Practical discretisation error estimates for data-driven materials modelling

Michael F. Herbst, EPFL

1 Current state of research in the field

1.1 Perspectives of discretisation error control in data-driven materials modelling

The development of novel materials is a key ingredient to tackle major 21st century challenges, such as health care, green chemistry, energy production and storage. Indeed, many of our current technological solutions are fundamentally limited by the physical and chemical properties of the materials we have at hand. To overcome these fundamental limits, the computational discovery of novel materials has become a vital tool, as exemplified by numerous success stories leading to the development of novel semiconductors and materials for electrocatalysis, hydrogen storage, Li-ion batteries [1–3]. In recent years data-driven approaches have brought considerable acceleration to computational materials discovery [4, 5] and radically changed the field. For example, machine-learned interatomic potentials trained on first-principle simulations — typically Kohn-Sham density-functional theory (DFT) — have increased the accessible size and time scales by orders of magnitude [6–11] while maintaining the accuracy level of the underlying DFT simulation.

However, a marked downside of data-driven methods is the need for large amounts of training data — sometimes requiring tens of millions of DFT simulations [12]. Moreover, since all common atomistic machine learning approaches model the error in the training data by small Gaussian noise, the conducted simulations need to be of homogeneously good quality. This in turn requires the choice of sufficiently good numerical parameters (discretisation basis, k-point sampling, tolerances) for the entire data generation process, such that the *numerical error* for *all* systems of the dataset is small. This drives up the computational cost considerably, which limits both the size of the training dataset as well as the size of the individual materials systems *in* the dataset. Subtle finite-size effects can therefore be sometimes baked into the trained model [13]. Furthermore even the selection of such numerical parameters which provide consistently good accuracy for large diverse datasets can be challenging: even some standard datasets of the atomistic machine learning community are known to feature artefacts and data inconsistencies. An example is the QM9 [14] dataset, where some structures are known to be "unreliable" and need to be excluded in order to not impact the quality of a trained surrogate [15]. Similarly the initial version of the MD17 dataset [16] turned out to be based on too loose numerical parameters, such it was later recomputed, leading to rMD17 [17].

When it comes to understanding and rigorously estimating the numerical error in plane-wave-based DFT simulations there has been remarkable progress in the mathematical community in recent years [18–24]. In particular with respect to the *discretisation error*, i.e. the error due to the chosen plane-wave basis set, a promising perturbation-based error estimation approach has been proposed. On selected materials systems and for selected pseudopotential models this approach has been shown to provide an efficient and accurate annotation of DFT energies and forces with their accompanying discretisation error [21]. From the point of view of data-driven materials modelling such error estimates unlock an unprecedented opportunity to move beyond a simple Gaussian-noise model and more quantitatively describe the error in the training data. If this can be exploited successfully during inference, this would overcome aforementioned limitations (a) by making statistical surrogates more resilient to heterogeneous and inconsistent training data and (b) by exploiting this resilience to save costs during training data generation — e.g. by purposefully employing smaller plane-wave basis sets when running DFT simulations on systems of previously unfeasible size.

Motivated by these opportunities the proposed work strives to achieve two main objectives. First, to develop

the perturbation-based error estimates into a routine tool for materials modelling and overcome their current limitations (see Sections 1.2.3 and 1.2.4). Amongst others this requires systematic testing on a broad range of materials systems as well as extension of the existing error estimation framework to include key DFT quantities of interest (QoI) — such as forces, stresses, relaxed structures and band structures. Second, to employ such quantitative error estimates to replace the assumption of a homogeneous error within Gaussian Process (GP) regression (Section 1.3.1). Here we will investigate (a) a heteroscedastic noise model as well as (b) an approach including the discretisation error as an additional quantity to regress. The resulting GP model will have the novel ability to propagate the DFT discretisation error through the surrogate — an aspect which we will exploit to reduce data generation cost while keeping the prediction accuracy of the final surrogate under control. To ensure our methods will be applicable for practical data-driven materials modelling we will employ standard datasets to guide and validate our developments throughout the research programme.

1.2 Practical *a posteriori* error estimation in density-functional theory

DFT methods are a common ingredient in materials modelling, since they are able to accurately simulate many optical, mechanical and chemical properties at an acceptable cost. The outcome of such a DFT simulation is naturally associated with a number of errors. This includes the *model error*, i.e. the error of the DFT approximation itself. Additionally there is the *numerical error* resulting from the basis set discretisation as well as the algorithmic procedure that is used to solve the underlying equations. Mathematically the DFT problem (see (1)) involves a non-convex Riemannian optimization of a non-linear energy functional with close linkage to a non-linear eigenvalue problem. Despite this complexity the past few years have seen considerable progress towards practical *a posteriori* error estimation techniques [18–24]. This includes some work by my collaborators and myself on linear Kohn-Sham models [19], i.e. where the non-linear terms in the energy are dropped. On the full DFT side a promising direction are approaches based on perturbative post-processing [18], including the recent work by my project partner to estimate the discretisation error for DFT energies and forces [21]. A first implementation and initial testing of these estimates has already been realised using DFTK, one of the core software on which this proposal relies (see Section 2). One thread of the proposed work will be to develop these estimates further into a readily applicable tool in practical materials modelling, overcoming remaining key limitations (see Sections 1.2.3 and 1.2.4).

1.2.1 Mathematical formulation of DFT

In the density matrix formalism the mathematical problem underlying DFT is the minimisation problem

$$\min_{P \in \mathcal{P}} E(P), \tag{1}$$

of the non-linear DFT energy functional $E(P)$ over the manifold $\mathcal{P}$ of density matrices (2). For notational simplicity we assume we model a periodic insulating system and further directly jump to the discretised setting: we employ a reference basis $\mathcal{X}_{\texttt{Eref}}$, consisting of a large number $\mathcal{N} = |\mathcal{X}_{\texttt{Eref}}|$ of plane-wave basis functions, i.e.

$$\mathcal{X}_{\texttt{Eref}} = \operatorname{span}\left\{ \frac{1}{\sqrt{\Gamma}} e^{iG\cdot x} \middle| \frac{1}{2}\|G\|^2 \leq \texttt{Eref}, \right\},$$

where $\texttt{Eref}$ is the kinetic energy cutoff, Γ the unit cell volume and G the frequency of the plane wave. Then

$$\mathcal{P} = \left\{ P \in \mathcal{H}, P^2 = P, \operatorname{tr}(P) = N \right\}, \qquad \mathcal{H} = \left\{ H \in \mathbb{C}^{\mathcal{N}\times\mathcal{N}}, H^\dagger = H \right\}, \tag{2}$$

where N is the number of electrons of the modelled system. Note that to solve problem (1) in practice one typically employs that all density matrices $P \in \mathcal{P}$ are rank N, suggesting a parametrisation $P = \sum_{i=1}^{N} \phi_i \phi_i^\dagger$ in

which the unknowns become the N orbitals $\phi_i \in \mathbb{C}^{\mathcal{N}}$. Either by directly minimising the energy (1) with respect to the orbitals $\{\phi_i\}_{i=1}^N$ or by satisfying its first-order optimality conditions — the so-called self-consistent field equations (SCF) — one then solves the problem iteratively. Based on the resulting minimiser P_* other quantities of interest (QoI) can be computed, such as the forces $F(P_*)$ (energy derivative wrt. atomic positions), the stresses $S(P_*)$ (energy derivative wrt. lattice parameters) or the band structure (eigenvalues of the Hamiltonian).

1.2.2 Perturbation-based error estimates for density-functional theory

A successful strategy to obtain a practical estimate of the discretisation error is based on a two-basis approach [19, 21, 22, 25]. On the smaller basis $\mathcal{X}_{\texttt{Ecut}}$ with $\texttt{Ecut} \ll \texttt{Eref}$ the full iterative DFT calculation is performed, which yields a converged density matrix we call P. For this outcome the discretisation error can be estimated by computing the first-order change under an increase of the discretisation. Assume on the reference basis $\mathcal{X}_{\texttt{Eref}}$ the solution is P_*, then the correction of P to P_* can be approximated as [21]

$$\delta P = P - P_* \simeq (\boldsymbol{\Omega} + \boldsymbol{f})^{-1} R \tag{3}$$

where $R = \Pi_P H(P)$ is the DFT residual of P, i.e. the Kohn-Sham Hamiltonian at P projected onto the tangent plane at P, which is non-zero in the reference basis $\mathcal{X}_{\texttt{Eref}}$. Further $\boldsymbol{f} = \Pi_P \nabla^2 E(P) \Pi_P$ and $\boldsymbol{\Omega}$ is a geometric component to the second derivative due to the manifold structure [21, 26]. Given δP as an approximation of the discretisation error of the density matrix, the error of other quantities of interest is again obtained, formally, by a first-order development, e.g.

$$\delta F = F(P) - F(P_*) = \left.\frac{dF}{dP}\right|_P \cdot \delta P \tag{4}$$

From a computational point of view (3) is not cheap — essentially as expensive as solving the minimisation problem (1) within $\mathcal{X}_{\texttt{Eref}}$, i.e. of asymptotic cost $\mathcal{O}\left(\texttt{Eref}^{3/2} \log(\texttt{Eref})\right)$. However, it turns out that for the forces the discretisation error is dominated by the low-frequency components $(\delta P)_1$ — i.e. the part of δP within the small $\mathcal{X}_{\texttt{Ecut}}$ basis [21]. This justifies further approximations [21] and one obtains, from a Schur complement,

$$(\delta P)_1 = (P - P_*)_1 \simeq (\boldsymbol{\Omega} + \boldsymbol{f})_{11}^{-1} \left(R_1 - (\boldsymbol{\Omega} + \boldsymbol{f})_{12} \boldsymbol{M}_{22}^{-1} R_2\right) \tag{5}$$

where in each case an index 1 indicates a projection into the small $\mathcal{X}_{\texttt{Ecut}}$ basis and a 2 a projection into $\mathcal{X}_{\texttt{Eref}} \setminus \mathcal{X}_{\texttt{Ecut}}$. Further $\boldsymbol{M}_{22}$ is a high-frequency approximation to $(\boldsymbol{\Omega} + \boldsymbol{f})_{22}$, which is cheap to invert [21]. On the full $\mathcal{X}_{\texttt{Eref}}$ basis this approach only requires the inversion of $\boldsymbol{M}_{22}$ and a single application of $(\boldsymbol{\Omega} + \boldsymbol{f})_{12}$. The expensive iterative inversion now only involves $(\boldsymbol{\Omega} + \boldsymbol{f})_{11}$, which is computed in the small $\mathcal{X}_{\texttt{Ecut}}$ basis. On the problems tested in [21] as well as in our preliminary work (Section 2.2) a tuning of `Eref` versus `Ecut` could be realised, such that obtaining an error estimate for a calculation based on $\mathcal{X}_{\texttt{Ecut}}$ required about the same additional cost as solving (1) itself.

Since the corrections δP and δF are not just obtained as error norms, but as full vector-valued quantities, one can also think of (4) as a **refinement strategy** for the forces beyond the results obtained in $\mathcal{X}_{\texttt{Ecut}}$. Provided a good tuning of `Eref` versus `Ecut` this offers an approach to reduce `Ecut` below the standard recommended values in large systems. Notably one can still (a) estimate the discretisation error (since corrections like δF can be computed) and (b) keep the accuracy loss minimal (as the refined $F + \delta F$ is available).

1.2.3 Practical limitation: Selection of cutoffs and tested pseudopotential models

The first order expansion underlying (3) (and implicitly (5)) is only meaningful if P is asymptotically close to P_*. Moreover the correction δP only provides a meaningful estimate of the discretisation error if P_* is close to the true (infinite basis) solution. This raises the question how the cutoff `Eref` should be chosen in relationship to

`Ecut`. Too large a value for `Eref` increases computational cost, too small a value may render the error estimates inaccurate. The original work [21] so far only performed preliminary tests on a few systems. Moreover they only employed Goedecker-type pseudopotentials [27, 28] — a pseudopotential family which is rarely considered in practice, since the resulting models are too "hard", i.e. require too large an `Ecut` for sufficient convergence. A validation of the perturbation-based error estimates for other systems and pseudopotential models as well as a determination of a good heuristics to choose `Eref` based on `Ecut` are thus required to make this estimation strategy feasible in practice. Some of this we started in preliminary work (see Section 2.2).

1.2.4 Practical limitation: Metallic systems and available DFT quantities

Another limitation for the applicability of the discretisation error estimates is the set of materials systems and the set of quantities of interest (QoIs), which can currently be treated. Since the error estimation framework of [21] assumes the DFT minimisation (1) to run over the manifold $\mathcal{P}$ of projectors, this requires the modelled material to be gapped, i.e. an insulator or semiconductor. For metallic systems the usual structure of (1) is different (minimisation runs over the convex hull of $\mathcal{P}$) and therefore needs additional developments. Regarding DFT quantities the developments of [21] currently only offer error estimates for energies, densities and forces. In this work we want to develop further perturbation-based error estimates for stresses and band structures as well as the error in the optimal structural parameters after a structure relaxation — that is the determination of the structural parameters (atomic positions and lattice parameters), which minimise the DFT energy. In principle (3) and (4) provide a framework where the error estimate of any $Q(P)$ could be computed as soon as the derivative $\frac{dQ}{dP}$ was available. However, to employ the computationally tractable expression (5) an additional assumption was needed, namely that the derivative $\frac{dQ}{dP}$ has its most dominant support on the low-frequency components, i.e. on the basis $\mathcal{X}_{\texttt{Ecut}}$. For the QoIs we target in this work this assumption remains to be checked.

1.3 Gaussian process regression in atomistic modelling

Machine learning techniques are well-established to replace density-functional theory (DFT) simulations by cheaper statistical surrogates. One popular tool in atomistic modelling — and the focus of this work — is Gaussian process (GP) regression. In the past GP regression has been successfully applied to accelerate a multitude of standard modelling tasks, such as the construction of machine-learned interatomic potentials [6, 29–31], structure relaxations [32] or global structure searches [33–36].

In a nutshell the goal of GP regression is to build an approximate statistical model for a mapping $X_* \mapsto Q(X_*)$, where one has only access to n noisy observations $\{X_i, Y_i\}_{i=1}^n$. When targeting an interatomic potential the $Q(X_i)$ are usually the atomic forces and the X_i a feature vector (such as SOAP [30, 37, 38]) to represent the atomistic structure of system i in a symmetry-adapted fashion [29]. In this proposal the observations Y_i will usually be the energies and forces resulting from a DFT simulation of system i. Compared to the "true" $Q(X_i)$ both the chosen DFT functional as well as the chosen numerical parameters induce an error, which is usually modelled as noise. A typical error model is

$$Y_i \;=\; Q(X_i) \;+\; \varepsilon_i, \qquad \varepsilon_i \;\sim\; N\left(0,\; \sigma^2\right), \tag{6}$$

where the noises ε_i are considered to be independent and to follow a normal distribution with mean zero and variance σ^2. In GP regression we additionally assume a Gaussian process prior

$$Q(X) \;\sim\; \mathrm{GP}\Big(\mu(X), k(X, X')\Big), \tag{7}$$

where $\mu(\,\cdot\,)$ is a chosen mean function — to describe the expected value of $Q(X)$ — and $k(\,\cdot\,,\,\cdot\,)$ a kernel function — to encapsulate our assumptions about the relationship between $Q(X_i)$ and $Q(X_j)$. We perform inference using

Bayes' theorem: Conditioned on our observed data $\{Y_i\}_{i=1}^n$ we want to predict the distribution of $Q(X_*)$ for an unseen system with features X_*. For a GP prior this can be done analytically leading to a multivariate Gaussian as the posterior distribution [39]

$$Q(X_*) \,\big|\, \boldsymbol{Y} \;\sim\; N\Big(\mu(X_*) \;+\; \boldsymbol{k}^T(\boldsymbol{K}+\boldsymbol{\Sigma})^{-1}\big(\boldsymbol{Y}-\boldsymbol{\mu}\big), \quad k(X_*,X_*) \;-\; \boldsymbol{k}^T(\boldsymbol{K}+\boldsymbol{\Sigma})^{-1}\boldsymbol{k}\Big), \tag{8}$$

where $\boldsymbol{\Sigma} = \mathrm{diag}(\sigma^2,\ldots,\sigma^2)$ is the diagonal matrix of n times σ^2 and we collected the vectors $\boldsymbol{\mu} = (\mu(X_1),\ldots,\mu(X_n))^T$, $\boldsymbol{Y} = (Y_1,\ldots,Y_n)^T$, $\boldsymbol{k} = (k(X_1,X_*),\ldots,k(X_n,X_*))^T$ as well as the kernel matrix $\boldsymbol{K}$ with elements $(\boldsymbol{K})_{ij} = k(X_i,X_j)$. By its probabilistic nature the trained GP surrogate thus not only yields a value, but a training-data-dependent posterior *distribution* (8). The posterior mean can be interpreted as the prediction for $Q(X_*)$ and the posterior variance as an associated uncertainty estimate. This appealing feature is the basis for many established uncertainty quantification schemes for GP-based surrogates for atomistic simulations [15, 40–42].

1.3.1 Current limitation: Homoscedastic noise and requirement for homogeneous data

Most if not all practical GP regression schemes in atomistic modelling employ the data model of (6), where the noise is assumed to be homoscedastic — i.e. of equal variance across the entire dataset. This effectively encodes an assumption of a homogeneous training data quality. If this is violated, both the quality of the GP's predictions as well as the usefulness of the posterior variance as an uncertainty estimate deteriorate.

In the GP literature a few approaches to overcome this limitation have been developed. This includes multi-fidelity [43, 44] or multitask [45] regression, in which a joint GP-based surrogate is constructed to model both multiple fidelity layers as well as the disparity between them. Strategies to define such fidelity layers can vary and include e.g. a definition as different sets of numerical parameters or different physical models. The appealing features of these models is to be able to improve the predictive quality at the highest fidelity level by employing larger quantities of cheaper low-fidelity data. Recently these ideas have seen initial applications in atomistic modelling [46–52] including my collaborative work to explore multitask regression for the first time [52]. However, most of these techniques still employ a homoscedastic noise model for the DFT error. They are thus similarly unable to benefit from quantitative error estimates — such as the perturbation-based estimates sketched in Section 1.2. To overcome this limitation the proposed work develops inference approaches for both single-task and multitask GP regression, which are able to take such quantitative error estimates into account — a first in atomistic modelling. The resulting models will become more resilient to heterogeneous training data and thus allow varying the plane-wave cutoff during data generation, which can be exploited to reduce cost.

2 Own contributions and preliminary work

In the past I have worked extensively on improving the efficiency and reliability of DFT-based materials simulations [19, 52–56]. This included both more robust algorithms based on the numerical analysis of DFT methods [19, 54–56] as well as more recently Gaussian process (GP) regression techniques tailored for training on heterogeneous data [52] — in this case data from a mixture of first-principle models. Both my project partner and myself have previously contributed to the development of error estimation techniques for Kohn-Sham models. This includes my work on guaranteed error estimates for bandstructures in linear Kohn-Sham [19] as well as Gaspard Kemlin's work on non-linear models [24] and DFT [21] — the latter being extended as part of the proposed programme.

A crucial component of our previous research at the edge of mathematics and materials modelling has been the Density-Functional ToolKit (DFTK) [53] — an open-source Julia-based [57] DFT package, which has been started about 5 years ago. In only 7500 lines of code DFTK supports a good range of standard methods (norm-

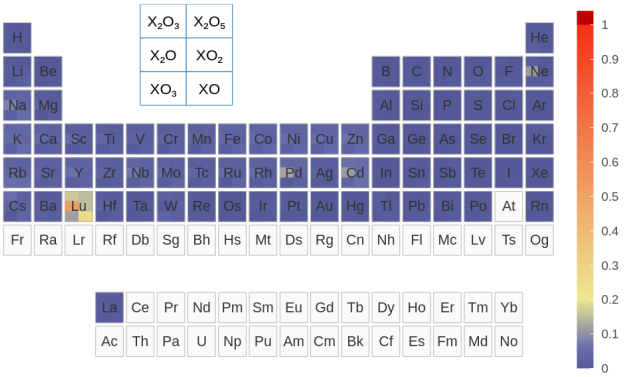


Figure 1: Verification of DFTK against the established Quantum Espresso [64] code using the automated workflows of [65] on the 432 virtual oxide structures of the **verification dataset** (see Section 3.2). Displayed is the ε metric from [65], which attests DFTK a consistent good agreement ($\varepsilon < 0.2$, deep blue colouring) on all structures save lutetium (Lu). This showcases the ability to use the AiiDA [66, 67] high-throughput engine to perform automated testing of the algorithms implemented in DFTK on a large number of materials systems.

conserving pseudopotentials, meta-GGA functionals) [58] as well as an integration with standard Julia HPC libraries (MPI, threading) enabling the simulation of systems up to 1000 electrons. At the same time simplified or customised models (e.g. linear Kohn-Sham, 1D or 2D setups) — often essential in mathematical research — are fully supported. A unique feature is DFTK's support for forward-mode **algorithmic differentiation (AD)**. This enables the computation of in principle any derivative of an output quantity of the code (bands, density of states, density, ...) wrt. any input quantity (atomic positions, lattice, DFT model, ...) — a first for plane-wave DFT. In the context of discretisation error estimation this has already been exploited in [21] to compute the derivative $\frac{dF}{dP}$ in (4) and we will also employ it in **WP2**. Overall DFTK's unique combination of a flexible *and* efficient code base has already attracted considerable community interest and has been instrumental in about a dozen publications in electronic structure theory, numerical analysis and mathematical physics [19, 21, 22, 24, 26, 54–56, 59–63] and further research is ongoing.

2.1 AiiDA+DFTK: Automated testing of mathematical algorithms

About a year ago my group and myself joined the efforts of the SNSF National Centre of Competence in Research "MARVEL". A first outcome of our involvement is a preliminary version of a plugin [68] to interface DFTK with the AiiDA workflow manager [66, 67]. In collaboration with Giovanni Pizzi (PSI) we employed this plugin to compare the DFT algorithms underlying DFTK against the established DFT package Quantum Espresso [64] in a *fully automated* fashion. The agreement is good to very good, see Figure 1. Performing such large-scale automated tests will be a central component in work packages **WP1** to **WP3** to overcome the current practical limitations of perturbation-based error estimates (see Sections 1.2.3 and 1.2.4).

2.2 Discretisation error estimates for other pseudopotentials

Previous numerical experiments of the perturbation-based error estimates [21] have been limited to DFT simulations employing Goedecker-type pesudopotentials [27]. The closed-form analytic expressions of these pseudopotentials makes them mathematically rather tractable and ensures that the asymptotic limit with respect to an increasing plane-wave discretisation basis $\mathcal{X}_{\texttt{Ecut}}$ is quickly established [19]. However, practical materials simulations more frequently employ other norm-conserving pseudopotential families, such as the pseudodojo [69] and the SG15 [70] potentials. In these potentials the functional form is only represented by data on a real-space grid, which is interpolated onto the plane-wave basis. As a result the convergence behaviour with

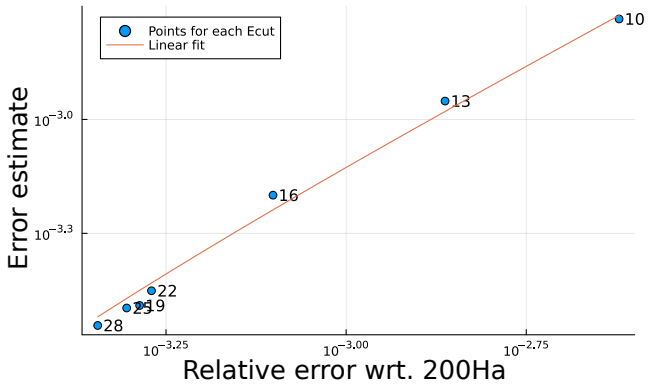


Figure 2: Excellent correlation of the relative l_2-error in the density between the "true" error (computed against a reference at 200 Hartree) and the perturbation-based error estimate of [21] employing the relationship `Eref` = `Ecut` + 30 Hartree. The chosen value of `Ecut` is indicated by small numbers next to each data point. We consider a silicon unit cell and the stringent pseudopotential from the pseudodojo database [69], thus demonstrating an extension of the error estimates of [21] beyond Goedecker-type pseudopotentials.

respect to increasing `Ecut` is more involved and it is not a priori clear whether perturbation-based estimates work well.

To test whether the perturbation-based estimates of [21] can still be employed on such practically relevant pseudopotentials we conducted DFT simulations of the silicon unit cell using both pseudodojo and SG15-type potentials. For both we obtained an excellent agreement between the "true" error and the estimated error, e.g. Figure 2. Based on this silicon system we also started tackling the other limitation of Section 1.2.3, namely to establish a good heuristics for choosing `Eref` based on `Ecut`. We found the relationship `Eref` = `Ecut` + 30 Hartree to provide a good strategy to balance the cost of solving (5) with the accuracy of the resulting error estimate. Moreover, if one employs this heuristic and for `Ecut` one uses the *low* recommended values of http://www.pseudo-dojo.org(intended only to give fast and crude results) then the refined density $(P + \delta P)$ and forces $(F + \delta F)$ turn out to have *smaller error* than a non-refined DFT simulation employing the *high* recommended `Ecut` of http://www.pseudo-dojo.org. Notably, in these experiments the computational time for solving both (1) plus (5) at low-`Ecut` level turns out to be comparable to the time for just solving (1) at high-`Ecut`. This demonstrates that the **perturbation-based error estimates** can work well **for pseudodojo-type potentials** and could even offer a novel **strategy to reduce the `Ecut`** employed in DFT simulations while **not compromising the quality** of the simulation outcome.

3 Research plan

3.1 Project goals

Our preliminary work (Section 2 and [21]) has already implemented first discretisation error estimation techniques within the density-functional toolkit (DFTK) and verified their applicability for selected materials, quantities of interest (QoIs) and pseudopotential models. We will develop these techniques further and integrate them into Gaussian Process regression to unlock novel opportunities for saving computational cost during data generation. Our research is guided and our outcomes tested employing datasets from materials modelling practice, see Section 3.2. We strive to achieve the following goals:

Turn discretisation error estimation into a practical tool (goal G1). Employ the representative set of chemical environments of the **verification dataset** to overcome the current practical limitations of the error estimates. That is: validate the discretisation error estimates over a representative range of chemical

environments and develop guidelines for choosing parameters balancing accuracy and cost.

Error estimation techniques for further key quantities of materials modelling (goal G2). Beyond the existing quantities (energies, densities and forces) we will develop error estimation strategies for stresses, band structures and the optimal structural parameters (atomic positions and lattice) determined by structure relaxation.

Error estimation for metallic systems (goal G3). Currently the available error estimation strategies assume idempotent density matrices (2), thus integer orbital occupations. We will extend the error estimation formalism to include the setting of non-integer occupation, i.e. a non-zero electronic temperature. This is the adequate regime when simulating metallic systems.

GP regression informed of discretisation error (goal G4). Develop a Gaussian process (GP) regression approach, in which the discretisation error is explicitly included, e.g. as part of a heteroscedastic noise model (see Section 1.3.1) or as an additional quantity to predict by regression. With this error-informed GP regression framework and with respect to the **silicon dataset** we will develop practical strategies to reduce data generation cost by tracking rigorously the effect of the discretisation error on the quality of the GP's predictions.

Error-informed multitask GP regression (goal G5). We will investigate the generalisation of **goal G4** to multitask GP regression approaches [45, 52]. The multiple tasks will be defined as the DFT quantities and error estimates resulting when multiple discretisation bases $\mathcal{X}_{\texttt{Ecut}}$ are used to simulate a structure. Since multitask models are able to exploit correlation between the tasks. This enables to make predictions at the quality of the highest-`Ecut` basis, but without the need to compute all training structures at this level. We will explore this flexibility to further reduce data generation cost.

3.2 Focus datasets

Throughout the proposal our methodological developments will be guided by two established datasets in the community. These datasets contain sufficient variety and are sufficiently representative, such that achieving our goals for these systems provides strong evidence that our methods work in general for practical materials modelling. Furthermore these systems are sufficiently simple, such that our fundamental research programme establishing a novel link between analytical and statistical techniques for error estimation is feasible.

Verification dataset. A recent verification study employed equation of state curves on 6 virtual oxides and 4 unary systems per element of the periodic table to compare DFT codes against all-electron references [65]. Employing only small and computationally tractable unit cells the structures of this study systematically cover a range of chemical environments. Since this dataset is integrated into the AiiDA workflow manager [66, 67] and thus readily accessible from DFTK (see Section 2.1), it provides an ideal automated validation framework when developing the error estimates in **WP1** to **WP3**.

Silicon dataset. Due to the practical relevance of silicon systems, the construction of machine-learned interatomic potentials for silicon structures is well-studied [6, 71–74]. In this work we will employ the dataset of [71] for three reasons. First, it has been constructed exactly for fitting potentials using Gaussian process (GP) regression [71], thus allowing an easy comparison of our developments (**WP4** and **WP5**) with existing results. Second, the dataset features silicon systems of varying sizes, from one to a few hundred atoms. Therefore it provides a good setup to investigate whether the development of error-informed GP regression indeed enables to lower the plane-wave cutoff `Ecut` for larger systems without impacting the quality of the potential. Finally, our preliminary results (Section 2.2) indicate that perturbation-based estimates work well for silicon unit cells.

Since the convergence behaviour with increasing `Ecut` mainly depends on the chosen pseudopotential model (see discussion in Section 3.4.1), we expect the estimates to work on this entire dataset as well.

3.3 Research team and key collaborators

The proposed project covers topics in numerical analysis, computational condensed matter physics and statistics. To ensure success a diverse team with multiple disciplinary backgrounds is hence needed. The following lists the profiles of the research team and new employees:

- **Doctoral candidate (DC):** The candidate has either a background in computational mathematics and an interest in running practical simulations *or* conversely prior experience with first-principle modelling as well as an interest in the mathematical analysis of DFT models. Main involvement in **WP1** to **WP3**.
- **Postdoctoral researcher (PR):** Main involvement in **WP4** and **WP5**, which feature the combination of a number of mathematical concepts. An experienced researcher in the development of advanced GP regression techniques (e.g. multitask, multi-fidelity), ideally in the context of machine learning for atomistic systems, is thus needed.
- **Gaspard Kemlin**, Université de Picardie Jules Verne: Project partner; expertise in the development of *a posteriori* discretisation error estimation techniques, main author of [21] (upon which this project builds) and DFTK contributor. Main involvement in **WP1** to **WP3** to support DC and guide developments.
- **Michael Herbst**: Applicant; interdisciplinary expertise in materials modelling, multitask GP models and efficient implementation of DFT methods in DFTK, main supervisor of employees and project coordination.

My key collaborators on topics related to this project are:

- **Giovanni Pizzi**, PSI: Joint work on integrating DFTK with AiiDA (related to **WP1**). Main developer of AiiDA and experience in running automated high-throughput materials simulations.
- **Michele Ceriotti**, EPFL: Collaboration on understanding uncertainty propagation between the scales of atomistic modelling (related to **WP4**). Broad experience in atomistic machine learning, including feature construction, GP regression and kernel methods.
- **Youssef Marzouk**, MIT: Joint work on multitask GP regression for atomistic machine learning (related to **WP5**). Experience in Bayesian statistics and statistical modelling of uncertainties in various physics and engineering applications.

3.4 Work packages

3.4.1 Systematic validation and parameter tuning of perturbation-based error estimates (WP1)

In this work package (WP) we will employ automated computations across the compounds of the **verification dataset** to overcome the limitations described in Section 1.2.3. In particular we will make use of the diversity of this dataset to determine reliable heuristics for choosing the cutoff `Eref` of the (large) basis $\mathcal{X}_{\texttt{Eref}}$ employed in the perturbative step (5) depending on the cutoff `Ecut` of $\mathcal{X}_{\texttt{Ecut}}$ used for solving the DFT problem (1). Since only **WP3** develops a treatment of metallic systems, we exclude the metallic unaries from the **verification dataset**. Note that this keeps some chemical diversity for each element, since all oxide structures are insulators.

Based on extending the existing DFTK-AiiDA plugin (see preliminary work in Section 2.1) we will first enable the automated computation of the perturbation-based estimates of Section 1.2 over large datasets using AiiDA. Using the **verification dataset** we then perform a parameter study of `Eref` depending on `Ecut`, where we track the correlation of the estimated and the "true" error as well as the required computational cost. The goal is to achieve for a variety of norm-conserving pseudopotential families (SG15, pseudodojo, Goedecker) and as large a

set of elements as possible a good correlation between the estimated and the "true" error, like in Figure 2. From our previous analysis on pseudopotential models [19, 24] we expect, that the ideal relation between `Eref` and `Ecut` is independent of the chemical environment, but only dependent on the decay of the Fourier coefficients of the potential — thus in turn only dependent on the smoothness near the nucleus. Therefore, similar to previous endeavours to tabulate recommended cutoffs for pseudopotentials [69, 75], one should be able to determine the ideal offset between `Eref` and `Ecut` once and for all for each pair of element and pseudopotential. Once achieved this **turns discretisation error estimation into a practical tool**.

Finally, we will investigate whether the obtained perturbative corrections (δE, δF, see Section 1.2.2) can be routinely employed as a **refinement strategy**. To check for possible limitations, biases or unphysical behaviour we consider the following:

- **Equation of state:** Based on automated AiiDA workflows we compute energy-volume curves (equations of state) of DFTK employing various basis sets $\mathcal{X}_{\texttt{Ecut}}$ with and without refinement of the energy. The resulting curves we compare against all-electron references following the procedure of [65] and investigate to what extend the agreement deteriorates for smaller cutoffs `Ecut` and whether the refinement can compensate.
- **Smooth DFT energy surface and forces:** For a sufficiently large plane-wave basis the DFT energy and accompanying forces vary smoothly with respect to atomic positions — an aspect, which is indeed exploited by machine-learning approaches. By distorting the geometry of systems from the **verification dataset** we will investigate whether this is still the case when employing a crude basis $\mathcal{X}_{\texttt{Ecut}}$ plus refinement.

Deliverables: Goal G1, software framework for automated validation of discretisation error estimates (via DFTK and AiiDA), publication discussing parameter selection and validation of perturbation-based estimates;

3.4.2 Error estimates for structures, stresses and bands (WP2)

The process of structure relaxation in materials modelling employs the forces $F(P, \mathcal{R}, \mathcal{L})$ (derivative of energy wrt. atomic positions $\mathcal{R}$) and stresses $S(P, \mathcal{R}, \mathcal{L})$ (derivative of energy wrt. lattice parameters $\mathcal{L}$) to minimise the energy wrt. $\mathcal{R}$ and $\mathcal{L}$ and find the minimum-energy structure. Clearly, if F and S suffer from a discretisation error, the optimal structure is erroneous as well. Knowing the discretisation errors δF and δS, the first-order error in the optimal structure can be determined by a Newton step as we will illustrate in the following. We collect the parameters as a vector $\theta = (\mathcal{R}, \mathcal{L})^T$ and the derivatives as $D(P, \theta) = \big(F(P, \mathcal{R}, \mathcal{L}), S(P, \mathcal{R}, \mathcal{L})\big)^T$. Having obtained the optimal positions and lattice in a given basis $\mathcal{X}_{\texttt{Ecut}}$ implies $D(P, \theta) = 0$. This defines locally an implicit function $\theta(P)$ around the optimal θ and corresponding solution P of (1). By implicit differentiation we formally obtain its first-order change $\delta\theta$ due to a density matrix change δP as $\delta\theta = \frac{d\theta}{dP}\delta P = -\left(\frac{\partial D}{\partial \theta}\right)^{-1}\left(\frac{\partial D}{\partial P}\right)\delta P$. Inserting the components of θ and D we expand as

$$\begin{pmatrix}\delta\mathcal{R}\\ \delta\mathcal{L}\end{pmatrix} = -\begin{pmatrix}\frac{\partial F}{\partial \mathcal{R}} & \frac{\partial F}{\partial \mathcal{L}}\\ \frac{\partial S}{\partial \mathcal{R}} & \frac{\partial S}{\partial \mathcal{L}}\end{pmatrix}^{-1}\begin{pmatrix}\frac{\partial F}{\partial P}\\ \frac{\partial S}{\partial P}\end{pmatrix}\delta P = -\begin{pmatrix}\frac{\partial F}{\partial \mathcal{R}} & \frac{\partial F}{\partial \mathcal{L}}\\ \frac{\partial S}{\partial \mathcal{R}} & \frac{\partial S}{\partial \mathcal{L}}\end{pmatrix}^{-1}\begin{pmatrix}\delta F\\ \delta S\end{pmatrix}, \tag{9}$$

where in the last equality we used (4), the equivalent expression for the stresses and noted that in this illustration F and S also depends on $\mathcal{R}$ and $\mathcal{L}$ (hence the partials).

Notably, (9) only needs to be solved once to estimate the error — namely after the structure relaxation has fully converged. Therefore the potentially demanding computations of the second energy derivatives $\frac{\partial F}{\partial \mathcal{R}}$, $\frac{\partial S}{\partial \mathcal{R}}$ are one-time costs and thus feasible — at least for the small structures of the **verification dataset**. Both for computing these derivatives as well as the derivative $\frac{\partial S}{\partial P}$ (which is needed to obtain δS in analogy to (4)) we employ the algorithmic differentiation capabilities of DFTK — following previous work [21]. Regarding the discretisation correction for the stresses δS a crucial aspect to be checked is whether the low-frequency components also dominate $\frac{\partial S}{\partial P}$, such that we can accurately approximate (3) by (5) — same as done for the

forces. If this is not possible an adequate approximation for (3) for the case of the stresses needs to be developed, such that the computationally infeasible iterative inversion of $\boldsymbol{\Omega} + \boldsymbol{f}$ on the full $\mathcal{X}_{\texttt{Eref}}$ basis can be avoided.

Given error estimates in the structural parameters, the total first-order discretisation error in a quantity QoI $Q(P, \mathcal{R}, \mathcal{L})$ can be obtained from the chain rule as

$$\delta Q = \frac{\partial Q}{\partial \mathcal{R}} \delta\mathcal{R} + \frac{\partial Q}{\partial \mathcal{L}} \delta\mathcal{L} + \frac{\partial Q}{\partial P} \delta P. \tag{10}$$

Besides tracking the error in the computation of the QoI itself (third term), this expression additionally allows the **propagation of error in the optimal structural parameters** ($\delta\mathcal{R}$ and $\delta\mathcal{L}$) to the QoI computation. Since structure relaxations are an additional iterative procedure around the DFT problem (1) (with one DFT problem solved per relaxation step) one sometimes employs a smaller cutoff `Ecut` for the relaxation than used for the subsequent computation of Q at the relaxed structure. Using (10) allows tracking and balancing respective error contributions in such a case.

The band structure is a common quantity Q, which is computed after a structure relaxation. Since the bands are the eigenvalues ϵ_n of the Hamiltonian $H(P)$, the first-order correction can be obtained from perturbation theory (with perturbation $H(P + \delta P) - H(P)$) as:

$$\delta\epsilon_n = \langle\psi_n | H(P + \delta P) - H(P) | \psi_n\rangle, \tag{11}$$

where ψ_n is the eigenvector corresponding to ϵ_n. Notably ψ_n only has support on `Ecut`, such that the cost-dominating computation of $H(P + \delta P)$ only needs to be performed in the small `Ecut` basis $\mathcal{X}_{\texttt{Ecut}}$, is thus independent of `Eref`.

All error estimates in this WP will be implemented into DFTK and integrated with the automated AiiDA-based testing framework of **WP1**. Using this framework and the **verification dataset** we develop an error estimation procedure for the common task of first performing a structure relaxation and then a band structure computation. Computational parameters (e.g. tolerance for solving (9)) will be tuned to ensure a good correlation of $\delta\mathcal{R}$, $\delta\mathcal{L}$ and $\delta\epsilon_n$ with the "true" error. Finally, opportunities to **save computational cost in the structure optimisation** from our ability to estimate and balance all error contributions in (10) will be explored.

Deliverables: Goal G2, publication on discretisation error estimates for stresses & relaxed structures, publication on propagation of structural error & error balancing in band computations, integration of developed estimates for use in high-throughput workflows (via DFTK and AiiDA-DFTK);

3.4.3 Extension of error estimates to metallic systems (WP3)

The perturbation-based discretisation error estimation strategy of [21] assumes that the DFT problem (1) involves a minimisation over the (non-convex) manifold $\mathcal{P}$. When representing a density matrix $P \in \mathcal{P}$ using $n > N$ orbitals, i.e. $P = \sum_{i=1}^{n} \alpha_i \phi_i \phi_i^\dagger$, a consequence is that the occupations $\alpha_i \in \{0, 1\}$ are always integer. For metallic systems the usual numerical setup is to employ a smearing function and a non-zero electronic temperature, leading to occupation values $0 \leq \alpha_i \leq 1$. The minimisation (1) thus runs over the convex hull of $\mathcal{P}$, which breaks the mathematical framework of the existing perturbation-based estimates [21].

In this WP we will develop a generalisation of [21] for metallic systems. We will stick to the successful idea of a two-grid approach, i.e. approximate the discretisation error of a solution obtained in $\mathcal{X}_{\texttt{Ecut}}$ by computing the first-order correction δP (3) for a residual R obtained in a reference basis $\mathcal{X}_{\texttt{Eref}}$. To obtain a tractable expression for R we will take ideas from the recent work on the convergence analysis for DFT simulations for metallic systems [76] as well as my collaborative work on an adaptive line search for DFT [55] — in which we construct a quadratic approximation of the energy for the general case including an electronic temperature. In the metallic

setting we also expect solving (3) to be numerically more challenging: in our recent analysis of density-functional perturbation theory (DFPT) [56] — which features a similar equation to (3) — we found that for metallic systems (a) employing an inadequate orbital-based representation of δP can lead to numerical instabilities and (b) that solving the Sternheimer equations (arising when applying $\boldsymbol{\Omega} + \boldsymbol{f}$) can become ill-conditioned. We then suggested a Schur-complement-based algorithm to overcome these problems [56]. However, further work is needed to include these ideas into the two-grid setting of the perturbation-based error estimates as well as to obtain a modified approximation like (5), in which most computations are restricted to the `Ecut` grid. If such an approximation can be successfully obtained and implemented we will proceed similar to **WP1**, namely to make use of the metallic unaries of the **verification dataset** (a) to tune computational parameters and (b) to verify the estimates based on automated computations of equations of state and DFT forces using DFTK and AiiDA.

Deliverables: Goal G3, publication about perturbation-based error estimates for metallic systems;

3.4.4 Error-informed Gaussian Process Regression (WP4)

In this WP we explore the integration of the discretisation error estimates of WP1 within GP regression. The goal is to overcome the need for homogeneous training data (see Section 1.3.1) and to develop inference procedures, which enable a quantitative assessment of the discretisation error in the GP's predictions. Our methods will be developed and validated by fitting a GP-based Gaussian Approximation Potential (GAP) of silicon following [71], from which the **silicon dataset** is taken. The first step of this WP will thus be to generate a heterogeneous dataset by computing the energies, forces and respective error estimates δE and δF of these silicon structures using a range of plane-wave cutoffs `Ecut`. For the smaller systems we will also compute reference energies and forces using even higher `Ecut` values. This resulting **multi-`Ecut` silicon dataset** we will make available to other researchers by publishing it on MaterialsCloud [77].

Based on this data we will first develop a **heteroscedastic noise model** [78–80], i.e. employ an ansatz $\varepsilon_i \sim N(0, \sigma_i^2)$ in (6), where the variance σ_i is different for each data point i. The main advantage of this approach is that inference is computationally no more expensive than in the homoscedastic case, since only the entries of the diagonal matrix $\boldsymbol{\Sigma}$ in the inference formula (8) are effected — now taking values $\sigma_1^2, \dots \sigma_n^2$. With respect to including the discretisation errors δE and δF as part of σ_i various options are reasonable and need to be tested. For example it is likely advantageous to still regularise the fit, e.g. employ a common hyperparameter $\tilde{\sigma}$ and set $\sigma_i = \tilde{\sigma} + \delta Q$, where δQ is the error estimate matching the data point (δE or respective component of δF). The resulting heteroscedastic GP model will be tested on the **multi-`Ecut` silicon dataset** in three steps:

- The noise model is validated and fine-tuned by comparing against the actual distribution of the errors in energy and forces. These errors are obtained as the difference of the values at low `Ecut` against the high-`Ecut` reference we computed as part of the dataset.
- Using data from only a single `Ecut` value (e.g. the 18 Hartree recommended for pseudodojo potentials) we test whether moving from the homoscedastic to the heteroscedastic noise model already has an effect on the quality of the GP's prediction.
- Finally we validate the resilience of the prediction quality of this error-informed GP to reducing the `Ecut` in selected data points. We will also investigate whether the variance of the posterior (8) reliably correlates with the discretisation error, i.e. that it increases in those regions of the feature space, where only *inexact* data is present. If successful we will investigate ways to exploit this model to reduce data generation cost, e.g. by tracking the posterior variance for quality control.

A disadvantage of the above hetereoscedastic noise approach is that it assumes discretisation error to be independent between data points, thus failing to capture that similar structures (close in feature space) tend to have similar discretisation errors as well. In a second step we will thus investigate constructing a **GP model to**

simultaneously predict a QoI and its discretisation error. On both the DFT quantity Q (either energy E or forces F) as well as the errors δQ we put GP priors

$$Q(X) \sim \mathrm{GP}\Big(\mu(X), k(X, X')\Big) \qquad \text{and} \qquad \delta Q(X) \sim \mathrm{GP}\Big(\mu_\delta(X), k_\delta(X, X')\Big), \tag{12}$$

however with potentially different mean and kernel functions. Since this effectively doubles the dimension of the data $\{Y_i\}_{i=1}^n$ inverting the kernel matrix during inference (8) becomes 8 times as expensive. Approximate inference strategies [81–84] are thus likely needed to treat the around 3000 structures of the **silicon dataset**. Based on (12) and our **multi-`Ecut` silicon dataset** we then learn multiple GP models — one for each `Ecut`— and validate the prediction of Q as well as δQ on unseen data. With this setup we will tackle the many practical questions, such as: (a) Should the heteroscedastic error model be kept for Q? (b) What are good kernels? (c) Is the data Q for low `Ecut` values even smooth enough to be learned accurately? (d) Does smoothness or prediction accuracy improve if instead of learning $(Q, \delta Q)$ we rather learn $(Q + \delta Q, \delta Q)$, i.e. the refined quantities and the error? Notably the resulting GP regression approach will be able to predict not only energies and forces, but also the expected discretisation error. For the first time this thus achieves a **quantitative propagation of the DFT error through a surrogate model**.

Deliverables: Goal G4, publication of the **multi-`Ecut` silicon dataset** as a heterogeneous dataset for advanced GP regression, publication on heteroscedastic GP regression as a way to exploit multi-`Ecut` training data, publication on GP regression approach to predict both quantities and their errors;

3.4.5 Error-informed multitask GP regression (WP5)

In this WP we extend the last part of **WP4** from single-task regression to multitask GP regression, i.e. where a GP model is trained with simulation data obtained using *multiple* `Ecut` cutoffs. For simplicity we will only discuss the case of regressing forces based on two cutoffs. The generalisation to m cutoffs follows by iterating L over the $m-1$ least accurate tasks in the following presentation. We denote by $(F^H(X_i), \delta F^H(X_i))$ the result of the high-`Ecut` DFT simulation and by $(F^L(X_i), \delta F^L(X_i))$ the result of a cruder low-`Ecut` simulation. Following our previous work [52] we assume the relationships

$$F^L(X_i) = \rho^L\, F^H(X_i) + d^L(X_i) \quad \text{and} \quad \delta F^L(X_i) = \rho^L_\delta\, \delta F^H(X_i) + d^L_\delta(X_i), \tag{13}$$

in our statistical model, where ρ^L, ρ^L_δ are correlation hyperparameters and $d^L(X_i)$ and $d^L_\delta(X_i)$ are disparity functions. Putting a GP prior on the disparity functions as well as $(F^L(X_i), \delta F^L(X_i))$ and $(F^H(X_i), \delta F^H(X_i))$ again results in an analytic expression for the posterior distribution of $(F^H(X_*), \delta F^H(X_*))$ given all data [52]. Exploiting the correlation between the DFT quantities at the various `Ecut` levels, the surrogate is thus able to make predictions of the forces and respective discretisation error *at the highest* `Ecut` level even though this level may not be available for all structures of the data set. Effectively the **surrogate is able to exploit convergence trends** leading to considerable flexibility which `Ecut` levels to consider for each structure during data generation.

We will investigate the opportunities from multitask GP regression to (a) **reduce the cost of training data generation** for fitting a silicon GAP potential while (b) still obtaining an **accurate prediction** of both the **forces** and their respective **discretisation error**. A challenge is not only to develop strategies for choosing the many hyperparameters and kernels, but also for selecting the `Ecut` levels at which a structure should be simulated. For this we will take inspiration from the optimal experimental design [85–88] and the Bayesian Optimisation literature [44, 89–91], where a number of strategies to balance GP prediction accuracy and data generation cost have already been proposed.

Deliverables: goal G5, publication on multitask GP regression for data with heterogeneous plane-wave cutoffs;

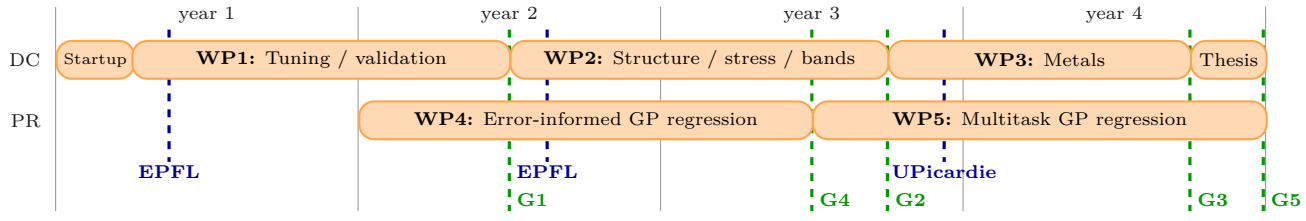
3.5 Risk analysis and quality assurance

A core component of the project is to take recent advances in the mathematical literature on the estimation of discretisation errors in density-functional theory (DFT) and to develop them into a readily applicable tool for data-driven materials modelling. While initial results and our preliminary work are promising, our aspired developments are ambitious and associated with considerable risks. Below, some key risks are listed along with our strategies for mitigation.

- **Error estimation techniques are too expensive or unreliable beyond simple systems (silicon).** Our preliminary work (Section 2.2) has already shown perturbation-based error estimation strategies to be promising to capture the discretisation error on silicon systems — both in terms of accuracy as well as computational cost. Even if **WP1** shows this does not generalise to the entire **verification dataset**, this risk is minimal for the **silicon dataset** (with only silicon structures). Based on the latter most planned work packages (WPs) can still be realised albeit with limitations regarding the generality of the obtained research outcomes. Furthermore, if shortcomings in the current perturbation-based estimation techniques are identified the strong support by my project partner Gaspard Kemlin and the systematic computational results on the **verification dataset** put us into an excellent position to overcome them.
- **Error estimates for the stresses are prohibitively expensive.** What makes error estimation practical for the forces is the fact that only the low-frequency components of δP via (5) are needed. If this approximation turns out to be unsuitable for the stresses in **WP2** and no alternative approximation of (3) can be developed, we will restrict our treatment of the error in the structure relaxation in **WP2** to the case where only atomic positions are optimised.
- **Error estimates for metallic systems cannot be obtained.** Developing a good error estimation scheme for metallic systems (**WP3**) is an open and challenging research problem. In the research plan therefore no other WP depends on the success of **WP3**. Moreover we plan a 6-week visit of the DC at the Université de Picardie to support **WP3**.
- **First-order error propagation does not yield accurate estimates.** Throughout this proposal errors are estimated approximately and propagated using first-order expansions, such as (3), (9) or (11). For more involved quantities or systems our discretisation error estimates may thus not be accurate. However, considering our aspired integration of such estimates within statistical models (**WP4** and **WP5**) one can very much treat the estimated errors themselves as noisy data — and then expect GP regression to capture the key trends while smoothening out inaccuracies. If we lower our goal from making quantitative error predictions to only exploiting the estimates to save computational cost, representing key trends in the discretisation error is likely sufficient — similar to the first active learning schemes for atomistic machine learning, which brought savings albeit being based on rather simple error models [40, 41].

4 Schedule and milestones

The following provides the planned timeline of WPs split into the two tracks mainly involving the DC & the PR. The achievement of goals (Section 3.1) is indicated by green lines and visits by blue lines. "EPFL" indicates a two-week visit of Gaspard Kemlin at EPFL and "UPicardie" a 6-week visit of the DC at Université de Picardie.



5 Relevance and impact

In interdisciplinary fields such as materials modelling groundbreaking advances are frequently enabled by establishing novel links between adjacent fields. Central to this proposal is indeed to establish such a novel link between recent mathematical advances on the estimation of discretisation errors for density-functional theory (DFT) [18–24] and modern data-driven materials modelling. We will not only develop such error estimation techniques into a routine tool, but also explore the opportunities of integrating such estimates with Gaussian process (GP) regression — a standard approach in atomistic machine learning. Crucially the resulting machine learning frameworks will be able to (a) provide estimates for the discretisation error in their own predictions and in this way (b) allow lowering the cost of data generation (by selectively using cruder basis sets) while keeping the impact on the prediction quality under control. In an age where DFT simulations are run in the millions [12] and occupy a noteworthy share of supercomputers (between 33% and 50% on Archer2 [92]) with their associated carbon footprint [93, 94] this has an enormous potential to save energy and accelerate materials discoveries in the years to come. Not to mention the impact practical error estimation techniques can have to (a) automatically setup calculations in a quality-controlled fashion and (b) to guide students or inexperienced users to what extend simulation outcomes can be trusted. Of note our error-informed GP regression approaches developed in **WP4** and **WP5** are in principle agnostic to the kind of error estimate they are presented with. In this proposal we target the recent perturbation-based discretisation error estimates [21], because these are currently the most mature mathematical approaches. However, other estimation techniques or even a combined approach based on estimates covering multiple error sources (e.g. the error due to insufficient k-point sampling) could be propagated through a GP model in follow-up research.

Beyond methodological advances, which will be disseminated in 7 research articles, we will also contribute to open-source software. This includes the integration of our error estimates developed in **WP1** to **WP3** into DFTK and the respective plugin of the AiiDA workflow engine, such that they can be readily employed for high-throughput materials discovery. Within the context of the NCCR MARVEL, in which my group is also involved in, this opens up possibilities for synergy, e.g. to lower data generation costs within the Machine Learning Pillar 2 or to make simulation errors more transparent to users of the MARVEL digital infrastructure (Pillar 3). Furthermore such DFTK-AiiDA workflows also contribute to closing the noteworthy gap between the mathematical community and practical materials simulations. Effectively it enables (a) the implementation of prototype algorithms in DFTK and (b) the subsequent *automated testing* on the diverse chemical environments on the **validation dataset** using AiiDA. Last but not least we will also publish our **multi-`Ecut` silicon dataset** of **WP4** on the MaterialsCloud community platform [77]. This dataset closes the gap of a rigorously annotated dataset to foster the development of more sophisticated multitask approaches in atomistic machine learning.

With the recent advances on error estimation techniques in the mathematical literature we are at a turning point: the numerical error of a DFT simulation no longer needs to be guessed, but can be rigorously estimated using mathematical principles. Even though further work is needed to make these methods completely mature, I am convinced these will unlock unprecedented opportunities to make DFT simulations (a) more robust (by tracking their errors) and (b) more efficient (by tuning the numerics exactly to deliver the required accuracy). With this work we contribute significantly to developing and advancing error estimation techniques for plane-wave DFT. We furthermore investigate first ideas how such techniques could be integrated and exploited within modern data-driven materials modelling workflows. With regards to future work on this exciting topic the proposed research provides a solid foundation.

Bibliography

[1] A. Jain, Y. Shin and K. A. Persson. Nature Review Materials, **1**, 15004 (2016).

[2] K. Alberi, M. B. Nardelli, A. Zakutayev et al. Journal of Physics D: Applied Physics, **52**, 013001 (2019).

[3] S. Luo, T. Li, X. Wang et al. Wiley Interdisciplinary Reviews: Computational Molecular Science, **11**, e1489 (2021).

[4] L. Himanen, A. Geurts, A. S. Foster and P. Rinke. Advanced Science, **6**, 1900808 (2019).

[5] R. Pollice, G. dos Passos Gomes, M. Aldeghi et al. Accounts of Chemical Research, **54**, 849 (2021).

[6] V. L. Deringer, A. P. Bartók, N. Bernstein et al. Chemical Reviews, **121**, 10073 (2021).

[7] J. Behler. Chemical Reviews, **121**, 10037 (2021).

[8] J. S. Smith, B. T. Nebgen, R. Zubatyuk et al. Nature Communications, **10**, 2903 (2019).

[9] P. O. Dral, A. Owens, A. Dral and G. Csányi. The Journal of Chemical Physics, **152**, 204110 (2020).

[10] S. M. Goodlett, J. M. Turney and H. F. Schaefer. The Journal of Chemical Physics, **159**, 044111 (2023).

[11] P. O. Dral, T. Zubatiuk and B.-X. Xue. In *Quantum Chemistry in the Age of Machine Learning*, 491–507. Elsevier (2023).

[12] L. Chanussot, A. Das, S. Goyal et al. ACS Catalysis, **11**, 6059 (2021).

[13] L. Gigli, A. Goscinski, M. Ceriotti and G. A. Tribello. *Modeling the ferroelectric phase transition in barium titanate with DFT accuracy and converged sampling* (2023). Arxiv:2310.12579.

[14] R. Ramakrishnan, P. O. Dral, M. Rupp and O. A. von Lilienfeld. Scientific Data, **1**, 140022 (2014).

[15] F. Musil, M. J. Willatt, M. A. Langovoy and M. Ceriotti. Journal of Chemical Theory and Computation, **15**, 906 (2019).

[16] S. Chmiela, A. Tkatchenko, H. E. Sauceda et al. Science Advances, **3**, e1603015 (2017).

[17] A. S. Christensen and O. A. von Lilienfeld. Machine Learning: Science and Technology, **1**, 045018 (2020).

[18] E. Cancès, G. Dusson, Y. Maday et al. Comptes Rendus Mathematique, **352**, 941 (2014).

[19] M. F. Herbst, A. Levitt and E. Cancès. Faraday Discussions, **224**, 227 (2020).

[20] G. Dusson and Y. Maday. IMA Journal of Numerical Analysis, **37**, 94 (2017).

[21] E. Cancès, G. Dusson, G. Kemlin and A. Levitt. SIAM Journal on Scientific Computing, **44**, B1312 (2022).

[22] G. Dusson, I. Sigal and B. Stamm. Mathematics of Computation, **92**, 217 (2023).

[23] G. Dusson and Y. Maday. *An overview of a posteriori error estimation and post-processingmethods for nonlinear eigenvalue problems* (2023). Arxiv:2303.01273.

[24] E. Cancès, G. Kemlin and A. Levitt. Journal of Scientific Computing, **98**, 25 (2024).

[25] E. Cancès, G. Dusson, Y. Maday et al. IMA Journal of Numerical Analysis, **41**, 2423 (2021).

[26] E. Cancès, G. Kemlin and A. Levitt. SIAM Journal on Matrix Analysis and Applications, **42**, 243 (2021).

[27] S. Goedecker, M. Teter and J. Hutter. Physical Review B, **54**, 1703 (1996).

[28] C. Hartwigsen, S. Goedecker and J. Hutter. Physical Review B, **58**, 3641 (1998).

[29] F. Musil, A. Grisafi, A. P. Bartók et al. Chemical Reviews, **121**, 9759 (2021).

[30] A. P. Bartók and G. Csányi. International Journal of Quantum Chemistry, **115**, 1051 (2015).

[31] M. Ceriotti, M. J. Willatt and G. Csányi. *Machine Learning of Atomic-Scale Properties Based on Physical Principles*, 1911–1937. Springer International Publishing (2020).

[32] A. Denzel and J. Kästner. The Journal of Chemical Physics, **148**, 094114 (2018).

[33] L. Chan, G. R. Hutchison and G. M. Morris. Journal of Cheminformatics, **11**, 32 (2019).

[34] M. Todorović, M. U. Gutmann, J. Corander and P. Rinke. npj Computational Materials, **5**, 35 (2019).

[35] E. Garijo del Río, J. J. Mortensen and K. W. Jacobsen. Physical Review B, **100**, 104103 (2019).

[36] R. Meyer and A. W. Hauser. The Journal of Chemical Physics, **152**, 084112 (2020).

[37] A. P. Bartók, M. C. Payne, R. Kondor and G. Csányi. Physical Review Letters, **104**, 136403 (2010).

[38] A. P. Bartók, R. Kondor and G. Csányi. Physical Review B, **87**, 184115 (2013).

[39] C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. the MIT Press (2006).

[40] R. Jinnouchi, F. Karsai and G. Kresse. Physical Review B, **100**, 014105 (2019).

[41] J. Vandermause, S. B. Torrisi, S. Batzner et al. npj Computational Materials, **6**, 20 (2020).

[42] G. Imbalzano, Y. Zhuang, V. Kapil et al. The Journal of Chemical Physics, **154**, 074102 (2021).

[43] M. Kennedy and A. O'Hagan. Biometrika, **87**, 1 (2000).

[44] A. I. Forrester, A. Sóbester and A. J. Keane. Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences, **463**, 3251 (2007).

[45] G. Leen, J. Peltonen and S. Kaski. Machine Learning, **1-2**, 157 (2012).

[46] G. Pilania, J. Gubernatis and T. Lookman. Computational Materials Science, **129**, 156 (2017).

[47] P. Zaspel, B. Huang, H. Harbrecht and O. A. von Lilienfeld. Journal of Chemical Theory and Computation, **15**, 1546 (2019).

[48] R. Batra, G. Pilania, B. Uberuaga and R. Ramprasad. ACS Applied Materials & Inference, **11**, 24906 (2019).

[49] A. E. Wiens, A. V. Copan and H. F. Schaefer. Chemical Physics Letters, **737**, 100022 (2019).

[50] T. Morishita and H. Kaneko. ACS Omega, **8**, 33032 (2023).

[51] V. Vinod, U. Kleinekathöfer and P. Zaspel. Machine Learning: Science and Technology, **5**, 015054 (2024).

[52] K. Fisher, M. Herbst and Y. Marzouk. *Multitask methods for predicting molecular properties from heterogeneous data* (2024). Arxiv:2401.17898.

[53] M. F. Herbst, A. Levitt and E. Cancès. Proceedings of the JuliaCon Conference, **3**, 69 (2021).

[54] M. F. Herbst and A. Levitt. Journal of Physics: Condensed Matter, **33**, 085503 (2020).

[55] M. F. Herbst and A. Levitt. Journal of Computational Physics, **459**, 111127 (2022).

[56] E. Cancès, M. F. Herbst, G. Kemlin et al. Letters in Mathematical Physics, **113**, 21 (2023).

[57] J. Bezanson, A. Edelman, S. Karpinski et al. SIAM Review, **59**, 65 (2017).

[58] *DFTK features documentation*. `https://docs.dftk.org/stable/features/` (2024). Accessed: 2024-03-29.

[59] E. Cancès, L. Garrigue and D. Gontier. Physical Review B, **107**, 155403 (2023).

[60] E. Cancès, M. Hassan and L. Vidal. Mathematics of Computation, **93**, 1203 (2023).

[61] E. Cancès and L. Meng. *Semiclassical analysis of two-scale electronic Hamiltonians for twisted bilayer graphene* (2023). Arxiv:2311.14011.

[62] T. Nottoli, I. Giannì, A. Levitt and F. Lipparini. Theoretical Chemistry Accounts, **142**, 69 (2023).

[63] J. Cazalis. Pure and Applied Analysis, **6**, 129 (2024).

[64] P. Giannozzi, O. Baseggio, P. Bonfà et al. The Journal of Chemical Physics, **152**, 154105 (2020).

[65] E. Bosoni, L. Beal, M. Bercx et al. Nature Reviews Physics, **6**, 45 (2023).

[66] S. P. Huber, S. Zoupanos, M. Uhrin et al. Scientific Data, **7**, 300 (2020).

[67] M. Uhrin, S. P. Huber, J. Yu et al. Computational Materials Science, **187**, 110086 (2021).

[68] *aiida-dftk: DFTK plugin for the AiiDA workflow engine.* `https://github.com/aiidaplugins/aiida-dftk` (2024). Accessed: 2024-03-06.

[69] M. van Setten, M. Giantomassi, E. Bousquet et al. Computer Physics Communications, **226**, 39 (2018).

[70] M. Schlipf and F. Gygi. Computer Physics Communications, **196**, 36 (2015).

[71] A. P. Bartók, J. Kermode, N. Bernstein and G. Csányi. Physical Review X, **8**, 041048 (2018).

[72] O. T. Unke, S. Chmiela, H. E. Sauceda et al. Chemical Reviews, **121**, 10142 (2021).

[73] V. L. Deringer, N. Bernstein, G. Csányi et al. Nature, **589**, 59 (2021).

[74] S. De, A. P. Bartók, G. Csányi and M. Ceriotti. Physical Chemistry Chemical Physics, **18**, 13754 (2016).

[75] G. Prandini, A. Marrazzo, I. E. Castelli et al. npj Computational Materials, **4**, 72 (2018).

[76] X. Dai, S. de Gironcoli, B. Yang and A. Zhou. Multiscale Modeling and Simulation, **21**, 777 (2023).

[77] L. Talirz, S. Kumbhar, E. Passaro et al. Scientific Data, **7**, 299 (2020).

[78] P. Goldberg, C. Williams and C. Bishop. In M. Jordan, M. Kearns and S. Solla (Eds.), *Advances in Neural Information Processing Systems*, vol. 10. MIT Press (1997).

[79] Q. V. Le, A. J. Smola and S. Canu. In *ICML '05: The 22nd International Conference on Machine Learning*. ACM Press (2005).

[80] K. Kersting, C. Plagemann, P. Pfaff and W. Burgard. In *ICML '07 & ILP '07: The 24th Annual International Conference on Machine Learning held in conjunction with the 2007 International Conference on Inductive Logic Programming*. ACM (2007).

[81] J. Quinonero-Candela and C. Rasmussen. Journal of Machine Learning Research, **6**, 1939–1959 (2005).

[82] H. Liu, Y.-S. Ong, X. Shen and J. Cai. IEEE Transactions on Neural Networks and Learning Systems, **31**, 4405 (2020).

[83] A. G. Wilson, C. Dann and H. Nickisch. *Thoughts on Massively Scalable Gaussian Processes* (2015). Arxiv:1511.01870.

[84] D. A. Cole, R. B. Christianson and R. B. Gramacy. Statistics and Computing, **31**, 1573 (2021).

[85] L. Le Gratiet and C. Cannamela. Technometrics, **57**, 418 (2015).

[86] P. Perdikaris, D. Venturi, J. O. Royset et al. Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences, **471**, 20150018 (2015).

[87] A. Gorodetsky and Y. Marzouk. SIAM/ASA Journal on Uncertainty Quantification, **4**, 796 (2016).

[88] A. P. Kyprioti, J. Zhang and A. A. Taflanidis. Structural and Multidisciplinary Optimization, **62**, 1135

(2020).

[89] D. Huang, T. T. Allen, W. I. Notz and R. A. Miller. Structural and Multidisciplinary Optimization, **32**, 369 (2006).

[90] M. Poloczek, J. Wang and P. Frazier. In I. Guyon, U. V. Luxburg, S. Bengio et al. (Eds.), *Advances in Neural Information Processing Systems*, vol. 30. Curran Associates, Inc. (2017).

[91] J. Wu, S. Toscano-Palmerin, P. I. Frazier and A. G. Wilson. In R. P. Adams and V. Gogate (Eds.), *Proceedings of The 35th Uncertainty in Artificial Intelligence Conference*, vol. 115 of *Proceedings of Machine Learning Research*, 788–798. PMLR (2020).

[92] *Software usage data from the ARCHER2 national service*. `https://github.com/ARCHER2-HPC/usage-data` (2024). Accessed: 2024-03-29.

[93] W. Feng and K. Cameron. Computer, **40**, 50 (2007).

[94] W. Feng, X. Feng and R. Ge. IT Professional, **10**, 17 (2008).