Technische Universität München
Fakultät für Chemie

Kernel-Based Machine Learning for Molecular Crystal Structure Prediction

Simon Peter Wengert

Vollständiger Abdruck der von der Fakultät für Chemie der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitzender: Prof. Dr. Steffen J. Glaser

Prüfer der Dissertation:

1. Prof. Dr. Karsten Reuter
2. Prof. Dr. Harald Oberhofer

Die Dissertation wurde am 29.06.2022 bei der Technischen Universität München eingereicht und durch die Fakultät für Chemie am 04.08.2022 angenommen.

Für meinen Papa.

Preface

The doctoral thesis at hand is publication-based and centers on Refs. 1, 2 and 3 published in international peer-reviewed scientific journals with associated articles being appended to this work.

The related scientific content is embedded in a broader context by introducing the reader into the underlying topic of this dissertation, as well as relevant concepts and methods that have been employed. This is followed by summaries of the published work together with assignments of individual author contributions and a concluding summary.

All work presented in this doctoral thesis was performed between June 2018 and July 2020 at the Chair of Theoretical Chemistry of the Technical University of Munich (TUM) and between August 2020 and June 2022 at the Fritz Haber Institute (FHI) of the Max Planck Society in Berlin under the supervision of Prof. Dr. Karsten Reuter. A research stay hosted by Prof. Dr. Gábor Csányi at the Engineering Laboratory at the University of Cambridge (UK) complements this work.

Berlin, June 2022

Abstract

For a systematic discovery of molecular crystal structures with customized properties, efficient search strategies are desired. Conveniently, these structure searches can be delegated to a machine by means of computational chemistry. Its tools allow to quantify the stability of an atomistic structure by computing its energy and experimentally observed crystals are associated with the most stable structural arrangements. As a consequence of this, successful *in silico* molecular crystal structure predictions (CSP) are associated with the solution to a global optimization problem. For a given molecule, reliable predictions of the most stable structural arrangements face a major challenge, though, which arises from the vast search spaces that need to be explored and the computationally expensive high levels of theory that need to be applied to resolve the typically small stability differences between crystal candidates.

To arrive at corresponding solutions in an efficient way structure-energy relationships need to be evaluated with both high accuracy and low computational costs. On that account, an approach has been developed in this work to generate accurate hybrid models for molecular crystals that feature short evaluation times. These hybrid models are composed of a computationally inexpensive physics-based description of long-range interactions at the density-functional tight-binding (DFTB) level and a short-range correction to reproduce highly accurate first-principles target methods (based on density-functional theory or wavefunction methods). The generation of the latter is achieved by a kernel-based supervised machine learning (ML) strategy developed to yield system-specific Δ-ML corrections that augment the DFTB baseline description.

Accounting for considerable computational costs associated with the evaluation of reference structures, the developed training procedure for Δ-ML models is characterized by a high data-efficiency. In this regard, the training benefits from the applied DFTB baseline as its description captures significant parts of interactions relevant to molecular crystals which circumvents the need to explicitly learn them from data. A diversity-driven selection of appropriate structures further reduces the number of required reference data representative for the intended application of the model to molecular CSP.

For single-component molecular crystals, the obtained hybrid models are shown to accurately reproduce the description of the high-level reference method at a fraction of the computational costs. Beyond that the models are differentiable which allows for efficient local structure optimization and, thus, gives rise to a significant reduction of the computationally most expensive part in typical molecular CSP studies. Conveniently, the approach has been shown to be broadly applicable to various types of single-component molecular crystals and corresponding interactions.

A developed extension of this approach provides a generalization to (neutral) multi-component crystals which are of great practical relevance for well-directed searches of materials featuring application-specific properties. In this context, the robustness of corresponding Δ-ML models is substantiated *inter alia* by performing molecular dynamics simulations at ambient conditions on co-crystal structures outside the scope of the reference structures used for their generation. Here, the obtained predictions of co-crystal densities have been verified by direct comparison with experimental measurements.

Apart from this, the versatile applicability of kernel-based unsupervised learning for gaining insights into data sets of atomistic structures and associated attributes has been illustrated. Here, atomic environments and entire structures have been described by a sophisticated representation while mutual relations between them have been measured and projected to a low-dimensional space by means of kernel principle component analysis. Tools for performing these mappings, subsequent visualization and interactive exploration are conveniently provided in course of the presented work along with illustrative examples to showcase various fields of application such as the analysis of molecular dynamics trajectories, the results of a crystal structure search or information associated with atomic environments in a well-established molecular database.

Zusammenfassung

Für ein systematisches Auffinden molekularer Kristallstrukturen mit bedarfsgerechten Eigenschaften werden effiziente Such-Strategien benötigt. Praktischerweise ist es möglich derartige Suchen nach geeigneten Strukturen auf Computer auszulagern indem Werkzeuge der berechnenden Chemie herangezogen werden. Da diese es ermöglichen die Stabilität einer atomistischen Struktur durch Berechnung der zugehörigen Energie zu quantifizieren und gleichzeitig die stabilsten Anordnungen mit experimentell beobachtbaren Kristallstrukturen assoziiert werden, erfolgt deren *in silico* Vorhersage durch Lösen eines globalen Optimierungsproblems. Für ein gegebenes Molekül sehen sich zuverlässige Vorhersagen über die stabilsten Anordnungen im Festkörper allerdings mit der bedeutenden Herausforderung konfrontiert, dass enorm große Such-Räume erkundet werden müssen und gleichzeitig der Einsatz rechenintensiver, hochgenauer Methoden der theoretischen Chemie nötig ist, um die üblicherweise geringen Stabilitätsunterschiede zwischen potentiellen Kristallstrukturen aufzulösen.

Für die effiziente Lösung des globalen Optimierungsproblems ist es daher nötig den Struktur-Energie-Zusammenhang für molekulare Kristalle mit hoher Genauigkeit und gleichzeitig schnell auszuwerten. Zu diesem Zweck wurde im Rahmen der vorliegenden Arbeit ein Vorgehen zur Erzeugung hybrider Modelle ausgearbeitet, welche die genannten Eigenschaften vereinen. Diese hybriden Modelle bestehen zum einen aus einer mit geringem Rechenaufwand verbundenen und auf physikalischen Gesetzmäßigkeiten beruhenden Beschreibung von langreichweitigen Wechselwirkungen auf Basis der dispersionskorrigierten *density-functional tight-binding* (DFTB) Methode, sowie einer kurzreichweitigen Korrektur, zur Nachbildung der Beschreibungen hochgenauer Methoden der theoretischen Chemie (basierend auf der Dichtefunktionaltheorie oder Wellenfunktionsmethoden). Für die Modellierung dieser Korrektur wurde, unter Verwendung einer Kernel-Methode des überwachten maschinellen Lernens (ML), eine Strategie ausgearbeitet, welche system-spezifische Δ-ML Modelle zur Verbesserung der DFTB Basislinienbeschreibung erzeugt.

Um dem erheblichen Rechenaufwand entgegenzutreten, der für die Auswertung von Referenzstrukturen benötigt wird, zeichnet sich das entwickelte Vorgehen zur Erzeugung von Δ-ML Modellen durch eine hohe Dateneffizienz aus. Diesbezüglich profitiert die Modellerzeugung insbesondere von der verwendeten DFTB Basislinie, in deren Beschreibung bedeutende Anteile der für molekular Kristalle wichtigen Wechselwirkungen bereits enthalten sind und die folglich nicht mehr aus Daten gelernt werden müssen. Indem auf eine breite Diversifikation bei der Auswahl von Referenzstrukturen geachtet wird, konnte die Anzahl an Daten zudem nochmals reduziert werden, die benötigten wird um die beabsichtigte Anwendung der Modelle zur Vorhersage von molekularen Kristallstrukturen repräsentativ abzubilden.

Für molekulare Kristalle, die aus einer Komponente aufgebaut sind, konnte aufgezeigt werden, dass die erhaltenen hybriden Modelle in der Lage sind die exakten Beschreibungen der Referenzmethoden mit hoher Genauigkeit und einem Bruchteil des Rechenaufwands nachzubilden. Darüber hinaus sind die Modelle differenzierbar und ermöglichen daher die effiziente Durchführung von lokalen Strukturoptimierungen, wodurch sich der benötigte Rechenaufwand des üblicherweise kostenintensivsten Aspekts der molekularen Kristallstrukturvorhersage erheblich reduzieren lässt. Für die Anwendung bei Kristallstrukturvorhersagen günstig ist zudem, dass das Vorgehen eine breite Einsetzbarkeit aufweist hinsichtlich unterschiedlichster Arten von molekularen Einkomponenten-Kristallen und den zugehörigen

Wechselwirkungen.

Durch die Weiterentwicklung dieses Vorgehens konnte zudem eine Verallgemeinerung auf (neutrale) Mehrkomponenten-Kristalle erreicht werden, welche von großer praktischer Bedeutung sind bei der gezielten Suche nach Materialien mit anwendungsspezifischen Eigenschaften. In diesem Zusammenhang konnte auch die Robustheit derartiger Δ-ML Modelle nachgewiesen werden, indem diese unter anderem zur Durchführung von Molekulardynamik Simulationen bei Umgebungsbedingungen an Co-Kristallstrukturen außerhalb der zur Modellentwicklung verwendeten Referenzstrukturen eingesetzt wurden. Dabei konnte verifiziert werden, dass sich die erhaltenen Dichten der Co-Kristalle in guter Übereinstimmung mit entsprechenden experimentellen Messergebnissen befinden.

Zusätzlich dazu wurde die vielseitige Einsetzbarkeit aufgezeigt für die Verwendung von Kernel-Methoden des unüberwachten Lernens, um Einblicke in Datensätze von atomistischen Strukturen und entsprechenden Attributen zu gewinnen. Dabei wurden atomare Umgebungen, sowie gesamte Strukturen mit Hilfe einer hochentwickelten Repräsentation beschrieben, während deren gegenseitige Beziehungen mittels Kernel-Hauptkomponentenanalyse gemessen und in niedrig-dimensionale Räume projiziert wurde. Entsprechende Hilfsmittel zur praktischen Durchführung dieses Verfahrens, sowie der anschließenden Visualisierung und interaktiven Erkundung der Daten wurden im Zuge dieser Arbeit bereitgestellt, gemeinsam mit veranschaulichenden Beispielen, welche verschiedene Anwendungsgebiete herausstellen, wie die Analyse von Trajektorien aus Molekulardynamik Simulationen, den Ergebnissen aus Kristallstruktur-Suchen oder der Untersuchung der verschiedenen atomaren Umgebungen in einer gängigen molekularen Datenbank.

Contents

List of Abbreviations

CN	coordination number
CSP	crystal structure prediction
DFT	density functional theory
DFTB	density-functional tight-binding
FPS	farthest point sampling
GAP	Gaussian approximation potential
GGA	generalized gradient approximation
GPR	Gaussian process regression
HF	Hartree-Fock
KPCA	kernel principal component analysis
KS	Kohn-Sham
LDA	local density approximation
MBD	many-body dispersion
MD	molecular dynamics
ML	machine learning
MLIP	machine learned interatomic potential
NN	neural networks
PBE	Perdew, Burke and Ernzerhof
PCA	principal component analysis
PES	potential energy surface
RKHS	reproducing kernel Hilbert space
SCF	self-consistent field
SCS	self-consistent screening
SOAP	smooth overlap of atomic positions
TS	Tkatchenko and Scheffler
XC	exchange-correlation

1 Introduction

Discovery and development of novel materials with tailored properties facilitates innovations and is fundamental for many technological advancements [4–6]. For this reason sophisticated strategies enabling systematic searches for materials with desired properties are among the perennial objectives in both academia and various sectors of industry. Approaches to achieve these targets take advantage of the close relation between structure and property [5]. For molecular crystals this relation enables searches for customized materials by exerting influence on the supramolecular assembly of its building blocks [7].

The atomistic structure of a molecular crystal is dependent on the nature of the molecule it is composed of [8] and in case of multi-component systems such as co-crystals [9], hydrates [10] and salts [11] also the relative proportion of involved molecular types. Moreover, the packing arrangement of molecules in the solid state is often subject to crystallization conditions which gives rise to the phenomenon of polymorphism [12] where multiple structures are found for crystals of identical composition. By taking advantage of the various influencing factors well-directed searches are able to discover molecular crystal structures that feature tailored properties for a certain applications [13, 14]. Experimental screening studies for such materials, however, tie up significant amounts of resources owing to the cost- and time-intensive process for synthesis and characterization [15].

Computational assistance for this process is therefore highly desired to enhance the efficiency of a targeted discovery of novel molecular crystal structures. Molecular crystal structure prediction (CSP) [16] based on information about its compounds solely can be accomplished by means of computational chemistry and its capability to model atomistic systems. The description of these systems is based on the Schrödinger equation and approximations to it [17–19] which are capable of relating structural arrangements with stabilities by means of the potential energy surface (PES) [20]. This relation provides guidance in revealing the actual supramolecular assemblies in the solid state which are associated with the global—or at least low-lying—minima on the respective landscape of the PES [21]. In simple terms, molecular CSP is concerned with a global optimization problem of the functional relation between the stability of a crystal and its underlying structure.

However, major challenges linked to these *in silico* structure predictions make it a long-standing issue in molecular modeling [22, 23]. Difficulties in minimizing the energy landscape of molecular crystals arise from the vast search spaces to be explored. These spaces are spanned by the coordinates of each atom and the lattice parameters defining the periodic cell used to represent crystal structures. Despite the existence of sophisticated search algorithms for navigating the complex landscape with its numerous minima the number of trial structures to be considered and locally optimized is still enormous [24–28]. This applies in particular to crystals composed of flexible molecules and co-crystals where the stoichiometric ratio constitutes an additional search dimension. At the same time, the supramolecular assembly in these systems is in many cases subject to a delicate balance between weak interactions such that sophisticated descriptions are required to reliably resolve low-lying minima on the energy landscape [29, 30]. Advanced methods like dispersion-corrected density-functional theory (DFT) are suitable for this task, but accompanied by significant computational demands [31, 32]. In molecular CSP this induces a trade-off between the accuracy to describe the landscape of the search space and its exploration which constitutes a current limitation in the predictive capability of practical search strategies [16].

Highly promising in this respect are opportunities offered by present achievements in adapting machine learning (ML) strategies to issues in modeling atomistic systems [33–35]. Most notably, supervised ML methods based on kernels [36, 37] or neural networks [38–42] feature outstanding interpolation strengths which enables the generation of accurate machine-learned interatomic potentials with an unrivaled accuracy/cost ratio [43]. These potentials generalize information about the energetic landscape from a representative set of training structures which have been evaluated at reference levels of theory appropriate for an intended application. Moreover, the success of ML strategies applied to chemical systems is greatly supported by the emergence of sophisticated representations of atomistic structures such as the smooth overlap of atomic positions (SOAP) [44] which are physically inspired and have certain fundamental aspects directly integrated [45].

This situation provides the basis for the cumulative dissertation at hand. In Ref. 1 we present a kernel-based supervised ML approach to generate accurate hybrid models for molecular crystals by augmenting the description of a low-cost semi-empirical baseline method with ML models trained on high-quality *ab initio* reference data and substantiate the additional value of these hybrid models in molecular CSP. The broad applicability of the approach is verified on a representative set of single-component molecular materials and with this confidence a generalization to co-crystal systems in presented in Ref. 2. Moreover, in the work published in Ref. 3 we combine kernel-based unsupervised learning strategies and structural representations to obtain low-dimensional projections of atomistic data sets suitable for visualization, exploration and analysis of the ever-increasing amount of data available.

The following chapters are intended to provide an overview of central concepts and methods employed in this thesis. Chapter 2 captures hierarchical simplifications to fundamental concepts of quantum mechanics that yield approximate practical methods suitable for an application to chemical systems. A clear focus will be on methods based on electron densities starting with the main aspects related to density-functional theory and established strategies for incorporating dispersion interactions into its description which are well-known to be crucial for molecular crystals. Afterwards, the chapter concludes with further approximations and concepts that yield the density-functional tight-binding method. Machine learning strategies for atomistic systems comprise various interconnected compounds which are presented in chapter 3 with focus on kernel-based methods. At first, the importance of appropriate representations for atomistic structures is discussed followed by corresponding similarity measurements based on kernel-functions. After that the integral parts of supervised learning strategies for interatomic potentials are presented. At last, the description of a kernel-based unsupervised learning strategy for dimensionality reduction and subsequent visualization of atomistic data sets ultimately concludes the methodological part. Chapter 4 comprises comprehensive summaries of the publications associated with this cumulative thesis together with assignments of contributions to the individual authors. Finally, the work concludes in chapter 5 summarizing the achieved contributions to the field.

2 Electronic Structure Theory

The treatment of atomistic systems on a fundamental level is possible by means of *ab initio* methods relying on the basic laws of quantum mechanics [46]. Interactions between the positively charged nuclei and negatively charged electrons of such systems are captured by the (non-relativistic) Hamiltonian

$$\hat{H} = \hat{T}_{\mathrm{e}} + \hat{T}_{\mathrm{N}} + \hat{V}_{\mathrm{ee}} + \hat{V}_{\mathrm{NN}} + \hat{V}_{\mathrm{Ne}}. \tag{2.1}$$

This operator represents the total energy of a system with kinetic energy contributions of the electrons ($\hat{T}_{\mathrm{e}}$) and nuclei ($\hat{T}_{\mathrm{N}}$) and potential energy contributions arising from electrostatic interactions between electrons and electrons ($\hat{V}_{\mathrm{ee}}$), nuclei and nuclei ($\hat{V}_{\mathrm{NN}}$), and nuclei and electrons ($\hat{V}_{\mathrm{Ne}}$).

In principle, solving the time-independent Schrödinger equation

$$\hat{H} \left|\boldsymbol{\psi}\right\rangle = E \left|\boldsymbol{\psi}\right\rangle \tag{2.2}$$

leads to a wave function $\boldsymbol{\psi}$ for the system containing all information about it. Unfortunately, directly solving Eq. (2.2) becomes too complicated for all but the simplest systems which makes the development of approximate practical methods desirable.

A fundamental assumption valid for a large part of chemistry, solid-state physics and materials science is the Born-Oppenheimer approximation [20] which exploits the tremendous mass difference between nuclei and electrons. This difference gives rise to a significant discrepancy in timescales for their respective motion which lets the nuclei appear static from the electronic point of view. Within the approximation the electrons are therefore assumed to adjust instantaneously to any change in nuclear coordinates by relaxing to the respective ground state. Eq. (2.2) therefore simplifies to the electronic Schrödinger equation with

$$\hat{H}_{\mathrm{el}} = \hat{T}_{\mathrm{e}} + \hat{V}_{\mathrm{ee}} + \hat{V}_{\mathrm{Ne}}, \tag{2.3}$$

where the kinetic energy operator for the nuclei and the nuclear-nuclear interaction potential have been removed and the nuclei positions now enter as parameters instead of variables. Repeatedly solving the electronic Schrödinger equation for different configurations of the nuclei in a system gives rise to the so-called potential energy surface (PES).

These surfaces represent the central function of many studies—and the present dissertation in particular—which rely on atomistic simulations of chemical systems. Stable and metastable configurations, for instance, are associated with global and local minima on these surfaces (as depicted in Fig. 2.1) which enables *in silico* predictions about—potentially not yet synthesised—chemical structures. Directly solving the electronic Schrödinger equation is, however, still too complex for most applications. Thus, further approximations are required in order to arrive at practical methods [17–19].

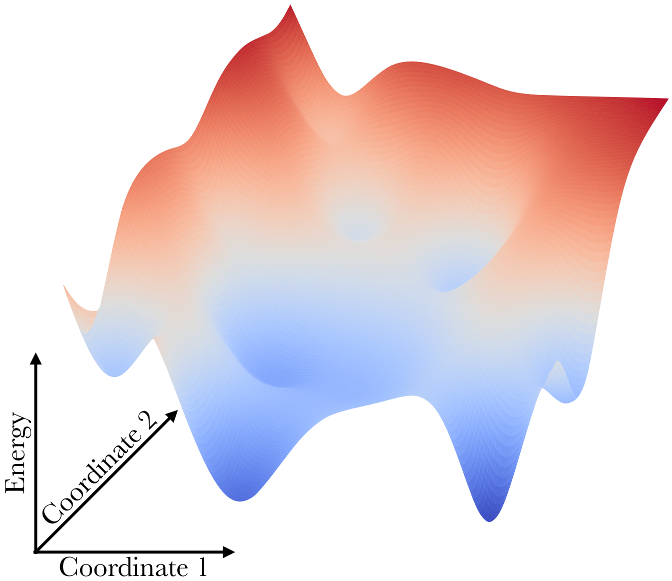


Fig. 2.1: *Exemplary potential energy surface*

2.1 Density Functional Theory

In density functional theory (DFT) the electron density $\rho(\boldsymbol{r})$—a function of three spatial coordinates only—constitutes the central quantity and, thus, replaces the high-dimensional wave function from above. A theoretical foundation for relying solely on the electron density is provided by the Hohenberg-Kohn theorems [47]. The first theorem states a one-to-one mapping between the ground state electron density $\rho_{\mathrm{g}}(\boldsymbol{r})$ and the electronic Hamiltonian and, thus, any ground-state property that can be derived from it. Most importantly, it is possible to write the electronic energy E_{g} of the ground state in terms of a functional of the ground state density. According to the second Hohenberg-Kohn theorem this functional is minimized by the ground state electron density which constitutes a variational principle

$$E[\rho(\boldsymbol{r})] \geq E_{\mathrm{g}}[\rho_{\mathrm{g}}(\boldsymbol{r})] \tag{2.4}$$

with great practical consequences.

2.1.1 Kohn-Sham Approach

The energy in terms of a functional of the electron density can be separated into individual contributions

$$E[\rho(\boldsymbol{r})] = T_{\mathrm{e}}[\rho(\boldsymbol{r})] + V_{\mathrm{ee}}[\rho(\boldsymbol{r})] + V_{\mathrm{Ne}}[\rho(\boldsymbol{r})]. \tag{2.5}$$

While an explicit density-based expression can be derived for the nuclei-electron interactions V_{Ne}, corresponding expressions for the kinetic energy of the electrons T_{e} and the electron-electron interaction V_{ee} are unknown.

On account of this situation, the approach by Kohn and Sham [48] is to separate out those parts from the two terms for which explicit expressions can be found. Although the remaining part is still unknown the fact that it represents a minor contribution to the overall energy allows for adequate results even if simpler approximations are applied to it. The separation reads as

$$E[\rho(\boldsymbol{r})] = T_{\mathrm{S}}[\rho(\boldsymbol{r})] + \frac{1}{2} \iint \frac{\rho(\boldsymbol{r})\rho(\boldsymbol{r'})}{|\boldsymbol{r}-\boldsymbol{r'}|} \mathrm{d}\boldsymbol{r}\mathrm{d}\boldsymbol{r'} + V_{\mathrm{Ne}}[\rho(\boldsymbol{r})] + E_{\mathrm{XC}}[\rho(\boldsymbol{r})], \tag{2.6}$$

where the first term represents the kinetic energy of non-interacting electrons (discussed in more details below) and constitutes a large fraction of the corresponding term for interacting electrons in Eq. (2.5). Similarly, the second term expresses classical electron-electron Coulombic interaction and captures major parts of the interaction in V_{ee}. The last term in Eq. (2.6) is called *exchange-correlation* (XC) functional. It comprises everything not yet included in the others and, thus, approximations have to be developed for it.

An expression for T_{S} can be found by constructing a fictitious reference system of non-interacting electrons. The individual electrons in this system are subject to an effective potential v_{eff} such that its overall electron density is identical to the real, interacting system. With a description based on wave functions it is possible to compute the exact kinetic energy by utilizing concepts known from Hartree-Fock (HF) [18] theory for the N electrons in the reference system with a Hamiltonian expressed as

$$\hat{H}_{\mathrm{KS}} = \sum_{i}^{N} \hat{h}_{\mathrm{KS},i} = \sum_{i}^{N} \left(-\frac{1}{2} \bigtriangledown_i^2 + v_{\mathrm{eff}}(\boldsymbol{r}) \right), \tag{2.7}$$

where $\bigtriangledown$ is a differential operator. The difficulty of finding a solution to the many-body problem of the real system has, thus, been converted into a single particle picture in the reference system. Individual particles in this system are described by the Kohn-Sham (KS) equations

$$\hat{h}_{\mathrm{KS},i} \left|\boldsymbol{\phi}_{\mathrm{KS},i}\right\rangle = \varepsilon_{\mathrm{KS},i} \left|\boldsymbol{\phi}_{\mathrm{KS},i}\right\rangle. \tag{2.8}$$

In contrast to single-particle orbitals obtained from HF theory, the resulting KS-orbitals of the fictitious system from Eq. (2.8) are not associated with a strict physical meaning, but represent mathematical objects solely. However, the electron density expressed as

$$\rho(\boldsymbol{r}) = \sum_{i}^{N} |\boldsymbol{\phi}_{\mathrm{KS},i}|^2, \tag{2.9}$$

is constructed to match with the real system.

Having obtained the single-particle wave functions of the non-interacting electrons the corresponding kinetic energy can be found by

$$T_{\mathrm{S}} = \sum_{i}^{N} -\frac{1}{2} \left\langle \boldsymbol{\phi}_{\mathrm{KS},i} \right| \bigtriangledown_i^2 \left|\boldsymbol{\phi}_{\mathrm{KS},i}\right\rangle. \tag{2.10}$$

A problem that still persists is that the effective potential in Eq. (2.7) itself depends on the entire electron density. Based on the second part of the Hohenberg-Kohn theorem—the variational principle—a self-consistent procedure can be applied to arrive at a solution.

Within the so-called *self-consistent field* (SCF) method a trial electron density is iteratively improved until a convergence criterion is reached. In principle, this strategy allows for obtaining the true ground state energy assuming an exact expression for the XC functional is available. This is, however, not the case—and probably never will be [49]—such that the method quality depends on adequate approximations for E_{XC}.

2.1.2 Approximate Exchange-Correlation Functionals

Various approximations for the XC functional exist which can roughly be organized in terms of associated computational complexity into local density approximations (LDA), generalized gradient approximations (GGA), meta-GGA and hybrid functionals. Details about the individual approximation strategies can be found in Refs. 18, 19 and 49.

Of particular relevance in context of the present dissertation are GGA and hybrid functionals employed in Refs. 1 and 2. GGA functionals take into account the (local) electron density, as well as its gradient which is important for the description of atomistic structures with an inhomogeneous electronic density such as molecular systems. An important representative of GGA functionals is the one developed by Perdew, Burke and Ernzerhof (PBE) [50]. Hybrid functionals make use of the fact that exact exchange energies can be obtained from HF theory. This energy is evaluated for the (occupied) KS-orbitals and used to augment the expressions of explicit XC functional. With this strategy hybrid functionals are capable of—at least approximately—curing deficiencies of these XC functionals, specifically the non-vanishing electron self-interaction that originates from an approximate treatment of the exchange energy. The improvements, however, are accompanied with significantly increased computational demands, particularly for periodic systems. A prominent hybrid functional is PBE0 [51] with a mixing according to

$$E_{\mathrm{XC}}^{\mathrm{PBE0}} = 0.75 \cdot E_{\mathrm{X}}^{\mathrm{PBE}} + 0.25 \cdot E_{\mathrm{X}}^{\mathrm{HF}} + E_{\mathrm{C}}^{\mathrm{PBE}}, \tag{2.11}$$

where the exchange energy obtained with PBE ($E_{\mathrm{X}}^{\mathrm{PBE}}$) is partially replaced with the corresponding KS-HF value ($E_{\mathrm{X}}^{\mathrm{HF}}$) and the correlation energy ($E_{\mathrm{C}}^{\mathrm{PBE}}$) is still completely described by PBE.

The exchange-correlation functionals that have been referred to in this section differ in terms of information included into the corresponding description. Additionally to the local electron density used in LDA computationally more complex models take into account less local information such as the corresponding gradient (GGA) and the Laplacian (meta-GGA). Nevertheless, the derived exchange-correlation functions still rely on (semi-)local information only and are, hence, not capable of providing proper descriptions of nonlocal effects.

2.1.3 Dispersion-Correction

Dispersion interaction is a nonlocal effect that arises from the correlation of electrons in motion which induce charge polarizations in the electron density. It leads to long-range attractive forces crucial for an accurate treatment of noncovalently bonded or condensed phase systems such as molecular crystals. Thus, dispersion-correction schemes have been developed to describe these interactions and couple their expression to an approximate DFT functional. Although it is possible to directly develop nonlocal density functions which are capable of describing these interactions—so-called dispersion-inclusive exchange-correlation functionals [52]—many popular schemes are post-SCF corrections that augment the results of underlying DFT calculations with dispersion contributions. A detailed insight into the conceptual understanding of dispersion interactions and mathematical frameworks for their modeling can be found in Ref. 53.

In the simplest form a pairwise interatomic model according to

$$E_{\mathrm{disp}} = -\frac{1}{2}\sum_{A,B}\frac{C_{6,AB}}{R_{AB}^{6}}f_{\mathrm{damp}}(R_{AB}) \tag{2.12}$$

is employed where dispersion contributions are described by dipole-dipole dispersion coefficients ($C_{6,AB}$) and distances (R_{AB}) between pairs of atoms (A, B). The R^6-dependency of the potential describes the well-known asymptotic attraction at large distances [54–56] while an empirical damping function is used to couple the long-range dispersion contribution to the short-range electron correlation captured by (semi-)local DFT functionals. More elaborate descriptions can be obtained by extending Eq. (2.12) for multipolar (e.g. dipole-quadrupole) and higher-order (e.g. three-body) dispersion contributions. Moreover, the $C_{6,AB}$ coefficients are system-dependent and various approaches for obtaining them have been developed. A mathematical expression for calculating the dispersion coefficient between atom A and B is provided by the Casimir-Polder relation [57]

$$C_{6,AB} = \frac{3}{\pi}\int_{0}^{\infty}\mathrm{d}\omega\alpha_A(\mathrm{i}\omega)\alpha_B(\mathrm{i}\omega), \tag{2.13}$$

where the dynamic polarizabilities at imaginary frequencies $\alpha(\mathrm{i}\omega)$ of the involved atoms are again system-dependent. This provides the basis for the three dispersion-correction schemes presented in the following.

For obtaining the $C_{6,AB}$ coefficients the D4 correction scheme developed by Grimme and co-workers [58, 59] relies on dynamic polarizabilities—pre-computed via time-dependent DFT [60]—of element specific reference systems and incorporates a dependency on the explicit chemical environment in two major steps. First, the atomic reference polarizabilities are scaled based on the environment dependent partial charges obtained for the system of interest (for instance from classical electronegativity equilibration [61]). These scaled reference polarizabilities are used in a second step to incorporate a geometric dependency based on the fractional coordination numbers (CN) of atoms. The CN of an atom in the system of interest is compared with element specific CNs in the reference systems by means of a Gaussian weighting function which further modifies the (charge-scaled) reference polarizabilities. Numerical integration of Eq. (2.13) finally yields system-dependent $C_{6,AB}$ coefficients.

An alternative way for obtaining system-dependent $C_{6,AB}$ coefficients is proposed by Tkatchenko and Scheffler [62]. In the so-called TS approach Eq. (2.13) is used to derive an expression for $C_{6,AB}$ coefficients which depends only on homonuclear parameters ($C_{6,AA}$, $C_{6,BB}$ and static polarizabilities α_A^0, α_B^0) for which free-atom reference values can be pre-computed—as in D4 via time-dependent DFT. Exploiting the direct relation between polarizability and volume, system-dependent homonuclear $C_{6,AA}$ coefficients are obtained by scaling the free-atom reference value based on the ratio of volumina associated with atom A in the system of interest and the free atom. Hirshfeld partitioning [63] allows to obtain volumina from the ground-state electron density such that the scaling can be performed by exploiting information from the underlying DFT calculations.

The many-body dispersion (MBD) [64] approach builds on TS and accounts for two additional effects. First of all, an atom located in a particular chemical environments will be affected by mutual interactions between fluctuating dipoles and the surrounding atoms will lead to an electrostatic screening of atomic polarizabilities. In the short-range XC effects are included in TS effective atomic polarizabilities by construction, but long-range electrostatic screening is not incorporated. In the MBD approach atomic polarizabilities containing both short- and long-range electrostatic screenings are obtained by modeling the surrounding in terms of a dipole field and solving the corresponding classical electrodynamic self-consistent screening (SCS) equations [65–67]. In a second step, the MBD approach uses the SCS results to construct a Hamiltonian that represents the atoms in a system by a collection of coupled isotropic three-dimensional quantum-harmonic oscillators, followed by (numerically) solving the Schrödinger equation [68, 69]. The fully nonadditive many-body dispersion contributions are then obtained as the difference between the zero-point energy of the coupled and uncoupled oscillators. Similar to the other approaches also MBD involves a damping function to couple the dispersion contribution to a (semi-)local DFT functional.

In noncovalently bonded systems each of the three approaches is capable of significantly improving the results of underlying DFT calculations. Individual characteristics, however, can make a certain approach more suitable for specific types of applications. The sophisticated description of MBD yields generally the most accurate results and has, thus, been used by us for a high-level treatment of molecular crystal structures in Refs. 1 and 2. Besides that both MBD and TS avoid introducing empiricism into the *ab initio* computations since all information is extracted from the electron density of the underlying DFT calculation (apart from the XC functional-dependent parameters required in the damping function). In contrast, there are numerous parameters entering the D4 framework, but evaluating the dispersion contribution is independent from the underlying electronic structure of a system. This characteristic makes it generally more suitable for coupling to methods other than DFT such as density-functional tight-binding.

2.2 Density-Functional Tight-Binding

The DFT approach presented in the previous section undoubtedly constitutes an eminent advancement for obtaining solutions to the Schrödinger equation in an efficient way. Nonetheless, substantial computational effort needs to be spent when screening vast databases or simulating large systems. Practical methods that allow to conduct such studies as well have been developed by introducing further approximations. Representatives of so-called semi-empirical methods are roughly three orders of magnitude faster than DFT owing to various approximations—such as relying on a small basis, and neglecting multi-center terms—and are still capable of yielding reasonable results as empirical tuning recovers most of the lost accuracy. While HF-based semi-empirical methods can be found in the literature [70–72], the purpose of this section is to provide insights into the DFT-based density-functional tight-binding (DFTB) method in its third order expansion as described in great detail in Refs. 73–78.

The development of DFTB models builds on a Taylor series of the total (KS) DFT energy expanded around a properly chosen reference density. Over the years, these models became more and more sophisticated by including higher order terms of the expansion. A compact expression that underlies the most recent extension to third order [78] is provided by

$$E^{\mathrm{DFTB}}[\rho_0+\delta\rho] = E^0[\rho_0] + E^1[\rho_0,\delta\rho] + E^2[\rho_0,(\delta\rho)^2] + E^3[\rho_0,(\delta\rho)^3] \tag{2.14}$$

with

$$E^0[\rho_0] = \underbrace{\frac{1}{2}\sum_{AB}\frac{Z_A Z_B}{R_{AB}} - \frac{1}{2}\iint\frac{\rho_0(\boldsymbol{r})\rho_0(\boldsymbol{r'})}{|\boldsymbol{r}-\boldsymbol{r'}|}\mathrm{d}\boldsymbol{r}\mathrm{d}\boldsymbol{r'} - \int V_{XC}[\rho_0]\rho_0(\boldsymbol{r})\mathrm{d}\boldsymbol{r} + E_{XC}[\rho_0]}_{E^{\mathrm{rep}}} \tag{2.15}$$

$$E^1[\rho_0,\delta\rho] = \underbrace{\sum_i n_i \langle\boldsymbol{\psi}_i|\,\hat{H}[\rho_0]\,|\boldsymbol{\psi}_i\rangle}_{E^{\mathrm{H}_0}} \tag{2.16}$$

$$E^2[\rho_0,(\delta\rho)^2] = \underbrace{\frac{1}{2}\iint\left(\frac{1}{|\boldsymbol{r}-\boldsymbol{r'}|} + \left.\frac{\delta^2 E_{XC}[\rho]}{\delta\rho(\boldsymbol{r})\delta\rho(\boldsymbol{r'})}\right|_{\rho_0}\right)\delta\rho(\boldsymbol{r})\delta\rho(\boldsymbol{r'})\mathrm{d}\boldsymbol{r}\mathrm{d}\boldsymbol{r'}}_{E^{\gamma}} \tag{2.17}$$

$$E^3[\rho_0,(\delta\rho)^3] = \underbrace{\frac{1}{6}\iiint\left.\frac{\delta^3 E_{XC}[\rho]}{\delta\rho(\boldsymbol{r})\delta\rho(\boldsymbol{r'})\delta\rho(\boldsymbol{r''})}\right|_{\rho_0}\delta\rho(\boldsymbol{r})\delta\rho(\boldsymbol{r'})\delta\rho(\boldsymbol{r''})\mathrm{d}\boldsymbol{r}\mathrm{d}\boldsymbol{r'}\mathrm{d}\boldsymbol{r''}}_{E^{\Gamma}}\,. \tag{2.18}$$

Here, the density of the ground state $\rho(\boldsymbol{r})$ is represented by the reference density $\rho_0(\boldsymbol{r})$ perturbed by density fluctuations $\delta\rho(\boldsymbol{r})$. For details about the terms E^0-E^3 and their derivation the reader is referred to the corresponding literature, for instance Refs. 73–75. Aiming for expressions of the individual terms that allow for a preferably fast evaluation of Eq. (2.14) the developers integrated several (further) approximations into the model. The derived simplified expressions E^{H_0}, E^{γ}, E^{Γ} and E^{rep} will be discussed in the following together with the main aspects of the underlying concepts.

The main part of the computational savings—responsible for 2-3 orders of magnitude compared to full DFT [75]—is related to the E^{H_0} term and realized by the interplay of several approximations applied in DFTB models. The derived simplified term reads

$$E^{\mathrm{H}_0} = \sum_i\sum_{AB}\sum_{\mu\in A}\sum_{\nu\in B} n_i c_{\mu i} c_{\nu i} H_{0,\mu\nu}, \tag{2.19}$$

where KS-orbitals $\boldsymbol{\psi}_i$ (compare Eq. (2.16)) are represented using a valence-only minimal basis $\boldsymbol{\phi}_\mu$

$$\boldsymbol{\psi}_i = \sum_\mu c_{\mu i}\boldsymbol{\phi}_\mu \tag{2.20}$$

in a linear-combination of atomic orbitals ansatz. The orbital basis set is obtained from numerical DFT calculations of individual atoms. In this case, the KS equations are, however, typically solved by applying an additional confinement potential which yields *compressed* atomic orbitals and, thus, atomic densities. The procedure is motivated by results of more sophisticated calculations on molecules and solids where it was found that electron densities can roughly be approximated as a superposition of compressed atomic densities [73]. Thus, without the confinement potential the resulting orbitals would be too diffuse to constitute an optimal basis. Moreover, a two-center approximation [79] is applied to the Hamiltonian

$$H_{0,\mu\nu} = \begin{cases} \epsilon^{\text{free atom}} & \text{if } \mu = \nu \\ \langle \phi_\mu | -\frac{1}{2}\nabla^2 + V[\rho_{0,A} + \rho_{0,B}] \, |\phi_\nu\rangle & \text{if } A \neq B \\ 0 & \text{if } A = B, \mu \neq \nu \end{cases} \tag{2.21}$$

where corresponding matrix elements can be pre-computed as function of interatomic distances between atom A and B for all element pairs. The numerical value related to the specific orientation of a *dimer* within an atomistic structure can be obtained by applying the Slater-Koster [80] combination rules which omits explicit evaluations of matrix elements during the runtime of a program.

For the term

$$E^{\gamma} = \frac{1}{2}\sum_{AB} \Delta q_A \Delta q_B \gamma^{\text{H}}_{AB}(R_{AB}) \tag{2.22}$$

density fluctuations are decomposed into atomic contributions, which in turn are approximated by applying multipole expansions where only the monopole terms are kept. As a result of this simplification the derived expression in Eq. (2.22) shows a dependency on atomic net charges Δq of individual atoms. These charges are obtained by employing a Mulliken charge analysis [81] based on the expansion coefficients of Eq. (2.20) defining the wave function ψ_i. This dependency triggers a self-consistent procedure to find the minimum of the overall DFTB energy which is enabled by applying the variational principle. By means of the obtained charges the function γ^{H}_{AB} is taking account for electron-electron interactions in E^{γ} and comprises two major aspects. For $A = B$ the function describes electron-electron interactions within one atom, namely on-site self-repulsion. This is expressed by means of the element-specific Hubbard parameter U_A (computed from DFT) which is twice the chemical hardness. This relationship can be understood using general concepts from atomic physics described in Ref. 82. In contrast, for $A \neq B$ and large distances γ^{H}_{AB} describes pure Coulomb interactions between two point charges Δq_A and Δq_B by reducing to basically R^{-1}_{AB}, i.e. contributions arising from the (semi-)local XC functional (compare Eq. (2.17)) are assumed to vanish. Moreover, the deviation from R^{-1}_{AB} is modelled by the covalent radii of the involved atoms which are determined by exploiting an inverse relationship to the corresponding Hubbard parameter. This relation is intuitive in that it implies a smaller chemical hardness for more diffuse atoms. An overall inverse relationship—as assumed by the DFTB framework—is obtained only for elements within one period though. Since the deviation is largest for hydrogen the γ^{H}_{AB} function is modified in case hydrogen is involved (indicated by the superscript H).

The simplified expression of the third order term (Eq. (2.18)) reads

$$E^{\Gamma} = \frac{1}{3}\sum_{AB} \Delta q_A^2 \Delta q_B \Gamma_{AB}(R_{AB}) \tag{2.23}$$

where the function Γ_{AB} corresponds to the derivative of γ^{H}_{AB} with respect to charge. Its incorporation into the DFTB methods induces a dependency on the charge state of an atom for the respective Hubbard parameter. Taking part in γ^{H}_{AB} (as discussed above) introduces a corresponding charge-dependency also to the E^{γ} term. Moreover, since the Hubbard parameter is related to the chemical hardness this enables a customized treatment dependent on the appearance of an atom as cation, anion or neutral species. Thus, the third order term improves the model applicability to systems with large net charges where local densities deviate significantly from the reference density.

The E^{rep} term comprises approximations of multiple terms (compare Eq. (2.15)), but is typically referred to as the *repulsion* term due to the dominating ion-ion repulsion at small distances. The collective contributions are expressed by

$$E^{\mathrm{rep}} = \frac{1}{2} \sum_{AB} V^{\mathrm{rep}}_{AB}(R_{AB}) \tag{2.24}$$

in terms of short-ranged repulsive potentials V^{rep}_{AB} between atom pairs which are specific to the combination of involved atom types. Dependency on the atom pair distance R_{AB} is obtained by fitting to experimental or high-level *ab initio* data—such as atomization energies, geometries, atomic forces and vibrational stretching frequencies—of representative reference systems. As a result of this practical aspect for constructing the repulsion term it bears a resemblance to XC functionals in DFT in that it effectively comprises multiple complex physical effects and uses simple functions for their description.

Besides the parameters required for constructing the repulsive potentials several other parameters entering the presented semi-empirical DFTB framework need to be determined. While some of the parameters (e.g. Hubbard parameters) can be obtained from atomic DFT calculations others are determined by means of specific parametrization schemes [83–85]. The 3ob parameter set, for instance, has been optimized for organic and biological application and was used in our studies on molecular crystals in Refs. 1 and 2.

Finally, since DFTB is derived from (semi-)local DFT it also inherits its deficiencies. As a result, dispersion interactions are not accounted for and correction schemes have been developed to introduce these effects. Recently, DFT dispersion-corrections schemes have been adapted to DFTB and *a posteriori* corrections according to D4, TS and MBD (discussed in the previous section) could be realized [78, 86]. The computational overhead of the dispersion-correction—which is typically negligible when combined with DFT—becomes more pronounced, however, when coupled with DFTB caused by the strongly reduced evaluation times of the underlying method. Aiming for a computationally efficient qualitative description of molecular crystal structures in Refs. 1 and 2 we therefore combined DFTB with TS and D4, respectively.

3 Machine Learning for Atomistic Systems

Approximate electronic structure methods like DFT and DFTB for modeling atomistic systems are physically or mathematically motivated. The applied approximations aim for a reduced level of complexity and associated computational costs which is typically achieved at the expense of predictive quality. An alternative way that allows to shortcut computational complexity is provided by means of machine learning (ML) and much effort has been spent to adopt the ML machinery to atomistic systems in the past two decades. In a more general sense, ML subsumes statistical algorithms whose performance enhance with supplied *experience* in a sense that regularities and patterns that underlie a data set can be *learned* and generalized. Owing to their universal formulation ML algorithms feature a broad spectrum of applications. On the other hand, this universal formulation induces the need for application-specific adjustment in order to use ML algorithms to full capacity and the endeavor of researchers in this respect enabled the realization of various applications for atomistic systems. An extensive overview about the current state of the field is provided in Refs. 33, 34, 45, 87. The following sections will focus on so-called *kernel*-based ML algorithms by introducing the most important (often interconnected) ingredients. This includes the construction of appropriate structural representations for their input, the definition of kernel-functions for corresponding similarity measurements and its subsequent application for modeling and revealing structure-property relationships in atomistic systems.

3.1 Representations

The way of representing chemical structures to ML algorithms plays a central role for their success. Dependent on the application at hand a suitable representation facilitates the model generation for structure-property mappings or the discovery of structure-related patterns in a data set. The process of converting information about atomic positions, chemical identities and lattice vectors into a suitable representation is called feature engineering [88].

This process can be used to provide additional information and incorporate some fundamental principles that apply to chemical systems which results in more robust, transferable and data-efficient ML models [38, 45]. Integrating known invariances directly into the representation, for instance, makes this information immediately available to the model. For the potential energy as one of the central properties corresponding invariances with respect to translation, rotation, as well as permutations of atoms of the same element can be incorporated in structural representations. Also other physical requirements such as smoothness for the structure-property mappings can be integrated.

Atom-centered representations—in contrast to global representation—describe a chemical structure by a set of local atomic environments. The physical significance of this representation lies in the assumption that properties of the chemical system can be constructed from individual atomic contributions. The validity of this additivity assumption is often justified as many properties are extensive in nature and contributions associated with a particular atom are mainly determined by its local environment. Moreover, the assumption offers several advantages for ML models such as a strongly reduced diversity within local environments—compared to the vast spaces associated with global structural arrangements—which decreases the required supply of corresponding training data (*vide infra*), as well as the risk of entering extrapolative regions. In addition to that the concept of locality increases the transferability as models can be applied to systems of different size and compositions more readily. A comprehensive overview of successfully applied representations is provided in Ref. 45.

The central representation used in the present dissertation is the so-called smooth overlap of atomic positions (SOAP) [44]. Here, the atomic environment $\boldsymbol{\chi}$ around a central atom is described by Gaussian functions located on each atom A in $\boldsymbol{\chi}$ via the neighborhood density

$$\rho_{\boldsymbol{\chi}_A}(\boldsymbol{R}) = \sum_{A\in\boldsymbol{\chi}} \exp\left(-\frac{|\boldsymbol{R}-\boldsymbol{R}_A|^2}{2\sigma_{\text{at}}^2}\right) \cdot f_{\text{cut}}(|\boldsymbol{R}|), \tag{3.1}$$

where the cutoff function f_{cut} ensures a smooth transition to zero at r_{cut} and the meaning of the involved length-related parameters (r_{cut} and σ_{at}) allows for a physically intuitive specification. An exemplary illustration of the neighborhood density is provided in Fig. 3.1.

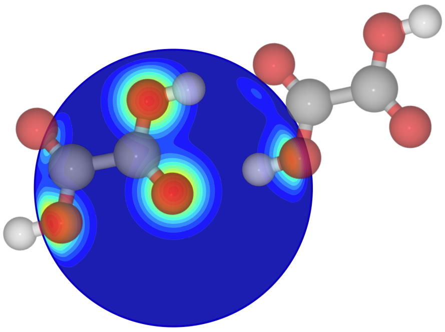


Fig. 3.1: *Visual depiction of a neighborhood density in SOAP for oxygen environments in an oxalic acid dimer. Red spheres: O, gray spheres: C, white spheres: H.*

Rotational invariance is achieved by expanding the neighborhood density in a basis of orthogonal radial functions and spherical harmonics and constructing the power spectrum

$$\theta_{nn'l}(\boldsymbol{\chi}) = \sqrt{\frac{8}{2l+1}} \sum_m (c_{nlm})^* c_{n'lm} \tag{3.2}$$

for the obtained expansion coefficients c_{nlm}. The spatial resolution of the neighborhood density is defined by the maximum values for l, n and n', respectively, used to construct the vector $\boldsymbol{\theta} = \{\theta_{nn'l}\}$. While vectors obtained in this way are suitable for representing single-component systems, multi-component systems can be described by separately constructing densities for each species β, computing power spectra $\theta_{nn'l}^{\beta\beta'}(\boldsymbol{\chi})$ for each pair of elements and concatenating the obtained values [33]. The resulting SOAP vectors are multi-body representations of atomic environments which are smooth with respect to atomic displacements and incorporate the invariances with respect to physical symmetries discussed above.

3.2 Kernel

Various ML applications applied in the context of chemical structures and corresponding local atomic environments are based on kernel functions. Generally, a kernel function $k(\boldsymbol{x}, \boldsymbol{x'})$ fulfills the requirement of being positive semidefinite and symmetric to swapping of its arguments $\boldsymbol{x}, \boldsymbol{x'}$. In context of chemical systems the arguments correspond to structural representations and the kernel function is used to measure the similarity between them. Similarities between passed arguments are returned by means of their inner product. The computation of the inner product, however, is performed (implicitly) in some higher dimensional space—called reproducing kernel Hilbert space (RKHS)—which is defined by the chosen kernel function. Based on this, so-called *kernel trick* it is possible to systematically introduce non-linearity in otherwise linear ML algorithms. An excellent tutorial introduction to kernel-based ML algorithms and its application to properties of small organic molecules is provided in Ref. 89.

Numerous kernel functions exist and appropriate choices are tailored to the learning task at hand. The SOAP kernel [35] is a prominent example for application to chemical systems and can be obtained from SOAP vectors (see previous section) as

$$k^{\mathrm{SOAP}}(\boldsymbol{\chi}, \boldsymbol{\chi'}) = \boldsymbol{\theta} \cdot \boldsymbol{\theta'} = \int_{\hat{R} \in \mathrm{SO}(3)} \mathrm{d}\hat{R} \left| \int \mathrm{d}\boldsymbol{R} \rho_{\boldsymbol{\chi}}(\boldsymbol{R}) \rho_{\boldsymbol{\chi'}}(\hat{R}\boldsymbol{R}) \right|^2 . \tag{3.3}$$

The equivalence to the overlap of neighbor densities integrated over 3D rotations emphasizes where the name SOAP has its origin. While Eq. (3.3) corresponds to a linear kernel many successful applications of the SOAP kernel employ low-order polynomial versions of it by raising the inner product to the power of ζ. The value selected for ζ defines the body order of the SOAP kernel and also influences the ability of the kernel to emphasis differences between atomic environments.

3.3 Machine Learning Interatomic Potentials

A subfield of ML are so-called supervised learning strategies which are capable of deriving functional relationships between inputs and outputs from underlying patterns in data sets. Dependent on the nature of the outputs a general distinction is made between classification for categorical values and regression for a continuous space of values. Machine-learned interatomic potentials (MLIP) [33–35] corresponds to the latter aiming for the mapping between representations of chemical systems and its potential energy surface.

3.3.1 Reference Data

Modern ML strategies achieve this mapping by generalizing information from a reference database $\{(y_i^{\mathrm{ref}}; \boldsymbol{x}_i^{\mathrm{ref}})\}_{i=1}^{M}$ storing representations of M input training structures $\boldsymbol{x}^{\mathrm{ref}}$ along with associated labels y^{ref} such as energies and forces. The database represents therefore a central component which is at the same time highly application specific. Its quality crucially impacts the effectiveness of the MLIP trained on it.

A model's capability to reproduce physical effects is defined *inter alia* by the level of theory employed to compute the labels. Numerical values are typically obtained from a computationally expensive high-level reference method. Alternatively, also energy and force differences between a high-level reference and a computationally much more efficient low-level method can be employed. Instead of modeling the PES directly such a Δ-ML [90, 91] approach provides a correction to be applied in combination with the low-level method which acts as a physical baseline. In the work we published in Refs. 1 and 2 we followed this strategy and present Δ-ML models of molecular crystals where the baseline resolves a conflict that arises from representing structures in terms of local atomic environments which complicates incorporating important long-range interactions (compare section 2.1.3) directly.

The computational expenses associated with evaluating the labels for the structures and related local environments stored in the reference database is a limiting factor. At the same time, however, these structures should correspond to a comprehensive representation for intended applications of the final MLIP as for extrapolative regions unphysical behaviour is expected. The generation of reference data is therefore a challenge on its own and several sampling strategies exist. Exploring the PES by means of molecular dynamics simulations at higher temperatures, for instance, prevents that models trained on such data enter extrapolative regions during production runs at lower temperatures. These samplings can be driven by *ab initio* methods, a computationally less expensive lower-level methods or in an iterative scheme with a preliminary MLIP initially trained on a small reference database [34]. In case a large pool of potential training structures is readily available a common task is to extract a concise set of structures from it for which computing labels on the reference level of theory is then performed. A reduction of redundant structural information to arrive at a set of diverse motifs can be achieved by applying a so-called farthest point sampling (FPS) [92, 93]. In FPS a maximization of structural diversity is accomplished by employing a kernel function (compare Section 3.2) to measure similarities and iteratively selecting the sample from the pool showing the greatest dissimilarity with respect to already drawn structures.

3.3.2 Loss Function

In search of a model capable of reproducing a target function the ultimate objective is to detect the function $\hat{f}$ from the space of all possible functions $\mathcal{F}$ where predicted output values y of related inputs $\boldsymbol{x}$ are in agreement with the target function at every single point in space [33], i.e.

$$\hat{f} = \arg\min_{f \in \mathcal{F}} \left[\int \mathcal{L}_{\mathrm{dev}}(f(\boldsymbol{x}), y) d(\boldsymbol{x}, y) \right], \tag{3.4}$$

where $\mathcal{L}$ measures deviations between predicted and actual function values. In consequence of the finiteness of available reference data and, thus, incomplete information about the target function the requirements on the model function $\hat{f}$ need to be modified. Besides predicting values in agreement with the provided reference data an optimal model function properly interpolates between them.

Predictions in agreement with the reference data is ensured for all ML models that minimize the function $\mathcal{L}$. The construction of ML models, however, starts from a highly general model class using so-called *universal approximators* with only minimal restrictions—such as smoothness—on the final functional form. Without taking countermeasures the learning process will therefore lead to overly complex models that provide a perfect description only for the reference data, but are virtually unusable for reliable predictions of any other input. In order to obtain models with a higher degree of generality loss functions are typically augmented by an additional term according to

$$\hat{f} = \arg\min_{f\in\mathcal{F}} \left[\sum_{i}^{M} \mathcal{L}_{\text{dev}}(f(\boldsymbol{x}_i^{\text{ref}}), y_i^{\text{ref}}) + \lambda\mathcal{R} \right] , \tag{3.5}$$

where the regularization $\mathcal{R}$ is used to penalize complexity and, thus, restrict the space of optimal solutions to simpler functions which are more likely to provide general models. As an example, Tikhonov regularization [94] favours solutions that will not heavily rely on individual input features and is achieved by means of the L^2-norm of coefficients defining the ML model (*vide infra*). The balance between the two terms in Eq. (3.5) critically influences the final model and can be controlled by $\lambda \geq 0$ which constitutes a hyperparameter. Insufficient regularization strength results in *overfitted* models showing a large *variance* between errors on training data and any other data. In contrast, disproportionate regularization causes *biased* or *underfitted* models which are too simplistic to provide reliable predictions. A graphical representation to visualize the effect of regularization is provided in Fig. 3.2. During model generation the *bias-variance trade-off* can by monitored by splitting a so-called validation set off from the training set and computing the prediction error on both sets. For a more general discussion on setting hyperparameters the reader is referred to Section 3.3.4.

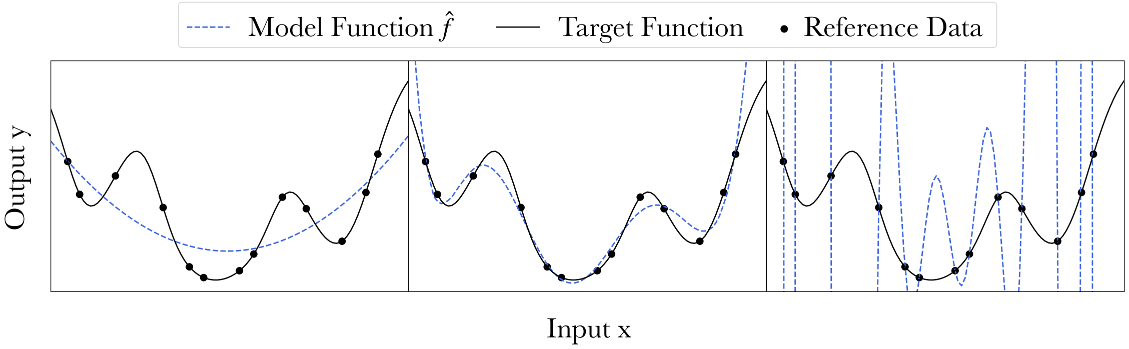


Fig. 3.2: *The influence of regularization on modeling a target function with a strongly regularized, underfitted model (left), a model with reasonable regularization (middle) and an overfitted model (right).*

3.3.3 Regression Tools

The process of modeling the functional relationship between inputs and outputs such that the loss function is minimized can be accomplished in various ways. A suitable strategy is to express the solution of this optimization problem in terms of kernel functions which has its theoretical basis in the representer theorem [95, 96]. It gives reason for approximating a target function in terms of a linear combination of kernel functions centered on the inputs of the training data and, thus, gives rise to the development of kernel-based methods. These methods are capable of describing complex nonlinear relationships between data points by means of the kernel-trick. Its application enables the conversion of any linear ML algorithm into a nonlinear version of it as long as outputs can be written in terms of inner products of the inputs. In this way, inputs are implicitly mapped to a higher-dimensional space (the RKHS) where the linear algorithm is applied and the result mapped back to original space is returned.

Gaussian process regression (GPR) is a prominent representative of kernel-based methods and in Refs. 1 and 2 we demonstrate its applicability to model the PES of molecular crystals and co-crystals by means of a Δ-ML scheme. This was accomplished by employing the Gaussian Approximation Potential (GAP) [36] developed by Bartók et al. which constitutes a general framework for GPR-based MLIPs. Insights into GPR will be provided in the following by means of two equivalent views on it (denoted as weight- and function-space view) that emphasise different aspects of the regression tool [34, 35, 89, 97, 98].

In weight-space view the starting point is to express the target function $f_\mathrm{t}(\boldsymbol{x})$—according to the representer theorem—in terms of scaled kernel functions

$$f_\mathrm{t}(\boldsymbol{x}) \approx \hat{f}(\boldsymbol{x}) = \sum_i^M s_i k(\boldsymbol{x}_i^\mathrm{ref}, \boldsymbol{x}) \tag{3.6}$$

where the sum iterates over the number of training set samples. Here, the kernel functions $k(\boldsymbol{x}_i^\mathrm{ref}, \cdot)$ can be viewed as basis functions which are centered at the training inputs and measure the similarity to any new input $\boldsymbol{x}$. The coefficients s_i associated with each of the M basis functions are obtained by minimizing the loss function

$$\mathcal{L} = \underbrace{\sum_i^M (y_i^\mathrm{ref} - \hat{f}(\boldsymbol{x}_i^\mathrm{ref}))^2}_{\mathcal{L}_\mathrm{dev}} + \underbrace{\sigma^2 \sum_{i,j}^M s_i k(\boldsymbol{x}_i^\mathrm{ref}, \boldsymbol{x}_j^\mathrm{ref}) s_j}_{\lambda\mathcal{R}} \tag{3.7}$$

where the training set samples $\{(y_i^\mathrm{ref}; \boldsymbol{x}_i^\mathrm{ref})\}_{i=1}^M$ enter the first term ($\mathcal{L}_\mathrm{dev}$), Tikhonov regularization is applied by means of the second term ($\mathcal{R}$) and the hyperparameter (λ) for controlling the balance between the terms is denoted by σ^2 (to facilitate comparison with function-space view later on). Note that in the regularization term coefficients of the *nonlinear* model are weighted by corresponding kernel elements which arises from penalizing complexity of the *linear* model in RKHS. Owing to the convexity of this optimization problem there exist a single solution to it and an analytical expression can be obtained by setting the derivative of $\mathcal{L}$ with respect to the coefficients to zero and re-arranging its terms for the coefficients. Written in matrix form this yields

$$\boldsymbol{s} = (\boldsymbol{K} + \boldsymbol{\Sigma})^{-1} \boldsymbol{y}^\mathrm{ref}, \tag{3.8}$$

where the kernel matrix $\boldsymbol{K} \in \mathbb{R}^{M \times M}$ comprises similarities between training inputs and $\boldsymbol{\Sigma} \in \mathbb{R}^{M \times M}$ is a diagonal matrix with elements σ^2. The elements in $\boldsymbol{\Sigma}$ can in principle also differ from one another with an associated meaning that is best understood from considering GPR in function-space view.

The function-space view corresponds to a probabilistic perspective on GPR where basis functions—in contrast to the weight-space view—are independent of the training data and will be used solely for defining a probability distribution of functions:

$$f_{\mathrm{t}}(\boldsymbol{x}) \approx \hat{f}(\boldsymbol{x}) = \sum_{h}^{H} w_h \Phi_h(\boldsymbol{x}). \tag{3.9}$$

This is achieved by considering each of the weights w_h to be drawn independently from a Gaussian probability distribution

$$P(w_h) \propto \mathcal{N}(0, \sigma_w^2). \tag{3.10}$$

The resulting probability distribution over functions is called a Gaussian process where σ_w^2 expresses prior beliefs about the overall variance of $f(\boldsymbol{x})$. In general, a Gaussian process corresponds to a collection of random variables with the characteristic that any finite number of it has a joint Gaussian distribution [97]. The function-space view employs this characteristic to infer predictions of new data points $\boldsymbol{x}$ based on the probability distribution of the reference observations $\{y_i^{\mathrm{ref}}\}_{i=1}^{M}$. The derivation of expression Eq. (3.8) from this perspective will provide additional insight into GPR and starts by considering the covariance between two observations

$$\left\langle \hat{f}(\boldsymbol{x})\hat{f}(\boldsymbol{x'}) \right\rangle = \int d\boldsymbol{w} P(\boldsymbol{w}) \sum_{h}^{H} w_h \Phi_h(\boldsymbol{x}) \sum_{h'}^{H} w_{h'} \Phi_{h'}(\boldsymbol{x'}) = \sigma_w^2 \sum_{h}^{H} \Phi_h(\boldsymbol{x}) \Phi_h(\boldsymbol{x'}). \tag{3.11}$$

The final expression in Eq. (3.11) is obtained from exploiting properties of the weights as drawing samples independently according to Eq. (3.10) gives rise to $\sigma_w^2 \delta_{hh'}$ when evaluating the integral. The scalar product of basis functions in this expression is used to define a kernel according to

$$k(\boldsymbol{x}, \boldsymbol{x'}) \equiv \sigma_w^2 \sum_{h}^{H} \Phi_h(\boldsymbol{x}) \Phi_h(\boldsymbol{x'}). \tag{3.12}$$

which induces kernel matrices build from scalar products in RKHS and corresponding properties that follow from its definition such as positive semidefiniteness. Moreover, for GPR only the kernel itself will be required to carry out the regression such that explicit knowledge about the basis functions is not essential. At the same time every basis gives rise to a different kernel such that the regression can be customized by choosing an appropriate kernel function.

The potential existence of noise on the reference observations $\{y_i^{\mathrm{ref}}\}_{i=1}^{M}$ is introduced by writing

$$\left\langle y_i^{\mathrm{ref}} y_j^{\mathrm{ref}} \right\rangle = k(\boldsymbol{x}_i^{\mathrm{ref}}, \boldsymbol{x}_j^{\mathrm{ref}}) + \delta_{ij}\sigma_i^2 \tag{3.13}$$

with the assumption that it is drawn independently for each sample from a Gaussian distribution with zero mean and variance σ_i^2. The possibility for applying (potentially) sample specific noise enables uncertainty to be assigned to individual training samples which allows for an imperfect fit to the data. The probability distributions for all reference observations is therefore expressed as multivariate Gaussian

$$P(\boldsymbol{y}^{\mathrm{ref}}) \propto \mathcal{N}(\mathbf{0}, \boldsymbol{K} + \boldsymbol{\Sigma}). \tag{3.14}$$

where the mean of the distribution is set to zero for notational simplicity (in accordance with the literature [35, 97] and justified since a prior guess for it could be subtracted from training data before fitting and added back after prediction). Predictions on new data points $\boldsymbol{x}$ are made by writing the joined Gaussian distribution with the training data and conditioning on the latter by means of Bayes' rule [99]

$$P(\boldsymbol{y}|\boldsymbol{y}^{\mathrm{ref}}) = \frac{P(\boldsymbol{y}^{\mathrm{ref}}, \boldsymbol{y})}{P(\boldsymbol{y}^{\mathrm{ref}})}. \tag{3.15}$$

which again yields a Gaussian distribution with mean $m(\boldsymbol{x})$ and variance $var(\boldsymbol{x})$ according to

$$m(\boldsymbol{x}) = \boldsymbol{k}^T \underbrace{(\boldsymbol{K} + \boldsymbol{\Sigma})^{-1}\boldsymbol{y}^{\mathrm{ref}}}_{\boldsymbol{s}} \qquad \text{and} \tag{3.16}$$

$$var(\boldsymbol{x}) = \boldsymbol{k} - \boldsymbol{k}^T(\boldsymbol{K} + \boldsymbol{\Sigma})^{-1}\boldsymbol{k}, \tag{3.17}$$

where $\boldsymbol{k}$ contains the kernel functions evaluated between $\boldsymbol{x}$ and the reference inputs $\{\boldsymbol{x}_i^{\mathrm{ref}}\}_{i=1}^M$. The function-space view on GPR therefore yields via Eq. (3.16) an expression for making predictions on a new data point in accordance with the weight-space view (compare Eq. (3.8)) and additionally delivers a measure for the uncertainty about this prediction by means of Eq. (3.17). Moreover, comparing both perspectives reveals an alternative interpretation for the hyperparameter σ^2 —regularization strength in weight-space and (potentially sample-resolved) uncertainty of reference observations in function-space—which provides guidance for its specification in practical applications.

Moreover, for modeling MLIPs with local atomic environments and, thus, local energies while *ab initio* methods provide energies for entire structures a customization of the general GPR process from above is required as described in detail in Refs. 35, 93 and 98. These citations additionally explain the incorporating of further information about the target function into the process which is typically available in terms of derivatives (forces and stresses) and is technically realized by means of kernel function derivatives.

Another aspect that needs to be considered in practical applications is the scaling behaviour in GPR. The scaling of *full* GPR (as presented above) with the number of training samples is (formally) cubic in terms of computational time and quadratic in terms of memory requirements, which makes it expensive for large data sets. *Sparse* GPR therefore pre-selects a representative set of samples for defining the locations of the kernels used as basis functions (in weight-space view). This reduces the number of coefficients that need to be determined (via matrix inversion in Eq. (3.8)) and, thus, regularized (Eq. (3.7)) and for predictions the kernel functions that need to be evaluated for each new sample (Eq. (3.16)). Ref. 35 provides a detailed description about sparse GPR and efficient strategies for practical implementation.

Finally it should be noted that besides kernel-based methods neural networks (NN) constitute another highly popular regression tool for generating MLIPs. In NNs a complex nonlinear function (e.g. the PES) is decomposed into a series of transformations which can be depicted by layers of connected neurons. An input (e.g. a structural representation) passing through these networks gets transformed by weights and biases that correspond to edges connecting the neurons. A neuron takes these weighted results, passes it through a nonlinear activation function and sends the obtained value to subsequently connected neurons. In this way, the original input gets mapped to several intermediate representations until a final output (e.g. an energy) is obtained. Optimal values for both types of parameters (weights and biases) need to be learned during the training process. Minimization of the loss function in NNs is typically carried out by random initialization of the parameters and subsequent optimization using stochastic gradient descent. Based on Behler and Parinello's work on high-dimensional neural networks potentials [38] numerous successful realization of MLIPs have been achieved such as ANI [39], PhysNet [40], GemNet [41] and SchNet [42] to name but a few.

3.3.4 Hyperparameters

Besides the model parameters to be determined by the regression algorithm during the learning process, ML models are typically accompanied by various hyperparameters. Values for these parameters are chosen beforehand and can be used to express prior beliefs about the data [33, 34]. In this way, the effectiveness of the model—its capability to generalize for instance—can be tuned. Consequences arising out of a particular value assigned to a hyperparameter strongly depend on its type. As an example, the regularization strength in a loss function is a hyperparameter that affects the complexity of a model (by affecting the magnitude of obtained regression coefficients). Others influence the composition of the model itself, for instance through choices about the type of kernel in GPR or the number of neurons and corresponding layers in a NN.

Interdependencies between hyperparameters and restrictions to a bound range or integers typically results in a rather complex space of values. Moreover, gradient-based methods for detecting optimal values for them is often not possible in an effective way due to the lack of analytical derivatives and the computational demands for numerical evaluations caused by the induced re-training of the model. At the same time, however, the model performance is typically fairly robust to small value changes for many hyperparameters. This allows for defining heuristics that are found to work reasonably well across different data sets. Alternatively, optimal values can be detected by means of random or grid searches combined with educated guesses for the search ranges.

When optimizing values for hyperparameters, each trial combination is used to train a separate model and subsequently selecting the one showing the best performance. Its quality is judged by the so-called generalization error obtained from predictions on a validation set which has been split off before training. More advanced methods such as k-fold cross-validation can be used to obtain better statistics on the generalization error. Here, multiple models are generated for each trial combination of hyperparameters by splitting the training data in k folds and performing the corresponding number of training rounds using $k-1$ folds while the remaining one servers as validation set.

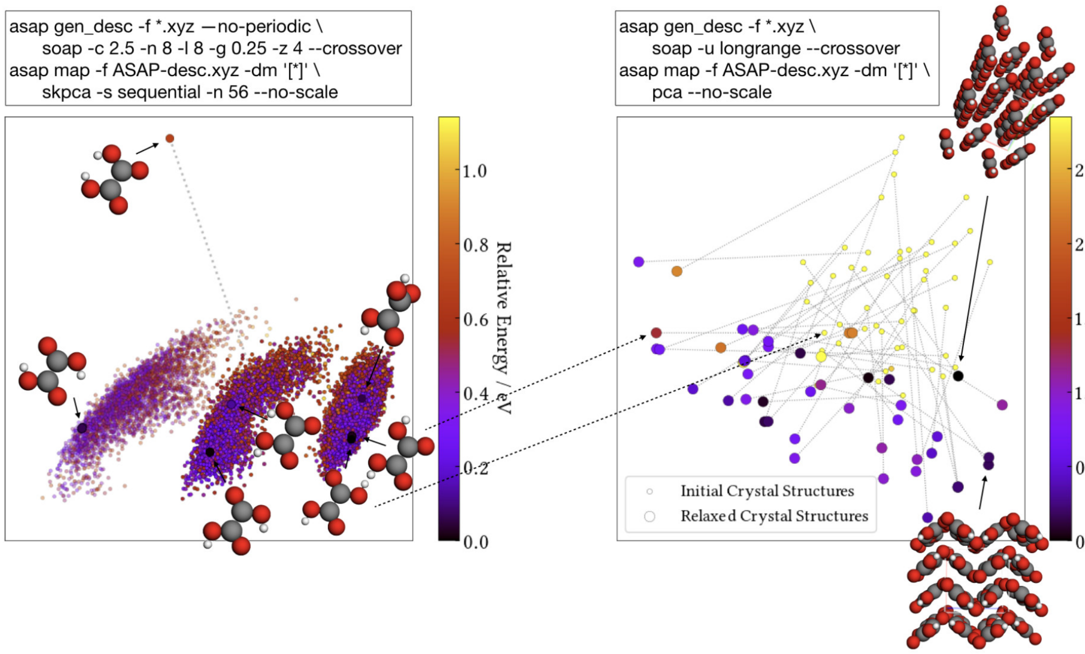
3.4 Kernel Principal Component Analysis

Gaining a higher-level understanding about chemical data sets is often difficult, due to the high dimensionality of structural representations like SOAP which complicates a visual representation and detecting underlying patterns. Preserving such patterns in a few meaningful dimensions is the goal of dimensionality reduction techniques which transform the original data in a way that keeps some relationship between points in high and low dimensional space intact. The scalar product between data points, for instance, constitutes the relationship preserved in principal component analysis (PCA) [100].

In PCA the initial step constitutes the construction of a covariance matrix from feature vectors of the data in original space. Based on the obtained matrix a new coordinate system for representing the data points is defined by taking advantage of associated eigenvalues and -vectors. While the eigenvectors constitute the axis of the new basis its eigenvalues serve as a measure for the associated information content which allows to arrange the axis accordingly. Eigenvectors corresponding to large eigenvalues point along directions with high variance and, thus, are assumed to be particularly suited for representing the data while the information content along directions with low variance is assumed to be insignificant. Most of the total variance in a data set is often explained by just a few directions. These so-called principal components—linear combinations of the axes of the original space—are then used to gain insight into the data set while remaining dimensions are typically ignored.

Since PCA is a linear technique, but relies on scalar products between feature vectors, it is possible to arrive at a non-linear version of it by applying the kernel trick (compare Section 3.2). Kernel principal component analysis (KPCA) [101] therefore uses the kernel matrix between data points—instead of the covariance matrix—for obtaining meaningful axes. Depending on the type of kernel applied in KPCA the resulting principal components will then be capable of revealing non-linear relationships in the set.

In the present dissertation both PCA and KPCA have been used to illustrate how low-dimensional projections of chemical structures and subsequent visualization can be utilized to gain insights in corresponding data sets. A reproduced example from Ref. 3 is shown in Fig. 3.3 for which KPCA has been applied in conjunction with the SOAP kernel. Finally, note that apart from the techniques presented in this section also others are frequently used for representing data sets of chemical structures such as t-distributed stochastic neighbor embedding [102], sketch-map [103] and the uniform manifold approximation and projection [104].



```
asap gen_desc -f *.xyz —no-periodic \
     soap -c 2.5 -n 8 -l 8 -g 0.25 -z 4 --crossover
asap map -f ASAP-desc.xyz -dm '[*]' \
     skpca -s sequential -n 56 --no-scale
```

```
asap gen_desc -f *.xyz \
     soap -u longrange --crossover
asap map -f ASAP-desc.xyz -dm '[*]' \
     pca --no-scale
```

Fig. 3.3: *(top) The two boxes show the commands used in combination with the ASAP code—which has been developed in conjunction with our work in Ref. 3—to generate the corresponding plots. (Left) KPCA map of gas-phase oxalic acid conformers (large points) that served as initial structures to sample configurations from various molecular dynamics (MD) simulations at 500 K (small points). Transparent points refer to configurations of MD simulations without transitions to other basins. (right) Randomly initialized crystal structures for oxalic acid (small yellow circles) along with its fully relaxed counterparts (large colored circles). The two visual representations of crystal structures correspond to the experimentally known α [105] (lower) and β [106] (upper) polymorph. Although structure initialization was conducted from the same gas-phase conformer for all trial crystals conformational changed upon relaxation have been observed in some cases (indicated by arrows across the panels). Reprinted with permission from Ref. 3. Copyright © 2020 American Chemical Society*

4 Publications

4.1 Data-efficient machine learning for molecular crystal structure prediction

Simon Wengert, Gábor Csányi, Karsten Reuter and Johannes T. Margraf

Chem. Sci. **12**, 4536 (2021)
DOI: https://doi.org/10.1039/D0SC05765G

4.1.1 Content

We present a data-efficient workflow to generate Δ-ML models for molecular crystals that feature high accuracy/cost ratios. To achieve this we enhance the description of a low-cost physical baseline by applying GPR-based ML corrections trained on high-quality *ab initio* reference data, most notably obtained from dispersion-corrected DFT (DFT+MBD). For the baseline we find DFTB in combination with TS dispersion-correction (DFTB+TS) to constitute an appropriate method for our approach since short evaluation times allow for extensive search space explorations in molecular CSP while reasonable descriptions of long-range interactions are provided.

This is of particular relevance since these interactions are important for an accurate description of molecular crystals, but are outside the range of our ML corrections where structures are represented in terms of local atomic environments using SOAP. Reference environments for generating ML corrections are based on a concise set of crystal structures obtained from a large pool by diversity-driven selections using FPS. These structures serve as templates to create training sets of two separate ML models for correcting intra- and intermolecular interactions, respectively. While the former is trained *inter alia* on the individual molecular conformations comprised in the crystal templates, extracted molecular clusters of varying size enter the training process of the latter. Separating the overall target function into regression problems for intra- and intermolecule interactions facilitates the generation of the individual models, while the complete avoidance of periodic reference structures enables a training on high-level methods also beyond DFT which would otherwise be computationally prohibitive. This is demonstrated by employing spin-component-scaled second-order Møller–Plesset theory (SCS-MP2) as the high-level reference in one case.

Moreover, we could show that our approach is broadly applicable to different molecular materials by considering various representative test systems. For each test system we verify the overall accuracy of obtained Δ-ML models for both crystal stabilities and relative stability rankings. Additionally, we demonstrate that our approach yields models that allow for reliable structure relaxations which can be conducted with a computational effort orders of magnitude smaller than the high-level target method even taking training costs into account.

4.1.2 Individual Contributions

The ideas underlying the presented workflow have been jointly conceived and constantly further developed by Johannes T. Margraf, Gábor Csányi and myself. Johannes T. Margraf and I wrote the manuscript which has been further edited by Gábor Csányi and Karsten Reuter. Gábor Csányi contributed particularly with substantial support regarding the GAP framework and SOAP. Johannes T. Margraf provided an early version of the code for gas-phase molecular dynamics simulations which has been further modified by myself to generate the molecular configurations that augment the intramolecular training set. Moreover, I created all additional code necessary to conduct the presented workflow. This includes *inter alia* the python-based MLtools package used to perform large parts of the workflow related to fitting the GAP models and the farthest point sampling, as well as extracting both single molecules and (unique) sets of molecular clusters from crystal structures. Additionally, I wrote an ASE-based calculator to combine the baseline method with the intra- and intermolecular ML corrections and to perform single-point calculations and structure optimizations with these Δ-ML models. Finally, I conducted all model generations, the subsequent analysis—including the timings—and created the figures.

4.2 A Hybrid Machine Learning Approach for Structure Stability Prediction in Molecular Co-crystal Screenings

Simon Wengert, Gábor Csányi, Karsten Reuter and Johannes T. Margraf

J. Chem. Theory Comput. **18**, 4586 (2022)
DOI: https://doi.org/10.1021/acs.jctc.2c00343

4.2.1 Content

On the basis of our approach described in Ref. 1 for single-component crystals we present an extended workflow that enables the generation of Δ-ML models for structure predictions in molecular co-crystal screenings. We demonstrate the applicability of this approach for co-crystals of variable composition consisting of a representative active pharmaceutical ingredient (paracetamol) and various co-former candidates (oxalic acid, naphthalene, phenazine and theophylline).

For each molecular type a separate model for intramolecular ML corrections is trained, while intermolecular interactions are corrected with a common ML model which captures the individual pairs between central pharmaceutical and co-formers including variations of the respective stoichiometric ratio. Both types of ML corrections are trained on energy and force differences using the GAP framework with dispersion-corrected hybrid DFT (denoted as PBE(0)+MBD) as the high-level reference method, while in final models (denoted as Δ-GAP) the combination with the dispersion-corrected DFTB baseline (DFTB+D4) delivers corrected energies, forces, as well as stresses.

We show that these Δ-GAP models reliably yield energies and forces in agreement with PBE(0)+MBD by means of a comprehensive set of test crystals. We further substantiate the robustness of our approach by explicit investigation of structures beyond the scope of the training set—in terms of packing density and stability—by considering the experimentally known co-crystals of our test systems. For each system we optimize atomic positions as well as unit cell parameters and find Δ-GAP structures reproducing PBE(0)+MBD results with high accuracy at a much lower computational cost. Moreover, we apply our Δ-GAP models to perform molecular dynamics simulations at ambient conditions and obtain co-crystal densities in agreement with experimental measurements.

4.2.2 Individual Contributions

The idea for extending the approach for single-component crystals presented in Ref. 1 to co-crystals has been jointly conceived by all contributing authors and Johannes T. Margraf, Gábor Csányi and myself have been involved in the realization process. Johannes T. Margraf and I wrote the manuscript which has been further edited by Gábor Csányi and proofread by Karsten Reuter. I created all code necessary for extending the workflow to co-crystals. This includes an ASE-based calculator for PBE(0)+MBD (defined in the manuscript) single-point calculations and structure optimizations. Moreover, based on the ASE calculator that has originally been written for single-component Δ-ML models, I implemented the enhancements required for co-crystal systems, as well as application of intra- and intermolecular ML corrections to stresses for performing simulations that involve variations in lattice parameters. Finally, I conducted all model generations, the subsequent analysis and created the figures.

4.3 Mapping Materials and Molecules

Bingqing Cheng, Ryan-Rhys Griffiths, Simon Wengert, Christian Kunkel, Tamas Stenczel, Bonan Zhu, Volker L. Deringer, Noam Bernstein, Johannes T. Margraf, Karsten Reuter, and Gábor Csányi

Acc. Chem. Res. **53**, 1981 (2020)
DOI: https://doi.org/10.1021/acs.accounts.0c00403

4.3.1 Content

This work focuses on the combination of unsupervised learning with atomic structure representations for visualization and analysis of molecular and materials data sets. With the ever-growing computational power the extent of such sets is constantly increasing and our work, thus, addresses the associated desire for exploring such data efficiently, thereby revealing underlying patterns and gaining physical and chemical insights.

In our work we rely on SOAP for describing local atomic environments and averaged versions of it for a global representation of entire systems. After having converted the entries of a data set accordingly, we couple these representations to PCA and KPCA in order to measure corresponding similarities with subsequent projection to low-dimensional spaces suitable for visualizations. We conveniently provide an automated and user-friendly command-line tool for these mappings via the Automatic Selection And Prediction tools for materials and molecules (ASAP) code. Complementary to it, we provide a web-browser based viewing tool that allows for interactive explorations of obtained maps by additionally displaying the 3D-structures along with attributes associated with a selected data point. Moreover, in the course of this work universal heuristics for setting SOAP hyperparameters have been formulated that are applicable to systems of arbitrary chemical composition which further facilitates an efficient exploration of atomistic structure data sets.

To showcase the usefulness of the data-driven framework presented we provide examples for a wide variety of systems along with the corresponding data which allows to reproduce the obtained maps readily. These example systems comprise crystalline and amorphous materials, as well as interfaces and data sets of organic molecules and cover fields of application such as exploring the results of random structure searches, inspecting the composition of training sets in ML and the analysis of molecular dynamics trajectories.

4.3.2 Individual Contributions

The idea underlying this collaborative work emerged from discussions between the authors that all work on machine learning and need to deal with large data sets of atomistic structures. In the course of preparing this manuscript all contributing authors jointly took part in improving on it by providing ideas, code or data sets of illustrative examples, as well as writing and editing the individual sections. Thus, a distinct assignment of all contributions to individual authors is generally not appropriate for this work. Nevertheless, the browser-based visualizer for atomistic structure data sets was written by Christian Kunkel (who also provided the basis for the code), Tamas Stenczel and myself. Additionally, I contributed with a section on visualizing molecular dynamics trajectories of oxalic acid molecules and full unit cell relaxations of corresponding crystal structures and partially contributed to the section on visualizing the well-established molecular QM9 data set, specifically the carbon environments part of it.

5 Summary, Conclusion and Outlook

In summary, in the work published in Ref. 1 we developed a kernel-based supervised learning approach to generate hybrid models for molecular crystals by augmenting a low-cost physics-based description of long-range interactions with short-range ML models trained on high-quality *ab initio* reference data. The approach features a high data-efficiency for generating these hybrid models for which, in turn, a large accuracy/cost ratio has been verified with computational demands mainly concerned with the applied semi-empirical baseline method. We could substantiate the applicability of these hybrid models to CSP by verifying the agreement with the high-level method in terms of stability and molecular arrangements upon structure optimization. The complete avoidance of periodic reference structures enables reference methods even beyond DFT which paves the way for applying correlated wave function methods in molecular CSP studies. Moreover, we showed that the presented approach is broadly applicable to different single-component molecular materials.

On this basis, extensions to our approach have been presented in Ref. 2 with focus on co-crystal systems which are predestined for property-driven screenings through a systematic variation of components. With this in mind, we combined a representative pharmaceutical with several co-forming molecules in various stoichiometric ratios to demonstrate the predictive capability of corresponding hybrid models. We could show that predictions agree with the high-level reference method even beyond the scope of the training structures and by means of molecular dynamics simulations at ambient conditions also with experimental measurements.

Additionally, we have been investigating the applicability of kernel-based unsupervised machine learning for dimensionality-reduction and subsequent visualization in Ref. 3. Combining (kernel) principal component analysis with sophisticated representations of atomistic structures based on SOAP could be shown to be applicable in an automated fashion and with great versatility by means of illustrative examples such as examining molecular dynamics trajectories, exploring results from crystal structure searches or revealing underlying patterns associated with atomic environments in an established molecular database.

To conclude, at the center of the dissertation at hand are kernel-based machine learning methods and their application to atomistic structures with strong focus on molecular systems, particularly molecular crystals. In the course of this work significant contributions have been made in obtaining accurate and computationally efficient models for application in molecular crystal structure prediction and associated global optimization problems by developing supervised learning schemes for their generation. The advanced capability of these models to explore the vast configurational spaces with high energetic precision, including the efficient optimization of all relevant degrees of freedom can be employed to facilitate the reliability of *in silico* structure predictions and ultimately corresponding screening studies for a well-directed discovery of novel materials. Here, the ongoing developments towards accounting for long-range interactions in machine learning models directly are highly promising since it provides a basis for reliable models without additional costs for the baseline method and, thus, a further enhanced efficiency in the assessment of crystal stabilities. Moreover, the emergence of differential models with high accuracy/cost ratio makes reliable stability comparisons based on free energies and the incorporation of experimental conditions into the assessment more readily accessible. On another note, considering the reduced demands for generating large data sets of atomistic structures the resulting desire for a simplified exploration and analysis has been addressed by demonstrating the applicability of kernel-based unsupervised learning to atomistic structure data sets and the development of convenient tools for their exploration and analysis. Altogether, the work underlying the present dissertation provides a valuable contribution to the rapidly evolving field of machine learning in computational chemistry which is expected to (further) enable numerous advances that have previously been out of reach.

Acknowledgments / Danksagung

First and foremost, I want to thank Karsten Reuter for his support throughout the course of this thesis. Thank you for having confidence in me and the selected direction to evolve my project, as well as your strategic advice whenever it was needed. Moreover, I am very happy about the opportunity to work with you and the whole group first at the Chair of Theoretical Chemistry in Munich and then at the Fritz-Haber-Institute here in Berlin.

Likewise, I want to thank Johannes Margraf for the intensive and great cooperation throughout all the years. I enjoy working with you very much and I undoubtedly benefit from your impressively extensive knowledge. Thanks a lot for your general constructive spirit and your continuous supporting role in all sorts of science-related questions and beyond. I further want to express my gratitude to Gábor Csányi for the many fruitful discussions and for all the valuable and inspiring ideas. I very much enjoyed my research stay at your group in Cambridge and I am grateful for the warm welcome you offered.

Above all, I want to thank Sina for the great time we have had together since we first met. Being able to share all the experiences of our common journey is a gift that I am very grateful for. Thank you also for your unconditional support in every situation. Von ganzem Herzen dankbar bin ich meinen Eltern, die mir so vieles im Leben erst möglich gemacht haben. Vielen Dank für all die schönen Momente, euren bedingungslosen Rückhalt und das Gefühl der Geborgenheit. Ganz besonderer Dank gilt dir, Papa, für deine Unterstützung und deinen Rat, auf den ich mich immer verlassen konnte. Ich bin unendlich dankbar für die wunderbare Zeit, die ich mit dir verbringen konnte, deine Warmherzigkeit und deine Liebe. Die Verbundenheit und die fröhlichen Momente mit der Familie ist etwas worüber ich ebenfalls sehr dankbar bin, sowie für die langjährige und tiefe Freundschaft mit Paul Bareiß und Florian Seibold.

I also want to express my gratitude to Dirk Flottmann. I very much enjoy our mutual assistance, your advice in our occasional meetings and your refreshing personality. Furthermore, I am very grateful for the friendly atmosphere within the groups at TUM and FHI and I want to thank all the people that contribute to this spirit. Particularly, I want to express my gratitude to the organizers of great events such as our group retreats and international food evenings. Moreover, I will remember all the good times together with Arobendo Mondal, Martin Deimel, Thorben Eggert, Hanna Türk and many more current and former group members, too many to name them all. I want to particularly acknowledge our running team and the mutual motivation, which made it possible for us to complete the Berlin Marathon. A big thank you goes to Julia Pach and Ruth Mösch for their helpfulness in terms of bureaucracy and I also want to thank Chiara Panosetti who initially proposed this project to me.

Finally, I want to acknowledge funding from Deutsche Forschungsgemeinschaft (DFG) under project RE1509/18-2, as well as the computational resources provided by the Leibniz Supercomputing Centre and the Max Planck Computing and Data Facility.

Bibliography

[1] S. Wengert, G. Csányi, K. Reuter, and J. T. Margraf, *Data-efficient machine learning for molecular crystal structure prediction*, Chem. Sci. **12**, 4536 (2021) (cit. on pp. i, 2, 6, 8, 11, 16, 18, 29, 33, 47).

[2] S. Wengert, G. Csányi, K. Reuter, and J. T. Margraf, *A Hybrid Machine Learning Approach for Structure Stability Prediction in Molecular Co-crystal Screenings*, J. Chem. Theory Comput. **18**, 4586 (2022) (cit. on pp. i, 2, 6, 8, 11, 16, 18, 33, 59).

[3] B. Cheng, R.-R. Griffiths, S. Wengert, C. Kunkel, T. Stenczel, B. Zhu, V. L. Deringer, N. Bernstein, J. T. Margraf, K. Reuter, and G. Csanyi, *Mapping Materials and Molecules*, Acc. Chem. Res. **53**, 1981 (2020) (cit. on pp. i, 2, 22, 23, 33, 69).

[4] K. S. Novoselov, A. K. Geim, S. V. Morozov, D. Jiang, Y. Zhang, S. V. Dubonos, I. V. Grigorieva, and A. A. Firsov, *Electric Field Effect in Atomically Thin Carbon Films*, Science **306**, 666 (2004) (cit. on p. 1).

[5] H. Furukawa, K. E. Cordova, M. O'Keeffe, and O. M. Yaghi, *The Chemistry and Applications of Metal-Organic Frameworks*, Science **341**, 1230444 (2013) (cit. on p. 1).

[6] B. Moulton and M. J. Zaworotko, *From Molecules to Crystal Engineering: Supramolecular Isomerism and Polymorphism in Network Solids*, Chem. Rev. **101**, 1629 (2001) (cit. on p. 1).

[7] J. W. Steed and J. L. Atwood, *Supramolecular chemistry* (John Wiley & Sons, 2022) (cit. on p. 1).

[8] J. W. Steed, *Should solid-state molecular packing have to obey the rules of crystallographic symmetry?* CrystEngComm **5**, 169 (2003) (cit. on p. 1).

[9] C. A. Gunawardana and C. B. Aakeröy, *Co-crystal synthesis: fact, fancy, and great expectations*, Chem. Commun. **54**, 14047 (2018) (cit. on p. 1).

[10] E. Jurczak, A. H. Mazurek, Ł. Szeleszczuk, D. M. Pisklak, and M. Zielińska-Pisklak, *Pharmaceutical Hydrates Analysis—Overview of Methods and Recent Advances*, Pharmaceutics **12**, 959 (2020) (cit. on p. 1).

[11] D. J. Berry and J. W. Steed, *Pharmaceutical cocrystals, salts and multicomponent systems; intermolecular interactions and property based design*, Adv. Drug Deliv. Rev. **117**, 3 (2017) (cit. on p. 1).

[12] A. J. Cruz-Cabeza, S. M. Reutzel-Edens, and J. Bernstein, *Facts and fictions about polymorphism*, Chem. Soc. Rev. **44**, 8619 (2015) (cit. on p. 1).

[13] J. Aaltonen, M. Allesø, S. Mirza, V. Koradia, K. C. Gordon, and J. Rantanen, *Solid form screening – A review*, Eur. J. Pharm. Biopharm. **71**, 23 (2009) (cit. on p. 1).

[14] S. Karki, T. Friščić, L. Fábián, P. R. Laity, G. M. Day, and W. Jones, *Improving Mechanical Properties of Crystalline Solids by Cocrystal Formation: New Compressible Forms of Paracetamol*, Adv. Mater. **21**, 3905 (2009) (cit. on p. 1).

[15] S. Domingos, V. André, S. Quaresma, I. C. B. Martins, M. F. Minas da Piedade, and M. T. Duarte, *New forms of old drugs: improving without changing*, J. Pharm. Pharmacol. **67**, 830 (2015) (cit. on p. 1).

[16] A. M. Reilly, R. I. Cooper, C. S. Adjiman, S. Bhattacharya, A. D. Boese, J. G. Brandenburg, P. J. Bygrave, R. Bylsma, J. E. Campbell, R. Car, D. H. Case, R. Chadha, J. C. Cole, K. Cosburn, H. M. Cuppen, F. Curtis, G. M. Day, R. A. DiStasio Jr, A. Dzyabchenko, B. P. van Eijck, D. M. Elking, J. A. van den Ende, J. C. Facelli, M. B. Ferraro, L. Fusti-Molnar, C.-A. Gatsiou, T. S. Gee, R. de Gelder, L. M. Ghiringhelli, H. Goto, S. Grimme, R. Guo, D. W. M. Hofmann, J. Hoja, R. K. Hylton, L. Iuzzolino, W. Jankiewicz, D. T. de Jong, J. Kendrick, N. J. J. de Klerk, H.-Y. Ko, L. N. Kuleshova, X. Li, S. Lohani, F. J. J. Leusen, A. M. Lund, J. Lv, Y. Ma, N. Marom, A. E. Masunov, P. McCabe, D. P. McMahon, H. Meekes, M. P. Metz, A. J. Misquitta, S. Mohamed, B. Monserrat, R. J. Needs, M. A. Neumann, J. Nyman, S. Obata, H. Oberhofer, A. R. Oganov, A. M. Orendt, G. I. Pagola, C. C. Pantelides, C. J. Pickard, R. Podeszwa, L. S. Price, S. L. Price, A. Pulido, M. G. Read, K. Reuter, E. Schneider, C. Schober, G. P. Shields, P. Singh, I. J. Sugden, K. Szalewicz, C. R. Taylor, A. Tkatchenko, M. E. Tuckerman, F. Vacarro, M. Vasileiadis, A. Vazquez-Mayagoitia, L. Vogt, Y. Wang, R. E. Watson, G. A. de Wijs, J. Yang, Q. Zhu, and C. R. Groom, *Report on the sixth blind test of organic crystal structure prediction methods*, Acta Cryst. B **72**, 439 (2016) (cit. on pp. 1, 2).

[17] A. Groß, *Theoretical Surface Science*, 2nd ed. (Springer, Berlin, Heidelberg, 2009) (cit. on pp. 1, 3).

[18] W. Koch and M. Holthausen, *A Chemist's Guide to Density Functional Theory*, 2nd ed. (Wiley-VCH, Weinheim, 2009) (cit. on pp. 1, 3, 5, 6).

[19] F. Jensen, *Introduction to Computational Chemistry*, 2nd ed. (John Wiley & Sons, Chichester, 2007) (cit. on pp. 1, 3, 6).

[20] M. Born and R. Oppenheimer, *Zur Quantentheorie der Molekeln*, Ann. Phys. **389**, 457 (1927) (cit. on pp. 1, 3).

[21] S. M. Woodley, G. M. Day, and R. Catlow, *Structure prediction of crystals, surfaces and nanoparticles*, Philos. Trans. R. Soc. A **378**, 20190600 (2020) (cit. on p. 1).

[22] J. Maddox, *Crystals from first principles*, Nature **335**, 201 (1988) (cit. on p. 1).

[23] S. L. Price, *Predicting crystal structures of organic compounds*, Chem. Soc. Rev. **43**, 2098 (2014) (cit. on p. 1).

[24] X. Li, F. S. Curtis, T. Rose, C. Schober, A. Vazquez-Mayagoitia, K. Reuter, H. Oberhofer, and N. Marom, *Genarris: Random generation of molecular crystal structures and fast screening with a Harris approximation*, J. Chem. Phys. **148**, 241701 (2018) (cit. on p. 1).

[25] R. Tom, T. Rose, I. Bier, H. O'Brien, Á. Vázquez-Mayagoitia, and N. Marom, *Genarris 2.0: A random structure generator for molecular crystals*, Comput. Phys. Commun. **250**, 107170 (2020) (cit. on p. 1).

[26] S. Yang and G. M. Day, *Exploration and Optimization in Crystal Structure Prediction: Combining Basin Hopping with Quasi-Random Sampling*, J. Chem. Theory Comput. **17**, 1988 (2021) (cit. on p. 1).

[27] H. Song, L. Vogt-Maranto, R. Wiscons, A. J. Matzger, and M. E. Tuckerman, *Generating Cocrystal Polymorphs with Information Entropy Driven by Molecular Dynamics-Based Enhanced Sampling*, J. Phys. Chem. Lett. **11**, 9751 (2020) (cit. on p. 1).

[28] D. H. Case, J. E. Campbell, P. J. Bygrave, and G. M. Day, *Convergence Properties of Crystal Structure Prediction by Quasi-Random Sampling*, J. Chem. Theory Comput. **12**, 910 (2016) (cit. on p. 1).

[29] A. G. Shtukenberg, Q. Zhu, D. J. Carter, L. Vogt, J. Hoja, E. Schneider, H. Song, B. Pokroy, I. Polishchuk, A. Tkatchenko, A. R. Oganov, A. L. Rohl, M. E. Tuckerman, and B. Kahr, *Powder diffraction and crystal structure prediction identify four new coumarin polymorphs*, Chem. Sci. **8**, 4926 (2017) (cit. on p. 2).

[30] N. Marom, R. A. DiStasio Jr., V. Atalla, S. Levchenko, A. M. Reilly, J. R. Chelikowsky, L. Leiserowitz, and A. Tkatchenko, *Many-Body Dispersion Interactions in Molecular Crystal Polymorphism*, Angew. Chem. Int. Ed. **52**, 6629 (2013) (cit. on p. 2).

[31] A. M. Reilly and A. Tkatchenko, *Understanding the role of vibrations, exact exchange, and many-body van der Waals interactions in the cohesive properties of molecular crystals*, J. Chem. Phys. **139**, 024705 (2013) (cit. on p. 2).

[32] J. Hoja, H.-Y. Ko, M. A. Neumann, R. Car, R. A. DiStasio, and A. Tkatchenko, *Reliable and practical computational description of molecular crystal polymorphs*, Sci. Adv. **5**, eaau3338 (2019) (cit. on p. 2).

[33] J. A. Keith, V. Vassilev-Galindo, B. Cheng, S. Chmiela, M. Gastegger, K.-R. Müller, and A. Tkatchenko, *Combining Machine Learning and Computational Chemistry for Predictive Insights Into Chemical Systems*, Chem. Rev. **121**, 9816 (2021) (cit. on pp. 2, 13–16, 21).

[34] O. T. Unke, S. Chmiela, H. E. Sauceda, M. Gastegger, I. Poltavsky, K. T. Schütt, A. Tkatchenko, and K.-R. Müller, *Machine Learning Force Fields*, Chem. Rev. **121**, 10142 (2021) (cit. on pp. 2, 13, 15, 16, 18, 21).

[35] V. L. Deringer, A. P. Bartók, N. Bernstein, D. M. Wilkins, M. Ceriotti, and G. Csányi, *Gaussian Process Regression for Materials and Molecules*, Chem. Rev. **121**, 10073 (2021) (cit. on pp. 2, 15, 18, 20).

[36] A. P. Bartók, M. C. Payne, R. Kondor, and G. Csányi, *Gaussian Approximation Potentials: The Accuracy of Quantum Mechanics, without the Electrons*, Phys. Rev. Lett. **104**, 136403 (2010) (cit. on pp. 2, 18).

[37] S. Chmiela, A. Tkatchenko, H. E. Sauceda, I. Poltavsky, K. T. Schütt, and K.-R. Müller, *Machine learning of accurate energy-conserving molecular force fields*, Sci, Adv. **3**, e1603015 (2017) (cit. on p. 2).

[38] J. Behler and M. Parrinello, *Generalized Neural-Network Representation of High-Dimensional Potential-Energy Surfaces*, Phys. Rev. Lett. **98**, 146401 (2007) (cit. on pp. 2, 13, 21).

[39] J. S. Smith, O. Isayev, and A. E. Roitberg, *ANI-1: an extensible neural network potential with DFT accuracy at force field computational cost*, Chem. Sci. **8**, 3192 (2017) (cit. on pp. 2, 21).

[40] O. T. Unke and M. Meuwly, *PhysNet: A Neural Network for Predicting Energies, Forces, Dipole Moments, and Partial Charges*, J. Chem. Theory Comput. **15**, 3678 (2019) (cit. on pp. 2, 21).

[41] J. Gasteiger, F. Becker, and S. Günnemann, *GemNet: Universal Directional Graph Neural Networks for Molecules*, in NeurIPS (2021) (cit. on pp. 2, 21).

[42] K. T. Schütt, P.-J. Kindermans, H. E. Sauceda, S. Chmiela, A. Tkatchenko, and K.-R. Müller, *SchNet: A Continuous-Filter Convolutional Neural Network for Modeling Quantum Interactions*, in Adv. Neural. Inf. Process. Syst. (2017) (cit. on pp. 2, 21).

[43] Y. Zuo, C. Chen, X. Li, Z. Deng, Y. Chen, J. Behler, G. Csányi, A. V. Shapeev, A. P. Thompson, M. A. Wood, and S. P. Ong, *Performance and Cost Assessment of Machine Learning Interatomic Potentials*, J. Phys. Chem. A **124**, 731 (2020) (cit. on p. 2).

[44] A. P. Bartók, R. Kondor, and G. Csányi, *On representing chemical environments*, Phys. Rev. B **87**, 184115 (2013) (cit. on pp. 2, 14).

[45] F. Musil, A. Grisafi, A. P. Bartók, C. Ortner, G. Csányi, and M. Ceriotti, *Physics-Inspired Structural Representations for Molecules and Materials*, Chem. Rev. **121**, 9759 (2021) (cit. on pp. 2, 13, 14).

[46] A. Szabo and N. S. Ostlund, *Modern Quantum Chemistry: Introduction to Advanced Electronic Structure Theory*, First (Dover Publications, Inc., Mineola, 1996) (cit. on p. 3).

[47] P. Hohenberg and W. Kohn, *Inhomogeneous Electron Gas*, Phys. Rev. **136**, B864 (1964) (cit. on p. 4).

[48] W. Kohn and L. J. Sham, *Self-Consistent Equations Including Exchange and Correlation Effects*, Phys. Rev. **140**, A1133 (1965) (cit. on p. 4).

[49] A. D. Becke, *Perspective: Fifty years of density-functional theory in chemical physics*, J. Chem. Phys. **140**, 18A301 (2014) (cit. on p. 6).

[50] J. P. Perdew, K. Burke, and M. Ernzerhof, *Generalized Gradient Approximation Made Simple*, Phys. Rev. Lett. **77**, 3865 (1996) (cit. on p. 6).

[51] C. Adamo and V. Barone, *Toward reliable density functional methods without adjustable parameters: The PBE0 model*, J. Chem. Phys. **110**, 6158 (1999) (cit. on p. 6).

[52] K. Berland, V. R. Cooper, K. Lee, E. Schröder, T. Thonhauser, P. Hyldgaard, and B. I. Lundqvist, *van der Waals forces in density functional theory: a review of the vdW-DF method*, Rep. Prog. Phys. **78**, 066501 (2015) (cit. on p. 7).

[53] J. Hermann, R. A. DiStasio, and A. Tkatchenko, *First-Principles Models for van der Waals Interactions in Molecules and Materials: Concepts, Theory, and Applications*, Chem. Rev. **117**, 4714 (2017) (cit. on p. 7).

[54] F. London, *Zur Theorie und Systematik der Molekularkräfte*, Z. Phys. **63**, 245 (1930) (cit. on p. 7).

[55] R. Eisenschitz and F. London, *Über das Verhältnis der van der Waalsschen Kräfte zu den homöopolaren Bindungskräften*, Z. Phys. **60**, 491 (1930) (cit. on p. 7).

[56] F. London, *The general theory of molecular forces*, Trans. Faraday Soc. **33**, 8b (1937) (cit. on p. 7).

[57] H. B. G. Casimir and D. Polder, *The Influence of Retardation on the London-van der Waals Forces*, Phys. Rev. **73**, 360 (1948) (cit. on p. 7).

[58] E. Caldeweyher, S. Ehlert, A. Hansen, H. Neugebauer, S. Spicher, C. Bannwarth, and S. Grimme, *A generally applicable atomic-charge dependent London dispersion correction*, J. Chem. Phys. **150**, 154122 (2019) (cit. on p. 7).

[59] E. Caldeweyher, J.-M. Mewes, S. Ehlert, and S. Grimme, *Extension and evaluation of the D4 London-dispersion model for periodic systems*, Phys. Chem. Chem. Phys. **22**, 8499 (2020) (cit. on p. 7).

[60] S. J. A. van Gisbergen, J. G. Snijders, and E. J. Baerends, *A density functional theory study of frequency-dependent polarizabilities and Van der Waals dispersion coefficients for polyatomic molecules*, J. Chem. Phys. **103**, 9347 (1995) (cit. on p. 7).

[61] S. A. Ghasemi, A. Hofstetter, S. Saha, and S. Goedecker, *Interatomic potentials for ionic systems with density functional accuracy based on charge densities obtained by a neural network*, Phys. Rev. B **92**, 045131 (2015) (cit. on p. 7).

[62] A. Tkatchenko and M. Scheffler, *Accurate Molecular Van Der Waals Interactions from Ground-State Electron Density and Free-Atom Reference Data*, Phys. Rev. Lett. **102**, 073005 (2009) (cit. on p. 7).

[63] F. L. Hirshfeld, *Bonded-atom fragments for describing molecular charge densities*, Theoret. Chim. Acta **44**, 129 (1977) (cit. on p. 7).

[64] A. Tkatchenko, R. A. DiStasio, R. Car, and M. Scheffler, *Accurate and Efficient Method for Many-Body van der Waals Interactions*, Phys. Rev. Lett. **108**, 236402 (2012) (cit. on p. 8).

[65] B. Felderhof, *On the propagation and scattering of light in fluids*, Physica **76**, 486 (1974) (cit. on p. 8).

[66] D. W. Oxtoby and W. M. Gelbart, *Collisional polarizability anisotropies of the noble gases*, Mol. Phys. **29**, 1569 (1975) (cit. on p. 8).

[67] B. Thole, *Molecular polarizabilities calculated with a modified dipole interaction*, Chem. Phys. **59**, 341 (1981) (cit. on p. 8).

[68] A. G. Donchev, *Many-body effects of dispersion interaction*, J. Chem. Phys. **125**, 074713 (2006) (cit. on p. 8).

[69] M. W. Cole, D. Velegol, H.-Y. Kim, and A. A. Lucas, *Nanoscale van der Waals interactions*, Mol. Simul. **35**, 849 (2009) (cit. on p. 8).

[70] M. Kolb and W. Thiel, *Beyond the MNDO model: Methodical considerations and numerical results*, J. Comput. Chem. **14**, 775 (1993) (cit. on p. 8).

[71] M. P. Repasky, J. Chandrasekhar, and W. L. Jorgensen, *PDDG/PM3 and PDDG/MNDO: Improved semiempirical methods*, J. Comput. Chem. **23**, 1601 (2002) (cit. on p. 8).

[72] J. J. P. Stewart, *Optimization of parameters for semiempirical methods V: Modification of NDDO approximations and application to 70 elements*, J. Mol. Model. **13**, 1173 (2007) (cit. on p. 8).

[73] D. Porezag, T. Frauenheim, T. Köhler, G. Seifert, and R. Kaschner, *Construction of tight-binding-like potentials on the basis of density-functional theory: Application to carbon*, Phys. Rev. B **51**, 12947 (1995) (cit. on pp. 8–10).

[74] M. Elstner, D. Porezag, G. Jungnickel, J. Elsner, M. Haugk, T. Frauenheim, S. Suhai, and G. Seifert, *Self-consistent-charge density-functional tight-binding method for simulations of complex materials properties*, Phys. Rev. B **58**, 7260 (1998) (cit. on p. 8).

[75] M. Gaus, Q. Cui, and M. Elstner, *DFTB3: Extension of the Self-Consistent-Charge Density-Functional Tight-Binding Method (SCC-DFTB)*, J. Chem. Theory Comput. **7**, 931 (2011) (cit. on pp. 8, 9).

[76] G. Seifert, *Tight-Binding Density Functional Theory: An Approximate KohnSham DFT Scheme*, J. Phys. Chem. A **111**, 5609 (2007) (cit. on p. 8).

[77] M. Elstner and G. Seifert, *Density functional tight binding*, Philos. Trans. R. Soc. A **372**, 20120483 (2014) (cit. on p. 8).

[78] B. Hourahine, B. Aradi, V. Blum, F. Bonafé, A. Buccheri, C. Camacho, C. Cevallos, M. Y. Deshaye, T. Dumitrică, A. Dominguez, S. Ehlert, M. Elstner, T. van der Heide, J. Hermann, S. Irle, J. J. Kranz, C. Köhler, T. Kowalczyk, T. Kubař, I. S. Lee, V. Lutsker, R. J. Maurer, S. K. Min, I. Mitchell, C. Negre, T. A. Niehaus, A. M. N. Niklasson, A. J. Page, A. Pecchia, G. Penazzi, M. P. Persson, J. Řezáč, C. G. Sánchez, M. Sternberg, M. Stöhr, F. Stuckenberg, A. Tkatchenko, V. W.-z. Yu, and T. Frauenheim, *DFTB+, a software package for efficient approximate density functional theory based atomistic simulations*, J. Chem. Phys. **152**, 124101 (2020) (cit. on pp. 8, 9, 11).

[79] G. Seifert and J.-O. Joswig, *Density-functional tight binding—an approximate density-functional theory method*, Wiley Interdiscip. Rev. Comput. Mol. Sci. **2**, 456 (2012) (cit. on p. 10).

[80] J. C. Slater and G. F. Koster, *Simplified LCAO Method for the Periodic Potential Problem*, Phys. Rev. **94**, 1498 (1954) (cit. on p. 10).

[81] R. S. Mulliken, *Electronic Population Analysis on LCAO–MO Molecular Wave Functions. I*, J. Chem. Phys. **23**, 1833 (1955) (cit. on p. 10).

[82] P. Koskinen and V. Mäkinen, *Density-functional tight-binding for beginners*, Comput. Mater. Sci. **47**, 237 (2009) (cit. on p. 10).

[83] M. Gaus, A. Goez, and M. Elstner, *Parametrization and Benchmark of DFTB3 for Organic Molecules*, J. Chem. Theory Comput. **9**, 338 (2013) (cit. on p. 11).

[84] Z. Bodrog, B. Aradi, and T. Frauenheim, *Automated Repulsive Parametrization for the DFTB Method*, J. Chem. Theory Comput. **7**, 2654 (2011) (cit. on p. 11).

[85] M. Gaus, C.-P. Chou, H. Witek, and M. Elstner, *Automatized Parametrization of SCC-DFTB Repulsive Potentials: Application to Hydrocarbons*, J. Phys. Chem. A **113**, 11866 (2009) (cit. on p. 11).

[86] M. Mortazavi, J. G. Brandenburg, R. J. Maurer, and A. Tkatchenko, *Structure and Stability of Molecular Crystals with Many-Body Dispersion-Inclusive Density Functional Tight Binding*, J. Phys. Chem. Lett. **9**, 399 (2018) (cit. on p. 11).

[87] K. T. Butler, D. W. Davies, H. Cartwright, O. Isayev, and A. Walsh, *Machine learning for molecular and materials science*, Nature **559**, 547 (2018) (cit. on p. 13).

[88] L. Himanen, M. O. Jäger, E. V. Morooka, F. Federici Canova, Y. S. Ranawat, D. Z. Gao, P. Rinke, and A. S. Foster, *DScribe: Library of descriptors for machine learning in materials science*, Comput. Phys. Commun. **247**, 106949 (2020) (cit. on p. 13).

[89] M. Rupp, *Machine learning for quantum mechanics in a nutshell*, Int. J. Quantum Chem. **115**, 1058 (2015) (cit. on pp. 15, 18).

[90] A. P. Bartók, M. J. Gillan, F. R. Manby, and G. Csányi, *Machine-learning approach for one- and two-body corrections to density functional theory: Applications to molecular and condensed water*, Phys. Rev. B **88**, 054104 (2013) (cit. on p. 16).

[91] R. Ramakrishnan, P. O. Dral, M. Rupp, and O. A. von Lilienfeld, *Big Data Meets Quantum Chemistry Approximations: The Δ-Machine Learning Approach*, J. Chem. Theory Comput. **11**, 2087 (2015) (cit. on p. 16).

[92] Y. Eldar, M. Lindenbaum, M. Porat, and Y. Zeevi, *The farthest point strategy for progressive image sampling*, IEEE Trans. Image Process. **6**, 1305 (1997) (cit. on p. 16).

[93] M. Ceriotti, M. J. Willatt, and G. Csányi, *Machine Learning of Atomic-Scale Properties Based on Physical Principles*, in *Handbook of Materials Modeling : Methods: Theory and Modeling* (Springer International Publishing, Cham, 2018) (cit. on pp. 16, 20).

[94] A. N. Tikhonov, A. Goncharsky, V. V. Stepanov, and A. G. Yagola, *Numerical Methods for the Solution of Ill-Posed Problems* (Kluwer Academic, Dordrecht, 1995) (cit. on p. 17).

[95] G. Wahba, *Spline Models for Observational Data* (Society for Industrial and Applied Mathematics, 1990) (cit. on p. 18).

[96] B. Schölkopf, R. Herbrich, and A. J. Smola, *A Generalized Representer Theorem*, in Computational Learning Theory (2001) (cit. on p. 18).

[97] C. Rasmussen and C. Williams, *Gaussian Processes for Machine Learning* (MIT Press, 2005) (cit. on pp. 18–20).

[98] A. P. Bartók and G. Csányi, *Gaussian approximation potentials: A brief tutorial introduction*, Int. J. Quantum Chem. **115**, 1051 (2015) (cit. on pp. 18, 20).

[99] T. Bayes and n. Price, *LII. An essay towards solving a problem in the doctrine of chances. By the late Rev. Mr. Bayes, F. R. S. communicated by Mr. Price, in a letter to John Canton, A. M. F. R. S*, Philos. Trans. R. Soc. **53**, 370 (1763) (cit. on p. 20).

[100] K. P. F.R.S., *LIII. On lines and planes of closest fit to systems of points in space*, The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science **2**, 559 (1901) (cit. on p. 22).

[101] B. Schölkopf, A. Smola, and K.-R. Müller, *Nonlinear Component Analysis as a Kernel Eigenvalue Problem*, Neural Comput. **10**, 1299 (1998) (cit. on p. 22).

[102] L. van der Maaten and G. Hinton, *Visualizing Data using t-SNE*, J. Mach. Learn. Res. **9**, 2579 (2008) (cit. on p. 22).

[103] M. Ceriotti, G. A. Tribello, and M. Parrinello, *Simplifying the representation of complex free-energy landscapes using sketch-map*, Proc. Natl. Acad. Sci. U.S.A. **108**, 13023 (2011) (cit. on p. 22).

[104] L. McInnes, J. Healy, and J. Melville, *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*, ArXiv e-prints 1802.03426 (2018) (cit. on p. 22).

[105] V. R. Thalladi, M. Nüsse, and R. Boese, *The Melting Point Alternation in a,ω-Alkanedicarboxylic Acids*, J. Am. Chem. Soc. **122**, 9227 (2000) (cit. on p. 23).

[106] J. L. Derissen and P. H. Smith, *Refinement of the crystal structures of anhydrous α- and β-oxalic acids*, Acta. Crystallogr. B. Struct. **30**, 2240 (1974) (cit. on p. 23).

Appendices

Paper # 1

Data-efficient machine learning for molecular crystal structure prediction
Simon Wengert, Gábor Csányi, Karsten Reuter and Johannes T. Margraf

Chem. Sci. **12**, 4536 (2021)
DOI: https://doi.org/10.1039/D0SC05765G

Chemical Science

ROYAL SOCIETY OF CHEMISTRY

EDGE ARTICLE

View Article Online
View Journal | View Issue

Check for updates

Cite this: *Chem. Sci.*, 2021, **12**, 4536

All publication charges for this article have been paid for by the Royal Society of Chemistry

Received 19th October 2020
Accepted 5th February 2021

DOI: 10.1039/d0sc05765g

rsc.li/chemical-science

Data-efficient machine learning for molecular crystal structure prediction†

Simon Wengert,[a] Gábor Csányi, [b] Karsten Reuter[ac] and Johannes T. Margraf *[ac]

The combination of modern machine learning (ML) approaches with high-quality data from quantum mechanical (QM) calculations can yield models with an unrivalled accuracy/cost ratio. However, such methods are ultimately limited by the computational effort required to produce the reference data. In particular, reference calculations for periodic systems with many atoms can become prohibitively expensive for higher levels of theory. This trade-off is critical in the context of organic crystal structure prediction (CSP). Here, a data-efficient ML approach would be highly desirable, since screening a huge space of possible polymorphs in a narrow energy range requires the assessment of a large number of trial structures with high accuracy. In this contribution, we present tailored Δ-ML models that allow screening a wide range of crystal candidates while adequately describing the subtle interplay between intermolecular interactions such as H-bonding and many-body dispersion effects. This is achieved by enhancing a physics-based description of long-range interactions at the density functional tight binding (DFTB) level—for which an efficient implementation is available—with a short-range ML model trained on high-quality first-principles reference data. The presented workflow is broadly applicable to different molecular materials, without the need for a single periodic calculation at the reference level of theory. We show that this even allows the use of wavefunction methods in CSP.

1 Introduction

The capability to reliably predict the structure of molecular crystals is considered one of the holy grails of molecular modeling.[1,2] Applications for such crystal structure prediction (CSP) methods range from finding new drugs with improved dissolution properties (and thus bioavailability) to organic semiconductors with novel optoelectronic properties.[3,4] CSP for these molecular materials is so elusive because both their properties and stabilities are critically determined by the interactions of their molecular building blocks in the condensed phase. Indeed, the competition of different interaction types (*e.g.* dispersion and hydrogen bonding) within molecular crystals often leads to the coexistence of multiple similarly stable crystal structures—so-called polymorphs—each exhibiting different physical properties.[5,6] The ability to predict these polymorphs from simulations would therefore allow the efficient exploitation of the great technological potential inherent in this structural diversity, but requires an unparalleled CSP accuracy/efficiency ratio to explore the vast configuration spaces with highest energetic precision.

In practice, this search requires the reliable assessment of the relative stability of different structures, as measured by the lattice energy:

$$E_{\text{latt}} = E_{\text{crys}}/N - E_{\text{iso}}, \tag{1}$$

where E_{crys} is the total energy of the crystal per unit cell, N is the number of molecules in the unit cell and E_{iso} is the energy of an isolated molecule in its most stable conformation. Here, the main challenge lies in the large number of possible polymorphs and the small energy differences between them.[5,7,8] In practice, there is thus a trade-off between the ability to screen a wide range of candidates (which requires a fast evaluation of free energy or other stability measures) and applying higher levels of theory that adequately describe the subtle interplay between different intermolecular interactions such as H-bonding, electrostatic, induction and dispersion effects. Many CSP approaches are therefore structured hierarchically using a computationally less demanding stability assessment for screening a large set of candidates, while more advanced methods (typically based on density-functional theory, DFT) are used for the final ranking of the most promising structures.[9,10]

In recent years, a range of methods have been developed for the approximate stability assessment in the initial screening step. Li *et al.*[11] for instance evaluate stabilities of trial

[a] *Chair of Theoretical Chemistry, Technische Universität München, 85747 Garching, Germany*

[b] *Engineering Laboratory, University of Cambridge, Cambridge CB2 1PZ, UK*

[c] *Fritz-Haber-Institut der Max-Planck-Gesellschaft, Faradayweg 4-6, 14195 Berlin, Germany. E-mail: margraf@fhi.mpg.de*

† Electronic supplementary information (ESI) available. See DOI: 10.1039/d0sc05765g

configurations by applying the Harris approximation to DFT, with crystal electron densities constructed from the superposition of frozen single molecule densities. Tailor-made empirical potentials have also been successfully used for the screening step, as demonstrated for instance by Neumann *et al.*[12] in the blind tests of organic crystal structure prediction organized by the Cambridge Crystallographic Data Center[9] (CCDC). Finally, semiempirical electronic structure methods like density-functional tight-binding (DFTB) have also emerged as promising tools to efficiently rank the stabilities of molecular crystal structures.[13,14] Note that the initial screening can itself be hierarchical, so that the overall CSP workflow often resembles a funnel of increasingly narrow and accurate selection schemes. Nevertheless, regardless of how the most promising candidates are selected, the final step of a hierarchical CSP workflow requires an accurate first-principles method that allows resolving the subtle stability differences between competing polymorphs, presently typically semi-local or hybrid DFT with a many-body dispersion correction (DFT+MBD).[10]

There are essentially two sources of error in such hierarchical CSP schemes. First, the initial screening may either not consider the true lowest-energy structure in the first place or discard it erroneously. Second, the high-level method in the final layer may not produce the correct ranking of the remaining candidates. Unfortunately, the obvious solutions to these issues preclude each other: on the one hand, the selection issue can be mitigated by starting with a larger set of candidates and less severe filtering. On the other hand, better ranking can be achieved with more elaborate methods, at a higher computational cost per evaluation. For a fixed computational budget one cannot do both of these things. What is worse, in general it is not clear at the outset which of the two is more critical.

A potential way out of this conundrum is offered by modern machine-learning (ML) techniques, which have been found to combine the accuracy required in many chemical applications with affordable computational costs (most of which is associated with the generation of training data rather than the actual application of the potential).[15–17] In particular, much progress has recently been made in the development of ML-models for high-dimensional potential energy surfaces such as Neural Network Potentials (NNPs) *via* the Generalized Neural-Network Representation of Behler and Parrinello[18] or the Gaussian Approximation Potentials (GAP) framework developed by Bartók *et al.*[19] A more comprehensive overview of ML techniques for the generation of interatomic potentials can be found elsewhere.[20–22]

The high flexibility of ML models—which can be considered the reason of their success—can also lead to unphysical results, however, if the model is forced to extrapolate beyond its training set. Consequently, robust and accurate ML potentials are often trained on tens of thousands of configurations, for which accurate reference data is required.[23] Fortunately, interatomic potentials need not necessarily be created from scratch. Instead, ML models have also been used to improve the description of an underlying baseline.[24,25] Ramakrishnan *et al.*[26] coined the expression Δ-ML for this approach and showed that one needs significantly fewer training examples in this case, compared to learning a complete interatomic potential. In the context of CSP, there is a further strong argument for Δ-ML: most ML potentials are inherently local, meaning that the energy is composed of atomic contributions that only depend on the immediate environment of each atom. Yet, intermolecular interactions like electrostatics and (many-body) dispersion can be quite long ranged. A local ML potential will neglect those contributions, whereas a Δ-ML approach can incorporate them in the baseline model without altering the ML framework.

In this paper we therefore develop a Δ-ML approach to CSP, yielding accurate models for the description of individual molecules and the corresponding molecular crystals. The approach is characterized by high data efficiency, meaning that the workflow is designed to keep the computational effort for training data generation as low as possible. This is achieved by using a robust and computationally efficient baseline method, a diversity-driven selection of training points and the complete avoidance of periodic calculations at the target level of theory (here full-potential DFT with a many-body dispersion correction or spin-component-scaled second order perturbation theory).

2 Theory

2.1 Levels of theory

Baseline method. We begin by defining an appropriate baseline method for our approach. Most importantly, this method should be computationally efficient (to allow application to a large set of test structures) and adequately describe the relevant intra- and intermolecular interactions (so as to minimize the required Δ-ML correction). In particular, it should provide a reasonable description of long-range interactions that are outside the range of the ML model. In our experience dispersion-corrected DFTB methods, in particular using the 3ob parameterization,[27] fulfill these criteria.

3ob is based on the expansion of the DFT total energy up to third-order in density-fluctuations (DFTB3), which provides a sophisticated description of electrostatics, charge transfer and polarization.[28] This leads to marked improvements in the description of organic and biomolecular systems and hydrogen bonding, compared to earlier variants. Since DFTB uses a minimal basis set and tabulated matrix elements, it provides speedups up to three orders of magnitude compared with semi-local DFT. We further apply the Tkatchenko–Scheffler (TS) correction,[14,29] which allows for an accurate incorporation of dispersion interactions at virtually no additional computational cost. Our baseline method is thus defined as DFTB3(3ob)+TS (called DFTB+TS in the following).

Target method. The primary high-level target method in this study will be semi-local DFT (using the PBE functional[30]) with a many-body dispersion correction.[31,32] This method (DFT+MBD in the following) is known to generate lattice energies in good agreement with experiment for the targeted molecular crystals. This can, *e.g.*, be seen by its excellent performance for the X23 database, which contains the experimental lattice energies of 23 crystals (obtained by back-correcting experimental enthalpies of sublimation).[33] Since X23 covers van der Waals (vdW)-bonded, hydrogen-bonded and mixed molecular crystals, this shows that DFT+MBD offers a balanced description of all interactions relevant for CSP. Moreover, relative stabilities of different polymorphs are also described well, as recently demonstrated by Shtukenberg *et al.*[34] for the rich polymorphism of coumarin. For comparison, the presented scheme is finally also applied to spin-component-scaled second-order Møller–Plesset theory (SCS-MP2) in one case.[35]

Δ-ML method. We now aim to learn a correction that fixes the shortcomings of our baseline method relative to the target method. This entails, among other things, multi-center contributions to the Hamiltonian, many-body dispersion effects and exchange–correlation contributions inadequately described by the two-center repulsive potential of DFTB.[36,37] To this end, we use Gaussian Process Regression *via* the Gaussian Approximation Potential (GAP) framework.[19] Kernel methods like GAP use a similarity measure between atomic configurations (the kernel) to infer the interatomic potential. Here, we use the smooth overlap of atomic positions (SOAP),[38] which is an inherently many-body representation of atomic environments, in line with the types of contributions we want to describe. As noted above, SOAP and related methods use a local representation, meaning that in the final Δ-ML model, all long-range physics are still described at the baseline level of theory. Full details about the fitting procedure are provided in the ESI.†

With the above definitions of the target (DFT+MBD) and baseline (DFTB+TS) methods and the Δ-ML approach (GAP) used to connect the two, the lattice energy as measure of crystal stability is written as:

$$E_{\text{latt}}^{\text{target}} \approx E_{\text{latt}}^{\text{baseline+GAP}} = E_{\text{crys}}^{\text{baseline}}/N - E_{\text{iso}}^{\text{baseline}} + \Delta E^{\text{GAP}} \tag{2}$$

where ΔE^{GAP} is the learned Δ-ML correction.

In the following, we further separate this Δ-ML contribution into intra- ($\Delta E^{\text{GAP(intra)}}$) and intermolecular ($\Delta E^{\text{GAP(inter)}}$) contributions. This has both theoretical and practical reasons. Firstly, the energetic contribution of, *e.g.*, stretching a covalent bond is orders of magnitude larger than the contribution of changing the distance between two molecules in a crystal by the same amount. Nonetheless, the intermolecular contributions are arguably much more important for CSP and final polymorph ranking, as evidenced by the wide application of CSP protocols with completely rigid molecules.[11,39,40] By fitting separate models, the intermolecular contributions are not overshadowed by the intramolecular ones. Secondly, data generation for an intramolecular correction is very cheap, as it only requires calculations on the gas-phase molecule. It is therefore practical to separate the two training processes.

Using this separation, we can rewrite eqn (2) as

$$E_{\text{latt}}^{\Delta\text{-ML}} = \left(E_{\text{crys}}^{\text{baseline}} + \Delta E_{\text{crys}}^{\text{GAP(inter)}} + \sum_{i}^{N} \Delta E_{\text{mol},i}^{\text{GAP(intra)}} \right) \Big/ N - \left(E_{\text{iso}}^{\text{baseline}} + \Delta E_{\text{iso}}^{\text{GAP(intra)}} \right) \tag{3}$$

where the sum runs over all molecules *i* in the unit cell, and only intramolecular corrections $\Delta E_{\text{iso}}^{\text{GAP(intra)}}$ appear, of course, for the isolated molecule.

2.2 Training data

The generation of training data is a crucial part of constructing any ML model. This data represents all knowledge about the target function that will be integrated into the fit. The required calculations at the target level of theory, however, typically also make this the most expensive part of any ML workflow. It is therefore essential to strike a balance between covering a wide range of configurations and requiring a manageable number of calculations.

To address this issue, we generate a large pool of trial configurations and subsequently select a maximally diverse subset using the farthest point sampling (FPS) method.[21,41] This entails the iterative selection of configurations so that each new datapoint is maximally dissimilar to the previously selected structures. In this context, the similarity between configurations is measured using the averaged SOAP kernel.[42]

Clearly, the most straightforward training data for the Δ-ML correction would be obtained from periodic calculations on the FPS crystals at the target level of theory (DFT+MBD in this case). However, these are precisely the kinds of expensive calculations that we would like to avoid by fitting a Δ-ML model. Furthermore, it would in principle be interesting to use even higher levels of theory (*e.g.* Coupled Cluster or Symmetry Adapted Perturbation Theory) as the target method, for which periodic calculations are either impossible or extremely demanding.

Fortunately, we found that it is possible to fit accurate Δ-ML models without using periodic calculations at the target level of theory at all. Specifically, we use crystal structures as templates to generate molecular clusters (called X-mers in the following), which reflect the diverse relative orientations of the molecules in a crystal, in addition to providing realistic monomer configurations (see Fig. 1).

The idea of using X-mer training data is reminiscent of a many-body expansion (MBE) of the lattice energy.[43] This is, however, notoriously difficult to converge for (polar) organic crystals and liquids, both in terms of length-scale and body-order.[44–46] For this reason, highly accurate MBE-based water models separate the description of long-range electrostatics from short-range interactions.[47] It is therefore highly beneficial to work in a Δ-ML framework herein, where long-range interactions are covered by the baseline method. Indeed, a ML correction for force-field lattice energies based solely on two-

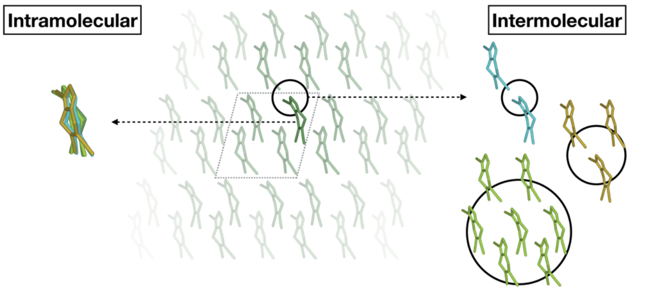


Fig. 1 Schematic separation of a crystal into monomers (entering the GAP(intra) learning workflow) and X-mers of various sizes (entering the GAP(inter) learning workflow).

body terms was recently reported by Day and coworkers.[48] In our work, we found that a pure two-body correction still displays significant errors in predicted lattice energies, and thus opted for the X-mer approach.

To this end, an initial pool of crystals is generated *via* the Genarris package.[11] Subsequently, we apply FPS to select 500 maximally diverse structures from this pool. These structures are then relaxed at the baseline level of theory, with fixed unit cells. Afterwards, a second FPS selection is performed on the relaxed crystals to obtain 250 training structures, while the rest are used for testing. Further details about training and test sets are given in the ESI.† Note that the training data for the intramolecular model is, *inter alia*, further supplemented with monomer configurations obtained from gas-phase MD simulations (see ESI† for details).

2.3 Model fitting

Using the above defined training data, we can now train separate GAP models for the intra- and intermolecular corrections. Specifically, we train the intramolecular correction on energy and force differences:

$$\begin{aligned}\Delta E^{\text{GAP(intra)}} &= E_{\text{mol}}^{\text{DFT+MBD}} - E_{\text{mol}}^{\text{DFTB+TS}}\\ \Delta F^{\text{GAP(intra)}} &= F_{\text{mol}}^{\text{DFT+MBD}} - F_{\text{mol}}^{\text{DFTB+TS}}\end{aligned} \tag{4}$$

The intermolecular correction is trained on differences in X-mer interaction energies:

$$\begin{aligned}\Delta E^{\text{GAP(inter)}} = &\left[E_{\text{X-mer}}^{\text{DFT+MBD}} - \sum_{i}^{X} E_{\text{mol},i}^{\text{DFT+MBD}}\right] - \Bigg[E_{\text{X-mer}}^{\text{DFTB+TS}}\\ &- \sum_{i}^{X} E_{\text{mol},i}^{\text{DFTB+TS}}\Bigg]\end{aligned} \tag{5}$$

The index i runs over all X molecules that constitute a cluster.

Details about the underlying concepts of SOAP and GAP are provided in the original literature.[19,38,49] A detailed listing of all hyperparameters and computational settings used in this work can be found in the ESI.†

3 Results and discussions

To illustrate the accuracy and efficiency of our Δ-ML approach, we will first separately discuss the accuracy reached for the intra- and intermolecular corrections, relative to their training targets. We then consider the accuracy of predicted lattice energies. For this we employ a representative set of four molecules and their molecular crystals, namely water (H_2O), pyrazine (C_4N_2), oxalic acid ($C_2O_4H_2$) and tetrolic acid ($C_4O_2H_4$).

3.1 Model performance: intramolecular Δ-ML

The accuracy of the intramolecular correction is assessed on monomer configurations extracted from the test and training crystals. Fig. 2 (top) shows the mean absolute error (MAE) of relative energies, compared to the high-level target method (DFT+MBD). For the DFTB+TS baseline, this MAE can be as high as 150 meV (for oxalic acid). This is a serious liability for CSP, where energy differences between polymorphs are often only tens of meV. In contrast, after the Δ-ML correction, the MAEs are reduced by orders of magnitude. Even in the most challenging case (oxalic acid) the corrected MAE is below 2 meV.

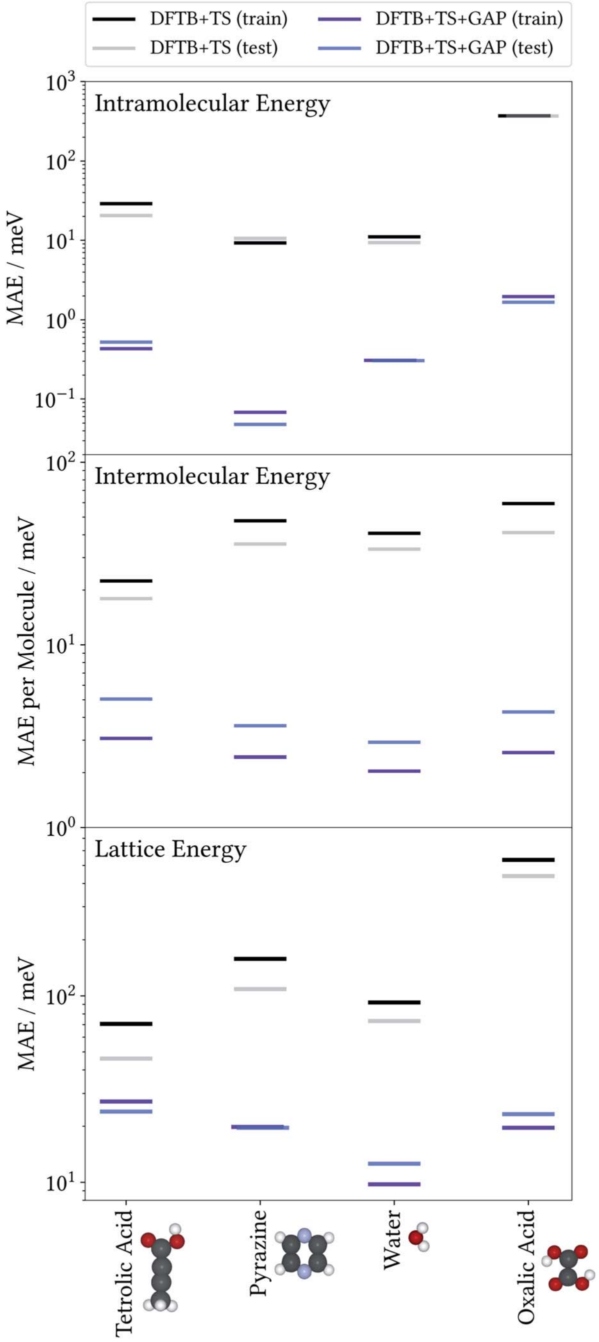


Fig. 2 Mean absolute error (MAE) of relative energies—with respect to the individual gas-phase global minimum—obtained with the baseline (DFTB+TS) and Δ-ML corrected (DFTB+TS+GAP) methods, against the DFT+MBD reference for monomer conformations from training and test crystals (top). Mean absolute error of intermolecular energies per molecule obtained with DFTB+TS and DFTB+TS+GAP against DFT+MBD for training and test X-mers (center). Mean absolute error for lattice energies of crystals entering the training and test crystals against the DFT+MBD reference (bottom). For details see text.

Moreover, the good agreement between training and test errors shows that the models are not overfitted. For the analysis on the accuracy of forces the reader is referred to the ESI.†

As a case in point, the excellent performance of the Δ-ML correction is confirmed when analyzing the seven predicted conformers of oxalic acid in detail. Indeed, conformer searches are themselves an integral part of molecular CSP studies, as gas-phase geometries are typically used as building blocks for the generation of trial crystals. Furthermore, the globally most stable gas-phase conformer is of special interest as the lattice energy is measured relative to it. Fig. 3 compiles the ranking of these seven conformers obtained at the different levels of theory, where we follow the nomenclature proposed in the literature[50] and refer to the conformers with a capital *C* (*cis*) or *T* (*trans*) depending on the relative orientation of the carboxylic acid groups, framed by lowercase c or t indicating whether the hydrogen atoms point to the inside or the outside. For the twisted conformer, where this nomenclature is not applicable, we use the symbol X.

For this highly sensitive test case, the Δ-ML method fully reproduces the energetic ordering of the target DFT+MBD method—which in turn is in agreement with the literature.[50,51] In contrast, the baseline DFTB+TS energies differ significantly and not even the lowest-energy conformer is correctly identified (reflected by the negative relative energy). In particular, DFTB+TS erroneously predicts most conformers to be rather close in energy, which could have severe consequences for an intended use as an initial screening method.

It is further revealing to consider the quality of the predicted geometries (see Fig. 3, bottom). For each conformer, the differences between geometries optimized with the low-cost methods (DFTB+TS or DFTB+TS+GAP) and the respective DFT+MBD reference is measured in terms of their root-mean-square deviation (RMSD). Similarly to the energies, the GAP-correction strongly improves the RMSD of all conformers—in most cases by more than an order of magnitude. At the same time it can be seen that DFTB+TS alone already provides quite accurate geometries in most cases. Here, the GAP correction cures only some subtle structural differences with respect to the DFT+MBD reference, as can be seen from the c*T*c overlay in Fig. 3, where the C–O–H angle in DFTB+TS is slightly too large. The exception to this is the t*C*t conformer. Here, DFTB+TS predicts a considerably different structure, which is brought into excellent agreement with the reference by the GAP-correction. This is again illustrated by the overlayed geometries, where DFT+MBD and DFTB+TS+GAP are almost indistinguishable.

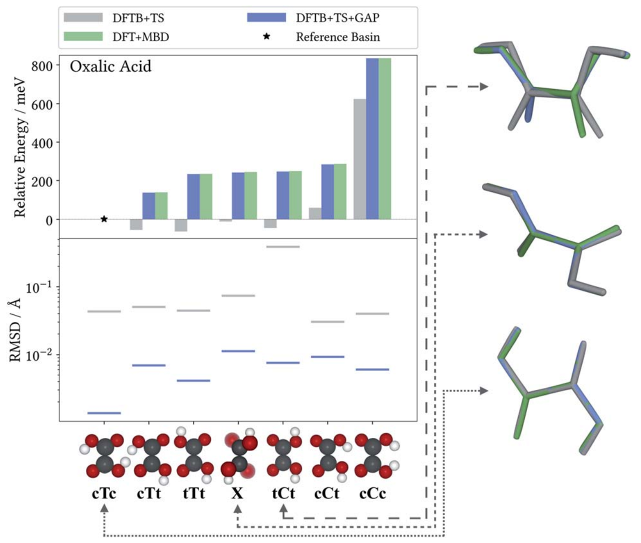


Fig. 3 Relative energies (top) and RMSDs (bottom) for oxalic acid gas-phase configurations relaxed with the DFTB+TS baseline, the DFTB+TS+GAP Δ-ML correction and the DFT+MBD reference. Structural overlays (right) for three conformers comparing the geometries predicted at the different levels of theory (see text).

3.2 Model performance: intermolecular Δ-ML

To evaluate the accuracy of the intermolecular Δ-ML contribution, we consider the intermolecular energies of X-mers, which are the training targets of this correction (see ESI† for a corresponding analysis of crystals). To this end, we consider X-mers of various sizes, again obtained from the training and test crystals. Fig. 2 (center) summarizes these results, in terms of the MAE, normalized by the number of molecules per X-mer. The Δ-ML method yields MAEs between 3 and 5 meV per molecule for test systems and slightly lower values for the training systems (1–2 meV per molecule). Again, the good agreement between test and training errors indicates that the proposed workflow yields Δ-ML models which generalize well beyond the training set.

Interestingly, tetrolic and oxalic acid show slightly larger MAEs, compared to pyrazine and water. We speculate that this is due to the higher flexibility of these molecules (see *e.g.* the oxalic acid conformers of Fig. 3), which causes a more diverse range of intermolecular arrangements. Overall, the GAP correction nevertheless improves the MAE per molecule by an order of magnitude (except for the tetrolic acid case, where the pure DFTB+TS description already yields a low MAE of around 20 meV per molecule).

3.3 Lattice energies

So far, we have analysed the accuracy of the intra- and intermolecular corrections on their respective training targets, and found large improvements relative to the baseline. However, the goal of the proposed method is to improve the description of crystal lattice energies. To evaluate this, we now benchmark the baseline and Δ-ML methods against the DFT+MBD target method for the lattice energies of molecular crystals. We again consider the crystals used to generate training and test sets separately. Note however, that even for the "training" crystals, the lattice energies were not used to fit the models. In this sense, all predictions in this section can be considered a validation of the Δ-ML model. Note that the lattice energies are referenced to the global gas-phase minimum of the molecule, calculated with the respective method. In the case of oxalic acid, the DFTB+TS lattice energies are therefore given with respect to a different gas-phase geometry.

The results are summarized in Fig. 2 (bottom). This figure shows that the improved description of intra- and intermolecular interactions also translates to an improved description of lattice energies, as expected. Specifically, the MAEs of the Δ-ML model lie between 12 and 24 meV per molecule, which in most cases corresponds to about an order of magnitude improvement. The exception is again tetrolic acid, which is already well described at the DFTB+TS level (but still improved by the GAP correction). These small MAEs also confirm our initial assumption, namely that the DFTB+TS baseline we employ adequately describes long-range interactions. This is further substantiated by considering the intermolecular contributions to the lattice energy separately, as shown in the ESI.†

From a CSP perspective, the lattice energies are arguably less important than the energetic ordering of the crystal structures, since we are more interested in which is the most stable crystal, rather than how stable it is in absolute terms. Fig. 4 therefore also includes the coefficients of determination (R^2) for the ranking order of the structures, which maps the correlation between reference and predicted data in a range between 0 (no correlation) and 1 (perfect correlation). Again, these are significantly improved by the GAP correction, with values between 0.967 and 0.995 indicating an excellent correlation between the energetic orderings of our Δ-ML model and the DFT+MBD target.

Importantly, errors for test crystals and the ones that (implicitly) enter the training are also in excellent agreement. This indicates a good generalization of the Δ-ML models beyond their training sets, also for the application to periodic systems. It is further notable that the MAEs for the baseline method are consistently larger for the training than the test set. This confirms that the workflow for training data selection leads to a set of particularly challenging and diverse systems. This can also be seen from the lattice energy correlation plots in Fig. 4, where the training structures cover the full range of lattice energies. In this context, it should be noted that the sampled range covers both negative and positive lattice energies. Although the focus of CSP is obviously on the systems with the most negative lattice energies, there are many trial crystals that need to be evaluated in the process. As these are not necessarily stable, creating a model that covers both ranges is actually desired, not least to be able to confidently discard unstable structures.

Fig. 4 provides more detailed insight into the performance of the baseline and Δ-ML models for the individual systems. As mentioned above, the baseline already provides a reasonable description of tetrolic acid. Nonetheless, there is significant scatter in the DFTB+TS correlation plot, which is also reflected in the energy ranking. Here, the GAP correction accounts for the subtle differences between baseline and target, leading to significant improvement.

In contrast, the lattice energy correlation plot for pyrazine displays a large systematic error, reflected in an erroneous slope (and consequently a large MAE). This deviation can be traced back to the fact that, for this system, unfavourable intermolecular interactions are less repulsive at the baseline level, compared to DFT+MBD (see ESI†). These systematic errors do not affect the ranking, however, which is in good agreement with DFT+MBD ($R^2 = 0.944$). The GAP correction is able to correct the systematic error in the lattice energies, leading to a strongly improved MAE. Importantly, however, the correction also further improves the energy ranking ($R^2 = 0.989$).

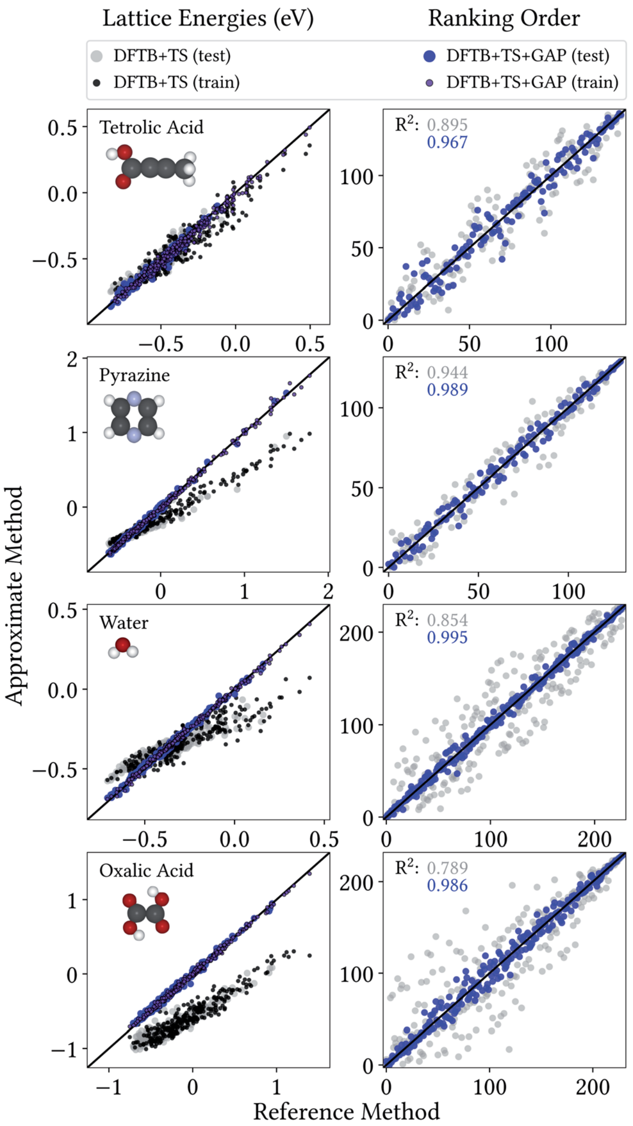


Fig. 4 Correlation plot for lattice energies of crystals entering the training and test crystals (left) and ranking order of test crystals (right), both with respect to DFT+MBD.

For water and oxalic acid, we observe both systematic errors and significant scatter in the predictions of the baseline method. Here, the GAP corrections need to account for a mixture of different effects simultaneously. The lattice energy correlation plots indicate different types of systematic deviations for these systems. While the slope for the water lattice energies is too small, oxalic acid additionally shows an offset of roughly 200 meV with respect to the DFT+MBD values. As with pyrazine, the erroneous slopes are explained by a systematic underestimation of repulsive intermolecular interactions (see ESI†). Meanwhile, the offset for oxalic acid is due to differences in intramolecular interactions at the baseline and target levels (compare Fig. 3). Here, the different predicted global minimum conformers result in a discrepancy of the intramolecular contributions to the lattice energy. As shown in Section 3.1 the GAP correction is very well suited to account for this situation. More generally, the GAP corrections lead to strongly improved lattice energies and ranking orders for both systems.

To quantify the error introduced by the X-mer approach, we further created an alternative set of Δ-ML models (see ESI†). Here, the intermolecular corrections were trained on FPS-selected crystals instead of the X-mers. Compared to the X-mer approach, these models display slightly improved lattice energies for most cases (by 4–6 meV per molecule) and are slightly worse in one case. The error incurred by the X-mer approach is thus small or non-existent for the systems considered herein.

3.4 Crystal structure prediction

To allow for a pointwise comparison of interaction potentials, the lattice energies in the previous section were computed *via* single point energy evaluations for frozen geometries (relaxed at the baseline level). Indeed, this strategy has also been employed in 'real' CSP applications.[14] However, the results in Section 3.1 show that the DFTB+TS baseline used herein can yield significantly erroneous geometries. This is an uncontrolled source of error, which will propagate through the entire CSP workflow. Fortunately, GAP models are differentiable, so that geometry relaxations at the Δ-ML corrected DFTB+TS+GAP level are also possible, at essentially no added cost. In this section, we will illustrate the benefit of this feature.

For this purpose, we consider target XXII of the most recent blind test of organic CSP.[9] It corresponds to the crystallized form of the tricyano-1,4-dithiino[*c*]-isothiazole (C_8N_4S) molecule. Notably, the six-membered ring in this molecule can be hinged, which induces a chiral-like character to the molecule and, thus, affects the number of space groups allowed in the solid state.

A Δ-ML model for target XXII was generated following the method detailed in Section 2. All results discussed in the following are for randomly generated trial crystal structures not included in the training process. Additionally, the known experimental crystal structure of the molecule is included,[52] to test whether it would have been correctly identified. Unlike in the previous section, all trial structures are relaxed at the baseline DFTB+TS and Δ-ML corrected DFTB+TS+GAP levels of theory, and validated with single point calculations at the target DFT+MBD level (see ESI† for an analysis as in Section 3.3). Fig. 5 shows the corresponding lattice energy correlation plot, as well as the ranking order.

The most striking feature of the lattice energy plot is a large offset between the baseline and target predictions. Similar to the oxalic acid case, this is—at least partly—explained by deviations in the intramolecular descriptions. DFT+MBD favours the two symmetry-equivalent conformations that exhibit a kink in the six-membered ring. Fig. 6 shows the DFT+MBD minimum energy path for the interconversion of these structures, obtained from a nudged elastic band (NEB) calculation. Here, the flat conformation of the molecule is found to be a saddle point, in agreement with previous reports.[9]

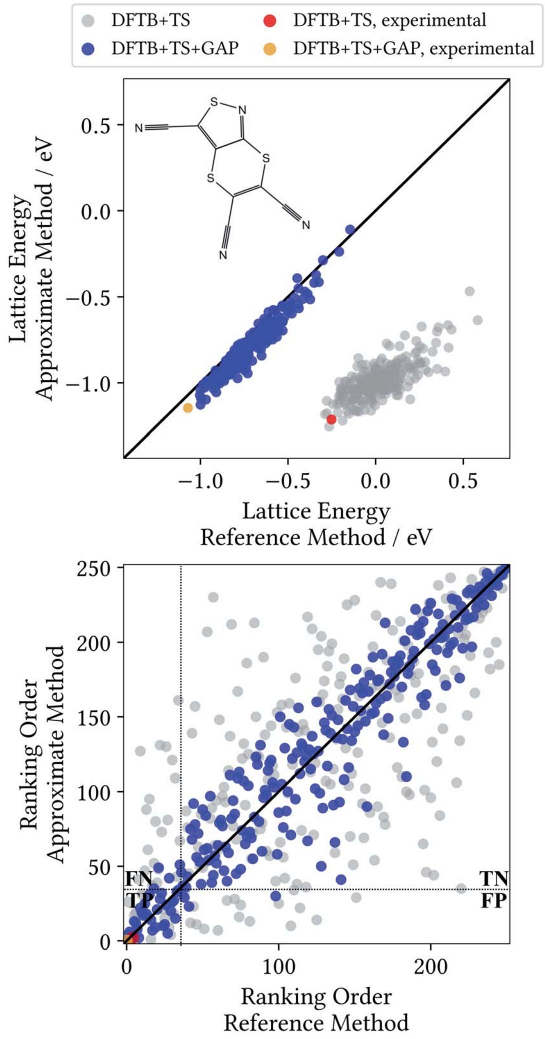


Fig. 5 Correlation plot for lattice energies of XXII crystals relaxed with the DFTB+TS baseline and the Δ-ML corrected DFTB+TS+GAP method against the respective DFT+MBD target level values (top) and corresponding ranking order (bottom) with separation into the four parts True-Positive (TP), True-Negative (TN), False-Positive (FP) and False-Negative (FN) – see text.

This profile changes dramatically when the minimum energy path is reevaluated with the baseline DFTB+TS method: the barrier turns into a broad valley. In fact, the gas-phase optimum found with DFTB+TS corresponds to the flat conformer, as can be seen from the overlay on the right-hand of Fig. 6. In combination with additional geometric deviations (*e.g.* a more acute C–S–N angle of the five-membered ring), this causes an energy difference of 670 meV between the gas-phase minima of the baseline and target methods (when evaluated at the DFT+MBD level). As can be seen in Fig. 5 and 6, the Δ-ML correction cures these discrepancies and largely eliminates the offset. More importantly, the correction also strongly improves the correlation in the energy ranking and correctly identifies the experimental structure to be the most stable.

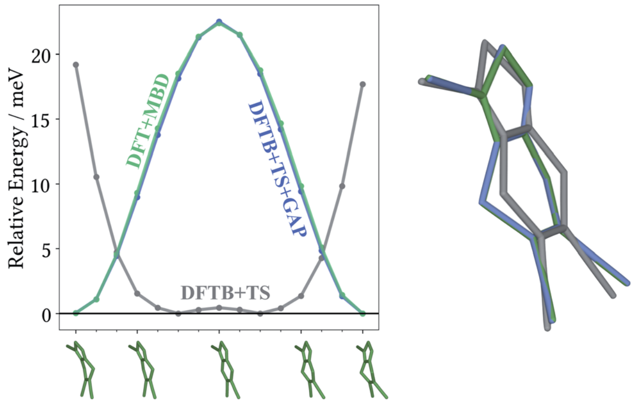


Fig. 6 Target DFT+MBD climbing image nudged elastic band results with baseline DFTB+TS and Δ-ML corrected DFTB+TS+GAP single-point evaluation (left). Energies are relative to the individual image with the lowest energy. Overlay of the gas-phase minimum geometries (right) obtained with DFT+MBD (green), DFTB+TS (gray) and DFTB+TS+GAP (blue).

In the CSP context, the most pertinent comparison of the two methods is provided by the ranking order plot in Fig. 5. Here, the baseline method displays a large scatter, with some structures that are deemed among the most stable by DFT+MBD being assigned high ranks (and *vice versa*). This results in a low coefficient of determination of 0.483. In contrast, the energetic ordering predicted by the Δ-ML model correlates very well with the DFT+MBD reference ($R^2 = 0.907$). This good agreement makes DFTB+TS+GAP a very promising method for CSP, particularly as a pre-screening method in hierarchical schemes. In this context, the most stable structures from the pre-screening would be further investigated with highly accurate (and expensive) methods, *e.g.* including vibrational contributions to the lattice free energy at the DFT+MBD level.

To illustrate the benefits of the GAP correction for this purpose, the ranking plot in Fig. 5 is divided in the style of a confusion matrix for the selection of the 35 most stable candidates. The resulting sectors indicate the true positive (TP), true negative (TN), false positive (FP) and false negative (FN) predictions. The quality of the selections made with the baseline and Δ-ML methods can now by visualized by the populations of the four sectors. DFTB+TS+GAP populates the most important sector, TP, with 32 out of 35 crystals. Uncorrected DFTB+TS, on the other hand, only yields 16 samples in this block. Furthermore, out of the three false positive predictions of DFTB+TS+GAP, two are very close to the dividing line.

As mentioned above, the experimentally determined crystal structure is indeed found to be the most stable structure at the Δ-ML level. Furthermore, the corresponding Δ-ML geometry is also found to be the most stable at the DFT+MBD level. In contrast, the baseline method predicts several other structures to be more stable than the experimental one. Critically, the experimental structure is not even the lowest energy one when DFT+MBD single point calculations are performed on DFTB+TS geometries. This is again due to significant deviations in the predicted geometries of DFTB+TS. Meanwhile there is excellent agreement between the predicted DFTB+TS+GAP crystal structure, and the one relaxed at the DFT+MBD level (see ESI†).

Finally, we return to the question of computational efficiency. As stated above, the main motivation for the presented Δ-ML approach is to avoid the large computational effort of calculations at the target level of theory. Most importantly, the savings of the Δ-ML model at prediction time should significantly outweigh the cost of generating the training data. To this end, the computational effort for generating the Δ-ML models and performing 10 000 crystal relaxations (a reasonable number for a CSP application)[9] is shown in Fig. 7. It can be seen that the cost of the training procedure is almost exclusively determined by reference calculations at the target level of theory (in particular for the X-mers).

For comparison, a Δ-ML model that exclusively uses the underlying crystals instead of X-mers requires *ca.* 5000 CPU hours for performing DFT+MBD reference calculations. At this level of theory, the cost for training with periodic crystal data is thus actually somewhat lower than with the X-mer approach. Note, however, that the accuracy of this model is actually slightly inferior to the X-mer approach (see ESI†). Furthermore, the growth in computational costs when including more training data will be steep, especially when considering higher reference levels of theory, as shown below.

Fig. 7 further shows that (once trained), the savings of the Δ-ML model at prediction time are substantial: 10 000 crystal relaxations at the target level of theory would require a staggering 30 million CPU hours, compared to just 80 000 CPU hours with the Δ-ML model. Furthermore, the costs for training

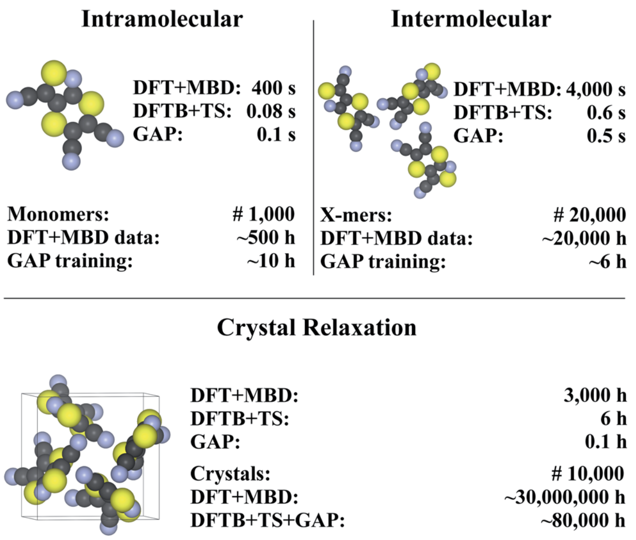


Fig. 7 Timings for generating the (intra- and intermolecular) model for XXII and crystal relaxations (as obtained on a Intel® Xeon® CPU E5-2697 v3 @ 2.60 GHz processor). The upper part in each of the tree sections illustrates timings for a single unit (monomer, X-mer or crystal), while the lower part corresponds to the time required for the specified number of training configurations (top) and an exemplary number of crystal relaxations (bottom). The costs for relaxations are included in the intramolecular timing (see text). Values are rounded to one significant digit (both in terms of time and number of geometries). For details see ESI.†

the Δ-ML models are roughly equivalent to the cost of explicitly relaxing just seven crystals at the DFT+MBD target level—an insignificant number compared to the requirements of a full-blown CSP study.

3.5 Crystal structure prediction beyond density functional theory

Dispersion corrected semi-local DFT is known to be quite accurate for noncovalent interactions, but it nevertheless displays some pathologies that can be problematic for CSP.[53] Most prominently, the self-interaction error in most functionals causes the over-delocalization of electrons, which leads to errors in the description of electrostatic potentials and charge transfer.[54]

In contrast, correlated wavefunction (WF) methods do not suffer from this problem. Furthermore, with these methods convergence to the exact result is, at least in principle, possible. Consequently, there has been much interest in applying WF theory to molecular crystals. This has been prohibitively expensive until recently, but new algorithms and hardware have made some benchmark calculations possible.[55–57] In this context, highly accurate (sub-kJ mol^{-1}) lattice energy predictions have been demonstrated, *e.g.* by Yang *et al.*[43] *via* a fragment strategy and by Zen *et al. via* diffusion quantum Monte Carlo.[58] While this highlights their potential for CSP, applying such methods to periodic systems is still far from routine and will not be feasible in a high-throughput context for the foreseeable future. The X-mer approach presented herein does not require periodic reference calculations, however, and thus opens the door to WF-based CSP.

To illustrate this, a modified version of the model from the previous section was developed, for which the intermolecular GAP was trained using spin-component-scaled second-order Møller–Plesset theory (SCS-MP2).[35,59] This highlights an additional feature of the presented approach, namely that different reference methods can be used for the intra- and intermolecular models. This can be particularly useful for flexible molecules, where an accurate prediction of torsional barriers, *e.g.* at the CCSD(T) level, may be required.[53]

To evaluate the new intermolecular model, the interaction energies for a test set of X-mers was considered. This reveals a MAE of 7 meV, slightly lower than the one obtained with the DFT+MBD reference (see ESI† for details). The corresponding full model was then used to relax the 251 trial crystals used in Section 3.4. While no periodic MP2 data is available for benchmarking in this case (for the reasons outlined above), the model correctly identifies the experimental geometry to be the most stable (see ESI† for details). The possibility of crystal relaxations with the ML model is particularly attractive in the context of WF methods, where gradients are much more expensive than single-point energy evaluations.[60]

As a final note, it should be mentioned that SCS-MP2 is not necessarily more accurate than DFT+MBD for this application. While the former offers a better description of electrostatics and Pauli repulsion (because the method is self-interaction free), the latter offers a true many-body description of dispersion, which is lacking at the (SCS-) MP2 level.[61] Nonetheless, this example demonstrates that the presented scheme can be used to apply correlated wavefunction methods in a CSP context. The computational costs to produce the SCS-MP2 X-mer training data lies at 190 000 CPU hours, while the direct application of SCS-MP2 for crystal relaxations in a molecular CSP study is simply not feasible.

4 Conclusions

In this work, we have presented a computationally efficient and accurate Δ-ML approach to CSP, using a low-cost baseline (DFTB+TS) that adequately describes long-range interactions. The method is characterized by addressing intra- and intermolecular corrections separately and features a high efficiency in terms of training costs. In particular, this is achieved by selecting diverse training configurations and completely avoiding periodic calculation for training data generation. The overall accuracy of lattice energies and relative stability rankings has been demonstrated on a representative set of test systems. Importantly, the approach yields models that allow for reliable structure relaxations, with a computational effort that is orders of magnitude smaller than the high-level target method (PBE+MBD or SCS-MP2), even taking training costs into account. To the best of our knowledge, this is the first generally applicable ML approach that allows structure relaxations in the context of CSP. This opens the door to a CSP workflow that allows screening large candidate pools with unprecedented accuracy.

We further note that the accuracy of the Δ-ML can, in principle, be further refined by including more data. Beyond this, the fact that no periodic calculations are required means that higher levels of theory, such as hybrid DFT or (correlated) wavefunction methods, can be used as the target method. Finally, having a differentiable model also allows the calculation of vibrational zero-point and free energy contributions to the crystal stability. This will be explored in future work.

5 Computational details

All DFT calculations were performed with FHI-aims,[62] using the PBE functional,[30] tier2 basis sets, tight integration grids and the MBD dispersion correction. DFTB3 calculations were performed using DFTB+[63] together with the 3ob parametrization[27] and TS dispersion correction.[14,29] For periodic calculations at both levels of theory, the *k*-grids were converged to obtain energetic accuracies of 1.5 meV per atom. SCS-MP2 (ref. 35 and 59) calculations were performed with ORCA[64,65] using the resolution of identity approximation.[66] GAP potentials were trained and evaluated with the QUIP package.[49] Candidate crystal structures were obtained with the Genarris package.[11] Additional tasks such as FPS and hyperparameter optimization were performed with the MLtools package available at https://github.com/simonwengert/mltools.git.

Conflicts of interest

G. C. is listed as an inventor on a patent filed by Cambridge Enterprise Ltd. related to SOAP and GAP (US patent 8843509, filed on 5 June 2009 and published on 23 September 2014).

Acknowledgements

S. W. and K. R. acknowledge funding from Deutsche Forschungsgemeinschaft (DFG) under project RE1509/18-2.

References

1 K. N. Houk and F. Liu, *Acc. Chem. Res.*, 2017, **50**, 539–543.
2 S. L. Price, *Chem. Soc. Rev.*, 2014, **43**, 2098–2111.
3 J. Hachmann, R. Olivares-Amaya, S. Atahan-Evrenk, C. Amador-Bedolla, R. S. Sánchez-Carrera, A. Gold-Parker, L. Vogt, A. M. Brockway and A. Aspuru-Guzik, *J. Phys. Chem. Lett.*, 2011, **2**, 2241–2251.
4 A. Pulido, L. Chen, T. Kaczorowski, D. Holden, M. A. Little, S. Y. Chong, B. J. Slater, D. P. McMahon, B. Bonillo, C. J. Stackhouse, A. Stephenson, C. M. Kane, R. Clowes, T. Hasell, A. I. Cooper and G. M. Day, *Nature*, 2017, **543**, 657–664.
5 A. J. Cruz-Cabeza, S. M. Reutzel-Edens and J. Bernstein, *Chem. Soc. Rev.*, 2015, **44**, 8619–8635.
6 T.-Q. Yu and M. E. Tuckerman, *Phys. Rev. Lett.*, 2011, **107**, 015701.
7 J. Nyman and G. M. Day, *CrystEngComm*, 2015, **17**, 5154–5165.
8 T. Beyer, T. Lewis and S. L. Price, *CrystEngComm*, 2001, **3**, 178–212.
9 A. M. Reilly, R. I. Cooper, C. S. Adjiman, S. Bhattacharya, A. D. Boese, J. G. Brandenburg, P. J. Bygrave, R. Bylsma, J. E. Campbell, R. Car, D. H. Case, R. Chadha, J. C. Cole, K. Cosburn, H. M. Cuppen, F. Curtis, G. M. Day, R. A. DiStasio Jr, A. Dzyabchenko, B. P. van Eijck, D. M. Elking, J. A. van den Ende, J. C. Facelli, M. B. Ferraro, L. Fusti-Molnar, C.-A. Gatsiou, T. S. Gee, R. de Gelder, L. M. Ghiringhelli, H. Goto, S. Grimme, R. Guo, D. W. M. Hofmann, J. Hoja, R. K. Hylton, L. Iuzzolino, W. Jankiewicz, D. T. de Jong, J. Kendrick, N. J. J. de Klerk, H.-Y. Ko, L. N. Kuleshova, X. Li, S. Lohani, F. J. J. Leusen, A. M. Lund, J. Lv, Y. Ma, N. Marom, A. E. Masunov, P. McCabe, D. P. McMahon, H. Meekes, M. P. Metz, A. J. Misquitta, S. Mohamed, B. Monserrat, R. J. Needs, M. A. Neumann, J. Nyman, S. Obata, H. Oberhofer, A. R. Oganov, A. M. Orendt, G. I. Pagola, C. C. Pantelides, C. J. Pickard, R. Podeszwa, L. S. Price, S. L. Price, A. Pulido, M. G. Read, K. Reuter, E. Schneider, C. Schober, G. P. Shields, P. Singh, I. J. Sugden, K. Szalewicz, C. R. Taylor, A. Tkatchenko, M. E. Tuckerman, F. Vacarro, M. Vasileiadis, A. Vazquez-Mayagoitia, L. Vogt, Y. Wang, R. E. Watson, G. A. de Wijs, J. Yang, Q. Zhu and C. R. Groom, *Acta Crystallogr., Sect. B: Struct. Sci., Cryst. Eng. Mater.*, 2016, **72**, 439–459.
10 J. Hoja, H.-Y. Ko, M. A. Neumann, R. Car, R. A. DiStasio and A. Tkatchenko, *Sci. Adv.*, 2019, **5**, eaau3338.
11 X. Li, F. S. Curtis, T. Rose, C. Schober, A. Vazquez-Mayagoitia, K. Reuter, H. Oberhofer and N. Marom, *J. Chem. Phys.*, 2018, **148**, 241701.
12 M. Neumann, F. Leusen and J. Kendrick, *Angew. Chem., Int. Ed.*, 2008, **47**, 2427–2430.
13 J. G. Brandenburg and S. Grimme, *J. Phys. Chem. Lett.*, 2014, **5**, 1785–1789.
14 M. Mortazavi, J. G. Brandenburg, R. J. Maurer and A. Tkatchenko, *J. Phys. Chem. Lett.*, 2018, **9**, 399–405.
15 V. L. Deringer and G. Csányi, *Phys. Rev. B*, 2017, **95**, 094203.
16 M. Rupp, *Int. J. Quantum Chem.*, 2015, **115**, 1058–1073.
17 J. S. Smith, B. T. Nebgen, R. Zubatyuk, N. Lubbers, C. Devereux, K. Barros, S. Tretiak, O. Isayev and A. E. Roitberg, *Nat. Commun.*, 2019, **10**, 2903.
18 J. Behler and M. Parrinello, *Phys. Rev. Lett.*, 2007, **98**, 146401.
19 A. P. Bartók, M. C. Payne, R. Kondor and G. Csányi, *Phys. Rev. Lett.*, 2010, **104**, 136403.
20 C. M. Handley and J. Behler, *Eur. Phys. J. B*, 2014, **87**, 152.
21 M. Ceriotti, M. J. Willatt and G. Csányi, *Handbook of Materials Modeling*, 2018, pp. 1–27.
22 P. O. Dral, *J. Phys. Chem. Lett.*, 2020, **11**, 2336–2347.
23 C. Schran, J. Behler and D. Marx, *J. Chem. Theory Comput.*, 2020, **16**, 88–99.
24 A. P. Bartók, M. J. Gillan, F. R. Manby and G. Csányi, *Phys. Rev. B*, 2013, **88**, 054104.
25 M. Stöhr, L. M. Sandonas and A. Tkatchenko, *J. Phys. Chem. Lett.*, 2020, **11**, 6835–6843.
26 R. Ramakrishnan, P. O. Dral, M. Rupp and O. A. von Lilienfeld, *J. Chem. Theory Comput.*, 2015, **11**, 2087–2096.
27 M. Gaus, A. Goez and M. Elstner, *J. Chem. Theory Comput.*, 2013, **9**, 338–354.
28 M. Gaus, Q. Cui and M. Elstner, *J. Chem. Theory Comput.*, 2011, **7**, 931–948.
29 M. Stöhr, G. S. Michelitsch, J. C. Tully, K. Reuter and R. J. Maurer, *J. Chem. Phys.*, 2016, **144**, 151101.
30 J. P. Perdew, K. Burke and M. Ernzerhof, *Phys. Rev. Lett.*, 1996, **77**, 3865–3868.
31 A. Tkatchenko, R. A. DiStasio, R. Car and M. Scheffler, *Phys. Rev. Lett.*, 2012, **108**, 236402.
32 A. Ambrosetti, A. M. Reilly, R. A. DiStasio and A. Tkatchenko, *J. Chem. Phys.*, 2014, **140**, 18A508.
33 A. M. Reilly and A. Tkatchenko, *J. Chem. Phys.*, 2013, **139**, 024705.
34 A. G. Shtukenberg, Q. Zhu, D. J. Carter, L. Vogt, J. Hoja, E. Schneider, H. Song, B. Pokroy, I. Polishchuk, A. Tkatchenko, A. R. Oganov, A. L. Rohl, M. E. Tuckerman and B. Kahr, *Chem. Sci.*, 2017, **8**, 4926–4940.
35 S. Grimme, *J. Chem. Phys.*, 2003, **118**, 9095–9102.
36 P. Koskinen and V. Mäkinen, *Comput. Mater. Sci.*, 2009, **47**, 237–253.
37 Y. Yang, H. Yu, D. York, Q. Cui and M. Elstner, *J. Phys. Chem. A*, 2007, **111**, 10861–10873.
38 A. P. Bartók, R. Kondor and G. Csányi, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 2013, **87**, 184115.

39 S. L. Price, M. Leslie, G. W. A. Welch, M. Habgood, L. S. Price, P. G. Karamertzanis and G. M. Day, *Phys. Chem. Chem. Phys.*, 2010, **12**, 8478–8490.

40 P. G. Karamertzanis and C. C. Pantelides, *J. Comput. Chem.*, 2005, **26**, 304–324.

41 A. P. Bartók, S. De, C. Poelking, N. Bernstein, J. R. Kermode, G. Csányi and M. Ceriotti, *Sci. Adv.*, 2017, **3**, e1701816.

42 S. De, A. P. Bartók, G. Csányi and M. Ceriotti, *Phys. Chem. Chem. Phys.*, 2016, **18**, 13754–13769.

43 J. Yang, W. Hu, D. Usvyat, D. Matthews, M. Schütz and G. K.-L. Chan, *Science*, 2014, **345**, 640–643.

44 C. Müller, D. Usvyat and H. Stoll, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 2011, **83**, 245136.

45 C. Müller and D. Usvyat, *J. Chem. Theory Comput.*, 2013, **9**, 5590–5598.

46 R. Podeszwa, B. M. Rice and K. Szalewicz, *Phys. Rev. Lett.*, 2008, **101**, 115503.

47 E. Lambros and F. Paesani, *J. Chem. Phys.*, 2020, **153**, 060901.

48 D. McDonagh, C. K. Skylaris and G. M. Day, *J. Chem. Theory Comput.*, 2019, **15**, 2743–2758.

49 A. P. Bartók and G. Csányi, *Int. J. Quantum Chem.*, 2015, **115**, 1051–1057.

50 A. Mohajeri and N. Shakerin, *J. Mol. Struct.: THEOCHEM*, 2004, **711**, 167–172.

51 J. Higgins, X. Zhou, R. Liu and T. T.-S. Huang, *J. Phys. Chem. A*, 1997, **101**, 2702–2708.

52 F. Curtis, X. Wang and N. Marom, *Acta Crystallogr., Sect. B: Struct. Sci., Cryst. Eng. Mater.*, 2016, **72**, 562–570.

53 J. Nyman, L. Yu and S. M. Reutzel-Edens, *CrystEngComm*, 2019, **21**, 2080–2088.

54 A. J. Cohen, P. Mori-Sanchez and W. Yang, *Science*, 2008, **321**, 792–794.

55 J. McClain, Q. Sun, G. K.-L. Chan and T. C. Berkelbach, *J. Chem. Theory Comput.*, 2017, **13**, 1209–1218.

56 A. Grüneis, M. Marsman and G. Kresse, *J. Chem. Phys.*, 2010, **133**, 074107.

57 I. Y. Zhang, A. J. Logsdail, X. Ren, S. V. Levchenko, L. Ghiringhelli and M. Scheffler, *New J. Phys.*, 2019, **21**, 013025.

58 A. Zen, J. G. Brandenburg, J. Klimeš, A. Tkatchenko, D. Alfè and A. Michaelides, *Proc. Natl. Acad. Sci. U. S. A.*, 2018, **115**, 1724–1729.

59 K. E. Riley, J. A. Platts, J. Řezáč, P. Hobza and J. G. Hill, *J. Phys. Chem. A*, 2012, **116**, 4159–4169.

60 E. A. Salter, G. W. Trucks and R. J. Bartlett, *J. Chem. Phys.*, 1989, **90**, 1752–1766.

61 P. Xu, M. Alkan and M. S. Gordon, *Chem. Rev.*, 2020, **120**, 12343–12356.

62 V. Blum, R. Gehrke, F. Hanke, P. Havu, V. Havu, X. Ren, K. Reuter and M. Scheffler, *Comput. Phys. Commun.*, 2009, **180**, 2175–2196.

63 B. Aradi, B. Hourahine and T. Frauenheim, *J. Phys. Chem. A*, 2007, **111**, 5678–5684.

64 F. Neese, *Wiley Interdiscip. Rev.: Comput. Mol. Sci.*, 2012, **2**, 73–78.

65 F. Neese, *Wiley Interdiscip. Rev.: Comput. Mol. Sci.*, 2018, **8**, e1327.

66 M. Feyereisen, G. Fitzgerald and A. Komornicki, *Chem. Phys. Lett.*, 1993, **208**, 359–363.

Paper # 2

A Hybrid Machine Learning Approach for Structure Stability Prediction in Molecular Co-crystal Screenings

Simon Wengert, Gábor Csányi, Karsten Reuter and Johannes T. Margraf

J. Chem. Theory Comput. **18**, 4586 (2022)
DOI: https://doi.org/10.1021/acs.jctc.2c00343

JCTC Journal of Chemical Theory and Computation

pubs.acs.org/JCTC Article

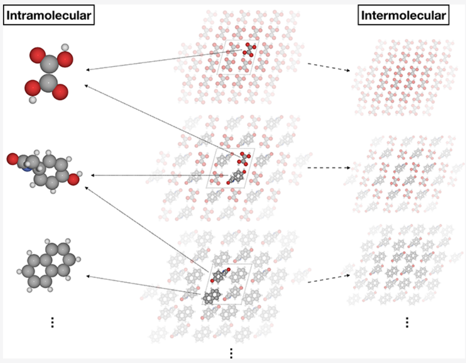
A Hybrid Machine Learning Approach for Structure Stability Prediction in Molecular Co-crystal Screenings

Simon Wengert, Gábor Csányi, Karsten Reuter, and Johannes T. Margraf*

Cite This: https://doi.org/10.1021/acs.jctc.2c00343 Read Online

ACCESS | Metrics & More | Article Recommendations | SI Supporting Information

ABSTRACT: Co-crystals are a highly interesting material class as varying their components and stoichiometry in principle allows tuning supramolecular assemblies toward desired physical properties. The *in silico* prediction of co-crystal structures represents a daunting task, however, as they span a vast search space and usually feature large unit cells. This requires theoretical models that are accurate and fast to evaluate, a combination that can in principle be accomplished by modern machine-learned (ML) potentials trained on first-principles data. Crucially, these ML potentials need to account for the description of long-range interactions, which are essential for the stability and structure of molecular crystals. In this contribution, we present a strategy for developing Δ-ML potentials for co-crystals, which use a physical baseline model to describe long-range interactions. The applicability of this approach is demonstrated for co-crystals of variable composition consisting of an active pharmaceutical ingredient and various co-formers. We find that the Δ-ML approach offers a strong and consistent improvement over the density functional tight binding baseline. Importantly, this even holds true when extrapolating beyond the scope of the training set, for instance in molecular dynamics simulations under ambient conditions.



1. INTRODUCTION

The physical properties of a molecular crystal are strongly dependent on the arrangement of its building blocks in the solid state.[1] In aggregate-induced emission, for instance, interactions in the crystalline phase (or even in concentrated solution) cause otherwise non-luminescent molecules to become emissive.[2] Similarly, piezochromic luminescent materials change the color of their emission when intermolecular arrangements in the solid state are altered by external mechanical stimuli.[3] Beyond these specific examples, the large variety of crystal forms detected and characterized for certain molecules reveals that the crystal structure impacts many other properties as well, such as aqueous solubility,[4] charge transport,[5] or plastic deformation[6] to name but a few.

Being able to control molecular arrangements in the solid state, consequently, enables tuning materials toward desired properties.[7] The design of multi-component molecular crystals, so-called co-crystals, is promising in this respect as it provides a versatile route to this goal.[8] Here, the molecule of interest crystallizes in the presence of another compound, a so-called co-former. Co-crystallization has garnered interest in both academia and industry as a strategy for the design of materials with improved performance. Applications include non-linear optics,[9] energetic materials,[10] and, most notably, pharmaceuticals.[11] Here, active pharmaceutical ingredients are often combined with co-formers to improve their bioavailabilty (e.g., by tuning the dissolution rate, solubility, compressibility, and thermal stability of the co-crystal).[12,13]

The space of possible co-formers is generally quite large. For pharmaceuticals, the "generally regarded as safe" (GRAS) list is often used, which contains hundreds of molecules considered as safe for human consumption. The synthesis of multi-component crystals thus provides a large design space. Unfortunately, the successful formation of a co-crystal from its compounds is by no means trivial.[14] Indeed, recrystallization is actually a common technique for purifying compounds, i.e., to separate them from one another. Moreover, the stability and structure of a potential co-crystal are hard to predict as they result from a delicate balance between relatively weak interactions.[15] Unlike conventional covalent chemistry, the synthesis of co-crystals is thus much more difficult to plan and often a game of trial and error. A more targeted approach would therefore be highly desirable. Here, computational methods could play an important role, e.g., by predicting whether a given co-former will lead to stable co-crystals and which structural motifs are likely to be formed for a given combination. This would allow narrowing the list of potential co-formers down to a few promising candidates and thus

Received: April 8, 2022

dramatically reduce the number of necessary experiments and associated costs.

The *in silico* search for molecular crystal structures faces some major challenges, however.[16] On one hand, the large search space of potential structures requires evaluating the stability of a large number of trial crystals. On the other hand, highly accurate (and thus computationally expensive) levels of theory need to be applied for a reliable prediction of crystal lattice energies.[17] Even for single-component crystals, this leads to a difficult trade-off between adequately exploring the space of possible structures and using sufficiently accurate methods to evaluate their stability. This situation is exacerbated on several fronts when screening for appropriate co-formers. First, a separate crystal structure search needs to be performed for each potential co-former. Second, the unit cells of co-crystals are typically significantly larger than those of single-component crystals as quantified by the number of molecules in the unit cell (Z) and the number of symmetry independent molecules (Z'). This means that there are more degrees of freedom to optimize ($Z' > 1$), while each energy evaluation is also more expensive (large Z). Finally, the stoichiometry of the stable co-crystal is typically unknown, which adds an additional dimension to the search space. As a consequence, computationally efficient and accurate potentials for crystal structure search and co-crystals in particular are highly desirable.

Owing to their outstanding accuracy-to-cost ratio, modern machine-learned (ML) potentials are in principle highly promising in this context. Challenges arise, however, from the importance of long-range contributions due to electrostatics or dispersion. Although recent advances in long-range ML potentials[18−22] bear good prospect for modeling condensed molecular systems, short-ranged ML potentials are still prevalent and, thus, generally less frequently applied in this context than for gas-phase molecules or ionic solids. As a notable exception, Montes-Campos et al. have nonetheless developed accurate ML potentials for molecular multicomponent systems and applied them to the related field of ionic liquids.[23] In this case, they benefited from the fact that the dynamics of liquids are only weakly influenced by long-range interactions, as is also the case for ion mobilities in solid electrolytes.[24] The importance of long-range interactions for the relative stabilities of molecular crystal polymorphs is well established, however.[25]

Kapil and Engel overcame this issue by using short-ranged ML potentials for sampling, in combination with additional *ab initio* calculations for stability ranking.[26] This allowed them to obtain highly accurate thermodynamic stabilities incorporating the combined effects from the electronic structure, quantum nuclear effects, and thermal contributions. In contrast, a Δ-ML[27] ansatz bypasses the need for subsequent *ab initio* calculations by combining local ML models with appropriate (long-ranged) baselines. This has proven to be highly useful for molecular crystal structure prediction (CSP).[28,29]

In a previous study, we presented a framework for the data-efficient generation of Δ-ML models for single-component molecular crystals, which benefits from a separate treatment of inter- and intramolecular interactions.[29] In this contribution, we present recent advances in extending this approach to co-crystals. Our approach is designed with the co-former screening setting in mind.[30] Consequently, we will consider a single active pharmaceutical ingredient (paracetamol) combined with four different co-formers, as shown in Figure 1. These systems have been proposed and extensively characterized by Karki et al.[13] Being one of the most common pharmaceuticals worldwide, paracetamol is a prototypical active pharmaceutical ingredient, while the co-formers oxalic acid (Oxa), naphthalene (Nap), phenazine (Phe), and theophylline (Thp) cover a wide range in terms of polarity, functional groups, and molecular shapes, inducing various types of intermolecular interactions and arrangements in the solid state.

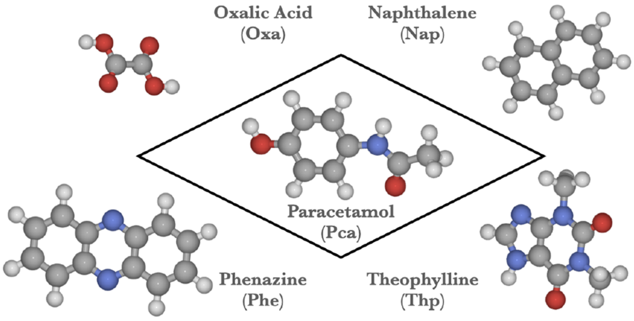


Figure 1. Central active pharmaceutical ingredient paracetamol (Pca) and the co-formers oxalic acid (Oxa), naphthalene (Nap), phenazine (Phe), and theophylline (Thp). Gray spheres: C, blue spheres: N, red spheres: O, white spheres: H.

2. METHODS

2.1. General Approach. The approach we previously developed[29] for single-component crystals has two main features. First, it combines a short-ranged ML potential with a long-ranged physical baseline (Δ-ML). Second, the ML potential is split into an intramolecular and intermolecular correction. The same idea was also used in local approximate models[31] for lattice energy minimizations of molecular crystals. We found this splitting to be advantageous because these interactions occur on different length scales. Additionally, reference data for the intramolecular correction can be generated cheaply from gas-phase calculations. It is even possible to use a different level of theory for this purpose. Below, we briefly summarize the main points of the method, highlighting the extensions that were developed for co-crystals.

2.2. Baseline Method. The dispersion-corrected density functional tight binding (DFTB) method represents an ideal baseline for CSP. First, it is efficient enough to be applied in a setting where several thousands of organic crystal structures need to be optimized.[32] In addition, the modern third-order variant of DFTB[33] combined with the 3ob[34] parameterization provides an accurate description of electrostatics, charge transfer, and polarization. Finally, the missing dispersion contributions can be corrected efficiently, e.g., via the D4 method.[35,36] The baseline method in this work is thus defined as DFTB3(3ob)+D4 (DFTB+D4 in the following).

2.3. Machine Learning Method. The intra- and intermolecular corrections to the baseline will be defined as Gaussian approximation potentials (GAP)[37,38] using the smooth overlap of atomic position (SOAP)[39] representation. These GAP models are fitted to both energies and forces. To account for the presence of different molecular building blocks in co-crystals, a separate intramolecular correction is fitted for each. In contrast, a single intermolecular correction is used to describe the interactions among paracetamol and the four co-formers. The energy expression of the combined DFTB+D4 and GAP model (termed Δ-GAP in the following) thus reads

$$E^{\Delta\text{-GAP}} = E_{\text{crystal}}^{\text{DFTB+D4}} + \Delta E_{\text{crystal}}^{\text{inter}} + \sum_{t}^{N_{\text{types}}} \sum_{i}^{N_t} \Delta E_i^{\text{intra},t} \tag{1}$$

where for each of the N_{types} possible components, the corresponding intramolecular GAP correction is applied to each molecule i (in which N_t is the number of molecules of type t present in the given unit cell). Note that intra- and intermolecular corrections are applied to energies, forces, and stresses. The models can thus be used for full unit cell relaxations and constant pressure molecular dynamics.

2.4. Target Method.

The high-level target method to which the correction is fitted will be hybrid DFT (using the PBE0[40] functional) with a many-body dispersion[25,41] (MBD) correction. PBE0+MBD provides a sophisticated description of the interactions relevant to organic solids. The importance of MBD contributions and hybrid functionals for the stability assessment of molecular crystals has been highlighted by Hoja and Tkatchenko.[42] For the X23 database, containing van der Waals (vdW)-bonded, hydrogen-bonded, and mixed molecular crystals, this combination has been shown to yield lattice energies within chemical accuracy (43 meV) when compared to (back-corrected) experimental enthalpies of sublimation.[43] Moreover, LeBlanc et al. found in their studies on multicomponent acid−base crystals that the exact-exchange mixing employed in hybrid DFT is essential to cure significant geometry errors introduced by the delocalization error of semilocal functionals.[44] Due to the prohibitive computational and memory requirements of PBE0+MBD with large basis sets, we define the target method—called PBE(0)+MBD hereafter—as a composite scheme: The intramolecular part is fully described by PBE0+MBD with a tightly converged basis of numerical atomic orbitals (NAO). The intermolecular part is described by PBE+MBD[45] with the same basis, plus the difference PBE+MBD to PBE0+MBD in a smaller NAO basis. A similar scheme was used by Hoja et al.,[17] who found it to yield lattice energies in excellent agreement with converged PBE0+MBD calculations.

2.5. Training Data.

The structures entering the training set ultimately define the information that is available about the target function. In the context of co-crystal screening studies, the training set should thus include combinations of the molecule of interest with all co-formers. To train the intermolecular model, we selected samples from a pool of ca. 10,000 trial structures created with the PyXtal package.[46] In this initial pool, a wide range of compositions was considered for each combination to span all possible stoichiometries. These trial candidates were locally relaxed at the DFTB+D4 level of theory. To obtain a diverse set of training structures from this pool, we then employed the farthest point sampling (FPS)[47] heuristic. Here, the SOAP kernel was used as a similarity measure between atomic environments and structures were sequentially added to the training set by selecting the most dissimilar structures to the current training set at each iteration. Note that there are several possibilities to define global similarity metrics between structures, given a local similarity metric like SOAP.[48] Herein, we simply used the maximal dissimilarity between any two atomic environments.[49] From this process, 1000 training structures were obtained, 250 for each crystal/co-former pair (including the corresponding single-component crystals).

We further included 77 structures corresponding to the experimentally known single-component crystals and randomly perturbed structures derived from them. The rationale behind this is that the experimental information about the single-component crystals is usually available in co-crystal studies. This allows us to include some additional information on highly stable interactions, though not for the important paracetamol/co-former contacts. The consequences of this bias in the training set will be discussed in detail below.

In contrast to the intermolecular correction, the training data for the intramolecular model is computationally cheap to generate as it only requires single-point calculations on monomer configurations in the gas phase. To obtain these configurations, monomer geometries were extracted from the training crystals. These were further supplemented, with configurations from gas-phase molecular dynamics simulations and local relaxations, to extensively cover the configurational space of each building block. Further details on the training sets and all training data are provided in the Supporting Information.

3. RESULTS AND DISCUSSION

To validate the presented approach, we will first test its performance on a diverse set of crystal structures as one would encounter in a CSP workflow. To this end, a test set of 1000 structures was generated in an analogous procedure to the training set generation. Here, the FPS selection included the training set to maximize the distance between test and training structures (see the Supporting Information for details). All test structures were subsequently relaxed at the Δ-GAP level. Lattice energies and force errors for this test set are summarized in Figure 2. For lattice energy calculation, we used

$$E_{\text{crystal}}^{\text{latt}} = (E_{\text{crystal}} - n_{\text{A}}E_{\text{gas,A}} - n_{\text{B}}E_{\text{gas,B}})/(n_{\text{A}} + n_{\text{B}}) \tag{2}$$

where the difference between the energy of the crystal, E_{crystal}, and the energies, E_{gas}, of its optimized molecular compounds is computed first and then normalized by the total number of compounds in the crystal unit cell. Note that lattice energies of single-component crystals have been calculated in the same way using $n_{\text{B}} = 0$.

In Figure 2 (top), Δ-GAP and DFTB+D4 predicted lattice energies are shown in comparison with the PBE(0)+MBD target values. The reference energies cover a broad range of ca. 1 eV per molecule and are mostly negative. This indicates that the random search in general leads to reasonable candidate structures, which are stable with respect to sublimation. The DFTB+D4 lattice energies are reasonably well correlated with this reference but display significant scatter. Furthermore, the lattice energies are systematically underestimated, leading to a mean absolute error (MAE) of 183 meV. Applying intra- and intermolecular corrections to this baseline in the Δ-GAP scheme strongly improves the agreement with the target, resulting in an overall MAE of only 34 meV. This is achieved both by eliminating the systematic underestimation of the lattice energies and by reducing the scatter in the predictions, as indicated by the significantly smaller standard deviation (STD) of the Δ-GAP errors (32 meV vs 83 meV). Indeed, the Δ-GAP energies actually show a slight offset toward more negative values due to the fact that the structures are minima on the Δ-GAP potential energy surface.

An even more substantial improvement is observed for force predictions (see Figure 2, bottom). Here, DFTB+D4 displays a broad error distribution and a correspondingly large MAE of 324 meV/Å. In contrast, the error distribution of predicted Δ-

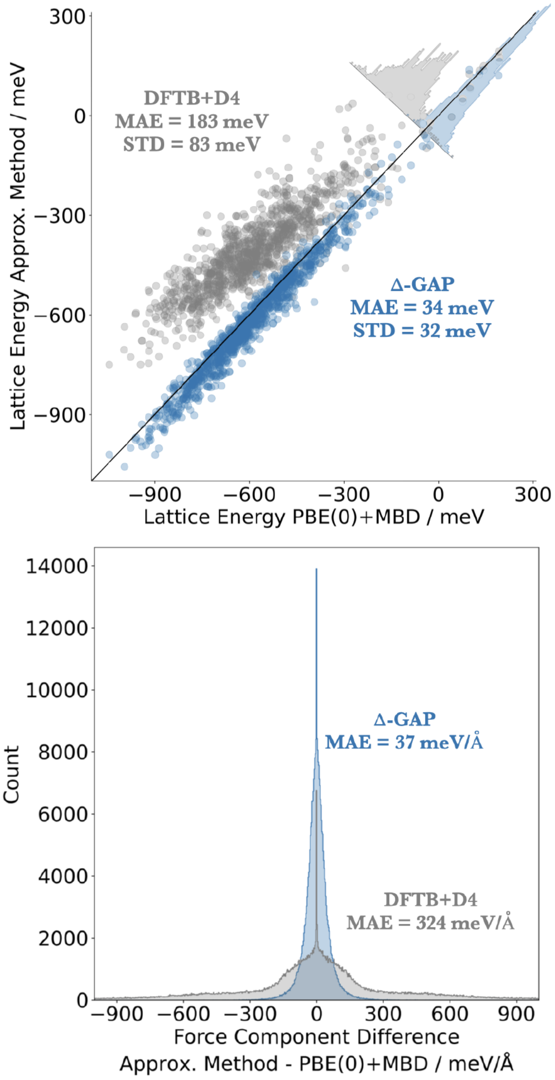


Figure 2. Correlation plot for the DFTB+D4 baseline and Δ-GAP lattice energies per molecule of PcaOxa, PcaNap, PcaPhe, and PcaThp test crystals (both single-component and co-crystals) against the PBE(0)+MBD target level of theory (top) and the corresponding differences in force components (bottom). Note that the slight shift of the Δ-GAP lattice energy distribution toward lower values compared to PBE(0)+MBD is due to the fact that the test set structures are minima on the Δ-GAP potential energy surface, while the training structures are minima on the DFTB+D4 surface (see text). The spike in the distributions of force component differences results from certain force components being zero by symmetry at all levels of theory.

GAP force components is much narrower and the MAE almost an order of magnitude lower. Importantly, while the lattice energy error of DFTB+D4 is fairly systematic, the force error cannot be corrected in a simple way and will lead to substantial deviations in the predicted structures. This is of particular relevance in the context of CSP, where accurate structure relaxations are often by far the most expensive component. Due to their small force errors, Δ-GAP relaxations should provide near PBE(0)+MBD quality structures at a fraction of the computational costs.

While the above results are promising, it should be emphasized that the training and test structures used herein are merely local minima. In particular, they are somewhat less dense and less stable than the known experimental structures for these co-crystals (see the Supporting Information). In future applications, this should be mitigated by using a more advanced CSP search algorithm (ideally together with an accurate ML potential as proposed herein) to generate more realistic structures. From the perspective of this paper, there is also a positive aspect to this discrepancy between training and experimental structures though, as it creates an opportunity to test the extrapolative capabilities of the presented approach. To this end, we test the accuracy of our method on the known experimental structures of each co-crystal.

For all experimental co-crystal structures, atomic positions and unit cell parameters were fully relaxed using the DFTB+D4 baseline, Δ-GAP model, and the PBE(0)+MBD target. For comparison, we also performed calculations at the PBE+MBD level, which is often used for relaxations instead of the more expensive hybrid PBE0 functional. These results are summarized in Figure 3.

Relative density deviations with respect to the PBE(0)+MBD geometry are shown in Figure 3a. We find that the DFTB+D4 structures are significantly contracted, in agreement with previous studies where this was attributed to insufficient Pauli-repulsion at longer distances.[32,50] In contrast, the Δ-GAP structures are in much better agreement, with only slightly higher densities. For comparison, PBE+MBD shows slightly larger but more systematic density deviations of around 3%. In contrast to Δ-GAP and DFTB+D4, this is due to systematically lower densities, which are likely a consequence of differences in the molecular electrostatic potentials predicted by semi-local and hybrid functionals.

On an atomistic level, crystal structures are typically compared with the RMSD_{15} metric,[51] as shown in Figure 3b. To this end, the root mean square deviation of the positions of non-hydrogen atoms in 15-molecule clusters extracted from the relaxed crystal structures is calculated. We again use the PBE(0)+MBD structures as the reference. As for the densities, the DFTB+D4 baseline displays the most significant structural discrepancies with the target. These are mostly due to reduced intermolecular distances, such as the spacings in the layered structures PcaOxa, PcaNap, and PcaThp and variations in molecular orientation (see Figure 4 and the Supporting Information for further examples). For PcaNap, additional discrepancy is caused by the intramolecular adjustment of paracetamol to the crystal environment. Here, the DFTB+D4 baseline predicts a weaker out-of-plane rotation of the C═O group, as highlighted in the inset. In all cases, these deviations are mitigated by the ML correction, though the effects are less distinct for PcaThp, which is already reasonably well described by the baseline. Finally, PBE+MBD is slightly more accurate and systematic than Δ-GAP, albeit at a much higher computational cost (by roughly 3 orders of magnitude, see the Supporting Information). Indeed, the structural discrepancies are in this case entirely due to the aforementioned density deviations, whereas the relative positions and orientations of the molecules are in good agreement with the PBE(0)+MBD relaxed structures.

In addition to these geometric comparisons, the relaxed structures were also evaluated from an energetic perspective. This is relevant when structures from the approximate method are used as inputs for single-point calculations or relaxations with higher level methods. Here, small structural deviations—bond distances for instance—can significantly impact predicted energies and energy differences. To evaluate the

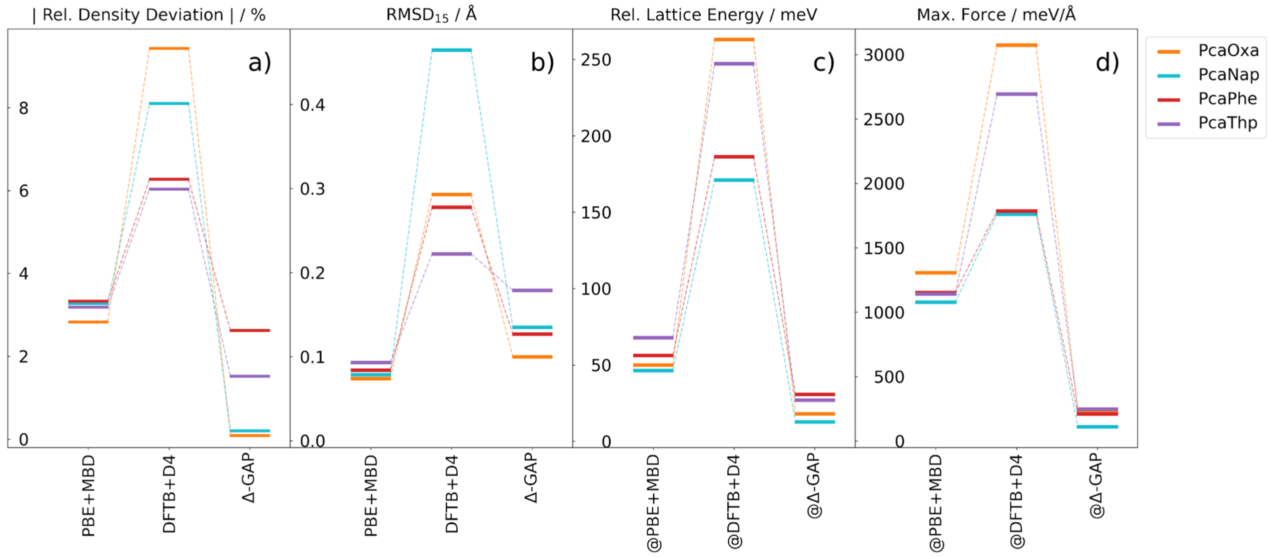


Figure 3. Comparison between PBE+MBD, the DFTB+D4 baseline, and Δ-GAP results on experimental co-crystals for PcaOxa, PcaNap, PcaPhe, and PceThp against the PBE(0)+MBD target level of theory in terms of the absolute values for percentage density deviations (a), the RMSDs between overlaying 15-mers sliced from crystal structures (b), lattice energies per molecule relative to PBE(0)+MBD optimized structures obtained from single-point calculations on structures optimized on the approximate levels of theory specified in the figure (c), and the corresponding maximum remaining PBE(0)+MBD forces (d).

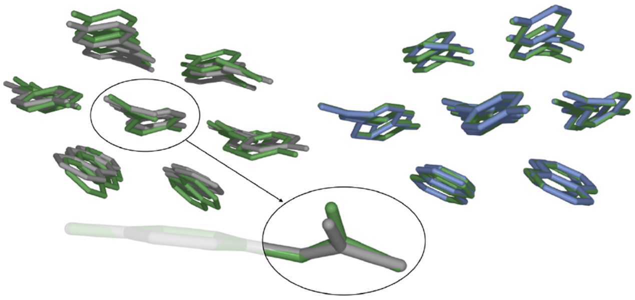


Figure 4. Overlay of the PBE(0)+MBD (green) optimized experimental PcaNap co-crystal with DFTB+D4 (gray) and Δ-GAP (blue). For DFTB+D4, a separate overlay is shown for paracetamol conformers extracted from the crystal environment.

quality of the structures in this context, single-point PBE(0)+MBD calculations were performed on the geometries predicted by the approximate levels of theory. Figure 3c illustrates the errors in lattice energies obtained from these calculations, while Figure 3d shows the corresponding maximum force. Here, the Δ-GAP values are lowest in all cases, indicating that they are closest to the PBE(0)+MBD minimum from an energetic perspective. The deviations of PBE+MBD are similarly systematic but significantly higher. Finally, the DFTB+D4 results are more scattered and generally poorer with maximum forces of up to 3 eV/Å for the putative minima and lattice energy errors of up to 250 meV.

Overall, the Δ-GAP model is thus a robust and significant improvement on DFTB+D4, even when applied outside the range of the training set. Perhaps surprisingly, it is even an improvement over the much more expensive PBE+MBD method in many respects, when comparing with the PBE(0)+MBD target. Of course, the ultimate test is comparison with experimental structures, however. Here, we somewhat unexpectedly found that the PBE+MBD densities are actually closer to the experimental values than the ones predicted by PBE(0)+MBD (and consequently also by Δ-GAP, see Figure 5).

These apparent deviations can be resolved by considering thermal effects, however. Computationally relaxed crystal structures correspond to the 0 K limit, whereas crystallographic experiments are usually performed at finite temperature and pressure. The over-contraction of PBE(0)+MBD will thus be counteracted by thermal expansion. An advantage of computa-

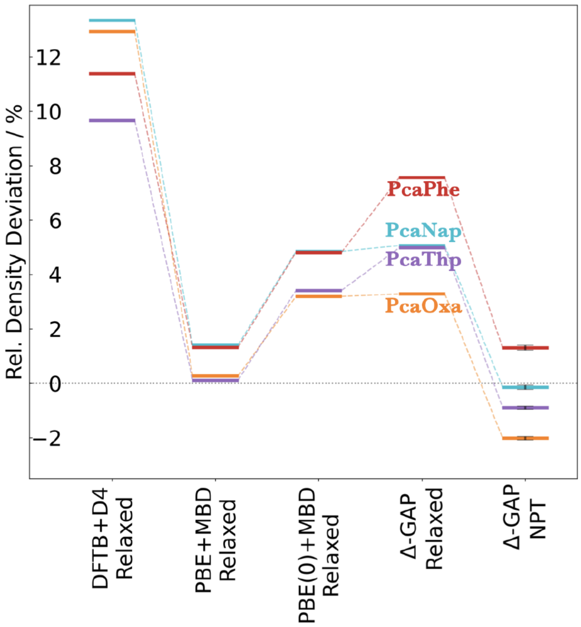


Figure 5. Percentage deviations from experimental measured densities for PcaOxa, PcaNap, PcaPhe, and PceThp co-crystals optimized with the DFTB+D4 baseline, PBE+MBD, the PBE(0)+MBD target level of theory, and Δ-GAP, as well as for densities obtained from Δ-GAP NPT simulations (298 K and 1 bar). For NPT, results corresponding to standard errors of the deviations are illustrated.

tionally efficient approaches like Δ-GAP is that they allow for including such effects in a straightforward manner by performing molecular dynamics in the NPT ensemble (at 298 K and ambient pressure). As shown in Figure 5, the average densities across these trajectories are indeed in very good agreement with the experiment. This also indicates that the PBE+MBD (0 K) densities are in fact fortuitously close to the experiment as the inclusion of thermal expansion effects would likely also cause them to decrease by ca. 5%.

Importantly, such finite temperature simulations would be computationally prohibitive on the hybrid DFT level. Being an efficient surrogate for PBE(0)+MBD, Δ-GAP thus allows performing simulations that would otherwise be impossible. These results also further underscore the robustness of our ML approach, given that the experimental structures are outside the scope of the training set and no crystal MD data was used for training at all. This is thanks to the strong physical prior that the DFTB+D4 baseline provides and the smoothness of the GAP correction. Additional improvements could be obtained by combining the current approach with more advanced structure search algorithms[52−54] and by iteratively refining the GAP correction in an active learning workflow.

4. CONCLUSIONS

We have presented an approach for Δ-ML potentials applicable to both pure crystals and co-crystals of variable composition. This Δ-GAP approach enables efficient global crystal structure searches with near hybrid DFT accuracy, at a much reduced cost. Building on a previous approach for single-component crystals, we fit separate intramolecular corrections for each component and a single intermolecular correction for all active molecule/co-former pairs. Our approach strongly reduces energy and force errors with respect to the baseline model.

Notably, the training structures used herein were generated with a simple random search procedure and consequently display markedly lower densities and stabilities than the known experimental co-crystals. Nevertheless, the Δ-GAP potentials are able to predict the structures of experimental polymorphs with high accuracy, outperforming PBE+MBD at a much lower computational cost. This shows that this approach is highly robust in an extrapolative regime. In future work, we aim to combine these potentials with more advanced CSP search algorithms.[52−54]

Finally, it should be noted that many-body dispersion can be rather long-ranged in some cases,[55] while our baseline method relies on the D4 correction, which lacks these effects. Since the intermolecular ML contributions are by construction short-ranged due to the use of a local representation, long-range many-body dispersion effects are thus currently neglected in our approach. This could be mitigated by including a physical many-body dispersion model in the baseline. An efficient ML-based MBD implementation that makes this computationally feasible has recently been reported.[56,57]

5. COMPUTATIONAL DETAILS

DFT calculations were performed with the all-electron code FHI-aims,[58] using the PBE[45] and PBE0[40] functionals. A post-SCF dispersion correction was applied using the MBD[25,41] method. Two accuracy levels with a large or small basis set have been used (compare Section 2). Large basis set calculations correspond to tier2 settings and tight integration grids, while small basis set calculations correspond to tier1 settings and light integration grids. DFTB3[33] calculations were performed using DFTB+[59] together with the 3ob[34] parametrization and the D4[35,36] dispersion correction without non-additive effects. For periodic calculations, the number of **k** points (n) in each direction is chosen as the smallest integer satisfying the relation $n \cdot a \geq x$, where a is the unit cell length along that direction and $x = 30$. GAP potentials were trained and evaluated using the QUIP[60] package. Candidate crystal structures were created with the PyXtal[46] package.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jctc.2c00343.

> Method and additional details on co-crystal stabilities, density dependence of the lattice energies for experimental co-crystals, structural overlay of optimized experimental co-crystals, density and lattice energy comparison between training and experimental co-crystals, molecular dynamics simulations, and comparison of computational costs (PDF)

AUTHOR INFORMATION

Corresponding Author

Johannes T. Margraf − *Fritz-Haber-Institut der Max-Planck-Gesellschaft, 14195 Berlin, Germany;* orcid.org/0000-0002-0862-5289; Email: johannes.margraf@ch.tum.de

Authors

Simon Wengert − *Fritz-Haber-Institut der Max-Planck-Gesellschaft, 14195 Berlin, Germany; Chair of Theoretical Chemistry, Technische Universität München, 85747 Garching, Germany*

Gábor Csányi − *Engineering Laboratory, University of Cambridge, Cambridge CB2 1PZ, United Kingdom*

Karsten Reuter − *Fritz-Haber-Institut der Max-Planck-Gesellschaft, 14195 Berlin, Germany*

Complete contact information is available at:
https://pubs.acs.org/10.1021/acs.jctc.2c00343

Funding

Open access funded by Max Planck Society.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

S.W. and K.R. acknowledge funding from Deutsche Forschungsgemeinschaft (DFG) under Project RE1509/18-2.

REFERENCES

(1) Aitipamula, S.; Chow, P. S.; Tan, R. B. H. Polymorphism in cocrystals: a review and assessment of its significance. *CrystEngComm* **2014**, *16*, 3451−3465.

(2) Hong, Y.; Lam, J. W. Y.; Tang, B. Z. Aggregation-induced emission. *Chem. Soc. Rev.* **2011**, *40*, 5361−5388.

(3) Sagara, Y.; Kato, T. Mechanically induced luminescence changes in molecular assemblies. *Nat. Chem.* **2009**, *1*, 605−610.

(4) Bernstein, J. *Polymorphism in Molecular Crystals*; International Union of Crystallography Monographs on Crystallography; Oxford University Press: Oxford, U.K., 2007.

(5) Jurchescu, O. D.; Mourey, D. A.; Subramanian, S.; Parkin, S. R.; Vogel, B. M.; Anthony, J. E.; Jackson, T. N.; Gundlach, D. J. Effects of

polymorphism on charge transport in organic semiconductors. *Phys. Rev. B* **2009**, *80*, No. 085201.

(6) Nichols, G.; Frampton, C. S. Physicochemical characterization of the orthorhombic polymorph of paracetamol crystallized from solution. *J. Pharm. Sci.* **1998**, *87*, 684−693.

(7) Aakeröy, C. B.; Sandhu, B. Solid Form Landscape and Design of Physical Properties. In *Engineering Crystallography: From Molecule to Crystal to Functional Form*; Roberts, K. J., Docherty, R., Tamura, R., Eds.; Springer: Dordrecht, The Netherlands, 2017; pp 45−56.

(8) Gunawardana, C. A.; Aakeröy, C. B. Co-crystal synthesis: fact, fancy, and great expectations. *Chem. Commun.* **2018**, *54*, 14047−14060.

(9) Choi, E.-Y.; Jazbinsek, M.; Lee, S.-H.; Günter, P.; Yun, H.; Lee, S. W.; Kwon, O.-P. Co-crystal structure selection of nonlinear optical analogue polyenes. *CrystEngComm* **2012**, *14*, 4306−4311.

(10) Aakeröy, C. B.; Wijethunga, T. K.; Desper, J. Crystal Engineering of Energetic Materials: Co-crystals of Ethylenedinitramine (EDNA) with Modified Performance and Improved Chemical Stability. *Chem. − Eur. J.* **2015**, *21*, 11029−11037.

(11) Steed, J. W. The role of co-crystals in pharmaceutical design. *Trends Pharmacol. Sci.* **2013**, *34*, 185−193.

(12) Schultheiss, N.; Newman, A. Pharmaceutical Cocrystals and Their Physicochemical Properties. *Cryst. Growth Des.* **2009**, *9*, 2950−2967.

(13) Karki, S.; Friščić, T.; Fábián, L.; Laity, P. R.; Day, G. M.; Jones, W. Improving Mechanical Properties of Crystalline Solids by Cocrystal Formation: New Compressible Forms of Paracetamol. *Adv. Mater.* **2009**, *21*, 3905−3909.

(14) Springuel, G.; Norberg, B.; Robeyns, K.; Wouters, J.; Leyssens, T. Advances in Pharmaceutical Co-crystal Screening: Effective Co-crystal Screening through Structural Resemblance. *Cryst. Growth Des.* **2012**, *12*, 475−484.

(15) Aakeröy, C. Is there any point in making co-crystals? *Acta Crystallogr., Sect. B: Struct. Sci., Cryst. Eng. Mater.* **2015**, *71*, 387−391.

(16) Reilly, A. M.; et al. Report on the sixth blind test of organic crystal structure prediction methods. *Acta Crystallogr., Sect. B: Struct. Sci., Cryst. Eng. Mater.* **2016**, *72*, 439−459.

(17) Hoja, J.; Ko, H.-Y.; Neumann, M. A.; Car, R.; DiStasio, R. A.; Tkatchenko, A. Reliable and practical computational description of molecular crystal polymorphs. *Sci. Adv.* **2019**, *5*, eaau3338.

(18) Grisafi, A.; Ceriotti, M. Incorporating long-range physics in atomic-scale machine learning. *J. Chem. Phys.* **2019**, *151*, 204105.

(19) Xie, X.; Persson, K. A.; Small, D. W. Incorporating Electronic Information into Machine Learning Potential Energy Surfaces via Approaching the Ground-State Electronic Energy as a Function of Atom-Based Electronic Populations. *J. Chem. Theory Comput.* **2020**, *16*, 4256−4270.

(20) Ko, T. W.; Finkler, J. A.; Goedecker, S.; Behler, J. A fourthgeneration high-dimensional neural network potential with accurate electrostatics including non-local charge transfer. *Nat. Commun.* **2021**, *12*, 398.

(21) Veit, M.; Wilkins, D. M.; Yang, Y.; DiStasio, R. A.; Ceriotti, M. Predicting molecular dipole moments by combining atomic partial charges and atomic dipoles. *J. Chem. Phys.* **2020**, *153*, No. 024113.

(22) Staacke, C. G.; Wengert, S.; Kunkel, C.; Csányi, G.; Reuter, K.; Margraf, J. T. Kernel charge equilibration: efficient and accurate prediction of molecular dipole moments with a machinelearning enhanced electron density model. *Mach. Learn.: Sci. Technol.* **2022**, *3*, 015032.

(23) Montes-Campos, H.; Carrete, J.; Bichelmaier, S.; Varela, L. M.; Madsen, G. K. H. A Differentiable Neural-Network Force Field for Ionic Liquids. *J. Chem. Inf. Model.* **2022**, *62*, 88−101. 34941253

(24) Staacke, C.; Heenen, H.; Scheurer, C.; Csányi, G.; Reuter, K.; Margraf, J. On the Role of Long-Range Electrostatics in Machine-Learned Interatomic Potentials for Complex Battery Materials. *ACS Appl. Energy Mater.* **2021**, *4*, 12562−12569.

(25) Ambrosetti, A.; Reilly, A. M.; DiStasio, R. A.; Tkatchenko, A. Long-range correlation energy calculated from coupled atomic response functions. *J. Chem. Phys.* **2014**, *140*, 18A508.

(26) Kapil, V.; Engel, E. A. A complete description of thermodynamic stabilities of molecular crystals. *Proc. Natl. Acad. Sci. U. S. A.* **2022**, *119*, No. e2111769119.

(27) Bartók, A. P.; Gillan, M. J.; Manby, F. R.; Csányi, G. Machine-learning approach for one- and two-body corrections to density functional theory: Applications to molecular and condensed water. *Phys. Rev. B* **2013**, *88*, 054104.

(28) McDonagh, D.; Skylaris, C.-K.; Day, G. M. Machine-Learned Fragment-Based Energies for Crystal Structure Prediction. *J. Chem. Theory Comput.* **2019**, *15*, 2743−2758.

(29) Wengert, S.; Csányi, G.; Reuter, K.; Margraf, J. T. Dataefficient machine learning for molecular crystal structure prediction. *Chem. Sci.* **2021**, *12*, 4536−4546.

(30) Aakeröy, C. B.; Grommet, A. B.; Desper, J. Co-crystal Screening of Diclofenac. *Pharmaceutics* **2011**, *3*, 601−614.

(31) Kazantsev, A. V.; Karamertzanis, P. G.; Adjiman, C. S.; Pantelides, C. C. Efficient Handling of Molecular Flexibility in Lattice Energy Minimization of Organic Crystals. *J. Chem. Theory Comput.* **2011**, *7*, 1998−2016.

(32) Iuzzolino, L.; McCabe, P.; Price, S.; Brandenburg, J. G. Crystal structure prediction of flexible pharmaceutical-like molecules: density functional tight-binding as an intermediate optimisation method and for free energy estimation. *Faraday Discuss.* **2018**, *211*, 275−296.

(33) Gaus, M.; Cui, Q.; Elstner, M. DFTB3: Extension of the Self-Consistent-Charge Density-Functional Tight-Binding Method (SCC-DFTB). *J. Chem. Theory Comput.* **2011**, *7*, 931−948.

(34) Gaus, M.; Goez, A.; Elstner, M. Parametrization and Benchmark of DFTB3 for Organic Molecules. *J. Chem. Theory Comput.* **2013**, *9*, 338−354.

(35) Hourahine, B.; et al. DFTB, a software package for efficient approximate density functional theory based atomistic simulations. *J. Chem. Phys.* **2020**, *152*, 124101.

(36) Caldeweyher, E.; Ehlert, S.; Hansen, A.; Neugebauer, H.; Spicher, S.; Bannwarth, C.; Grimme, S. A generally applicable atomic-charge dependent London dispersion correction. *J. Chem. Phys.* **2019**, *150*, 154122.

(37) Bartók, A. P.; Payne, M. C.; Kondor, R.; Csányi, G. Gaussian Approximation Potentials: The Accuracy of Quantum Mechanics, without the Electrons. *Phys. Rev. Lett.* **2010**, *104*, 136403.

(38) Deringer, V. L.; Bartók, A. P.; Bernstein, N.; Wilkins, D. M.; Ceriotti, M.; Csányi, G. Gaussian Process Regression for Materials and Molecules. *Chem. Rev.* **2021**, *121* (16), 10073−10141.

(39) Bartók, A. P.; Kondor, R.; Csányi, G. On representing chemical environments. *Phys. Rev. B* **2013**, *87*, 184115.

(40) Adamo, C.; Cossi, M.; Barone, V. An accurate density functional method for the study of magnetic properties: the PBE0 model. *J. Mol. Struct.: THEOCHEM* **1999**, *493*, 145−157.

(41) Tkatchenko, A.; DiStasio, R. A.; Car, R.; Scheffler, M. Accurate and Efficient Method for Many-Body van der Waals Interactions. *Phys. Rev. Lett.* **2012**, *108*, 236402.

(42) Hoja, J.; Tkatchenko, A. First-principles stability ranking of molecular crystal polymorphs with the DFT+MBD approach. *Faraday Discuss.* **2018**, *211*, 253−274.

(43) Reilly, A. M.; Tkatchenko, A. Understanding the role of vibrations, exact exchange, and many-body van der Waals interactions in the cohesive properties of molecular crystals. *J. Chem. Phys.* **2013**, *139*, No. 024705.

(44) LeBlanc, L. M.; Dale, S. G.; Taylor, C. R.; Becke, A. D.; Day, G. M.; Johnson, E. R. Pervasive Delocalisation Error Causes Spurious Proton Transfer in Organic Acid−Base Co-Crystals. *Angew. Chem., Int. Ed.* **2018**, *57*, 14906−14910.

(45) Perdew, J. P.; Burke, K.; Ernzerhof, M. Generalized Gradient Approximation Made Simple. *Phys. Rev. Lett.* **1996**, *77*, 3865−3868.

(46) Fredericks, S.; Parrish, K.; Sayre, D.; Zhu, Q. PyXtal: A Python library for crystal structure generation and symmetry analysis. *Comput. Phys. Commun.* **2021**, *261*, 107810.

(47) Ceriotti, M.; Willatt, M. J.; Csányi, G. Machine Learning of Atomic-Scale Properties Based on Physical Principles. *Handb. Mater. Model.* **2018**, 1−27.

(48) De, S.; Bartók, A. P.; Csányi, G.; Ceriotti, M. Comparing molecules and solids across structural and alchemical space. *Phys. Chem. Chem. Phys.* **2016**, *18*, 13754−13769.

(49) Timmermann, J.; Lee, Y.; Staacke, C. G.; Margraf, J. T.; Scheurer, C.; Reuter, K. Data-efficient iterative training of Gaussian approximation potentials: Application to surface structure determination of rutile IrO_2 and RuO_2. *J. Chem. Phys.* **2021**, *155*, 244107.

(50) Mortazavi, M.; Brandenburg, J. G.; Maurer, R. J.; Tkatchenko, A. Structure and Stability of Molecular Crystals with Many-Body Dispersion-Inclusive Density Functional Tight Binding. *J. Phys. Chem. Lett.* **2018**, *9*, 399−405.

(51) Chisholm, J. A.; Motherwell, S. *COMPACK*: a program for identifying crystal structure similarity using distances. *J. Appl.Crystallogr.* **2005**, *38*, 228−231.

(52) Case, D. H.; Campbell, J. E.; Bygrave, P. J.; Day, G. M. Convergence Properties of Crystal Structure Prediction by Quasi-Random Sampling. *J. Chem. Theory Comput.* **2016**, *12*, 910−924.

(53) Tom, R.; Rose, T.; Bier, I.; O'Brien, H.; Vázquez-Mayagoitia, Á.; Marom, N. Genarris 2.0: A random structure generator for molecular crystals. *Comput. Phys. Commun.* **2020**, *250*, 107170.

(54) Song, H.; Vogt-Maranto, L.; Wiscons, R.; Matzger, A. J.; Tuckerman, M. E. Generating Cocrystal Polymorphs with Information Entropy Driven by Molecular Dynamics-Based Enhanced Sampling. *J. Phys. Chem. Lett.* **2020**, *11*, 9751−9758.

(55) Hoja, J.; Reilly, A. M.; Tkatchenko, A. First-principles modeling of molecular crystals: structures and stabilities, temperature and pressure. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2017**, *7*, No. e1294.

(56) Poier, P. P.; Lagardère, L.; Piquemal, J.-P. $O(N)$ Stochastic Evaluation of Many-Body van der Waals Energies in Large Complex Systems. *J. Chem. Theory Comput.* **2022**, *18*, 1633−1645.

(57) Poier, P. P.; Jaffrelot Inizan, T.; Adjoua, O.; Lagardère, L.; Piquemal, J.-P. Accurate Deep Learning-Aided Density-Free Strategy for Many-Body Dispersion-Corrected Density Functional Theory. *J. Phys. Chem. Lett.* **2022**, *13*, 4381−4388.

(58) Blum, V.; Gehrke, R.; Hanke, F.; Havu, P.; Havu, V.; Ren, X.; Reuter, K.; Scheffler, M. Ab initio molecular simulations with numeric atom-centered orbitals. *Comput. Phys. Commun.* **2009**, *180*, 2175−2196.

(59) Aradi, B.; Hourahine, B.; Frauenheim, T. DFTB, a Sparse Matrix-Based Implementation of the DFTB Method. *J. Phys. Chem. A* **2007**, *111*, 5678−5684.

(60) Bartók, A. P.; Csányi, G. Gaussian approximation potentials: A brief tutorial introduction. *Int. J. Quantum Chem.* **2015**, *115*, 1051−1057.

https://doi.org/10.1021/acs.jctc.2c00343

Paper # 3

Mapping Materials and Molecules

Bingqing Cheng, Ryan-Rhys Griffiths, Simon Wengert, Christian Kunkel, Tamas Stenczel, Bonan Zhu, Volker L. Deringer, Noam Bernstein, Johannes T. Margraf, Karsten Reuter, and Gábor Csányi

Acc. Chem. Res. **53**, 1981 (2020)
DOI: https://doi.org/10.1021/acs.accounts.0c00403

ACCOUNTS
of chemical research

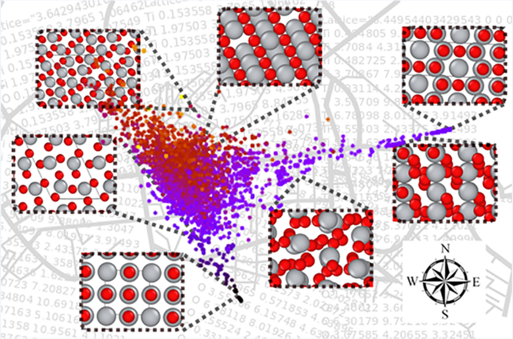
pubs.acs.org/accounts Article

Mapping Materials and Molecules

Bingqing Cheng,* Ryan-Rhys Griffiths, Simon Wengert, Christian Kunkel, Tamas Stenczel, Bonan Zhu, Volker L. Deringer, Noam Bernstein, Johannes T. Margraf, Karsten Reuter, and Gabor Csanyi

Cite This: *Acc. Chem. Res.* 2020, 53, 1981−1991 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information



CONSPECTUS: The visualization of data is indispensable in scientific research, from the early stages when human insight forms to the final step of communicating results. In computational physics, chemistry and materials science, it can be as simple as making a scatter plot or as straightforward as looking through the snapshots of atomic positions manually. However, as a result of the "big data" revolution, these conventional approaches are often inadequate. The widespread adoption of high-throughput computation for materials discovery and the associated community-wide repositories have given rise to data sets that contain an enormous number of compounds and atomic configurations. A typical data set contains thousands to millions of atomic structures, along with a diverse range of properties such as formation energies, band gaps, or bioactivities.

It would thus be desirable to have a data-driven and automated framework for visualizing and analyzing such structural data sets. The key idea is to construct a low-dimensional representation of the data, which facilitates navigation, reveals underlying patterns, and helps to identify data points with unusual attributes. Such data-intensive maps, often employing machine learning methods, are appearing more and more frequently in the literature. However, to the wider community, it is not always transparent how these maps are made and how they should be interpreted. Furthermore, while these maps undoubtedly serve a decorative purpose in academic publications, it is not always apparent what extra information can be garnered from reading or making them.

This Account attempts to answer such questions. We start with a concise summary of the theory of representing chemical environments, followed by the introduction of a simple yet practical conceptual approach for generating structure maps in a generic and automated manner. Such analysis and mapping is made nearly effortless by employing the newly developed software tool ASAP. To showcase the applicability to a wide variety of systems in chemistry and materials science, we provide several illustrative examples, including crystalline and amorphous materials, interfaces, and organic molecules. In these examples, the maps not only help to sift through large data sets but also reveal hidden patterns that could be easily missed using conventional analyses.

The explosion in the amount of computed information in chemistry and materials science has made visualization into a science in itself. Not only have we benefited from exploiting these visualization methods in previous works, we also believe that the automated mapping of data sets will in turn stimulate further creativity and exploration, as well as ultimately feed back into future advances in the respective fields.

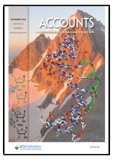
KEY REFERENCES

- Bartok, A. P.; De, S.; Poelking, C.; Bernstein, N.; Kermode, J.; Csanyi, G.; Ceriotti, M. Machine Learning Unifies the Modelling of Materials and Molecules. *Science Advances* **2017**, *3*, e1701816.[1] *Showcases SOAP and Gaussian process regression for machine learning in a variety of material and molecular prediction tasks.*
- Reinhardt, A.; Pickard, C. J.; Cheng, B. Predicting the phase diagram of titanium dioxide with random search and pattern recognition. *Phys. Chem. Chem. Phys.* **2020**, *22*, 12697−12705.[2] *Uses the automatic maps for crystal structure predictions.*
- Stuke, A.; Kunkel, C.; Golze, D.; Todorović, M.; Margraf, J. T.; Reuter, K.; Rinke, P.; Oberhofer, H. Atomic structures and orbital energies of 61,489 crystal-forming organic molecules. *Sci. Data* **2020**, *7*, 58.[3] *Example of emerging role of "Big Data" in molecular chemistry.*

INTRODUCTION

We are experiencing dramatic growth of data in chemistry, physics, and materials science, thanks to the ever-increasing

Received: June 24, 2020
Published: August 14, 2020



https://dx.doi.org/10.1021/acs.accounts.0c00403

computational power available, advances in electronic structure methods and algorithms, and community-wide data repositories. Exploiting the "big data" efficiently and effectively using traditional tools is not easy: data sets often contain thousands to millions of atomistic structures, along with diverse properties. Consequently, machine learning (ML) methods are increasingly employed to handle the large and complex data sets.[1,4−8] Often, data visualization is an initial and a final step of these data-driven studies. A low-dimensional map shows a condensed view of the data set and reveals underlying patterns, such as clusters, outliers, and correlations, allowing researchers to gain first insights from visual inspections.[9,10] During the final stage, visualization is essential and efficient in communicating results.

However, most of these papers focus on data generation or ML predictions while displaying visualizations without much interpretation or explanation. This is somewhat unsatisfactory, as many chemical representations and embedding methods for generating these maps are available. Furthermore, it may be unclear in what ways the maps are helpful and what kind of physical insights they provide. To fill in this gap, this Account summarizes the underlying principles of the visualization and showcases its applicability to a wide variety of physical systems.

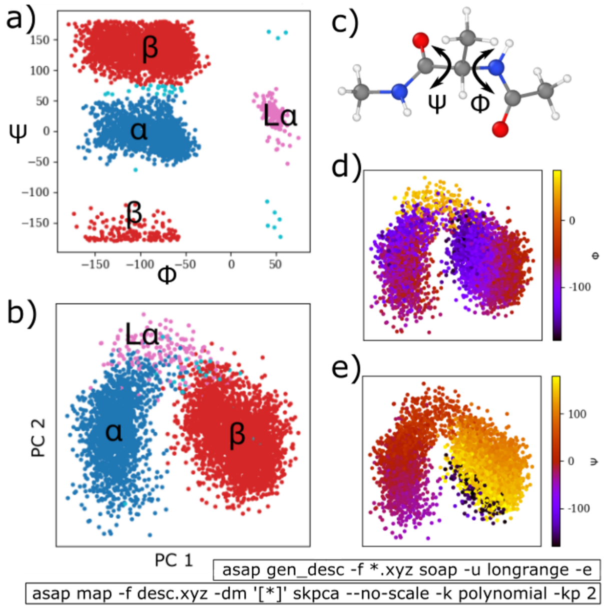
To largely automate the mapping task, we have developed user-friendly software packages: the Automatic Selection And Prediction tools for materials and molecules (ASAP) is a Python-based command-line tool that enables automatic analysis and mapping using just a couple of simple commands and options. We display such commands in snippets below when showing figures generated using the ASAP. To explore a data set interactively, we rely on a web-browser based viewing tool that can display the 3D-structure corresponding to each data point, together with its attributes.

■ ESSENTIAL CONCEPTS AND METHODS FOR MAPPING ATOMIC STRUCTURE

Low-Dimensional Embedding

The geometrical configuration of a molecule or material is intrinsically high dimensional, $3n$ for n atoms. To visualize the relationship between the structures in a data set, we need to represent each structure as a point in a low-dimensional space, typically the two dimensions of paper or a computer screen. This high ($3n$) to low dimensional transformation is called *dimensionality reduction* or *embedding*. Such embedding is common and crucial for analyzing simulation results or structural databases. Traditionally, it usually requires human insights for selecting appropriate low-dimensional coordinates, often referred to as collective variables (CVs). A textbook embedding example is the Ramachandran plot that visualizes energetically favorable regions for backbone dihedral angles (Φ and Ψ) of amino acid residues in a protein structure. The plot in Figure 1a illustrates an alanine dipeptide molecule with 66 geometric degrees of freedom using just two torsion angles. Most configurations concentrate in three distinct clusters, associated with common secondary structure elements (the α-helix, β-sheet, and left-handed α-helix).

The Ramachandran and similar plots provide powerful insight into high-dimensional structural data, but they typically require domain knowledge to hand-craft the CVs for every specific system. In contrast, automatic and system-agnostic embedding methods for atomistic structures do not rely on system-specific information. In general, embedding procedures



```
asap gen_desc -f *.xyz soap -u longrange -e
asap map -f desc.xyz -dm '[*]' skpca --no-scale -k polynomial -kp 2
```

Figure 1. (a) Ramachandran plot of 5000 configurations of an alanine dipeptide selected from a molecular dynamics trajectory,[11] with respect to the two dihedral angles indicated in panel c. The snapshots are classified according to the canonical structural motifs. (b,d,e) KPCA projections using the SOAP descriptors, colored according to classifications, Φ, and Ψ, respectively.

preserve some relationships between the points in high and low dimensional space. Loosely speaking, points that are "close" to each other in high dimension should remain so on the low-dimensional map. Embedding methods differ in the definition of "closeness", whether calculated for all points or just a subset, and in the numerical algorithms employed. A particularly simple method is *principal component analysis* (PCA), which defines closeness as the scalar product between the vectors pointing to the points.[12] Consequently, the axes of the low-dimensional map are just the first few eigenvectors of the *design matrix*, formed by concatenating the high dimensional coordinates. Alternatively, if the closeness is defined using pairwise Euclidean distances, the method is called multidimensional scaling.[13] Other definitions of closeness yield *t*-distributed stochastic neighbor embedding (t-SNE),[14] sketch-map,[15] the uniform manifold approximation and projection (UMAP),[16] etc.

Therefore, the critical first step in designing a successful embedding method for materials and molecules is to decide how to compare atomic structures, that is, by defining a distance metric. Several methods have been proposed over the past decade to describe structures, primarily for predicting atomic scale properties using machine learning.[17−24] They all respect the appropriate physical symmetries; many are based on atomic densities, and these are essentially equivalent in some limit, differing only in the basis onto which the density is projected.[25] Here we focus on the Smooth Overlap of Atomic Positions (SOAP) descriptor,[20] coupled with *kernel* PCA (KPCA),[26] which defines a scalar product in high dimensions with respect to a metric, as given by a user-supplied *kernel function*.

The computational cost of the whole process of constructing a map, computing descriptors, and then using PCA or a sparse version of KPCA as implemented in ASAP scales linearly with the total number of atoms in the data set. The workflow usually takes only a few seconds on laptops for moderately sized data sets, and less than a few minutes even for the largest set considered in this Account. To make the method suitable for even larger sets, the ASAP code is made parallelizable and contains tools to sparsify data sets (i.e., select a representative subset) as well.

Returning to the first example, an automatic mapping of the alanine dipeptide configurations using this method is shown in Figure 1b. Similarly to the standard Ramachandran plot in Figure 1a, the structures with different motifs are clearly separated on the KPCA map. (Note that in PCA or KPCA the first few eigenvectors of the design matrix, which form the axes of the plot, are also called “principal components”, PCs.) Panels d and e show the same KPCA projection but with the points colored according to Φ and Ψ. The strong horizontal color gradient in panel e suggests that PC1 is essentially equivalent to Ψ, with the additional advantage that the β cluster does not split. The vertical (PC2) axis is well correlated with the Φ angle at the top of the plot, where the Lα cluster is separated from the others. As such the KPCA map provides the same or even improved view compared with the conventional Ramachandran plot, but without relying on the prior domain knowledge.

Describing and Comparing Atomic Environments

The automatic comparison and mapping of materials and molecules starts with describing each *atomic environment*, $\mathcal{X}$, which consists of the atoms (chemical species and position) within a sphere of radius r_{cut} centered at a specific atom. A good descriptor of $\mathcal{X}$ should be invariant to translation, rotation, and permutation of atoms of the same species, because these operations do not change physical properties. Many traditional descriptors used in cheminformatics are based on the covalent connectivity of atoms, such as simple valence counting and common neighbor analysis,[27] the presence or absence of predefined atomic fragments (e.g., the Morgan fingerprints[28]), or orientational order parameters.[29] These are relatively low dimensional descriptors and lose much geometric information. We opt to retain all geometric information when representing atomic environments and structures, and then rely on the dimensionality reduction of the embedding to arrive at a low-dimensional map.

To construct SOAP descriptors, first consider an atomic environment $\mathcal{X}$ that contains only one atomic species, and place a Gaussian function of width σ centered on each atom i in $\mathcal{X}$ to make an atomic density function:

$$\rho_{\mathcal{X}_i}(\mathbf{r}) = \sum_{i\in\mathcal{X}} \exp\left[-\frac{|\mathbf{r}-\mathbf{r}_i|^2}{2\sigma^2}\right] f_{\text{cut}}(|\mathbf{r}|) \tag{1}$$

where $\mathbf{r}$ denotes a point in Cartesian space, $\mathbf{r}_i$ is the position of atom i relative to the central atom of $\mathcal{X}$, and the cutoff function, f_{cut}, smoothly decays to zero beyond the radius r_{cut}. This density representation ensures invariance with respect to translations and permutations of atoms of the same species, but not rotations. To obtain a rotationally invariant descriptor, we expand the density in a basis of spherical harmonics, $Y_{lm}(\hat{\mathbf{r}})$, and a set of orthogonal radial functions, $g_n(r)$, as

$$\rho_{\mathcal{X}}(\mathbf{r}) = \sum_{nlm} c_{nlm} g_n(|\mathbf{r}|) Y_{lm}(\hat{\mathbf{r}}) \tag{2}$$

and construct the *power spectrum* of the density using the expansion coefficients,

$$\psi_{nn'l}(\mathcal{X}) = \sqrt{\frac{8}{2l+1}} \sum_m (c_{nlm})^* c_{n'lm} \tag{3}$$

Then we obtain a vector of descriptors $\boldsymbol{\psi} = \{\psi_{nn'l}\}$ by considering all components $l \leq l_{\max}$ and $n, n' \leq n_{\max}$, which act as band limits, controlling the spatial resolution with which the atomic density is resolved. The generalization to more than one chemical species is straightforward:[4] we construct separate densities for each of n_{sp} species α, and compute power spectra $\psi_{nn'l}^{\alpha\alpha'}(\mathcal{X})$ for each pair of elements α and α', where the two species indices correspond to the c^* and c coefficients, respectively. The $n_{\text{sp}}(n_{\text{sp}}+1)/2$ vectors corresponding to each of the $\alpha-\alpha'$ pairs are then concatenated to obtain the descriptor vector of the complete environment. In some cases, we might choose to neglect the cross terms ($\alpha \neq \alpha'$) and obtain a much shorter descriptor vector. The ASAP tool uses the DScribe python library to compute the SOAP descriptors.[30]

Subsequent dimensionality reduction needs a distance metric to compare atomic environments or, equivalently, a positive semidefinite similarity kernel K (the latter should take its maximum value for a pair of identical environments and be smaller but positive for different environments). A natural similarity kernel between atomic densities is the overlap integrated over all 3D rotations, and it turns out that computing it is easy once we have the SOAP vectors,[20]

$$K(\mathcal{X}, \mathcal{X}') = \int_{\hat{R}\in\text{SO}(3)} \text{d}\hat{R} \left| \int \text{d}r \rho_{\mathcal{X}}(\mathbf{r}) \rho_{\mathcal{X}'}(\hat{R}\mathbf{r}) \right|^2 = \boldsymbol{\psi}^{\text{T}}\boldsymbol{\psi} \tag{4}$$

When considering a large number of atomic environments, we collect their descriptor vectors into a *design matrix*, $\boldsymbol{\Psi}$, whose rows are the descriptor vectors $\boldsymbol{\psi}$. For N environments, each described by a descriptor vector of length D, the design matrix has size $N \times D$. From the design matrix, we can form the *kernel matrix* of size $N \times N$, whose elements are given by the similarity kernel between each environment. The simplest linear kernel is $\mathbf{K} = \boldsymbol{\Psi}\boldsymbol{\Psi}^{\text{T}}$, for which PCA and KPCA are equivalent; other options are available.[12] A common choice together with the SOAP representation is to raise the above kernel elements to a small integer power, giving rise to a polynomial kernel. If ones needs an explicit distance between two environments, it can be defined by

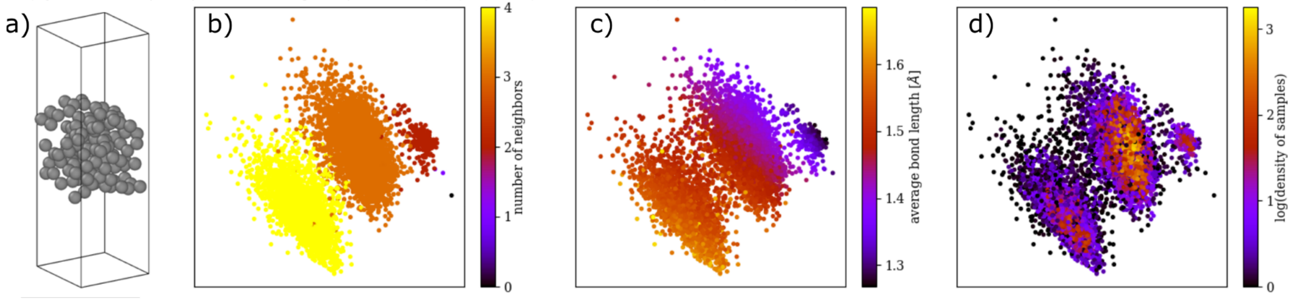
$$\begin{aligned} d(\mathcal{X}, \mathcal{X}') &= \sqrt{(\boldsymbol{\psi} - \boldsymbol{\psi}')^2} \\ &= \sqrt{K(\mathcal{X}, \mathcal{X}) + K(\mathcal{X}', \mathcal{X}') - 2K(\mathcal{X}, \mathcal{X}')} \end{aligned}$$

Notice that for nonlinear kernels, one can thus define the distance using just the kernel, bypassing explicit descriptors entirely.

Universal SOAP Hyperparameters

The length-scale hyperparameters (r_{cut} and σ) for constructing the SOAP vectors can be fine-tuned for any given application.[31] While to date this was done case by case, we have now formulated general heuristics for choosing the SOAP hyperparameters for a system with arbitrary chemical composition. The radial resolution is related to σ and $r_{\text{cut}}/n_{\max}$, and the angular resolution is determined by $2\pi/l_{\max}$, as well as σ/r at each shell of radius r. As such, using a set of fixed

https://dx.doi.org/10.1021/acs.accounts.0c00403



```
asap gen_desc -f *.xyz soap -c 2 -n 8 -l 8 -g 0.2 -pa
asap map -f desc.xyz -dm '[*]' -ua skpca -k polynomial -kp 2 --no-scale
asap kde -f desc.xyz -dm '[*]' -ua kde_internal
```

Figure 2. (a) Snapshot of an amorphous carbon thin film.[36] (b−d) KPCA projections of the atomic environments from 50 snapshots of the system with 125 carbon atoms,[36] colored according to coordination number (b), average bond length (c), and the logarithm of the relative probability of each atomic environment (d). The rightmost point is absent in panel c, as the corresponding atom has no neighbors.

hyperparameters is inefficient, because different systems have distinct length scales and varying spatial complexity. Furthermore, a system with many different chemical elements can contain a wide range of length scales, so using multiple sets of SOAP descriptors with different hyperparameters can be advantageous.[1,32]

Our universal heuristics are based on the characteristic bond lengths in the system, which in turn depend on the chemical species involved. For each atomic species Z, we calculate six structures (dimer, graphite, diamond, β-Sn, body-centered cubic, and face-centered cubic) spanning coordination from 1 to 12, minimizing the total energy with respect to uniform isotropic strain of each structure. The bond length in the lowest energy structure is defined as r^Z_{typ}, and the shortest bond length of any local minimum structure is r^Z_{min}. We then use these species-specific bond lengths to choose the SOAP hyperparameters for a given system with a set of species. The specific rules for doing this and the resulting length scales for two examples are included in the Supporting Information. In the ASAP tool, the usage of these hyperparameters is simply activated by the "−universal" or "-u" flag.

Comparing Molecules and Crystal Structures

So far we have described how to represent atomic environments. Frequently, however, we would like to represent, compare, and map *entire structures*. This requires descriptors for whole structures instead of environments. To do this, for structure A, one can combine all the descriptors for the environments $\mathcal{X}_i$ of all N_A atoms, and the most straightforward way is to simply take the average,

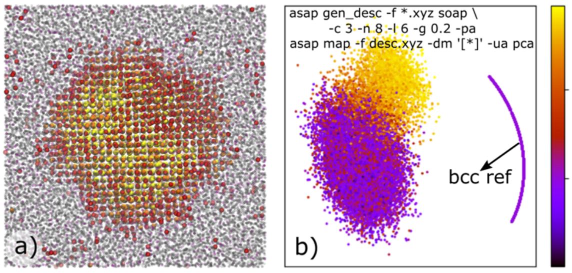
$$\Phi(A) = \frac{1}{N_{\text{A}}} \sum_{i \in A}^{N_{\text{A}}} \psi(\mathcal{X}_i) \tag{5}$$

Alternative constructions that lose less information are described elsewhere.[4,33] In the presence of multiple chemical species, one can apply a single sum or first average separately for each species and then concatenate the species-specific averaged vectors. From the descriptor vector for each structure, one can then construct the design matrix and the kernel matrix, analogously to the procedure for environments.

■ EXAMPLES

Amorphous Carbon

Here we show an example application on tetrahedral amorphous carbon (*ta*-C) films, which have intricate local environments.[34−36] The KPCA maps in Figure 2b−d show 2D projections based on the atomic SOAP descriptors of local environments in *ta*-C (illustrated in Figure 2a). Carbon atoms with different coordination numbers are automatically separated into clusters on the maps, reminiscent of the traditional classification of carbon environments as "sp", "sp^2", and "sp^3". In addition, the KPCA maps show continuous distributions of different environments within the sp and sp^2 clusters: there is significant variability in bond lengths that is strongly correlated to the vertical axis. The implication of such variability is discussed in-depth by Caro et al. in terms of reactivity (hydrogenation energy) and the classification of carbon bonds.[5] KPCA does not further separate the points within the coordination clusters as shown by the single density peak of each cluster in Figure 2d, which suggests there is no clear-cut way to subdivide the sp and sp^2 clusters.

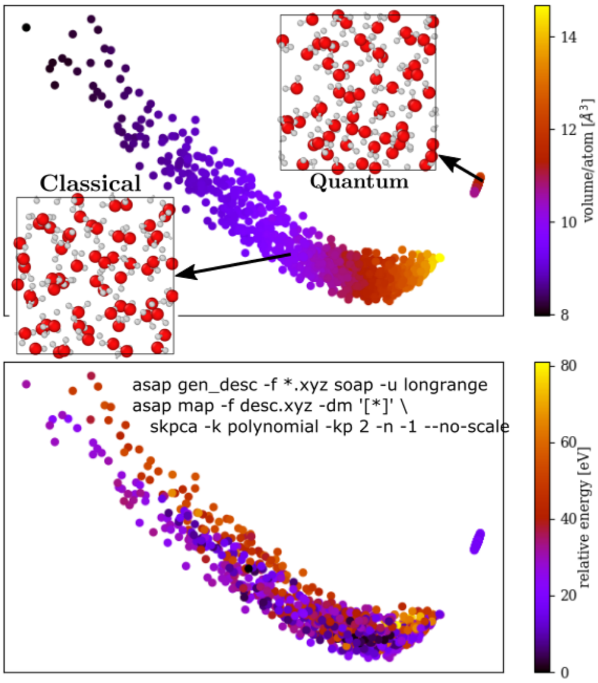


```
asap gen_desc -f *.xyz soap \
        -c 3 -n 8 -l 6 -g 0.2 -pa
asap map -f desc.xyz -dm '[*]' -ua pca
```

Figure 3. Snapshot of a Lennard-Jones system of 23 328 atoms containing a solid nucleus surrounded by undercooled liquid. Atoms are colored according to the similarity of their environment to fcc[39,40] (yellow means very similar, black and purple mean dissimilar). (a) Real space view; (b) PCA map for the atomic environments. In addition, we show the location of the bcc atomic environments (from perfect bcc crystals at a range of molar volumes) on the PCA map.

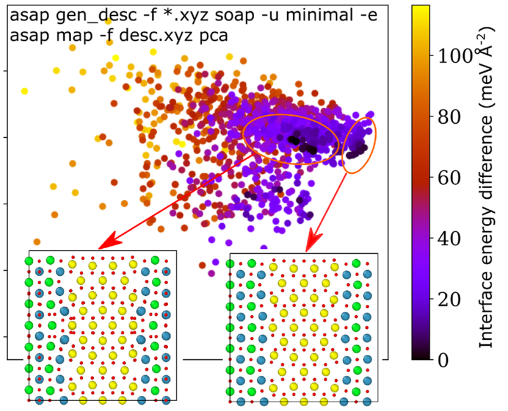
The Nucleation of a Crystal from the Liquid State

We now show the use of the automatic mapping in understanding the structural heterogeneity of nucleation. Solidification of materials starts with a small crystal nucleating from the melt. Despite a multitude of atomistic simulation studies, it is still a matter of debate whether body-centered cubic (bcc) ordering exists at the surface of the nuclei of face-centered cubic (fcc) crystals.[37] This controversy arises because the physical definition of bcc ordering is somewhat ambiguous and also because the commonly used local bond order parameters[38] do not distinguish between bcc and interface atoms.



```
asap gen_desc -f *.xyz soap -u longrange
asap map -f desc.xyz -dm '[*]' \
  skpca -k polynomial -kp 2 -n -1 --no-scale
```

Figure 4. KPCA maps of liquid water configurations (1000 classical and 593 quantum mechanical structures) from a training set,[42] colored according to volume (upper panel) and the relative energy of each configuration (lower panel).



```
asap gen_desc -f *.xyz soap -u minimal -e
asap map -f desc.xyz pca
```

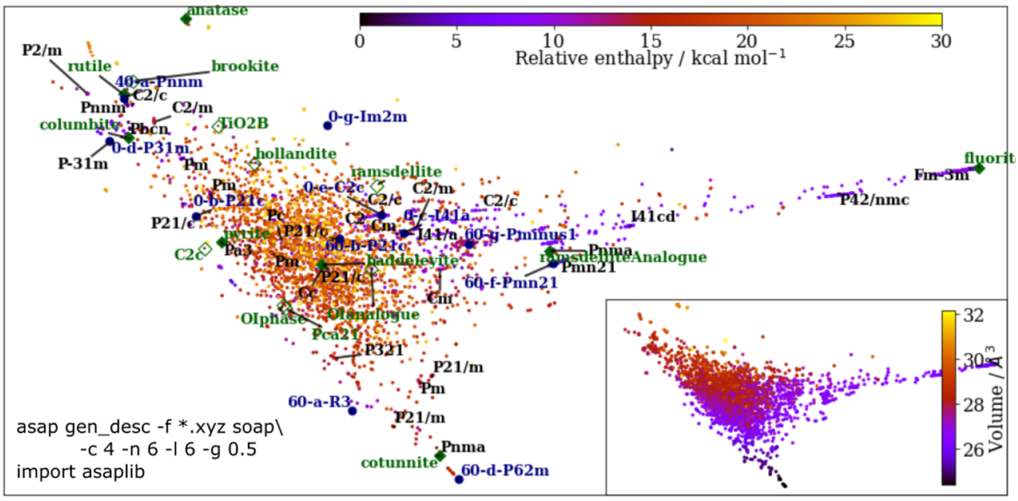
Figure 6. PCA map for 1332 STO/CeO_2 (100)/(110) interface structures relaxed from randomly generated structures.[45] These structures consist of four layers on each side of the interface with a mirror plane at the middle. The red ovals indicate two distinct lowest-energy groups.

In Figure 3, we show the PCA map based on SOAP descriptors of each atom-centered environment inside a Lennard-Jones system consisting of a solid nucleus surrounded by undercooled liquid. Environments are colored according to how similar they are to fcc using a conventional fcc order parameter that was used for enhanced sampling.[39,40] Figure 3 reveals a smooth and gradual transition between the center of the nucleus and the bulk liquid, with two blobs of data points corresponding to the fcc and liquid-like motifs. There is no clear indication of an extra density peak that is associated with the bcc local ordering. Furthermore, the reference bcc environments are clearly separated on the map. The embedding thus severely questions the existence of bcc ordering at the surface of the forming nuclei.

Liquid Water Structure

The compositions of training sets are crucial for the quality of machine learning models for chemistry and materials. Mapping atomic structures is useful for examining and understanding the training configurations, particularly when curating or expanding an existing data set. One such task is the fitting of machine learning interatomic potentials, which are increasingly popular as they can be both accurate and efficient.[41]

Here we visualize the training set of a recent potential for bulk liquid water.[42] First 1000 structures of liquid water were



```
asap gen_desc -f *.xyz soap\
        -c 4 -n 6 -l 6 -g 0.5
import asaplib
```

Figure 5. KPCA map for TiO_2 structures generated from random structure searches at 20 GPa:[2] each dot indicates a crystal structure, with the known and new phases found in ref 2 shown using blue or green markers, respectively, and annotated by their names. If a certain phase is found in the search, it is marked as a solid symbol and otherwise a hollow symbol (e.g., *C2c*, TiO_2B, and ramsdellite). Only the space groups of structures that have low energy and have appeared multiple times are indicated. The plot is generated using a Python notebook, by importing ASAP as a library.

```
asap gen_desc --no-periodic -f atoms_qm9.xyz soap -u minimal
asap map -f ASAP-desc.xyz -dm [SOAP-n4-l3-c2.4-g0.3] -c atomization_energy_eV
  skpca -k linear --no-scale --n_sparse 1000
```

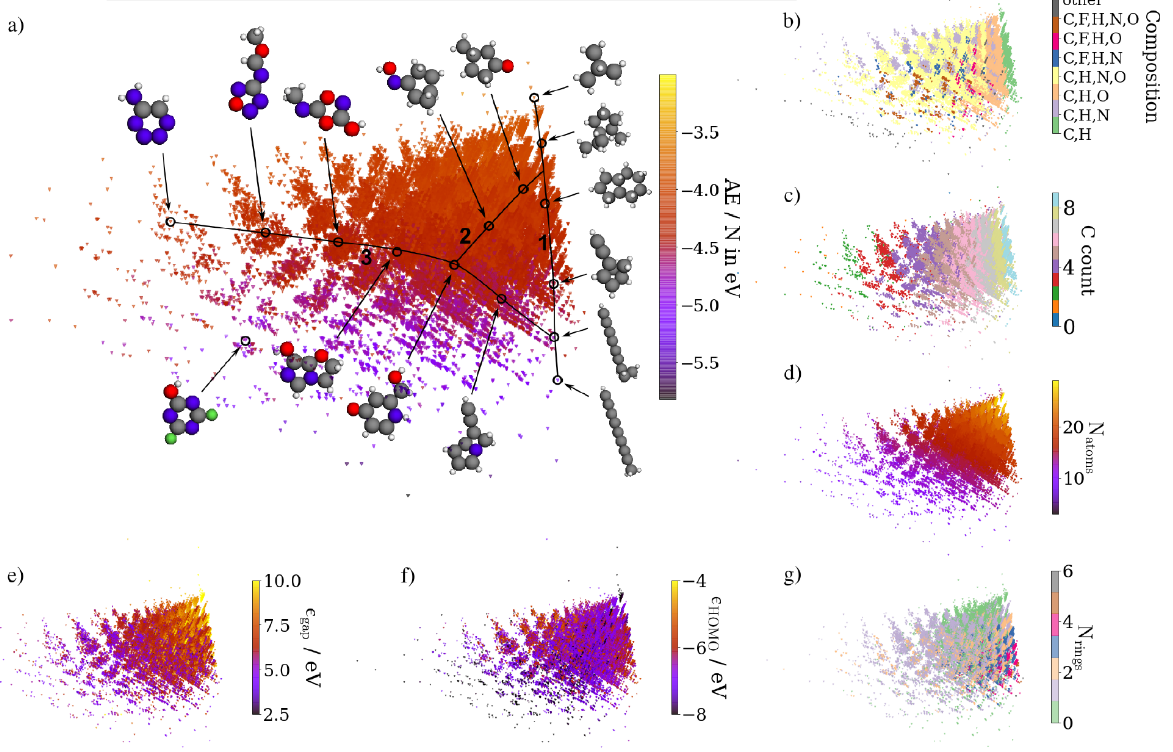


Figure 7. KPCA maps of the QM9 database using a global SOAP kernel. The frames are color-coded according to structural descriptors (b, c, d, g) and quantum mechanical properties (a, e, f).

harvested from classical molecular dynamics simulations at 1000 K and densities between 0.7 and 1.2 g/mL, and augmented with lower energy configurations obtained after a few steps of geometry optimization. The remaining configurations were extracted from path-integral molecular dynamics (PIMD) simulations at ambient pressure and 300 K, which account for the quantum mechanical nature of hydrogen nuclei. The difference between classical and quantum mechanical water is not apparent from inspecting atomic snapshots by eye, cannot be captured using conventional metrics such as oxygen radial distribution functions,[42] and has only subtle manifestation in hydrogen bond analysis.[43]

However, in the KPCA maps of the training set (Figure 4), the distinction is obvious: the classical and quantum water form two well-separated clusters. It is further revealed that the classical water configurations have a relatively wide spread in both energy and molar volume, and both quantities are correlated with the axes of the plot. Such spread in the training set is important for constructing a potential that is stable at a range of pressures and elevated temperatures.

Crystal Structure Search: Titanium Dioxide

Ab initio random structure search[44] is a very productive tool of materials discovery. To demonstrate the use of visualization in this domain, we show an example of mapping the TiO_2 crystalline polymorphs[2] that were produced from random searches.[44] This data set includes thousands of distinct TiO_2 structures with different atomic coordinates, cell shapes, and numbers of formula units in the cell. Even though the knowledge of space groups, molar volumes, and energies of the structures provides hints on how to classify them, it is still a formidable task to sort through them manually. The KPCA map in Figure 5 instead directly gives an overview of the structural similarities between 4690 locally stable structures of titanium dioxide. Properties such as the relative enthalpy or unit cell volume vary smoothly across the figure, and regions of high density or stability are revealed. We project the known (marked in blue) and newly discovered phases (marked in green) of TiO_2 on the map,[2] so one can immediately spot if a particular phase has been found in the random search, instead of having to rely on the traditional identifications such as the space groups. Indeed, as also shown on the map, different structures can adopt an identical space group, while atomic configurations that are structurally similar were classified to have distinct symmetries.

Structure of Heterogeneous Interfaces

Structure searches can be extended to systems with interfaces to reveal the stable configurations that are hard to obtain otherwise.[45] The data analysis for this is even more challenging compared with bulk phases, because the presence of the interface breaks the crystallographic symmetry, so the traditional space group analysis is often ineffective. The extended, often low-symmetry nature of interfaces also makes visual inspections more difficult. Hence, automatic maps become extremely desirable in this case.

Figure 6 shows the PCA map of $SrTiO_3$ and CeO_2 (STO/CeO_2) (100)/(110) interface structures. Each point represents a configuration at a local energy minimum. The relative energies, used as the color scale, strongly correlate with the

```
asap gen_desc -f *.xyz --no-periodic soap -u minimal -pa
asap map -f ASAP-desc.xyz -dm '[*]' -ua --only_use_species 6 pca --no-scale
```

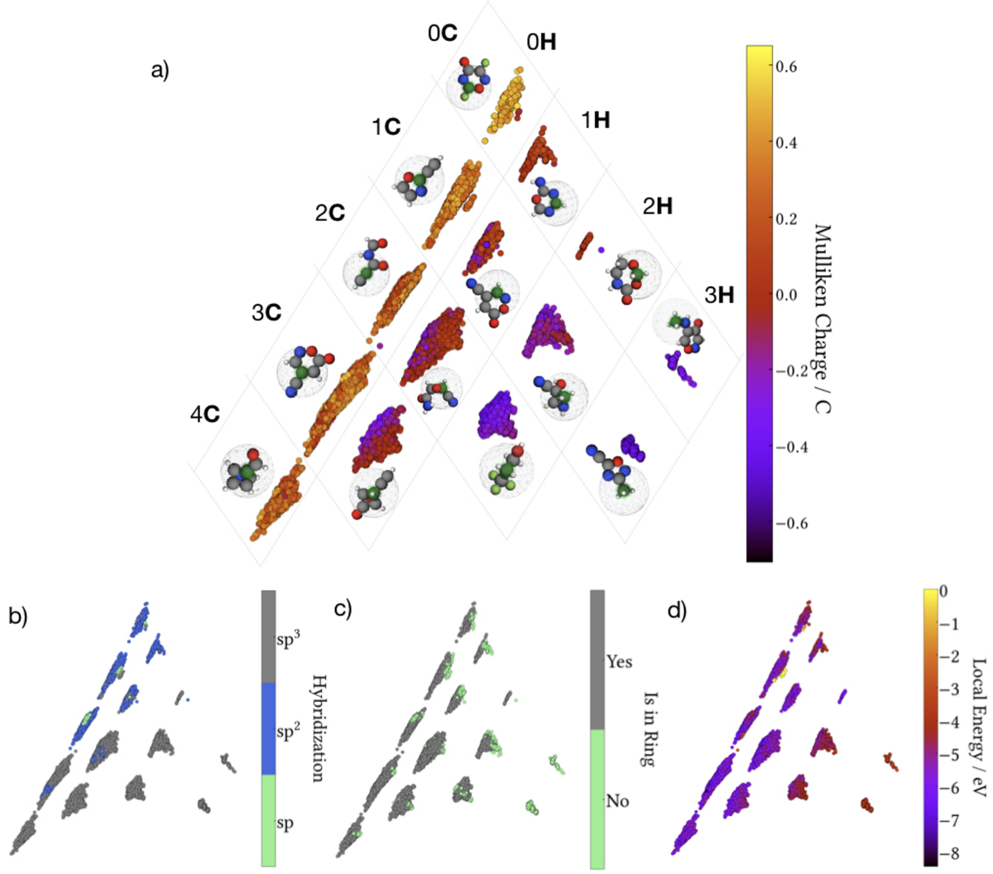


Figure 8. KPCA maps of carbon atom environments in the QM9 database. Maps are color-coded according to Mulliken charges (a), hybridization (b), whether the atoms are in rings (c), and local energies predicted by a machine learning potential (d).

horizontal axis of the map. This means that while the interfacial energies are not used to construct the map, PCA identifies them automatically, presumably just from the distortion of the interface regions. We identified two clusters with low energies: group A consists of structures similar to the ideal interface that forms by simply joining the bulk phases, while group B contains the reconstructed structures.

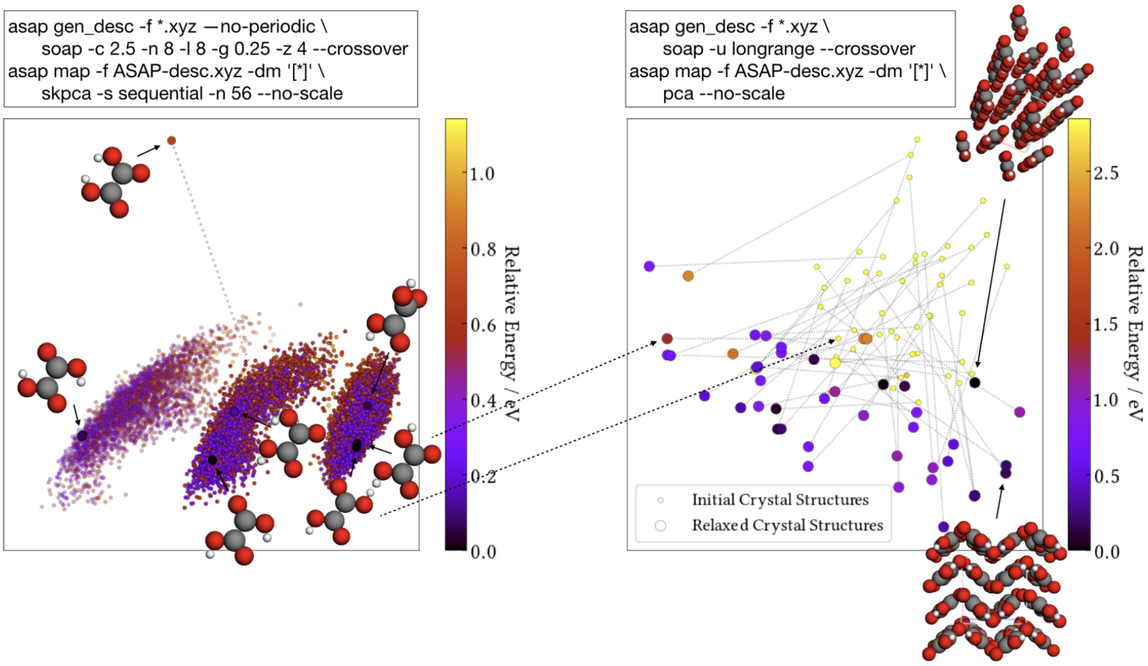
Organic Molecules

The QM9 data set,[46] which contains 133 885 organic molecules composed of H and up to nine heavy atoms (C, N, O, and F), has become a standard benchmark for ML-based property prediction. Here we compare molecular structures using average SOAP descriptors (eq 5) and then use a sparse version of KPCA for dimensionality reduction as the data set is large. We use the resulting map (Figure 7) to navigate the QM9 set and exploit the interactive viewer to observe molecules along various "paths" through the map (illustrated in Figure 7a).

Color-coding the points on the map using elemental compositions (Figure 7b) shows that pure hydrocarbons and other compositions (e.g., C, H, O or C, H, N, O) form separate clusters. Together panels c and d of Figure 7 show that different carbon and total atom counts cause further splitting of these clusters. The key features of the map are thus mainly defined by molecular composition. Furthermore, systems with different numbers of rings also form distinct clusters across the map (Figure 7g).

Molecular properties correlate both with the axes of the map and with the molecular compositions. The atomization energy per atom (Figure 7a) scales inversely with the total number of atoms (Figure 7d).[4,47] The reason is that most molecules in QM9 contain 9 non-hydrogen atoms, so molecules with fewer overall atoms tend to have more double and triple bonds. This also explains the trend in the HOMO−LUMO gap, ϵ_{gap} (Figure 7e): unsaturated compounds tend to have lower gaps. On the other hand, HOMO energies, ϵ_{HOMO} (Figure 7e), are less systematic, presumably because the electronegativities of the contained elements and structural features like π-conjugation have a strong influence.

As a complementary way to visualize QM9, we consider the atomic environments of all the carbon atoms (Figure 8). Upon inspection, the clusters of environments are found to reflect different numbers of neighboring carbon and hydrogen atoms (strongly correlated with the vertical and horizontal axes of the plot, respectively). The clusters thus correspond to atom-types, reflecting the fundamental concept behind classical bio-organic



```
asap gen_desc -f *.xyz —no-periodic \
     soap -c 2.5 -n 8 -l 8 -g 0.25 -z 4 --crossover
asap map -f ASAP-desc.xyz -dm '[*]' \
     skpca -s sequential -n 56 --no-scale
```

```
asap gen_desc -f *.xyz \
     soap -u longrange --crossover
asap map -f ASAP-desc.xyz -dm '[*]' \
     pca --no-scale
```

Figure 9. (left) KPCA map of oxalic acid conformers in the gas-phase (large points) and configurations from different MD simulations at 500 K initialized at the conformer geometries (small points). Configurations belonging to the MD for which no transitions to other basins are observed are shown as transparent points. (right) Randomly generated oxalic acid unit cells (small yellow circles) and the corresponding fully relaxed crystals (large colored circles). The experimentally known α[48] (lower structure) and β[49] (upper structure) polymorphs are highlighted. All random structures were initialized from the same gas-phase conformer, but in some cases, the conformer changed upon relaxation (highlighted by arrows across the panels).

force-fields, which define different atom-types according to the basic bonding topology of a molecule. Each cluster displays a fairly homogeneous Mulliken charge, with a large number of hydrogen neighbors leading to a negative partial charge on the carbon atom (and vice versa). Within each cluster, different realizations of the C/H neighborhoods cause further difference in the charges. For example, both a carboxylic acid and a $-CF_3$ group attached to a hydrocarbon contain a central carbon atom with a single carbon and no hydrogen neighbors. Such subclusters are illustrated in Figure 8b,c according to the hybridization and whether a carbon atom is part of a ring. This also serves as a warning that dimensionality reduction may obscure some relevant structural features of the data. In this case, carbon and hydrogen are the most abundant elements in the data set so they dominate the embedding, whereas the role of heteroatoms is not immediately clear. Color-coding using additional properties and inspecting representative structures can fill this gap.

Visualizing atomic environments also helps understand and interpret ML potentials. We consider a SOAP-based GAP model trained on QM9 energies,[1] in which total energies are expressed as the sums of local atomic energies. Figure 8d is color-coded using these local energies and shows systematic trends of similar energies within each cluster and a smooth variation of energies between clusters. This shows how the local energies are related to the specific environments.

Polymorphs and Conformers of Oxalic Acid Crystals

KPCA maps can also be used to emphasize differences in conformations or crystal polymorphs for systems with fixed composition. This is illustrated for oxalic acid (OA) in Figure 9. In the left panel, seven conformers of OA (each a (meta-)stable structure on the potential energy surface) are shown. The map intuitively arranges these structures according to the orientation of the protons (from left to right, in–in, in–out, and out–out), and its vertical axis correlates with the relative energy.

Additionally, configurations sampled from a series of MD trajectories initialized from each conformer geometry are shown. Note that the highest energy conformer is not thermally stable and almost immediately rearranges during the MD. These configurations are arranged in larger basins separated by energetic barriers, while conformers within a basin are connected by low energy paths. In particular, the leftmost conformer (with corresponding points indicated by partial transparency) does not rearrange during the MD simulation due to the high kinetic stability afforded by the two intramolecular hydrogen bonds, whereas all other conformers are connected by the MD trajectories.

In the right panel, a similar KPCA plot is shown for 48 bulk crystal structures of OA, which were generated using random structure search. The initial random structures (small yellow circles) and the corresponding fully relaxed configurations (large colored circles) are connected by gray lines. All molecules in the structure search were initialized from the bottom-right conformer (out–out, trans) in the left panel, but in some cases the monomers in the relaxed structures belong to a different conformer (indicated by dashed arrows between the two panels). The random and optimized crystal structures stay on distinct regions of the plot. Almost all optimized structures are well-separated, indicating that there are many stable minima for the crystal, unlike in the gas-phase. Besides the two experimentally formed polymorphs, several other crystalline structures of OA with comparable energies are also found. This multitude of low energy local minima makes organic crystal structure prediction difficult.

The two maps in Figure 9 highlight different aspects of molecular structure: the intramolecular aspects (mainly proton orientation) on the left, and the differences in intermolecular interactions on the right. Considering both thus allows a more complete understanding of the structural factors underpinning molecular crystal formation.

■ CONCLUSION

Automating the mapping of diverse classes of materials and molecules yields physical and chemical insights, saves human effort, and provides a data-driven perspective on large atomistic data sets. Because of this utility and the software packages that are now available, we believe that these maps will become a standard tool for the wider computational chemistry and materials science community. From the perspective of methodology, there is certainly room for improvement. For example, a systematic comparison of the maps produced using different descriptors and dimensionality reduction algorithms would be useful, as would be the development of new schemes that have better scaling with respect to the number of atomic species in the data set. As in most works using ML for chemistry and materials to date, we have neglected long-range interactions and correlations. Incorporating descriptions of these may improve the ability of maps to discern and tease out such effects, for example, in ionic solutions and large protein complexes.

All in all, beyond visualization being a valuable tool for molecular modeling, it is also becoming a science in itself. Without any doubt, this Account will not be the final word in this new science, and we anticipate exciting new developments.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.accounts.0c00403.

Details on the universal SOAP heuristics (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Bingqing Cheng − *Department of Chemistry, University of Cambridge, Cambridge CB2 1EW, United Kingdom;* orcid.org/0000-0002-3584-9632; Email: bc509@cam.ac.uk

Authors

Ryan-Rhys Griffiths − *Cavendish Laboratory, University of Cambridge, Cambridge CB3 0HE, United Kingdom*

Simon Wengert − *Chair for Theoretical Chemistry and Catalysis Research Center, Technische Universität München, D-85747 Garching, Germany*

Christian Kunkel − *Chair for Theoretical Chemistry and Catalysis Research Center, Technische Universität München, D-85747 Garching, Germany*

Tamas Stenczel − *Engineering Laboratory, University of Cambridge, Cambridge CB2 1PZ, United Kingdom*

Bonan Zhu − *Department of Materials Science and Metallurgy, University of Cambridge, Cambridge CB3 0FS, United Kingdom*

Volker L. Deringer − *Department of Chemistry, Inorganic Chemistry Laboratory, University of Oxford, Oxford OX1 3QR, United Kingdom;* orcid.org/0000-0001-6873-0278

Noam Bernstein − *Center for Materials Physics and Technology, U.S. Naval Research Laboratory, Washington, D.C. 20375, United States*

Johannes T. Margraf − *Chair for Theoretical Chemistry and Catalysis Research Center, Technische Universität München, D-85747 Garching, Germany*

Karsten Reuter − *Chair for Theoretical Chemistry and Catalysis Research Center, Technische Universität München, D-85747 Garching, Germany;* orcid.org/0000-0001-8473-8659

Gabor Csanyi − *Engineering Laboratory, University of Cambridge, Cambridge CB2 1PZ, United Kingdom*

Complete contact information is available at:
https://pubs.acs.org/10.1021/acs.accounts.0c00403

Notes

The authors declare no competing financial interest.
The data sets and scripts for the visualization are uploaded to a public repository at https://github.com/BingqingCheng/Mapping-the-space-of-materials-and-molecules. The ASAP code and the interactive viewing tool are also available: https://github.com/BingqingCheng/ASAP; https://github.com/chkunkel/projection_viewer. An alternative interactive viewing tool[50] developed by another research group is at https://chemiscope.org. The output of ASAP can be directly used as the input of either viewing tool.

Biographies

Bingqing Cheng is a junior research fellow in Cambridge, and her work focuses on theoretical predictions of material properties.

Ryan-Rhys Griffiths is a Ph.D. student in Cambridge, working on machine learning methodology for scientific applications.

Simon Wengert is a Ph.D. student at TU Munich, and his work focuses on machine-learning assisted crystal structure prediction.

Christian Kunkel is a Ph.D. student at TU Munich, working on machine-learning based organic materials discovery.

Tamás K. Stenczel is a student at the University of Cambridge, and his research work focuses on machine-learning modeling of reactive chemical systems.

Bonan Zhu is a Ph.D. student at the University of Cambridge, and his research work focuses on predicting and modeling oxide interfaces.

Volker Deringer is Associate Professor of Theoretical and Computational Inorganic Chemistry at the University of Oxford.

Noam Bernstein is a research physicist the U.S. Naval Research Laboratory.

Johannes T. Margraf is a research group leader at TU Munich, working on molecular machine learning.

Karsten Reuter is Professor of Theoretical Chemistry at TU Munich.

Gábor Csányi is Professor of Molecular Modelling at the Engineering Laboratory, University of Cambridge.

■ ACKNOWLEDGMENTS

B.C. acknowledges funding from Swiss National Science Foundation (Project P2ELP2-184408). V.L.D. acknowledges a Leverhulme Early Career Fellowship. S.W. acknowledges funding from Deutsche Forschungsgemeinschaft (DFG), and

J.M. and C.K. are grateful for support by DFG through TUM International Graduate School of Science and Engineering (IGSSE), GSC 81. The work of N.B. was supported by the U.S. Office of Naval Research through the U.S. Naval Research Laboratory's fundamental research base program.

■ REFERENCES

(1) Bartok, A. P.; De, S.; Poelking, C.; Bernstein, N.; Kermode, J.; Csanyi, G.; Ceriotti, M. Machine Learning Unifies the Modelling of Materials and Molecules. *Science Advances* **2017**, *3*, No. e1701816.

(2) Reinhardt, A.; Pickard, C. J.; Cheng, B. Predicting the phase diagram of titanium dioxide with random search and pattern recognition. *Phys. Chem. Chem. Phys.* **2020**, *22*, 12697−12705.

(3) Stuke, A.; Kunkel, C.; Golze, D.; Todorović, M.; Margraf, J. T.; Reuter, K.; Rinke, P.; Oberhofer, H. Atomic structures and orbital energies of 61,489 crystal-forming organic molecules. *Sci. Data* **2020**, *7*, 58.

(4) De, S.; Bartók, A. P.; Csányi, G.; Ceriotti, M. Comparing molecules and solids across structural and alchemical space. *Phys. Chem. Chem. Phys.* **2016**, *18*, 13754−13769.

(5) Caro, M. A.; Aarva, A.; Deringer, V. L.; Csanyi, G.; Laurila, T. Reactivity of amorphous carbon surfaces: rationalizing the role of structural motifs in functionalization using machine learning. *Chem. Mater.* **2018**, *30*, 7446−7455.

(6) Butler, K. T.; Davies, D. W.; Cartwright, H.; Isayev, O.; Walsh, A. Machine learning for molecular and materials science. *Nature* **2018**, *559*, 547−555.

(7) Bernstein, N.; Csányi, G.; Deringer, V. L. De novo exploration and self-guided learning of potential-energy surfaces. *npj Computational Materials* **2019**, *5*, 99.

(8) Huang, J.-X.; Csányi, G.; Zhao, J.-B.; Cheng, J.; Deringer, V. L. First-principles study of alkali-metal intercalation in disordered carbon anode materials. *J. Mater. Chem. A* **2019**, *7*, 19070−19080.

(9) Isayev, O.; Fourches, D.; Muratov, E. N.; Oses, C.; Rasch, K.; Tropsha, A.; Curtarolo, S. Materials Cartography: Representing and Mining Materials Space Using Structural and Electronic Fingerprints. *Chem. Mater.* **2015**, *27*, 735−743.

(10) Ceriotti, M. Unsupervised Machine Learning in Atomistic Simulations, between Predictions and Understanding. *J. Chem. Phys.* **2019**, *150*, 150901.

(11) Nüske, F.; Wu, H.; Prinz, J.-H.; Wehmeyer, C.; Clementi, C.; Noé, F. Markov state models from short non-equilibrium simulations—Analysis and correction of estimation bias. *J. Chem. Phys.* **2017**, *146*, 094104.

(12) Friedman, J.; Hastie, T.; Tibshirani, R. *The elements of statistical learning*; Springer series in statistics; Springer: New York, 2001; Vol. *1*.

(13) Cox, M. A.; Cox, T. F. *Handbook of data visualization*; Springer, 2008; pp 315−347.

(14) Maaten, L. v. d.; Hinton, G. Visualizing data using t-SNE. *Journal of machine learning research* **2008**, *9*, 2579−2605.

(15) Ceriotti, M.; Tribello, G. A.; Parrinello, M. Simplifying the representation of complex free-energy landscapes using sketch-map. *Proc. Natl. Acad. Sci. U. S. A.* **2011**, *108*, 13023−13028.

(16) McInnes, L.; Healy, J.; Melville, J. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint* arXiv:1802.03426 2018, https://arxiv.org/abs/1802.03426.

(17) Behler, J. Atom-centered symmetry functions for constructing high-dimensional neural network potentials. *J. Chem. Phys.* **2011**, *134*, 074106.

(18) Huo, H.; Rupp, M. Unified representation for machine learning of molecules and crystals. *arXiv preprint* arXiv:1704.06439 2017, https://arxiv.org/abs/1704.06439.

(19) Bartók, A. P.; Payne, M. C.; Kondor, R.; Csányi, G. Gaussian Approximation Potentials: The Accuracy of Quantum Mechanics, without the Electrons. *Phys. Rev. Lett.* **2010**, *104*, 136403.

(20) Bartók, A. P.; Kondor, R.; Csányi, G. On representing chemical environments. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2013**, *87*, 184115.

(21) Faber, F. A.; Christensen, A. S.; Huang, B.; Von Lilienfeld, O. A. Alchemical and structural distribution based representation for universal quantum machine learning. *J. Chem. Phys.* **2018**, *148*, 241717.

(22) Sadeghi, A.; Ghasemi, S. A.; Schaefer, B.; Mohr, S.; Lill, M. A.; Goedecker, S. Metrics for measuring distances in configuration spaces. *J. Chem. Phys.* **2013**, *139*, 184118.

(23) Barker, J.; Bulin, J.; Hamaekers, J.; Mathias, S. *Scientific Computing and Algorithms in Industrial Simulations*; Springer, 2017; pp 25−42.

(24) Hansen, K.; Biegler, F.; Ramakrishnan, R.; Pronobis, W.; Von Lilienfeld, O. A.; Muller, K.-R.; Tkatchenko, A. Machine learning predictions of molecular properties: Accurate many-body potentials and nonlocality in chemical space. *J. Phys. Chem. Lett.* **2015**, *6*, 2326−2331.

(25) Willatt, M. J.; Musil, F.; Ceriotti, M. Atom-density representations for machine learning. *J. Chem. Phys.* **2019**, *150*, 154110.

(26) Schölkopf, B.; Smola, A.; Müller, K.-R. Nonlinear Component Analysis as a Kernel Eigenvalue Problem. *Neural Computation* **1998**, *10*, 1299−1319.

(27) Tsuzuki, H.; Branicio, P. S.; Rino, J. P. Structural characterization of deformed crystals by analysis of common atomic neighborhood. *Comput. Phys. Commun.* **2007**, *177*, 518−523.

(28) Rogers, D.; Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742−754.

(29) Lechner, W.; Dellago, C. Accurate determination of crystal structures based on averaged local bond order parameters. *J. Chem. Phys.* **2008**, *129*, 114707.

(30) Himanen, L.; Jäger, M. O. J.; Morooka, E. V.; Federici Canova, F.; Ranawat, Y. S.; Gao, D. Z.; Rinke, P.; Foster, A. S. DScribe: Library of Descriptors for Machine Learning in Materials Science. *arXiv preprints* arXiv:1904.08875, 2019, https://arxiv.org/abs/1904.08875.

(31) Deringer, V. L.; Pickard, C. J.; Csányi, G. Data-Driven Learning of Total and Local Energies in Elemental Boron. *Phys. Rev. Lett.* **2018**, *120*, 156001.

(32) Bernstein, N.; Bhattarai, B.; Csanyi, G.; Drabold, D. A.; Elliott, S. R.; Deringer, V. L. Quantifying Chemical Structure and Machine-Learned Atomic Energies in Amorphous and Liquid Silicon. *Angew. Chem., Int. Ed.* **2019**, *58*, 7057−7061.

(33) Mavracic, J.; Mocanu, F. C.; Deringer, V. L.; Csanyi, G.; Elliott, S. R. Similarity Between Amorphous and Crystalline Phases: The Case of TiO2. *J. Phys. Chem. Lett.* **2018**, *9*, 2985−2990.

(34) Deringer, V. L.; Csányi, G. Machine learning based interatomic potential for amorphous carbon. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2017**, *95*, 094203.

(35) Caro, M. A.; Deringer, V. L.; Koskinen, J.; Laurila, T.; Csányi, G. Growth Mechanism and Origin of High sp^3 Content in Tetrahedral Amorphous Carbon. *Phys. Rev. Lett.* **2018**, *120*, 166101.

(36) Deringer, V. L.; Caro, M. A.; Jana, R.; Aarva, A.; Elliott, S. R.; Laurila, T.; Csanyi, G.; Pastewka, L. Computational Surface Chemistry of Tetrahedral Amorphous Carbon by Combining Machine Learning and Density Functional Theory. *Chem. Mater.* **2018**, *30*, 7438−7445.

(37) Ten Wolde, P. R.; Ruiz-Montero, M. J.; Frenkel, D. Numerical evidence for bcc ordering at the surface of a critical fcc nucleus. *Phys. Rev. Lett.* **1995**, *75*, 2714.

(38) Lechner, W.; Dellago, C.; Bolhuis, P. G. Role of the prestructured surface cloud in crystal nucleation. *Phys. Rev. Lett.* **2011**, *106*, 085701.

(39) Cheng, B.; Tribello, G. A.; Ceriotti, M. Solid-liquid interfacial free energy out of equilibrium. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2015**, *92*, 180102.

(40) Cheng, B.; Ceriotti, M. Bridging the gap between atomistic and macroscopic models of homogeneous nucleation. *J. Chem. Phys.* **2017**, *146*, 034106.

(41) Deringer, V. L.; Caro, M. A.; Csányi, G. Machine Learning Interatomic Potentials as Emerging Tools for Materials Science. *Adv. Mater.* **2019**, *31*, 1902765.

(42) Cheng, B.; Engel, E. A.; Behler, J.; Dellago, C.; Ceriotti, M. Ab initio thermodynamics of liquid and solid water. *Proc. Natl. Acad. Sci. U. S. A.* **2019**, *116*, 1110−1115.

(43) Wang, L.; Ceriotti, M.; Markland, T. E. Quantum fluctuations and isotope effects in ab initio descriptions of water. *J. Chem. Phys.* **2014**, *141*, 104502.

(44) Pickard, C. J.; Needs, R. Ab initio random structure searching. *J. Phys.: Condens. Matter* **2011**, *23*, 053201.

(45) Zhu, B.; Schusteritsch, G.; Lu, P.; MacManus-Driscoll, J. L.; Pickard, C. J. Determining Interface Structures in Vertically Aligned Nanocomposite Films. *APL Mater.* **2019**, *7*, 061105.

(46) Ramakrishnan, R.; Dral, P. O.; Rupp, M.; von Lilienfeld, O. A. Quantum chemistry structures and properties of 134 kilo molecules. *Sci. Data* **2014**, *1*, 140022.

(47) Jung, H.; Stocker, S.; Kunkel, C.; Oberhofer, H.; Han, B.; Reuter, K.; Margraf, J. T. Size-Extensive Molecular Machine Learning with Global Representations. *ChemSystemsChem.* **2020**, *2*, e1900052.

(48) Thalladi, V. R.; Nüsse, M.; Boese, R. The Melting Point Alternation in α,ω-Alkanedicarboxylic Acids. *J. Am. Chem. Soc.* **2000**, *122*, 9227−9236.

(49) Derissen, J. L.; Smith, P. H. Refinement of the crystal structures of anhydrous α- and β-oxalic acids. *Acta Crystallogr., Sect. B: Struct. Crystallogr. Cryst. Chem.* **1974**, *30*, 2240−2242.

(50) Fraux, G.; Cersonsky, R.; Ceriotti, M. Chemiscope: Interactive Structure-Property Explorer for Materials and Molecules. *Journal of Open Source Software* **2020**, *5*, 2117.