Introduction

*Development of quantum theory of matter

Discovery of laws

Discovery of faster computation

*Quantum simulation and its limitations

*The data revolution

Example of astonishing achievements

Common problem: curse of simensionality

*Use of data for quantum simulation

*Thesis objectives

Atomistic simulation

The beginigging of a new paradigm for building fast classical models which exploit heavy quantum calculations

In the remainder of this introductory section we briefly review the relative strengths and weaknesses of standard parametrized (P-FFs) and machine learning force fields (ML-FFs). We then consider how accurate P-FFs are hard to develop but eventually fully exploit useful knowledge on the systems, while GP-based ML-FFs offer a general mathematical framework for handling training and validation, but are significantly slower (Section I A). These shortcomings motivate an analysis of how prior knowledge such as symmetry has been so far incorporated in GP kernels (Section I B) and points to features still missing in ML kernels, which are commonplace in the more standard, highly efficient P-FFs based on truncated n-body expansions (Section I C). This suggests the possibility of defining a series of n-body GP kernels (Section II B), providing a scheme to construct them (Section II C and D) and, after the best value of n for the given target system has been identified with appropriate testing (Section II E), exploiting their dimensionally-reduced feature spaces to massively boost the execution speed of force prediction (Section III).

Parametrized and machine learning force fields

Producing accurate and fully transferable force fields is a remarkably difficult task. The traditional way to do this involves adjusting the parameters of carefully chosen analytic functions in the hope of matching extended reference data sets obtained from experiments or quantum calculations Stillinger:1985zz,Tersoff:1988jt. The descriptive restrictiveness of the parametric functions used is both a drawback and a strength of this methodology. The main difficulty is that developing a good parametric function requires a great deal of chemical intuition and patient effort, guided by trial and error steps with no guarantee of success Brenner:2000uh. However, for systems and processes in which the approach is fruitful, the development effort is amply rewarded by the opportunity to provide extremely fast and accurate force models Mishin:2004bh, vanDuin:2001ig,Cisneros:2016hi,Reddy:2016ic. The identified functional forms will in these cases contain valuable knowledge on the target system, encoded in a compact formulation that still accurately captures the relevant physics. Such knowledge is furthermore often transferable to novel (while similar) systems as a prior piece of information, i.e., it constitutes a good working hypothesis on how these systems will behave. When QM data on the novel system become available, this can be simply used to fine-tune the parameters of the functional form to a new set of best-fit values that maximise prediction accuracy.

Following a different approach, nonparametric ML force fields can be constructed, whose dependence on the atomic position is not constrained to a particular analytic form. An implementation and tests exploring the feasibility of ML to describe atomic interactions can be found, e.g., in pioneering work by Skinner and Broughton Skinner:1995ea that proposed using ML models to reproduce first-principles potential energy surfaces. More recent works implementing this general idea have been based on Neural Networks Behler:2007fe, Gaussian Process (GP) regression Bartok:2010fj or linear regression on properly defined bases Shapeev:2016kn. Current work aims at making these learning algorithms both faster and more accurate Li:2015eb,Glielmo:2017dj,Botu:2015kb, Ferre:2016uc,Podryabinkin:2017jp,Takahashi:2017th.

As processing power and data communication bandwidth increase, and the cost of data storage decreases, modeling based on ML and direct inference promises to become an increasingly attractive option, compared with more traditional classical force field approaches. However, although ML schemes are general and have been shown to be remarkably accurate interpolators in specific systems, so far they have not become as widespread as it might have been expected. This is mainly because standard classical potentials are still orders of magnitude faster than their ML counterpart Boes:2016im. Moreover, ML-FFs also involve a more

complex mathematical and algorithmic machinery than the traditional compact functional forms of P-FFs, whose arguments are physically descriptive features that remain easier to visualize and interpret.

Prior knowledge and GP kernels

These shortcomings provide motivation for the present work. The high computational cost of many ML models is a direct consequence of the general inverse relation between the sought flexibility and the measured speed of any algorithm capable of learning. Highly flexible ML algorithms by definition assume very little or no prior knowledge of the target systems. In a Bayesian context, this involves using a general prior kernel, typically aspiring to preserve the full universal approximator properties of e.g., the square exponential kernel Williams:2006vz,Bishop:998831. The price of such a kernel choice is that the ML algorithm will require large training databases Kearns:1994wq, slowing down computations as the prediction time grows linearly with the database size.

Large database sizes are not, however, unavoidable, and any data-intensive and fully flexible scheme to potential energy fitting is suboptimal by definition, as it exploits no prior knowledge of the system. This completely agnostic approach is at odds with the general lesson from classical potential development, indicating that it is essential for efficiency to incorporate in the force prediction model as much prior knowledge of the target system as can be made available. In this respect, GP kernels can be tailored to bring some form of prior knowledge to the algorithm.

For example, it is possible to include any symmetry information of the system. This can be done by using descriptors that are independent of rotations, translations and permutations Li:2015eb,Rupp:2015ep,Deringer:2017ea,Fabe Alternatively, one can construct scalar-valued GP kernels that are made invariant under rotation (see e.g., Bartok:2013cs,Glielmo:2017dj) or matrix-valued GP kernels made covariant under rotation (Glielmo:2017dj, an idea that can be extended to higher-order tensors Bereau:2017vq,Grisafi:2017tw). Invariance or covariance are in these cases obtained starting from a non-invariant representation by appropriate integration over the $SO(3)$ rotation group Glielmo:2017dj, Bartok:2013cs.

Symmetry aside, progress can be made by attempting to use kernels based on simple, descriptive features corresponding to low-dimensional feature spaces. Taking inspiration from parametrized force fields, these descriptors could e.g., be chosen to be interatomic distances taken singularly or in triplets, yielding kernels based on 2- or 3-body interactions Glielmo:2017dj, Szlachta:2014jh,Huo:2017ta. Since low-dimensional feature spaces allow efficient learning (convergence is reached using small databases), to the extent that simple descriptors capture the correct physics, the GP process will be a relatively fast, while still very accurate, interpolator.

Scope of the present work

There are, however, two important aspects that have not as yet been fully explored while trying to develop efficient kernels based on dimensionally reduced feature spaces. Both aspects will be addressed in the present work.

First, a systematic classification of rotationally invariant (or covariant, if matrix valued) kernels, representative of the feature spaces corresponding to n-body interactions is to date still missing. Namely, no definition or general recipe has been proposed for constructing n-body kernels, or for identifying the actual value (or effective interval of values) of n associated with already available kernels. This would be clearly useful, however, as the discussion above strongly suggests that the kernel corresponding to the lowest value of n compatible with the physics of the target system will be the most informationally efficient one for carrying out predictions: striking the right balance between speed and accuracy.

Second, for any ML approach based on a GP kernel and a fixed database, the GP predictions for any target configuration are also fixed once and for all. For an n-body kernel, these predictions do not need, however, to be explicitly carried out as sums over the training dataset, as they could be approximated with arbitrary precision by mapping the GP prediction on a new representation based on the underlying n-body feature space. We note that this approximation step would make the final prediction algorithm independent of the database size, and thus in principle as fast as any classical n-body potential based on functional forms, while still parameter free. The remainder of this work explores these two issues, and it is structured as follows.

In the next Section II, after introducing the terminology and the notation (II A), we provide a definition of an n-body kernel (II B) and we propose a systematic way of constructing n-body kernels of any order n, showing how previously proposed approaches can be reinterpreted within this scheme (II C and D). We