bar{U}_{ee}[\rho] - \frac{1}{2}\int\int \frac{\rho(\mathbf{r}_1)\rho(\mathbf{r}_2)}{\mathbf{r}_{12}} d\mathrm{r}_1 d\mathrm{r}_2. \tag{13}$$

Here, $\mathbf{r}_{12}$ is the distance between two points in space, r_1 and r_2 and $\int\int \rho(\mathbf{r}_1)\rho(\mathbf{r}_2)/\mathbf{r}_{12} d\mathrm{r}_1 d\mathrm{r}_2$ is the classical electrostatic repulsion energy of a continuous charge density ρ. With the definitions (12) and (13), equation (10) can be rewritten as

$$E_v[\rho] = \int \rho(\mathbf{r})v(\mathbf{r})d\mathbf{r} + \bar{T}_{e,s}[\rho] + \frac{1}{2}\int\int \frac{\rho(\mathbf{r}_1)\rho(\mathbf{r}_2)}{\mathbf{r}_{12}} + \Delta\bar{T}_e[\rho] + \Delta\bar{U}_{ee}[\rho] \tag{14}$$

Note that the functionals $\Delta\bar{T}_e$ and $\Delta\bar{U}_{ee}$ are unknown. By defining the exchange-correlation energy functional $E_{xc}[\rho]$ as

$$E_{xc}[\rho] := \Delta\bar{T}_e[\rho] + \Delta\bar{U}_{ee}[\rho] \tag{15}$$

we can rewrite equation (14) as

$$E_v[\rho] = \int \rho(\mathbf{r})v(\mathbf{r})d\mathbf{r} + \bar{T}_{e,s}[\rho] + \frac{1}{2}\int\int \frac{\rho(\mathbf{r}_1)\rho(\mathbf{r}_2)}{\mathbf{r}_{12}} + E_{xc}[\rho]. \tag{16}$$

We can introduce single-particle orbitals θ_i^{KS} for the non-interacting system that allow us to write the density ρ as

$$\rho = \sum_{i=1}^{N} |\theta_i^{KS}|^2. \tag{17}$$

Using equation (17) and

$$\bar{T}_{e,s}[\rho] = -\frac{1}{2}\sum_i \langle \theta_i^{KS}(1)|\nabla_1^2|\theta_i^{KS}(1)\rangle \tag{18}$$

we can find the ground state energy density ρ_0 by finding the Kohn-Sham orbitals θ_i^{KS}. These orbitals satisfy the Schrödinger-like equation

$$\left[-\frac{1}{2}\nabla_1^2 - \sum_k \frac{Z_k}{r_{1k}} + \int \frac{\rho(\mathbf{r}_2)}{r_{12}}d\mathbf{r}_2 + v_{xc}(1)\right]\theta_i^{KS}(1) = \epsilon_i^{KS}\theta_i^{KS}(1). \tag{19}$$

Here, ϵ_i^{KS} are the Kohn-Sham orbital energies and $v_{xc}(1)$ is the exchange correlation potential. The latter is defined as

$$v_{xc}(\mathbf{r}) := \frac{\delta E_{xc}[\rho(\mathbf{r})]}{\delta\rho(\mathbf{r})}. \tag{20}$$

Equation (19) can be rewritten as

$$\hat{h}^{KS}(1)\theta_i^{KS} = \epsilon_i^{KS}\theta_i^{KS}(1), \tag{21}$$

introducing the one-electron Kohn-Sham Hamilton operator $\hat{h}^{KS}$.

In practice, a Kohn Sham DFT calculation starts with an initial guess of the Kohn-Sham wavefunctions θ_i^{KS}, that can be used to calculate ρ. Subsequently, the exchange-correlation potential $v_{xc}(\boldsymbol{r})$ is calculated using equation (20). Now that the Kohn-Sham Hamiltonian operator $\hat{h}^{KS}$ is fully determined, an updated version of the Kohn-Sham wavefunctions can be calculated with equation (19). These updated wavefunctions can then be used to perform the mentioned calculations again. The described iterative procedure is then carried out until a predefined convergence criterion is met.

A major issue with the Kohn-Sham method is that the exchange-correlation functional $E_{xc}[\rho]$ is unknown. However, there are several approaches to approximate $E_{xc}[\rho]$, like the local-density approximation (LDA) [6,26] and the generalized-gradient approximation (GGA) [27,28].

2.4 Hubbard Correction

As mentioned above, the exact expression of the exchange-correlation functional $E_{xc}[\rho]$ and therefore of the total energy functional is unknown. To overcome this problem, approximations like LDA and GGA are necessary. Unfortunately approximate functionals that are used in today's DFT are generally not able to fully capture the complexity of the quantum many-body problem. In fact, approximate exchange-correlation functionals show a tendency to delocalize valence electrons too much and to over-stabilize metallic ground states [29, 30]. Typical examples used to illustrate the effects of that error are Mott insulators. According to classical band theory Mott insulators should be classified as metals [31]. Using approximate DFT functionals generally leads to the same picture [29]. However, Mott insulators, as the name already says, show insulating behavior in experiments. The insulating nature of Mott insulators is caused by strong Coulomb repulsion between electrons. This repulsion forces the electrons to localize in atomic-like orbitals (Mott localization) [29, 31]. Due to this localization the electrons can not move freely between the atoms. Instead, they jump between neighboring atoms through a hopping mechanism. When the electrons do not have enough energy to move between different atom sites, the system is insulating.

The localization of orbitals that among other things leads to the insulating character of Mott insulators can be modeled by the introduction of the Hubbard correction [32]. The Hubbard correction is applied by adding an "on-site" Coulomb repulsion with strength U to the relevant partially filled orbitals. More precisely, the addition of the Hubbard correction leads to a new expression of the total energy, that can be formulated as (Dudarev's correction, adapted from [33]):

$$E_{DFT+U} = E_{DFT} + \frac{U}{2}\sum_{\sigma}(n_{m,\sigma} - n^2_{m,\sigma}). \tag{22}$$

Here, $n_{m,\sigma}$ is the occupation number of the mth state of the interacting shell (most of the time d or f shell) with a given projection of spin σ.

The parameter U enters the DFT+U calculation as an input variable. The numerical value of the parameter has to be chosen carefully and with good understanding of the approximations. Often this is done in a semi-empirical way, by comparing the results of DFT calculations with different values of U with experimental data for example on the band gap. Finally a numerical value is chosen for the parameter U that leads to a best possible agreement with the experiments. However, there also exist more empirical methods to estimate the numerical value of U for a given material system, such as the constrained local density approximation approach [34, 35] or the constrained random-phase approximation approach [36].

2.5 Constrained magnetic moment

In a magnetic material, there exist atoms with non-vanishing magnetic moments. In theory, the orientation of the magnetic moment may be different for each atom (see section 4.2). All possible combinations of single atomic magnetic moments are contained in the magnetic configuration space of a material. In order to explore the magnetic configuration space of magnetic materials it is necessary to be able to impose an arbitrary magnetic configuration within DFT, including configurations that do not represent the ground state of the system. This can be achieved in VASP by adding an additional penalty contribution to the energy expression that forces the local magnetic moments to get close to or take the desired configuration $\hat{\boldsymbol{M}}_I^0$ [37]. This penalty term adds to the original energy expression E_0 as follows:

$$E = E_0 + \sum_I \lambda \left[\boldsymbol{M}_I - \hat{\boldsymbol{M}}_I^0 (\hat{\boldsymbol{M}}_I^0 \cdot \boldsymbol{M}_I) \right]^2 \tag{23}$$

Here the index I determines the specific atom in the supercell, $\hat{\boldsymbol{M}}_I^0$ is the normalized desired magnetic moment and $\boldsymbol{M_I}$ is the magnetic moment integrated in a sphere Ω_I with a radius equivalent to the Wigner-Seitz radius of the atom. Consequently, $\boldsymbol{M_I}$ can be expressed as

$$\boldsymbol{M}_I = \int_{\Omega_I} \boldsymbol{m}(\boldsymbol{r}) F_I(|\boldsymbol{r}|) d\boldsymbol{r}. \tag{24}$$

The function $F_i(|\boldsymbol{r}|)$ in equation (24) has the value one in the center of the sphere and monotonously goes to zero at the boundary of the sphere. The parameter λ from equation (23) is a Lagrange multiplier that can be used to lower the penalty energy term E_p at the minimum of the constrained energy functional. It can be shown that E_p is inversely proportional to the Lagrange multiplier λ. A proof of this claim and further details of the constraining mechanism implemented in VASP can be found in [37]. For the calculations performed in this work λ was chosen in such a way that the penalty energy contribution E_p does not exceed 10^{-5} eV.

3 Bayesian Optimization

Bayesian optimization is an active machine learning scheme based on Gaussian processes. Contrary to other machine learning approaches that train with a fixed set of data, active machine learning methods actively acquire new data. The goal of Bayesian optimization is to create model functions of possibly complicated, hard to evaluate functions and dependencies. Bayesian optimization is designed for cases where these unknown underlying functions are so hard to evaluate that the limited resources only allow their evaluation for a small number of input values. It can be especially useful to use Bayesian optimization if the underlying function has a multi-dimensional input space.

One can roughly say that Bayesian optimization consists of two separate, but interconnected parts. One part is designed to create the surrogate model functions based on the already acquired data points without the use of a predefined "fitting function". For the realization of this task Gaussian processes are used. The other part is designed to determine the input data for the next function evaluation that presumably contributes the most to the understanding of the underlying unknown function. This decision is made with the help of the so-called acquisition functions. Besides modeling an unknown function, Bayesian optimization can also be used to efficiently search for the global minimum of an unknown function, even if this unknown function has several local minima.

In this thesis work Bayesian optimization is used in the form of the program *Bayesian optimization Structure Search* (BOSS) [13]. BOSS was developed by the Computational Electronic Structure Theory group at Aalto university in Finland. It is maintained and updated continuously.

3.1 Gaussian Processes

Let us consider an arbitrary continuous function f, analytical or not. As for any function, it is possible to discretize this example function, so that the resulting discrete function can be seen as a collection of points. Since the original function was continuous, one would expect that the function values at neighboring points $f(x_i)$ and $f(x_{i+1})$ are not independent from each other. In other words, the covariance between both points would be positive. On the other hand, looking at two points $f(x_i)$ and $f(x_{i+n})$, where n is large, one would not expect to see similar function values. The covariance of both points would most likely be low.

One example, where the covariance between two points x_i and x_j plays an important role is Brownian motion. Brownian motion describes the random motion of particles suspended in a liquid or gaseous medium. The motion of such a particle is determined by the random forces caused by the random fluctuations of the neighboring particles. Therefore, Brownian motion can be seen as a continuous random walk. For more information on Brownian motion see for example [38]. It is possible to simulate such a process in one space dimension d, which describes the displacement of the particle over time t by starting at $d(t = 0) = 0$ and adding a step width δd_t for every timestep δt. The step width δd_t is drawn from a given probability distribution at every times step. The result of this procedure is a discrete function $d(t)$ that is determined by $d(t + \delta t) = d(t) + \delta d_t$. Three different possible "trajectories" $d(t)$ of a hypothetical particle can be seen in figure 1. The positions of the particle at different points in time are not independent from each other within a small time interval, as indicated above. However, the specific change of the particle's position within a certain amount of time is determined by a random, stochastic law. Because of the described stochastic nature of Brownian motion, it is an example of a *stochastic process*. A stochastic process is a mathematical object that describes an infinite sequence of joint random variables. In fact, the trajectory $d(t)$ of a particle that obeys Brownian motion can be seen as a result of the stochastic process that describes the Brownian motion. To be more precise, if we look at the time interval $I := [0, T]$, each $t \in I$ can be seen as a random variable. Since I is continuous, there is an infinite number of possible t and therefore an infinite number of random variables. However, the particle positions $d(t)$ for different $t \in I$ are correlated. Hence we have an infinitive sequence of joint random variables. Another important aspect of stochastic processes can be illustrated using the example of Brownian motion. In figure 1, three different functions or particle trajectories are shown. All three of those functions result from the same joint probability distribution meaning that they result from the same stochastic process. Obviously, an infinite number of particle trajectories are imaginable theoretically. This example shows, that a stochastic process does not represent only one single resulting function, but rather a family of possible functions. Stochastic processes can therefore also be seen as distributions over functions. Individual functions that are sampled from a stochastic process are called *realizations* of the stochastic process.

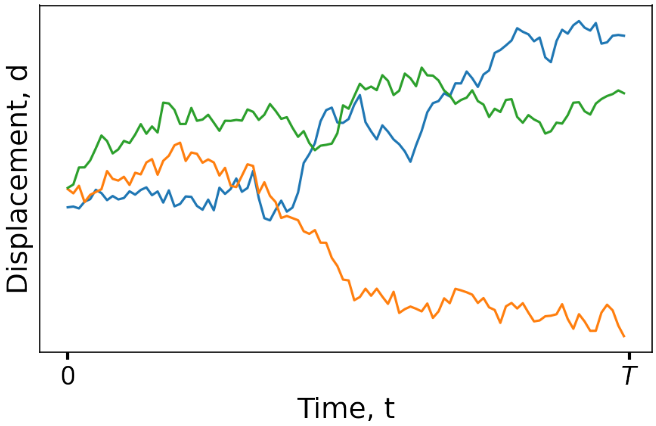

Figure 1: Three examples of one dimensional Brownian motion. Shown is the displacement d of a hypothetical particle from the initial position over time t.

In the example of Brownian motion, a displacement δd_t of the particle is drawn from a stochastic distribution at every time t. If the stochastic distribution is a

Gaussian distribution, the corresponding stochastic process is called a *Gaussian process.* Corresponding to the previous definition of a stochastic process, one can define a Gaussian process as an infinite sequence of joint random variables $x \in X$, where any finite number of which have a joint Gaussian distribution [39, p.13]. Since a Gaussian process is generally defined for a infinite-dimensional set of random variables, it is not possible to define a covariance matrix. However it is possible to define a covariance function $k(x, x')$, also called a *kernel function.* The covariance function can be interpreted as a notion of "closeness" between two function values $f(x)$ and $f(x')$ of the points x and x', averaged over all possible functions f that can be sampled from the Gaussian process. It can therefore be defined in the following way:

$$k(x, x') = \mathbb{E}[(f(x) - m(x))(f(x') - m(x'))] \tag{25}$$

The covariance function is an important part of Gaussian process regression, because it contains information about the assumptions on the data. A valid kernel function needs to be positive semi-definite [39, p.79]. If a covariance function is translational invariant, meaning that it can be written as a function of $x - x'$, then it is called a *stationary* covariance function. If it can furthermore be written as a function of $|x - x'|$ the covariance function is called *isotropic.* Additional to the kernel function, one more thing is necessary to unambiguously specify a Gaussian process, namely its mean function $m(x)$. The mean function of a Gaussian can be defined as

$$m(x) = \mathbb{E}[f(x)] \tag{26}$$

The expectation value in equation (26) is taken over all functions $f(x)$ that can be sampled from the Gaussian process. Consequentially we can write a Gaussian process as

$$f(x) \sim \mathcal{GP}(m(x), k(x, x')). \tag{27}$$

The definition of a Gaussian process automatically yields a consistency requirement, also called *marginalization property* [39, p.13]. Let us consider the covariance matrix Σ of a finite subset A of the input domain X. The consistency requirement states that if the Gaussian process specifies $(y_1, y_2 \sim \mathcal{GP}(m(A), \Sigma)$, then it must also specify $y_1 \sim \mathcal{GP}(m(B), \Sigma_{11})$, where Σ_{11} is the relevant submatrix of Σ that corresponds to $B \subseteq A$. In other words, examining a broader set of variables has no impact on the distribution of the smaller set.

3.1.1 The squared exponential kernel

One of the most often used kernel functions is the *squared exponential kernel*, which can be defined as

$$k(x, x') = \sigma^2 \exp\left(-\frac{|x - x'|^2}{2\ell^2}\right). \tag{28}$$

As one can easily see, the squared exponential function is an isotropic covariance function. The hyperparameters σ and ℓ are called *variance* and *lengthscale*, respectively. The effects of these hyperparameters are demonstrated in figure 2. In figures 2a, 2b and 2c, the squared exponential covariance function is shown as a function of $x - x'$ with different values of the hyperparameters. Figures 2d, 2e and 2f show three realizations of a corresponding Gaussian processes, respectively. In all three cases, the mean function $m(x) = 0$ was chosen for the definition of Gaussian process. The kernel function and the realizations that can be seen in the

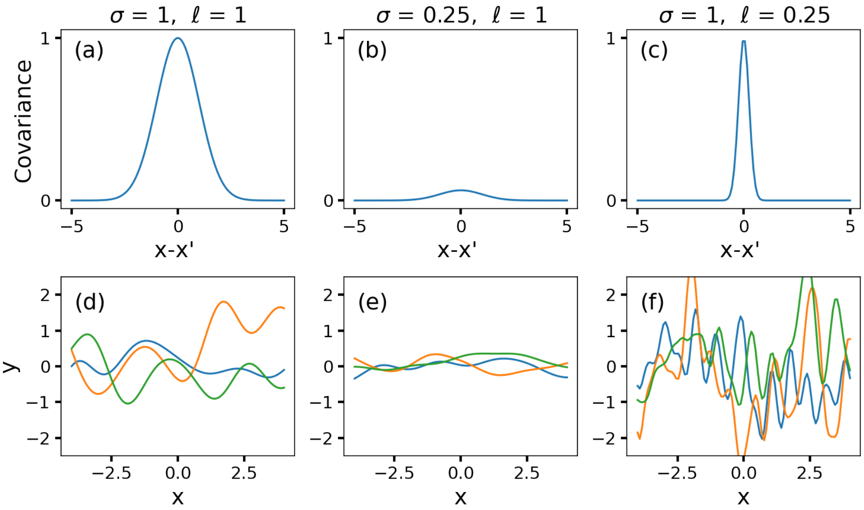

Figure 2: Effect of different values for the hyperparameters of the squared exponential kernel (eq. (28)). The figures a, b and c show the kernel function as a function of $x - x'$. Three different settings were chosen. The exact values for the variance σ and the lengthscale ℓ can be found above the figures. Figures d, e and f shows three different realizations of the corresponding kernel function above them.

figures a and d correspond to the hyperparameters $\sigma = 1$ and $\ell = 1$. By lowering the value of the variance σ from 1 to 0.25, the covariance function changes (see figure 2b. More specifically, the function value of the covariance function decreases for all $x - x'$. This can also be derived from the mathematical formulation of the squared exponential kernel in equation (28). As a result, the corresponding realizations tend to deviate less from the mean function of the Gaussian process. Note that the plotted y-axis interval of the figures is the same in all of the three cases.

The covariance function that can be seen in figure 2c was created using the same value for σ that was used in figures 2a and 2d. However, the lengthscale ℓ is reduced to 0.25. The functions that were sampled from the Gaussian process which uses this low value for ℓ (figure 2f) vary more rapidly. Thus, the lengthscale can be seen as a correlation length, meaning that it gives information about the maximal distance $|x - x'|$, below which the function values $f(x)$ and $f(x')$ are close to each other.

During Gaussian process regression, the hyperparameters of a chosen kernel are optimized for the specific set of observations. More on this process in section 3.4.

3.1.2 The standard periodic kernel

Often Bayesian optimization is used to characterize systems that are previously almost or completely unknown. However, in some cases information about the periodicity of the objective function exists. One can, for example, often assume a 2π-periodicity of a function whose argument is a rotation angle. In these cases it may be helpful to use a kernel that already "expects" periodicity. One such kernel is the so called *standard periodic kernel*:

$$k(x, x') = \sigma^2 \exp\left(-\frac{2}{\ell^2}\sin^2\left(\pi\frac{|x - x'|}{p}\right)\right) \tag{29}$$

The standard periodic kernel has some similarities with the squared exponential kernel such as the existence of the hyperparameters variance σ and lengthscale ℓ, that play a similar role as in the squared exponential covariance function. These two hyperparameters also have comparable effects on the kernel and resulting realizations as in the case of the squared exponential kernel that are exemplary depicted in figure 2. A major difference between the two kernels is that the argument of its exponential function contains a squared sine function. This periodic argument introduces a periodicity to the whole covariance function. The period of the standard periodic kernel is given by the hyperparameter p. Examples of the standard periodic covariance function and corresponding realizations for different values of the hyperparameters ℓ and p can be seen in figure 3.

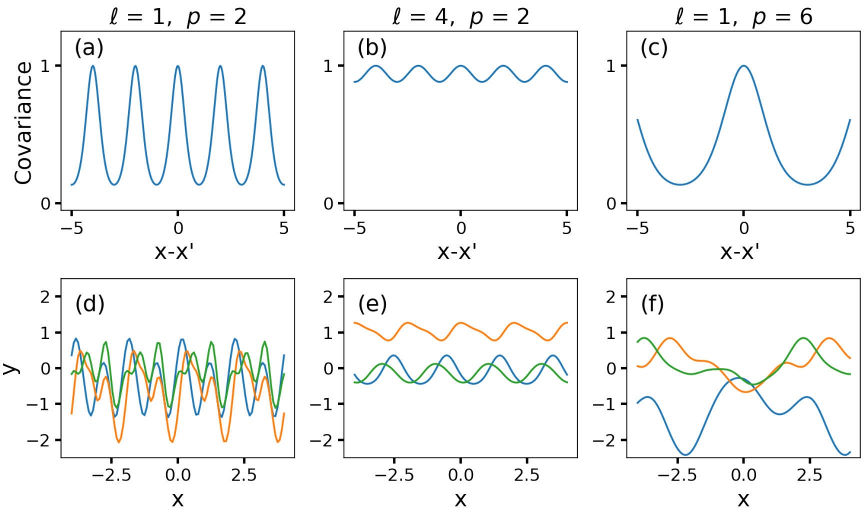

Figure 3: Examples of the standard periodic kernel (eq. (29)) with different hyperparameters and three corresponding realizations respectively. Figures a, b and c show the covariance as a function of $x - x'$. Figures d, e and f show the corresponding realizations. Three different settings were chosen. The exact values of the chosen hyperparameters can be found above the corresponding column. The variance σ was always set to 1.

As one can see from equation (29) and figure 3 the covariance function is periodic. It is always highest when $x - x' + ap$ with $a \in \mathbb{Z}$ is small. As discussed before, this high covariance leads to similar function values $f(x)$ and $f(x')$. Hence the realizations of a Gaussian process with the standard periodic kernel are also periodic. The standard periodic covariance functions that can be seen in figures 3a and 3b differ only by the value of the hyperparameter lengthscale ℓ, resulting in higher values for the covariance for every value of $x - x'$, while preserving the periodicity. The realizations that correspond to this covariance function (figure 3e) are "smoother" than the ones that correspond to a lower value of ℓ (figure 3d), meaning that they show fewer specific features and can be more easily compared to simple trigonometric functions. The effect of the hyperparameter p can be seen by comparing figures 3a and 3d with figures 3c and 3f. In the former case, the period p is set to two and in the latter case it is set to six, while keeping the other two hyperparameters σ and ℓ at the same value. The comparison between the figures shows that varying p changes the period of both the covariance function and the realizations, as could be expected from equation (29).

3.1.3 The Matérn class of covariance functions

The Matérn covariance functions are named by Stein [40] after the forestry statistician Bertil Matérn [41], though they were deduced earlier by Whittle in 1954 [42]. The Matérn covariance functions form a class of covariance functions that are parameterized by the positive smoothness parameter ν. Often, like in [39, p.84] they are defined as

$$k_{\text{Matérn}}(r) = \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\sqrt{2\nu r}}{\ell} \right)^{\nu} K_{\nu} \left(\frac{\sqrt{2\nu r}}{\ell} \right), \tag{30}$$

where Γ is the gamma function, K_ν is a modified Bessel function of order ν [43, S. 444 *et seq.*] and ℓ is the hyperparameter lengthscale. Here we use $r = |x - x'|$ for the distance between two points x and x'. This is possible because the Matérn covariance functions are isotropic. An advantage of the Matérn covariance functions is their great flexibility. The positive smoothness parameter ν can be used to adapt the smoothness of the resulting functions to the existing problem. When ν is small the functions that can be sampled from the resulting process tend to be more rough, and when ν is large the resulting functions are more smooth. In fact, for specific values of ν the Matérn covariance functions match several other models. For example, it can be shown that for $\nu \to \infty$ the Matérn function represents a Gaussian model [39, 44]. The expression (30) simplifies when $\nu = p + 1/2$ if p is a non-negative integer. With this condition fulfilled, the Matérn covariance functions can be rewritten as a product of an exponential and a polynomial of order p [39, 44]:

$$k_{\nu=p+1/2}(r) = \exp\left(-\frac{\sqrt{2\nu r}}{\ell} \right) \frac{\Gamma(p+1)}{\Gamma(2p+1)} \sum_{i=0}^{p} \frac{(p+1)!}{i!(p-1)!} \left(\frac{\sqrt{8\nu r}}{\ell} \right)^{p-i}. \tag{31}$$

Figure 4a shows the Matérn covariance function with the smoothness parameters $\nu = 1/2$, $\nu = 3/2$ and $\nu = 5/2$, respectively, comparing them to the squared

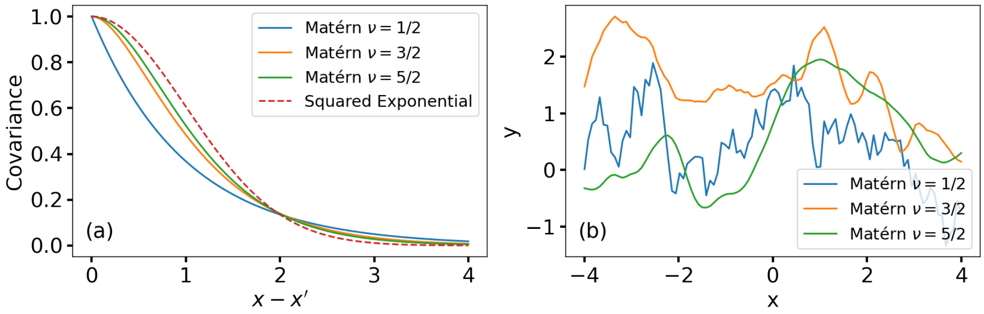

Figure 4: Comparison of the Matérn kernels with smoothness parameter $\nu = 1/2$, $\nu = 3/2$ and $\nu = 5/2$ and $\ell = 1$. Figure a shows the kernel functions of the three Matérn kernels and the kernel function of the square exponential kernel for comparison. Figure b shows a random realization of the three Matérn kernels, respectively.

exponential covariance function. As the figure indicates, the covariance between close points x and x' is higher for high values of ν. This ultimately leads to more smooth realizations (see figure 4b). Using $\nu = 1/2$, the resulting process becomes relatively rough, probably too rough for most applications (see figure 4b). Furthermore, it may thereby lead to a too high ability to adapt to random noise in the original data. These problems are less severe for $\nu = 3/2$ and $\nu = 5/2$. The processes resulting from the Matérn kernels using $\nu \geq 7/2$ are very hard to distinguish from each other and from the smooth squared exponential. Therefore the probably most interesting Matérn covariance functions are the ones that correspond to $\nu = 3/2$ and $\nu = 5/2$ [39, p.85]. They read

$$k_{\nu=3/2}(r) = \left(1 + \frac{\sqrt{3}r}{\ell}\right) \exp\left(-\frac{\sqrt{3}r}{\ell}\right), \tag{32}$$

$$k_{\nu=5/2}(r) = \left(1 + \frac{\sqrt{5}r}{\ell} + \frac{5r^2}{3\ell^2}\right) \exp\left(-\frac{\sqrt{5}r}{\ell}\right). \tag{33}$$

Examples of realizations of these kernels are shown in figure 4b. Since $\nu = 3/2$ and $\nu = 5/2$ allow more rough functions than the squared exponential covariance functions, they may be more suitable when there are sudden jumps in the objective function.

3.2 Gaussian Process Regression

Regression is a machine learning discipline with the goal to predict real values for previously unknown inputs. Gaussian processes can be used for regression. To understand the mechanisms of Gaussian process regression we need to introduce the notions *prior* and *posterior*. In Bayesian statistics, a prior distribution describes the subjective beliefs that one has about a specific system, before taking into account available data. It is a measure of how “plausible” each parameter value is considered at this stage [45]. A prior in Gaussian process regression is defined by a Gaussian process and therefore by the mean function and the covariance function with given hyperparameters before taking into account available data. A posterior distribution is the probability distribution that one assigns to each value of a parameter after analyzing the available data. It is built based on a prior distribution and the observed data [45]. We will later see that the posterior distribution for all values of the input parameter x can also be described by a Gaussian process, that was determined using the available information $f(\boldsymbol{X})$ for a finite number of points $\boldsymbol{X} = x_1, ..., x_n$ on an objective function f. Even if the data $f(\boldsymbol{X})$ is noisy it is possible to determine a Gaussian process posterior as we will see in section 3.3.

Suppose that n observations $(\boldsymbol{X}, \boldsymbol{y})$ have been made with $\boldsymbol{X} = (x_1, ..., x_n)$ and $\boldsymbol{y} = (y_1, ..., y_n)$. It is then possible to make predictions for the new inputs $\boldsymbol{X_*}$ by drawing $\boldsymbol{f_*}$ from the posterior distribution $p(f|(\boldsymbol{X}, \boldsymbol{y}))$. Obviously, the previous observations $\boldsymbol{y}$ and the new predictions $\boldsymbol{f_*}$ follow a joint Gaussian distribution. This distribution can be written as follows [46]:

$$\begin{bmatrix} \boldsymbol{y} \\ \boldsymbol{f_*} \end{bmatrix} \sim \mathcal{N}\left(\boldsymbol{0}, \begin{bmatrix} K(\boldsymbol{X}, \boldsymbol{X}) & K(\boldsymbol{X}, \boldsymbol{X_*}) \\ K(\boldsymbol{X_*}, \boldsymbol{X}) & K(\boldsymbol{X_*}, \boldsymbol{X_*}) \end{bmatrix}\right) \tag{34}$$

Here, $\mathcal{N}(m, C)$ is the notion for a joint Gaussian distribution with mean m and covariance matrix C. Furthermore, $K(\boldsymbol{X}, \boldsymbol{X})$ and $K(\boldsymbol{X_*}, \boldsymbol{X_*})$ are the covariance matrix between all previously observed points and the covariance matrix between all newly introduced points, respectively. $K(\boldsymbol{X}, \boldsymbol{X_*})$ and $K(\boldsymbol{X_*}, \boldsymbol{X})$ are the covariance matrices between the previously observed points and the newly introduced points and vice versa. It can be shown [47, chapter VIII, section 9.1] that the conditional distribution $p(\boldsymbol{f_*}|\boldsymbol{X}, \boldsymbol{y}, \boldsymbol{X_*})$ also is a multivariate normal distribution. This resulting multivariate normal distribution has the mean

$$K(\boldsymbol{X_*}, \boldsymbol{X})K(\boldsymbol{X}, \boldsymbol{X})^{-1}\boldsymbol{y} \tag{35}$$

and covariance matrix

$$\Sigma = K(\boldsymbol{X_*}, \boldsymbol{X_*}) - K(\boldsymbol{X_*}, \boldsymbol{X})K(\boldsymbol{X}, \boldsymbol{X})^{-1}K(\boldsymbol{X}, \boldsymbol{X_*}). \tag{36}$$

A prove of this statement can be found in [47, chapter VIII, section 9.3].
Note that if $[y, \boldsymbol{f_*}]^T$ is assumed to be continuous the multivariate normal distribution can be seen as a Gaussian process with mean function

$$m_{post}(\boldsymbol{x}) = K(\boldsymbol{x}, \boldsymbol{X})K(\boldsymbol{X}, \boldsymbol{X})^{-1}\boldsymbol{y} \tag{37}$$

and kernel

$$k_{post}(\boldsymbol{x}, \boldsymbol{x'}) = k(\boldsymbol{x}, \boldsymbol{x'}) - K(\boldsymbol{x}, \boldsymbol{X})K(\boldsymbol{X}, \boldsymbol{X})^{-1}K(\boldsymbol{X}, \boldsymbol{x'}). \tag{38}$$

Figure 5a shows five exemplary realizations of the Gaussian process posterior as well as the known data points $\boldsymbol{X}$ (black dots). For this example the test function

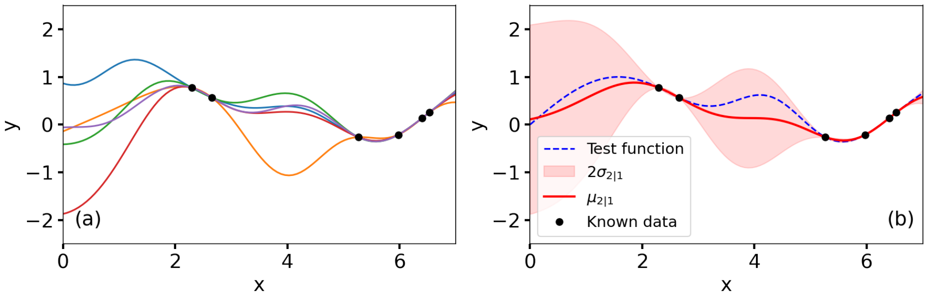

Figure 5: A Gaussian process posterior. Figure a shows known data (black dots) and five randomly sampled realizations of the Gaussian process that results from the Gaussian process regression. Figure b shows the test function (dashed blue), the mean function (dark red), the confidence region (light red) and the previously acquired data points (black dots). The acquired data points were sampled from the test function.

$f_{test}(x) = \sin(x) + 1.5 \cdot \exp(-(x - 4.3)^2)$ was used and six points were chosen at random to be used as "known data". In some regions (*e.g.* right third) the five sampled functions agree well and in other regions (*e.g.* left third) they diverge between the known data points. All realizations of the Gaussian process shown in figure 5a run threw the known data points. This is not a coincidence since they are sampled using the conditional probability distribution defined earlier. In figure 5b, both the test function (dashed blue line) as well as the randomly chosen "known data points" are shown. Also plotted is the resulting mean function $m(x)$ (dark red) and the confidence region of the mean function (red shadow). This region is defined for every point x of the shown x-axis interval as $[m(x) - 2\sigma_x, m(x) + 2\sigma_x]$. Here σ_x is the standard deviation of the Gaussian distribution of the random variable x. The value of the standard deviation can be extracted from the covariance matrix as (see *e.g.* [48, p.113 *et seqq.*]):

$$\sigma_{x(i)} = \sqrt{(\Sigma)_{ii}}, \tag{39}$$

where i is the index of the multivariate distribution that corresponds to x. The fact that it is possible to determine a standard deviation for every point of the mean function using Gaussian process regression is a great advantage of that approach. It allows to use smart sampling strategies, that are also part of Bayesian optimization (see section 3.5). As shown in figure 5b, the mean function agrees well with every known data point. At these points, uncertainty, *i.e.* the standard deviation drops to zero. This is an accurate assumption, because this model assumes that observations are noiseless. The right figure also shows that the mean function only deviates from the test function in regions, where the uncertainty is high. In this example, but also in most other cases, the test function (*i.e.* the objective function) lies within the 2σ-range of the mean function. Therefore, in Gaussian process regression the mean function can serve as the surrogate model for the unknown objective function.

Figure 5 also gives a graphic interpretation of how Gaussian process regression works. From all possible functions that can be sampled from a Gaussian process prior, as seen for example in the figures 2d - 2f, those are chosen for the posterior that fulfill the condition that they run threw the acquired data points (figure 5a). From these sampled functions, the surrogate model $m(x)$ and the variance $\sigma(x)$ are calculated (figure 5b). Although the described procedure would yield correct results, it is not used for real applications of Gaussian process regression. Sampling the necessary number of functions would simply be much more computationally heavy than the procedure of calculating the conditional probabilities described above. However, the idea of selecting the functions that fulfill the necessary requirements transports an accurate and helpful image.

3.3 Noisy observations

Until now we have assumed that the observed data is noiseless. Admittedly this is an assumption that cannot be made in general. However, in most cases it is possible to anticipate a certain level of noise. This information can be used to get more accurate predictions by Gaussian process regression and it helps the model not to get "confused" by inaccurate data. For the implementation of Gaussian process regression that expects a certain level of noise it is necessary to update the model that was introduced so far. This updated model assumes that there is a continuous objective function $f(x)$, so that the output of the data acquisition process y can be written as

$$y = f(x) + \epsilon, \tag{40}$$

where $\epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2)$ is a normal distributed random variable with variance σ_ϵ [46]. Following the calculation that previously led to equation (34), this assumption yields the following expression of the joint Gaussian distribution of $\boldsymbol{y}$ and $\boldsymbol{f}_*$:

$$\begin{bmatrix} \boldsymbol{y} \\ \boldsymbol{f}_* \end{bmatrix} \sim \mathcal{N}\left(\boldsymbol{0}, \begin{bmatrix} K(\boldsymbol{X},\boldsymbol{X}) + \sigma_\epsilon^2 \boldsymbol{I} & K(\boldsymbol{X},\boldsymbol{X}_*) \\ K(\boldsymbol{X}_*,\boldsymbol{X}) & K(\boldsymbol{X}_*,\boldsymbol{X}_*) \end{bmatrix}\right) \tag{41}$$

Here, $\boldsymbol{I}$ is the identity matrix. Equivalent to the noise-free case, it is possible to express the joint Gaussian distribution as a Gaussian process. Here, the resulting Gaussian process has the mean function

$$m_{post}(\boldsymbol{x}) = K(\boldsymbol{x},\boldsymbol{X})[K(\boldsymbol{X},\boldsymbol{X}) + \sigma_\epsilon^2 \boldsymbol{I}]^{-1} \boldsymbol{y} \tag{42}$$

and the covariance function

$$k_{post}(\boldsymbol{x},\boldsymbol{x}') = k(\boldsymbol{x},\boldsymbol{x}') - K(\boldsymbol{x},\boldsymbol{X})[K(\boldsymbol{X},\boldsymbol{X}) + \sigma_\epsilon^2 \boldsymbol{I}]^{-1} K(\boldsymbol{X},\boldsymbol{x}'). \tag{43}$$

Therefore it is possible to perform Gaussian process regression with noisy data using equations (42) and (43), if the level of noise σ_ϵ is known [46].

3.4 Fitting kernel functions

As described in section 3.2 the mean function of a Gaussian process posterior can be used as a model of a wanted objective function. However, there are several different possible posteriors for a single set of data. This comes from the fact that the resulting posterior does not only depend on the data set, but also on the chosen kernel and the choice of hyperparameters in particular. This is demonstrated in figure 6 using an exemplary data set. The subfigures of figure 6 were created using the same set of data with gaussian noise and the squared exponential kernel. For each subfigure the hyperparameter lengthscale ℓ was set to a different value. The individual value can be found above the corresponding subfigure. With these settings, Gaussian process regression was performed, anticipating the same level of gaussian noise as was used for the creation of the data set. As one can see in the three subfigures, the choice of the hyperparameter ℓ has a big impact on the mean function (displayed in dark red). The mean function of figure 6b ($\ell = 1$) seems to capture the overall trend of the data best. It touches almost all of the data points and every point lies within the claimed 2σ confidence region (red shadow). The mean function displayed in figure 6a ($\ell = 0.2$) touches every data point, but it does not seem to reproduce the overall trend of the data points as well as figure 6b. The mean function of figure 6c ($\ell = 5$) does not touch most of the points and also does not seem to recreate the overall trend of the data.

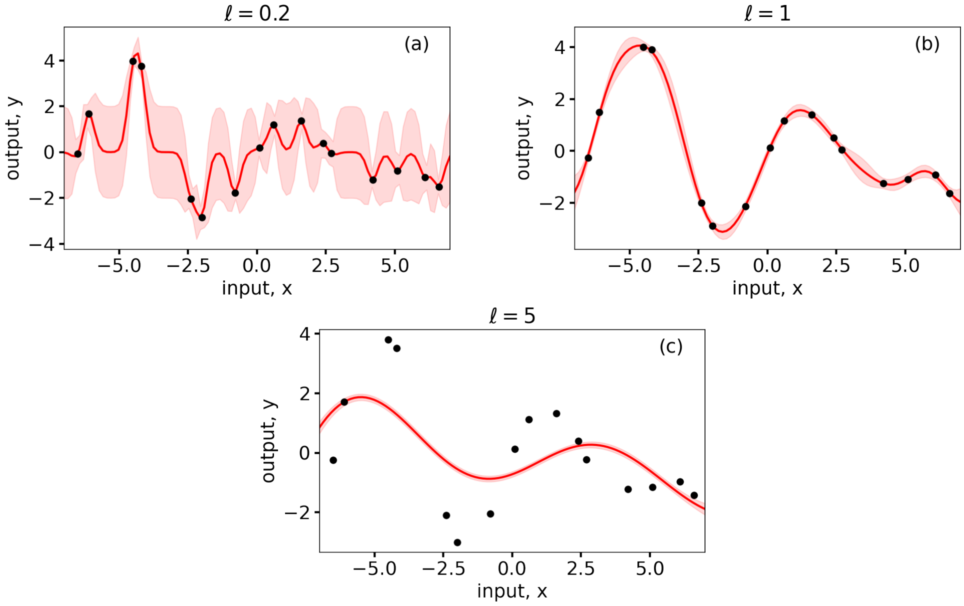

Figure 6: Gaussian process posterior of set of points using the squared exponential kernel and three different values for the hyperparameter lengthscale ℓ. The explicit values are written directly above the subfigures. In each subfigure the mean function and the 2σ confidence region are displayed.

The examples displayed in figure 6 clearly shows that it is important to optimize the hyperparameters of a kernel function in order to obtain the best possible model. Within BOSS, this is done by maximizing the marginal likelihood $p(\boldsymbol{y}|\boldsymbol{X}, \theta)$, where θ is the collection of hyperparameters of the kernel and $(\boldsymbol{X}, \boldsymbol{y})$ is the observed set of data. The optimization is performed using a vectorized random scoring of candidate solutions followed by a single conjugate gradient refinement [49].

3.5 Acquisition functions

As explained above, Gaussian process regression can be used to obtain reliable surrogate models based on limited data sets, even if nothing else is known about the underlying function. One of the advantages of Gaussian process regression compared to other comparable methods is that it gives information about the calculated uncertainty of the model for every point of the input domain. This information can be used to define an element of the input domain that is “most interesting”. Defining this “point of greatest interest” is an important part of Bayesian optimization. In general one can define Bayesian optimization as an iterative scheme that starts by performing Gaussian process regression with the already known data points. Then, the information of the current surrogate model and the uncertainty at every point is used to find the “most interesting point”. The unknown objective function is then evaluated at this point. After adding the new data point to the data set, the new model is calculated with the updated data set. This procedure can then be iterated until convergence.

Of course there are several different thinkable strategies that one can use to define the “most interesting” element of the input domain. It is important to note that this choice can have great impact on the performance of the Bayesian optimization approach, since it determines which input values are used for the evaluation of the objective function. In general these values are determined by assigning an “interestingness” value to every element of the considered input domain and then choosing the extremum. The functions that are used to assign that property to every element are called *acquisition functions*.

Very often the objective that one has when using Bayesian optimization is finding the global minimum of an unknown function. In these cases it makes sense to acquire more information about the regions, where the global minimum is expected to be based on the already acquired data. This strategy is called *exploitation*. However, it may also be advantageous to acquire information about regions that have not been sampled much yet in order not to miss the global minimum. This strategy is called *exploration*. If an acquisition function favors exploration too much, it may take long to find the global minimum. If an acquisition function favors exploitation too much, the global minimum may be missed. It is therefore important to find a good compromise between both strategies.

A rather simple, but often powerful acquisition function that is also embedded in BOSS is the *explore* acquisition function. It can be defined in the following way:

$$\mathcal{A}_t^{explore}(x) = -\sigma_t(x) \tag{44}$$

Here, $\sigma_t(x)$ is the variance calculated using equation (39) for input value x at iteration step t. This acquisition function is called explore acquisition function, because it fully incorporates the exploration strategy and does not follow the exploitation strategy. It is minimized by the input value that has highest uncertainty. This minimizing value can then be used as an input value for the next data acquisition. The explore acquisition function is most advantageous when there is no interest in special points like the global minimum. Instead it has its strength in efficiently modeling the overall trend of the objective function.

The explore acquisition function is not a good choice when one is looking for the global minimum of an objective function. Instead it is more advisable to employ an acquisition function that also uses the earlier mentioned exploitation strategy. One such acquisition function is the *lower confident bound* (elcb) acquisition function. It can be defined as

$$\mathcal{A}_t^{elcb}(x) = \mu_t(x) - \eta_t \sigma_t(x), \tag{45}$$

where $\mu_t(x)$ is the mean function evaluated at iteration step t evaluated at x, $\sigma_t(x)$ again is the variance at iteration step t evaluated at x and η_t is a parameter that controls the level of exploration. The influence of the parameter η_t can be seen in figure 7.

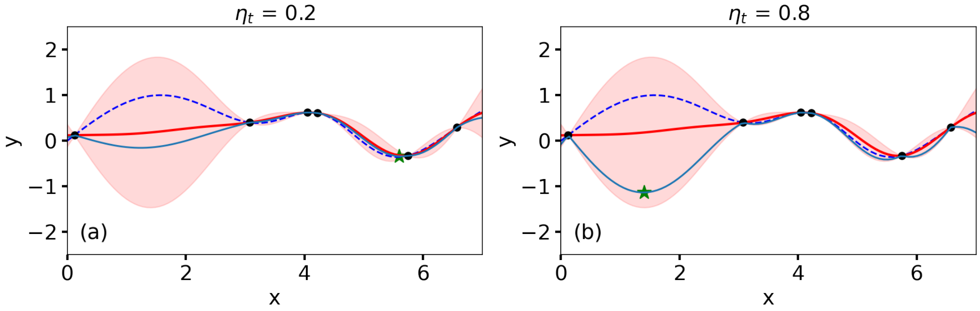

Figure 7: Depicted in the two images is the same Gaussian process posterior with the lower confidence bound acquisition function. The blue dotted line shows the objective function. The mean function is shown as a solid dark red line and the confidence area is shown as a light red region. The black dots represent the data set used for the Gaussian process regression. In figure a the parameter η_t is set to 0.2, whereas in figure b it is set to 0.8. The minima of the two acquisition functions are mareked with a green star, respectively. Because of the different values of η_t, the minima have different positions.

Although both images visible in the figure show the Gaussian process regression of the same data set using the same kernel function, the proposed position for the next data acquisition (green stars) differs. This difference comes from the different choice for the parameter η_t. In figure 7a, η_t is set to 0.2, whereas in figure 7b it is set to 0.8. The higher value used in the right figure leads to an increased significance of the areas with high uncertainty thus resulting in a different minimum of the acquisition function. In BOSS, the parameter η_t is set by default to $\eta_t^2 = 2\log[t^{d/2+2}\pi^2/(3\epsilon_\eta)]$, where d is the number of input dimensions and ϵ_η is a small positive number (usually 0.1) [50]. This approach is called *exploratory lower confidence bound* (elcb). It is the standard acquisition function within BOSS. With the given definition, η_t increases with increasing number of iteration steps. As a consequence, the strategy of exploration is used more and more with an increasing number of iterations. The advantage of this approach is that the algorithm can not “get stuck” at a local minimum by ignoring regions with high uncertainty. The acquisition functions lower confidence bound and exploratory lower confidence bound have their strength in efficiently finding the global minimum of an objective function.

3.6 Symmetries

In section 3.1.2 the standard periodic kernel was discussed. It makes sense to use this kernel if it is known that the objective function possesses a certain periodicity. The surrogate model will then possess the same periodicity at every iteration step. There is another, seemingly more simple way to impose a symmetry on the surrogate model that is not limited to periodicities. The idea is the following: If a function has any kind of symmetry, for every element of the input space x there must exist one more elements $x', x'', ...$ that are the symmetric images of the original element x. the elements $x', x'', ...$ must then have the same function value as x. This means:

$$f(x) = f(x') = f(x'') = ... \tag{46}$$

It is possible to use this principle within the BOSS program. When the objective function is evaluated at position x, the function value $f(x)$ will then not only be assigned to this position, but also to the symmetric images of x. Any kind of symmetry can be defined within BOSS. Figure 8 shows an example of the use of this feature. Here, the objective function has a mirror symmetry with a mirror

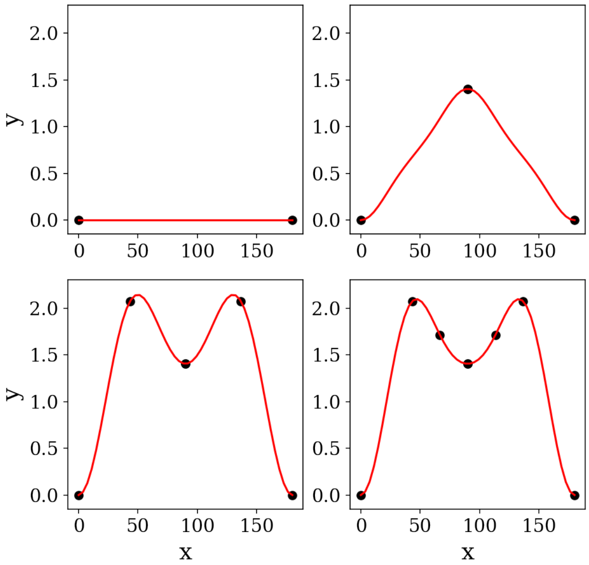

Figure 8: Four iterations of a Bayesian optimization run. The objective function possesses a mirror symmetry with a mirror axis at $x = 90$. Due to this symmetry two data points can be added to the data set at each iteration step.

axis at $x = 90$. This symmetry is then imposed on the surrogate model by defining

$$x' = 180 - x. \tag{47}$$

In figure 8 four iteration steps of Bayesian optimization can be seen. With each step, the original new data point and its mirror image can be added to the data set. With this procedure the surrogate model always has the wanted mirror symmetry.

3.7 Performance evaluation

Due to the stochastic nature of Gaussian processes and noisy data acquisitions, no Bayesian optimization surrogate model can assumed to be perfect. However, the difference between the surrogate function and the objective function can be expected to generally converge towards a low value. Since performing a large number of iteration steps to reach this low value would contradict the main purpose of Bayesian optimization, namely being computationally effective, it is important to define a certain measure of convergence of the surrogate model.

There are several possible measures that can be used to construct a convergence criterion. Generally two types of convergence criteria can be distinguished. The first type depends on solid knowledge about the true objective function. If the underlying objective function is known, it is possible to define a loss function that determines the difference between the current and the true model. The second type of convergence parameters does not depend on knowledge about the objective function. Instead, these convergence measures are based on the change of the surrogate model per iteration step. However, if the true model is unknown, it is often difficult to find a guarantee that a model is converged. This is why convergence criteria of the second type are generally less reliable.

Let us first look at the first type of convergence parameters. A versatile loss function that can be used as such a convergence criterion is the mean squared error (MSE). The MSE can be defined as follows:

$$MSE((x_n, y_n)_{n=1}^{N}, t) = \frac{1}{N} \sum_{n=1}^{N} (y_n - f_t(x_n))^2 \tag{48}$$

Here, $(x_n, y_n)_{n=1}^{N}$ is a set of N data points that can be assumed to be correct and f_t is the model at iteration step t. In this work, literature data were assumed to be correct. Therefore, the MSE measures the mean square deviation of the surrogate model from literature data in this case.

Often the position of the global minimum or the global maximum of a function are especially interesting. Hence, another meaningful approach to build a convergence criterion is to track the distance between the extremum position of the surrogate model an true extremum position.

There are two possible ways to define a threshold for the convergence criterion of the first type. One way is to choose an acceptable distance between the current model and the true model. The other way is to define a threshold for the difference of the convergence indicator between two consecutive iteration steps. This second way can also be used for the second type of convergence indicators.

Within the scope of this thesis convergence criteria of this first type are especially useful since a goal of the thesis is to meaningfully compare Bayesian optimization to more classical approaches. In section 5 we will therefore compare models acquired with Bayesian optimization to literature data, which we can assume as true.

However, performing Bayesian optimization on a function that is completely known is obviously not useful for materials research. In fact, Bayesian optimization is best used if there is close to or completely no knowledge about the system under investigation. It is therefore crucial to have convergence criteria that do not depend on prior knowledge about the objective function. These convergence criteria make the the second type of convergence criteria. A helpful convergence criterion of the second type that is specific to Bayesian optimization is based on the convergence of the hyperparameters of the kernel function. As mentioned in section 3.4, the hyperparameters of the kernel functions can be optimized for a specific set of data. Since the collection of acquired data grows with each iteration step of Bayesian optimization, it is reasonable to update the hyperparameters regularly. In this work, the hyperparameters were always optimized at every iteration step. The convergence behavior of the hyperparameters can therefore be used as an indicator for the convergence of the surrogate model. However, it is important to be careful with the interpretation of this convergence measure. As the example in figure 9 shows, it is possible that two surrogate models agree very well although the hyperparameters of the corresponding covariance functions have different values. Both models shown in the figure coincide well. Plotting them in one single graph would make it difficult to distinguish them by naked eye. Their mean squared deviation is $2.8 \cdot 10^{-4}\,\mathrm{meV}^2$. Despite this good agreement the hyperparameters of the used squared exponential kernel differ significantly. The exact value of the hyperparameters are indicated in the inset of the corresponding figure. This example shows that a good agreement between two models created using Gaussian process regression does not necessarily mean that the used kernel hyperparameters

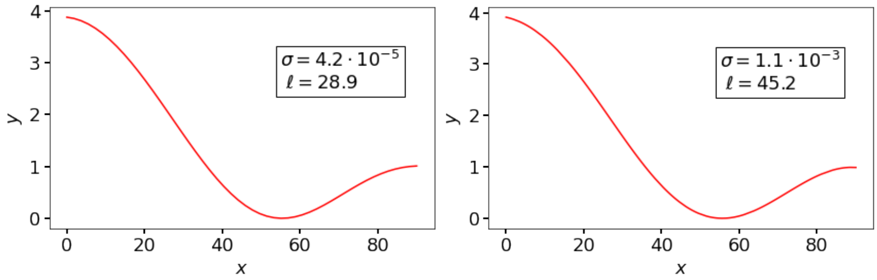

Figure 9: Two surrogate models created with Gaussian process regression. Despite the different hyperparameter values (shown as insets in the two figures) the models agree very well.

have a similar value. Concluding one can say that although the convergence of hyperparameters can be a useful indicator for convergence, it should not be seen as an absolute measure.

One goal of this work is to compare functions that are known from literature to models that were acquired using Bayesian Optimization within the scope of this work. These functions describe the total energy of a material as a function of a certain degree of freedom. This degree of freedom often describes the change of the magnetization direction of the material (see section 4.2). However, within this thesis, we additionally study even broader parts of the magnetic configuration space of certain benchmark materials by taking into account more degrees of freedom. This incremented number of degrees of freedom is reached by varying the direction of the magnetic moments of certain groups of atoms separately from each other. Due to the lack of literature data that can be used for comparison, it makes sense to use a convergence criterion of the second type in these cases. If a model is converged sufficiently, it should not change much if more data points are added to the data set. Thus the squared difference between two consecutive models can be used as a convergence indicator. Such an indicator can be defined as

$$MSE_{step}((y_n^{t-1})_{n=1}^N, (y_n^t)_{n=1}^N) = \frac{1}{N} \sum_{n=1}^{N} \left(y_n^{t-1} - y_n^t\right)^2 , \tag{49}$$

where $(y_n^t)_{n=1}^N$ is the surrogate model at iteration step t, which consists of N data points. In this work, M_{step} is called M_{2D} and M_{4D} respectively, according to the number of degrees of freedom that were taken into account.

The question if a model is converged sufficiently always depends on the task that one wants to accomplish. If one is mainly interested in finding a global energy minimum, it might be enough to look at the convergence behavior of the current global minimum. If one is interested in more detailed features of the total energy as a function of the magnetic configuration, it might for example be more important to look at the convergence behavior of the kernel hyperparameters.

4 Magnetic Materials

4.1 Magnetism basics

On a quantum level, there are two main reasons for the existence of magnetic moments in atoms, electron spin and electron angular momentum [51, p.669]. The magnetic moment of one electron that results from the electron's spin μ_S is described by the formula

$$\boldsymbol{\mu}_s = -g_s \mu_B \frac{\boldsymbol{s}}{\hbar}, \tag{50}$$

where g_s is the electron spin g-factor, $\boldsymbol{s}$ is the three-component spin angular momentum, $\hbar$ is the reduced Planck constant, $\mu_B = e\hbar/2m_e$ is the Bohr magneton, e is the elementary charge and m_e is the mass of an electron. Another cause for magnetic moments in atoms is the electron angular momentum. The magnetic moment of an electron caused by this can be written as

$$\boldsymbol{\mu}_l = -g_l \mu_B \frac{\boldsymbol{l}}{\hbar}. \tag{51}$$

$\boldsymbol{l}$ Is the three-component vector of the quantized orbital angular momentum and g_l the corresponding g-factor.

An effect that is important for the understanding of many magnetic materials is spin-orbit coupling (SOC). SOC is a relativistic effect that describes the coupling between the spin $\boldsymbol{s}$ of an electron and its orbital angular momentum $\boldsymbol{l}$. An expression of the SOC energy can be obtained in the small-velocity limit of the Dirac equation [52], which fulfills the requirements of special relativity (for a derivation, see *e.g.* [53]). Furthermore, it can be derived in a semi-classical picture as a relativistic correction to the Schrödinger equation [51, p.707], that also provides a physical intuition of the effect. However, it has to be noted that the Schrödinger equation is not fully relativistic, even with the mentioned correction term.

Let us consider an electron orbiting a nucleus with charge Ze. In an inertial rest frame comoving with the electron, the nucleus appears to be orbiting the electron. This moving charge, which is nothing else than a circular current I induces a magnetic field $\boldsymbol{B} = (\mu_0 I/2r)\hat{\boldsymbol{n}}$ where μ_0 is the vacuum permeability. The unit vector $\hat{\boldsymbol{n}}$ is orthogonal to the area that is enclosed by the circular orbit of the nucleus. The classical energy of this system can be expressed as

$$E_{SOC} = -\boldsymbol{\mu}_s \cdot \boldsymbol{B}. \tag{52}$$

Using the general definition $\boldsymbol{\mu}_l = \pi r^2 I \hat{\boldsymbol{n}}$ and the definition $\boldsymbol{\mu}_l = (Ze/2m_e)\boldsymbol{l}$, where Z is the number of protons in the nucleus and m_e is the mass of the electron, the magnetic field that interacts with the spin magnetic moment can be expressed as

$$\boldsymbol{B} = \frac{\mu_0 \boldsymbol{\mu}_l}{2\pi r^3} = \frac{\mu_0 Ze}{4\pi m_e r^3}\boldsymbol{l}. \tag{53}$$

Now using equation (50) and $\mu_0 = 1/\epsilon_0 c^2$ with the vacuum permittivity ϵ_0 and the speed of light c the contribution of spin-orbit coupling to the total energy can be quantified as

$$E_{SOC} = -\boldsymbol{\mu}_s \cdot \boldsymbol{B} \approx \frac{Ze^2}{4\pi\epsilon_0 m_e^2 c^2 r^3} \boldsymbol{l} \cdot \boldsymbol{s}, \tag{54}$$

assuming $g_s \approx 2$. Since we have that $\langle r^{-3} \rangle \propto Z^3$ [54, p. 210], the spin-orbit interaction energy is proportional to Z^4. Hence, SOC is especially relevant for heavy atoms.

In the absence of a spin-orbit interaction, the energy eigenstates can be labeled by the combined angular momentum of all electrons in the atom $\boldsymbol{L}$ and the combined spin of all electrons in the atom $\boldsymbol{S}$. With spin orbit coupling, these eigenstates split into a number of energy labels with differing J, that can take values between $|L - S|$ and $L + S$. This is known as *fine structure* splitting. Each of these levels has degeneracy $2J+1$, that can be further split by applying a magnetic field [54, p.31 *et seqq.*].

Until now we have treated spin-orbit coupling as a weak perturbation and the main energy was mainly determined by the electrostatic interactions that control the values L and S. This model is called *L-S coupling* or *Russell-Saunders coupling*. For atoms with high atomic number Z the spin-orbit interaction energy is dominant, since it is proportional to Z^4, as mentioned above. Thus for heavy atoms it works better to use the so-called *j-j coupling* scheme, where the spin and orbital angular momentum of each electron are coupled separately.

4.2 Noncollinear magnetism

When there are multiple atoms with non-vanishing magnetic moments present in a material, the magnetic moments tend to orient in a collective pattern when the temperature is low enough. In general, there are several different magnetic orderings thinkable. In many materials the magnetic moments of the atoms are parallel or anti-parallel with respect to each other. This behavior is called *collinear* magnetism. A parallel ordering is called *ferromagnetism* (FM), an antiparallel ordering is called *antiferromagnetism* (AFM) (see figure 10). Iron, for example, is a ferromagnetic material, whereas manganese(II) oxide shows antiferromagnetic behavior. However, there are materials that show neither ferromagnetic nor antiferromagnetic ground states. Instead the magnetic moments are tilted by a certain

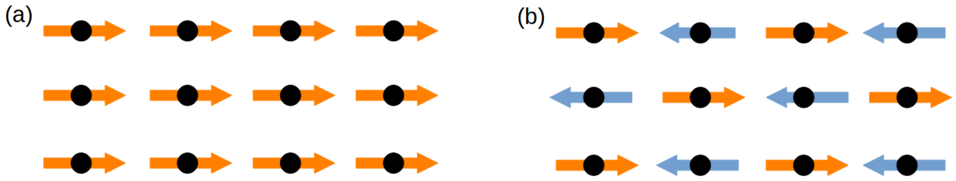

Figure 10: Sketch of the collinear magnetic orderings ferromagnetism (figure a) and antiferromagnetism (figure b).

angle, the so-called *canting angle* Φ (see figure 11a). This behavior is called noncollinear magnetism. A reason for a noncollinear magnetic ground state can for example be strong spin-orbit coupling. The exact magnetic ground sate of such a noncollinear magnet can be found with computational approaches like DFT by varying the canting angle Φ and searching for the energy minimum. A typical curve that describes the magnetic canting energies $\Delta E(\Phi)$ could look like figure 11b. In this work, we use the term *magnetic energy curve* for one-dimensional functions like the one shown figure 11b, that model the dependence of the total energy on one or more canting angles. In the case of multi-dimensional functions, we use the term *magnetic energy landscape*. Often there is a distinction made between canted ferromagnetism (cFM) and canted antiferromagnetism (cAFM). When the magnetic moments of the atoms are closer to being parallel than antiparallel ($\Phi < 45°$) the state is referred to as cFM. The opposite case is referred to as cAFM.

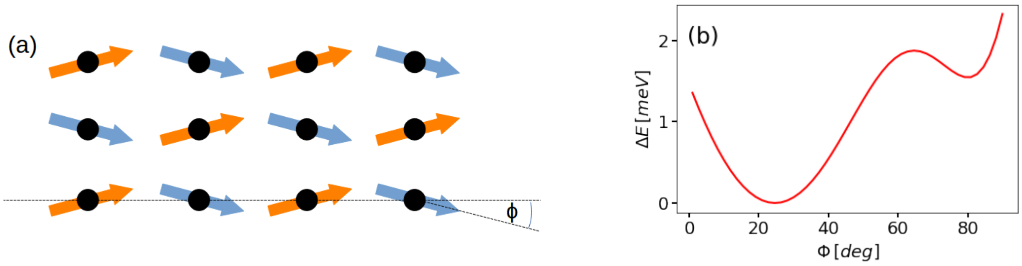

Figure 11: Canted antiferromagnetism as an example of noncollinear magnetism. The magnetic configuration of noncollinear magnets is sketched in figure a. An example of a curve that shows the total energy as a function of the canting angle is shown in figure b.

4.3 Materials

In this work, Bayesian optimization is used to create models that describe the dependence of the total energy on a canting angle for three different materials. The goal is to compare these results to literature data that were acquired without the help of Bayesian optimization to determine if and how much Bayesian optimization improves the efficiency of the calculations. The three materials that were used for the investigation, Ba_2NaOsO_6, UO_2 and Sr_2IrO_4 are introduced here.

4.3.1 Ba_2NaOsO_6

Ba_2NaOsO_6 (BNOO) is a transition metal double perovskite that shows strong spin-orbit coupling and electronic correlation effects [55]. BNOO is a $5d^1$ Mott insulator [56] with an effective total angular momentum $J_{eff} = 3/2$ [55]. Below the Curie-Temperature $T_C = 6.8$K it shows an ordered spin moment of $\sim 0.2\mu_B$ [57]. The sodium atoms and osmium atoms are located at octahedral sites. This means that they are in the center of the oxygen octahedra NaO_6 and OsO_6 respectively [58]. Figure 12a shows the simplified crystal structure of BNOO, where the barium and sodium ions have been omitted to provide a more clear picture. The red and blue arrows that can be seen in figure 12 indicate the magnetic moments of each osmium atom. The magnetic state shown in the figure is the magnetic ground state according to an experimental nuclear magnetic resonance (NMR) study [59]. As the figure shows there are two magnetic sites in BNOO. They define two sublattices (red plane and blue plane), that in itself are ferromagnetic. According to the mentioned NMR study, the magnetic moments of the two sublattices align with a 67° angle relative to the [110] axis, which is the easy axis of BNOO [61].

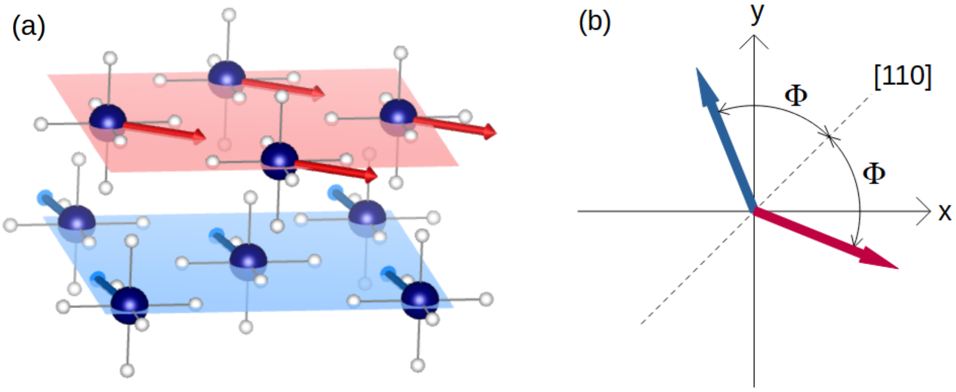

Figure 12: Figure a shows the simplified crystal structure of BNOO omitting barium and sodium atoms. The arrows indicate the magnetic ground state of BNOO according to experimental results [59]. Figure b describes the magnetic ground state of BNOO in more detail and defines the canting angle Φ with respect to the easy axis [110]. The figures are inspired by [60].

Fiore Mosca *et al.* [60] showed that this experimentally verified ground state can not be reproduced with first principles methods using the pristine, undistorted atomic configuration of BNOO. Instead, the canted noncollinear ground state of BNOO is concomitant with Jahn-Teller distortions of the OsO_6 octahedra. All Os-O bond lengths are changed due to the distortions. In fact, in the distorted lattice, the lengths of the in-plane Os-O bonds depend on the corresponding lattice plane. In figure 13 the relevant bonds are sketched to illustrate the distortions. Following the notation of Van Vleck [62], the distortions can be classified as $Q_1 = 0.031$ and $\tan^{-1}(Q_2/Q_3) = 79.35°$ or $\tan^{-1}(Q_2/Q_3) = -79.35°$ for the two planes, respectively. In this work, the BNOO crystal without the Jahn-Teller distortions is referred to as *undistorted BNOO*. The BNOO crystal that holds the mentioned Jahn-Teller distortions, on the other hand, is referred to as *distorted BNOO*.

According to Fiore Mosca *et al.* the total energy as a function of the canting angle Φ as defined in figure 12b can be described by the following formula:

$$E(\phi) = A_1 \cos(2\phi) + A_2 \cos(4\phi) + A_3 \cos(6\phi) + B_1 \sin(2\phi) + B_2 \sin(4\phi) \tag{55}$$

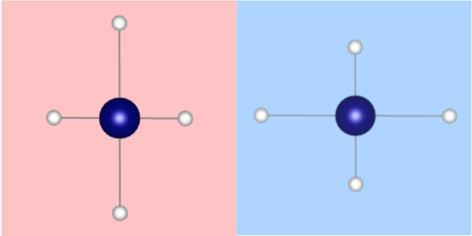

Figure 13: Exaggerated distortions of the octahedra in BNOO. According to Fiore Mosca *et al.* these distortions are necessary to reproduce the experimentally verified magnetic ground state of BNOO. The figure is inspired by [60].

The detailed derivation of this formula can be found in the supplementary material of [60]. Formula (55) describes the total energy as a function of the canting angle for both the undistorted and the distorted BNOO. In the case of the undistorted BNOO, the coefficients B_1 and B_2 vanish.

For the DFT calculations of this work, a Coulomb interaction potential of $U = 3.4$ and a 6x6x6 k-mesh have been used in accordance with the procedure of [60]. Also in accordance with this reference, energies and energy differences of BNOO in this work are given for two formula units.

4.3.2 UO$_2$

Despite having a fairly simple crystal structure uranium dioxide shows complex magnetic behavior [63]. It shows interesting magnetic phases and phase transitions at finite temperatures and also shows an interesting magnetic ground state below the transition temperature of $\sim 30\,\mathrm{K}$ [64]. At this temperature, uranium dioxide possesses four different magnetic sites and a noncollinear ground state. This has been shown experimentally [65, 66] and was confirmed computationally using first principles methods [67]. The magnetic ground state of UO$_2$ is a so-called 3-k ground state that can be seen in figure 14 [65, 68–71]. In the 3-k configuration the canting angle Φ that is defined in the figure has the value $54.7°$. Dudarev *et al.* [72] were able to recreate the magnetic ground state configuration of uranium dioxide with first principles methods. Furthermore they derived an expression for the total energy of uranium dioxide as a function of the canting angle Φ as defined in figure 14:

$$\begin{aligned}\Delta E(\Phi) =& B_0 + B_1\cos(\Phi) + B_2\cos(2\Phi) + B_3\cos(3\Phi) + B_4\cos(4\Phi) + \\ & C_1\sin(\Phi) + C_2\sin(\Phi) + C_4\sin(4\Phi)\end{aligned} \tag{56}$$

The factors B_i and C_i are specified in [72]. In accordance with their computational procedure, a 6x6x6 k-mesh and the Coulomb interaction potential $U = 3.46\,\mathrm{eV}$ were used in the work of this thesis work for all DFT calculation. In the mentioned reference, the energy differences are given for one formula unit. In order to gain the same results, this is also done in this work.

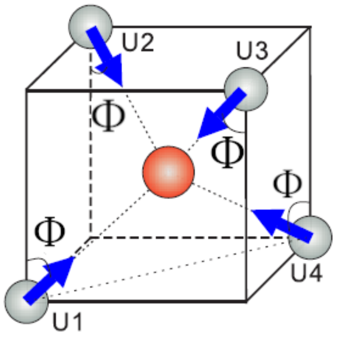

Figure 14: Crystal structure of uranium dioxide. The four gray atoms are uranium atoms and the red atom is the oxygen atom. The blue arrows indicate the vector of the magnetic moments. Shown is the magnetic ground state of uranium dioxide at a temperature of $0\,\mathrm{K}$. The figure originates from Dudarev *et al.* [72].

4.3.3 Sr_2IrO_4

Sr_2IrO_4 is an interesting and well-studied iridate and belongs to the Ruddelsden-Popper series. It attracted special attention because it was the first reported case where the interplay between electron correlation and strong spin orbit coupling leads to the opening of a Mott band gap in a $J_{eff} = 1/2$ band [73, 74]. This makes it a $J_{eff} = 1/2$ Dirac-Mott insulator. J_{eff} refers to the effective total angular momentum.

The crystal lattice of SIO can be seen in figure 15a. The atoms with non-vanishing magnetic moments are the Ir atoms. As the figure shows, they are located at the center of IrO_6 octahedra. Experimental results [76] indicate that the IrO_6 octahedra are tilted around the c-axis with an angle of $\alpha = 11.5°$. On the other hand, DFT calculations suggest a value of $\alpha = 13.2°$ [75]. For this work,

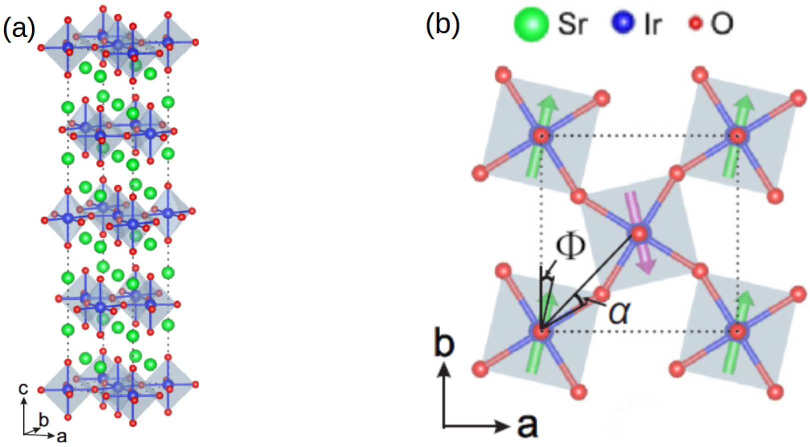

Figure 15: Lattice structure of Sr_2IrO_4. Figure a shows the side view and figure b shows the top view. Figure b also indicates the definition of the rotation angle α and the canting angle Φ. The figures were taken from [75] and adapted.

α was fixed at 13.2°. Figure 15b also shows the definition of the magnetic canting angle Φ. Experiments suggest that in the ground state $\Phi = 12.2°$ [77]. In the DFT study that serves as a comparison in this work [75], the canting angle is given by $\Phi = 12.3°$. According to [75], the total energy can be modeled as a function of the canting angle Φ by the following equation:

$$\Delta E = 16JS^2\cos(2\Phi) + 8K\cos(4\Phi) - 16D_zS^2\sin(2\Phi) \tag{57}$$

Here, J is the nearest neighbor isotropic exchange, K is the single-site anisotropy and D_z is the Dzyaloshinskii-Moriya coupling parameter.

For the DFT calculations of this work on SIO, a Coulomb interaction potential $U = 1.8\,\mathrm{eV}$ and a 7x7x3 k-mesh have been chosen in accordance with the procedure used by the study that serves as a comparison to the results of this thesis [75]. In accordance with this reference, the energy differences of this work are given for eight formula units.

5 Results

In this work, the BOSS program is used to test if Bayesian optimization works as a tool to build magnetic curves for different magnetic materials. Therefore, a bash-script was set up that connects BOSS with VASP. The purpose and functionalities of the script are sketched in figure 16. This script receives information from BOSS about the magnetic configuration that is determined as most interesting by the acquisition function. This information is generally expressed in the form of one or more canting angles. The script then translates this information into magnetic moment vectors for each atom of the material. Afterwards it generates the necessary input files for VASP and copies them into a newly created directory, where the DFT calculation is started subsequently. After the termination of the calculation, the script reads the calculated total energy from the relevant VASP output file and returns it to BOSS. With this new information, BOSS determines the next interesting magnetic configuration according to the procedure described in section 3.

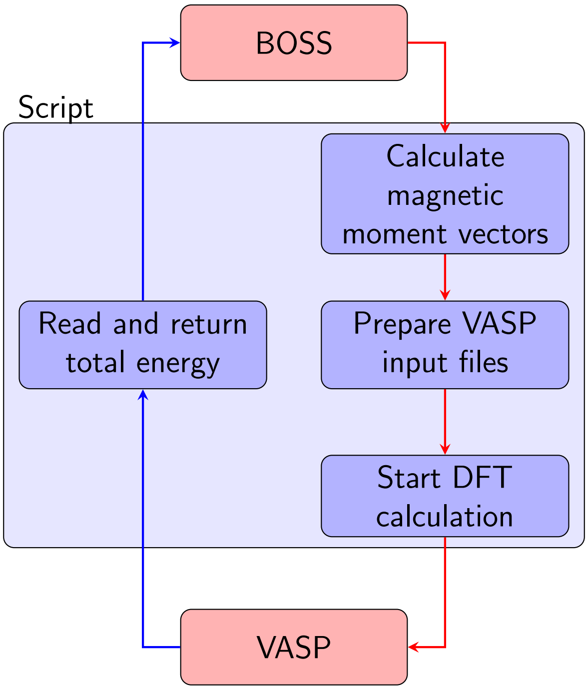

Figure 16: Sketch of the functionality of the bash script that connects BOSS with VASP. Information about the wanted magnetic configuration is transported by the script from BOSS to VASP (red arrows). After the DFT calculation, the calculated total energy is read and returned to BOSS (blue arrows).

5.1 One-dimensional models

An important goal of this work is to recreate existing models of the one-dimensional magnetic energy curve, with the canting angle Φ defined individually for each material. All Bayesian optimization models shown in the section 5.1 were created using the squared exponential covariance function and the exploratory lower confidence bound acquisition function. These functions are used in BOSS as default settings and were chosen for all models in this section, so that the performance of BOSS on the different materials can be compared easily.

5.1.1 Undistorted Ba_2NaOsO_6

The first material under investigation is Ba_2NaOsO_6 in its undistorted phase, which means that it does not hold the Jahn-Teller distortions mentioned in section 4.3.1. Figure 17 compares the existing literature data from [60] (blue points) with a model constructed by BOSS (red line) after the acquisition of seven DFT calculations. The acquired data points that were used to create the model are depicted as green stars. Note that for the literature curve 36 DFT calculations were performed. Despite the difference in the number of performed DFT calculations the results agree well. The question of a meaningful convergence criterion can be answered in different ways, as already mentioned in section 3.7. One way to determine the convergence of a Bayesian optimization model is the convergence of the kernel hyperparameters σ and ℓ. Figure 18 gives details of the evolution of the hyperparameter values as a function of the number of iterations. Note that the figures concerning the variance σ (18a and 18b) are displayed using a logarithmic scale to give a more clear picture. Taking a look at absolute values of the hyperparameters (figures 18a and 18c) gives a good intuition for their convergence. It becomes clear that both hyperparameters show solid convergence behavior meaning that the absolute values of the hyperparameters change only minimally after a number of iterations. The latest hyperparameter values are $\sigma_{\text{latest}} = 7.49 \cdot 10^{-3}$ eV and $\ell_{\text{latest}} = 17.32°$. The figures 18b and 18d show the change of the absolute values of the hyperparameters between two consecutive iteration steps, $\Delta\sigma$ and $\Delta\ell$, respectively. After the tenth data acquisition the change of the hyperparameter

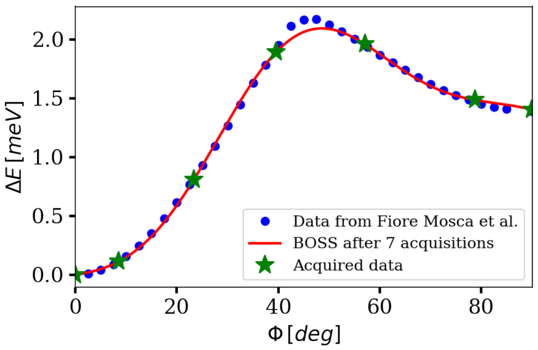

Figure 17: One-dimensional magnetic energy curve of undistorted BNOO. Shown is the literature data from Fiore Mosca *et al.* [60] (blue points) as well as the model created by the BOSS program after seven DFT calculations (red line). The data points that were acquired in order to construct the model are shown as green stars.

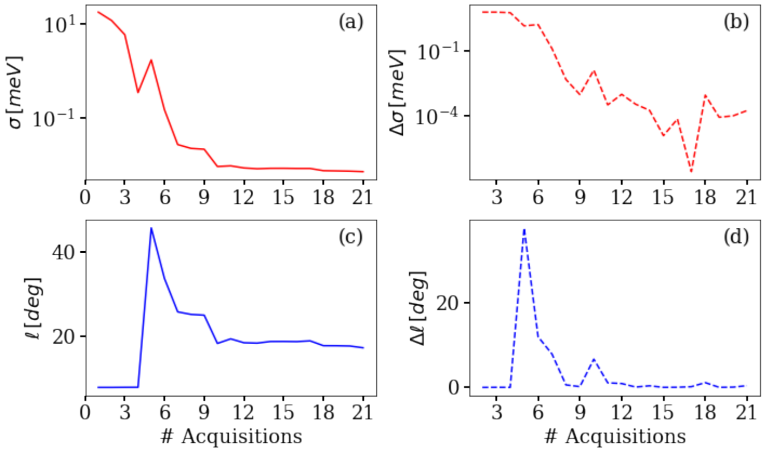

Figure 18: Evolution of the hyperparameters σ (figures a and b) and ℓ (figures c and d) for 1D BNOO. Figures a and c show the absolute value of the hyperparameters after a given number of acquisitions. Figures b and d show the difference between two iteration steps.

σ is always less then 10^{-3} meV between two consecutive iteration steps. Furthermore, the change of the hyperparameter ℓ is always less then after the tenth data acquisition 2°. For both σ and ℓ the change of the hyperparameter between two consecutive steps is multiple orders of magnitude lower than its absolute value after only a few data acquisitions.

Concerning the overall shape of the magnetic energy curve, another meaningful convergence measure is the mean squared error (MSE) of the model created by BOSS with respect to the model proposed by Fiore Mosca *et al.* [60], which was assumed to be correct. The evolution of the MSE with each iteration step can

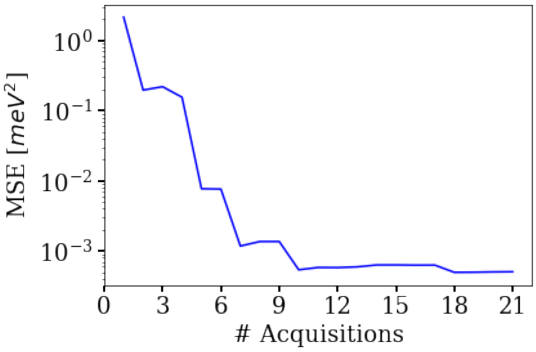

Figure 19: Evolution of the mean squared error of the surrogate model that was built for the magnetic energy curve of undistorted BNOO using BOSS.

be seen in figure 19. The MSE drops and stays below $10^{-2}\,\mathrm{meV}^2$ after the fifth DFT calculation. This can already be seen as a satisfactory value. Subsequently, the MSE decreases even further and reaches values below $10^{-3}\,\mathrm{meV}^2$ after the acquisition of ten DFT data points.

5.1.2 Distorted Ba_2NaOsO_6

Jahn-Teller distortions have influence on the magnetic curve of BNOO. Only slight distortions can even lead to the formation of "new" local minima. This effect was reported in the work of Fiore Mosca *et al.* [60]. With certain Jahn-Teller distortions described in section 4.3.1, Fiore Mosca *et al.* were able to recreate and explain the experimentally observed magnetic ground state that is characterized by a canting angle of 67°. In figure 20a the 36 DFT data points that were acquired and used by Fiore Mosca *et al.* can be seen as well as a fit of this data using the fitting function (55), which was also used in their work. Figure 20b shows the BOSS surrogate model after the acquisition of nine DFT data points. The acquired data is also displayed in the figure. Figure 20 shows that also in the case of distorted BNOO the two curves appear to be very similar despite a different numbers of used DFT calculations. Even in a direct comparison (figure 21) the two model appear to agree very well. The two models shown in figure 20 have a mean squared

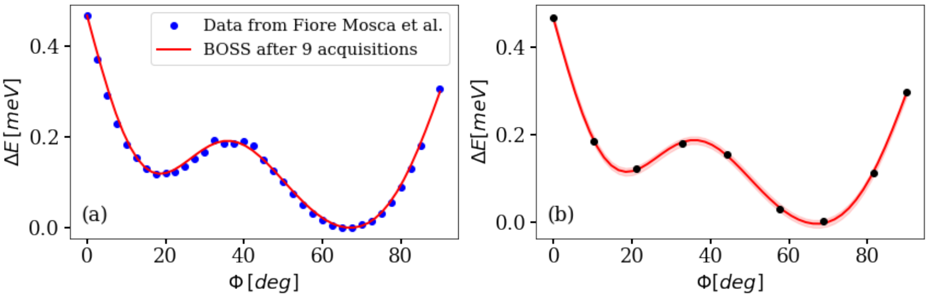

Figure 20: One dimensional magnetic curve of distorted BNOO. Figure a shows the literature data given in [60]. Figure b shows the surrogate model given by BOSS after nine data acquisitions.

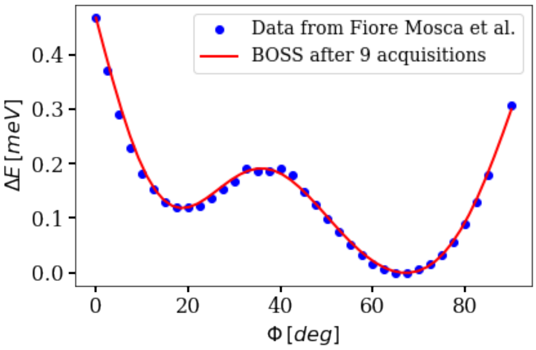

Figure 21: Direct comparison between the literature data (blue dots) and the model created by BOSS after nine DFT calculations (red line).

difference of $5.48 \cdot 10^{-5}\,\mathrm{meV}^2$. This value drops even further for an increasing number of iterations figure 22 shows. The MSE drops below a value of $10^{-3}\,\mathrm{meV}^2$ already after the eighth data acquisition. It later saturates at a value of around $3 \cdot 10^{-3}\,\mathrm{meV}^2$. Finding the energy minimum in distorted BNOO with Bayesian optimization is less trivial than finding it in undistorted BNOO, since in distorted BNOO the canting angle that corresponds to the minimal energy configuration is different form zero. As mentioned before, the magnetic ground state of BNOO is characterized by a canting angle of $\Phi = 67°$. Figure 23a shows the difference between the minimum position of the BOSS surrogate model and the literature value d_{min} as a function of iteration steps. Since this value of the canting angle at the ground state is given in the literature with an accuracy of 1°, it is not surprising that d_{min} does not reach values below 10^{-1} degrees. The lowest value of d_{min} that can be expected is the order of accuracy of the literature, which is 1° as mentioned before. This value is first reached by a BOSS model after five data acquisitions and finally remains below this value after seven data acquisitions.

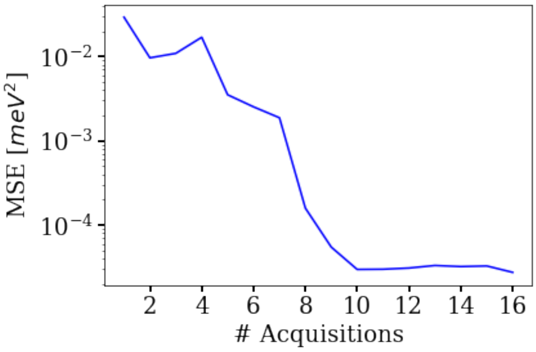

Figure 22: Evolution of the mean squared error of the surrogate model that was built for the magnetic energy curve of distorted BNOO using BOSS.

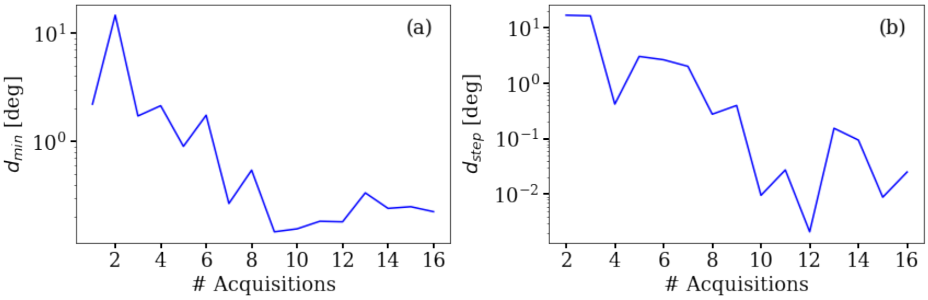

Figure 23: Figure a shows the absolute difference between the minimum position of the surrogate model and the literature value (67°). In figure b the difference between the minimum position of two consecutive iteration steps is depicted.

Further insights into the convergence of the minimum prediction can be provided by the difference of the proposed minimum position between two consecutive iteration steps d_{step}. This quantity does not rely on the information of the true minimum position and therefore does not depend on its accuracy. The value of d_{step} at every iteration step is shown in figure 23b. As this figure indicates, d_{step} first falls below 10^{-1} degrees after nine data acquisitions and mostly stays below this value afterwards. Only once, after the twelfth acquisition, the predicted minimum position changes by more than 0.1°. At this iteration step the value of d_{step} is 0.15°.

The evolution of the hyperparameter values can be seen in figure 24. As the figures 24a and 24c suggest, the absolute value for both hyperparameters σ and ℓ come close to the final value after approximately ten data acquisitions. The changes of the hyperparameter values between two consecutive iterations are shown in the figures 24b and 24d. Note that both figures have a logarithmic scale. They also suggest that the hyperparameters do not change considerably anymore after ten or 11 iterations.

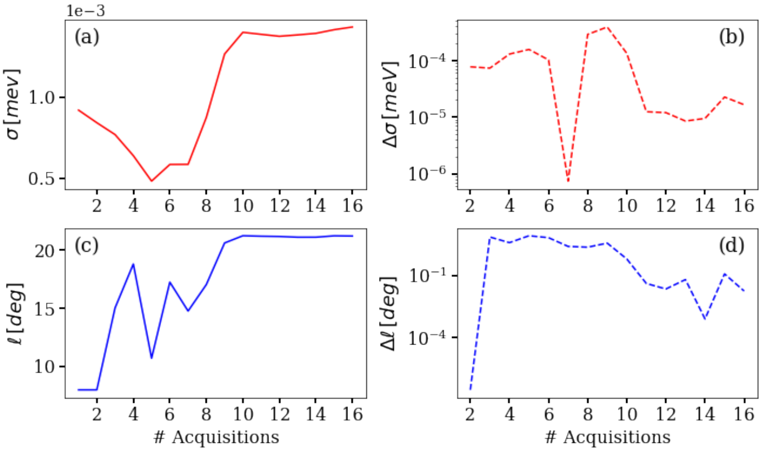

Figure 24: Evolution of the hyperparameters σ (a and b) and ℓ (c and d) for the one dimensional magnetic curve of distorted BNOO. Figures a and c show the absolute value of the hyperparameters after a given number of acquisitions. Figures b and d show the differences of the absolute values between two consecutive iteration steps.

5.1.3 UO_2

As mentioned in section 4.3.2, the magnetic ground state of uranium dioxide corresponds to a canting angle of $\Phi = 54.7°$. The total energy of UO_2 as a function of the canting angle according to literature data can be seen in figure 25a. The figure shows the original data from Dudarev *et al.* (blue points) as well as a fit of the data with function (56) (red line). The data set used by Dudarev *et al.* consists of 46 DFT calculations. Figure 25b shows the surrogate model created by BOSS based on six DFT calculations. The energy values that correspond to the six DFT calculations are indicated in the figure as black points. Despite the different size of

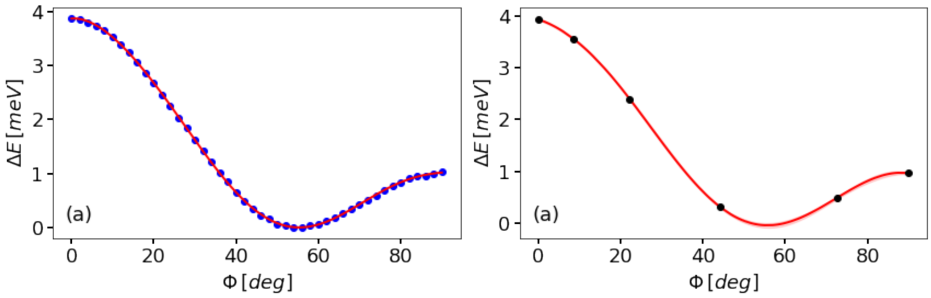

Figure 25: One dimensional magnetic curve of uranium dioxide. Figure a shows the literature data given in [72]. Figure b shows the surrogate model given by BOSS after six data acquisitions

the used data sets, the literature model and the BOSS surrogate model show the same features good agreement in general. To demonstrate this good agreement, the BOSS surrogate model and the literature data set are shown in the same graph in figure 26. The mean squared difference between the BOSS model that was created

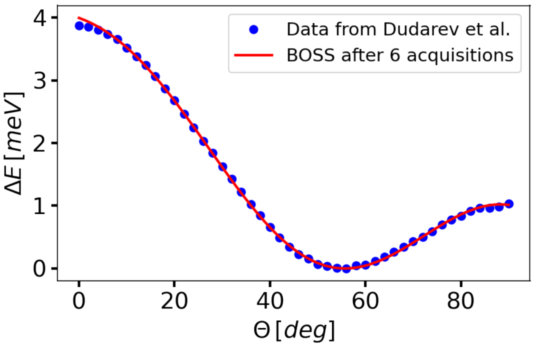

Figure 26: Direct comparison between the literature data (blue dots) and the model created by BOSS after six DFT calculations (red line).

using six DFT acquisitions and the literature model is $4.6 \cdot 10^{-4}\,\mathrm{meV}^2$. This value quantifies the mentioned good agreement. The evolution of the MSE can be seen in figure 27a. This evolution again can be used as an indicator for the convergence of the surrogate model. The MSE first falls under the value of $10^{-2}\,\mathrm{meV}^2$ after five DFT acquisitions. With a further increasing number of iterations, the MSE appears to roughly stay at the same level. This stagnation of the MSE suggests that the surrogate model does not change much anymore after this point. The small remaining discrepancy between the literature model and the model created with BOSS is most certainly caused by the small but finite error that is inherent to the performed DFT calculations. The minimum position of the surrogate model of BOSS changes with the number of acquired data points. The development of the the difference between this position and the predicted minimum position of 54.74° can be seen in figure 27b. After the acquisition of five data points the min-

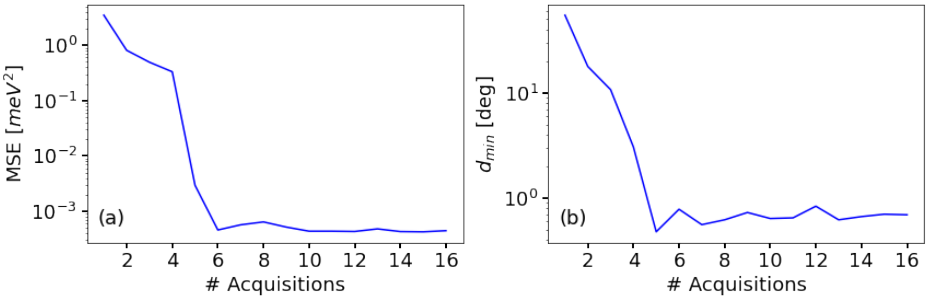

Figure 27: Convergence indicators of the one-dimensional model of the magnetic curve of UO_2. Figure a shows the evolution of the mean squared error. Figure b shows the evolution of the distance between the minimum position of the BOSS surrogate model and the verified minimum position of $54.7°$.

imum position that is predicted by BOSS is closer to the theoretical value than $1°$. The evolution of d_{min} may also be interpreted as the mentioned stagnation of convergence, supporting the theory of noise in the data acquisition process.

5.1.4 Sr_2IrO_4

The literature data for the total energy as a function of the canting angle Φ as defined in section 4.3.3 can be seen in figure 28a together the the fit of equation (57) to this data. Figure 28b shows the model of the curve created by BOSS after five DFT calculations. The BOSS model shows good agreement with the literature curve as the comparison between figure 28a and 28b indicates. The good agreement between both models can again be demonstrated by a direct comparison, which is shown in figure 29. The data set provided by Liu *et al.* consists of nine DFT calculations, whereas the BOSS surrogate model shown in figures 28 anbd 29 (red line) was created using five DFT calculations. Assuming the literature curve as true, the mean squared error of the model shown in figure 28b is $6.2 \cdot 10^{-3}\,\mathrm{meV}^2$. The mean squared error as a function of the number of performed DFT calculations can be seen in figure 30a. As the figure suggests, the MSE first drops below

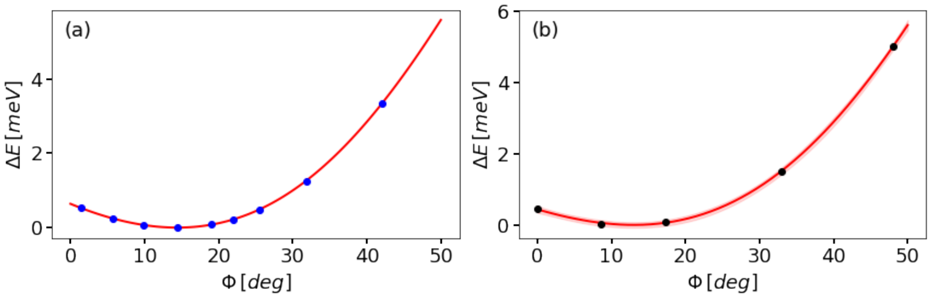

Figure 28: One-dimensional magnetic curve of SIO. Figure a shows the literature data given in [75]. Figure b shows the model given by BOSS after five data acquisitions.

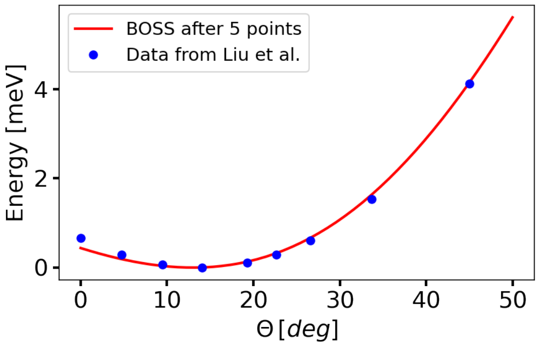

Figure 29: Direct comparison between the literature data (blue dots) and the model created by BOSS after five DFT calculations (red line).

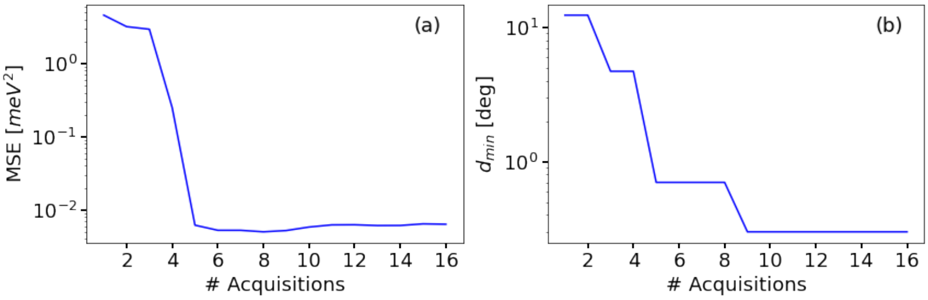

Figure 30: Convergence indicators of the one-dimensional model of the magnetic curve of SIO. Figure a shows the evolution of the mean squared error. Figure b shows the evolution of the distance between the minimum position of the BOSS surrogate model and the experimentally verified minimum position of $12.2°$.

$10^{-2}\,\mathrm{meV}^2$ after five iterations. The MSE stays below this value also for the subsequent iteration steps. Just like in the case of uranium dioxide, the evolution of the MSE shows a early stagnation of the convergence of the surrogate model at reasonably low values of the MSE. The difference between the experimentally verified canting angle in the magnetic ground state and the minimum position given by the surrogate BOSS model d_{min} is shown in figure 30 for each iteration step. This difference drops below $1°$ after five iteration steps. The value of d_{min} also stagnates after a low number of DFT data acquisitions. This may be connected to the mentioned stagnation of the MSE. On the other hand d_{min} can not be expected to take values of an order lower than 10^{-1} degrees, since this is the given level of accuracy of the experimental value.

5.2 Multi-dimensional models

In section 5.1 the magnetic moments of the noncollinear magnetic configurations are defined in such a way that there is only one degree of freedom, namely the canting angle Φ. Obviously, this is a rather drastic limitation of the space of all possible magnetic configurations. In most cases this limitation is justified by physical arguments. For example, in the search for a global energy minimum there may be hints in the form of symmetry arguments or experimental findings that show that the energy minimum must be in a specific lower-dimensional slice or subspace of the complete configuration space. Although this method of limiting the configuration space often is a successful approach, it only works if the assumption that justifies the limitation of the configuration space is valid. This makes this approach prone to errors. The risk of using a flawed assumption about the system makes it interesting to investigate a larger part of the configuration space and allow more degrees of freedom. Allowing more degrees of freedom comes with a serious drawback often referred to as the *Curse of Dimensionality.* Since the number of dimensions of the resulting model increases with the number of degrees of freedom, the number of DFT calculations needed to build such a model increases dramatically. Because of this drawback investigating noncollinear magnetism with spin-orbit coupling within density functional theory is mostly done allowing only one degree of freedom. Bayesian optimization is designed to build reliable model functions with especially small data sets. This makes it an interesting tool to allow more degrees of freedom and to get a more complete picture of magnetic effects in quantum materials. Being able to build reliable multi-dimensional models is also interesting for the investigation of completely unknown material systems. In order to find a global energy minimum in these cases, it is necessary to allow several degrees of freedom.

5.2.1 Undistortorted Ba_2NaOsO_6

As already mentioned in section 4.3.1, there exist two magnetic sites in Ba_2NaOsO_6. These two magnetic sites can be seen in figure 12a. Nevertheless we defined the magnetic moments of both sites using only one single degree of freedom, namely the canting angle Φ. The magnetic moments were then defined according to the two subfigures of figure 12. Decoupling the magnetic moments of both lattice sites leads to two degrees of freedom instead of one. The new canting angles Φ_1 and Φ_2 are defined as shown in figure 31. Using these definitions allows us to investigate a broader range of the magnetic energy landscape.

There exist several symmetries in undistorted BNOO that can be exploited to efficiently build surrogate models following the scheme presented in section 3.6. The first symmetry corresponds to interchanging the magnetic moments of the two magnetic sites. This does not affect the total energy of the material. The second symmetry corresponds to changing the signs of both canting angles Φ_1 and Φ_2. This is a symmetry operation, because the bonds of the OsO_6 octahedrons that lie in the x-y plane have the same length. Note that this is only the case for undistorted BNOO, not for distorted BNOO. The third symmetry is a combination of the first two symmetries. The total energy does not change if both the absolute values and the signs of the two canting angles are swapped. Due to these symmetries it is possible to add four data points to the data set after only one DFT calculation. The resulting model of the magnetic energy landscape of undistorted

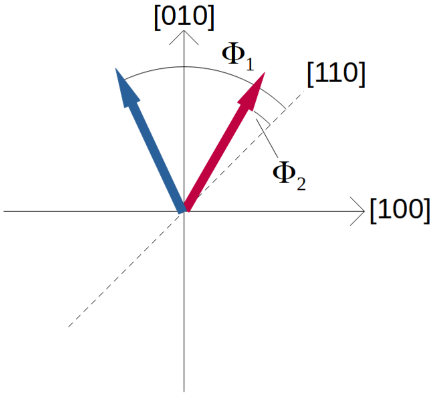

Figure 31: Definition of the two canting angles Φ_1 and Φ_2 for the two magnetic sites defined in figure 12a. As the figure shows, both angles are defined with respect to the [110] axis of the crystal. Both magnetic moments lie within the (001) plane of the Ba_2NaOsO_6 crystal.

BNOO using the above definitions of Φ_1 and Φ_2 and exploiting the described symmetries is shown in figure 32. For the creation of the model, the Matérn covariance

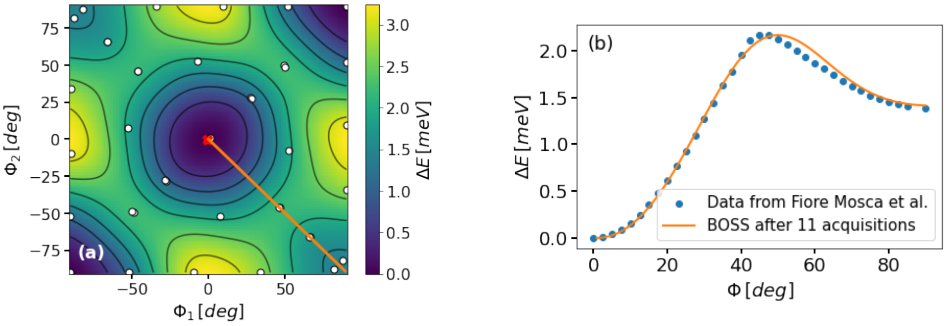

Figure 32: Surrogate model of the energy landscape of undistorted Ba_2NaOsO_6 after 11 DFT calculations using existing symmetries. Figure a shows a two dimensional model that uses Φ_1 and Φ_2 as defined in figure 31 as input parameters. The red cross marks the global minimum of the model. Figure b shows a one-dimensional slice of the same model as well as corresponding literature data from Fiore Mosca *et al.* [60]. The position of the slice in the two-dimensional model a is highlighted with an orange line.

function with $\nu = 5/2$ and the explore acquisition function were used. Figure 32a shows the model of the full two-dimensional energy landscape. For the creation of the model 11 DFT calculations were used. Because of the existing symmetries, four times more data points could be added to the data set. Thus the total data set consists of forty-four data points. The used data points are marked in figure 32a as white dots. The global minimum of the model is with $\Phi_1 = -0.9°$ and $\Phi_2 = -0.1°$ close to the ferromagnetic position $\Phi_1 = \Phi_2 = 0°$, which is said to be the global DFT energy minimum of this structure [60]. Figure 32b shows a one-dimensional slice of the two-dimensional model. The position of the slice agrees with the previous definition of Φ shown in figure 12b that was used for the creation of the one-dimensional models in section 5.1. The slice is marked in the two-dimensional model shown in figure 32a as an orange line. The one-dimensional slice of the model shows good agreement with literature data from Fiore Mosca *et al.* which is also shown in figure 32b. The mean squared error between the one-dimensional slice of the model and the literature data is $4.6 \cdot 10^{-3}\,\text{meV}^2$. The evolution of the MSE with increasing number of DFT calculations is shown in figure 33. The MSE falls under the value of $10^{-2}\,\text{meV}^2$ after a total of 11 DFT calculations. Also for higher iteration numbers the MSE stays below this value. The MSE is a good measure for the convergence of the model in the one-dimensional slice that corresponds to the reference data, but it can not be seen as a measure for the convergence of the whole two-dimensional model. However, there are a number of convergence measures that do not depend on previous knowledge about the magnetic properties of the system, as discussed in section 3.7. The evolution of the kernel hyperparameters can always give information of the convergence of a model created with Bayesian optimization. In figure 34 the hyperparameters σ and ℓ are shown as a function of the number of acquisitions. Although the covariance function that is used to describe the two-dimensional model of the energy surface has two lengthscale hyperparameters, only one is shown in the figure because both curves agree to large extent. As the figure shows, the hyperparameters do not seem to have converged completely after 36 DFT calculations. This also means that the whole two-dimensional model may not be completely converged at this stage. However the changes of the model do not seem to take place within the one-dimensional region marked as an orange line in figure 32a since the surrogate model does not seem to change much there, as figure 33 shows. The dissatisfying evolution of the hyperparameters may also be explained by the questionable significance of this convergence measure, as discussed in section 3.7. Another, probably more reliable convergence measure that does not depend on the existence of reference data is

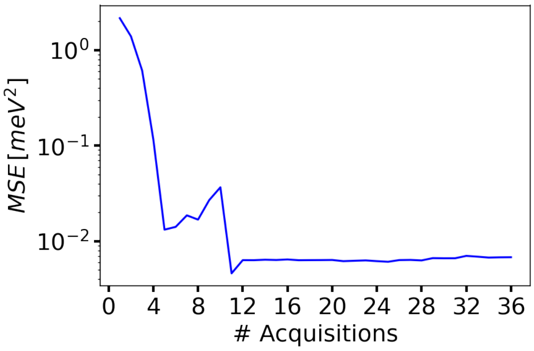

Figure 33: Mean squared error (MSE) of the one-dimensional slice of the BOSS model compared to literature data shown as a function of the number of performed DFT calculations.

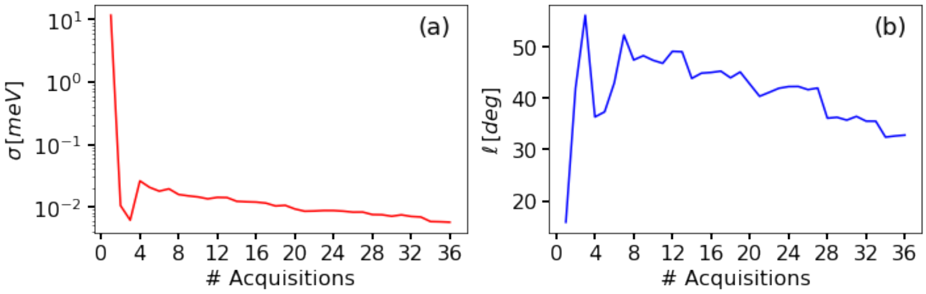

Figure 34: Evolution of the hyperparameters σ (figure a) and ℓ (figure b) for the two-dimensional model of undistorted Ba_2NaOsO_6. The evolution of the hyperparameter σ is shown in log-scale for clarity reasons.

the mean squared difference of every point of the input domain between two consecutive iteration steps, MSE_{2D}. The evolution of MSE_{2D} can be seen in figure 35. The figure shows that the absolute values of MSE_{2D} quickly drop to relatively low values. However, even after more than 20 data acquisitions the MSE_{2D} values oscillate considerably. These ambiguous observations make it difficult to decide if the model can be seen as accurate or not. On the one hand the good agreement with the model with literature data shown in figure 32b and the low levels of the corresponding mean squared error shown in figure 33 suggest a reliable, accurate model. On the other hand the fact that the hyperparameters are still changing after 36 data acquisitions (shown in figure 34) and the ambiguous evolution of the MSE_{2D} (shown in figure 35) do not support this optimistic interpretation. It would be helpful to have reliable reference data on the energy landscape of undistorted BNOO also for the complete two-dimensional subspace of the configuration space.

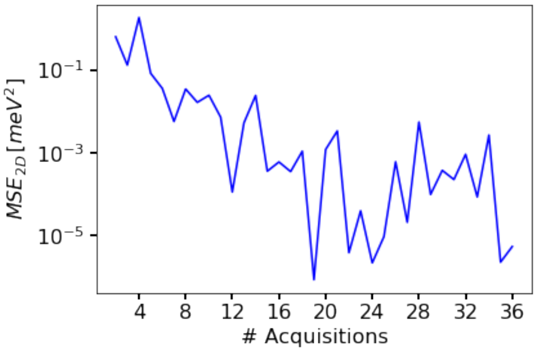

Figure 35: Mean squared difference of the full two-dimensional model between two consecutive iteration steps MSE_{2D} as a function of the number of performed DFT calculations

5.2.2 Distorted Ba_2NaOsO_6

In distorted BNOO the new canting angles Φ_1 and Φ_2 can be defined in the same way as in the undistorted case (see figure 31). However, due to the distortion of the octahedrons around the Osmium atoms, only one symmetry exists. This symmetry corresponds to swapping both the magnetic sites and the sign of the canting angle. Figure 36 shows a model of the two-dimensional magnetic energy landscape. The model was created using the explore acquisition function and the Matérn kernel with $\nu = 5/2$. Furthermore the above mentioned symmetry was exploited using the scheme explained in section 3.6. Figure 36a shows the full model using 26 DFT calculations. The region of the configuration space that corresponds to the definition of the canting angle Φ, which was used for the one-dimensional models is indicated by an orange line in the lower right corner of the figure. The red cross, which indicates the global minimum of the two-dimensional model does not lie on that orange line. Instead it is close to the ferromagnetic position that can be characterized by $\Phi_1 = \Phi_2 = 90°$. Therefore the global energy minimum, according to this two-dimensional model, is not part of the limited configuration space that was used before. The energy difference between the global minimum and the local minimum on the orange line is 0.84 meV. This is a larger energy difference than the difference between the lowest and the highest energy within the limited one-dimensional region of configuration space. The magnetic configuration corresponding to the energy minimum that was found in this work does not agree with experimental findings on BNOO (see section 4.3.1). The results of the NMR study of Lu *et al.* [59] clearly show that there must exist two different lattice sites with different magnetic moment. This is in conflict with the almost collinear nature of the magnetic ground state reported here. Figure 36b shows the one-dimensional slice of the two-dimensional model corresponding to the previous definition of canting angle Φ. It shows good agreement with the literature data,

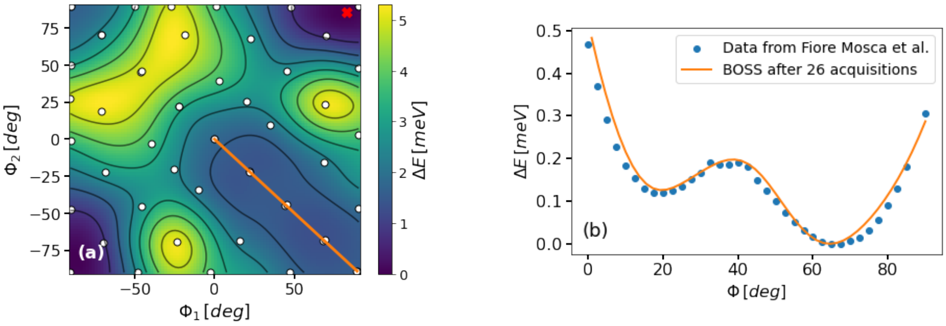

Figure 36: Surrogate model of the energy landscape of distorted BNOO after 26 DFT calculations using existing symmetries. Figure a shows a two dimensional model using Φ_1 and Φ_2 as defined in figure 31 as input parameters. The red cross marks the global minimum of the model. Figure b shows a one-dimensional slice of the same model as well as corresponding literature data from Fiore Mosca *et al.* [60]. The position of the slice in the two-dimensional model a is highlighted with an orange line.

which is depicted in the figure. This can be interpreted as a sign for a trustable two-dimensional model and good convergence. However, the mean squared deviation between the one-dimensional slice of the model and the literature data yields a more differentiated picture (figure 37). It shows that after performing more than 26 DFT calculations, the squared difference between the BOSS model and the literature model increases. It eventually reaches a maximum of $9.8 \cdot 10^{-3}\,\mathrm{meV}^2$ after the fortieth iteration. After this maximum, it gradually decreases again to lower values. This evolution of the MSE lowers the confidence in the correctness of the two-dimensional model.

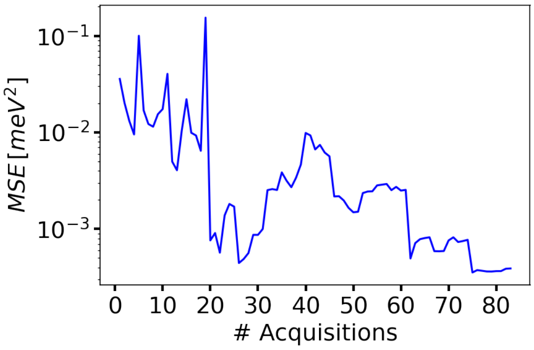

Figure 37: Mean squared error (MSE) of the one-dimensional slice of the BOSS model compared to literature data shown as a function of the number of performed DFT calculations.

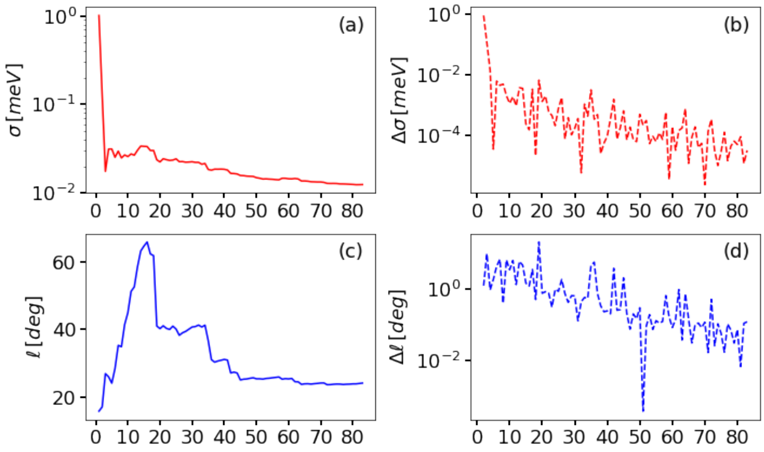

Figure 38: Evolution of the hyperparameters σ (figures a and b) and ℓ (figures c and d) for the two-dimensional magnetic curve of distorted BNOO. Figures a and c show the absolute value of the hyperparameters after a given number of acquisitions. Figures b and d show the difference between two consecutive iteration steps.

The evolution of the kernel hyperparameters can be seen in figure 38. The figure shows that the hyperparameters do not change significantly any more after about 50 iterations. This indicates that the models might be more trustworthy and therefore more stable after this point. The difference of each hyperparameter between two consecutive iterations (figures 38b and 38d) seem to decrease almost linearly in the logarithmic figure, which can be interpreted as an exponential decrease and as a good sign for hyperparameter convergence. However, the evolution of the MSE_{2D} shows a trend that can be seen as more ambiguous. It is shown in figure 39. Although the general trend of the MSE_{2D} is decreasing, the fluctuations of this quantity are relatively high. Therefore the evolution of the MSE_{2D} does not necessarily support the claim that the model is converging.

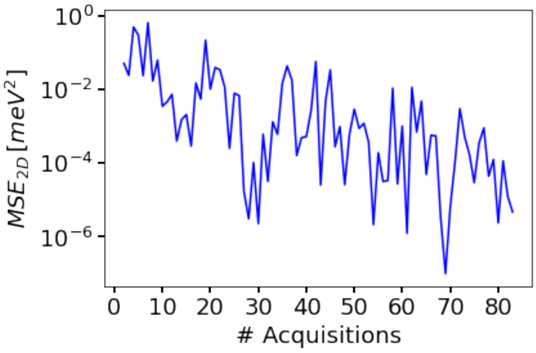

Figure 39: Mean squared difference of the full two-dimensional model between two consecutive iteration steps MSE_{2D} as a function of the number of performed DFT calculations.

5.2.3 UO_2

As described in section 4.3.2 uranium dioxide has four inequivalent magnetic sites. For the creation of the one-dimenional models, the canting angle Φ was defined as shown in figure 14. Due to this definition, the canting angles at the four lattices sites were all changed in a concerted way. Decoupling the canting angles at all four lattice sites increases the number of degrees of freedom from one to four. This definition makes it possible to create a four-dimensional model of the energy landscape, allowing to search for the magnetic ground state of UO_2 in a much larger configuration space. Since a four-dimensional model is difficult to display, it is necessary to create a lower-dimensional slice of the model. Figure 40 shows a two-dimensional slice of the four-dimensional model. The slice was created using the conditions $\Phi_1 = \Phi_3$ and $\Phi_2 = \Phi_4$. The indices correspond to the numbering of the atoms as it is shown in figure 14. The model was crated using the exploratory lower confidence bound acquisition function and the Matérn kernel with $\nu = 3/2$. It is based on 25 DFT calculations. Figure 40a shows the mentioned two-dimensional slice of the model. The red cross indicates the position of the global minimum of the slice. Shown as an orange line is the one-dimensional subspace of the full configuration space that corresponds to the literature data from [72]. In figure 40b the corresponding one-dimensional slice of the full four-dimensional model is

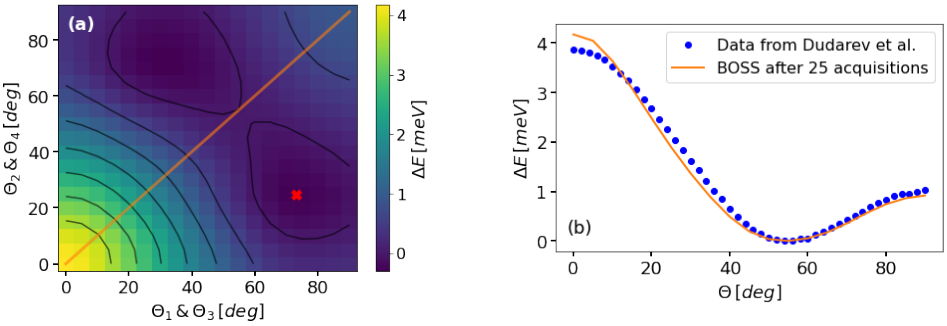

Figure 40: Surrogate model of the magnetic energy landscape of UO_2 after 25 DFT calculations using existing symmetries. Figure a shows a two dimensional slice of the four dimensional model using the condition $\Phi_1 = \Phi_3$ and $\Phi_2 = \Phi_4$. The red cross marks the global minimum of the model. Figure b shows a one-dimensional slice of the same model as well as corresponding literature data from Dudarev *et al.* [72]. The position of the one-dimensional slice in the two-dimensional slice a is highlighted with an orange line.

compared to the literature data. As the figure shows, both models agree to good extent. The mean squared error of the BOSS model is $3.9 \cdot 10^{-2}\,\mathrm{meV}^2$. This means that the model is almost correct, at least in the area around the one-dimensional slice. The evolution of this mean squared error can be seen in figure 41c. As the figure shows, the mean squared error suddenly decreases after approximately 20 data acquisitions. After this drop the value of the MSE stays at a relatively low level of $\sim 1.5 \cdot 10^{-2}\,\mathrm{meV}^2$. It is worth noting that the numerical global minimum (red cross) of the two-dimensional slice does not lie in the part of the configuration space used to build the one-dimensional model. Instead the four-dimensional model seems to have four minima that are almost equal in absolute energy. Two of those minima can be seen in figure 40a, the other two lie in a different subspace of the four-dimensional model. None of those four minima coincides with the 3-k state, that appears to be the magnetic ground state of uranium dioxide, according to experiments (see section 4.3.2). The one-dimensional slice of the configuration space that was investigated in the literature is defined by the condition $\Phi_1 = \Phi_2 = \Phi_3 = \Phi_4$. The mentioned finding suggest that the assumption that the global minimum position fulfills this condition might be questionable.

Although the good agreement between literature data and the model after 25 DFT calculations (figure 40b) can be seen as a good sign for convergence of the model in this particular one-dimensional slice, this can not be said about the rest of the model with certainty. Finding a good convergence criterion for the complete model is difficult because the known data for this material is limited to only a small part of the relevant configuration space. The hyperparameters and the squared change between two consecutive iteration steps MSE_{4D} can be used as convergence measures for the full four-dimensional model, but they are limited in their information value. These convergence measures are shown in the figures

41a, 41b and 41d, respectively. Although the covariance function that is used here has four lengthscale hyperparameters, only one is shown in the figure, since the curves agree to large extent. The hyperparameter σ seems to be converged

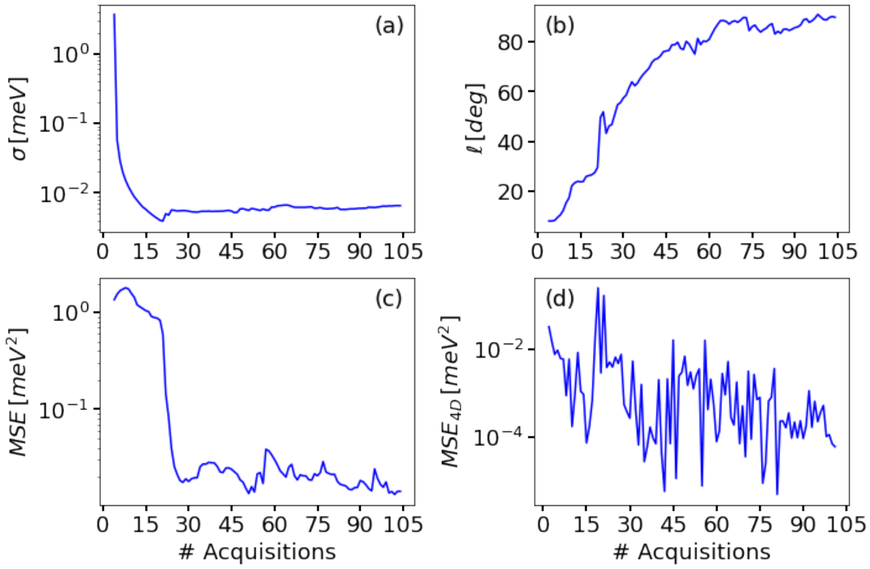

Figure 41: Convergence measures of the four-dimensional model of the magnetic energy landscape of UO_2. Figure a and b show the evolution of the kernel hyperparameters σ and ℓ, respectively. Figure c shows the MSE of the model in the one-dimensional subspace that is defined by $\Phi_1 = \Phi_2 = \Phi_3 = \Phi_4$. Figure d shows the mean squared change between the four-dimensional surrogate models of two consecutive iteration steps.

already after about 25 iterations, according to the results visible in figure 41a. The lengthscale on the other hand seems to reach convergence later (see figure 41b). Similar to the case of distorted BNOO, the convergence measure MSE_{4D} generally decreases with an increasing number of iteration steps, but shows relatively large fluctuations. This can be seen in figure 41c. In total the convergence measures give an unclear picture of the state of the convergence of the model. It is quite likely that the full four-dimensional model did not completely converge to an exact model. This is emphasized by the fact that the one-dimensional literature model can not be reproduced with one hundred DFT calculations when using the explore acquisition function. This raises the suspicion that the four-dimensional model is only exact in the region around the global minimum. This hypothesis might be checked by comparing a one-dimensional slice of the model to a reliable one-dimensional model that was created with a sufficient number of DFT calculations.

5.3 Settings analysis

There are a variety of kernel functions and acquisition functions that can be used for Bayesian optimization. The functions that were used in this work are described in sections 3.1 and 3.5, respectively. The choice of kernel function and acquisition function, as well as the used symmetries (see section 3.6) can affect the performance of the approach. In this section we analyze the performances of Bayesian optimization runs with different settings to determine which approaches show the highest efficiency. Tabular summaries of the performances of different settings for all three materials can be found in appendix A.

5.3.1 Undistorted Ba_2NaOsO_6

Using the settings presented in section 5.1.1 it was possible to create models of the magnetic curve of undistorted BNOO that had a mean squared error lower than $10^{-3}\,\mathrm{meV}^2$. Table A.1 summarizes the results of the quality evaluations based on the MSE for different settings of the Bayesian optimization approach. Although for all settings good agreements with the literature data could be reached, the MSE did not take values below $10^{-3}\,\mathrm{meV}^2$ for reasonable numbers of acquired data in some cases. There is a striking distinction between the runs investigating the range between 0° and 90° and the runs investigating the range between 0° and 180° exploiting the periodicity and or the symmetry of the underlying function. Using the smaller input interval it was always possible to reach a MSE below $10^{-3}\,\mathrm{meV}^2$ while using the larger input interval it was never possible. On the other hand, using the larger input interval while exploiting the periodicity of the function led to better agreement with literature data after very small numbers of acquisitions (≤ 5), as table A.1 shows. Therefore the question whether it is advantageous to use the standard periodic kernel depends on the level of accuracy that one wants to achieve. Comparing pairs of two runs that only differ in the use of the acquisition function (elcb or explore) suggests that using the explore acquisition function is advantageous for reconstructing the complete underlying function. On average, one acquisition less was needed to reach a MSE of 10^{-2}. This make sense, since the explore acquisition function is designed just for that task, whereas the exploratory lower confidence bound (elcb) acquisition function is primarily designed to efficiently find the global minimum of an objective function. When using a periodic kernel and the larger 180° input range, it seems to be advantageous to also use the existing mirror symmetry of undistorted BNOO. Independent from the choice of acquisition function, the number of required DFT calculations necessary to reach a low MSE value is lower when using the symmetry. Using the elcb acquisition function, four less data acquisitions were necessary to reach a MSE below $10^{-2}\,\mathrm{meV}^2$, while using the explore acquisition function, two less acquisitions were necessary. In order to reach the acceptable value of $10^{-2}\,\mathrm{meV}^2$ for the mean squared error, the choice of acquisition function, kernel and used symmetry does not seem to have a big impact.

5.3.2 Distorted Ba_2NaOsO_6

We investigated the performance of the proposed approach for the investigation of BNOO with a distorted crystal lattice using different kernel functions, acquisition functions and symmetries. The number of DFT calculations necessary to reach different levels of accuracy with respect to the mean squared error and the global minimum position for the different settings are shown in table A.2. Two different acquisition functions and three different kernel functions have been used in all six possible combinations. The expected positive effect of using the explore acquisition function on the MSE can not be observed with great clearness. The average numbers of necessary DFT calculations to reach a MSE below $10^{-3}\,\mathrm{meV}^2$ and $10^{-4}\,\mathrm{meV}^2$ is smaller by only 0.5 if the explore acquisition function is used. In general, the squared exponential kernel appears to perform slightly better than the two Matérn kernels. This holds true for the mean squared error as well for the localization of the minimum position. Using the squared exponential kernel in combination with the explore acquisition function the proposed minimum position is closer to the true minimum position than 1° after five DFT calculations. With the same settings, an acceptable mean squared error below $10^{-3}\,\mathrm{meV}^2$ can be reached after six DFT calculations.

5.3.3 UO_2

Details on the performance of our approach on uranium dioxide using different settings can be found in table A.3. Using an augmented range of 180° instead of 90° did not lead to significant improvements in any case. In fact, using the standard periodic kernel without imposing a mirror symmetry on the surrogate models worsened the performance regarding the mean square error. However, it was possible to decrease the mean squared error already after four data acquisitions below $10^{-2}\,\mathrm{meV}^2$ when using the existing mirror symmetry in combination with the standard periodic kernel. The number of acquisitions necessary to find the global minimum with good accuracy does not seem to be significantly dependent on the choice of the kernel and acquisition functions. In every case it was possible to find the minimum with an error of less than 1° by using between five and seven DFT calculations.

5.3.4 Sr_2IrO_4

In section 5.1.4, the successful creation of a satisfactory model for the dependence of the total energy of the canting angle in SIO is demonstrated. This was done by making use of the squared exponential kernel in combination with the exploratory lower confidence bound acquisition function. Good agreement with the function reported in literature could be reached with this approach. After five DFT calculations the mean squared error dropped below $10^{-2}\,\mathrm{meV}^2$. However, this low level could not be reproduced for different settings within a reasonable number of

DFT calculations (see table A.4). Since the overall energy differences in SIO are comparably large (up to $\sim 5\,\text{meV}$) this difference are not very large on a relative scale. Often there was a slight mismatch between both functions for low values of the canting angle. This could be a sign for slightly different computational setups. The mean squared errors do not change significantly after six data acquisitions in any case. this may also be explained by small errors that get introduced systematically due to different computational approaches. The slight mismatch in acquired data could also explain the mismatch in the minimum positions that can be seen in table A.4, especially since the curve only shows a moderate gradient around the minimum. With all tested settings it was possible to roughly identify the correct canting angle of the magnetic ground state ($d_{min} < 2°$). A better approximation ($d_{min} < 1°$) could only be reached in three cases. Using the Matérn kernel with $\nu = 3/2$ and the explore acquisition function, this was even possible after performing as little as four DFT calculations. However, it is not apparent from the analysis which settings are best used for the identification of the global minimum position in the case of SIO.

5.3.5 Undistorted Ba_2NaOsO_6 2D

As reported in section 5.2.2, the magnetic energy surface was successfully explored beyond the one-dimensional subspace that was reported in literature. More details on the definition of this two-dimensional subspace of configurations space can be found in section 5.2.1. Satisfying results could be achieved with several settings of the Bayesian optimization run. The number of DFT calculations necessary to reach a certain threshold of the mean squared error is given in table A.5 for different settings. In some cases, an existing symmetry was enforced on the model. As the table shows, Bayesian optimization does not always succeed to reach a good level of agreement with literature data in the sense that the MSE can be reduced to a value under $10^{-2}\,\text{meV}^2$ within a reasonable number of data acquisitions. Apparently, in this specific case, the choice of kernel seems to be a critical factor. Using the squared exponential kernel it was never possible to reach a mean squared error below $10^{-2}\,\text{meV}^2$. Another seemingly helpful factor was using the existing symmetries in undistorted BNOO. Only while exploiting the symmetries it was possible reach a mean squared error below $10^{-2}\,\text{meV}^2$. On average it was possible to reach a mean squared error below $10^{-1}\,\text{meV}^2$ using 6.5 DFT calculations while exploiting the symmetries. Without this, 18.5 DFT calculations were necessary on average. The choice of acquisition function did not seem to have a significant impact on the performance.

5.3.6 Distorted Ba_2NaOsO_6 2D

A two-dimensional model of the energy landscape of distorted BNOO could be built with the help of Bayesian optimization, so that the one-dimensional slice agrees well with the literature data. However, In order to reach a convincing level of convergence (below $10^{-3}\,\text{meV}^2$) within a reasonable number of DFT calculations

(< 80) special settings had to be made. As table A.6 shows, this level of accuracy could not be reached using the squared exponential kernel. Furthermore it was necessary to exploit the existing symmetry in distorted BNOO. The best performance could be seen by using the Matérn kernel with $\nu = 5/2$ and the exploratory lower confidence bound acquisition function while imposing the mentioned symmetry on the surrogate models. In general, the choice of acquisition function does not seem to play a significant role. On average, it was necessary to perform 28 DFT calculations to reach a mean squared error below $10^2\,\mathrm{meV}^2$ when using the exploratory lower confidence bound, while 26 calculations were necessary when using the explore acquisition function. Most likely this can partially be explained by the generally high acquisition numbers. Due to the definition of the acquisition function (equation (45)), the exploration strategy is used more and more while increasingly neglecting the exploitation strategy with an increasing number of iterations. Therefore the differences between the two acquisition functions vanish at high iteration numbers to some extent. Although it was only in some cases possible to recreate the one-dimensional literature curve with convincing accuracy, all runs showed the same general features of the two-dimensional magnetic energy landscape. Most interestingly, the global minimum of the BOSS surrogate models does not correspond to the experimentally verified magnetic ground state of BNOO.

6 Discussion and Conclusion

In this work Bayesian optimization was used to build model functions of the magnetic energy surface of noncollinear magnets and to investigate their magnetic ground states. It has been studied which effect the use of different kernel functions and acquisition functions as well as imposed model symmetries have on the efficiency of the approach. The mean squared error was used as the main measure of the quality of each model. Additionally, the difference between the minimum of the model and the literature value for the minimum, the change of the model between two consecutive iteration steps and the evolution of the kernel hyperparameters were used to assess the convergence behavior of the surrogate models. The noncollinear magnets Ba_2NaOsO_6, UO_2 and Sr_2IrO_4 were used as benchmark materials in this thesis. Literature models of the energy surface of all three materials were successfully reproduced using Bayesian optimization. Compared to the number of DFT calculations used in the corresponding literature the results were reproduced using between 44% and 91% fewer calculations.

The literature model of the magnetic energy landscape of undistorted Ba_2NaOsO_6 could be reproduced with good accuracy. For several combinations of kernel functions and acquisition functions, only five single DFT calculations were necessary to reproduce the literature function with a mean squared error below $10^{-2}\,\mathrm{meV}^2$ (see table A.1). Therefore, this methodology significantly outperforms the DFT calculation count used in the referenced study by requiring a remarkable 86% fewer DFT calculations to achieve the mentioned level of accuracy. However, if the theoretical model that accurately describes the relation between canting angle and the total energy (equation (55)) is known, this number of DFT calculations does not pose a significant advantage. This can be shown by performing DFT calculations for equidistant values of the canting angle and performing a least-squares fit of the model function (55). The result of such a simplistic procedure that uses five equidistant DFT data points can be seen in figure 42. The figure shows that the

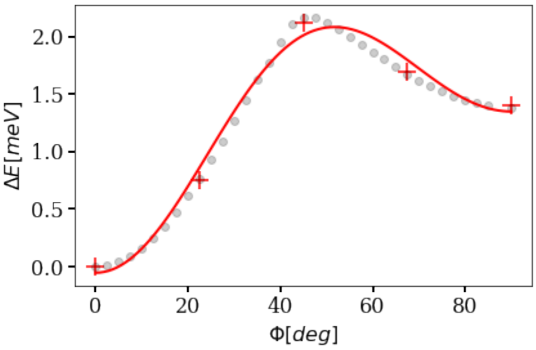

Figure 42: Comparison of a naive curve fitting approach with literature data. Five DFT calculations at equidistant angles were used.The red crosses represent the acquired data and the red curve shows the corresponding fit of the data using function (55).The gray dots represent the literature data.

simplistic approach already works moderately well. This is not surprising since assuming noise-less data acquisition, the number of data points necessary to unambiguously define the parameters of a fit function equals the number of those parameters. The mean squared error of this simplistic approach is $6.6 \cdot 10^{-3}\,\mathrm{meV}^2$. However, a crucial condition for the success of this method is the knowledge of the theoretical model (equation (55)).

The literature function that describes the total energy of Ba_2NaOsO_6 with Jahn-Teller distortions as a function of a magnetic canting angle could also be reproduced with good accuracy using Bayesian optimization. Since the magnetization-dependent energy differences of the distorted BNOO are smaller than the energy differences of the undistorted BNOO, it is necessary to reach higher levels of accuracy, in order to accurately model all features of the function. It was possible to reach a mean squared error below $10^{-3}\,\mathrm{meV}^2$ after six DFT calculations using the squared exponential kernel and the explore acquisition function, which are 83% fewer DFT calculations compared to the number of calculations that were used in the relevant study. The result of the simplistic approach of generating a model function that was described above can be seen in figure 43. Here, six DFT calculations are used as well. This model also agrees well with the literature data, as

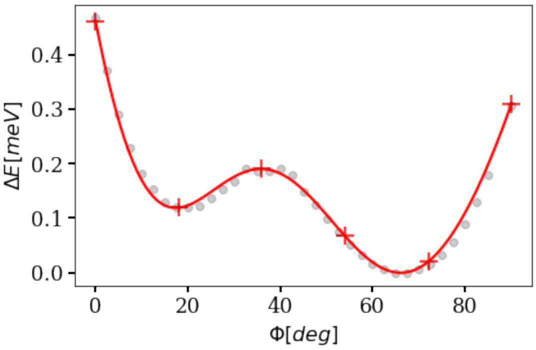

Figure 43: Comparison of a naive curve fitting approach with literature data. Six DFT calculations at equidistant angles were used.The red crosses represent the acquired data and the red curve shows the corresponding fit of the data using function (55).The gray dots represent the literature data.

figure 43 shows. The mean squared error of this model is $1.0 \cdot 10^{-4}$. Therefore, the use of Bayesian optimization only poses a significant advantage in this case if the relevant fitting function is unknown. However, by using Bayesian optimization it was possible to identify the canting angle that corresponds to the energy minimum with an error below 1° using only five DFT calculations. This was done using the squared exponential covariance function and the explore acquisition function. These settings appear to work best for the investigation of distorted Ba_2NaOsO_6.

The one dimensional function that describes the total energy of uranium dioxide as a function of the canting angle could also be recreated. In fact, different combinations of kernel functions, acquisition functions and symmetries lead to satisfying results. Good agreement with the literature model (mean squared error smaller than $10^{-2}\,\mathrm{meV}^2$) could already be reached with only four DFT calculations by exploiting the mirror symmetry of the objective function in combination with the standard periodic kernel. This number is lower by 91% compared to the number of DFT calculations used in the study that first reported the model. Taking a more classical approach, that means performing DFT calculations at equidistant angles and fitting this data to the relevant function (equation (56)) would not work with this small number of DFT calculations. Equation (56) has eight free parameters that need to be determined. This is only possible with at least eight data points. There is an intuitive way to justify the last statement. When the equation (56) and the four data points are known, one can see this as a system of four linear equations. A set of of linear equations can only be fully determined if there are as many equations as there are free parameters in each equation.

The model capturing the relationship between the canting angle and the total energy of Sr_2IrO_4 was successfully replicated as well. A satisfying mean squared error below $10^{-2}\,\mathrm{meV}^2$ could be reached by using five DFT calculations and employing the squared exponential kernel in combination with the exploratory lower confidence bound acquisition function. Therefore, 44% fewer DFT calculations were necessary to reach the given level of accuracy than were used in the original publication. It was possible to identify the correct magnetic ground state of Sr_2IrO_4 with an error below 1° using only four DFT calculations in the best case.

In general, the effect of using different acquisition functions is rather small for one-dimensional models. In most cases it was not possible to distinguish the effect from probabilistic uncertainties. The kernel functions that worked best in accelerating the model creation appeared to depend on the specific material, the required accuracy and the specific research question.

In this work Bayesian optimization was not only used for the recreation of existing one-dimensional magnetic curves, but also for the creation of multi-dimensional models of the energy landscape of the magnets. This was done by defining two independent degrees of freedom in the case of BNOO and four independent degrees of freedom in the case of uranium dioxide. Naturally, more DFT calculations are needed for the construction of multi-dimensional models compared to one-dimensional models. Additionally, the convergence behavior of multi-dimensional models appeared to be more ambiguous than in the one-dimensional case.

Nevertheless, it was possible to build two-dimensional models of the energy-landscape of both undistorted and distorted BNOO that agree well with the existing literature data. Interestingly, the two-dimensional models show that neither the undistorted BNOO crystal nor the distorted BNOO crystal can be used to explain the experimentally observed magnetic ground state of BNOO. Unlike reported by Fiore Mosca *et al.*, the specific Jahn-Teller distortions lead within DFT to a magnetic ground state that does not agree with experimental data. Therefore, more research is necessary to explain the experimental observations concerning magnetism in BNOO.

As mentioned in section 5.2.3 it was possible to create a model of the total energy of uranium dioxide that takes into account four degrees of freedom. Naturally, the accuracy of this four-dimensional model can only be judged with certainty along the one-dimensional slice reported in literature. As figure 40b shows, the model agrees well with the existing literature in this region of the configuration space. This can be seen as a good sign for the correctness of the model. Surprisingly, according to the four-dimensional model created with Bayesian optimization, the minima of the energy landscape of uranium dioxide do not lie in the one-dimensional slice analyzed in literature. Instead, the previously reported minimum corresponds to a saddle point in the four-dimensional model. Symmetrically arranged around the saddle point, four different energy minima appear to exist. The existence of these "new" minima within DFT is backed by DFT calculations that were automatically performed during the Bayesian optimization run. By allowing more degrees of freedom the minimal DFT energy that was calculated is smaller than the DFT energy calculated at the minimum position of the one-dimensional slice by 1.8 meV. However, it is generally assumed that the 3-k ordering mentioned in section 4.3.2 is the magnetic ground state of uranium dioxide. Previously, other magnetic orderings, namely 1-k and 2-k orderings were taken into account for the explanation of experimental results. It was shown [78], however, that the 3-k ordering best explains the results of neutron scattering experiments under an external magnetic field [65]. Furthermore Ikushima *et al.* argued that the results of their NMR study [70] agree better with a 3-k structure than to a 1-k or 2-k structure. Furthermore, Blackburn *et al.* [69] performed spherical neutron spin polarimetry and found that the observed polarization behavior can be explained best by a 3-k structure. However, the "new" ground state observed in this work was not explicitly mentioned or considered. On the other hand, an explanation for the observed ground state may also be found in the choice of exchange-correlation functional. The work of Zhou *et al.* [79] shows that different exchange-correlation functionals can lead to different calculated magnetic ground states in uranium dioxide. Nevertheless, a convincing explanation for the mentioned observations still needs to be found.

In general, the correctness of the entire multi-dimensional models presented in this work may be disputable, since the convergence behavior of the Bayesian optimization surrogate models is ambiguous. Furthermore, the data that can be taken as a reference only cover a small subspace of the models. However, these models show that Bayesian optimization has a great potential to be used as a versatile tool in the research on quantum magnetism. In principle, the approach can easily be expanded to study the impact of even more degrees of freedom on the total energy. The only remaining difficulty is to determine a necessary number of data points to decrease the inaccuracies of the surrogate model to an acceptable minimum. By investigating two degrees of freedom for each relevant lattice site, all possible canting angles, in-plane and out-of-plane can be covered. Therefore, Bayesian Optimization may be used to study black box systems and to cross-validate previous assumptions about magnetic ground states and further properties of magnetic materials.

7 Acknowledgments

First and foremost, I want to thank Prof. Cesare Franchini, who made this thesis possible in the first place. Besides developing the topic of the thesis, he always happily helped with any kind of problem and gave helpful inputs at the crucial points of the work. Furthermore he established the contact to the Computational Electronic Structure Theory group at Aalto university.

Moreover I want to thank Prof. Dieter Süss for being the second examiner of my work.

I also want to express my gratitude to my co-supervisor Lorenzo Celiberti for his perpetual support. In many fruitful briefings and discussions he shared his physical expertise and taught me how to work with VASP.

During the work of this thesis I had the pleasure to learn about Bayesian Optimization and the BOSS program from as. Prof. Milica Todorović. I am very thankful for her readiness to give detailed explanations for every question that I had.

I am also grateful to Dr. Dario Fiore Mosca for the stimulating discussions and the interesting insights that he gave me during our common time in Vienna.

Finally, I must express my profound gratitude to the whole Computational Materials Physics group of the university of Vienna. It was a great pleasure to spend the time of my Master's thesis within this group of nice people. I really enjoyed the time with everyone at Kolingasse.

A Tabular summary of performances

A.1 Undistorted Ba_2NaOsO_6 1D

Table A.1: Summary on the performance of Bayesian optimization on the recreation of one-dimensional magnetic curves for undistorted Ba_2NaOsO_6 . The table shows the number of DFT calculations that are necessary to reach a certain level of convergence. This level of convergence is defined as falling below and staying below a certain threshold value for the MSE. The threshold values are given in the first line of the table.

Settings				MSE		
Kernel	**Acquisition function**	**Range**	**symmetry**	$< 10^{-1}\,\mathrm{meV}^2$	$< 10^{-2}\,\mathrm{meV}^2$	$< 10^{-3}\,\mathrm{meV}^2$
SE	elcb	$0° - 90°$	no	5	5	10
SE	explore	$0° - 90°$	no	5	5	9
Matérn ($\nu = 3/2$)	elcb	$0° - 90°$	no	5	7	8
Matérn ($\nu = 3/2$)	explore	$0° - 90°$	no	5	5	8
Matérn ($\nu = 5/2$)	elcb	$0° - 90°$	no	5	6	7
Matérn ($\nu = 5/2$)	explore	$0° - 90°$	no	4	5	9
stdp	elcb	$0° - 180°$	no	3	9	-
stdp	explore	$0° - 180°$	no	4	7	-
stdp	elcb	$0° - 180°$	1x	3	5	-
stdp	explore	$0° - 180°$	1x	3	5	-
Matérn ($\nu = 3/2$)	explore	$0° - 180°$	1x	3	5	-
Matérn ($\nu = 5/2$)	explore	$0° - 180°$	1x	5	5	-

A.2 Distorted Ba_2NaOsO_6 1D

Table A.2: Summary on the performance of Bayesian optimization on the recreation of one-dimensional magnetic curves for distorted Ba_2NaOsO_6 . The table shows the number of DFT calculations that are necessary to reach a certain level of convergence. This level of convergence is defined as falling below and staying below a certain threshold value for the MSE. The threshold values are given in the first line of the table.

Settings			MSE			d_{min}
Kernel	**Acquisition function**	**Range**	$\mathbf{< 10^{-2}\,meV^2}$	$\mathbf{< 10^{-3}\,meV^2}$	$\mathbf{< 10^{-4}\,meV^2}$	$\mathbf{< 1^\circ}$
SE	elcb	$0^\circ - 90^\circ$	5	8	9	7
SE	explore	$0^\circ - 90^\circ$	2	6	9	5
Matérn $(\nu = 3/2)$	elcb	$0^\circ - 90^\circ$	2	8	10	-
Matérn $(\nu = 3/2)$	explore	$0^\circ - 90^\circ$	3	7	11	9
Matérn $(\nu = 5/2)$	elcb	$0^\circ - 90^\circ$	5	8	10	12
Matérn $(\nu = 5/2)$	explore	$0^\circ - 90^\circ$	3	8	9	9

A.3 UO_2 1D

Table A.3: Summary on the performance of Bayesian optimization on the recreation of one-dimensional magnetic curves for UO_2. The table shows the number of DFT calculations that are necessary to reach a certain level of convergence. This level of convergence is defined as falling below and staying below a certain threshold value for the MSE. The threshold values are given in the first line of the table.

Settings				MSE			d_{min}
Kernel	**Acquisition function**	**Range**	**symmetry**	$< \mathbf{10^{-1}\,meV^2}$	$< \mathbf{10^{-2}\,meV^2}$	$< \mathbf{10^{-3}\,meV^2}$	$< 1^\circ$
SE	elcb	$0^\circ - 90^\circ$	no	5	5	6	5
SE	explore	$0^\circ - 90^\circ$	no	5	5	7	7
Matérn $(\nu = 3/2)$	elcb	$0^\circ - 90^\circ$	no	5	5	9	5
Matérn $(\nu = 3/2)$	explore	$0^\circ - 90^\circ$	no	5	6	9	6
Matérn $(\nu = 5/2)$	elcb	$0^\circ - 90^\circ$	no	5	5	10	5
Matérn $(\nu = 5/2)$	explore	$0^\circ - 90^\circ$	no	5	5	-	5
stdp	elcb	$0^\circ - 180^\circ$	no	4	8	-	5
stdp	elcb	$0^\circ - 180^\circ$	1x	3	4	-	6
stdp	explore	$0^\circ - 180^\circ$	1x	3	4	7	6
Matérn $(\nu = 3/2)$	explore	$0^\circ - 180^\circ$	1x	5	5	9	5
Matérn $(\nu = 5/2)$	explore	$0^\circ - 180^\circ$	1x	5	5	8	5

A.4 Sr_2IrO_4 1D

Table A.4: Summary on the performance of Bayesian optimization on the recreation of one-dimensional magnetic curves for SIO. The table shows the number of DFT calculations that are necessary to reach a certain level of convergence. This level of convergence is defined as falling below and staying below a certain threshold value for the MSE. The threshold values are given in the first line of the table.

Settings		MSE		d_{min}	
Kernel	**Acquisition function**	$< \mathbf{10^{-1}\,meV^2}$	$< \mathbf{10^{-2}\,meV^2}$	$< \mathbf{2^\circ}$	$< \mathbf{1^\circ}$
SE	elcb	5	5	5	5
SE	explore	5	-	5	-
Matérn ($\nu = 3/2$)	elcb	6	-	15	-
Matérn ($\nu = 3/2$)	explore	4	-	4	4
Matérn ($\nu = 5/2$)	elcb	5	-	15	-
Matérn ($\nu = 5/2$)	explore	5	-	5	13

A.5 Undistorted Ba_2NaOsO_6 2D

Table A.5: Summary of the performance of the creation two-dimensional models of undistorted BNOO. The performance is evaluated on the basis of the model's ability to recreate the one-dimensional data from [60]. This ability is measured in the form of the mean squared error. The threshold was considered not reached at all (-) when the threshold was not reached within 36 DFT calculations.

Settings			MSE		d_{min}
Kernel	**Acquisition function**	**Symmetry**	$< \mathbf{10^{-1}}$	$< \mathbf{10^{-2}}$	$< \mathbf{1.8^\circ}$
SE	elcb	no	8	-	14
SE	elcb	3x	6	-	6
SE	explore	no	19	-	19
SE	explore	3x	11	-	4
Matérn ($\nu = 3/2$)	elcb	3x	7	13	7
Matérn ($\nu = 3/2$)	elcb	no	25	-	21
Matérn ($\nu = 3/2$)	explore	3x	4	15	23
Matérn ($\nu = 3/2$)	explore	no	21	-	24
Matérn ($\nu = 5/2$)	elcb	no	17	-	17
Matérn ($\nu = 5/2$)	elcb	3x	6	13	6
Matérn ($\nu = 5/2$)	explore	no	21	-	25
Matérn ($\nu = 5/2$)	explore	3x	5	11	3

A.6 Distorted Ba_2NaOsO_6 2D

Table A.6: Summary of the performance of the creation two-dimensional models of distorted BNOO. The performance is evaluated on the basis of the model's ability to recreate the one-dimensional data from [60]. This ability is measured in the form of the mean squared error. The threshold was considered not reached at all (-) when the threshold was not reached within 80 DFT calculations.

Settings			MSE	
Kernel	**Acquisition function**	**Symmetry**	$\mathbf{< 10^{-2}\,meV^2}$	$\mathbf{< 10^{-3}\,meV^2}$
SE	elcb	no	-	-
SE	elcb	1x	31	-
SE	explore	no	18	-
SE	explore	1x	38	-
Matérn ($\nu = 3/2$)	elcb	no	36	-
Matérn ($\nu = 3/2$)	elcb	1x	27	-
Matérn ($\nu = 3/2$)	explore	no	28	-
Matérn ($\nu = 3/2$)	explore	1x	24	63
Matérn ($\nu = 5/2$)	elcb	no	30	-
Matérn ($\nu = 5/2$)	elcb	1x	14	28
Matérn ($\nu = 5/2$)	explore	no	21	-
Matérn ($\nu = 5/2$)	explore	1x	20	62

B Abstract/Zusammenfassung

Abstract

Density functional theory (DFT) is the most established theoretical framework for the study of quantum materials in the context of first principles electronic structure methods. However, these calculations may require a significant computational cost, especially for large systems with strong spin-orbit coupling. These materials often show complex magnetic properties which make them interesting for both fundamental and applied research. Therefore, devising more efficient methods to approach their complex energy landscape is particularly desired and useful. One tool that has proven to be efficient in modeling unknown functions and finding global minima is the active machine learning scheme Bayesian Optimization. However, so far Bayesian Optimization has not been used for the investigation of noncollinear spin structures. The main goal of this thesis is to test if and how well Bayesian Optimization can be used as a tool to accelerate the exploration of the magnetic energy surface and to find the global magnetic ground states of selected noncollinear transition metal oxides. Specifically, we focus on Ba_2NaOsO_6, UO_2 and Sr_2IrO_4, three magnetic materials that exhibit strong spin-orbit coupling and have already been studied both experimentally and computationally. However, the existing computational studies that investigate the energy landscape of the three materials are limited to one-dimensional subspaces of the complete configuration spaces. This means that only one degree of freedom was taken into account in the search for the global energy minimum. Compared to the number of DFT calculations used in the corresponding literature, we were able to reproduce the results using between 44% and 91% fewer DFT calculations. Furthermore, we were able to investigate broader, previously unexplored parts of the configuration space by taking into account up to four degrees of freedom. The work presented in this thesis shows that Bayesian Optimization can be used to cut computational costs for noncollinear magnetic calculations within DFT. Creating models of a magnetic energy landscape that take into account more than one degree of freedom, which was previously often unfeasible due to limited computational resources, becomes possible with the help of Bayesian Optimization. This opens exciting new possibilities for the investigation of magnetic materials.

Zusammenfassung

Die Dichtefunktionaltheorie (DFT) ist eine der am weitesten verbreiteten ab initio Methoden zur Erforschung von Festkörpern auf quantenphysikalischer Ebene. Allerdings erfordern DFT-Rechnungen oft einen erheblichen Rechenaufwand, insbesondere für große Materialsysteme mit starker Spin-Bahn-Kopplung. Diese Materialien zeigen oft spannende magnetische Eigenschaften, die sie sowohl für grundlegende als auch für angewandte Forschung interessant machen. Daher ist es besonders wünschenswert und nützlich, effiziente Methoden zur Erforschung ihrer komplexen Energielandschaft zu entwickeln. Ein Werkzeug, das sich als ef-

fizient bei der Modellierung unbekannter Funktionen und der Suche nach globalen Minima erwiesen hat, ist Bayes'sche Optimierung, welches zum Bereich des aktiven maschinellen Lernens zählt. Bisher wurde die Bayes'sche Optimierung jedoch noch nicht für die Untersuchung von nicht kollinearen magnetischen Materialsystemen verwendet. Das Hauptziel dieser Arbeit ist es, zu testen, ob und wie gut die Bayes'sche Optimierung als Werkzeug zur Beschleunigung der Erforschung der magnetischen Energielandschaft und der magnetischen Grundzustände ausgewählter nicht-kollinearer Übergangsmetalloxide verwendet werden kann. Konkret konzentrieren wir uns auf Ba_2NaOsO_6, UO_2 und Sr_2IrO_4, drei magnetische Materialien, die eine starke Spin-Bahn-Kopplung aufweisen und bereits sowohl experimentell als auch computergestützt untersucht wurden. Die bestehenden computergestützten Studien, die die Energielandschaft der drei Materialien untersuchen, sind jedoch auf eindimensionale Unterräume der vollständigen Konfigurationsräume beschränkt. Das bedeutet, dass jeweils nur ein Freiheitsgrad bei den Untersuchungen berücksichtigt wurde. Im Vergleich zur Anzahl der in der entsprechenden Literatur verwendeten DFT-Rechnungen konnten wir die Ergebnisse unter Verwendung von 44% bis 91% weniger DFT-Rechnungen reproduzieren. Darüber hinaus ist es uns gelungen größere, zuvor unerforschte Teile des Konfigurationsraums durch die Berücksichtigung von bis zu vier Freiheitsgraden zu untersuchen. Die in dieser Arbeit vorgestellten Ergebnisse zeigen, dass die Bayes'sche Optimierung zur Reduzierung der Rechenkosten von DFT-Untersuchungen der magnetischen Eigenschaften von Festkörpern verwendet werden kann. Insbesondere das Erstellen von mehrdimensionalen Modellen einer magnetischen Energielandschaft wird durch die Verwendung der Bayes'schen Optimierung vereinfacht. Dies eröffnet aufregende neue Möglichkeiten für die Untersuchung magnetischer Materialien.

References

[1] Allan A. Mills. The lodestone: History, physics, and formation. *Annals of Science*, 61:273–319, 7 2004.

[2] M. Blackman. The lodestone: A survey of the history and the physics. *Contemporary Physics*, 24:319–331, 1983.

[3] Niels Bohr. II. The Doctor's Dissertation (Text and Translation). *Niels Bohr Collected Works*, 1:163–393, 1 1972.

[4] H-J Van Leeuwen. Problemes de la théorie électronique du magnétisme. *J. phys. radium*, 2(12):361–377, 1921.

[5] P. Hohenberg and W. Kohn. Inhomogeneous electron gas. *Physical Review*, 136:B864–B871, 11 1964.

[6] W. Kohn and L. J. Sham. Self-consistent equations including exchange and correlation effects. *Physical Review*, 140:A1133–A1138, 11 1965.

[7] Robin Haunschild, Andreas Barth, and Werner Marx. Evolution of DFT studies in view of a scientometric perspective. *Journal of Cheminformatics*, 8:52, 12 2016.

[8] Jonas Močkus. On bayesian methods for seeking the extremum. In *Optimization Techniques IFIP Technical Conference: Novosibirsk, July 1–7, 1974*, pages 400–404. Springer, 1975.

[9] J. Mockus. *The Bayesian approach to global optimization.* Springer-Verlag, 1989.

[10] Yunxing Zuo, Mingde Qin, Chi Chen, Weike Ye, Xiangguo Li, Jian Luo, and Shyue Ping Ong. Accelerating materials discovery with Bayesian optimization and graph deep learning. *Materials Today*, 51:126–135, 12 2021.

[11] Maituo Yu, Shuyang Yang, Chunzhi Wu, and Noa Marom. Machine learning the Hubbard U parameter in DFT+U using Bayesian optimization. *npj Computational Materials*, 6:180, 11 2020.

[12] Tatsuhito Ando, Naoto Shimizu, Norihisa Yamamoto, Nobuyuki N. Matsuzawa, Hiroyuki Maeshima, and Hiromasa Kaneko. Design of Molecules with Low Hole and Electron Reorganization Energy Using DFT Calculations and Bayesian Optimization. *The Journal of Physical Chemistry A*, 126:6336–6347, 9 2022.

[13] Milica Todorović, Michael U. Gutmann, Jukka Corander, and Patrick Rinke. Bayesian inference of atomistic structure in functional materials. *npj Computational Materials 2019 5:1*, 5:1–7, 3 2019.

[14] G. Kresse and J. Furthmüller. Efficient iterative schemes for *ab initio* total-energy calculations using a plane-wave basis set. *Physical Review B*, 54:11169, 10 1996.

[15] G. Kresse and J. Furthmüller. Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set. *Computational Materials Science*, 6:15–50, 7 1996.

[16] D. Hobbs, G. Kresse, and J. Hafner. Fully unconstrained noncollinear magnetism within the projector augmented-wave method. *Physical Review B*, 62:11556, 11 2000.

[17] D. R. Hartree. The Wave Mechanics of an Atom with a Non-Coulomb Central Field. Part II. Some Results and Discussion. *Mathematical Proceedings of the Cambridge Philosophical Society*, 24:111–132, 1 1928.

[18] V. Fock. “Selfconsistent field” mit Austausch für Natrium. *Zeitschrift für Physik*, 62:795–805, 11 1930.

[19] Antoine Georges and Gabriel Kotliar. Hubbard model in infinite dimensions. *Physical Review B*, 45:6479–6483, 3 1992.

[20] Jozef Noga and Rodney J. Bartlett. The full CCSDT model for molecular electronic structure. *The Journal of Chemical Physics*, 86:7041–7050, 6 1987.

[21] J. Herrin and J. S. Howland. The Born–Oppenheimer Approximation: Straight-Up and with a Twist. *Reviews in Mathematical Physics*, 09:467–488, 5 1997.

[22] M. Born and R. Oppenheimer. Zur Quantentheorie der Molekeln. *Annalen der Physik*, 389:457–484, 1927.

[23] P Gombás and P Gombás. Das statistische Modell von Thomas und Fermi. *Die Statistische Theorie des Atoms und ihre Anwendungen*, pages 30–76, 1949.

[24] Tanja Van Mourik, Michael Bühl, and Marie Pierre Gaigeot. Density functional theory across chemistry, physics and biology. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 372, 3 2014.

[25] Á. Nagy. Density functional. theory and application to atoms and molecules. *Physics Reports*, 298:1–79, 5 1998.

[26] M.D. Segall, P. J. Lindan, M.A. Probert, C.J. Pickard, P.J. Hasnip, S.J. Clark, and M.C. Payne. First-principles simulation: Ideas, illustrations and the CASTEP code. *Journal of physics: condensed matter*, 14(11):2717, 2002.

[27] John P. Perdew, John A. Chevary, Sy H. Vosko, Koblar A. Jackson, Mark R. Pederson, Dig J. Singh, and Carlos Fiolhais. Atoms, molecules, solids, and surfaces: Applications of the generalized gradient approximation for exchange and correlation. *Physical review B*, 46(11):6671, 1992.

[28] Axel D. Becke. Density-functional exchange-energy approximation with correct asymptotic behavior. *Physical review A*, 38(6):3098, 1988.

[29] Burak Himmetoglu, Andrea Floris, Stefano De Gironcoli, and Matteo Cococcioni. Hubbard-corrected DFT energy functionals: The LDA+U description of correlated systems. *International Journal of Quantum Chemistry*, 114(1):14–49, 2014.

[30] Matteo Cococcioni. The LDA+U approach: a simple Hubbard correction for correlated ground states. *Correlated Electrons: From Models to Materials Modeling and Simulation*, 2, 2012.

[31] Sarah A. Tolba, Kareem M. Gameel, Basant A. Ali, Hossam A. Almossalami, and Nageh K. Allam. The DFT+U: Approaches, accuracy, and applications. *Density Functional Calculations-Recent Progresses of Theory and Application*, 1:5772, 2018.

[32] John Hubbard. Electron correlations in narrow energy bands. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, 276:238–257, 11 1963.

[33] S. L. Dudarev, G. A. Botton, S. Y. Savrasov, C. J. Humphreys, and A. P. Sutton. Electron-energy-loss spectra and the structural stability of nickel oxide: An LSDA+U study. *Physical Review B*, 57:1505–1509, 1 1998.

[34] VI Anisimov and O Gunnarsson. Density-functional calculation of effective Coulomb interactions in metals. *Physical Review B*, 43(10):7570, 1991.

[35] Matteo Cococcioni and Stefano De Gironcoli. Linear response approach to the calculation of the effective interaction parameters in the LDA+ U method. *Physical Review B*, 71(3):035105, 2005.

[36] M Springer and F Aryasetiawan. Frequency-dependent screened interaction in Ni within the random-phase approximation. *Physical Review B*, 57(8):4364, 1998.

[37] Pui-Wai Ma and SL Dudarev. Constrained density functional for noncollinear magnetism. *Physical Review B*, 91(5):054420, 2015.

[38] Peter Mörters and Yuval Peres. *Brownian Motion.* Cambridge University Press, 1 2001.

[39] Carl Edward. Rasmussen and Christopher K. I. Williams. *Gaussian processes for machine learning.* MIT Press, 2006.

[40] Michael L. Stein. *Interpolation of Spatial Data: Some Theory for Kriging.* Springer, 1999.

[41] Bertil Matérn. *Spatial Variation*, volume 49, No.5. Springer-Verlag, second edition, 1986.

[42] Peter Whittle. On stationary processes in the plane. *Biometrika*, pages 434–449, 1954.

[43] Milton Abramowitz and Irene A. Stegun. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables.* U.S. Department of Commerce, National Bureau of Standards, 10th printing edition, 1972.

[44] Budiman Minasny and Alex B. McBratney. The Matérn function as a general model for soil variograms. volume 128, pages 192–207, 10 2005.

[45] William M. Bolstad and James M. Curran. *Introduction to Bayesian Statistics, Third Edition.* John Wiley & Sons, Inc., 8 2016.

[46] Eric Schulz, Maarten Speekenbrink, and Andreas Krause. A tutorial on Gaussian process regression: Modelling, exploring, and exploiting functions. *Journal of Mathematical Psychology*, 85:1–16, 8 2018.

[47] Richard von Mises and Hilda Geiringer. *Mathematical theory of probability and statistics.* Acad. Pr., 1964.

[48] Geoffrey Grimmett and Dominic JA Welsh. *Probability: An introduction.* OUP, second edition, 2014.

[49] BOSS developers team. Hyperparameter optimization. *https://cest-group.gitlab.io/boss/manual/keywords.html#hyperparameter-optimization*, 8 2023.

[50] Michael U. Gutmann and Jukka Corander. Bayesian optimization for likelihood-free inference of simulator-based statistical models. *Journal of Machine Learning Research*, 17:1–47, 2016.

[51] Rudolf Gross and Achim Marx. *Festkörperphysik.* De Gruyter, 1 2018.

[52] P.A.M. Dirac. The quantum theory of the electron. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, 117:610–624, 2 1928.

[53] Jairo Sinova and AH MacDonald. Theory of spin–orbit effects in semiconductors. *Semiconductors and Semimetals*, 82:45–87, 2008.

[54] Stephen Blundell. *Magnetism in condensed matter.* OUP Oxford, 2001.

[55] Shruba Gangopadhyay and Warren E. Pickett. Spin-orbit coupling, strong correlation, and insulator-metal transitions: The $J_{eff} = 3/2$ ferromagnetic Dirac-Mott insulator Ba_2NaOsO_6. *Physical Review B - Condensed Matter and Materials Physics*, 91, 1 2015.

[56] Dario Fiore Mosca, Hermann Schnait, Lorenzo Celiberti, Markus Aichhorn, and Cesare Franchini. The Mott transition in the $5d^1$ compound Ba_2NaOsO_6 : a DFT+DMFT study with PAW non-collinear projectors. *arXiv preprint arXiv:2303.16560*, 3 2023.

[57] A. S. Erickson, S. Misra, G. J. Miller, R. R. Gupta, Z. Schlesinger, W. A. Harrison, J. M. Kim, and I. R. Fisher. Ferromagnetism in the Mott insulator Ba_2NaOsO_6. *Physical Review Letters*, 99, 7 2007.

[58] Katharine E. Stitzer, Mark D. Smith, and Hans-Conrad zur Loye. Crystal growth of Ba_2MOsO_6 (M= Li, Na) from reactive hydroxide fluxes. *Solid State Sciences*, 4(3):311–316, 2002.

[59] L. Lu, M. Song, W. Liu, A.P. Reyes, P. Kuhns, H.O. Lee, I.R. Fisher, and V.F. Mitrović. Magnetism and local symmetry breaking in a Mott insulator with strong spin orbit interactions. *Nature communications*, 8(1):14407, 2017.

[60] Dario Fiore Mosca, Leonid V. Pourovskii, Beom Hyun Kim, Peitao Liu, Samuele Sanna, Federico Boscherini, Sergii Khmelevskyi, and Cesare Franchini. Interplay between multipolar spin interactions, Jahn-Teller effect, and electronic correlation in a $J_{eff} = 3/2$ insulator. *Physical Review B*, 103:104401, 3 2021.

[61] Hiroaki Ishizuka and Leon Balents. Magnetism in $s = \frac{1}{2}$ double perovskites with strong spin-orbit interactions. *Physical Review B*, 90(18):184422, 2014.

[62] J. H. Van Vleck. The Jahn-Teller Effect and Crystalline Stark Splitting for Clusters of the Form XY6. *The Journal of Chemical Physics*, 7:72–84, 1 1939.

[63] Paolo Giannozzi and Paul Erdös. Theoretical analysis of the 3-k magnetic structure and distortion of uranium dioxide. *Journal of magnetism and magnetic materials*, 67(1):75–87, 1987.

[64] R. A. Cowley and G. Dolling. Magnetic excitations in uranium dioxide. *Physical Review*, 167:464–477, 3 1968.

[65] P. Burlet, J. Rossat-Mignod, O. Vogt, J.C. Spirlet, J. Rebivant, et al. Neutron diffraction on actinides. *Journal of the Less Common Metals*, 121:121–139, 1986.

[66] R. Caciuffo, Giuseppe Amoretti, Paolo Santini, Gerard Heath Lander, J. Kulda, and P. de V. Du Plessis. Magnetic excitations and dynamical jahn-teller distortions in uo 2. *Physical Review B*, 59(21):13892, 1999.

[67] Robert Laskowski, Georg KH Madsen, Peter Blaha, and Karlheinz Schwarz. Magnetic structure and electric-field gradients of uranium dioxide: An ab initio study. *Physical Review B*, 69(14):140408, 2004.

[68] G. Amoretti, A. Blaise, R. Caciuffo, J. M. Fournier, M. T. Hutchings, R. Osborn, and A. D. Taylor. $5f$ -electron states in uranium dioxide investigated using high-resolution neutron spectroscopy. *Physical Review B*, 40:1856–1870, 7 1989.

[69] Elizabeth Blackburn, Roberto Caciuffo, Nicola Magnani, Paolo Santini, PJ Brown, M Enderle, and GH Lander. Spherical neutron spin polarimetry of anisotropic magnetic fluctuations in UO_2. *Physical Review B*, 72(18):184411, 2005.

[70] K. Ikushima, S. Tsutsui, Y. Haga, H. Yasuoka, R.E. Walstedt, N.M. Masaki, A. Nakamura, S. Nasu, and Y. Ōnuki. First-order phase transition in UO_2: ^{235}U and ^{17}O NMR study. *Physical Review B*, 63(10):104404, 2001.

[71] S.B. Wilkins, R. Caciuffo, C. Detlefs, Jean Rebizant, Eric Colineau, Franck Wastin, and GH Lander. Direct observation of electric-quadrupolar order in UO_2. *Physical Review B*, 73(6):060406, 2006.

[72] S.L. Dudarev, P. Liu, D.A. Andersson, C.R. Stanek, T. Ozaki, and C. Franchini. Parametrization of LSDA+U for noncollinear magnetic configurations: Multipolar magnetism in UO_2. *Physical Review Materials*, 3(8):083802, 2019.

[73] B.J. Kim, Hosub Jin, S.J. Moon, J.-Y. Kim, B.-G. Park, C.S. Leem, Jaejun Yu, T.W. Noh, C. Kim, S.-J. Oh, et al. Novel $J_{eff} = 1/2$ Mott state induced by relativistic spin-orbit coupling in Sr_2IrO_4. *Physical review letters*, 101(7):076402, 2008.

[74] S.J. Moon, Hosub Jin, Kyung Wan Kim, W.S. Choi, Y.S. Lee, Jaejun Yu, G. Cao, A. Sumi, H. Funakubo, Christian Bernhard, et al. Dimensionality-Controlled Insulator-Metal Transition and Correlated Metallic State in $5d$ Transition Metal Oxides $Sr_{n+1}Ir_nO_{3n+1}$ (n = 1,2, and ∞). *Physical review letters*, 101(22):226402, 2008.

[75] Peitao Liu, Sergii Khmelevskyi, Bongjae Kim, Martijn Marsman, Dianzhong Li, Xing-Qiu Chen, DD Sarma, Georg Kresse, and Cesare Franchini. Anisotropic magnetic couplings and structure-driven canted to collinear transitions in Sr_2IrO_4 by magnetically constrained noncollinear DFT. *Physical Review B*, 92(5):054428, 2015.

[76] M.K. Crawford, M.A. Subramanian, R.L. Harlow, J.A. Fernandez-Baca, Z.R. Wang, and D.C. Johnston. Structural and magnetic studies of Sr_2IrO_4. *Physical Review B*, 49(13):9198, 1994.

[77] Stefano Boseggia, H.C. Walker, J. Vale, R. Springell, Z. Feng, R.S. Perry, M. Moretti Sala, Henrik M. Rønnow, S.P Collins, and Desmond Francis McMorrow. Locking of iridium magnetic moments to the correlated rotation of oxygen octahedra in Sr_2IrO_4 revealed by x-ray resonant scattering. *Journal of Physics: Condensed Matter*, 25(42):422202, 2013.

[78] Paolo Santini, Stefano Carretta, Giuseppe Amoretti, Roberto Caciuffo, Nicola Magnani, and Gerard H. Lander. Multipolar interactions in f-electron systems: The paradigm of actinide dioxides. *Reviews of Modern Physics*, 81(2):807, 2009.

[79] Shuxiang Zhou, Hao Ma, Enda Xiao, Krzysztof Gofryk, Chao Jiang, Michael E. Manley, David H. Hurley, and Chris A. Marianetti. Capturing the ground state of uranium dioxide from first principles: Crystal distortion, magnetic structure, and phonons. *Physical Review B*, 106(12):125134, 2022.