EPFL

Thèse 7977

Characterising Structure and Stability of Materials using Machine Learning

Présentée le 6 novembre 2020

à la Faculté des sciences et techniques de l'ingénieur
Laboratoire de science computationnelle et modélisation
Programme doctoral en science et génie des matériaux

pour l'obtention du grade de Docteur ès Sciences

par

Andrea ANELLI

Acceptée sur proposition du jury

Prof. C. Hébert, présidente du jury
Prof. M. Ceriotti, directeur de thèse
Prof. F. Pietrucci, rapporteur
Prof. G. Day, rapporteur
Prof. N. Marzari, rapporteur

École polytechnique fédérale de Lausanne

2020

Acknowledgements

First and foremost, I would like to express my deepest gratitude to my thesis director, prof. Michele Ceriotti. His natural ability in blending attention to details and long term vision will always be a source of inspiration on how to carry out scientific research capable of pushing the boundaries of state of the art.
I want to dedicate a special thank to Dr Edgar Engel, friend and colleague whose enthusiasm and unparalleled productivity has helped in transforming many ambitious targets into proud achievements.
The laboratory of computational science and modelling (COSMO) has been my academic family for the past four years, growing from a handful of members into an almost thirty people group. Each of the group members who have gone through the third floor of the MXG building has taught me something valuable, sparking scientific discussions and collaborations that continue well beyond the breadth of the PhD experience. Among the members of the group, I would like to express my gratitude to Edoardo, Gabriele, Daniele, Venkat and Piero for being my first points of contact with the small world of Lausanne, the science of COSMO, and managed to create a homely feeling since the very beginning.
Embarking on a PhD is a long journey, which I was lucky enough to have started with Giulio, Felix, Andrea, a smaller subgroup in COSMO with which I have shared projects, hopes and achievements- to them I want to express a sincere "Thank you" for having coped with me for all these years.
A special thanks to Kevin and Federico, for their unwavering support and mentoring throughout all the thesis work, and their impossibly low activation barrier in helping realise projects stemming from our daily discussions. Finally, I would like to thank Natasha, Chiheb, Jigyasa for their impressive support, advice and interactions outside of the laboratory.
Beyond my academic microcosm(o), I would like to thank Anne Roy, for making the group hospitable, greener and feel like a (swiss) family.
I want to express my exceptional appreciation to Alexandra, for her infinite love, support and patience. You are a source of strength and have helped make every insurmountable obstacle look smaller and manageable.
Finally, as a constant source of positivity and support, I would like to express my gratitude to my family.
Grazie a tutti, è stato divertente.

Lausanne, 5 June 2020 A. A.

Abstract

The search of novel materials using in-silico high-throughput screening is emerging as a fundamental step in the pipeline of materials discovery, but its low yields in terms of synthesisable structures limit its effectiveness.
In order to isolate configurations that show promise to be stable at experimental conditions, scientists have traditionally relied on a convex hull construction, which finds the phases that are favourable under known thermodynamic boundary conditions (i.e. pressure, compositions). While this scheme is robust and computationally inexpensive, it is severely limited: it can only isolate phases which exhibit stability to well known thermodynamic variables, and does not provide a systematic way of dealing with the inevitable uncertainties in the computationally determined formation energies of the candidate phases, or with the presence of multiple similar phases that only differ because of disorder or defects.
In this thesis, we introduce a novel thermodynamic framework which aims at enhancing the exploratory power of structure searches by generalising the convex hull construction (GCH). This scheme relies on a set of data-driven, structural descriptors which correlate with the systems' responses to external thermodynamic conditions. Moreover, it is probabilistic, thus offering a way to assign a probability to every configuration of becoming stabilisable by application of a given thermodynamic constraint, rather than classifying in a binary manner synthesisable and non-synthesisable phases.
We first benchmark the predictive power of this scheme in a series of increasingly complex structure searches tasks, ranging from high-pressure hydrogen crystals to detection of molecular crystals stabilisation through chemical substitution. We then show how the probabilistic character of our proposed scheme increases the robustness of the convex hull constructions to differences in the choice of potential energy surfaces and atomistic description of the system. Finally, we show the results of a fully explorative search on the phase diagram of ice. The GCH constructions re-discovers all the known 17 phases included in the set and further proposes a set of candidates for meta-stability. To test the novel phases stability range we analyse their enthalpies at various pressures, finding regimes where novel candidates appear to be favourable over the already known phases.
The results presented in this thesis show that the GCH can effectively enhance structure searches schemes with a robust and far-sighted probe for potentially synthesisable configuration.

Key words: Machine learning, Thermodynamics, High-throughput screening, DFT, Crystal

structure prediction, Generalised convex hull, Molecular Crystals

Abstract

La ricerca di nuovi materiali attraverso simulazioni high-throughput è un approccio emergente nella scoperta di nuovi materiali, ma la sua efficacia è minata dallo scarsa capacità di individuare strutture effettivamente sintetizzabili.
L'individuazione di configurazioni che dimostrano un alto potenziale di stabilizzazione sperimentale si basa tipicamente sull'utilizzo di convex hulls, capaci di trovare le fasi termodinamicamente favorite sotto specifiche condizioni esterne.
Sebbene questo schema sia robusto ed efficiente in termini di costo computazionale, non è esente da difetti fondamentali: può solo isolare le fasi stabilizzate da condizioni termodinamiche note a priori, e non contiene alcun modo sistematico di considerare le inevitabili incertezze contenute nei cacoli delle energie di formazione delle fasi sotto screening.
In questa tesi introduciamo un nuovo approccio che mira ad aumentare il potere esplorativo delle ricerche di strutture generalizzando la costruzione del convex hull. Questo schema si basa su un insieme di descrittori strutturali estratti automaticamente dai dati, che possono correlare con qualsiasi meccanismo di stabilizzazione che influisca sulle configurazioni atomiche delle fasi.
Inoltre, ha una natura probabilistica, e offre un modo per assegnare una probabilità ad ogni configurazione di risultare stabilizzabile sotto l'applicazione un determinato vincolo termodinamico, piuttosto che fornire una classificazione binaria fra fasi stabilizzabili e non stabilizzabili. Il potere predittivo di questo schema viene inizialmente testato in una serie di ricerche strutturali di complessità incrementale, partendo da idrogeno cristallino ad alta pressione, fino alla caratterizzazione della stabilizzazione di cristalli molecolari attraverso sostituzioni di specie chimiche.
Successivamente, mostriamo come la natura probabilistica del nostro schema proposto aumenti la robustezza del convex hull rispetto alla scelta di potenziali differenti e nella scelta della descrizione atomistica del sistema.
Infine, mostriamo i risultati di una ricerca completamente esplorativa sul diagramma di fase del ghiaccio. Riscoprendo tutte le 17 fasi note incluse nel set, proponiamo inoltre una serie di candidati alla metastabilità, e dimostriamo la validità delle fasi proposte testando la loro stabilità rispetto a diverse condizioni termodinamiche. I risultati presentati in questa tesi dimostrano come il GCH possa aumentare, in maniera efficace, schemi di ricerca di strutture con una estensione robusta e flessibile per la ricerca di fasi con alto potenziale di stabilizzazione.

Parole chiave: Machine learning, Termodinamica, High-throughput screening, DFT, Predizio-

ne di struttura cristallina, Generalised convex hull, cristalli molecolari

Contents

List of Figures

List of Tables

1 Introduction

When nature finishes to produce its own species, man begins using natural things in harmony with this very nature to create an infinity of species

Leonardo Da Vinci

1.1 Materials Design

Materials science is a relatively modern branch of science which deals with the modelling and fabrication of materials with a desired set of properties. However, the history of materials research and discovery has played a key role in the development of human civilisation: many fundamental technological breakthroughs have come through the discovery of new materials. The fundamental role played by materials in history is made clear from the naming of ages of civilisations – the stone, iron and bronze ages – with each new technological era being brought about by a new material[1], as shown in Fig. 1.1. Notably, one can observe a trend where subsequent discoveries of a revolutionary material involve a reduction of length scale involved in their production process. The core business of materials science has thus evolved throughout its history, from a practice strongly focused on chemical and industrial process, to a fully interdisciplinary field of study aimed at uncovering the complex relations governing a macroscopic material's properties from its atomic constituents[2]. While a relevant portion of materials research continues to strive for advanced and efficient production of existing compounds, materials design from first principles has picked up an immense momentum[3]. With the advent of cutting edge experimental techniques like chemical vapour deposition (CVD), focused ion beam (FIB) milling and etching it has become increasingly feasible to synthesise materials with an atomistic resolution. Similarly, it is possible to characterise their structure through Scanning Electron Microscopy (SEM), X-ray diffraction (XRD), nuclear magnetic resonance (NMR) or Raman Spectroscopy (to name a few). This wealth of experimental

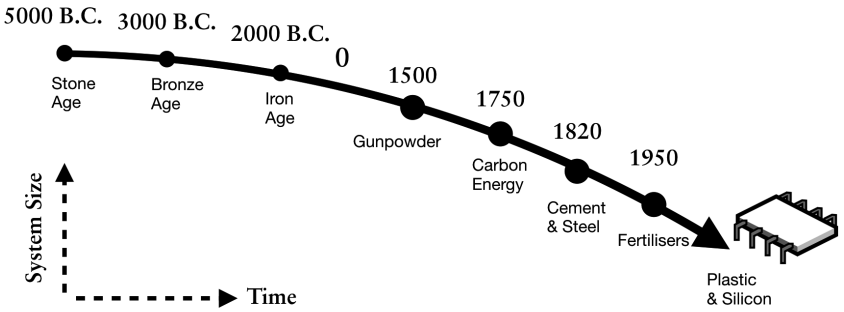


Figure 1.1 – The history of technological ages defined by their fundamental materials

techniques has paved the way for the synthesis and characterisation of novel compounds.
In parallel to the evolution of experimental techniques, the theoretical modelling of materials properties has seen a fundamental surge in all of its fields, bridging the wide gap between the macroscopic, experimentally observable behaviour of materials and their physics at the atomistic level.
Materials properties are determined by the relative arrangements of its single atoms constituents, which are governed by their corresponding many-body Schrödinger equations. While this is a conceptually well defined problem, its numerical solution proves to be prohibitive for any system with more than a few electrons, and thus theoretical materials scientist resort to numerical solutions of simpler, ideal systems which approximate the compounds of interest[4]. This approach, combined with the progressive refinement of the approximated models, has rendered the route of a fully computational experiment (in-silico) a viable route to achieve the discovery of novel compounds and has paved the way for what is considered today to be computational materials modelling[5, 6].
A fundamental branch of research has thus emerged from this materials discovery pipeline, one that aims to performing as many exploratory computational experiments as possible, to explore materials spaces (ensemble of compounds with specific stoichiometries) which can yield materials with industrial value. Such practice is often called "high-throughput" computational materials science, and it is generally a resource-intensive preliminary screening step in the materials design process[7, 8].
A current issue that limits the effectiveness of high-throughput searches is the efficiency which existing algorithms offer, i.e. number of number of candidate materials that survive the property screening (often calculated at a computationally costly quantum mechanical level) over the number of compounds screened. It has become thus of imperative importance to enhance further the explorative power of materials high-throughput search by raising its efficiency, allowing for computational searches schemes which are more likely to find novel, breakthrough materials. The bottlenecks often faced by the high-throughput search are often of two kinds : the low number of configurations exhibiting the desired properties, and most importantly, their experimental realisability.
The objective of this thesis is to empower computational structure searches with a framework

capable to identify in an unbiased and automatic manner the structures which have the highest likelihood of being synthesised in an experimental condition, and thus of industrial interest.

1.2 Modelling Materials Stability

The task of computational materials modelling consists in characterising a materials behaviour under a certain set of external constraints, which are modelled to mimic the effect of the experimental conditions it could be synthesised in. To achieve such a description at an atomistic level, the coordinates of the nuclei defining a particular material's building block are modelled according to the thermodynamic ensemble the user wants to reproduce. These ensembles define a probability distribution of arrangements of the atomic positions. Thus, to measure an observable of interest one must obtain the values of the observable over the ensemble of structures sampled, each weighted by its corresponding probability. The configuration space accessible by a collection of atoms depends on the atoms involved in the material and the physics determining their interactions.
To achieve a physically relevant modelling of a material's behaviour one has to strike a balance between an exhaustive sampling of the corresponding ensemble and an accurate description of the potential energy surface of the atomic arrangements defining it. While a correct treatment of a compound's behaviour should be founded on a first-principles (ab-initio) quantum mechanical description of the nuclei-nuclei, nuclei-electron and electron-electron interactions, the computational cost associated to the solution of their underlying Schrödinger equation is often prohibitive. To sidestep this limitation, approximate methods ranging from semi-empirical models to fully tabulated interactions (force fields) are often employed in place of the rigorous ab-initio potential energy surfaces. While first principle interactions are general and can be applied to any atomic arrangements, tabulated potentials can fall onto unphysical descriptions of systems when exploring regions of phase space (atomic arrangements and momenta) which lie outside of the conditions they have been parametrised for. Further, a different number of configurations' energies must be calculated depending on the in-silico experiment of choice, which correspond to either particularly long simulations or system sizes comprising intractably large collections of atoms. In practice the assessment of a material's behaviour from an atomistic simulation point of view requires a compromise between accuracy of description of its atoms' interactions and extent of sampling of the ensemble of choice. The evaluation of materials' stability is often practiced finding a balance between the two axes of sampling accuracy, and as a result different compromises are chosen depending on the resources available or the level of accuracy required.
A common way to assess a material's stability is whether its arrangement of atoms corresponds to a minimum on its system's potential energy surface (PES). While there are often a very large number of configurations for which the net force acting on each atom is zero, their enumeration defines a sound starting point for the discovery of structures that can exhibit

stability to thermodynamic perturbations. The robustness of a minimum configuration is correlated to its lifetime in an experimental realisation, and measured by the depth of its well in the PES. Every point that is not the global minimum of the PES is considered to be meta-stable.
The great limitation of such approach is that it disregards the effects that any temperature and pressure related perturbation can have on the stability of the local configurations. To give a concrete example, it is only by including vibrational degrees of freedom that the hexagonal packing of ice results to be more stable than its cubic arrangement[9].
To establish a stronger link between experiments and simulations, it is thus fundamental to include the contributions that arise from assessing the stability of the configuration in the ensemble of interest, requiring in practice evaluations of their system's free energies. These energies require extensive sampling of the configurations interactions and thus become feasible only for a handful of candidates that exhibit promising features.

1.3 Machine Learning in Atomistic Modelling

The advent of machine learning (ML) has introduced a fundamental paradigm change in almost all the fields of science, promoting the role of data into that of a powerful catalyst for the discovery of previously unknown insights. In parallel, the combined growth of computational resources and materials modelling techniques has enabled materials scientists to sidestep many costly experiments by approximating them with atomistic simulations [10]. This approach allows obtaining data containing far more information than the one pursued by the computational experiment, as it encodes the physically relevant phenomena behind it. The outcome of each computational experiment can be thus rethought as a collection of data containing fundamental correlations which shed light on the atomic-scale behaviour of their corresponding materials.
The application of machine learning techniques to data coming from atomistic simulations has allowed computational materials scientist to achieve a broad spectrum of methodological and theoretical breakthroughs ranging from fast surrogate ML models predicting ab-initio grade properties at a fraction of the cost [11] to generative models which can autonomously design molecules showing a desired set of properties [12].
The field of atomistic machine learning, however, has seen a late bloom when compared to other fields of computational science. This delay was due to the need of finding an accurate translation of atomistic systems into features, i.e. arrays of numbers commonly used as input for ML schemes. The choice of a descriptor function used to "featurise" the molecular inputs needs to satisfy a number of symmetries which lie at the core of the properties one wants to characterise: two molecules of benzene should be described by the same feature irrespective of their relative orientation, translation of reference axis, or permutation of labels of their atoms.
With the introduction of symmetry-adapted descriptors, such as Bag of Bonds [13], Coulomb

matrices [14] and SOAP fingerprints [15], there has been a prolific development of ML models aimed at producing accurate predictors of a variety of molecular properties: electronic charge density predictions on organic molecules [16], NMR chemical shifts in molecular crystals [17] and machine-learning interatomic potentials [18] to list a few.
Since these set of descriptors aim at capturing the structural similarity between different molecular inputs, one can extend their use to navigating extensive collections of data coming from molecular dynamics or structures searches to recover recurring patterns hidden in the data. This task is typically called unsupervised learning, and it has been used successfully to identify novel definitions of hydrogen bonds in condensed matter [19] or discover the main building blocks in zeolite frameworks [20]. These agnostic approaches can be combined with training labels to obtain the flexibility of unsupervised learning with supervised schemes by performing combined learning: the new patterns that are extracted in an agnostic approach from the set can find their validation using the properties available in the training set.
The development of ML schemes to model materials stabilities has traditionally focused on building supervised models capable of interpolating the energies of an atomic system using a set of reference calculations. The goal of these machines is to cut the cost of a computationally expensive, high accuracy energy evaluation by training a predictive model which infers a new configuration's energy based on its structure, as seen by its descriptors. Starting from simple isolated molecule's energies predictions [14], the community has developed strategies to extend the predictive models to solids [21, 22], complex compounds spaces [23] and more recently, charged systems [24].
These schemes can be a solid starting point for performing high-throughput screening of materials with the desired range of stabilities. However, they can only provide static information of the configuration under study. In order to include insight into the effect that thermal excitation can have on the systems, ML potentials (MLP) can be trained to perform accurate simulations at a reduced computational cost. While similar to the aforementioned static energy regressors, MLIPs require a subtler and costlier training procedure, as they are built to model the potential energy surfaces and their derivatives, thus requiring many out of equilibrium configurations. Despite their training requirements, MLPs are a very promising means to endow costly calculations (either due to sheer system size or sampling time) with an accuracy comparable to the one reached by ab-initio methods. Successful examples of their use are for example, free energy calculations for achieving quantitative comparisons of phase stability in ice [25], modelling of structural transition in nanometer scale dense disordered silicon [26] or estimation of proton transfer's free energy at the water-TiO_2 interface [27]. While these methods are very effective at performing comparisons of configurations well represented by their training set, they have not proven to be reliable enough to single-handedly guide a structure search exploration[28]. This observation limits the promise of ML aided structure discovery heavily, requiring a different angle to enhance the efficiency of crystal structure prediction routines.
Over the last decade, a tremendous amount of work has been dedicated to understanding the interplay between the short-range interactions and long-range ordering effects in hydrogen-bonded systems, showing how much of its behaviour is dominated by competing

local structural motifs [29]. Similarly, the weak intermolecular forces causing polymorphism in molecular crystals is determined mainly from the local interactions between the molecular moieties [30]. These investigations proved that a detailed description of local environments in such molecular systems could provide an insightful starting point for capturing the mechanisms governing a system's stability.
The overarching goal of this thesis is to use unsupervised ML techniques to discover the link between structural patterns in crystals and stabilising mechanisms and leverage them to widen the breadth of crystal structure search to probe stabilisation schemes beyond the conventional ones. The result of this research is the introduction of a generalised convex hull construction (GCH), a data-driven redefinition of the well established convex hull method which is capable of extracting potentially stable phases arising from crystal structure searches in an entirely agnostic manner.

1.4 Summary

The thesis is organised as follows: We provide a short introduction to machine learning techniques in atomistic simulations in chapter II. In Chapter III, after a basic introduction on geometrical approaches to determine phase stability, we introduce the theoretical framework behind the generalised convex hull construction. In chapter IV, we show applications of the method to solid hydrogen at high pressure, the hydrogen-oxygen binary system at different stoichiometries and finally, in molecular materials. In chapter V we show an explorative application of the GCH framework to aid the discovery of novel phases of ice.
Finally, in Chapter VI, we propose some future outlooks while drawing our conclusions. Chapter III is partially adapted from refs.[31], while chapter IV is adapted from refs.[31]. Chapter V is adapted from [32].

Theory Part I

2 Unsupervised Machine Learning

2.1 Introduction

Machine learning (ML) is a branch of Artificial Intelligence, whose scope is to provide computers with a set of instructions that allow them to infer correlations from training data that can be generalised to new, unseen inputs.
ML is a fundamentally interdisciplinary field which has seen massive progress over the past decade, bringing breakthrough discoveries bringing breakthrough discoveries from IBM's Project Debate [33] to the automatic interpretation and translation of text [34]. The underlying power of ML approaches is their ability to produce models that describe complex phenomena by only training machines with examples, bypassing the need of either finding their analytical models or running their costly underlying calculations when available.
The process of learning in an ML algorithm corresponds to the model's performances increase upon gaining experience, which in this analogy correspond to the amount of data. It is in fact in the data that all the correlations between the input (i.e. the hypothesis) and the outputs (the results) are encoded: the scope of a successful ML algorithm is to adapt its functional form to model the input-output relation in the most accurate way possible, so to produce a confident estimate of new outputs given an unseen set of inputs.
The practice of learning given these additional forms of information is known as supervised learning (SL), in which a predictive model is trained to extrapolate an unknown property for a new, unlabeled data entry. Alternatively, schemes targeting the extraction of recurring patterns are commonly referred to as unsupervised learning (UL) and are typically used in cases where no information on training data, such as classifications or dependent outputs, is available.

2.1.1 Supervised Learning

The typical setting of a Supervised Learning algorithm starts from an ensemble of input-output pairs $(\mathbf{x}, \mathbf{y})$ with $x \in \mathbb{R}^D$ and $y \in \mathbb{R}^M$, to produce a $\mathbf{y}^*$ given a new input $\mathbf{x}^*$. One can see how such a formulation allows extreme flexibility in terms of the problem statement since the input

x can be either an image (e.g. encoded in their RGB arrays) or a vector, while the output can be either an array of floats or a set of binaries (multi-output classification).
The function $\mathscr{F}$ is a mapping between the inputs and the outputs, and depends on a series of parameters often called "model": $\{w_i\}_{i\in\text{model}} = \mathbf{w}$. In order to adapt the function to fit the training points, it is customary to find the optimal $\mathbf{w}^*$ by minimising the model's loss function:

$$\mathbf{w} = \min_w \mathscr{L}\left(\mathbf{w}, \{\mathbf{x}_i, y_i\}_{i\in\text{validation}}\right)$$

Over a series of values coming from a validation set, which is extracted from the dataset and used as a reference. Depending on the model of choice, many different functional forms can be adopted, and the choice of their optimal weights extracted either analytically or through global minimisation.
The early approaches stemmed from extending linear regression to multi-dimensional inputs of the kind $\mathbf{X} = \{w_i\}_{i\in\text{dataset}}$, consisting of a linear filter in the form of:

$$\mathscr{P}(\mathbf{X}) = \mathbf{w}^*\mathbf{X}$$

Overfitting arises when the model is trained to match perfectly the training inputs, at the cost of its generality. To compensate, one can regularise the loss functions so to prevent its solutions from following too closely the input data: the most common example is the Tikhonov regularisation (L^2)[35] or a Lasso regularisation (L^1)[36]. Another merit of regularised approaches is to produce more stable estimators, as very often the large dimensional spaces used as inputs suffer from multicollinearity (i.e. a linear dependence between them).
In modern supervised approaches many branches of modelling have been used to produce efficient regressors: Boltzmann machines, decision trees, support vector machines, neural networks and more recently, graph networks, are ubiquitous examples [37–39].
A very recent promising trend involves the use of deep learning schemes, where the architecture of the models is often composed of multiple interconnected layers of non-linear functions. These schemes leverage on the complexity of their structure (i.e. the raw number of fitting parameters, often in the order of millions and above) to build models of extreme accuracy, at the cost of long training times and requirement of massive training datasets. Examples of these structures are deep Neural networks such as AlexNet (or VGG net), deep graph networks and extreme learning [40, 41].

2.1.2 Unsupervised Learning

The other fundamental branch of ML is called unsupervised learning (UL), and it is commonly used with cases where correlations have to inferred from unlabelled data.
The purpose of UL algorithms is two-fold:

1. explore the dataset in search of hidden correlations between the input points (dimensionality reductions, outliers analysis).

2. find rules to group them by (clustering analysis)

Since a great deal of the development work in this thesis lies its foundations on UL algorithms, the scope of the following sections is to guide the reader throughout some critical algorithms in both dimensionality reductions and clustering techniques. For the readers interested in a general overview of ML, two interesting references are *Information Theory, Inference, and Learning Algorithms* by DJC. MacKay [42] and *Pattern Recognition and Machine Learning* by C. Bishop [43].

2.2 Dimensionality reduction

Dimensionality reduction (DR) aims at finding a lower-dimensional projection of a high dimensional set such that a metric of choice is preserved. One can, in general, define their dataset as a collection of points $\mathbf{X} = \{\mathbf{x}_i\}_{i=1..N}$ and $\mathbf{x}_i \in D$, indicating a collection of N points populating a D-dimensional "feature" space.
When a population of points comes from some form of sampling of its underlying distribution, it is common for it to lie in a subspace of its full feature space. The effective dimensionality of this lower-dimensional manifold is often called "intrinsic dimensionality" (ID). An accurate estimation of a set's ID is key in determining potential for compression in signals [44], fractals' dimensions [45] and reconstructing a protein sequence [46]. DR techniques are devoted to finding the optimal $d < D$ dimensional space which preserves the relevant topology from the full-dimensional space. Colloquially stated, DR is often used to provide a reduction of high dimension data projected into human-interpretable dimensions, i.e. less than or equal to three. The reduction of the dimensionality of the data is often used to aid its visualisation for rapid detection of outliers or clusters, although it can be used to uncover much more information. Further examples are the removal of unimportant degrees of freedom, identification of slow modes in high dimensional time-series [47].

2.2.1 Principal Component Analysis

One of the most traditional approaches to DR is principal component analysis (PCA) [48]. PCA applies an orthogonal transformation to the feature space that reduces the number of its (possibly) correlated dimensions into a (smaller) number of uncorrelated directions called principal components (PC).
The rationale guiding the reduction is to produce a lower dimensionality embedding whose dimensions are uncorrelated and ordered so that the first few retain most of the variation present in the original variables. Although there are multiple ways to extract the PC from a set, the practical calculation of the principal components often amounts solving an eigenvalue-eigenvector problem for a positive-semidefinite symmetric matrix.
If one considers the set of N, D-dimensional points $\mathbf{X} = \{\mathbf{x}_i\}_{i=1..N}$, it is possible to extract its PC

by calculating the eigenvectors of the covariance matrix associated to the centred distribution

$$\tilde{\mathbf{X}} = \mathbf{X} - \frac{1}{N}\sum_{i=0}^{N}\mathbf{x}_i \tag{2.1}$$

Hereon the " " accent will be dropped for simplicity and all input sets "$\mathbf{X}$" should be assumed centered. We can calculate the covariance matrix $\mathbf{C} = \mathbf{X}^T\mathbf{X}$ and the simply rewrite every points' coordinate as :

$$\mathbf{x}_i = \frac{1}{N}\sum_{i=0}^{N}\mathbf{x}_i + \sum_{k=1}^{N} p_{i,k}\mathbf{e}_k \tag{2.2}$$

The $\mathbf{e}_k$ are the eigenvectors of the matrix C and are D dimensional orthonormal vectors.
The scalar $p_{i,k}$ is the projection of the point $\mathbf{x}_i$ on the dimension k. At this stage, there still has not been any compression, however, if one sorts the eigenvectors in descending order of their corresponding eigenvalues, the new point's coordinate can be approximating by cutting its expansion until the d large enough eigenvalues.
This thresholding effectively allows to approximate:

$$\mathbf{x}_i \sim \sum_{k=1}^{d} p_{i,k}e_k$$

PCA is extremely general and is often used to produce a rapid overview of the order of modality of the distribution of inputs or filter the data representation from undesired noise.

2.2.2 The kernel trick

It is often the case that the relationship between points is best represented in terms of a kernel measure of their similarity. Such a mathematical construction allows to bypass explicit referencing to their underlying Hilbert space, provided the existence of kernel function associated to it, as depicted in Fig. 2.1. This approach is often commonly called "kernel trick" and allows to describe a distribution of points in a Hilbert space by knowing the values of the inner product between their features. If we introduce a non-empty set $\mathbf{X}$, one can define a kernel k as any symmetric, real-valued function acting on $\mathbf{X}\times\mathbf{X}$ that is positive definite (psd). This condition amounts to finding a symmetric function $k : X \times X \rightarrow \mathrm{R}$ that satisfies:

$$\sum_{i,j}^{n} c_i c_j k(x_i, x_j) \geq 0 \tag{2.3}$$

for any $n \in \mathrm{N}$, and $\mathbf{x}_1, .., \mathbf{x}_n \in \mathbf{X}$ and $c_1, .., c_n \in \mathrm{R}$.
A useful property of a psd kernel is that it can be thought of as an inner product between two feature veectors of its arguments[49]. This results in simple linear kernels by considering the Gram Matrix obtained on the feature space, or consider higher order correlations between the

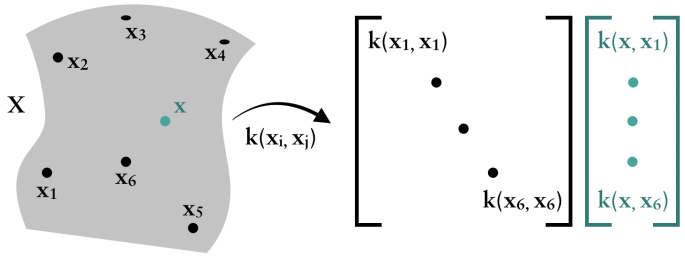


Figure 2.1 – If one considers the set of points $\{x_i\}_{i=1..6}$ lying in the space X, one can either characterise them by their coordinates $x_1, x_2, x_3..x_6$ or adopt a kernel representation of their pairwise similarity. In this case, by using the 6×6 dimensional positive-definite symmetric matrix $k(x_i, x_j)$ depicted above one encode the same spatial correlations. To formulate the predictions on a new point x, one will compare its similarity with respect to the initial embedding set, by building a new similarity matrix between the new input and the training set.

data by adopting kernel functions, such as Gaussian, radial basis or polynomial kernels.
The advantage of this formulation is that it enables the highly non-linear mapping of the input data and allows saving storage or memory space when representing datasets lying in dimensions D far larger than their set size N.
For a more detailed discussion on kernel methods, we suggest the reading of reference [3, 49].
Alternatively, it is extremely common to represent a distribution of points through a measure of their pairwise distance. In general, to define a distance metric d, acting on the same set X, one would need to define a function satisfying the following properties:

- $d(x_i, x_j) \geq 0$ and $d(x_i, y_j) = 0$ if and only if $i = j$ (non-negativity)
- $d(x_i, x_j) = d(x_j, x_i)$ (symmetric)
- $d(x_i, x_k) \geq d(x_i, x_j) + d(x_j, x_k)$ (triangular inequality)

for any x_i, x_j, x_z of X.
In case one is able to define a kernel measure between a distribution of points, it is possible to produce a valid corresponding distance function by calculating:

$$d(x_i, x_j) = \sqrt{k(x_i, x_i) + k(x_j, x_j) - 2k(x_i, x_j)} \tag{2.4}$$

This equation allows one to seamlessly switch the two metrics, and effectively allows the use of many-distance based techniques starting from kernels representations.

2.2.3 Multidimensional scaling

A widely adopted strategy used to perform DR is to find the low dimensional Cartesian projection that best reproduces the pairwise distances in the high dimensional space. Such approaches fall in the family of multi-dimensional scaling (MDS). In the case of metric MDS, a stress function $\mathscr{S}(x_1, x_2, .., x_N)$ is defined, which accumulates the residual of the lower dimensional embedding compared to the high dimensional space in the form :

$$\mathscr{S}(x_1, x_2, .., x_N) = \frac{1}{N}\sqrt{\sum_{i,j}(d_{ij}^p - \|x_i - x_j\|)^2} \tag{2.5}$$

The p term is the parameter involved in the embedding, and can be tuned to weight with different intensity the distances in the full-dimensional space.
The formulation of metric MDS is extremely flexible, and can be naturally extended to introduce elements of nonlinearity by distorting the distances in the stress function.
The distortion of the distance functions serves the purpose of characterising the topology of the high dimensional space one is trying to approximate. While keeping a Euclidean distance can be useful in simple manifolds, it can in general reproduce a deceiving picture of the folding of the points distribution. A powerful alternative consists finding the best embedding which preserves the geodesic distances between the points in high dimension as sketched in Fig.2.2, by following the Isomap algorithm [50]. Isomap aims to capture the complex shapes of a high dimensional manifold M by minimising a distance metric which encodes the connectivity of the full distribution of points.
To find the optimal embedding $\mathbf{y}^*$, one has to follow three steps:

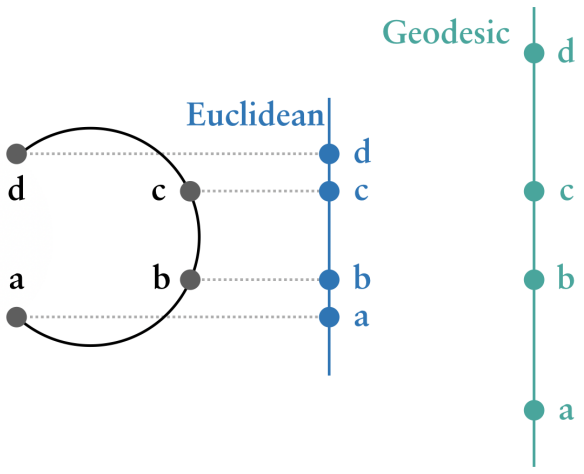


Figure 2.2 – In figure, a simple depiction of the difference between a projection preserving the pairwise Euclidean distance between points, and on the right the corresponding isometric mapping procedure.

1. Calculate the Euclidean distance $d(x_i, x_j)$ between each pair of points in the set and use it to build a corresponding (weighted) connectivity graph G. The resulting graph is obtained by thresholding the distance matrix according to a fixed neighbouring radius ε. The values of the edges between neighbouring points is their Euclidean distance.

2. The pairwise geodesic distance matrix $d_M(\mathbf{x}_i, \mathbf{x}_j)$ is approximated by the shortest path distance matrix $d_G(\mathbf{x}_i, \mathbf{x}_j)$. To obtain d_G, one sets all $d_G(\mathbf{x}_i, \mathbf{x}_j)$ elements corresponding to neighbours on the connectivity graphs to their euclidean distance, and the remaining ones to an arbitrarily high number (∞ in the original article). Finally, one fills every entry remaining entry of the d_G matrix by cycling throughout every point of the set $k = 1..N$: $d_G(\mathbf{x}_i, \mathbf{x}_j) = min\{d_G(\mathbf{x}_i, \mathbf{x}), d_G(\mathbf{x}_i, \mathbf{x}_j) + d_G(\mathbf{x}_k, \mathbf{x}_j)\}$

3. Finally, classical MDS is applied to the matrix $D_G = \{d_G(x_i, x_j)\}_{i,j \in N}$, by optimising the stress function $\mathscr{X}$ on the coordinate vectors y_i of the resulting Euclidean embedding space:

$$\mathscr{X} = \|\tau(D_G) - \tau(D_Y)\|_{L^2} \tag{2.6}$$

   Where D_Y represents the Euclidean distance matrix in the new, low dimensional space Y and τ transforms the distance matrices into inner products [50].

The power of ISOMAP lies on its ability to produce a mapping which aims at unfolding the global distribution of points in the high dimensional space, which allows for visualisation of global pathways linking the main modes of the data.

2.2.4 Sketch-map

Another very effective way to introduce nonlinearity in case of data coming from atomistic simulations has been introduced in the Sketch-map algorithm [51]. To obtain a lower dimensional embedding one can minimize the stress function :

$$\mathscr{X}^2 = \sum_{i \neq j} \frac{1}{w_i w_j} \sum_{i \neq j} w_i w_j \left[F(R_{ij}) - f(r_{ij})\right]^2 \tag{2.7}$$

Where w_i is the weight of i-th point, $R_{ij} = \|\mathbf{X}_i - \mathbf{X}_j\|$ with $\mathbf{X} \in \mathbb{R}^D$ and $r_{ij} = \|\mathbf{x}_i - \mathbf{x}_j\|$ with $x \in \mathbb{R}^d$, with $D >> d$. The two distances are filtered by two independent sigmoid functions $F = S_{A,B,\sigma}$ and $f = S_{a,b,\sigma}$,which maps monotonically $\mathbb{R}^+$ to $[0,1)$, and uses the following functional form:

$$S_{a,b,\sigma}(r) = 1 - \frac{1}{[1 + (2^{\frac{a}{b}} - 1)(\frac{r}{\sigma})^a]^{\frac{b}{a}}} \tag{2.8}$$

This formulation allows the user to select through the parameter σ which distance to be considered as a threshold for perceptual distance. The parameter a (b) tunes the steepness

with which points falling before (after) σ are mapped to zero (one). The combination of σ with the two steepness parameters allows the user to highlight similarity at the desired length scale, disregarding fluctuations arising from the intrinsic noise in the data or suppressing slow modes hidden in the samples.
Another advantage of MDS schemes is their scaling with the size of the database (i.e. the number N of points to embed). While the direct stress function minimisation scales $O(N) \sim N^2$ with N points, it is possible to approximate the embedding into a two-step problem, thereby reducing the scaling to $O(N) \sim NM$.
To build such a reduction, one can produce an explicit embedding for a reduced set of representative configurations M (landmarks), and minimise the stress function only in terms of distances of every other X point from the M landmarks :

$$\mathscr{X}^2(\mathbf{x}) = \left(\sum_{i=1}^{M} w_i\right)^{-1} \sum_{i=1}^{M} w_i [F(\|\mathbf{X}-\mathbf{X}_i\|) - f(\|\mathbf{x}-\mathbf{x}_i\|)]^2 \tag{2.9}$$

Where w_i is the weight assigned to the i-th landmark.
This formulation allows fast and non-linear embedding for millions of points and has been used successfully in many molecular dynamics simulation data, where thermal vibrations introduce often unimportant degrees of freedom. Relevant examples show its application in describing the potential energy surface (see Chapter 2 for further details) of a Lennard Jones system [52] or describe the conformations of aspartic acid in solution [53].

2.2.5 Kernel Principal Component Analysis

The PCA framework has been extended to kernels [54], allowing for an easier visualisation of the high dimensional manifolds reproduced by the kernel trick.
Given that the kernel matrix corresponds to an inner product in its reproducing kernel Hilbert space (RKHS)[55], considering two arbitrary elements $\mathbf{x}$ and $\mathbf{x}'$, one can write that :

$$K(\mathbf{x},\mathbf{x}') = \langle \boldsymbol{\phi}(\mathbf{x}), \boldsymbol{\phi}(\mathbf{x}')\rangle = \boldsymbol{\phi}(\mathbf{x})\boldsymbol{\phi}(\mathbf{x}')^T \tag{2.10}$$

Where the function is a nonlinear mapping $\boldsymbol{\phi} : \mathbb{R}^{\mathbb{D}} \rightarrow F$, projecting the dataset into a potentially infinite dimensional space. If we assume that the data is centered in F, one can calculate the covariance matrix associated with the mapped distribution as

$$\mathbf{C}_{\boldsymbol{\phi}} = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{\phi}(\mathbf{x}_i)^T \boldsymbol{\phi}(\mathbf{x}_i) \tag{2.11}$$

The solution to the eigenvalue problem will be given by all the eigenvectors $\mathbf{V}$ and their corresponding eigenvalues λ. By construction, all the solutions included in $\mathbf{V}$ can be spanned from the collection of $\{\boldsymbol{\phi}(\mathbf{x}_i)\}_N$ with linear combination weights α_i, which allows us to rephrase

the solution to the eigenvalue problem in terms of :

$$\lambda(\boldsymbol{\phi}(\mathbf{x}_i)\cdot\mathbf{V}) \;=\; (\boldsymbol{\phi}(x_i)\cdot\mathbf{C}_{\boldsymbol{\phi}}\mathbf{V}) \qquad \text{and} \qquad \mathbf{V}=\sum_{i=1}^{N}\alpha_i\boldsymbol{\phi}(\mathbf{x}_i) \tag{2.12}$$

By combining these two conditions one can write the :

$$\lambda\sum_{i}^{N}\alpha_i(\boldsymbol{\phi}(\mathbf{x}_k)\boldsymbol{\phi}(\mathbf{x}_i)) \;=\; \frac{1}{N}\sum_{i}^{N}\alpha_i(\boldsymbol{\phi}(\mathbf{x}_k)\cdot\sum_{j}^{N}\boldsymbol{\phi}(\mathbf{x}_j))(\boldsymbol{\phi}(\mathbf{x}_j)\boldsymbol{\phi}(\mathbf{x}_i)) \quad \text{for } \; k=1..N \tag{2.13}$$

By recalling that the matrix element of the kernel matrix is $K_{ij}=\boldsymbol{\phi}(\mathbf{x}_i)\boldsymbol{\phi}(\mathbf{x}_j)^T$, one can write the equation in a matrix form $N\lambda\mathbf{K}\boldsymbol{\alpha}=\mathbf{K}^2\boldsymbol{\alpha}$.
The solution of such a system are found by solving the eigenvalue problem in $\boldsymbol{\alpha}$ of $N\lambda\boldsymbol{\alpha}=\mathbf{K}\boldsymbol{\alpha}$.
If we consider all the eigenvalues λ_{iD} of $\mathbf{K}$ from above, and $\boldsymbol{\alpha}$ their corresponding eigenvectors, one can impose that their non-zero eigenvalues corresponding vectors in F are normalised:

$$(\mathbf{V}^k\cdot\mathbf{V}^k)=1=\sum_{i,j}\alpha_i^k\alpha_j^kK_{ij}=(\alpha^k\cdot K\alpha^k)=\lambda(\alpha^k\cdot\alpha^k) \tag{2.14}$$

When applying principal component analysis, it is customary to extract only the first d projections on the eigenvectors $\mathbf{V}^k$ in F, thus, given a new test point with an image $\boldsymbol{\phi}(\mathbf{x}^{test})$, one can write

$$(\mathbf{V}^{\mathbf{k}}\cdot\boldsymbol{\phi}(\mathbf{x}^{\text{test}})) \;=\; \sum_{i}^{D}\alpha_i^k(\boldsymbol{\phi}(\mathbf{x}_i)\cdot\boldsymbol{\phi}(\mathbf{x}^{\text{test}})) \;=\; \alpha\cdot k(\mathbf{x},\mathbf{x}^{\text{test}}) \tag{2.15}$$

This result extend the use of PCA to any non-linear mapping of the input data, and is a fundamental result of dimensionality reduction theory.

2.2.6 Other Dimensionality Reduction Schemes

Multidimensional scaling is not the only way to capture non-linear spatial features, as many other approaches have been developed which suit different scenarios: Hessian based locally linear embedding (He-LLE) [56], t-SNE [57], and diffusion maps [58].
A modern approach to dimensionality reduction is to train neural networks schemes called autoencoders [59][60]. An autoencoder is an artificial neural network which aims at encoding the input vector into an identical output vector. It can be simply turned into a dimensionality reduction scheme by reducing an inner hidden layer number of units with an arbitrarily lower number of nodes. The input vectors are thus mapped into a latent space which is used as a lower-dimensional embedding of the input data. The rest of the network carries the task of remapping the latent space inputs to the full high dimensional space, and is often called a decoder.
Reducing dimensionality can be useful for characterising variability, removing unimportant

degrees of freedom, and discovering better representations of complex data. In atomistic simulations, for instance, the degrees of freedom characterising a system can be on the order of millions, which often makes the reduction of dimensionality necessary for obtaining an intuitive picture of the physics hidden in the data.

2.3 Clustering

A second family of unsupervised learning algorithms deals with the problem of automatically partitioning points into groups. Such separation is usually carried out to reflect the similarity between points belonging to the same groups and at the same time, meaningful difference between points belonging to different groups.
The aim of such separation is to glance at the dominating traits in the dataset, which usually reflect into a finite and clear number of clusters. This it plays a fundamental role in many modern schemes and real-life scenarios: automatic detection of fake news [61] and fraudulent behaviour [62].
As in most machine learning algorithms, at the core of a clustering scheme, there is a cost function measuring the quality of grouping achieved at the iteration t as a function of every point's assignment $\mathscr{L}(\mathbf{X}) = \{l(\mathbf{x}_i)\}_{i=1..N}$. The function l maps from $\mathbb{R}^D \rightarrow \mathbb{R}^{N_{clusters}}$ and stores the assignments of each point $\mathbf{x}_i \in \mathbf{X}$ to the j-th cluster, either with a boolean value (hard clustering) or with a probability (soft clustering).

2.3.1 k-means

The simplest algorithm for cluster analysis, k-means, is considered to be a partitioning scheme, in that every point has to be assigned to a group, as shown in Fig.2.3.
The idea behind k-means is to partition n samples into k clusters in which each data point

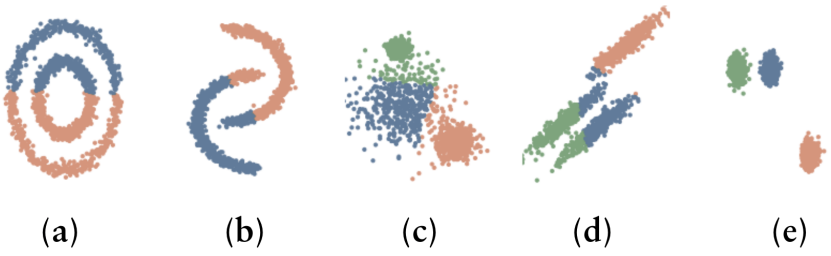


Figure 2.3 – k-means clustering assignment applied to a set of five different tests. The two (three in case (c),(d),(e)) different colours define the clusters found. One can already see how k-means excels in presence of globular data (i.e. (c) and (e)) and finds it difficult to follow the structure of the data whenever the differet modes become more anisotropic.

belongs to the cluster with the nearest mean, serving as a prototype of the cluster.
If one considers a feature matrix $\mathbf{X} = \{\mathbf{x}_i\}_{i=1..N}$, with $x \in \mathbb{R}^D$, the k-means objective is to

partition it into a set $\mathbb{S} = S_1, S_2..., S_k$ that minimises the within-cluster sum of squares :

$$\arg\min_{\mathbf{S}} \sum_{i=1}^{k} \sum_{\mathbf{x}\in S_i} \|\mathbf{x} - \mu_i\|^2 \tag{2.16}$$

The vector $\mu_i = \frac{1}{|S_i|}\sum_{x_i\in S_i}$ represents the average value of the points belonging to the cluster i, it can be seen as a virtual input around which every cluster is centered.
To solve this problem, one can use Lloyd's algorithm (Voronoi iteration) to obtain a solution iteratively by following a two-step approach (often called Expectation-Maximisation):

1. Expectation: assign every point to a set:

$$S_i^{(t)} = \{x_p : \|x_p - \mu_i^{(t)}\|^2 \leq \|x_p - \mu_j^{(t)}\|^2 \forall\, j, 1 \leq j \leq k\}$$

   where x_p is assigned exclusively to one cluster per t iteration

2. Maximisation: update the mean positions by averaging over the points in the newly built sets:

$$\mu_i^{(t+1)} = \frac{1}{\|S_i^{(t)}\|} \sum_{x_j\in S_i^{(t)}} x_j$$

The algorithm converges with the assignments and their corresponding. This solution, however, does not always guarantee to coincide with the optimal solution: the local minimum curse is one of the key weak spots of the k-means clustering.
This issue is strictly dependant on the choice of the initial means, but can be compensated by either running multiple initialisations, or by choosing them systematically.
A popular fix to this has been proposed in k-means++ [63], which augments the speed of convergence from $O(N\log N)$ to $O(\log N)$. The idea behind k-means++ revolves around the choice of an optimal starting point for the mean vectors, which aims at maximising the coverage over the distribution of points.
By defining a minimum distance of a data point from its closest centre $D(x)$, one can obtain an optimal set of k centres by using the following recipe:

1. Choose the first μ_0 randomly from X
2. The next centre μ_t is the $x \in X$ with the highest probability $p(x)$ defined as follows :

$$p(x) = \frac{D(x)^2}{\sum_{x\in X} D(x)^2}$$

3. Repeat step 2 until k centres have been defined

This approach allows to set up an efficient starting point for the classical k-means algorithm defined before and produces consistently faster convergence to a lower minimum of the associated cost function.

2.3.2 Gaussian mixture models

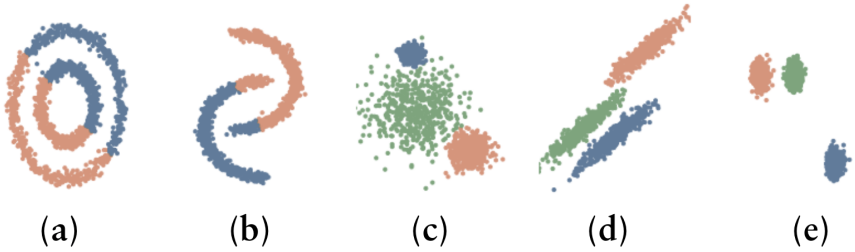


Figure 2.4 – Gaussian mixture model clustering assignment applied to a set of five different tests. The two (three in case (c),(d),(e)) different colours define the clusters found. When compared to the results obtained through k-means, one can see how GMM suffers of substantially similar drawbacks. The capacity of employing anistropic gaussian mixtures allows it to follow accurately the shape of data of case (d).

K-means is traditionally considered to be a hard clustering approach since every point can be assigned to one and only one cluster. Such model can often be limiting, especially in cases where distributions are scattered with different densities, or in the presence of unclear boundaries between nuclei.
To alleviate for such condition one imagine the assignment vectors as probabilities of each point of being part of a cluster S_i. In Gaussian mixture models, the user supposes that the data is distributed according to a multimodal Gaussian distribution, which can be decomposed in sums of multiple Gaussians.
Each centre of these normal distributions represents a centre of a cluster, and their covariance matrix is optimised so to maximise the likelihood of the spread of points around them.
In GMM, every point is assigned a probability $P(x)$ that depends on the sum of the k models:

$$P(x) = \sum_{i=1}^{k} \phi_i \mathcal{N}\left(x \mid \mu_i, \Sigma_i\right) \qquad \text{with} \qquad \sum_{i=1}^{k} \phi_i = 1 \tag{2.17}$$

Where ϕ_i is each model's weight, and $\mathcal{N}\left(x \mid \mu_i, \Sigma_i\right) = \frac{1}{\sqrt{(2\pi)^k |\Sigma_i|}} \exp\{(-\frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1}(x-\mu_i))\}$ represents a multivariate Normal distribution.
To find the optimal parameters the ML estimation is carried out numerically by using a two-step expectation maximisation (EM) process, much like in k-means. The first step is called expectation step and involves the calculation of the component assignments's expectation C_k for each $x_i \in X$ given a set of model parameters $(\phi_k, \mu_k \sigma_k)$.
In the second step, often called maximisation step, the objective is to maximise the expectations coming from the previous step over the variables costituing the model parameters. In

practice the result of this step will be the update the model parameters to $(\phi_{k+1}, \mu_{k+1}\sigma_{k+1})$. This sequence of operations create an iterative process which continues until a convergence of the model parameters is reached, giving a maximum likelihood estimate. Intuitively, the algorithm works because knowing the component assignment C_k for each x_i makes solving for $(\phi_k, \mu_k\sigma_k)$ easy, while knowing $(\phi_k, \mu_k\sigma_k)$ makes inferring $p(C_k|x_i)$ trivial.
The expectation step corresponds to the latter case, while the maximisation step corresponds to the former. Thus, by alternating between which values are assumed fixed, or known, maximum likelihood estimates of the non-fixed values can be calculated efficiently.
To initialise the parameters, we extract from X to the k list of means $\mu_{i\,i=1..k}$, and define a constant variance for all the components equal to the variance in the set and fix every component weight to $\phi_i = 1/k$. To find the optimal $(\phi_k, \mu_k\Sigma_k)$ one iterates through the EM steps as follows, for a univariate case:

1. E-Step, where one calculates $\gamma_{i,k}$, probability of a point i to be in cluster k, $\forall\ i, k$

$$\gamma_{i,k} = \frac{\phi_k \mathcal{N}\left(x_i \mid \mu_k, \sigma_k\right)}{\sum_{j=1}^{k} \phi_j \mathcal{N}\left(x_i \mid \mu_j, \sigma_j\right)}$$

2. M-Step, which determines the update of the models' parameters :

$$\phi_k = \sum_{i=1}^{N} \frac{\gamma_{i,k}}{N} \quad , \quad \mu_k = \frac{\sum_{i=1}^{N} \gamma_{i,k} x_i}{\sum_{i=1}^{N} \gamma_{i,k}} \quad , \quad \sigma_k^2 = \frac{\sum_{i=1}^{N} \gamma_{i,k}(x_i - \mu_k)^2}{\sum_{i=1}^{N} \gamma_{i,k}}$$

The extension to a multivariate case is trivial and allows the user to capture clusters which present anisotropic distributions.
Finally, to calculate the probability of a point i to belong to cluster k, one can calculate:

$$P(S_k|x_i) = \frac{\phi_k \mathcal{N}(x \mid \mu_k, \sigma_k)}{\sum_{j=1}^{k} \phi_k \mathcal{N}(x \mid \mu_j, \sigma_j)} \tag{2.18}$$

The advantage of GMM over k-means clustering lies in the ability to perform soft assignments and the ability to detect cluster with non-globular shapes, thanks to the anisotropy introduced in the multivariate covariance matrices.
Similarly to k-means, it still requires in input the number k of clusters, which is practically dealt with by running multiple iterations of the algorithm with increasing k, and plotting the trend of a measure of the quality of their resultant assignments.

2.3.3 Validating the clustering

The assessment of the assignment can be carried out either internally (without relying on external information) or by leveraging on pre-existent knowledge about the input data (external, or semi-supervised). A simple and effective way to validate the clustering efficacy internally is

the silhouette coefficient[64], although several other metrics are available: Davies-Bouldin index or the Dunn index are two common examples.
The silhouette value is a measure of how similar an object is to its own cluster (cohesion) compared to other clusters (separation), and can be measured by calculating first:

- The mean inter-cluster distances $a(i)$, for each data point $i \in S_i$, where S_i is cluster i:

$$a(i) = \frac{1}{|S_i| - 1} \sum_{j \in S_i, i \neq j} d(i,j)$$

- The minimum among the mean intra-cluster distance, seen from point $i \in S_i$ to any other cluster S_k (it can be thought of as a proxy of the distance from the closest cluster):

$$b(i) = \min_{k \neq i} \frac{1}{|S_k|} \sum_{j \in S_k} d(i,j)$$

- Finally, the silhouette is calculated as :

$$S(i) = \frac{b(i) - a(i)}{\max\{b(i), a(i)\}}$$

The value of $S(i)$ peaks at 1 when the point would not be better assigned to any other cluster, and is equal to -1 in the opposite scenario.

2.3.4 Density based Clustering

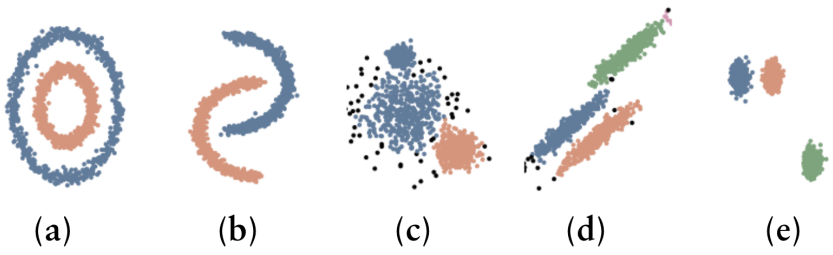


Figure 2.5 – DBSCAN clustering assignment applied to a set of five different tests. The two (three in case (c),(d),(e)) different colours define the clusters found. The advantage of following the density of points to trace the boundaries of the clusters renders DBSCAN particularly performing in the examples shown so far. The points being part of boundary regions are coloured black and labelled as noise. The main weak point of DBSCAN clustering scheme is its reliance on a constant density of points: one can see how already in (c), the sparser sampling in the regions of the main globe cannot be detected different from noise.

There are cases in which a-priori estimation of the number of clusters can be ill-defined or defeat the purpose of the clustering in the first place. In these scenarios, it is common to

adopt schemes that separate regions of points depending solely on their density distribution. Such approaches have the advantage of being able to find the number of distinct clusters automatically, although they need a parametric reference to the minimum density required for a cloud of points to qualify as a cluster.

A popular algorithm to perform density-based clustering is DBSCAN[65], which labels points closely packed together into different clusters and detects outlying samples in sparse regions as noise. The critical parameters in DBSCAN are ϵ and minPts, which indicate the cutoff distance connecting two points and the minimum number of points defining a dense neighbourhood, respectively. Combining these two parameters with the pairwise distances between the dataset's samples, one can tag every point on whether they belong to core points, reachable points (boundaries) and noise (outliers). A core point is defined as a point which has at least minPts within ϵ distance from itself, while a reachable point is one which contains at least one core point in its neighbourhood. Finally, all points with less than minPts, non-reachable points, below ϵ are considered to be noise.

Once every point is labelled, a cluster is formed by all the core points and their corresponding reachable points.

The advantage of clustering through DBSCAN is that one can group points belonging to manifolds of any shape, provided they are sufficiently connected. It is a popular choice for clustering due to its inherent robustness to outliers and the ease of interpretability of its parameters. A limitation of DBSCAN lies in its ability to deal with distribution with regions of varying densities. The solution to this problem has been introduced in the HDBSCAN algorithm [66], which bypasses the definition of a ϵ distance by performing a cluster that is solely dependant on the number of minPts defined.

2.4 Machine learning in atomistic systems

The techniques introduced so far have been developed for a general class of problems, but they all share the assumption that a particular set of inputs can be described in terms of collection of vectors of numerical properties, often called feature matrix, "**X**". One of the first critical issues in applying the techniques seen so far to a system described by a collection of atoms lies in finding an efficient description of its features in terms of input vectors.
The choice of the optimal representation is a crucial step in any machine learning approach since it establishes which degrees of freedom of the systems the algorithm will capture. As such, the descriptors play an active role in the analysis of the data and must provide an injective mapping from the input to the feature space (i.e. different inputs corresponds to different features). Before attempting to define approaches to extract a descriptor matrix from atomistic simulation outputs, it is important to define what type of data structure is used commonly. The vast majority of the data which is of interest of atomistic modelling of materials deals with a collection of molecules or atoms described in terms of their Cartesian positions and the species (elements) they contain.
The description of solids makes use of the periodicity of the system to represent the whole material in terms of its unit cell[1], while a finite-size system (often called "cluster") is defined as a set of coordinates without an explicit box bounding them.
We can thus define a configuration (or frame) as $\mathbf{P} = \{(x_i, y_i, z_i), z_i\}_{i=1..N}$, where i denotes the i-th atom, (x_i, y_i, z_i) its i-th cartesian coordinates and z_i its species atomic number. In presence of a unit cell denoting an infinite periodic system, one will expect an additional term $\mathbf{h} = (\mathbf{a}, \mathbf{b}, \mathbf{c})$ which contains the three cartesian vectors spanning the unit cell volume.
The scope of this thesis revolves around the characterisation of patterns existing in a material's atomistic description using machine learning techniques. Still, before one can apply the existing algorithms out of the box, it is important to understand how to best represent (i.e. vectorise) an ensemble of atoms into a feature matrix **X**. In the following subsections, we list a group of successful approaches to featurisation of atomistic inputs, together with their limits and advantages.

2.4.1 Cartesian coordinates and symmetries

The most basic form of feature one can assign to a configuration can be obtained by considering the list of its atomic coordinates. This approach maps any arbitrary configuration $\mathbf{P}_i \rightarrow \mathbb{R}^{3N_i}$, where N_i corresponds to the number of atoms contained in the i-th frame. If the analysis one has to carry out expects the same set of atoms in the same order and reference frame, then it is possible to define a measure of similarity between two elements as the root mean square displacement of each atom.

[1] Disordered systems are usually considered to be periodic as well, although with a large enough unit cell to minimise the spatial correlations between each replica. In case of a crystal, the unit cell is chosen as its crystal's primitive cell.

While simple in its implementation, such an approach is still widely used as an elementary fingerprint to characterise structural motifs in biomolecular complexes [67]. However, this representation presents several limitations: structures with different numbers of atoms lie in different spaces and rigid translations or rotations of the reference system produce different features for the same effective configuration.
It is clear that in most practical cases, one finds it useful to compare configurations where it can be hard to define a fixed common reference set, or where in general , the number of atoms and species can change. Depending on the machine learning task at hand, the choice of the most suitable descriptor is often non-trivial.

The essential difference between domain-oriented machine learning models and general algorithms is that often, the potential energy surface (i.e. the feature-to-property relation) one wants to reproduce respects several well-known properties. These properties can be encoded in the feature vectors themselves, allowing to model to focus only on the degrees of freedom we are interested in modelling.
To offer a concrete example: when comparing two identical molecules of water, our algorithm should be able to consider them identical irrespective of where the origin of their Cartesian reference is, whether one is rotated with respect to another, or whether atoms labels are swapped in order. Such set of conditions make the choice of Cartesian coordinates as a set of descriptors a rather undesirable choice, and gave rise to a particularly prolific branch of research devoted to finding symmetry adapted descriptors for atomistic systems.

2.4.2 Local descriptors

The concept of *chemical bond* introduced by Pauling [68] and Kohn's principle of *near sightedness* [69] have in common the observation that fundamental local properties (such as the electron density n(r)) are most affected by their surrounding potential only at nearby points. This important notion makes it possible to reconsider the exercise of reconstructing the property of a larger ensemble of atoms in terms of a combination of contributions coming from the local environments that constitute it.
Inspired by this fundamental principle, many approaches to obtain a description of atomic systems start from breaking down the configurations into collections of so-called "environments", i.e. atom centred subvolumes of the atomic arrangements which cover until the desired cutoff length r_c. For the sake of simplicity, although many schemes exist to subsample a portion of space containing a distribution of points, we cover only the ones using spherical environments, centred around each of the frames' atoms.
Before we start entering into the details of some of the most used schemes in the field of local descriptors, we introduce a series of terms that are going to use in this work.
What we called so far as an "environment" $\mathscr{X}_i$, is defined as the collection of atomic coordinates appearing within a sphere of cutoff r_c centred in the atom i, while to denote atoms from a specific species, we use lower case greek letters. One of the first approaches to describe local

environments comes from the development of a metric capable of distinguishing local order in polycrystalline or amorphous systems.

Coordination number and Steinhardt Parameters

At its most basic implementation, one could already grasp the features of the particular system at hand by merely counting the number of atoms contained in the environments of interest. By calculating what is commonly referred to as "coordination number" $N(\mathscr{X})$ for an environment $\mathscr{X}$, it is possible to distinguish crystalline configurations or, in the context of solidification of materials, separate environments being part of the different phases coexisting in the simulation cell.
Such a simple scheme, however effective, fails to grasp differences between similarly dense environments, where local ordering becomes a matter of particular angular distributions, such as the case of water and ice polymorphs.
To this end, Steinhardt et al. introduced a set of local bond order parameters which are based on the complex vectors $q_{lm}(i)$:

$$q_{lm}(\mathscr{X}) = \frac{1}{N(\mathscr{X})} \sum_{j=1}^{N(\mathscr{X})} Y_{lm}\left(\mathbf{r_{ij}}\right) \tag{2.19}$$

In this expression, Y_{lm} denotes the spherical harmonics with quantum numbers l and $m \in [-l, l]$, and $\mathbf{r_{ij}}$ describes the distance vector joining the central atom i to the j-th atom in the environment $\mathscr{X}$.
By calculating the root mean square over the l numbers, one can finally calculate the environment's Steinhardt order parameter :

$$q_l(\mathscr{X}) = \sqrt{\frac{4\pi}{2l+1} \sum_{-l}^{l} |q_{lm}(\mathscr{X})|^2} \tag{2.20}$$

These descriptors report solely information on the angular correlations contained within their cutoffs, whose sensitivity can be tuned according to the l of choice. This information is, however, extremely efficient in distinguishing ordered, crystalline environments from disordered ones, making Steinhardt parameters still an effective collective variable choice for enhanced sampling techniques focusing on simple phase transitions.

Atom Centered Symmetry Functions

The full characterisation of a 3D distribution of points is far from a trivial computational task in that it needs to capture two essential correlations between each of the points: their radial and angular distributions. One of the first descriptors developed to quantify both radial and

angular correlations came from the work of Behler and Parrinello [70], as an attempt to find a succinct representation of atomic systems to be used as input in a Neural Network scheme to model systems potential energy surfaces. In this work, the authors introduced a family of functions acting on the atomic coordinates called "symmetry functions" (SF), which capture correlations between atoms in a neighbourhood of a central atom indexed i. All the functions involve the use of a cutoff functions which weighs every atomic contribution with a decaying function of its distance from the center of choice, up until a cutoff radius R_c, here reported according to the formalism introduced in [70]:

$$f_c\left(R_{ij}\right) = \begin{cases} \tanh^3\left(1-\frac{R_{ij}}{R_c}\right) & \text{for } R_{ij} \leq R_c \\ 0 & \text{for } R_{ij} > R_c \end{cases} \tag{2.21}$$

We define the distance R_{ij} as the euclidean distance between the center atom i with the j-th atom of the frame. Once defined such a cutoff function, all the remaining symmetry functions combine it with different functional forms that probe the angular and radial correlations within the spherical environment around the atomic centres. While several SFs have been introduced over the years, we limit ourselves to the description of the two most used ones, the so-called G^2 and G^3 forms.
The G^2 function calculated for atom i is defined as:

$$G_i^2 = \sum_{j=1}^{N_{atom}} e^{-\eta\left(R_{ij}-R_s\right)^2} \cdot f_c(R_{ij}) \tag{2.22}$$

The role of this radial descriptor is to provide an overview of the density of atoms distributed at a distance from the central atom R_s by means of a Gaussian probe of width depending on η.It is a 2-body function that requires calculation of the pairwise distances of the atoms present in the frame, and depends on solely 2 continuous parameters.
To appreciate the angular correlations present in an environment, one needs to compute the 3-body correlation function G^3, which functional form is defined as follows:

$$G_i^3 = 2^{1-\zeta} \sum_{j\neq i} \sum_{k\neq i,j} \left[\left(1+\lambda\cdot\cos\theta_{ijk}\right)^{\zeta} \cdot e^{-\eta\left(R_{ij}^2+R_{ik}^2+R_{jk}^2\right)} \cdot f_c(R_{ij}) \cdot f_c(R_{ik}) \cdot f_c(R_{jk})\right] \tag{2.23}$$

Similarly to the radial counterpart, this SF focuses on a portion of space which is Gaussian-weighted (through the parameter η) on the pairwise distances between the three atoms involved in the triplet. Moreover, it introduces an angular sensitivity through the introduction of the angle θ_{ijk} formed by the scalar product between distance vectors $\hat{\mathbf{R}}_{ij} \cdot \hat{\mathbf{R}}_{ik}$. The angular resolution can be tweaked by modifying the two parameters $\eta \in \mathbb{R}$, which regulates the angular width of the probe, and $\lambda \in \{-1, 1\}$, which switches the octants where cosine function hits its maximum. The calculation of the G^3 function is computationally more involved as it requires the enumeration of the triplets present in the desired environment, and depends on three parameters. While a single SF cannot capture all the spatial correlations encoded in a collection of frames, it is customary to use a collection of different parametrisations of both G^2 and G^3 to produce an ensemble of angular and radial descriptions of each atomic

neighbourhood.
The result of such an approach is an atomic feature vector $\mathbf{X}_{SF}$ which contains N entries, corresponding to the number of symmetry function choices found to be suitable for the system. In practice, the SF selection can be seen as selective non-uniform gridding of the neighbourhood of every atom. This consideration implies the need for the users to know which angular and radial lengthscales encode the correlations relevant to the applications.
The SFs have been introduced to model atomic environments to be provided as inputs to feed-forward neural network-based machine learning potentials (MLP) [70], and later have been used to successfully build a classifier for local environments in polymorphic systems [71]. Nowadays they still serve a central role in the development of MLPs owing to an efficient implementation offered in the n2p2 package [72], and an overhauled functional form that reduces computational cost related to the scaling with the species space [73].
While SFs can produce an intuitive description of an atom's neighbourhood, they both require an a-priori knowledge of the systems' relevant length scales, and a large number of parameters to perform model selection onto. A recent solution to automatise the search of a relevant set of SF has been proposed by Imbalzano et al [74], where a deterministic CUR selection scheme is used to subsample the set of SFs that capture the most variance throughout the configurations available in the set.

Smooth Overlap of Atomic Positions

A different approach that has been introduced by Bartok et al. [15] aims at reproducing a quantitative description of the similarity kernel between two different environments by building probability density fields as proxies of atomic arrangements, and later compare them by measuring their overlaps.
The result of this scheme is called smooth overlap of atomic probabilities (SOAP), and allows to accurately map the correlations between the atoms in an environment onto a high dimensional space which is sensitive to both the angular and radial information. To build the SOAP kernel, the starting hypothesis is to define the atomic density $\rho_{\mathscr{X}}$for an environment $\mathscr{X}$ centred around an atom i as a sum of Gaussian density distributions centred around each atom's positions with width σ:

$$\rho_{\mathscr{X}}(\mathbf{r}) = \sum_{i\in\mathscr{X}} \exp\left\{\left(-\frac{\|\mathbf{r}-\mathbf{r}_i\|^2}{\sigma^2}\right)\right\} \tag{2.24}$$

This represents a smooth proxy of the atomic arrangements and which is both permutationally and translationally invariant. It serves as building block to build a measure of similarity between two environments $\mathscr{X}$,$\mathscr{X}'$, which can be obtained by calculating the overlap between the two densities across all the possible 3D rotations through the operator $\hat{\mathbf{R}}$, as depicted in Fig.2.6:

$$k(\mathscr{X},\mathscr{X}') = \int d\hat{\mathbf{R}} \left| \int \rho_{\mathscr{X}}(\mathbf{r})\rho'_{\mathscr{X}}(\hat{\mathbf{R}}\mathbf{r})d\mathbf{r} \right|^n \tag{2.25}$$

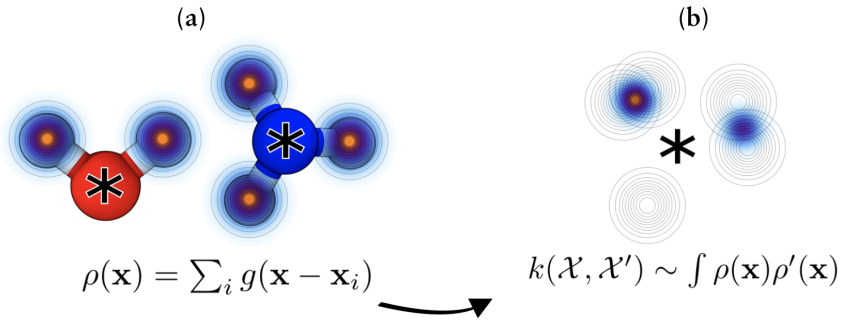


Figure 2.6 – The smooth overlap of atomic position description assumes to decorate each atomic coordinate with a species dependent gaussian distribution. This way one can represent the collection of atoms within an environment in terms of a density field (a). The overlap between these densities evaluated over all the possible pairwise rotations defines thus the similarity between the two structures (b).

It is important to note that, although the rotational integral introduces a sensitivity to the angular correlations, to maintain them in the final kernel, one has to impose an exponent $n \geq 2$.[2] It is customary to redefine a normalised kernel so that the self-similarity maxes out to 1, by writing simply:

$$\hat{k}(\mathcal{X}, \mathcal{X}') = \frac{k(\mathcal{X}, \mathcal{X}')}{\sqrt{k(\mathcal{X}, \mathcal{X})k(\mathcal{X}', \mathcal{X}')}} \tag{2.26}$$

While the numerical evaluation of the integral across all rotations is far from trivial, it can be calculated analytically by representing the density field in terms of orthogonal radial basis function and spherical harmonics.
By rewriting the atom density as :

$$\rho_{\mathcal{X}}(\mathbf{r}) = \sum_{i \in \mathcal{X}} \exp\left\{\left(-\frac{\|\mathbf{r} - \mathbf{r}_i\|^2}{\sigma^2}\right)\right\} = \sum_{nlm} c_{nlm} g_n(\|\mathbf{r}\|) Y_{lm}(\hat{\mathbf{r}}) \tag{2.27}$$

This allows to evaluate the similarity kernel in the form of a dot product of the rotational invariant power spectra:

$$p(\mathcal{X})_{nn'l} = \sum_m (c_{nlm})^\dagger c_{n'lm} \tag{2.28}$$

The unit vector $\mathbf{p}(\mathcal{X})$ obtained by collecting all its components is the SOAP feature vector, with the SOAP kernel amounting to the linear kernel calculated between two different envi-

[2] It can be trivially seen that, in case of an order $n = 1$ kernel, the integral order can be swapped, and thus the rotational dependence integrated out. The result of such construction would be a radial kernel, akin to the overlap of the two system's radial distribution functions.

ronments' power spectra:

$$k(\mathscr{X},\mathscr{X}') = \mathbf{p}(\mathscr{X})\cdot\mathbf{p}\left(\mathscr{X}'\right) \tag{2.29}$$

While this formulation is in theory exact, in practice one limits the depth of the radial and angular expansion to the number of components that can fit into the memory requirements, obtaining an approximation of the overlap between the two atomic arrangements.
The cases analysed so far have been derived for a class of systems where there is no chemical degree of freedom: there is no difference in species types between the environments under scrutiny.
In more realistic scenarios, one usually needs to perform a comparison between environments which can include many different chemical species, here defined $\alpha^1,\alpha^2..\alpha^N$. In these scenarios, De et al [21] have introduced a strategy to account for chemical diversity, by first establishing the total number of different elements N_e present in the dataset.
Once this is available, one can build a collection of "partial" power spectra called $p^{\alpha,\alpha'}(\mathscr{X})$ which are based on the density spawned by the species pairs α,α'.To then build the final environmental kernel it suffices to combine them all:

$$k(\mathscr{X},\mathscr{X}') = \sum_{\alpha\alpha'}\mathbf{p}^{\alpha\alpha'}(\mathscr{X})\cdot\mathbf{p}^{\alpha\alpha'}\left(\mathscr{X}'\right) \tag{2.30}$$

In which the new partial power spectra element is defined as :

$$p(\mathscr{X})^{\alpha\alpha'}_{bb'l} = \sum_m \left(c^{\alpha}_{b,lm}\right)^{\dagger} c^{\alpha'}_{b'lm} \tag{2.31}$$

The key advantage brought forth by SOAPs is their physical interpretability, as the parameters underlying the descriptors generation require solely the definition of the dimension of the environment (their cutoff length) and the width of the Gaussian caps deposited into every atomic site.
SOAPs found their first and principal use as the foundation for the Gaussian approximation potentials [75], a class of MLP which models systems PES by interpolating between the training set information using Gaussian process regression. The success of SOAPs has however gone beyond their use in MLP, becoming a key element used for the regression of local properties of atomic systems such as NMR shieldings in molecular crystals [17] or in the screening of corrosive materials [76] to name a few recent examples.

Atom density representations

Both the methods described thus far attempt at a description of a local environment by sampling the density distribution of its atoms: these approaches, in fact, fall into the category of the so-called density-based descriptors [77].
Over the past decade, several different schemes involving sampling an environments density

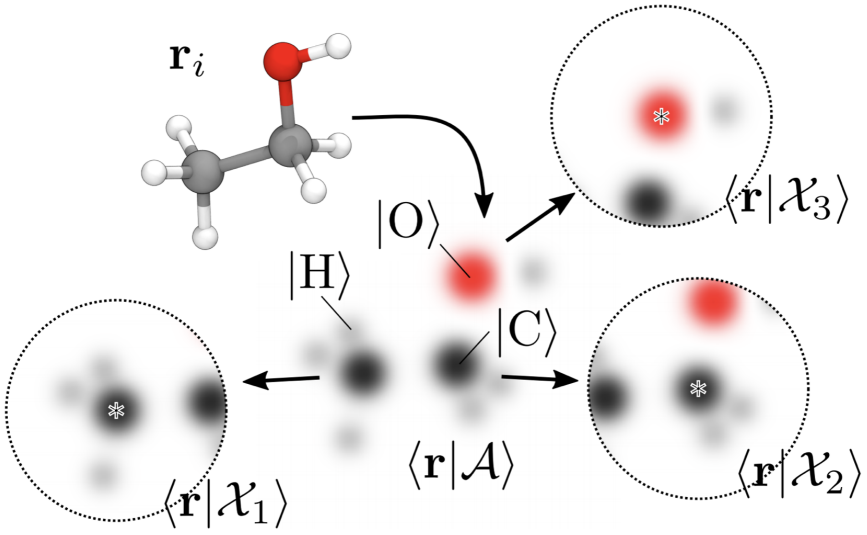


Figure 2.7 – The molecule of ethanol can be represented in terms of the atom density descriptors of the environments it is composed of in the real-space $\langle r|$ basis. The separation of the different chemical channels is achieved by decorating the functions with elemental kets.

distribution have emerged, but only recently they have been shown to be representable in the same formalism, allowing for a fully general density representation scheme to be developed. In the framework of the density-based descriptors [78], one can define the usual frames containing all the atoms as the ket $|\mathscr{A}\rangle$. If we place a smooth localised function (i.e. a gaussian) on every atom i and decorate it with an orthonormal ket $|\alpha_i\rangle$ to denote each species (as shown in Fig.2.7, one can represent $\mathscr{A}$ in the position space as :

$$\langle \mathbf{r}|\mathscr{A}\rangle = \sum_{i\in\mathscr{A}} g\left(\mathbf{r_i}-\mathbf{r}\right)|\alpha_i\rangle \tag{2.32}$$

By the same token, one can represent an environment $\mathscr{X}_i$ centered around an atom i as :

$$\langle \mathbf{r}|\mathscr{X}_i\rangle = \sum_j f_c\left(r_{ij}\right) g\left(\mathbf{r}_{ij}-\mathbf{r}\right)\left|\alpha_j\right\rangle \tag{2.33}$$

Such environment can be represented in any basis set of choice, here shown for example the case of an expansion on spherical harmonics and radial basis functions $g_n(r)$:

$$\langle \alpha nlm|\mathscr{X}_i\rangle = \int d\mathbf{r}\,\langle nlm|\mathbf{r}\rangle\,\langle \alpha\mathbf{r}|\mathscr{X}_i\rangle = \int dr d\hat{\mathbf{r}} r^2 g_n(r)\, Y_m^l(\hat{\mathbf{r}})\psi^{\alpha}_{\mathscr{X}_i}(r\hat{\mathbf{r}}) \tag{2.34}$$

Where α denotes the species channel (i,e. H, C..etc) and $\psi^{\alpha}_{\mathscr{X}_i}(r\hat{\mathbf{r}})$ the density field generated by the distribution of atoms of type α. So far, this description presented in eq 2.33 is not invariant to rotation of reference system.
An operative way to introduce the invariance to the desired operator is to symmetrise the

descriptor by performing a Haar integration (i.e. an averaging over the symmetry group $\hat{S}$):

$$\left|\mathscr{A}^{(1)}\right\rangle_{\hat{S}} = \int \hat{S}|\mathscr{A}\rangle \, d\hat{S} \tag{2.35}$$

To achieve a rotational invariant description of an atomic environment $\mathscr{X}$, it is sufficient to peform the Haar integral of its ket $|\mathscr{X}\rangle$ over the $SO(3)$ rotation group:

$$\left|\mathscr{X}^{(1)}\right\rangle_{\hat{R}} = \int d\hat{R}\hat{R}|\mathscr{X}\rangle \tag{2.36}$$

The drawback of performing such averaging operation directly on the descriptor's ket is to end up losing all the angular (or in general, higher body-order) correlations present the environment. It has been shown in ref [78] that, in order to include high order correlations between atomic positions one can take tensor products of the structural ket before performing the Haar integral.

$$\left|\mathscr{X}^{(\nu)}\right\rangle_{\hat{R}} = \int d\hat{R} \prod_{i}^{\nu} \hat{R} \otimes \left|\mathscr{X}^{(i)}\right\rangle \tag{2.37}$$

The result of this operation is a symmetrised representation vector, which encodes correlations up to the $(\nu)+1$ body orders. To obtain a descriptor akin to the SOAP introduced in the previous examples, one needs to produce a description capturing angular information thus needs to calculate the Haar integral using an exponent of $\nu = 2$. To obtain the previously described power spectra p between two chemical species α, α', the integral becomes:

$$K^{(2)}\left(\mathscr{X}_i, \mathscr{X}_j\right) = \int d\hat{R} |\sum_{\alpha n l m} \langle \mathscr{X}_i | \alpha n l m \rangle \left\langle \alpha n l m | \hat{R} \middle| \mathscr{X}_j \right\rangle|^2 = \sum_{\alpha n \alpha' n' l} \left\langle \mathscr{X}_i^{(2)} \middle| \alpha n \alpha' n' l \right\rangle \left\langle \alpha n \alpha' n' l \middle| \mathscr{X}_j^{(2)} \right\rangle \tag{2.38}$$

In this expression, one can distinguish the right hand form to be identical to the dot product of the previously defined SOAP power spectra $p^{\alpha\alpha'}_{nn'l}(\mathscr{X})$ between two species α, α' on the envirnment $\mathscr{X}$.
The advantage of such notation is to provide both a flexible environment where to represent multiple different density descriptors, and to allow for an operative way to introduce symmetries and desired body order correlations in the descriptors in a clear manner.

2.4.3 Global descriptors

The methods described so far have been developed to characterise inherently short-ranged correlations, and are defined around the atoms composing the structures under study.
The vast majority of properties of interest of materials science are however global in nature, that is, depend on the concerted interactions of the overall configuration at hand: measures of stability of a particular crystalline system, or the radius of gyration of a supramolecular

polymer to list a couple of simple examples.
This consideration leads many of the local schemes to be impractical for direct probing of features that happen at scales that farther sighted than their cutoffs, and require either complex manipulations of local representations or different approaches altogether.

Coulomb Matrix

An early attempt at featurising a global arrangement of atoms was introduced by Rupp et al[14] with the use of Coulomb Matrices.
In their work, the authors propose a molecular descriptor based on the atoms coordinates and their nuclear charges which is based on the matrix of the Coulomb repulsions existing between the atoms in the configuration:

$$m_{ij} = \begin{cases} 0.5Z_i^{2.4} & \text{for } i = j \\ \frac{Z_i Z_j}{\|R_i - R_j\|} & \text{for } i \neq j \end{cases} \tag{2.39}$$

The diagonal terms of this matrix are polynomial forms of the nuclear charge fitted onto the atomic energies while the off-diagonal terms represent the pairwise Coulomb interaction between the atoms of the molecule .
This formulation is, however, still incomplete, as still susceptible to changes in case of permutation of the atoms' labels. To overcome this limitation, the authors propose to either adopt the eigenvalue list ϵ as a feature vector or taking the flattened upper triangle of the M coulomb matrix, padding with zeros every vector where the number of atoms is different. While both the solutions address the issue of permutational invariance, they introduce two drawbacks respectively: the eigenvalues vector is not guaranteed to be unique, while the sorting introduces discontinuities in the representation.

Bag of Bonds

Hansen et al introduced an incremental improvement to the Coulomb matrix. [13], by redistributing the previous descriptors' values into a feature vector called "Bag of Bonds".
In this formulation, every pair of elements $\alpha\beta$ forms a "bag" (i.e. HO, HH..etc), that is a bin which contains the values of the CM corresponding to every $\alpha\beta$ pair present in the molecule, in descending order. Every bag is of a fixed size, established by the greatest bag size across the set, requiring zero-padding for every molecule smaller than the reference size.
Once the bags are filled, they are concatenated to form the Bag of Bond of a given molecule. To build this in an automatically sorted manner :

$$f_{BoB}(x, z_1, z_2) = \sum_{i,j=1}^{N_a} \delta\left(x - Z_i Z_j d_{i,j}^{-1}\right) \delta(z_1, Z_i) \delta(z_2, Z_j) \tag{2.40}$$

By running over the $N_a \times N_a$ pairs of atoms i, j in the frame, the bag of bonds descriptor accumulates the dirac deltas of all the inverse distances into a continuous function, whenever the selected Z_i, Z_j pairs are matching the bag of choice. By simply sampling the non-zero elements per pair, one can obtain the discretised representation used in ref [13].

Many body tensor representation

A recent substantial upgrade to the BoB descriptors has generalised their formulation to higher body order pairs, introducing the so called Many Body tensor representation[79]. One can already see how the BoB defined above can be collected into a $N_e \times N_e \times X_{\text{step}}$ tensor, where N_e represents the number of elements in the system, and X_{step} the size of the discretised radial distances axis.
They first introduce a spatial smoothing function $\mathscr{D}(x, g_2(i, j))$ in place of the Dirac delta's, capturing any functional relation $g_2(i, j)$ between the pair of atoms i, j.
Further, rather than using a Kronecker δ to select the desired species pairing defining the 2 body bag, a matrix of elemental correlation $C \in N_e \times N_e$, with N_e being the number of chemical species present in the dataset.
The final form of the MBTR descriptor, acting on a series of bags $\mathbf{z} = z_1, z_2, .., z_n$ is defined as follows:

$$f_{\text{MBTR}}(x, \mathbf{z}) = \sum_{\mathbf{i}}^{N_a} w_k(\mathbf{i}) \mathscr{D}\left(x, g_k(\mathbf{i})\right) \prod_{j=1}^{k} C_{z_j, Z_{i_j}} \tag{2.41}$$

The function $w(\mathbf{i})$ can be used to map a scalar weight to different bags composed of the tuples $\mathbf{z} \in \mathrm{N}^k$ (with corresponding indices $\mathbf{i} = (i_1, .., i_k)$, with different functional forms depending on the choice of the body order k. [3]
Once the descriptors are calculated, they are first discretised on the x axis into a sufficient number of bins N_x, so that one can calculate the similarity between two configurations by applying any desired kernel function to the resultant tensors of $N_e^k \times N_x$ dimensions.

Local Environments Mixtures

A different approach in generating a global description for any given configuration relies on creating it from a mixture of local descriptors of its environments.
Such strategy allows to extend the use and the accuracy of the techniques of subsec. 2.4.2 to the description of molecules and solids.
One of the most successful approaches to produce global features starting from environmental ones was developed by De et al [21], by creating e measure of global similarity between two

[3] The function g_k is usually suggested to count atoms in single body descriptors, inverse distance in $k = 2$, angles for $k = 3$ and dihedrals in $k = 4$

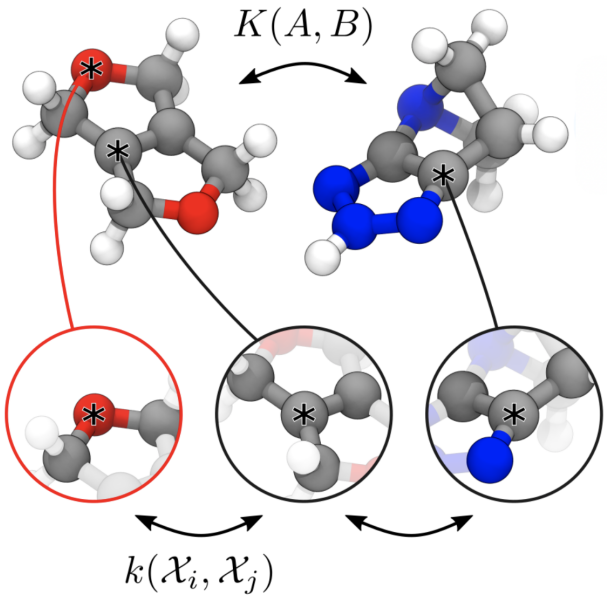


Figure 2.8 – The similarity between the two molecules in figure is casted in function of the pairwise similarity between the environments they are composed of. Depending on the choice mixing rule adopted to combine the kernels between the local structures one can obtain different global similarity kernel K.

structures using the SOAP fingerprints of their respective environments. In their work, the authors first introduce several ways to combine the environments into structural kernels (see Fig.2.8) and then show their efficacy in both regression tasks and partitioning of different crystalline configurations.
By first considering two different configurations A, B composed of atoms N_A and N_B, one can define the (generally) rectangular correlation matrix:

$$C_{ij}(A,B) \;=\; k(\mathscr{X}_i^A, \mathscr{X}_j^B) \tag{2.42}$$

Where $\mathscr{X}_i^A$ ($\mathscr{X}_j^B$) defines the i-th (j-th) environment of frame A (B). The first, straightforward way to combine all the information contained in $C_{A,B}$ consists of taking the average of all the pairwise similarity of their respective environments. Such approach defines an average structural kernel :

$$\bar{K}(A,B) \;=\; \frac{1}{N_A N_B}\sum_{i,j} C_{ij}(A,B) \;=\; \left[\frac{1}{N_A}\sum_i \mathbf{p}\left(\mathscr{X}_i^A\right)\right]\cdot\left[\frac{1}{N_B}\sum_j \mathbf{p}\left(\mathscr{X}_j^B\right)\right] \tag{2.43}$$

There are many advantages to this formulation. Firstly, one can save a considerable amount of memory by simply storing the average SOAP power spectrum per structure and secondly, it ensures a proper metric since it's defined in terms of a dot product of two virtual power spectra.
Another physical advantage of such description lies in its linearity with respect to the atomic

composition of the two configurations, a feature that is desirable when building predictive models of properties which can be thought as composition of local contributions. The obvious drawback of such formulation lies in its lack of uniqueness with respect to two different configurations that appear to have the same average features.
To bypass these limitations, the authors introduced a regularised entropy match (REMatch) kernel, which exploits optimal transport theory[55] to define a parametrised kernel mixing rule that can tune the similarity between a fully averaged model to a best matching scheme (i.e. where similarity between two configurations is assessed based on the most similar environments).
In the simple case of a system with an equal number of atoms (although extendable to differently sized systems in [21]), one can start by defining the Best match kernel as :

$$\hat{K}(A,B) = \max_{\mathbf{P}\in\mathscr{U}(\mathscr{N}_A,\mathscr{N}_B)} \sum_{i,j} C_{ij}(A,B)\,P_{ij} \tag{2.44}$$

In this equation the search is carried out in set $\mathscr{U}(N_A,N_B)$, containing all the $N_A \times N_B$ doubly stochastic matrices, with normalized rows and columns so that they sum to $1/N_A, 1/N_B$.
In the REMatch kernel, the final similarity measure is obtained by performing a weighted sum between the elements of the C_{ij} matrix:

$$\hat{K}^{\gamma}(A,B) = \left[\mathrm{Tr}\{\mathbf{P}_{\gamma} C(A,B)\}\right]^{\zeta} \tag{2.45}$$

The ζ factor is an exponent that can increase the sensitivity of the measure to slight perturbation in positions,while the mixing $\mathbf{P}$ term is obtained by using the Sinkhorn algorithm to find the argument that minimises a regularised deviation between matching environments:

$$\mathbf{P}_{\gamma} = \underset{\mathbf{P}\in\mathscr{U}(\mathscr{N}_A,\mathscr{N}_B)}{\mathrm{argmin}} \sum_{i,j} P_{ij}\left(1 - C_{ij}(A,B) + \gamma \log P_{ij}\right) \tag{2.46}$$

The regularisation is given by the information entropy of the doubly stochastic matrix (i.e. $S(\mathbf{P}) = -\sum P_{ij}\log P_i j$), while the parameter γ allows tuning between a best match scheme (by switching the regularisation to zero) and an average kernel, in the limit of $\gamma -> \infty$. The work presented in the results section of this thesis makes extensive use of SOAP-based global similarity measures, and in case the reader is interested in a detailed description of the methods illustrated so far, we refer to [80, 81].

3 Phase Stability in Materials Science

3.1 Introduction

The fundamental knowledge that it is possible to obtain the ground state configuration of a material by finding its corresponding global minimum in its Gibbs free energy landscape (i.e. for any predetermined set of external variables) has empowered chemists and engineers with an operative way to discover novel, technologically relevant, materials.
While such minima are by definition the most stable configuration a system can achieve, it is often the case that many configurations exhibiting a meta-stable character can find critical industrial applications [82].
Although higher in energy than the equilibrium ground state configurations, many meta-stable states can be accessed under particular thermodynamic conditions, and if surrounded by sufficiently high barriers, these can become kinetically persistent on a finite timescale. This observation has extended the field of search for novel compounds beyond the global minimisation of a system's Gibbs free energy, requiring extensive enumeration and characterisation of meta-stable phases spanning remote areas of compounds' phase diagrams.
While the aforementioned approach becomes analytically intractable for complex systems, many numerical approaches have been developed with the aim of obtaining efficient enumeration of energy landscapes' local minima.
In the following chapter, we cover the thermodynamic principles underlying the concept of equilibrium between different phases, and the most common approaches to analyse it. Further, particular emphasis is placed on the geometrical approach to solving the chemical and phase equilibria problem utilising a convex hull (CH) construction. Finally, a series of improvements of the conventional CH formula are presented and explained in details. A considerable amount of the description presented in this chapter has been extracted from [31]. The author's contribution to the project was part of the theoretical development, the numerical implementation and its application to a set of examples.

3.2 Phase and chemical equilibrium

The problem of predicting the structure of a mixture at given thermodynamic conditions has a long history, and finds its modern foundation in the works of Gibbs[83].
Finding which kind of phase would appear stable at a specific temperature, pressure and composition is, in fact, a central task in chemical engineering, one that is usually solved by finding the configuration for which the system's Gibbs free energy G is minimal, under a fixed temperature T and pressure p.
To be more concrete in our formulation, let's assume to work with a closed system Σ (one where no mass transfer with the outside is allowed), with Λ being the list of chemical compounds present in the mixture and starting moles per compound $\mathbf{n^0} = \{n_i^0\}_{i\in\Lambda}$. Once the system reaches equilibrium, there will be Π different phases coexisting, each of them with a number of moles n_i^π, for the phase $\pi \in \Pi$ and i-th component. Within this framework, the goal of solving for phase and chemical equilibrium amounts to finding the unique distribution of n_i^π that rises from a given starting condition $\mathbf{n}^0, T, p$, by performing a constrained optimisation on G_Σ. The total Gibbs free energy of such a system depends on each component's chemical potential μ_i^π and can be defined as follows:

$$G_\Sigma\left(\mathbf{n}, p, T\right) = \sum_{\pi\in\Pi}\sum_{i\in\Lambda} n_i^\pi \mu_i^\pi \tag{3.1}$$

The constraint to optimise under comes from enforcing conservation of mass, which can be written in a concise manner introducing a formula matrix notation, by imposing the total number of atoms of each element as constant.
One can define b_e as the total number of moles of element e in the system, and E an array containing all the elements necessary to form the chemical compounds in Λ. The vector of $\mathbf{b} = \{b_e\}_{e\in E}$ is composed of constant values which depend in turn from the initial distribution of n_i^0. By introducing $a_{e,i}$ (often called formula matrix element) as an atom counter for the element e of the i component, one can impose the mass balance as :

$$b_e = \sum_{\pi\in\Pi}\sum_{i\in\Lambda} a_{e,i} n_i^\pi \tag{3.2}$$

And combine it with the non negativity of the mole numbers $n_i^\pi \geq 0$ to obtain two sets of linear constraints to the phase and chemical equilibrium optimisation problem. It is common to shorten the notation of the set of linear constraints by aggregating them into a succinct matrix formulation:

$$\mathbf{An} = \mathbf{b} \tag{3.3}$$

$$\mathbf{n} \geq 0 \tag{3.4}$$

Where the **b** vector contains the total number of moles of every element in the system, **A** is the formula matrix containing all the $a_{e,i}$ elements and **n** contains all the n_i^π moles per compound i in phase π.

In order to calculate the total Gibbs free energy, we need to define the chemical potential μ_i^π of every component as:

$$\mu_i^\pi = \mu_i^{0,\pi} + RT \log c_i^\pi x_i^\pi \tag{3.5}$$

Which adjusts the chemical potential of the species i in phase k from its value in a pure phase $\mu_i^{0,\pi}$ with a mixing term which depends on its molar fraction x_i^π through a coefficient c_i^π. Depending on the nature of the interaction governing the mixture of phases one can adopt different expressions for the molar fraction coefficients. In the case of ideal liquid solutions it is usually taken as the unit, while in general, in a mixture of non-ideal liquids it takes the form of an activity coefficient γ_i^π.
From this point onward, we have cast the problem of finding the equilibrium configuration of a mixture of species in terms that are addressable by standard algorithms of constrained optimisation (e.g. Lagrangian multipliers):

$$\begin{aligned} G_\Sigma(\mathbf{n}, p, T) &= \sum_{\pi\in\Pi}\sum_{i\in\Lambda} n_i^\pi \mu_i^\pi \\ s.t.\ \ \mathbf{An} &= \mathbf{b} \\ \mathbf{n} &\geq 0 \end{aligned} \tag{3.6}$$

A crucially useful trait of thermodynamic potentials is that they have been shown to be concave functions of their intensive variables and convex functions of their extensive variables [84], rendering the Gibbs free energy convex on the mole numbers $\mathbf{n}$, and thus the molar Gibbs free energy convex on the partial composition $\mathbf{x}$.
The combination of the properties mentioned above have spawned several optimisations approaches that are based on convex optimisation, formalised in ref [85], but in general require the assumption of ideal liquids mixtures, i.e. liquids whose activity coefficient is equal to one. Typically once the equations of state (EOS) are set, prior knowledge of composition or equilibrium phases can be used to compute numerically the multiphase equilibrium configuration of a variety of systems either linear programming approaches [86, 87], Quasi-Newton successive substitution [88], particle swarm [89] and genetic algorithms [90].

3.3 Geometrical solution of phase equilibrium

The intrinsic difficulties in minimising G arise from the fact that while the problem statement can be cast in an optimisation friendly form, conventional approaches can be ineffective at finding solutions. This stems from the fact G is often multi-valued or can possess multiple function definitions within the same thermodynamic boundaries (different EOS). Further, while these schemes provided excellent solutions to many industrial problems, they are not suitable to be applied in conjunction with computational structure searches, as in these cases one finds it useful to obtain a relative comparison of stability between local minima of a

system's potential energy surface, with no information available in any configuration's EOS. An alternative and simpler approach to predicting equilibrium structures can be borrowed from the works of Hildebrandt and Glasser [91], which aims at solving the minimisation problem of finding G using a geometrical approach.
The conceptual foundations of such approach stem from the original work of Gibbs on *Thermodynamic properties of substances* [83], where the author studies the "primitive surfaces" associated to the $G(\mathbf{n})_{p,T}$ (i.e. with p,T fixed) surface and determines the "derived surfaces", convex hulls of the primitive surfaces, as the loci of the equilibrium configurations. To show how one can employ such construction to obtain a solution substantially equivalent to that of a non-linear optimisation problem (i.e. Eq. 3.6), we consider a mixture containing n different species, fully described by the $n-1$ vector of molar fractions $\mathbf{X} = [x_1, ..., x_{n-1}]$. Working in an isobaric, isothermal condition, our system's Gibbs free energy at $\mathbf{X}_i$ is $G(\mathbf{X}_i)$[1]. If we consider a composition at $\mathbf{X}_i$, one would expect to find the equilibrium phase as the minimum in $G(\mathbf{X}_i)$, however, the system may be able to decrease G further by undergoing a phase decomposition (i.e. by splitting into two or more phases). To be more concrete, one can consider a system at composition $\mathbf{X}$ which splits in two phases X_1 and X_2 at fractions f_1 and $f_2 = 1 - f_1$.
Since the system's Gibbs free energy is extensive and homogeneous to its arguments, one can calculate the split system's G as :

$$G^{sys}(\mathbf{X}) \;=\; f_1 G(\mathbf{X}_1) + f_2 G(\mathbf{X}_2) \tag{3.7}$$

It is apparent that in case $G^{sys}(\mathbf{X}) < G(\mathbf{X})$, it is more convenient for the system to split into a mixture of the two phases. While simple, this example shows how, when minimising G of a system at the desired point, one must consider all the possible splits that could lower the total free energy. In practice, this amounts to consider the boundaries of the linear envelope of G, by considering its convex hull "conv$[G(\mathbf{X})]$":

$$\sum_{i=1}^{m} f_i G(\mathbf{X}_i) \;\in\; \text{conv}[G(\mathbf{X})] \tag{3.8}$$

Where the fractions f_i are non-negative and sum up to one , the hull has $m \geq 1$ vertices, and $\partial\{\text{conv}[G(\mathbf{X})]\}$ defines its boundary.
We will prove now that a system's Gibbs free energy is minimised at phase equilibrium if and only if $G^{sys}(\mathbf{X})$ lies on the boundary of the convex hull of $G(\mathbf{X})$. To prove this, we follow the proof proposed by Hildebrandt [91], and define a $\mathbf{Y} = [G, x_1, ..., x_{n-1}]$ n-dimensional object containing the composition and the $n-1$ dimensional hypersurface $G(\mathbf{X})$. The $\partial\{\text{conv}[G(\mathbf{X})]\}$ is an $n-1$ dimensional hypersurface enveloping the n dimensional polytope conv$[G(\mathbf{X})]$.
If we consider the support hyperplane $\mathbf{n}$ to $\partial\{\text{conv}[G(\mathbf{X})]\}$ in $\mathbf{X}$, three different scenarios can happen:

- $\mathbf{n}$ touches $\partial\{\text{conv}[G(\mathbf{X})]\}$ only in $\mathbf{X}$. In this case $\mathbf{X}$ is a vertex of the convex hull and G^{sys} will be minimised by a pure phase of such composition, as shown in Fig. 3.1(a).

[1] As usual, G can be multi-valued, depending on the number of phases present in our system

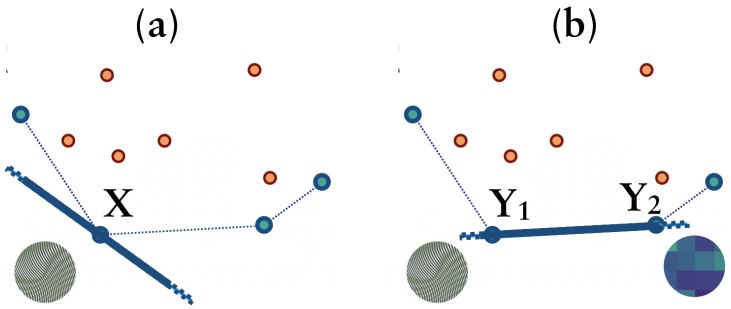


Figure 3.1 – Schematic depiction of the different way in which the support hypersurface can touch the sides of the boundaries of the convex hull of G. In case (a) there is a single point of contact, indicating a single solution corresponding to a stable pure phase X. In figure (b), in case the hypersurface touches twice the convex hull, the resulting equilibrium configuration will be a mixture of the two phases Y_1, Y_2

- **n** touches $G(\mathbf{X})$ in two points, then the corresponding line $\mathbf{Y}_2 - \mathbf{Y}_1$ lies in $\partial\{\text{conv}[G(\mathbf{X})]\}$, shown in Fig. 3.1(b). G^{sys} will be minimised by a mixture of the two phases $\mathbf{Y_1}, \mathbf{Y_2}$.

- **n** touches $G(\mathbf{X})$ in m points $\{\mathbf{X}_j\}_{j=1..m}$ with $m < n$, then the associated $\{Y_j\}_{j=1..m}$ span a $m-1$-dimensional hyperplane $\mathbf{S}_m \in \partial\{\text{conv}[G(\mathbf{X})]\}$, which leads to G^{sys} being minimised by a mixture of m different phases.

Let's consider the third case, and call $\mathbf{S}$ the $n-1$ dimensional hyperplane containing $\mathbf{S}_m$ with normal vector $\mathbf{N}$. Such normal vector is by construction orthogonal to all the vectors contained in the plane $\mathbf{S}_m$:

$$\mathbf{N} \cdot (\mathbf{Y}_1 - \mathbf{Y}_i) = 0 \text{ for } i = 2, .., m \tag{3.9}$$

$$\nabla [G - G(\mathbf{X}_i)] = s\mathbf{N} \text{ for } i = 1, .., m \tag{3.10}$$

Where s is a scalar, and $\{\mathbf{Y}_i\}_{i=1..m} \in \partial\{\text{conv}[G(\mathbf{X})]\}$. By recombining the two one can write:

$$\nabla [G - G(\mathbf{X}_i)] \cdot \mathbf{Y}_i = \text{constant for } i = 1, .., m \tag{3.11}$$

Which is equivalent to:

$$G - \sum_{i=1}^{n-1} \frac{\partial G(\mathbf{X})}{\partial x_i} x_i = \text{constant} \tag{3.12}$$

The above equation can be satisfied only when, across all the m phases and the $n-1$ compo-

nents:

$$\frac{\partial G(\mathbf{X}_1)}{\partial x_i} = \frac{\partial G(\mathbf{X}_2)}{\partial x_i} = ... = \frac{\partial G(\mathbf{X}_m)}{\partial x_i} \quad \text{for} \quad i = 1, .., n-1 \tag{3.13}$$

At phase equilibrium, we know that the chemical potential $\mu_i(\mathbf{X}_j)$ of species i involved in a phase j has to be identical across all the m phases:

$$\mu_i(\mathbf{X}_j) = G(\mathbf{X}_j) - \sum_{k=1}^{n-1} x_k \frac{\partial G(\mathbf{X}_j)}{\partial x_k} + \frac{\partial G(\mathbf{X}_j)}{\partial x_i} \text{ for } i < n \tag{3.14}$$

$$\mu_n^j(\mathbf{X}_j) = G(\mathbf{X}_j) - \sum_{k=1}^{n-1} x_k \frac{\partial G(\mathbf{X}_j)}{\partial x_k} \tag{3.15}$$

We can see that equations 3.11 and 3.14 lead to the same result, allowing for a purely geometric solution to be a viable necessary (and sufficient) condition to find phase stability. While mathematically sound, the geometrical approach to Gibbs free energy minimisation became quickly uncommon, due to the computational complexity behind the scaling of convex hull algorithms with the systems' complexity.

In light of the modern rise of computational resources and the development of efficient methods to compute convex hulls of curves lying in n-dimensional spaces [92], the use of geometrical solutions to the phase stability problem has become a widespread practice in the field of high-throughput materials discovery, from the materials project platform [93], leading to the discovery of superionic states[94] and novel meta-stable high pressure phases[95, 96]. These examples are just the tip of a long list of scientific works which exploited this construction. Such a construction is well suited to theoretical materials discovery because it provides a sufficient set of geometrical conditions to screen at once potentially stable phases on the basis of their energetics.

The mission of high throughput materials searches consists usually in identifying the ground states configuration that would emerge under a given set of thermodynamic constraints, by performing crystal structure predictions (CSP) algorithms to explore a system's configurational degrees of freedom.

One critical aspect of CSP approaches is that they often rely on global optimisations of the system's potential energy (or at best, its enthalpy) as a cheaper proxy to the Gibbs free energy, disregarding thermal effects altogether. Once several local minima of the potential energy surface have been sampled, one constructs a convex hull of all phases on composition space by building a formation energy (E-x) convex hull, and can thus determine the stable phases as the ones forming its vertices.

When introducing pressure $p > 0$, it is useful to include the molar volume information to build a CH (E-v), which can distinguish which phases would be stable at the desired pressure (reflected in molar volume). One can combine the two information to build a multidimensional hull where both the features are considered at once, allowing to find the composition which is most stable at a specific pressure.

3.4 Generalising the convex hull construction

Convex hull constructions have proven useful in numerous structure searching applications such as Refs. [97–103]. However, the conventional form has some crucial limitations. The choice of one particular feature, such as molar volume, on which the CH is constructed, relies on experimental evidence or preconceived notions of which thermodynamic constraints may stabilise structures of interest. It limits which stabilisable structures are identified, and is generally insufficient to explore the structural diversity that can be accessed experimentally through complex thermodynamic constraints such as pressure, composition, doping with guest molecules, substitution of portions of organic compounds, electric or magnetic fields, etc. (for instance, see Ref. [104]). Secondly, the conventional CH construction neglects inevitable inaccuracies in the computed (free) energies and geometries, which render the CH probabilistic in nature.

To overcome the above limitations, we introduce a probabilistic generalised CH (GCH) frame-

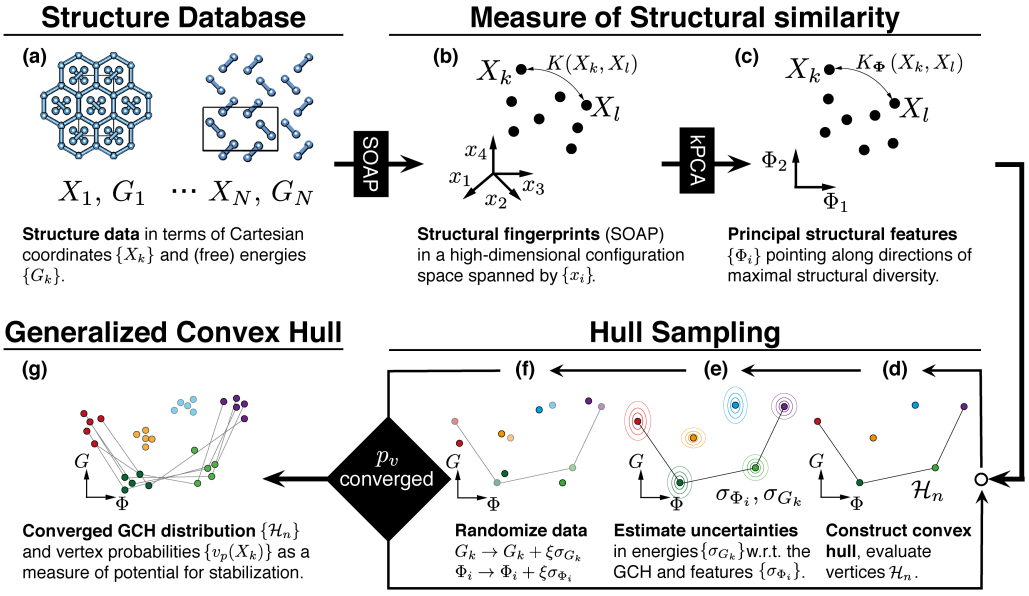


Figure 3.2 – Schematic representation of the GCH framework. X_k denotes structure k with (free) energy G_k and the associated (SOAP) structural descriptors $x_i(X_k)$ and (PCA) principal features $\mathbf{\Phi}(X_k) = \{\Phi_i(X_k)\}$. $\mathscr{H}_n$, σ_{G_k}, and σ_{Φ_i} denote the nth convex hull, the uncertainty in the (free) energy of X_k with respect to the current convex hull, and the uncertainty in Φ_i, respectively. ξ are normally distributed random numbers and $p_v(X_k)$ denotes the fraction of the sampled hulls for which $X_k \in \mathscr{H}$ (as a measure of the stabilisability of X_k).

work for evaluating the probabilities of structures being stabilised by general thermodynamic constraints. A schematic representation of this framework is shown in Fig. 3.2. It (i) quantifies the uncertainty arising from the inevitable errors in the underlying energies and structures, and (ii) rests on geometric fingerprints $\mathbf{\Phi} = \{\Phi_i\}$, which reflect the full structural diversity of the dataset. While there is considerable freedom in choosing such fingerprints, they must

exhibit an additive behaviour, that guarantees that a macroscopic sample X, which is a phase-separated mixture of different components X_k with molar fractions w_k, has a fingerprint $\mathbf{\Phi}(X) = \sum_k w_k \mathbf{\Phi}(X_k)$. A simple way to guarantee that Φ_i fulfills this requirement is to choose descriptors that are consistent with an atom-based decomposition, $\mathbf{\Phi}(X) = \sum_{\mathscr{X} \in X} \boldsymbol{\phi}(\mathscr{X}) / N_X$. Here $\boldsymbol{\phi}(\mathscr{X})$ are the fingerprints of the N_X atom-centered, local environments $\mathscr{X}$ within the structure X. Additivity ensures that any structure with features inside a convex region of D-dimensional feature-space can be decomposed into a phase-separated mixture of the $D+1$ vertices of the convex region, without changing the corresponding D features of the fingerprint describing the system (although the resultant fingerprint may differ in the remaining features). By considering the molar free energy as a function of a set of D features Φ_i, one can thus generalise the CH construction to identify the structures that are stable with respect to decomposition subject to the abstract "thermodynamic constraint" defined by a given set of D features.

3.4.1 Data-driven structure fingerprints.

For a given dataset $\{X_k\}$ we extract a small set of key data-driven features that captures most of its structural diversity by performing a kernel principal component analysis (KPCA) on a kernel measure of similarity $K(X_k, X_l)$ between pairs of structures X_k and X_l. I.e. we compute the eigenvalues λ_i and eigenvectors $\mathbf{u}^i$ of the kernel matrix, $K_{kl} = K(X_k, X_l)$, and evaluate the features of a structure X_k as

$$\Phi_i(X_k) = \sum_l u^i_l \sqrt{\lambda_i} K_{kl}. \tag{3.16}$$

These features are additive, provided that:

$$K(X_k, X_l) = \sum_{\mathscr{X}_k \in X_k, \mathscr{X}_l \in X_l} \frac{k(\mathscr{X}_k, \mathscr{X}_l)}{N_{X_k} N_{X_l}} \tag{3.17}$$

where $k(\mathscr{X}_k, \mathscr{X}_l)$ is a kernel measure of similarity between pairs of local environments $\mathscr{X}_k$ and $\mathscr{X}_l$. Loosely speaking, the resultant KPCA features $\Phi_i(X_k)$ are orthonormal measures of the similarity of the structure X_k to a particular combination of all structures in the dataset, dominated by the structurally most distinct configurations. Although the applications presented in the rest of the thesis involve the use of similarity kernels, the GCH construction is not restricted to KPCA features.

In general, any structural descriptor that is linear to its environments can be used either directly as a structure fingerprint. Alternatively, it can be cumulated in a feature matrix from which to extract, through PCA, its resulting structure fingerprints (e.g. using collections of Behler-Parrinello symmetry functions). The choice of structural descriptor is a fundamental degree of freedom of the GCH construction, and should be chosen so to capture all the structural variability the database can offer.

From a practical point of view, the GCH built in this work are always based on SOAP similarity kernels (as described in Chapter 2), by constructing a global similarity kernel obtained by

simple averaging of the databases' environmental kernels.

3.4.2 Feature selection and interpretation.

The abstract nature of these KPCA features begs the question of (i) how to identify which among them have the potential to stabilise structures and should thus be included in the GCH construction, and (ii) how to relate them to experimentally-realisable conditions. Note that the ambition to identify structures that can be stabilised by manipulating electronic properties is put on a firm basis by the Hohenberg-Kohn theorem [105], which guarantees that electronic ground-state properties correlate with structural features. When no prior knowledge of the system is available, the KPCA eigenvalue spectrum provides an indication of the maximum intrinsic dimensionality of the structure data at hand [106]. It can thus be used to choose the dimensionality of the GCH such that the full structural diversity of the dataset is explored. Even in this worst-case scenario, the resultant pool of candidates is typically orders of magnitude smaller than the underlying structure database, rendering it possible to further investigate the relations between the features of the candidates and physical observables (or thermodynamic constraints) such as density, composition, etc. This can not only help to translate abstract structural features into practically realisable synthetic protocols but also to refine the selection of features on which the GCH is constructed a posteriori to those which couple strongly to experimentally realisable conditions and thus have the greatest potential for stabilising structures.

3.4.3 Probabilistic GCH and uncertainty quantification.

So far, the GCH framework neglects the inevitable uncertainties in (computed) free energies, lattice parameters and atomic positions, and thus in the determination of the hull vertices. We, therefore, propose a probabilistic extension in which the GCH probability distribution is sampled by constructing many possible convex hulls based on free energies and geometries, which have been randomised according to their respective uncertainties. In practice we take the typical model errors on the energies ϵ and Cartesian coordinates (for example, due to the choice of density functional in density-functional-theory (DFT) calculations or the absence of quantum nuclear effects) to be known from experience or benchmarks. We estimate the resultant errors in the energies relative to the instantaneous hull, σ_{G_k}, exploiting structural correlations to account for correlations between the errors in $\{G_k\}$. In particular, we ensure that σ_{G_k} vanishes for the vertex structures and any structure that is a phase-separated mixture of the vertices of its associated simplex (its "parent phases"), while otherwise reflecting how different a given non-hull structure is from the parent phases. The rationale is that the typical errors are not random, but correlate with the structural features. Consider for instance a phase-separated mixture X_k composed of molar fractions w_{kl} of the parent phases X_l with calculated energies $G_l+\epsilon_l$. Its calculated energy is identical to the corresponding combination of the energies of the parent phases, including their errors, $\sum_l w_{kl}(G_l+\epsilon_l)$. This is exactly the

definition of the convex hull energy constructed on $G_l + \epsilon_l$, so that the energy of X_k relative to the hull will be zero regardless of the errors. Hence, σ_{G_k} should vanish.
Let us introduce a practical definition that satisfies this requirement. We estimate σ_{G_k} as the fraction of the total error ϵ associated with the deviation of the features $\Phi_i(X_k)$ from the ideal interpolation in terms of the parent phases, $\Phi_i^{\mathrm{GCH}}(X_k) \equiv \sum_{X_j \in \mathscr{H}} w_{kj}\Phi_i(X_j)$

$$\sigma_{G_k} = \epsilon \sqrt{\frac{1}{\sigma_G^2} \sum_{i=1} \left[g_i \left(\Phi_i(X_k) - \Phi_i^{\mathrm{GCH}} \right) \right]^2}. \tag{3.18}$$

Here g_i is the energetic response to changes in Φ_i, which we learn by ridge regression from a machine-learning model of G_k, and σ_G^2 is the variance of G over the entire dataset. Due to additivity, for a physical mixture, $\Phi_i(X_k) = \Phi_i^{\mathrm{GCH}}(X_k)$ for all the features, including those that are not used for the GCH construction, which ensures that $\sigma_{G_k} = 0$. On the contrary, for each point that is not a physical mixture of hull points, only the features used to build the GCH will coincide with $\Phi_i^{\mathrm{GCH}}(X_k)$. In this case, σ_{G_k} scales with the residual structural diversity that is not captured by the GCH coordinates.
Note that the dependence of the uncertainties σ_{G_k} on the instantaneous hull implies that the hull distribution must be sampled "self-consistently".
The randomisation of the features Φ_i requires knowledge of how the uncertainty in the underlying atomic coordinates and lattice parameters (or "structure parameters") propagates to uncertainties in the features, σ_{Φ_i}. At each iteration, we also randomise the features Φ_i to reflect the uncertainty in the underlying atomic coordinates. We estimate σ_{Φ_i} by randomising the structure parameters of n_r reference configurations X_r, that we choose by farthest point sampling. In practice, we randomize each reference structure n_s times, compute the features for the randomized structures $\Phi_i(X_r^s)$, and evaluate

$$\sigma_{\Phi_i} = \sqrt{\frac{1}{n_s n_r} \sum_{r}^{n_r} \sum_{s}^{n_s} \left(\Phi_i(X_r^s) - \Phi_i(X_r) \right)^2}. \tag{3.19}$$

After an extensive sampling of the GCH distribution, the rate with which each structure occurs as a vertex $p_{\mathrm{vertex}}(X_k)$ roughly quantifies how trustworthy the identification of the structure X_k as stabilisable is, and its average distance from the hull provides a measure of its (meta-)stability. One can interpret the uncertainties introduced thus far as a simplified thermal motion which allows individuation of configurations exhibiting near degenerate energies and structures. The advantage offered by such a scheme is to effectively introduce an approximate measure of the effect of entropy in the different phases' stability.

3.4.4 Coarse-graining of the GCH vertices.

In cases where large numbers of very similar structures (for example owing to stacking faults or partial disorder) compete for stability, each candidate exhibits a small individual probability

of becoming stable. However, collectively such a cluster of structures represents a stable phase. For convenience, we reduce the full list of potential vertices to representatives of each cluster, i.e. of each stable phase. These are identified by sequentially eliminating the N lowest probability candidates with a cumulative probability $\sum_{k=1}^{N} p_{\text{vertex}}(X_k) < 1$ (which guarantees that no complete cluster of structures that constitutes one stabilisable structure gets eliminated entirely in one step) from the dataset and resampling the GCH for the thus reduced dataset. This procedure is repeated until the lowest $p_{\text{vertex}}(X_k)$ is above a set threshold of 0.5. This "coarse-graining" ensures that the surviving candidates correctly accumulate the probability of becoming stable associated with their respective clusters of similar structures. Even though we only consider these marginal probabilities, the GCH directly samples the full hull distribution, which can further be used to investigate, for instance, which structures compete for stability. While the identification of experimentally-synthesisable compounds is the focus of this work, the generalised CH (GCH) framework proposed in the following also translates (at negligible computational cost) input energies into a far better measure of structural stability, namely the energy relative to the GCH. The latter can be used in place of bare energies in diverse applications, such as experimental crystal structure determination protocols or as the fitness function driving structure searches *in situ*.

3.4.5 Code Availability

The method presented thus far has been implemented in python3 and is publicly available on github[2]. The code follows a two step process, where a first routine (`gch_init.py`) initialises the GCH construction parameters (errors in cartesian coordinates and energies, as well as desired vertices minimum probability) and generates a set of new rattled configurations for the configurational uncertainty estimate. Once the shaken configurations have been successfully generated, the user is required to produce a similarity kernel between the novel shaken configurations, and the database under study. To finally start the GCH construction is thus sufficient to feed the similarity kernels generated for the novel rattled configurations into the `gch_run.py` and find the resulting stable configurations contained in the dataset.

[2] https://github.com/andreanelli/GCH

Applications Part II

4 Applications to structure searches

4.1 Introduction

The generalised convex hull construction provides a succinct and robust approach to enumeration of stable configurations, and is presented hereon as a natural complement to structure searches investigations. The key parameter in all the GCH applications is the intrinsic dimensionality of database under investigation, which reflects the breadth of structural sampling obtained by the CSP algorithm of choice, and the complexity of the underlying crystal structure landscape.

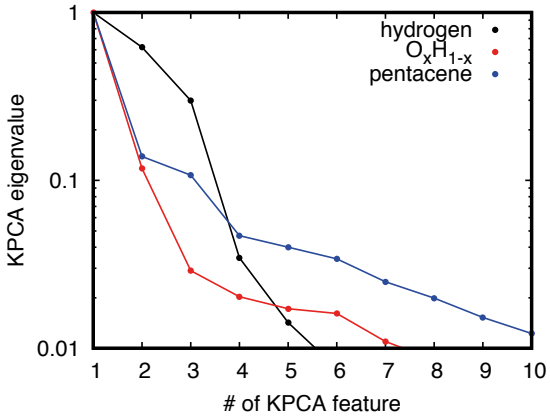


Figure 4.1 – KPCA eigenvalues for the applications we discuss in this work, namely: hydrogen (black), H_xO_{1-x} (red), and pentacene (blue), obtained from SOAP similarity kernels with $r_c = 2$ Å, 5 Å, and 5 Å, respectively.

The dimensionality of the convex hulls built in the following examples is chosen to be determined by the number of principal components features that are shown to have a sufficiently large corresponding eigenvalue. In order to obtain them, we first calculate the SOAP kernel of the whole database, and then extract the eigenvalue spectrum associated with its PCA decomposition. The result of this operation is shown in figure 4.1.

The aim of the estimation of the correct intrinsic dimensionality is to choose the minimum number of components needed to capture the largest amount of structural variance in the dataset.
To showcase the range of applicability of the GCH construction, we apply it to four problems of increasing complexity, namely a database of hydrogen solid phases at gigapascal pressure, a set of oxygen-hydrogen binary crystal structures, a subset of this database for which we demonstrate how a GCH can predict oxygen phases that are stabilized by magnetism, and a set of crystalline polymorphs of pentacene for which we investigate the effect of chemical substitutions. Finally, the last section of the chapter covers the stability of the GCH to errors in the input energy data and choice of descriptors hyperparameters.

Author Contributions

This chapter reproduces, with minor adjustments, the content of Ref. [31]. The contribution of the author of this thesis to this work is the development and the design of the method, while the calculations of the magnetisation in oxygen dimers and relaxations of the molecular crystals have been obtained by Dr. Edgar Engel.

4.2 Hydrogen at gigapascal pressure.

Hydrogen at high pressures exhibits a strongly polymorphic character, transitioning from structures formed by ordered hydrogen dimers to fully monoatomic arrangements with fundamentally different electronic structure properties.
In this jungle of monoatomic polymorphs, it has been long theorised the existence of a metallic phase which could be used as a stable room temperature superconductor [107, 108]. Over the past decade several attempts, both experimental and theoretical aimed at uncovering this relatively unknown region of its phase diagram, with the discovery of a considerable number of novel meta-stable phases ranging from phase-II to phase-VIb [109–114].
While novel experimental techniques such as diamond anvil processes have helped a great deal in realising many of the theorised configurations [115], accurate simulation of hydrogen electronic structure and its dynamics pose still a technological challenge [116]. These limits hinder substantially the investigations of the possible dynamic paths that can lead phase transitions across the phase space, and exacerbate an already complex gap existing between solid hydrogen's experimental characterisation and its modelling.
One of the most successful techniques in discovering hydrogen's meta-stable phases has been the systematic investigation of its potential energy surface using CSP algorithms, which we use here onward as a starting basis for the following GCH constructions. Given the interest in the high pressure region of hydrogen's phase diagram and availability of CSP data, we use this system as our first benchmark.
We analyze 7,964 locally-stable hydrogen structures from an ab initio random structure search

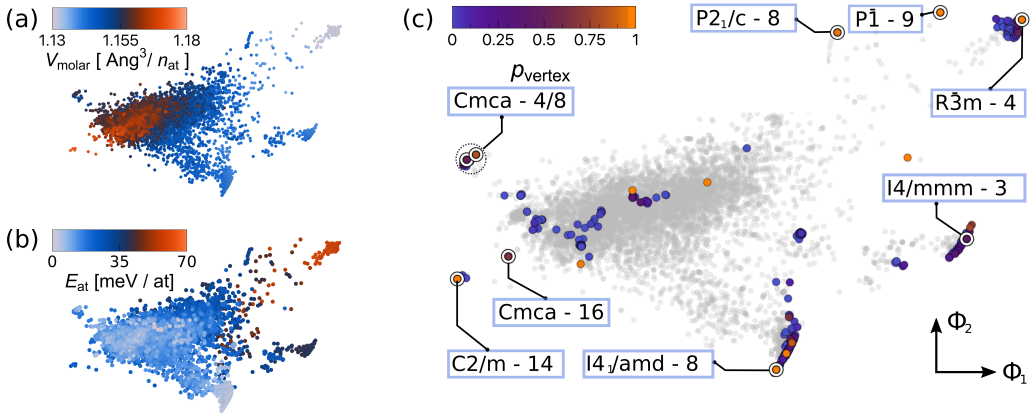


Figure 4.2 – Maps of 7,594 hydrogen structures spanned by the two dominant KPCA features, Φ_1 and Φ_2. Due to their abstract nature (Eqs. (3.16) to (**??**)) the numerical value of Φ_1 and Φ_2 is not shown. Each point corresponds to a structure in the dataset. The maps on the left are colored according to molar volume (top) and to the molar energy (bottom). One can see the clear correlation between the KPCA coordinates, and structural and energetic properties. The larger map highlights structures with non-negligible probability p_{vertex} of being part of the GCH, which is represented as a color scale. Candidates surviving an additional "coarse-graining" step down to the point where all remaining structures have $p_{\text{vertex}} = 1$ are labelled according to space group and number of atoms per unit cell. By comparison with the map colored according to molar energy, one sees that the convex hull identifies clusters of configurations that are low in energy and/or extremal in structure.

(AIRSS) [100, 117] at 500 GPa based on DFT geometry optimizations using the PBE functional [118], where extensive experimental and theoretical literature [119–123] provides a detailed reference of stabilisable structures.

Guided by the indication of the intrinsic dimensionality of the dataset provided by the KPCA eigenvalue spectrum of a SOAP kernel ($r_c = 2$ Å) (see Fig. 4.1) we construct the GCH on the dominant four KPCA features, thus identifying 81 candidate structures. In the process, we successfully recover the high-pressure molecular $I4_1amd$ and atomic $R\bar{3}m$ phases of hydrogen, as well as analogues of the lower-pressure phases II to IV. The result of this search are presented in Fig. 4.2, where one can see how the many of the stable phases populate the boundaries of the KPCA map. Notably, phase II and IV are not stable at the simulated conditions, so being able to find very similar structures among the candidates is a testament to the long-sightedness of AIRSS and the predictive power of the GCH.

To achieve the same feat using a conventional energy-volume CH, structures up to around 8 meV/atom above the CH have to be retained, leaving a disproportionately larger pool of more than 2,000 potentially stabilisable structures. When constructed on just the first four KPCA features the GCH framework identifies 81 stabilisable structures, including analogues of the $Pc2_1$-24, the $C2/c$-24 and $Cmca$-12 [120], and the $Cmca$-4 and Pc-48 [102] candidates for phases II, III, and IV, respectively shown in Fig. 4.3.

The aforementioned figure shows, albeit partially, the robustness of the construction to detection of phases in presence of distortions or analogues.

In order to elucidate stabilisation mechanisms it is necessary to understand which experimentally-realizable thermodynamic constraints the KPCA features couple to. For instance, Fig. 4.4 highlights the strong correlation between the KPCA features and molar volume, indicating that pressure may be exploited to stabilize various identified candidate structures. This is, of course, consistent with the experimental fact that hydrogen phases I to V can consecutively be stabilized by applying increasing pressure. This validation scheme offers an unsupervised strategy to shed light on potentially novel mechanisms of stabilisation, by testing correlation (e.g. R^2 scores) measures with other response properties of the system.

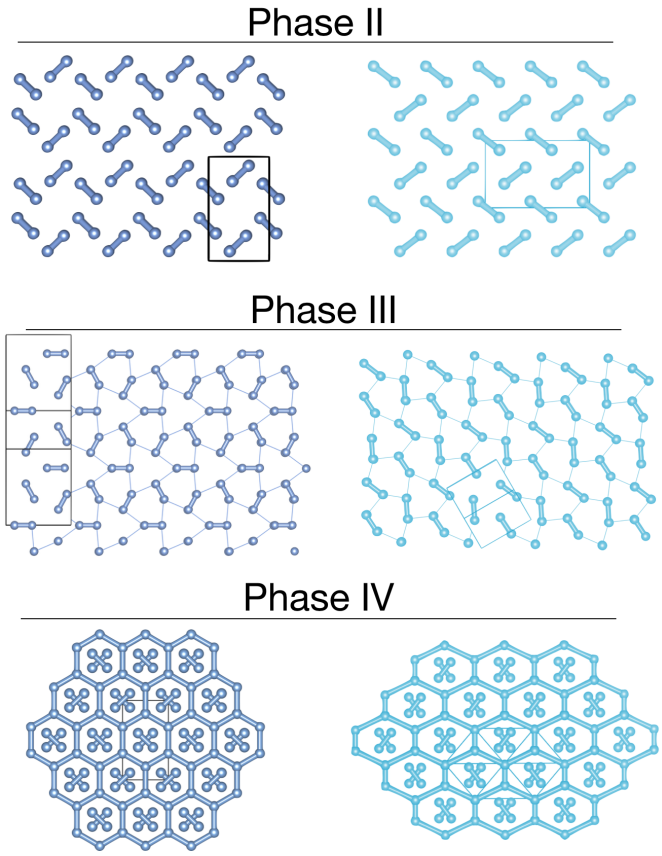


Figure 4.3 – Comparison of literature candidate structures for hydrogen phases II to IV (left, dark blue) and the high-pressure analogues identified as stabilisable by applying the GCH framework to a dataset of hydrogen structures at 500 GPa (right, light blue).

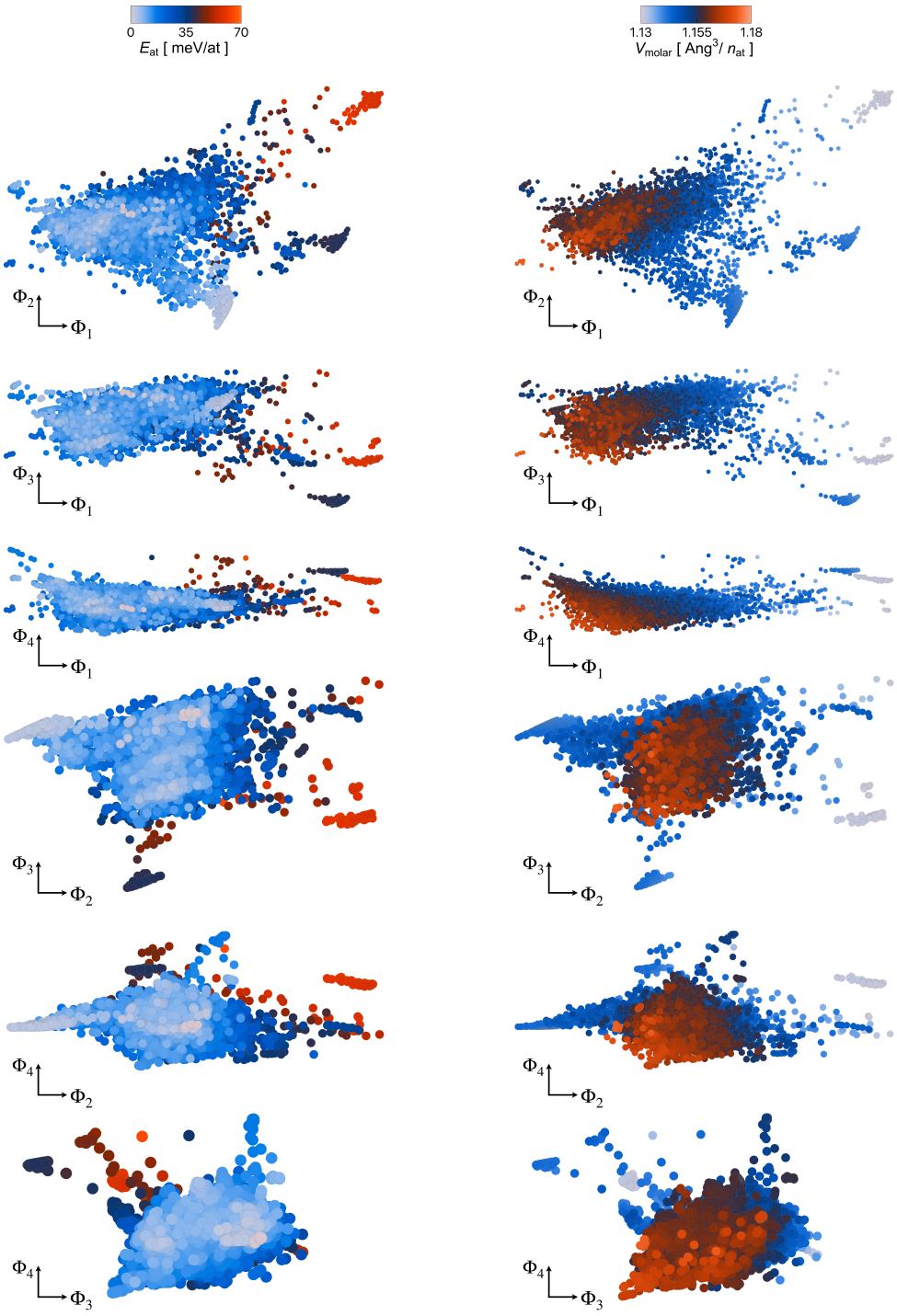


Figure 4.4 – On the left side, correlation between the KPCA features Φ_1 to Φ_4 and energy in meV/atom, on the right side, correlation between the KPCA features Φ_1 to Φ_4 and molar volume in Å^3/atom.

4.3 Oxygen-Hydrogen binary compounds.

The next level of complexity in computational materials discovery involves the modelling of multi-component systems: in the case of 51,376 locally-stable H_xO_{1-x} configurations from an *ab initio* random structure search (AIRSS) [100, 117] at 20 GPa (again based on DFT geometry optimizations using the PBE functional [118]) the GCH framework must resolve the most stable stoichiometries, while at the same time recovering various hydrogen, ice and oxygen phases.
This example is of fundamental importance, as it shows how the construction can be applied to tackle both chemical and structural degrees of freedom at once.
The KPCA eigenvalues based on a SOAP kernel (r_c = 5 Å) decay by more than an order of magnitude after the first feature (see Fig. 4.1). This reflects the dominant role of composition in determining structural diversity and forecloses the identification of the first KPCA feature with composition (see Fig. 4.5 (a)).
Along this principal axis, one identifies the expected stable oxygen, hydrogen, and ice structures, but also crystalline hydrogen peroxide, ice phases with different fractions of intercalated hydrogen molecules and crystalline molecular hydrogen and oxygen phases with guest water molecules.
The latter are unstable in the absence of other stabilizing fields as highlighted by an energy-composition CH construction. Their stability on the GCH arises because the first KPCA feature (while predominantly describing composition) also measures molar volume as an additional stabilizing factor.
When constructed on the first two KPCA features the GCH framework identifies 171 stabilisable structures, differing in both stoichiometry and geometry (see Fig. 4.5 (b)). Among nine hydrogen structures are phase I, the $Pc2_1$-24 candidate for phase II, and the $Cmca$-4 candidate for phase IV [102]. Reassuringly, the three ice phases include the experimentally stable ice VII/VIII and the $Pmc2_1$ high-pressure candidate phase of Hermann *et al.* [124].

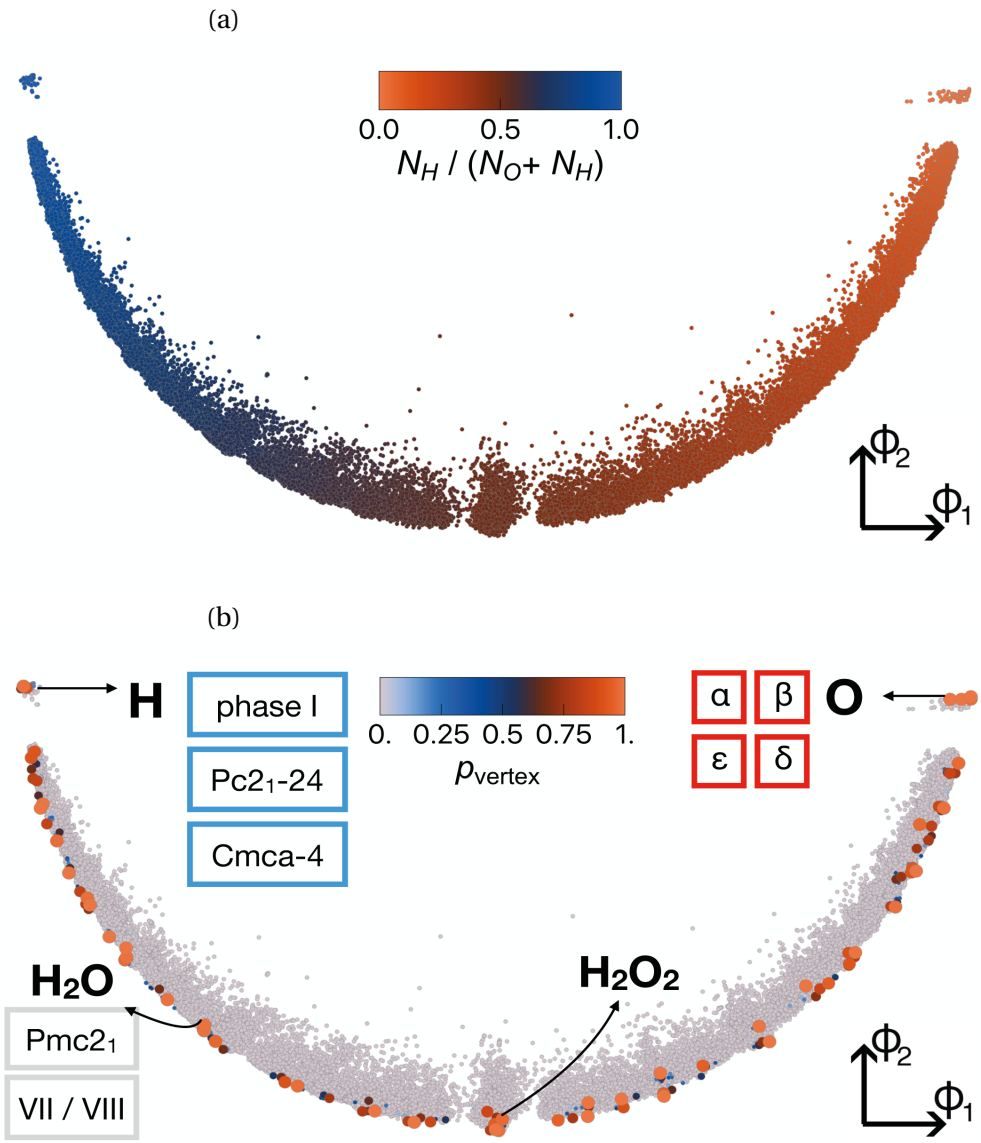


Figure 4.5 – Map of 51,376 H_xO_{1-x} structures spanned by the two dominant KPCA features, Φ_1 and Φ_2. The structures are colored according to (a) composition, and (b) their probability, p_{vertex}, of constituting a vertex of the CH of $E(\Phi_1,\Phi_2)$. The positions of experimentally-confirmed and proposed hydrogen, ice, hydrogen peroxide, and oxygen structures are highlighted. Proposed structures are labelled according to their symmetry group.

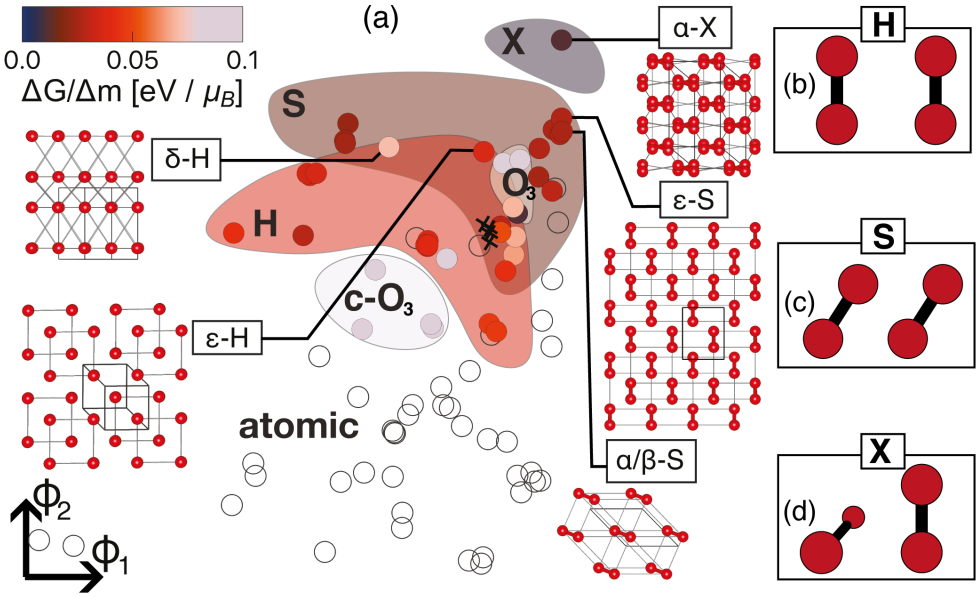


Figure 4.6 – PCA projection of the subset of 84 pure oxygen structures onto Φ_1 and Φ_2 as obtained for the full dataset of 51,376 H_xO_{1-x} structures. Diamagnetic molecular structures (filled circles) are colored according to $\Delta G/\Delta m$. Atomic and ferromagnetic molecular structures are shown as empty circles and crosses, respectively. The shaded regions highlight molecular structures in the H, S, and X configurations ((b) to (d)), and are colored according to the respective mean values of $\Delta G/\Delta m$. This highlights the correlation between $\Phi_{1,2}$, molecular tilts, and energetic response to magnetization $\Delta G/\Delta m$ as a proxy of the potential for stabilisation by magnetic fields.

4.4 Magnetically-stabilized phases of oxygen.

The six oxygen structures deserve a more detailed discussion, as they demonstrate that the GCH is capable of revealing subtle mechanisms of stabilisation, which have barely been touched upon in literature, such as the stabilisation of unconventional molecular oxygen phases by external magnetic fields. Using the nomenclature introduced in Refs. [125–127], the six oxygen structures include the conventional α/β and ε [128] phases, in which the O_2 molecules align in the so called "H"-state (Fig. 4.6 (b)). The GCH further detects α/β and δ phases with uniformly-tilted O_2 molecules ("S" state, Fig. 4.6 (c)) and an α phase, in which the molecules display an alternating tilt pattern ("X" state, Fig. 4.6 (d)). Experimental evidence suggests that these may be stabilized by strong magnetic fields [127, 129], which we further substantiate using spin-polarized DFT calculations using Quantum Espresso [130] (see Fig. 4.6, Fig. 4.7 and Fig. 4.8). This demonstrates (i) that structural features do indeed correlate with subtle responses to manipulations of the electronic structure of a configuration and (ii) how one can verify the coupling between abstract structural coordinates and experimentally-realizable thermodynamic constraints.

The KPCA eigenvalue spectrum obtained for the full set of 51,376 locally-stable O_xH_{1-x} struc-

tures of different stoichiometry reflects that composition is the central feature distinguishing the 51,376 O_xH_{1-x} structures.
This implies that the KPCA features obtained for the full dataset do not optimally distinguish between different oxygen phases and begs the question whether the selection of stabilisable oxygen phases presented in the global case survives a closer, separate inspection of the 84 oxygen phases.
We therefore reconstruct the KPCA and GCH on the "reduced dataset", which only contains the 84 pure oxygen phases. Fig. 4.7 (a) highlights the resultant pronounced correlation between

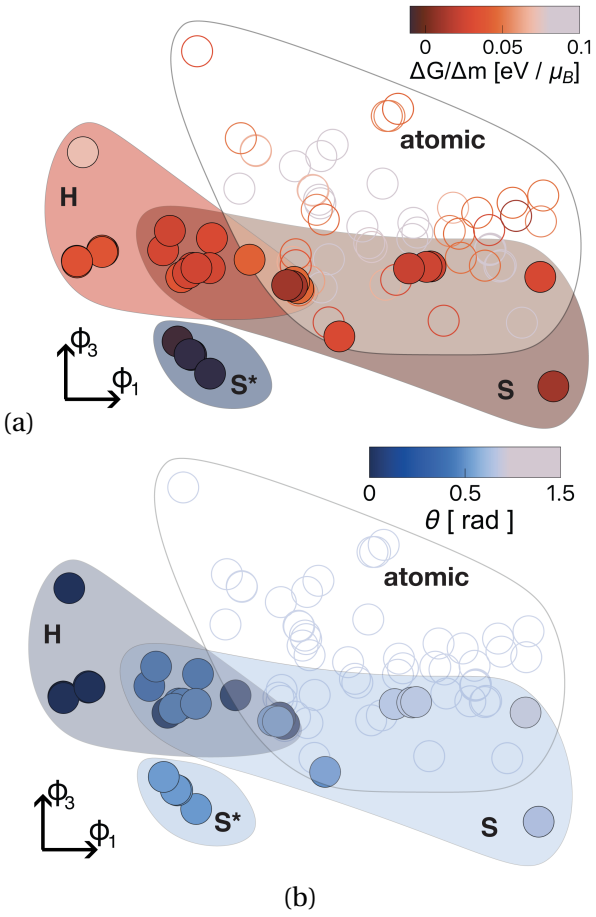


Figure 4.7 – Dedicated KPCA projections of the 84 pure oxygen structures onto the Φ_1 and Φ_3 features. Molecular and atomic structures are shown as disks and circles, respectively. In (a) they are colored according to their energetic response to magnetization $\Delta G/\Delta m$ (as a proxy of their potential for stabilisation using external magnetic fields), while in (b) they are colored according to the tilt of the molecular axes with respect to the molecular planes. The shaded regions highlight molecular structures in the conventional H, and the S and X configurations, and are colored according to the mean value of $\Delta G/\Delta m$ and the mean tilt across structures of a given tilt configuration, respectively.

the third KPCA feature and the response to magnetization (as a proxy for the potential for stabilisation by magnetic fields). Reassuringly, the identification of magnetic fields as a viable

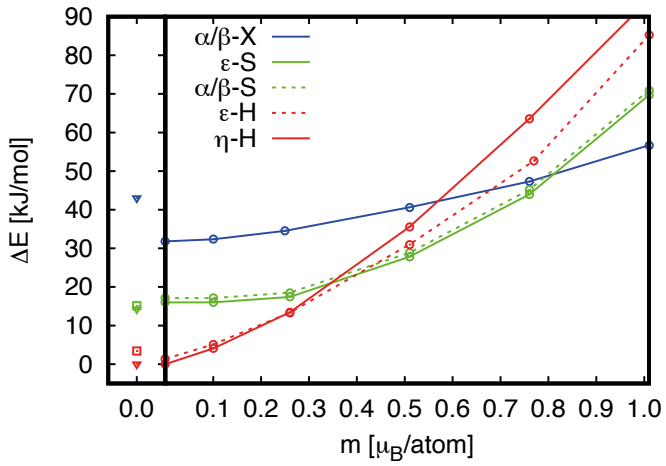


Figure 4.8 – The left panel shows the relative lattice energies. ΔE of the pure oxygen configurations constituting vertices of the GCH for the O_xH_{1-x} dataset. The right panel shows the dependence of ΔE on magnetization m as a proxy for the potential for stabilisation by external magnetic fields. The differences between ΔE for $m = 0\,\mu_B$ using the Quantum Espresso implementation of PBE-DFT and those underlying the GCH construction highlights the size of typical uncertainties in input energies.

stabilisation mechanism and the classification of (a) atomic and molecular phases and (b) "H" and "S" type molecular phases is consistent with the structure of Fig. 4.6.

We also observe that this overall structure is retained when the similarity kernel is replaced with a SOAP kernel with a substantially reduced cut-off radius of 3.25 Å, despite the differences in the resulting structural similarity measure.

This sanity check is motivated by the observation that the original similarity kernel represents a compromise required to simultaneously rationalize the different stoichiometries in the O_xH_{1-x} dataset.

Even though the (single) "X" configuration is is not identified as structurally extremal, it is among the 33 candidates identified by a three-dimensional GCH construction. One might speculate that the alternating tilt pattern characterizing the "X" configuration is not a good discriminator for the remaining 83 oxygen structures and therefore not among the dominant KPCA features.

Its succinct identification in the full O_xH_{1-x} set may indicate that angular correlations in structural environments play a more important role in distinguishing the 51,376 O_xH_{1-x} structures than the 84 oxygen structures.

To probe the stabilizing potential of external magnetic fields and the importance of molecular oxygen in the "X" state we calculate the lattice energies of all oxygen structures as a function of magnetization using first-principles, spin-polarized DFT calculations using Quantum Espresso [130] with the PBE functional [118].

Fig. 4.8 shows the lattice energy of the most promising oxygen candidates as a function of magnetization. It demonstrates that magnetic fields have the potential to stabilize molecular

phases in the "S" configuration (with a uniform tilt of the O_2 molecular axes) and ultimately those in the "X" configuration (with an alternating tilt pattern) with respect to their counterparts in the conventional "H" configuration.
An important subtlety concerns the classification of "H", "S", and "X" type molecular phases. The nomenclature introduced in Ref. [127] only uniquely defines the relative arrangement of O_2 molecules in an O_2-O_2 dimer (or an $(O_2)_3$ triplet). In order to generalize it to crystalline systems it is necessary to identify molecular layers. In analogy with the Heisenberg Hamiltonian originally used to rationalize the emergence of ferro- and antiferromagnetism in molecular oxygen system [125, 126], these are identified as containing the nearest neighbour molecules. This is equivalent to maximizing the inter-layer spacing, and, heuristically, minimizing the inter-layer coupling. A tilt-angle can then be uniquely defined as the angle of the molecular axes with respect to the plane-normal associated with the molecular layers (see Fig. 4.7 (b)).
Despite the slowly decaying KPCA eigenvalue spectrum in Fig. 4.9, which forecloses a compar-

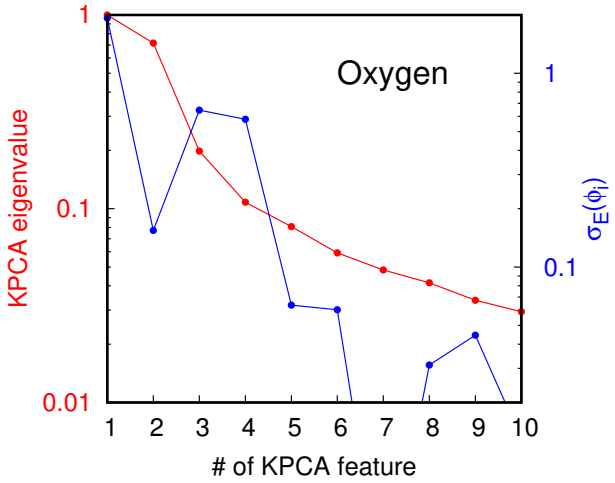


Figure 4.9 – PCA eigenvalues, ε_i, (red) and estimates of the contributions to the energetic variance of the dataset due to each feature, $\sigma_E(\Phi_i)$, (blue) obtained from the SOAP similarity kernel with $r_c = 5$ Åfor 84 locally-stable oxygen structures at 20 GPa.

atively large number of stabilisable structures, a two-dimensional GCH already encompasses the pool of stabilisable structures identified before.
The results for GCH constructions using up to three KPCA features are collected in Table 4.1. For completeness, Table 4.1 also summarizes the results of the GCH constructed on the SOAP kernel using $r_c = 3.25$ Å.

	$r_c = 5$ Å			$r_c = 3.25$ Å		
	1	2	3	1	2	3
α/β-H	−	+	+	+	+	+
α/β-S	+	+	+	+	+	+
δ-H	−	+	+	−	−	+
δ-S	−	+	+	+	+	+
ϵ-H	+	+	+	−	+	+
ϵ-S	−	−	+	−	−	−
ζ	+	+	+	+	+	+
ozone	−	−	+	−	−	+
chain	−	+	+	−	−	+
total	5	14	33	5	16	33

Table 4.1 – Recovery of known and proposed phases of high-pressure oxygen on the basis of GCH constructions based on a SOAP similarity kernel for the oxygen-only dataset using 5 Å and 3.2 Å cut-off radii, respectively. Note that the three-dimensional GCH based on the 3.25 Å kernel also identifies the "X" state of α/β oxygen as stabilisable.

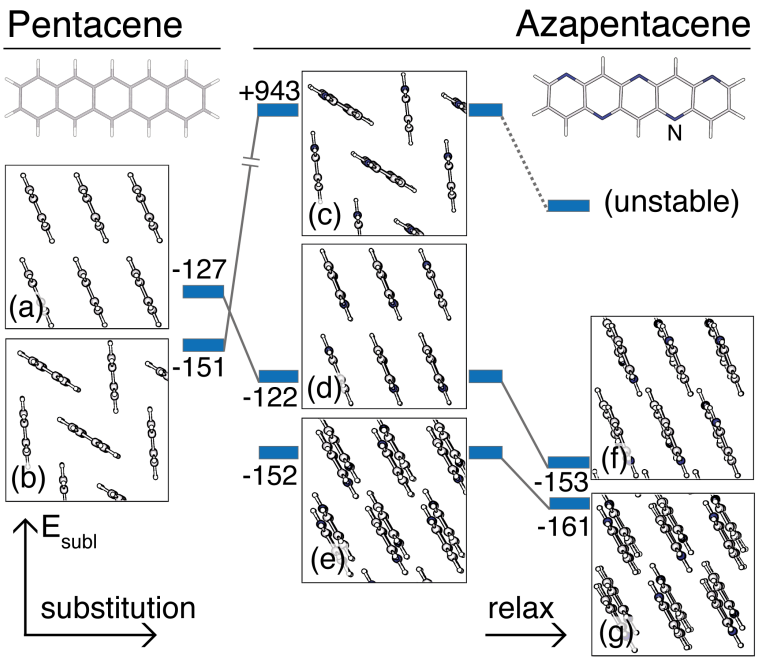


Figure 4.10 – Sublimation energies, E_{subl}, of different pentacene configurations in kJ/mol before (left) and after 5A nitrogen substitution (center), and after subsequent geometry optimization (right). (a) is among the most unstable pentacene configurations in the dataset. (e) is the most stable 5A substituted azapentacene configuration among 594 configurations from an independent structure search [131]. The E_{subl} computed for the Campbell bulk phase (b) of 151.019 kJ/mol agrees with the experimental values of 154.5 [132] and 156.9 ± 13.6 kJ/mol [133] to within the errors.

4.5 Nitrogen substitution in pentacene.

As a final example, we analyze a database of 564 locally-stable arrangements of pentacene molecules. This application beyond high-pressure physics demonstrates how the GCH can suggest suitable starting points for structure searches studies involving chemical substitution. A key problem in performing molecular scale crystal structure searches lies in the prohibitive computational cost associated to a thorough sampling of the disruptive changes in properties that chemical substitutions can offer.

In an ideal workflow, one would aim to perform a single reference potential energy surface survey (i.e. for an unsubstituted molecular solid), to obtain an understanding of the quantitative effect a single substitution would have on the system.

In reality this is often impossible as the potential energy surface of the substituted crystals shares little in common with their underlying reference's one, requiring an extensive layer of additional sampling.

In this example we aim to show how much can be inferred by structural patterns emerging from a CSP exercise on an "unsubstituted" reference system (pentacene crystals) and leverage

it to discover configurations that can become energetically favourable upon substitution of chemical species.
The configurations were obtained by a systematic structure search [131], based on rigid, DFT-optimized molecular units interacting via the W99 force-field [134]. In Ref. [131], this structure search is accompanied by independent searches for 5A (see Fig. 4.10) and 5B nitrogen-substituted molecules: this is a fundamental requirement since the stability of a given molecule is rarely a good predictor of the behavior of its substituted counterparts [135].
We first perform a KPCA of the pentacene dataset using the same SOAP kernel ($r_c = 5$ Å and $\sigma = 0.3$ Å) which has previously proven suitable for energy regressions [136].
Alongside conventional, energetically favourable herringbone configurations, such as the Campbell bulk phase [137], the GCH constructed on the two dominant KPCA features identifies five energetically unfavourable configurations with planar, colinear arrangements of molecules as stabilisable.
This finding leads us to consider the emerging structural patterns as potentially stabilisable by a chemical substitution, since comparable with energetically favourable stacking observed in the azapentacene independent structure search.
To put this intuition under test, we apply a nitrogen substitution that aims to bridge the pentacene convex hull structures with the azapentacene's set (generated according to a 5A substitution scheme), observing their resultant stability.
Whereas nitrogen substitution of the global minimum pentacene configuration leads to a high-energy, unstable structure, several of the GCH vertices that are much higher in energy, which would therefore be discarded in a conventional analysis, retain their geometry upon nitrogen substitution and relaxation (see Fig. 4.10). Moreover, they exhibit competitive energies compared to the most stable 5A substituted configuration.
The GCH framework has thus effectively identified pentacene configurations with potential for stabilisation by nitrogen substitution.

4.6 Sensitivity to errors in energetics.

The probabilistic sampling of the GCH does not only provide a robust strategy to eliminate redundant structures and for uncertainty quantification. It also significantly reduces the sensitivity to errors in input energies compared to conventional deterministic CH constructions.

	TP	FP	FN	$\tilde{d}$	RMSE
E-ρ CH	3	0	2	0.0139	0.22
d-GCH (1D)	5	1	2	0.0227	0.11
GCH (1D)	8	2	11	0.0168	0.10
GCH (1D) cg	5	1	2	0.0046	
d-GCH (3D)	51	9	13	0.0087	0.07
GCH (3D)	113	12	35	0.0066	0.07
GCH (3D) cg	50	8	12	0.0009	

Table 4.2 – Sensitivity analysis of the (conventional) energy-density (E-ρ CH) hull, and deterministic (d-GCH), and probabilistic hulls (GCH) constructed on the first (1D) and first three (3D) KPCA features (before and after coarse-graining (cg)). Different metrics of the similarity of different CH constructions are evaluated on the basis of W99 and DFT sublimation energies for the 564 pentacene configurations from Ref. [131]: (i) the numbers of true (TP) and false positive (FP), and false negative (FN) identifications of stabilisable pentacene configurations on the basis of W99 energies, (ii) the distance $\tilde{d}$ between the W99 and DFT based hulls as defined in Eq. (4.2), and (iii) the RMSE in kJ/mol in the W99 convex hull energies $\{E_k^{\mathrm{W99}}\}$ compared to "reference" DFT convex hull energies $\{E_k^{\mathrm{DFT}}\}$ (for the full dataset).

To assess how sensitive different CH constructions are with respect to the details of the input energies, we calculate DFT sublimation energies using Quantum Espresso [130] with the PBE functional and a Grimme-D2 dispersion correction [1] for all 564 pentacene configurations for comparison with those obtained from the W99 force-field.
The DFT and W99 sublimation energies exhibit substantial differences (resulting in a root-mean-square error (RMSE) with respect to each other of the 0.15 kcal/mol after subtracting the respective averages), including a different global energy minimum structure.
As shown in Table 4.2,one can reduce effectively the discrepancies by computing energies relative to the convex hull:

$$E_k^{\mathrm{DFT/W99}} = G_k^{\mathrm{DFT/W99}} - \sum_l w_{kl}^{\mathrm{DFT/W99}} G_l^{\mathrm{DFT/W99}} \tag{4.1}$$

This is a consequence of the fact that energy errors are correlated, which we also exploit in our probabilistic hull construction.
The set of structures that are tagged as "synthesizable" is perhaps even more important than

[1] We use a plane-wave energy cut-off of 100 Rydberg, a Monkhorst-Pack **k**-point grid [**monkhorst_1976**] spacing of less than $2\pi \times 0.07$ Å^{-1}, and the ultrasoft C.pbe-n-kjpaw_psl.0.1.UPF, H.pbe-kjpaw_psl.0.1.UPF, and N.pbe-n-kjpaw_psl.0.1.UPF pseudopotentials from *http://www.quantum-espresso.org*

the estimate of the instability of the other candidates. Since different, structurally very similar configurations, for example only differing in proton or stacking (dis-)order, can be equivalently valid representatives of the same (stabilisable) phase, one cannot simply compare the indices of the structures identified as vertices. To determine whether two hulls $\mathscr{H}_{\mathrm{DFT}}$ and $\mathscr{H}_{\mathrm{W99}}$ constructed on the basis of DFT and W99 energies, $\{G_k^{\mathrm{DFT}}\}$ and $\{G_k^{\mathrm{W99}}\}$, respectively, contain structurally similar vertices, we define a "distance" between hulls as the mean minimum Euclidean distance between their respective vertices

$$\begin{aligned} \tilde{d} &= \frac{1}{2}\left(d_{\mathrm{DFT}}^{\mathrm{W99}} + d_{\mathrm{W99}}^{\mathrm{DFT}}\right) \\ d_{\mathrm{DFT}}^{\mathrm{W99}} &= \sqrt{\frac{1}{N_{\mathrm{DFT}}} \sum_{X_i \in \mathscr{H}_{\mathrm{DFT}}} \min_{X_j \in \mathscr{H}_{\mathrm{W99}}} \left|\mathbf{\Phi}(X_i) - \mathbf{\Phi}(X_j)\right|^2}. \end{aligned} \tag{4.2}$$

The results of this analysis, shown in Table 4.2 confirm that the GCH construction reduces the sensitivity of both the vertex selection and the measure of stability compared to a conventional construction. Increasing the dimensionality of the fingerprint space on which the hull is constructed, sampling probabilistically different realizations of the hull, and eliminating redundant structures in the database, all lead to a more robust determination of stabilisable structures that should be considered for further theoretical or experimental investigation.

4.7 Sensitivity to the construction of the similarity kernel

What we have shown so far relies on the choice of a short ranged descriptor which cuts off any interaction involving distances larger than r_c. The choice of such cutoff can often be non trivial and its influence on the resulting GCH outcome must be benchmarked.
In this section we show how one can substantially obtain a consistent set of results from GCH constructions built on different starting descriptions provided they offer a comparable description. The critical parameter in the SOAP framework is the cutoff radius r_c, which determines the extension of the environments under study, and thus the spatial correlations relevant to the application.
To define substantially different descriptions of our systems we assume that it is sufficient to span a large enough number of cutoff radii, and finally calculate each of their corresponding sets of convex hulls to assess their effect on the construction..
For the purpose of this example we perform a benchmark by building multiple generalised convex hull analysis on the same solid hydrogen dataset, changing solely the cutoff radius of the SOAP descriptor employed to generate the similarity kernels.
Fig. 4.11 shows the average radial distribution function (RDF) of the 7,594 hydrogen structures, which suggests that the second, third, and fourth coordination shells typically sit at radii of around 1.6, 2.7, and 3.7 Å.
This motivates the SOAP cut-off radius $r_c = 2$ Å used in the construction, since choosing r_c in between coordination shells reduces the sensitivity of the kernel to marginal differences

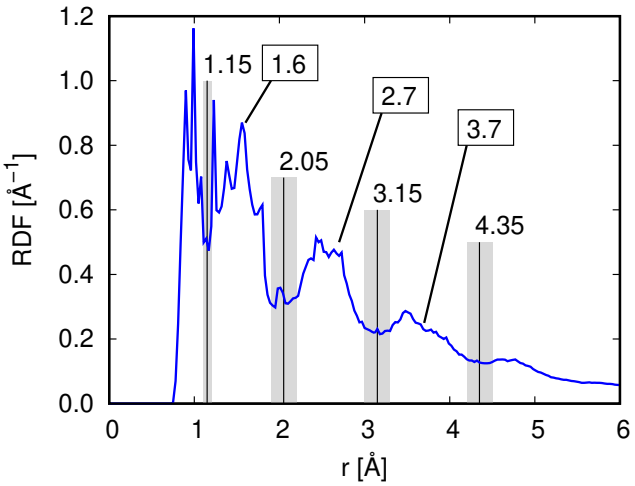


Figure 4.11 – Average RDF of 7,594 hydrogen structures. The mean radii between coordination shells are indicated by solid black lines.

in coordination radii between very similar structures. The lack of a unique choice of r_c is mitigated by the observation that the GCH construction is relatively insensitive to the choice of r_c.

Fig. 4.12 (a) shows that the fraction of matching candidates relative to the number of unique candidates in hulls based on slightly different r_c (defined as $\frac{2|\mathscr{H}_{r_{c1}} \cap \mathscr{H}_{r_{c2}}|}{|\mathscr{H}_{r_{c1}}|+|\mathscr{H}_{r_{c2}}|}$) is a substantial 81% (95%) for differences in r_c of 0.15 Å(0.05 Å).
Fig. 4.12 (b) shows that the hulls obtained on the basis of kernels based on r_c between (1) the first and second and (2) the second and third coordination shells also match to about 80%.

In analogy to the sensitivity analysis in the above section, Table 4.3 further provides the typical best match distance between candidates obtained using different cut-off radii, $\tilde{d}(\mathscr{H}_{r_{c1}}, \mathscr{H}_{r_{c2}})$. In practice, $\tilde{d}$ is computed based on each of the two sets of KPCA features, $\mathbf{\Phi}^{r_{c1}}$ and $\mathbf{\Phi}^{r_{c2}}$, and averaged. For reference, we also report (i) the root mean square distance between structures:

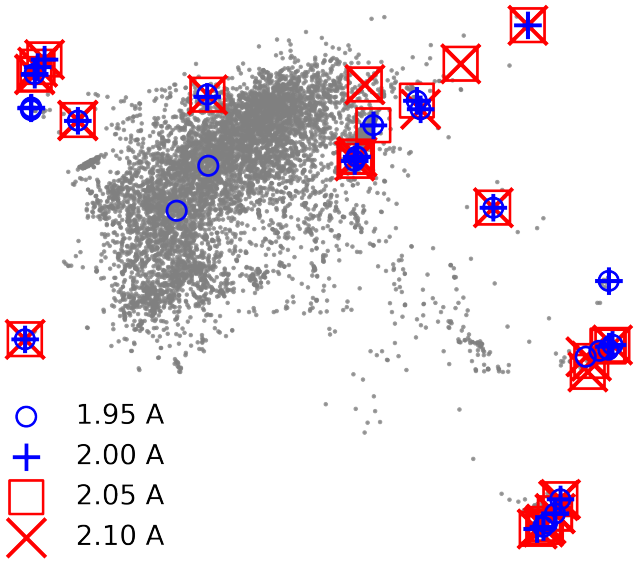
$$\begin{aligned} \tilde{\mathscr{D}}(r_{c1}, r_{c2}) &= \frac{(\mathscr{D}(r_{c1}) + \mathscr{D}(r_{c2}))}{2} \\ \mathscr{D}(r_c) &= \sqrt{\frac{1}{N^2} \sum_{i,j=1}^{N} \left| \mathbf{\Phi}^{r_c}(X_i) - \mathbf{\Phi}^{r_c}(X_j) \right|^2} \end{aligned} \tag{4.3}$$

and (ii) the mean distance between the hulls $\mathscr{H}_{r_c}$ and 1,000 randomly selected sets of structures of same size as $\mathscr{H}_{r_c}$, $\tilde{\mathscr{D}}(\mathscr{H}_{r_c}) \sim 0.09$. Table 4.3 shows that $\tilde{d}$ is generally much smaller than both $\tilde{D}$ and $\tilde{\mathscr{D}}$, and furthermore comparable to the values obtained in the previous examples for hulls constructed on input energies from different levels of theory. Further to Fig. 4.3 this suggests that the uncertainty in the hulls due to the kernel construction is comparable to that arising from the errors in the underlying structure data and that the hulls are relatively insensitive to the details of the kernel construction.

r_{c1}	r_{c2}	$\tilde{d}$	$\tilde{\mathscr{D}}(r_{c1}, r_{c2})$
1.95	2.00	0.009	0.130
1.95	2.05	0.018	0.128
1.95	2.10	0.019	0.125
2.00	2.05	0.010	0.125
2.00	2.10	0.012	0.123
2.05	2.10	0.013	0.120
2.05	3.10	0.025	0.111
2.05	4.25	0.024	0.095
3.10	4.25	0.015	0.083

Table 4.3 – The typical distance between best match structures from two hulls $\tilde{d}(\mathscr{H}_{r_{c1}}, \mathscr{H}_{r_{c2}})$ is much smaller than the average pairwise distance between the structures in the dataset (Eq. (4.3)), suggesting that, if constructed on a physically meaningful similarity kernel, the GCH framework identifies similar sets of stabilisable structures, irrespective of the details of the kernel construction.

(a) GCH vertices obtained on the basis of kernels with r_c = 1.95Å(blue circles), 2.0 Å(blue pluses), 2.05 Å(red squares), and 2.1 Å(red crosses).



(b) GCH vertices obtained on the basis of kernels with r_c = 2.05 Å(red circles), 3.1 Å(blue pluses), and 4.25 Å(red crosses).

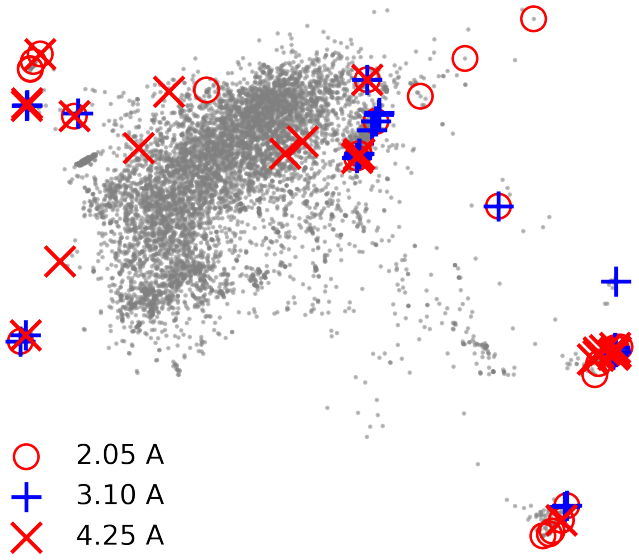


Figure 4.12 – Map of the 7,594 locally-stable hydrogen structures spanned by the features $\Phi_1^{r_c}$ and $\Phi_2^{r_c}$ obtained from a kernel with r_c = 2.0 Å. GCH vertices obtained on the basis of kernels with r_c between 1.95 Åand 4.25 Åare highlighted, showing that the GCH selection is relatively robust to substantial changes to the similarity kernel.

4.8 Discussion

These examples clearly evidence the wide spectrum of thermodynamic constraints which can be rationalized using the GCH framework and serve to showcase the remarkable versatility and transferability of the GCH framework, which reflect its data-driven nature and conceptual simplicity.
The construction is only weakly dependent on the details of the kernel, and its probabilistic nature renders it robust to errors in the determination of the (free)-energies of different phases, which is very important given the harsh compromises one has to make between the accuracy and thoroughness of high-throughput structure searches.
Moreover, it is capable of eliminating redundant configurations in a physically meaningful way and of providing estimates of stability regimes in terms of experimentally-realizable thermodynamic constraints.
The GCH framework provides a robust, data-driven, method- and error-insensitive evolution of the convex hull construction, one of the most essential tools to predict and rationalize the stability of materials, and to identify experimentally stabilisable structures among large numbers of potential, locally stable configurations.
If paired with a regression scheme based on the same descriptor, one can aim to predict an estimate of a geometry's energy and then observe it's potential position with respect to the GCH built with the training set of structures.
This scheme could allow for a rapid screening technique when biasing the configuration space over which the CSP is performed, and potentially skew the search towards regions that either unexplored or at lower energied.

5 Navigating the phase diagram of ice

5.1 Introduction

Ice is a complex system of interest across much of science, ranging from astrophysics to biology. On the Earth's surface and in its atmosphere it plays a central role in determining climate and in countless natural processes and technological applications. Ice is also a key constituent of the Earth's crust and mantle. Its phase diagram and properties have been investigated across a wide range of temperatures and pressures by experimentalists and theoreticians alike.

A total of 18 crystalline ice phases have been formed under various conditions [138], seven of which are metastable [139]. In addition, a number of hypothetical ice phases [140–151] have been predicted and characterised using computer simulations. All of these phases are molecular crystals that fulfil the "Bernal-Fowler ice rules" [152] and form four-connected networks. In most ice phases the distinct ways of dressing the oxygen sublattice with hydrogen atoms within the ice rules (the so-called "proton-orderings") are quasi-energetically degenerate [153]. The corresponding proton-disorder typically "freezes out" at low temperatures. Disregarding proton-order ice only forms eight topologically-distinct oxygen networks under the pressure and temperature conditions that have been explored experimentally thus far. Theoretical studies have also suggested structures of water ice under ultrahigh pressures of up to many terapascals and its eventual decomposition [154, 155].

The phase-diagram of ice has recently received renewed interest, firstly, because the theoretical discovery of the s-III clathrate hydrate [150] and the experimental description of ice XVII [156] and of two-dimensional forms of ice [151, 157, 158] have demonstrated that our understanding of ice is far from complete. Secondly, it has become apparent that the nucleation and melting of ice are complex processes in which meta-stable ice phases play a role [159, 160]. For example, stable hexagonal ice (Ih), meta-stable cubic ice (Ic), and stacking-disordered ice (Isd) have been shown to play important roles in ice nucleation in molecular dynamics and lattice switching Monte Carlo simulations [159, 160]. In classical nucleation theory an interfacial free energy advantage of a few percent will lead to preferential nucleation of metastable phases with free energies up to around 10 meV/molecule above the stable phase [161]. Despite

valiant efforts using structure searching methods such as *ab initio* random structure searching (AIRSS) [162], to our knowledge no comprehensive study of (meta-)stable ice phases and their formation has been published to date. The problem is twofold: firstly, the enormous configuration space must be explored in a reasonably comprehensive manner. Secondly, in order to render the structure search relevant to experiment, the large number of theoretical (meta-)stable structures generated in the process must be reduced to those which can be formed experimentally. This refinement must be a-priori and quantitative. Finally, different stabilising factors – such as the absorption of guest molecules [163] – can be investigated further, and methods such as forward flux sampling [164–167] and enhanced sampling metadynamics [168–170] may be used to identify possible synthetic pathways.

This work aims for a comprehensive study of crystalline ice phases, focussing on the exploration of configuration and the reduction of the resulting intractably large amount of structure data to a small number of structures, which are likely to be accessible experimentally. In the results (section 5.2) we exploit the isomorphism between ice and silica networks [149] to explore the relevant parts of the configuration space of ice using databases of theoretically-enumerated, four-connected networks. In section 5.4 we rationalise the resultant structural data on the basis of purely energetic considerations and thereby identify structures which can be stabilised under pressure. By design this approach cannot identify structures stabilised by thermodynamic and kinetic constraints other than pressure, such as temperature, electric fields, concentrations of guest molecules, etc. In section 5.5 we overcome this limitation by applying the generalised convex hull construction. Moreover, we use the sketch-map algorithm [51] to construct a navigable map of the configuration space of ice. This primarily serves as an aide in developing an intuitive understanding of structural relationships. However, it also shows potential for helping to identify formation pathways for new candidate ice phases.

Author Contributions

This chapter reproduces, with minor adjustments, the content of Ref.[32]. The contribution of the author of this thesis to this work is the application of the generalised convex hull construction to the set of ice configurations detailed in Section 5.5, calculation of the convex hulls using different density functionals in 5.6 and the clustering scheme proposed in section 5.7.

5.2 Exploring configuration space

The strong isomorphism between ice and silica networks has previously been explored in Ref. [149] and arises because both silica and water preferentially form four-connected networks composed of corner-sharing tetrahedral units. For instance, the low-pressure ice structures, ice Ih and ice Ic are isomorphs of tridymite (lonsdaleite net) and cristobalite (diamond net), respectively, and clathrate hydrate cage structures like the sI and sII clathrate hydrates [171–173], which are stabilised by absorption of small inert guest molecules [174], have direct

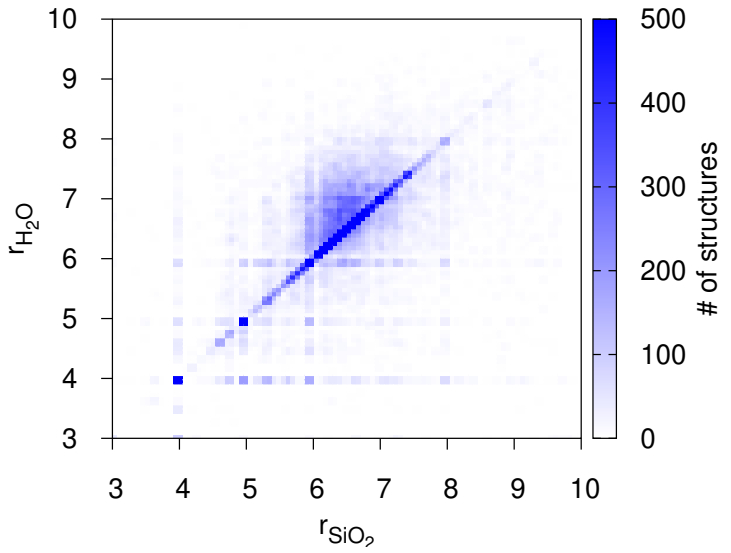


Figure 5.1 – Correlation between the average ring sizes, r, of SiO_2 and H_2O polymorphs. More than 1/3 of the ice polymorphs retain the ring statistics of their counterpart SiO_2 network.

analogues in the world of alumino–silicates (or "zeolites") [175–179]. The basic building blocks of silica and water ice are so similar that it is even possible to form silica/water hetero-networks in which silicate oligomers form part of the hydrate lattice [180, 181].

There is a vast literature on four-connected structures, including an atlas describing the underlying networks of porous crystalline zeolites [182] and a number of very large databases of theoretically-enumerated networks, such as the databases of Treacy [183] and Deem [184]. Graph network enumeration has previously been applied to crystal structure prediction [185] and, in particular, to sp^2- and sp^3-carbon [186, 187]. The above databases have proven to be a valuable resource in searching for sp^3 allotropes of carbon [188] and constitute a comprehensive source of four-connected networks from which topologically-distinct phases of ice can be constructed and geometry optimised to the respective associated local minimum energy structures using conjugate gradient methods. Recently, the search for computationally stable ultralow-density ices [189] on the basis of the atlas of zeolites [182] has hinted at the potential of this approach, despite its more limited scope and despite only considering stabilisation under (effective negative) pressure. The search for (meta-)stable ice phases is facilitated by the strong correspondence between zeolite and ice structures. Fig. 5.1 shows the strong correlation between the average ring sizes of SiO_2 structures and their counterpart H_2O polymorphs after geometry optimisation, indicating that structurally distinct SiO_2 networks generally translate into structurally distinct H_2O networks.

The large size of the databases of hypothetical zeolites necessitates some preselection of structures. Tribello *et al.* [149] show that the energies and densities of low-density SiO_2 networks and their counterpart H_2O networks are correlated, but this correlation does not carry across to structures of densities comparable to and higher than that of ice Ih.

Consequently, neither SiO_2 lattice energies nor densities can be used for preselection. Since all known ice phases (with the exception of ice V/XIII) have unit cells containing no more than 16 molecules, applying a cut-off to the unit cell size provides a reasonable method for preselection, which can be improved systematically by including structures with larger unit cells. In practice, we preselect only networks with unit cell volumes of less than 800Å^3 and without 3-rings, which would normally induce excessive strain in an ice structure. Out of 331,172 (Deem) and 5,389,408 (Treacy) zeolites, this leaves 74,731 structures. This selection contains duplicates since the databases are not mutually exclusive. Low density structures with low SiO_2 lattice energies are added back in by including the experimentally synthesised zeolite networks from the IZA database [190].

5.3 Treatment of the database

Geometry optimisation of the resulting 74,963 structures using first-principles quantum-mechanical methods is viable. However, at this stage only rough lattice energies are required to identify the low energy sectors of configuration space. The definition of low energies is provided by the differences between the lattice energies of different proton-orderings and between the quantum vibrational corrections of different structures, which are both of order $\sim 10\,\mathrm{meV/H_2O}$ [191]. To obtain a coarse energetic filter the authors have chosen to calculate the lattice energies of the unfiltered pool of candidates using a ReaxFF force field[192, 193] rather than a fixed geometry approach such as TIP-4p.
The choice is consistent with the initial goal of rendering the search unbiased with respect to hydrogen bonding network geometries.
In fact, structures such as ice X and higher pressure configurations could potential be disregarded given their distinct intermolecular distances.
After removing high energy configurations the geometries of the remaining structures are refined using PBE-DFT. Removing duplicates leaves 15,882 distinct structures.
The duplicates were identified by applying the “crysim” tool from the AIRSS method [162] to the oxygen sublattices.

5.4 Phase stability and characterisation of structures

The large pool of candidate structures highlights the central challenge of computational structure searches: the number of theoretical (meta-)stable configurations which can be constructed increases exponentially with system size, but only those that can be observed experimentally are of interest. Their selection must take into consideration the uncertainty in the computational framework, the possibility of kinetic and/or surface effects promoting the formation of metastable phases, and the (de-)stabilisation of phases by different thermodynamic boundary conditions, such as pressure.

To identify the polymorphs which are most likely to form at different pressures we first consider

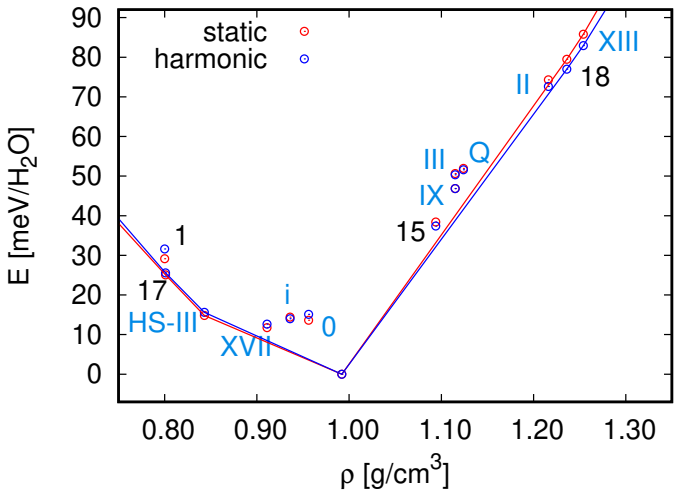


Figure 5.2 – Energy-density convex hull of PBE-DFT static lattice energies (red) and free energies including harmonic vibrations (blue) relative to ice Ih for known ice phases (blue labels) and energetically competitive phases (black labels). The labels of the novel energetically competitive phases correspond to the numbering scheme in Fig. 5.3. The energy-density convex hulls at the static lattice and harmonic vibrational levels are indicated by red and blue solid lines, respectively.

a well-established approach based on a convex-hull construction. The convex hull of energy (as a proxy for free energy) as a function of density, $E_{\mathrm{ch}}(\rho)$, is formed by structures which are stable against decomposition into two or more structures with lower average energy at the same average density, and the so called "tie-lines" that connect them. In the absence of kinetic effects, the only phases that can be observed by manipulating the density of the system (for example through pressure) are exactly those that constitute the vertices of the energy-density convex hull (see Fig. 5.2). In analogy with the Bell-Evans-Polanyi principle, which states that highly exothermic chemical reactions have low activation energies, the stability of a given metastable structure can be assessed by the free energy of decomposition into stable structures. We refer to this as the "dressed energy". Plainly put, the proximity of a metastable structure to the convex hull is a measure of its stability. The "dressed energy" is calculated by subtracting the convex-hull energy at the corresponding density ρ from the lattice energy E (as a proxy for free energy), $E_{\mathrm{dr}} = E - E_{\mathrm{ch}}(\rho)$. Ultimately only structures with E_{dr} less than 10 meV are retained, for which kinetic, entropic and/or surface effects may plausibly lead to preferential formation during nucleation.

Setting aside all prior knowledge of ice, this procedure identifies the theoretical i, 0, and quartz phases, and the known Ih/XI, II, III/IX, V/XIII, VII/VIII, and X phases of ice. Moreover it identifies the structure that has since been identified experimentally as the porous ice XVII [156, 194]. This clearly demonstrates the potential of our structure searching approach. However, not all known ice phases are classified as synthesisable, which highlights the limitations of

the established convex hull approach: it fails to identify synthesisable metastable structures (such as ice IV and XII/XIV) and structures which can be stabilised and made synthesisable by thermodynamic and kinetic constraints other than pressure (such as XVI which initially forms by absorption of H_2 guest molecules). These limitations will be addressed in Section 5.5.

In addition, the ice counterparts of the zeolites with network codes IRR, IWV, SGT, and DDR and three hypothetical zeolites are identified as prime candidates for stabilisation by varying the system density. The counterparts of the IRR and IWV (not shown in Fig. 5.2), 207_1_4435, and DDR zeolites (labelled 1 and 17 in Fig. 5.2) are excellent candidates for stabilisation under negative pressure or by inclusion of guest molecules. The DDR counterpart, in particular, has previously been proposed as a possible clathrate hydrate [149]. The IRR and IWV counterparts exhibit substantially lower densities than the known CS-I, CS-II, and HS-III clathrate hydrates [171–174], suggesting that they may only become stable at large negative pressures. Conversely, the counterparts of the PCOD8172143 and 11_2_15848 zeolites (labelled 15 and 18 in Fig. 5.2) may be stabilised under positive pressure.

When comparing structures whose stabilities lie within a few meV/H_2O of each other, anharmonic quantum nuclear effects (QNE) must be accounted for, as highlighted by the stabilisation of ice Ih with respect to Ic by anharmonic quantum nuclear vibrations [191], as well as by effects of similar magnitude observed in other H-bonded crystals [195]. Anharmonic QNE in particular stabilise ice XVII and the HS-III clathrate by a few meV/molecule with respect to Ih. Their resultant zero pressure free energies exceed that of Ih by only 6.8 and 7.8 meV/molecule at the PBE-DFT level, respectively. The relative stability of the counterparts of the PCOD8172143 and 11_2_15848 zeolites with respect to Ih, on the other hand, is affected very little.

5.5 Using machine-learning to navigate the structural landscape

An analysis based on the energy-density convex hull as in Section 5.4 identifies candidate structures that can be stabilised by pressure. However, this does not address several crucial issues: (a) obtaining a global picture of configuration space from which one can gather an intuitive understanding of the relations between different polymorphs; (b) assessing the effectiveness of the structure search, identifying more or less obvious “gaps”; (c) selecting candidates stabilised by thermodynamic constraints other than pressure, such as absorption of guest molecules, electric fields, etc. All of these problems can be tackled effectively within a framework that borrows ideas from the machine-learning community. Points (a) and (b) are addressed by constructing an abstract, unbiased and general two-dimensional representation of configuration space in terms of the similarity relations between structures. Point (c) is addressed by generalising the conventional convex hull construction.

The first key ingredient of an intuitive representation of configuration space is a measure of the similarity of different configurations.
We use the smooth overlap of atomic positions (SOAP) kernel [15], combined with an entropy-

regularized matching (REMatch) approach [21]. This captures the fundamental symmetries of the problem, such as invariance to alternative representations of the same periodic structure, particle labelling, and rigid rotations and translations of the atomic coordinates. Based on the kernel-induced distance, we apply the sketch-map algorithm [51] to obtain a two-dimensional representation that reproduces as accurately as possible the (non-linearly transformed) distance between each pair of structures.
The construction and its parameters were designed to assess the oxygen lattice while being insensitive to proton disorder and hydrogen-bonding defects.
To assess the structural similarity between the configurations in the database under study we used a REMatch-SOAP kernel, as implemented in the `glosim.py` package (http://cosmo-epfl.github.io), with the following choice of hyperparameters controlling the description of atomic environments:

```
/src/glosim/glosim.py -n 9 -l 6 -c 5 -g 0.5 -periodic -nocenter 1 -kernel
rematch -gamma 0.01 -nonorm
```

Hydrogen atoms were included in the definition of the atom-density overlap, but were not considered as environment centers, so as to de-emphasise proton (dis)order in the definition of structural similarity. The choice of cut-off radius was tuned to achieve a clear separation between the known phases of ice in the database.

The non-linear `sketch-map` dimensionality reduction scheme was then applied to the SOAP kernel measure of similarity for 400 farthest-point-sampled landmark structures following the procedure described in reference [51] and using the following parameters: $\sigma = 0.12$, $A = 2$, $B = 4$, $a = 2$, $b = 2$

The resulting map is shown in Fig. 5.3 and provides a much-needed global picture of the lie of the land. Notably it is spanned by collective coordinates measuring abstract structural features, which in general cannot be related to single conventional observables such as density in a meaningful way. Consequently, their numerical values are not shown in Fig. 5.3.

Several observations highlight the heuristic value of such a representation: (1) The positions on the map correlate well with both density and lattice energy (see Fig. 5.4); (2) Structures related by proton-disorder, such as Ih/XI, III/IX, and VII/VIII, are clustered together; (3) Structures related by stacking disorder, such as Ih, Ic, and Isd, are clustered together; (4) The spread in energy at a given point on the map is comparable to the energy scale of stacking defects and H-bonding defects. H-bonding defects and different proton-orderings develop during the geometry optimisation of the ice structures, which (in analogy with their SiO_2 parent structures) are initialised with bond-centered protons.

Furthermore, the general structure of the map is consistent with the strategy we followed to construct our set of structures. The upper portion of the map, corresponding to tetrahedral ices and silica-like networks is densely sampled, with structures clustered in partially-overlapping regions. The lower part of the map, corresponding to very dense (e.g., ice X) and very open

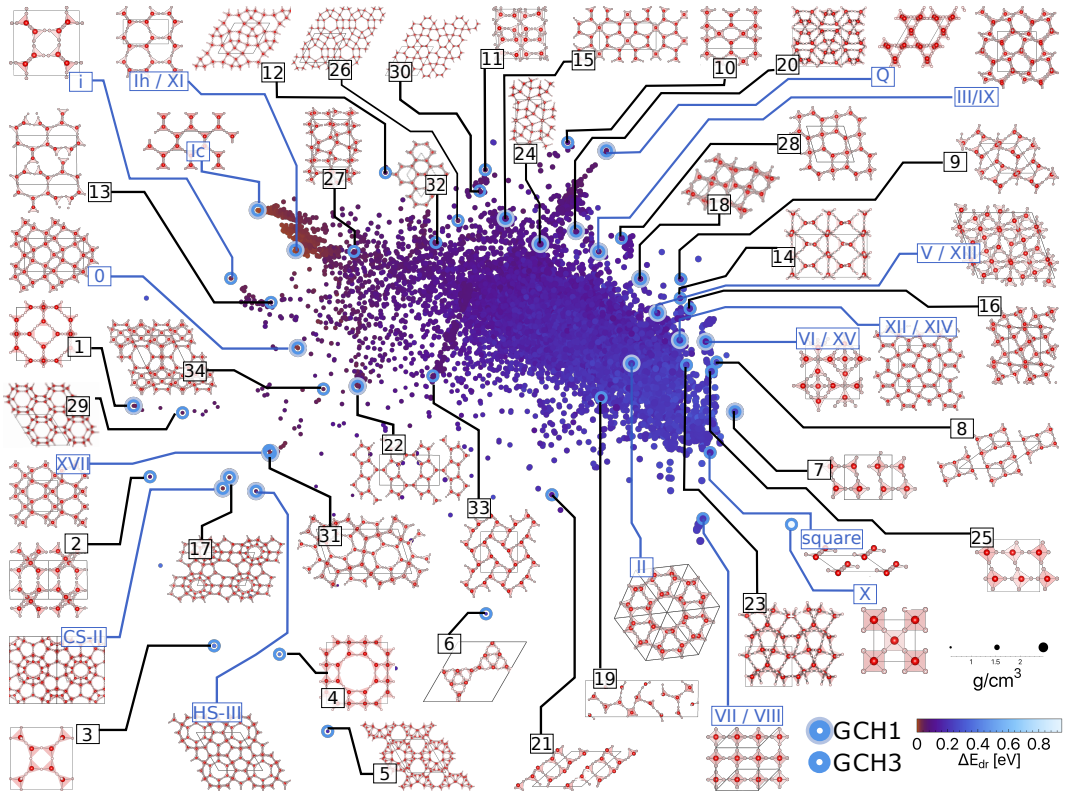


Figure 5.3 – Sketch-map of the structural similarity of 15,882 distinct PBE-DFT geometry-optimised ice structures. The sketch-map coordinates correlate strongly with density and configurational energy, but ultimately measure abstract structural features, which leaves their numerical value without intuitive meaning (therefore not shown in the axes). Instead the density and static lattice energy of each structure is encoded by the size and colour of the respective point on the map. Known ice phases are labelled in blue. The 34 new candidates are labelled in black and numbered in order of increasing dressed energy relative to the GCH3. Their atomic structures are shown to highlight their structural diversity.

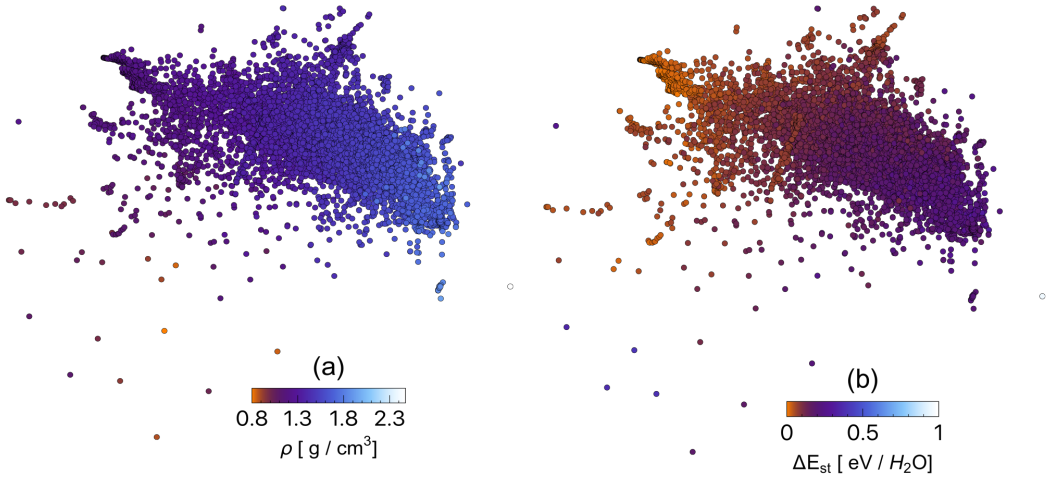


Figure 5.4 – The sketch-map coordinates correlations with (a) energy and (b) density

structures (e.g., those originating from the IZA zeolite dataset) is sparse. At high density, this sparsity results from the increasing importance of geometric constraints, which limit structural diversity and prohibit the formation of energetically feasible "mixed phases" containing structural patterns from two or more low-energy configurations. At low density our preselection strategy leads to sparse sampling. Sketch-map therefore provides indications of the quality of configuration-space sampling, which can be used to focus the structure search on the regions that need it most.

Finally, structures with low E_{dr} are projected onto the periphery of the map, whereas the central region is largely populated by defective, "mixed phase" structures that lie far from the energy-density convex hull. This suggests that a GCH construction could be used to identify configurations that can be stabilised (and made "synthesisable" [196]) by the application of appropriate thermodynamic constraints beyond the conventional molar volume manipulation. Conversely, the GCH allows us to identify structures, which can be stabilised by imposing thermodynamic and/or kinetic constraints that couple to the abstract structural features. In analogy with E_{dr} one can then define a generalised dressed energy $E_{\mathrm{dr}}^{(n)}$, that quantifies the stability of a given configuration subject to constraints that couple to the n structural features $\phi_1 \ldots \phi_n$. This approach lends itself naturally to the objective of the search, as it allows to probe all the structural breadth offered by a configuration search spanning a large pressure range.
Crucially, the KPCA components (unlike the collective variables defining the highly non-linear sketch-map projection) form a vector space in which the notion of convexity is well-defined. To maintain linearity with respect to the underlying structure compositions however, the global SOAP kernel must be reconstructed using an averaged descriptor.

By increasing the number n of features considered, the screening becomes progressively more

inclusive, since multiple axes of structural diversity are considered simultaneously.

Including three KPCA descriptors in the GCH construction, we identify 50 structures within 20 meV of the GCH (see Fig. 5.3), which include all of the known ice phases except ice IV. Ice IV is not classified as synthesisable due to its particularly high lattice energy, which is consistent with the experimental observation that ice IV is metastable and only forms occasionally upon slow heating of high-density amorphous ice before annealing to ice III, V, or VI.

The 50 structures also include the theoretical i, 0, quartz, and square phases, the CS-II clathrate hydrate (which is identical in structure to ice XVI), and the HS-III clathrate hydrate. Furthermore, we identify 34 new configurations which are excellent candidates for experimental formation and which we propose as candidates for ices XVIII through LI.
Among them are, in particular, the ice counterparts of the DDR, SGT, and NON zeolites, which were previously suggested as promising candidates for clathrate hydrates by Tribello *et al.* [149], and two structures reminiscent of a high pressure structure with $Pbcm$ symmetry proposed by Hermann *et al.* [124].
Notably, while the most promising candidates for experimental formation (as indicated by their ordering in Fig. 5.3) are low-density ice counterparts of different zeolite networks, the counterpart of the ITT network, which was suggested as the most stable "aeroice" structure below around -0.4 GPa in Ref. [189], is dynamically unstable at the employed level of theory. For reference, using the rPW86-vdW2 exchange-correlation functional ITT ice is still much less stable than IRR ice proposed as stabilizable in this work.
It is worth noting that the counterpart of the LTA zeolite (structure 4 in Fig. 5.3) has also most recently received attention as an "ultralow" density clathrate ice in Ref. [197].

5.6 Sensitivity to choice of density functional

While the GCH generally depends on the kernel, our choice of kernel representation is very general and rather unbiased, which is reflected by the weak dependence of this selection of structures on the choice of hyperparameters for the SOAP kernel (as shown in Chapter 4).
Notably, the GCH is also remarkably insensitive to the details of the underlying (free) energy calculations.
To showcase its robustness to the choice of comparable potential energy surfaces here we present a set of comparisons of the convex hulls obtained by recalculating the best 57 structures found in the PBE based search using other standard choices of functionals.
In this exercise one can see how the use of the PBE xc-functional is further justified by investigating how the relative stability of the known and novel structures proposed for synthesis is affected by the choice of xc-functional. Santra *et al.* [198] have shown that the rPW86-vdW2 xc-functional [199] produces particularly accurate relative stabilities for the known phases of ice compared to different semi-local xc-functionals. More recently the SCAN functional of Sun *et al.* [200] has been shown to reproduce said relative stabilities in equivalently good agreement with experiment [201]. The PBE, rPW86-vdW2, and SCAN energies of the 50 synthesisable

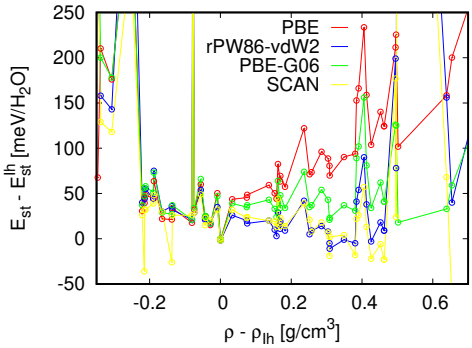


Figure 5.5 – E_{st} with respect to ice Ih in meV/H_2O as a function of ρ in g/cm^3 for the 50 synthesisable PBE-relaxed structures highlighted in the sketch-map (and 7 additional structures just more than 20 meV/H_2O above the GCH) obtained using different xc-functionals. The E_{st} of the PBE-relaxed structures were calculated using the PBE [118] functional, the PBE functional with the Grimme (G06) dispersion correction [202], the rPW86-vdW2 functional [199], and the SCAN functional [200].

PBE-relaxed structures highlighted and 7 additional structures just more than 20 meV/H_2O above the GCH (calculated using Quantum Espresso [130][1]) is shown as a function of density in Fig. 5.5. However, the more stringent and representative test in the context of this work is the comparison of relative stabilities after geometry optimisation with the respective xc-functional (see Fig. 5.6). Both figures highlight the importance of dispersion effects on the static lattice energies of different ice structures and show that the energy-density CH (only explicitly shown in Fig. 5.6) is sensitive to the choice of xc-functional, even though the set of structures within $\sim$ 10 meV/H_2O is insensitive to it. Once again, the GCH construction is largely unaffected by the choice of xc-functional, rendering the central features of the configuration space map shown in Fig.5.3 insensitive to the choice of xc-functional. This is most strikingly demonstrated by re-evaluating the GCH constructed on the first three KPCA descriptors for the 136 most promising candidates, but using the E_{st} obtained using the rPW86-vdW2 [199] xc-functional. Out of 38 vertices obtained using PBE all but one are recovered using rPW86-vdW2 and only three additional vertices emerge. To sum it up, 37 out of 38 structures are still identified as GCH *vertices* when the lattice energies of the structures highlighted in Fig. 5.3 are computed using the dispersion-corrected rPW86-vdW2 exchange-correlation functional [199] instead of the PBE functional, despite significant differences with respect to the PBE lattice energies. The rPW86-vdW2 functional has been shown to be particularly accurate for the known phases of ice [198].
In contrast the energy-density CH depends more sensitively on the choice of exchange-correlation functional. One may wonder whether unoptimised structures might provide a sufficient structural database, thereby eliminating the expensive geometry-optimisation

[1] with a plane-wave energy cut-off of 40 Rydberg, the same k-point grids employed in the original PBE-DFT calculations, and the O.pbe-rrjkus.UPF, H.pbe-rrjkus.UPF, O.pbe-hgh.UPF, and H.pbe-hgh.UPF pseudopotentials from http://www.quantum-espresso.org.

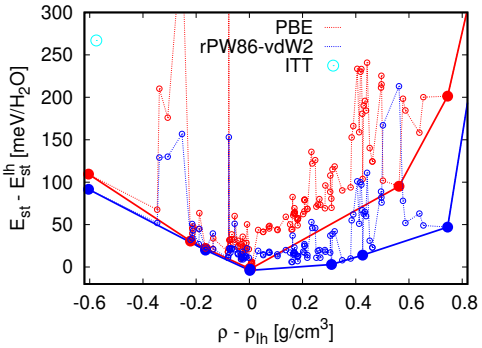


Figure 5.6 – $E_{st}(\rho)$ with respect to ice Ih in meV/H_2O for the 136 structures within 20 meV/H_2O of the original GCH after full geometry-optimisation using the PBE and rPW86-vdW2 functionals, respectively. The respective energy-density CH are shown as solid lines. The CH vertices are highlighted as thick, filled circles. The ice counterpart of the ITT zeolite network, which was suggested as the most stable "aeroice" structure below around -0.4 GPa in Ref. [189], is highlighted in cyan, but is unstable at the PBE level of theory and still far from stable at the rPW86-vdW2 level of theory.

step, which also provides energetic information as a proxy for stability. Indeed, a GCH construction based on structural information alone successfully identifies most of the known ice phases. However, this apparent success crucially relies on identifying synthesisable configurations as "extremal". This in turn requires geometry optimisation in order to collapse the large configuration space volumes associated with synthesisable configurations onto the corresponding deep minima in the free energy surface. The correlation between the depth of a minimum and the configuration space volume it attracts has been shown explicitly for Lennard-Jones clusters [203, 204].

5.7 The novel phases

In Table 5.1 summarises the structural and energetic details of the structures highlighted on the sketch-map in Fig.5.3. A typical characteristic of the GCH construction is that an increasing the number of kPCA components produces an increasingly inclusive selection. In Fig. 5.7 shows GCH selections with one and three components, setting a cutoff threshold on the dressed energies at 20 meV. Points below this threshold are coloured according to their GCH dressed energies.

In both cases the number of structures with low dressed energies is large, since the database includes configurations that are only distinguished by the presence of proton disorder or stacking faults, which are expected to change the binding energy by just a few meV/H_2O. While it would be possible to inspect structures manually to identify those that are genuinely structurally distinct, it is easier to further reduce the number of proposed phases by an additional clustering step, based on the high dimensional similarity matrix.

Index	Original label	$\frac{H_2O}{\text{cell}}$	S.G.	ρ	ρ_{ref}	E_{dr}^{ρ}	E_{dr}^{GCH-1}	E_{dr}^{GCH-3}
1	207_1_4435	12	$Im\bar{3}m$	0.965	(0.772)	0.00	19.64	0.00
2	12_2_29187	6	$I4/m$	0.998	0.797	38.13	44.20	0.00
3	ACO	16	$Im\bar{3}m$	1.109	(0.887)	364.22	338.53	0.00
4	LTA	24	$Pm\bar{3}m$	0.933	(0.764)	349.61	341.04	0.00
5	BSV	48	$Ia\bar{3}d$	0.879	(0.703)	127.77	131.70	0.00
6	169_2_7915	12	$P6_122$	0.848	(0.678)	155.48	141.81	0.00
7	53_3_726600	16	$Cmcm$	1.682	(1.346)	141.80	0.00	0.00
8	20_2_26425	6	$P2_1$	1.592	(1.274)	165.75	58.47	0.00
9	12_2_32449	6	$C2/c$	1.577	(1.262)	100.96	65.68	0.00
10	84_2_1419	6	$P4_2/m$	1.349	1.089	57.14	37.74	0.00
11	61_2_8842	16	$R3$	1.339	1.085	26.36	20.96	0.00
12	169_2_10608	12	$C222_1$	1.177	0.946	50.76	38.10	0.00
13	PCOD8047078	12	$P2_1$	1.131	0.920	53.84	60.63	0.00
14	67_2_1563	16	$Pbma$	1.570	1.171	88.64	51.37	0.00
15	PCOD8172143	10	$Pnn2$	1.344	1.092	19.50	9.22	1.52
16	152_2_118474	9	$P3_121$	1.599	1.270	89.78	61.98	1.79
17	DDR	40	$C2/m$	0.996	0.801	19.02	19.83	2.40
18	11_2_15848	8	$C2/c$	1.535	1.236	32.31	7.85	4.36
19	91_2_8335121	16	$P1$	1.688	(1.350)	17.15	29.05	5.44
20	PCOD8301974	16	$I4_1/a$	1.443	1.110	31.52	14.31	8.99
21	PCOD8045578	8	$R\bar{3}m$	1.422	(1.138)	83.74	69.60	10.48
22	58_2_511	12	$Cmcm$	1.112	0.908	23.09	17.57	13.41
23	151_2_4949650	9	$P3_112$	1.638	1.313	64.30	24.97	13.54
24	PCOD8007225	16	$P\bar{1}$	1.438	(1.150)	30.20	19.83	15.54
25	2_2_342692	4	$Pbnm$	1.681	(1.345)	128.00	37.64	15.76
26	PCOD8321499	18	$C222_1$	1.260	1.024	35.38	21.64	15.97
27	PCOD8047931	16	$P2_1$	1.219	0.984	40.78	36.67	17.39
28	15_2_201714	6	$Ibam$	1.472	1.068	49.18	21.46	17.76
29	MAR	72	$P2_1/m$	0.971	0.798	16.10	34.48	17.93
30	PCOD8324623	18	$P2_1$	1.323	1.081	38.29	28.75	18.20
31	SGT	32	$I4_1/amd$	0.972	0.794	9.97	23.46	19.02
32	20_2_28176	6	$C2$	1.262	(1.010)	38.23	27.73	19.19
33	14_2_48453	8	$Pmnn$	1.355	(1.084)	36.73	33.36	19.64
34	NON	22	$Fmmm$	1.050	0.860	16.09	21.19	19.64

Table 5.1 – Structure data for novel candidate ice phases. Columns one and two provide the mapping between the structure indices and the original labels of the corresponding four-connected networks in the databases of Treacy *et al.* [183, 205, 206] and Deam *et al.* [184] and the IZA atlas of zeolites [182]. The next columns provide the number of molecules per unit cell, the space group, and the initial and refined PBE-DFT density, ρ and ρ_{ref} (measured in [g/cm^3]). Brackets indicate values that have been estimated noting that the refined PBE-DFT densities are consistently around 20 % smaller than those from the initial PBE-DFT calculations. The last three columns contain the dressed energies relative to the CH built on the density (E_{dr}^{ρ}), a generalised convex hull (GCH) with one principal component (E_{dr}^{GCH-1}), and three components (E_{dr}^{GCH-1}). All energies are expressed in meV/H_2O.

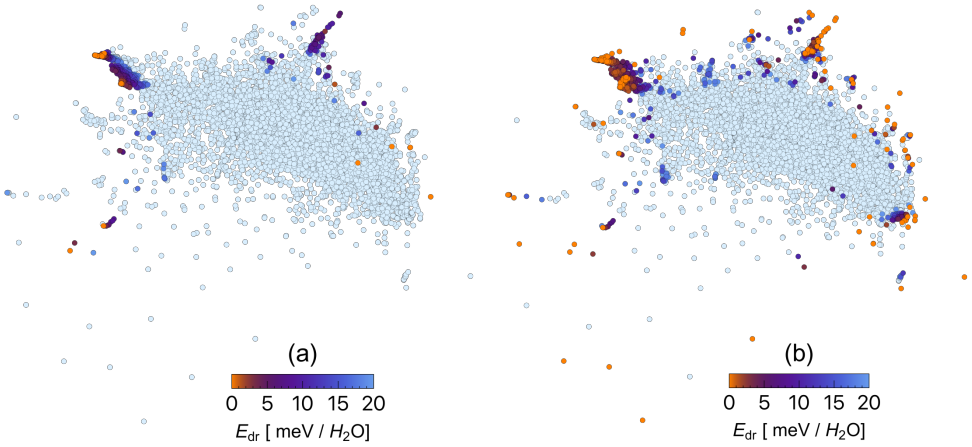


Figure 5.7 – Sketch-map coloured according to distance from the GCH as a measure of stabilisability. Structures are coloured according to their dressed energies with respect to the GCH constructions using (a) one and (b) three kPCA components, respectively. Structures shown in bright blue are more than 20 meV/H_2O above the hulls, while the rest is coloured according to the legends.

5.8 Discussion

The success of the GCH construction in discovering the known ice phases and clathrate hydrates entirely *a priori* highlights that, although kinetic factors play an important role in determining which ice phases are formed in practice, structural and simple energetic considerations can provide a great deal of physical insight. More importantly, it demonstrates that the GCH approach does not simply discern structurally diverse configurations, but very effectively selects configurations which can be formed in experiment. It thereby provides strong support for the 34 proposed, new, structurally diverse candidates for ices XVIII to LI. This should spur experimental efforts to ratify our predictions.

At this stage the candidates are embedded in a human-readable sketch-map of configuration space mainly as an aide in developing an intuitive understanding of the relation of the proposed candidates to the known ice phases. However, highlighting the phase transitions between the known ice phases suggests that proximity on the sketch-map is a good indicator for the existence a viable transition pathway. In conjunction with the GCH construction the sketch-map therefore provides a tractably small and yet structurally diverse set of synthesisable candidate structures and a means of identifying endpoints and suitable reaction coordinates for further investigation of formation pathways, for example, using umbrella sampling [207], forward flux sampling [164] or enhanced sampling metadynamics approaches [208], which have already proven successful in simulating the nucleation of ice [165–169]. The relation of the KPCA descriptors in the GCH construction to conventional quantities, such as density, vibrational spectra, and concentrations of different types of guest molecules, promises to provide more direct guidance in identifying experimental formation pathways. However, this goes beyond the scope of this study.

In addition, the approach demonstrated in this work sheds light on the energetics of proton-order/disorder, stacking-disorder, as well as H-bonding and planar defects, and also provides a glimpse of the preferred (quasi-) two-dimensional forms of ice.

In its current state the biggest limitation of our structure search is the preselection cut-off on system size. Relaxing this cut-off will drive the structure search towards completeness, which is the obvious next step. More generally, the connection between structural patterns and configurational energy exposed by the sketch-map dimensionality reduction suggests an expedient recipe for even more extensive database-driven searches. These need not be limited to crystalline water ice, but could range from other tetrahedrally coordinated systems, such as silica or the carbon allotropes, to liquid, disordered, and glassy systems.

Conclusions Part III

6 Conclusions

The high-throughput search for novel phases of materials is nowadays an integral part of the materials discovery pipeline, and its current limits can be alleviated by the application of novel machine learning techniques.
The advent of accurate machine learning potentials has made it increasingly feasible to bypass the computational costs involved in an ab-initio calculation, thus allowing to concentrate resources to compute accurate measures of systems' free energies and properties. However, the exponential complexity associated with the exploration of a compound's space and the complex transition between its minima remain a fundamental challenge for efficient in-silico materials design.
Understanding the key thermodynamic mechanisms underlying the transitions between metastable phases of materials can provide experimentalists a fundamental starting point for the successful synthesis of simulated materials, and represents a key milestone in the roadmap towards a property guided materials design pipeline.
In this thesis, we have introduced a generalisation of the convex hull construction, a tool which is commonly used to screen configurations that exhibit stability under a known thermodynamic drive. The GCH employs as thermodynamic drives a set of fully agnostic, data-driven structural features which in turn encode the effects of any general thermodynamic force. We show that the GCH recovers phases that could be stabilised by applications of unconventional fields, like magnetic fields (H_xO_{1-x} binary compounds), the substitution of a chemical species (pentacenes molecular crystals) as well as negative pressures (zeolitic phases of ice).
Further, it introduces a probabilistic measure for each structure to be a vertex, and thus proposing a quantitative measure of a candidate's stabilisability. In order to obtain these probabilities, the structures are lightly randomised in their coordinates, and their corresponding hulls are sampled until their their probabilities of being vertices converge.
This mechanism allows firstly to account, albeit very approximately, for thermal fluctuations, and secondly, to allow competing configurations to lower each other probabilities, so to allow for the pruning of similar configurations at every iteration.
The result of the GCH construction is a list of configurations that constitute its vertices, and thus exhibit high potential for stabilisation. We further suggest a scheme to decode which

thermodynamic field could couple to the extracted structural descriptors, by looking at correlations with pre-existing calculated properties.
The interpretation of the abstract structural variables has much potential for further improvements: the stabilisation mechanisms, however subtle, appear to have similar structural effects across different compounds; thus a transfer learning scheme could be used a starting step to suggest unexplored stabilisation mechanism for novel systems. As a more concrete example, one can notice how concerted reorientation of dimer sub-units can be linked to an application of an electromagnetic field, irrespective of the species involved.
A byproduct of a GCH construction is what we called the dressed energy E_{hull}, which measures a structure's distance from the constructed convex hull(s). This quantity represents more than the direct regression of a system's energy, having a direct implication on the configuration proximity to becoming stabilisable.
If one considers the construction of GCH as a training over a set of configurations labelled by their energies, and E_{hull} as the output of the prediction, it is easy to consider the GCH method as a supervised machine learning model and use it to predict the potential stability of a novel structure, given its position on the GCH space. This quantity could be used to enhance crystal structure searching schemes to point searches towards unsampled, high energy regions, and rendering the search more efficient. Moreover, one could use this scheme to early stop structure minimisations which show low potential to get close, or surpass, the surfaces of the GCH. The results of these investigations were built to produce a quantitative benchmark of our models, and as such, many of the phases suggested by the GCH constructions haven't been further explored or tested for further stability studies.
The search for novel phases of hydrogen under high pressure is currently an active field of research, with a strong effort aimed at finding the Wigner-Huntington transition leading to a metallic, superconducting structure. A point of continuation of this study could be to characterise in more details the free energies of configurations emerged from the high-pressure hydrogen set, and test for the presence of potential candidates which exhibit metallic behaviour.
The study of the effects of chemical substitutions on the pentacene sets links the effect of changes of molecular fragments to the system's underlying potential energy surface. This application has the potential to pave the way for a novel approach to characterise the stabilising effects that specific moieties can produce when packed in a novel molecular crystal cell.
Finally, we show the exploratory power of the GCH construction by using it to lead a crystal structure search aimed at uncovering unexplored regions of ice's phase diagram. The result of this work serves both as a benchmark - in that we recover all the known phases of ice contained in the set - and a proof of the versatility of our model in dealing with a database spanning a phase space ranging from negative pressures to units of GPa. This is substantiated by testing the regime of stability of the proposed phases and showing them to be competitive with the known phases at different ranges of pressures, showing promise for further stability studies.
To summarise, we have developed a method to facilitate the discovery of potentially stable configurations coming from crystal structure searches. We showed its applicability in a wide range of complex systems and used it to suggest or (re)discover their underlying stabilisation

mechanisms. This novel approach can be used to both to uncover subtle, unknown structure to property relations, and to extract meta-stable configurations that would have escaped conventional screening procedures.

Bibliography

[1] National Research Council. “Materials Science and Engineering – Volume I, The History, Scope, and Nature of Materials Science and Engineering”. In: *Materials and Man's Need* (1975), p. 288. DOI: 10.17226/10436.

[2] Jonathan Wood. “The top ten advances in materials science”. In: *Materials Today* 11.1-2 (Jan. 2008), pp. 40–45. DOI: 10.1016/S1369-7021(07)70351-6.

[3] Thomas Hofmann, Bernhard Schölkopf, and Alexander J. Smola. “Kernel methods in machine learning”. In: *Ann. Statist.* 36.3 (June 2008), pp. 1171–1220. DOI: 10.1214/009053607000000677.

[4] Yehuda B. Band and Yshai Avishai. “Approximation Methods”. In: *Quantum Mechanics with Applications to Nanotechnology and Information Science*. Elsevier, Jan. 2013, pp. 303–366. DOI: 10.1016/B978-0-444-53786-7.00007-1.

[5] G Ceder HL Chen G Hautier. “Synthesis, computed stability, and crystal structure of a new family of inorganic compounds: carbonophosphates”. In: *J. Am. Chem. Soc.* 134 (2012), pp. 19619–19627. DOI: 10.1021/ja3040834.

[6] R Gautier. “Prediction and accelerated laboratory discovery of previously unknown 18-electron ABX compounds”. In: *Nat. Chem.* 7 (2015), pp. 308–316. DOI: 10.1038/nchem.2207.

[7] Wilhelm F. Maier, Klaus Stowe, and Simone Sieg. *Combinatorial and high-throughput materials science*. 2007. DOI: 10.1002/anie.200603675.

[8] R Catlow SM Woodley. “Crystal structure prediction from first principles”. In: *Nat. Mater.* 7 (2008), pp. 937–946. DOI: 10.1038/nmat2321.

[9] Y. Paul Handa, D. D. Klug, and Edward Whalley. “Difference in energy between cubic and hexagonal ice”. In: *The Journal of Chemical Physics* 84.12 (June 1986), pp. 7009–7010. DOI: 10.1063/1.450622.

[10] Kresten Lindorff-Larsen et al. “Picosecond to Millisecond Structural Dynamics in Human Ubiquitin”. In: *The Journal of Physical Chemistry B* 120.33 (Aug. 2016), pp. 8313–8320. DOI: 10.1021/acs.jpcb.6b02024.

[11] V. Botu et al. “Machine learning force fields: Construction, validation, and outlook”. In: *Journal of Physical Chemistry C* 121.1 (2017), pp. 511–522. DOI: 10.1021/acs.jpcc.6b10908.

[12] Rafael Gómez-Bombarelli et al. “Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules”. In: *ACS Central Science* 4.2 (Feb. 2018), pp. 268–276. DOI: 10.1021/acscentsci.7b00572.

[13] Katja Hansen et al. “Machine Learning Predictions of Molecular Properties: Accurate Many-Body Potentials and Nonlocality in Chemical Space”. In: (2015). DOI: 10.1021/acs.jpclett.5b00831.

[14] Matthias Rupp et al. “Fast and accurate modeling of molecular atomization energies with machine learning”. In: *Physical Review Letters* 108.5 (Jan. 2012). DOI: 10.1103/PhysRevLett.108.058301.

[15] Albert P. Bartók, Risi Kondor, and Gábor Csányi. “On representing chemical environments”. In: *Phys. Rev. B* 87 (18 May 2013), p. 184115. DOI: 10.1103/PhysRevB.87.184115.

[16] Andrea Grisafi et al. “Symmetry-Adapted Machine Learning for Tensorial Properties of Atomistic Systems”. In: *Physical Review Letters* 120.3 (Jan. 2018), p. 036002. DOI: 10.1103/PhysRevLett.120.036002.

[17] Federico M Paruzzo et al. “Chemical Shifts in Molecular Solids by Machine Learning”. In: *Nature communications* 9.1 (2018), pp. 1–10. DOI: 10.1038/s41467-018-06972-x.

[18] Alireza Khorshidi and Andrew A. Peterson. “Amp: A modular approach to machine learning in atomistic simulations”. In: *Computer Physics Communications* 207 (Oct. 2016), pp. 310–324. DOI: 10.1016/J.CPC.2016.05.010.

[19] Piero Gasparotto and Michele Ceriotti. “Recognizing molecular patterns by machine learning: An agnostic structural definition of the hydrogen bond”. In: *The Journal of Chemical Physics* 141.17 (Nov. 2014), p. 174110. DOI: 10.1063/1.4900655.

[20] Benjamin A. Helfrecht et al. “A new kind of atlas of zeolite building blocks”. In: *Journal of Chemical Physics* 151.15 (Oct. 2019), p. 154112. DOI: 10.1063/1.5119751.

[21] S. De et al. “Comparing molecules and solids across structural and alchemical space”. In: *Physical Chemistry Chemical Physics* 18 (2016), p. 13754. DOI: 10.1039/C6CP00415F.

[22] Tian Xie and Jeffrey C Grossman. “Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties”. In: *Physical review letters* 120.14 (2018), p. 145301. DOI: 10.1103/PhysRevLett.120.145301.

[23] Felix A. Faber et al. “Machine Learning Energies of 2 Million Elpasolite (ABC_2D_6) Crystals”. In: *Phys. Rev. Lett.* 117 (13 Sept. 2016), p. 135502. DOI: 10.1103/PhysRevLett.117.135502.

[24] Andrea Grisafi and Michele Ceriotti. “Incorporating long-range physics in atomic-scale machine learning”. In: *Journal of Chemical Physics* 151.20 (Nov. 2019), p. 204105. DOI: 10.1063/1.5128375.

[25] Bingqing Cheng et al. “Ab initio thermodynamics of liquid and solid water”. In: *Proceedings of the National Academy of Sciences of the United States of America* 116.4 (Jan. 2019), pp. 1110–1115. DOI: 10.1073/pnas.1815117116.

[26] Volker L. Deringer et al. *Structural transitions in dense disordered silicon from quantum-accurate ultra-large-scale simulations*. 2019. DOI: 1912.07344.

[27] Marcos F. Calegari Andrade et al. "Free energy of proton transfer at the water–TiO2 interface from ab initio deep potential molecular dynamics". In: *Chem. Sci.* 11 (9 2020), pp. 2335–2341. DOI: 10.1039/C9SC05116C.

[28] Volker L. Deringer et al. "Data-driven learning and prediction of inorganic crystal structures". In: *Faraday Discussions* 211.0 (Oct. 2018), pp. 45–59. DOI: 10.1039/c8fd00034d.

[29] Rui Shi and Hajime Tanaka. "Microscopic structural descriptor of liquid water". In: *The Journal of Chemical Physics* 148.12 (Mar. 2018), p. 124503. DOI: 10.1063/1.5024565.

[30] Jack Yang et al. "Large-Scale Computational Screening of Molecular Organic Semiconductors Using Crystal Structure Prediction". In: *Chemistry of Materials* 30.13 (July 2018), pp. 4361–4371. DOI: 10.1021/acs.chemmater.8b01621.

[31] Andrea Anelli et al. "Generalized convex hull construction for materials discovery". In: *Physical Review Materials* 2.10 (Oct. 2018). DOI: 10.1103/PhysRevMaterials.2.103804.

[32] Edgar A. Engel et al. "Mapping uncharted territory in ice from zeolite networks to ice structures". In: *Nature Communications* 9.1 (Dec. 2018), p. 2173. DOI: 10.1038/s41467-018-04618-6.

[33] Shachar Mirkin et al. "Listening Comprehension over Argumentative Content". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, Oct. 2018, pp. 719–724. DOI: 10.18653/v1/D18-1078.

[34] Guillaume Lample et al. *Phrase-Based and Neural Unsupervised Machine Translation*. 2018. DOI: arXiv:1804.07755.

[35] David L. Phillips. "A Technique for the Numerical Solution of Certain Integral Equations of the First Kind". In: *J. ACM* 9.1 (Jan. 1962), pp. 84–97. DOI: 10.1145/321105.321114.

[36] Robert Tibshirani. "Regression Shrinkage and Selection via the Lasso". In: 58 (1996), pp. 267–288. DOI: 10.2307/2346178.

[37] Alexandra Nagy and Vincenzo Savona. "Variational Quantum Monte Carlo Method with a Neural-Network Ansatz for Open Quantum Systems". In: *Phys. Rev. Lett.* 122 (25 June 2019), p. 250501. DOI: 10.1103/PhysRevLett.122.250501.

[38] Thorsten Joachims. *Learning to classify text using support vector machines*. Vol. 668. Springer Science & Business Media, 2002. DOI: 10.1007/978-1-4615-0907-3.

[39] Thomas N Kipf and Max Welling. "Semi-supervised classification with graph convolutional networks". In: *arXiv preprint arXiv:1609.02907* (2016). DOI: arXiv:1609.02907.

[40] Karen Simonyan and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition". In: *arXiv preprint arXiv:1409.1556* (2014). DOI: arXiv:1409.1556.

[41] Gao Huang et al. “Semi-supervised and unsupervised extreme learning machines”. In: *IEEE transactions on cybernetics* 44.12 (2014), pp. 2405–2417. DOI: r10.1109/TCYB.2014.2307349.

[42] David J. C. MacKay. *Information Theory, Inference Learning Algorithms*. USA: Cambridge University Press, 2002. DOI: 10.5555/971143.

[43] Christopher M. Bishop. *Pattern recognition and machine learning*. Springer, 2006, p. 738. DOI: 10.5555/1162264.

[44] R. Bennett. “The intrinsic dimensionality of signal collections”. In: *IEEE Transactions on Information Theory* 15.5 (Sept. 1969), pp. 517–525. DOI: 10.1109/TIT.1969.1054365.

[45] Colleen D. Cutler. “A Review of the theory and estimation of fractal dimension”. In: *Dimension Estimation and Models*, pp. 1–107. DOI: 10.1142/9789814317382_0001.

[46] Elena Facco et al. “The intrinsic dimension of protein sequence evolution”. In: *PLOS Computational Biology* 15.4 (Apr. 2019), pp. 1–16. DOI: 10.1371/journal.pcbi.1006767.

[47] Yusuke Naritomi and Sotaro Fuchigami. “Slow dynamics in protein fluctuations revealed by time-structure based independent component analysis: The case of domain motions”. In: *The Journal of Chemical Physics* 134.6 (2011), p. 065101. DOI: 10.1063/1.3554380.

[48] Ian T. Jolliffe and Jorge Cadima. “Principal component analysis: a review and recent developments”. In: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 374.2065 (2016), p. 20150202. DOI: 10.1098/rsta.2015.0202.

[49] Alain Berlinet and Christine Thomas-Agnan. “Theory”. In: *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Boston, MA: Springer US, 2004, pp. 1–54. DOI: 10.1007/978-1-4419-9096-9_1.

[50] J B Tenenbaum, V de Silva, and J C Langford. “A global geometric framework for nonlinear dimensionality reduction.” In: *Science (New York, N.Y.)* 290.5500 (Dec. 2000), pp. 2319–23. DOI: 10.1126/science.290.5500.2319.

[51] Michele Ceriotti, Gareth A. Tribello, and Michele Parrinello. “Simplifying the representation of complex free-energy landscapes using sketch-map”. In: *Proceedings of the National Academy of Sciences* 108.32 (2011), pp. 13023–13028. DOI: 10.1073/pnas.1108486108.

[52] “Demonstrating the Transferability and the Descriptive Power of Sketch-Map”. In: *Journal of Chemical Theory and Computation* 9.3 (Mar. 2013), pp. 1521–1532. DOI: 10.1021/ct3010563.

[53] “Mapping the conformational free energy of aspartic acid in the gas phase and in aqueous solution”. In: *The Journal of Chemical Physics* 146.14 (Apr. 2017), p. 145102. DOI: 10.1063/1.4979519.

[54] B. Schölkopf, A. Smola, and K. Müller. “Nonlinear Component Analysis as a Kernel Eigenvalue Problem”. In: *Neural Computation* 10.5 (July 1998), pp. 1299–1319. DOI: 10.1162/089976698300017467.

[55] Marco Cuturi. “Sinkhorn Distances: Lightspeed Computation of Optimal Transport”. In: (). DOI: arXiv:1306.0895.

[56] David L. Donoho and Carrie Grimes. “Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data”. In: *Proceedings of the National Academy of Sciences of the United States of America* 100.10 (May 2003), pp. 5591–5596. DOI: 10.1073/pnas.1031596100.

[57] Laurens Van Der Maaten and Geoffrey Hinton. “Visualizing Data using t-SNE”. In: 9 (2008), pp. 2579–2605. DOI: http://www.jmlr.org/papers/v9/vandermaaten08a.html.

[58] R. R. Coifman et al. “Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps”. In: *Proceedings of the National Academy of Sciences of the United States of America* 102.21 (May 2005), pp. 7426–7431. DOI: 10.1073/pnas.0500334102.

[59] Yoshua Bengio, Martin Monperrus, and Hugo Larochelle. “Nonlocal estimation of manifold structure”. In: *Neural Computation* 18.10 (Oct. 2006), pp. 2509–2528. DOI: 10.1162/neco.2006.18.10.2509.

[60] Mayu Sakurada and Takehisa Yairi. “Anomaly detection using autoencoders with nonlinear dimensionality reduction”. In: *ACM International Conference Proceeding Series*. Vol. 02-December-2014. Association for Computing Machinery, Dec. 2014, pp. 4–11. DOI: 10.1145/2689746.2689747.

[61] *Unsupervised Content-Based Identification of Fake News Articles with Tensor Decomposition Ensembles*. Tech. rep. 2018. DOI: 10.475/123_4.

[62] Yong Ge et al. “A taxi driving fraud detection system”. In: *Proceedings - IEEE International Conference on Data Mining, ICDM*. 2011, pp. 181–190. DOI: 10.1109/ICDM.2011.18.

[63] David Arthur and Sergei Vassilvitskii. “k-means++: The Advantages of Careful Seeding”. In: *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms* (). DOI: 10.1145/1283383.1283494.

[64] Peter J. Rousseeuw. “Silhouettes: A graphical aid to the interpretation and validation of cluster analysis”. In: *Journal of Computational and Applied Mathematics* 20.C (1987), pp. 53–65. DOI: 10.1016/0377-0427(87)90125-7.

[65] Martin Ester et al. “A density-based algorithm for discovering clusters in large spatial databases with noise”. In: *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining* (1996), pp. 226–231. DOI: 10.5555/3001460.3001507.

[66] Ricardo J.G.B. Campello et al. “Hierarchical density estimates for data clustering, visualization, and outlier detection”. In: *ACM Transactions on Knowledge Discovery from Data* 10.1 (July 2015). DOI: 10.1145/2733381.

[67] Oliver Fleetwood et al. “Molecular insights from conformational ensembles via machine learning”. In: *Biophysical Journal* (Dec. 2019). DOI: 10.1016/j.bpj.2019.12.016.

[68] Linus. Pauling. “The nature of the chemical bond. Application of results obtained from the quantum mechanics and from a theory of paramagnetic susceptibility to the structure of molecules”. In: *Journal of the American Chemical Society* 53.4 (Apr. 1931), pp. 1367–1400. DOI: 10.1021/ja01355a027.

[69] E. Prodans and W. Kohn. “Nearsightedness of electronic matter”. In: *Proceedings of the National Academy of Sciences of the United States of America* 102.33 (Aug. 2005), pp. 11635–11638. DOI: 10.1073/pnas.0505436102.

[70] Jörg Behler and Michele Parrinello. “Generalized neural-network representation of high-dimensional potential-energy surfaces”. In: *Physical Review Letters* 98.14 (Apr. 2007). DOI: 10.1103/PhysRevLett.98.146401.

[71] Philipp Geiger and Christoph Dellago. “Neural networks for local structure detection in polymorphic systems”. In: *Journal of Chemical Physics* 139.16 (Oct. 2013). DOI: 10.1063/1.4825111.

[72] Andreas Singraber, Jörg Behler, and Christoph Dellago. “Library-Based LAMMPS Implementation of High-Dimensional Neural Network Potentials”. In: *Journal of Chemical Theory and Computation* 15.3 (Mar. 2019), pp. 1827–1840. DOI: 10.1021/acs.jctc.8b00770.

[73] M. Gastegger et al. “WACSF - Weighted atom-centered symmetry functions as descriptors in machine learning potentials”. In: *Journal of Chemical Physics* 148.24 (June 2018). DOI: 10.1063/1.5019667.

[74] Giulio Imbalzano et al. “Automatic selection of atomic fingerprints and reference configurations for machine-learning potentials”. In: *The Journal of chemical physics* 148.24 (2018), p. 241730. DOI: 10.1063/1.5024611.

[75] Albert P. Bartók and Gábor Csányi. “Gaussian approximation potentials: A brief tutorial introduction”. In: *International Journal of Quantum Chemistry* 115.16 (Aug. 2015), pp. 1051–1057. DOI: 10.1002/qua.24927.

[76] Tim Würger et al. “Data Science Based Mg Corrosion Engineering”. In: *Frontiers in Materials* 6 (Apr. 2019). DOI: 10.3389/fmats.2019.00053.

[77] Kevin Rossi and James Cumby. “Representations and descriptors unifying the study of molecular and bulk systems”. In: *International Journal of Quantum Chemistry* (Dec. 2019). DOI: 10.1002/qua.26151.

[78] Michael J Willatt, Félix Musil, and Michele Ceriotti. “Atom-density representations for machine learning”. In: *The Journal of Chemical Physics* 150 (2019), p. 154110. DOI: 10.1063/1.5090481.

[79] Haoyan Huo and Matthias Rupp. “Unified Representation of Molecules and Crystals for Machine Learning”. In: (Apr. 2017). DOI: arXiv:1704.06439.

[80] Michele Ceriotti, Michael J Willatt, and Gábor Csányi. "Machine Learning of Atomic-Scale Properties Based on Physical Principles". In: (). DOI: 10.1007/978-3-319-42913-7_68-1.

[81] Michele Ceriotti. *Unsupervised machine learning in atomistic simulations, between predictions and understanding*. Apr. 2019. DOI: 10.1063/1.5091842.

[82] Zhiming Li et al. "Metastable high-entropy dual-phase alloys overcome the strength-ductility trade-off". In: *Nature* 534.7606 (May 2016), pp. 227–230. DOI: 10.1038/nature17981.

[83] W.J. Gibbs. "A method of geometrical representation of the thermodynamic properties of substances by means of surfaces". In: *Transaction of the Connecticut Academy* (1873). DOI: https://www3.nd.edu/~powers/ame.20231/gibbs1873b.pdf.

[84] L. Galgani and A. Scotti. *Further remarks on convexity of thermodynamic functions*. Apr. 1969. DOI: 10.1016/0031-8914(69)90015-9.

[85] H. Greiner. "The chemical equilibrium problem for a multiphase system formulated as a convex program". In: *Calphad* 12.2 (1988), pp. 155–170. DOI: 10.1016/0364-5916(88)90017-X.

[86] C. C.R.S. Rossi, L. Cardozo-Filho, and R. Guirardello. "Gibbs free energy minimization for the calculation of chemical and phase equilibrium using linear programming". In: *Fluid Phase Equilibria* 278.1-2 (Apr. 2009), pp. 117–128. DOI: 10.1016/j.fluid.2009.01.007.

[87] L. G. Bullard and L. T. Biegler. "Iterated linear programming strategies for non-smooth simulation: A penalty based method for vapor-liquid equilibrium applications". In: *Computers and Chemical Engineering* 17.1 (1993), pp. 95–109. DOI: 10.1016/0098-1354(93)80007-A.

[88] Long X. Nghiem and Yau Kun Li. "Computation of multiphase equilibrium phenomena with an equation of state". In: *Fluid Phase Equilibria* 17.1 (1984), pp. 77–95. DOI: 10.1016/0378-3812(84)80013-8.

[89] H. Zhang. "A Review on Global Optimization Methods for Phase Equilibrium Modeling and Calculations". In: *The Open Thermodynamics Journal* 5.1 (Nov. 2011), pp. 71–92. DOI: 10.2174/1874396x01105010071.

[90] Gade Pandu Rangaiah. "Evaluation of genetic algorithms and simulated annealing for phase equilibrium and stability problems". In: *Fluid Phase Equilibria* 187-188 (Sept. 2001), pp. 83–109. DOI: 10.1016/S0378-3812(01)00528-3.

[91] D. Hildebrandt and D. Glasser. "Predicting phase and chemical equilibrium using the convex hull of the Gibbs free energy". In: *The Chemical Engineering Journal and The Biochemical Engineering Journal* 54.3 (1994), pp. 187–197. DOI: 10.1016/0923-0467(94)00202-9.

[92] C. Bradford Barber, David P. Dobkin, and Hannu Huhdanpaa. “The Quickhull Algorithm for Convex Hulls”. In: *ACM Transactions on Mathematical Software* 22.4 (1996), pp. 469–483. DOI: 10.1145/235815.235821.

[93] Shyue Ping Ong et al. “Li - Fe - P - O2 phase diagram from first principles calculations”. In: *Chemistry of Materials* 20.5 (Mar. 2008), pp. 1798–1807. DOI: 10.1021/cm702327g.

[94] Cong Liu et al. “Multiple superionic states in helium–water compounds”. In: *Nature Physics* 15.10 (Oct. 2019), pp. 1065–1070. DOI: 10.1038/s41567-019-0568-7.

[95] Maximilian Amsler et al. “Exploring the High-Pressure Materials Genome”. In: *Phys. Rev. X* (2018). DOI: 10.1103/PhysRevX.8.041021.

[96] Ion Errea et al. “High-pressure hydrogen sulfide from first principles: A strongly anharmonic phonon-mediated superconductor”. In: *Physical Review Letters* 114.15 (Apr. 2015). DOI: 10.1103/PhysRevLett.114.157004.

[97] I. Errea et al. “High-pressure hydrogen sulfide from first principles: a strongly anharmonic phonon-mediated superconductor”. In: *Physical Review Letters* 114 (2015), p. 157004. DOI: 10.1103/PhysRevLett.114.157004.

[98] A. P. Drozdov et al. “Conventional superconductivity at 203 kelvin at high pressures in the sulfur hydride system”. In: *Nature* 525 (2015), p. 73. DOI: 10.1038/nature14964.

[99] M. Mayo et al. “Ab Initio Study of Phosphorus Anodes for Lithium- and Sodium-Ion Batteries”. In: *Chemistry of Materials* 28 (2016), p. 2011. DOI: 10.1021/acs.chemmater.5b04208.

[100] C. J. Pickard, M. Martinez-Canales, and R. J. Needs. “Density functional theory study of phase IV of solid hydrogen”. In: *Physical Review B* 85 (2012), p. 214114. DOI: 10.1103/PhysRevB.85.214114.

[101] S. Azadi et al. “Dissociation of high-pressure solid molecular hydrogen: a quantum monte carlo and anharmonic vibrational study”. In: *Physical Review Letters* 112 (2014), p. 165501. DOI: 10.1103/PhysRevLett.112.165501.

[102] N. D. Drummond et al. “Quantum Monte Carlo study of the phase diagram of solid molecular hydrogen at extreme pressures”. In: *Nature Communications* 6 (2015), p. 7794.

[103] B. Monserrat et al. “Hexagonal structure of phase III of solid hydrogen”. In: *Physical Review B* 94 (2016), p. 134101. DOI: 10.1103/PhysRevB.94.134101.

[104] A. Pulido et al. “Functional materials discovery using energy–structure–function maps”. In: *Nature* 657 (2017), p. 543.

[105] P. Hohenberg and W. Kohn. “Inhomogeneous electron gas”. In: *Physical Review* 136 (1964), B864. DOI: 10.1103/PhysRev.136.B864.

[106] K. Fukunaga and D. R. Olsen. “An Algorithm for Finding Intrinsic Dimensionality of Data”. In: *IEEE Transactions on Computers* 20 (1971), p. 176. DOI: 10.1109/T-C.1971.223208.

[107] E. Wigner and H. B. Huntington. “On the possibility of a metallic modification of hydrogen”. In: *The Journal of Chemical Physics* 3.12 (Dec. 1935), pp. 764–770. DOI: 10.1063/1.1749590.

[108] N. W. Ashcroft. “Metallic Hydrogen: A High-Temperature Superconductor?” In: *Phys. Rev. Lett.* 21 (26 Dec. 1968), pp. 1748–1749. DOI: 10.1103/PhysRevLett.21.1748.

[109] R. J. Hemley and H. K. Mao. “Phase transition in solid molecular hydrogen at ultrahigh pressures”. In: *Physical Review Letters* 61.7 (Aug. 1988), pp. 857–860. DOI: 10.1103/PhysRevLett.61.857.

[110] T. W. Barbee, Marvin L. Cohen, and José Luís Martins. “Theory of high-pressure phases of hydrogen”. In: *Physical Review Letters* 62.10 (Mar. 1989), pp. 1150–1153. DOI: 10.1103/PhysRevLett.62.1150.

[111] Ho Kwang Mao and Russell J. Hemley. “Ultrahigh-pressure transitions in solid hydrogen”. In: *Reviews of Modern Physics* 66.2 (1994), pp. 671–692. DOI: 10.1103/RevModPhys.66.671.

[112] P. Loubeyre et al. “X-ray diffraction and equation of state of hydrogen at megabar pressures”. In: *Nature* 383.6602 (1996), pp. 702–704. DOI: 10.1038/383702a0.

[113] Chris J. Pickard and Richard J. Needs. “Structure of phase III of solid hydrogen”. In: *Nature Physics* 3.7 (July 2007), pp. 473–476. DOI: 10.1038/nphys625.

[114] Sam Azadi and Thomas D Kühne. “Unconventional phase III of high-pressure solid hydrogen”. In: *Physical Review B* 100 (2019), p. 155103. DOI: 10.1103/PhysRevB.100.155103.

[115] Ranga P. Dias and Isaac F. Silvera. “Observation of the Wigner-Huntington transition to metallic hydrogen”. In: *Science* 355.6326 (Feb. 2017), pp. 715–718. DOI: 10.1126/science.aal1579.

[116] Sam Azadi and W. M.C. Foulkes. “Fate of density functional theory in the study of high-pressure solid hydrogen”. In: *Physical Review B - Condensed Matter and Materials Physics* 88.1 (July 2013), p. 014115. DOI: 10.1103/PhysRevB.88.014115.

[117] C. J. Pickard, M. Martinez-Canales, and R. J. Needs. “Erratum: Density functional theory study of phase IV of solid hydrogen”. In: *Physical Review B* 86 (2012), p. 059902. DOI: 10.1103/PhysRevB.86.059902.

[118] J. P. Perdew, K. Burke, and M. Ernzerhof. “Generalized Gradient Approximation Made Simple”. In: *Physical Review Letters* 77 (1996), p. 3865. DOI: 10.1103/PhysRevLett.77.3865.

[119] M. I. Eremets and I. A. Troyan. “Conductive dense hydrogen”. In: *Nature Materials* 10 (2011), p. 927. DOI: 10.1038/nmat3175.

[120] J. M. McMahon and D. M. Ceperley. “Ground-State Structures of Atomic Metallic Hydrogen”. In: *Physical Review Letters* 106 (2011), p. 165302. DOI: 10.1103/PhysRevLett.106.165302.

[121] J. M. McMahon et al. “The properties of hydrogen and helium under extreme conditions”. In: *Reviews of Modern Physics* 84 (2012), p. 1607. DOI: 10.1103/RevModPhys.84.1607.

[122] M. I. Eremets, I. A. Troyan, and A. P. Drozdov. “Low temperature phase diagram of hydrogen at pressures up to 380 GPa. A possible metallic phase at 360 GPa and 200K”. In: (2016). DOI: arXiv:1601.04479.

[123] P. Dalladay-Simpson, R. T. Howie, and E. Gregoryanz. “Evidence for a new phase of dense hydrogen above 325 gigapascals”. In: *Nature* 529 (2016), p. 63. DOI: 10.1038/nature16164.

[124] A. Hermann, N. W. Ashcroft, and R. Hoffmann. “High pressure ices”. In: *Proceedings of the National Academy of Sciences* 109 (2012), p. 745. DOI: 10.1073/pnas.1118694109.

[125] G. C. DeFotis. “Magnetism of solid oxygen”. In: *Physical Review B* 23 (1981), p. 4714. DOI: 10.1103/PhysRevB.23.4714.

[126] M. C. van Hemert, P. E. S. Wormer, and A. van der Avoird. “*Ab Initio* Calculation of the Heisenberg Exchange Interaction between 0_2 Molecules”. In: *Physical Review Letters* 51 (1983), p. 1167. DOI: 10.1103/PhysRevLett.51.1167.

[127] R. Kitaura et al. “Formation of a One-Dimensional Array of Oxygen in a Microporous Metal-Organic Solid”. In: *Science* 298 (2002), p. 2358. DOI: 10.1126/science.1078481.

[128] Y. A. Freiman and H. J. Jodl. “Solid oxygen”. In: *Physics Reports* 401 (2004), p. 1. DOI: 10.1016/j.physrep.2004.06.002.

[129] T. Nomura et al. “Novel Phase of Solid Oxygen Induced by Ultrahigh Magnetic Fields”. In: *Physical Review Letters* 112 (2014), p. 247201. DOI: 10.1103/PhysRevLett.112.247201.

[130] P. Giannozzi et al. “QUANTUM ESPRESSO: a modular and open-source software project for quantum simulations of materials”. In: *Journal of Physics: Condensed Matter* 21 (2009), p. 395502.

[131] J. E. Campbell, J. Yang, and G. M. Day. “Predicted energy-structure-function maps for the evaluation of small molecule organic semiconductors”. In: *Journal of Materials Chemistry C* 5 (2017), p. 7574. DOI: 10.1039/C7TC02553J.

[132] C.G. De Kruif. “Enthalpies of sublimation and vapour pressures of 11 polycyclic hydrocarbons”. In: *Journal of Chemical Thermodynamics* 12 (1980), p. 243.

[133] V. Oja and E. M. Suuberg. “Vapor Pressures and Enthalpies of Sublimation of Polycyclic Aromatic Hydrocarbons and Their Derivatives”. In: *Journal of Chemical Engineering Data* 43 (1998), p. 486. DOI: 10.1021/je970222l.

[134] D. E. Williams. “Improved intermolecular force field for crystalline hydrocarbons containing four- or three-coordinated carbon”. In: *Journal of Molecular Structure* 485 (1999), p. 321. DOI: 10.1016/S0022-2860(99)00092-7.

[135] I. Giangreco, J. C. Cole, and E. Thomas. “Mining the Cambridge Structural Database for Matched Molecular Crystal Structures: A Systematic Exploration of Isostructurality”. In: *Cryst. Growth Des.* 17 (2017), p. 3192. DOI: 10.1021/acs.cgd.7b00155.

[136] Félix Musil et al. “Machine learning for the structure–energy–property landscapes of molecular crystals”. In: *Chemical Science* 9.5 (Jan. 2018), pp. 1289–1300. DOI: 10.1039/C7SC04665K.

[137] R. B. Campbell, J. M. Robertson, and J. Trotter. “The crystal and molecular structure of pentacene”. In: *Acta Crystallographica* 14 (1961), p. 705. DOI: 10.1107/S0365110X61002163.

[138] P. V. Hobbs. *Ice physics*. Oxford: Oxford University Press, 2010. DOI: 10.3189/S0022143000030847.

[139] V. F. Petrenko and R. W. Whitworth. *Physics of Ice*. Oxford: Oxford University Press, 1999. DOI: 10.1093/acprof:oso/9780198518945.001.0001.

[140] Sherwin J. Singer et al. “Hydrogen-Bond Topology and the Ice VII/VIII and Ice Ih/XI Proton-Ordering Phase Transitions”. In: *Physical Review Letters* 94 (2005), p. 135701.

[141] Chris Knight and Sherwin J. Singer. “Prediction of a Phase Transition to a Hydrogen Bond Ordered Form of Ice VI”. In: *Journal of Physical Chemistry B* 109 (2005), pp. 21040–21046. DOI: 10.1021/jp0540609.

[142] Jer-Lai Kuo. “The low-temperature proton-ordered phases of ice predicted by ab initio methods”. In: *Physical Chemistry Chemical Physics* 7 (2005), pp. 3733–3737. DOI: 10.1039/b508736h.

[143] Jer-Lai Kuo and Werner F. Kuhs. “A First Principles Study on the Structure of Ice-VI: Static Distortion, Molecular Geometry, and Proton Ordering”. In: *Journal of Physical Chemistry B* 110 (2006), pp. 3697–3703. DOI: 10.1021/jp055260n.

[144] Chris Knight and Sherwin J. Singer. “Hydrogen bond ordering in ice V and the transition to ice XIII”. In: *Journal of Chemical Physics* 129 (2008), p. 164513. DOI: 10.1063/1.2991297.

[145] G. A. Tribello, B. Slater, and C. G. Salzmann. “A Blind Structure Prediction of Ice XIV”. In: *Journal of the American Chemical Society* 128 (2006), pp. 12594–12595. DOI: 10.1021/ja0630902.

[146] John Russo, Flavio Romano, and Hajime Tanaka. “New metastable form of ice and its role in the homogeneous crystallization of water”. In: *Nature Materials* 13 (2014), pp. 733–739. DOI: 10.1038/nmat3977.

[147] Christopher J. Fennell and J. Daniel Gezelter. “Computational Free Energy Studies of a New Ice Polymorph Which Exhibits Greater Stability than Ice Ih”. In: *Journal of Chemical Theory and Computation* 1 (2005), pp. 662–667. DOI: 10.1021/ct050005s.

[148] Igor M. Svishchev and Peter G. Kusalik. “Quartzlike polymorph of ice”. In: *Physical Review B* 53 (1996), R8815. DOI: 10.1103/physrevb.53.r8815.

[149] Gareth A. Tribello et al. “Isomorphism between ice and silica”. In: *Physical Chemistry Chemical Physics* 12 (2010), pp. 8597–8606. DOI: 10.1039/b916367k.

[150] Y. Huang et al. “A new phase diagram of water under negative pressure: The rise of the lowest-density clathrate s-III”. In: *Science Advances* 2 (2016), e1501010. DOI: 10.1126/sciadv.1501010.

[151] C. Ji et al. “Two dimensional ice from first principles: Structures and phase transitions”. In: *Physical Review Letters* 116 (2016), p. 025501. DOI: 10.1103/PhysRevLett.116.025501.

[152] J. D. Bernal and R. H. Fowler. “A Theory of Water and Ionic Solution, with Particular Reference to Hydrogen and Hydroxyl Ions”. In: *Journal of Chemical Physics* 1 (1933), p. 515. DOI: 10.1063/1.1749327.

[153] L. Pauling. “The structure and entropy of ice and of other crystals with some randomness of atomic arrangement”. In: *Journal of the American Chemical Society* 57 (1935), pp. 2680–2684. DOI: 10.1021/ja01315a102.

[154] J. Hama and K. Suito. “Physics and Chemistry of Ice”. In: ed. by N. Maeno and T. Hondoh. Sapporo: Hokkaido University Press, 1992. DOI: 10.1039/9781847557773.

[155] C. J. Pickard, M. Martinez-Canales, and R. J. Needs. “Decomposition and Terapascal Phases of Water Ice”. In: *Physical Review Letters* 110 (2013), p. 245701. DOI: 10.1103/PhysRevLett.110.245701.

[156] L. del Rosso, M. Celli, and L. Ulivi. “New porous water ice metastable at atmospheric pressure obtained by emptying a hydrogen-filled ice”. In: *Nature Communications* 7 (2016), p. 13394. DOI: 10.1038/ncomms13394.

[157] G. Algara-Siller et al. “Square ice in graphene nanocapillaries”. In: *Nature* 519 (2015), p. 443. DOI: 10.1038/nature14295.

[158] C. Ji et al. “Double-layer ice from first principles”. In: *Physical Review B* 95 (2017), p. 094121. DOI: 10.1103/PhysRevB.95.094121.

[159] Amir Haji-Akbari and Pablo G. Debenedetti. “Direct calculation of ice homogeneous nucleation rate for a molecular model of water”. In: *Proceedings of the National Academy of Sciences of the USA* 112 (2015), pp. 10582–10588.

[160] D. Quigley. “Communication: Thermodynamics of stacking disorder in ice nuclei”. In: *Journal of Chemical Physics* 141 (2014), p. 121101. DOI: 10.1063/1.4900772.

[161] David Quigley, Dario Alfé, and Ben Slater. “Communication: On the stability of ice 0, ice i, and Ih”. In: *Journal of Chemical Physics* 141 (2014), p. 161102.

[162] C. J. Pickard and R. J. Needs. “*Ab initio* random structure searching”. In: *Journal of Physics: Condensed Matter* 23 (2011), p. 053201.

[163] G. A. Tribello and B. Slater. “A theoretical examination of known and hypothetical clathrate hydrate materials”. In: *Journal of Chemical Physics* 131 (2009), p. 024703. DOI: 10.1063/1.3142503.

[164] Allen et al. “Forward flux sampling-type schemes for simulating rare events: Efficiency analysis”. In: *J. Chem. Phys.* 124 (2006), p. 024102. DOI: 10.1063/1.2198827.

[165] Li et al. “Ice nucleation at the nanoscale probes no man’s land of water”. In: *Nat. Comm.* 4 (2013), p. 1887. DOI: 10.1038/ncomms2918.

[166] Haji-Akbari and Debenedetti. “Direct calculation of ice homogeneous nucleation rate for a molecular model of water”. In: *PNAS* 112 (2015), p. 10582. DOI: 10.1073/pnas.1509267112.

[167] Bi et al. “Enhanced heterogeneous ice nucleation by special surface geometry”. In: *Nat. Comm.* 8 (2017), p. 15372. DOI: 10.1038/ncomms15372.

[168] Quigley and Rodger. “A metadynamics-based approach to sampling crystallisation events”. In: *Molec. Sim.* 35 (2009), p. 613. DOI: 10.1080/08927020802647280.

[169] Giberti et al. “Metadynamics studies of crystal nucleation”. In: *IUCrJ* 2 (2015), p. 256. DOI: 10.1107/S2052252514027626.

[170] S. Pipolo et al. “Navigating at Will on the Water Phase Diagram”. In: *Physical Review Letters* 119 (2017), p. 245701. DOI: 10.1103/PhysRevLett.119.245701.

[171] M. von Stackelberg and H. Müller. “Zur Struktur der Gashydrate”. In: *Naturwissenschaften* 38 (1951), p. 456. DOI: 10.1007/BF00641166.

[172] W. Claussen. “Suggested structures of water in inert gas hydrates”. In: *Journal of Chemical Physics* 19 (1951), p. 259. DOI: 10.1063/1.1748187.

[173] L. Pauling and R. Marsh. “The structure of chlorine hydrate”. In: *Proceedings of the National Academy of Sciences of the USA* 38 (1952), p. 112. DOI: /10.1073/pnas.38.2.112.

[174] G. A. Jeffrey. “Inclusion Compounds 1”. In: *Hydrate Inclusion Compounds*. Ed. by J. L. Atwood, J. E. Davies, and D. D. MacNicol. New York: Academic Press, 1984. DOI: 10.1007/BF00656757.

[175] H. Gies, H. Gerke, and F. Liebau. “Chemical composition and synthesis of melanophlogite, a clathrate compound of silica”. In: *N. Jb. Miner. Mh.* (1982), pp. 119–124. DOI: 10.1088/0953-8984/26/10/10320.

[176] J. L. Schlenker et al. “Crystal structure of a synthetic high silica zeolite - ZSM-39”. In: *Nature* 294 (1981), pp. 340–342. DOI: 10.1038/294340a0.

[177] J. V. Smith and C. S. Blackwell. “Nuclear magnetic resonance of silica polymorphs”. In: *Nature* 303 (1983), pp. 223–225. DOI: 10.1038/303223a0.

[178] H. Gies. “Studies of Clathrasils. VI. Crystal structure of dodecasil 3C, another synthetic clathrate compound of silica”. In: *Z. Kristallogr.* 167 (1984), pp. 73–82. DOI: 10.1524/zkri.1984.167.14.73.

[179] Y. Long et al. “Single crystal growth, morphology, and structure of ZSM-39 and its variation CF4”. In: *J. Incl. Phenom.* 5 (1987), pp. 355–362. DOI: 10.1007/BF00665368.

[180] J. Emmer and M. Wiebcke. “Heteronetwork clathrates with three-dimensional mixed silicate-water host frameworks and channel systems”. In: *Journal of the Chemical Society, Chemical Communications* (1994), p. 2079. DOI: 10.1039/C39940002079.

[181] M. Wiebcke. “Structural links between zeolite-type and clathrate hydrate-type materials”. In: *Journal of the Chemical Society, Chemical Communications* (1991), pp. 1507–1508. DOI: 10.1016/0927-6513(93)80062-Y.

[182] C. Baerlocher, W. M. Meier, and D. H. Olson. *Atlas of Zeolite Framework Types*. Amsterdam: Elsevier, 2007. DOI: 10.1002/aic.690350523.

[183] M. M. J. Treacy et al. "Enumeration of periodic tetrahedral frameworks. II. Polynodal graphs". In: *Microporous Mesoporous Materials* 74 (2004), p. 121. DOI: 10.1524/zkri.1997.212.11.768.

[184] D. J. Earl and M. W. Deem. "Toward a Database of Hypothetical Zeolite Structures". In: *Ind. Eng. Chem. Res.* 45 (2006), p. 5449. DOI: 10.1021/ie0510728.

[185] B. Winkler et al. "Systematic prediction of crystal structures". In: *Chemical Physics Letters* 337 (2001), p. 36. DOI: 10.1016/S0009-2614(01)00126-9.

[186] B. Winkler et al. "Prediction of nanoporous sp^2-carbon framework structure by combining graph theory with quantum mechanics". In: *Chemical Physics Letters* 312 (1999), p. 536. DOI: 10.1016/S0009-2614(99)00943-4.

[187] R. T. Strong et al. "Systematic prediction of crystal structures: An application to sp^3-hybridized carbon polymorphs". In: *Physical Review B* 70 (2004), p. 045101. DOI: 10.1103/PhysRevB.70.045101.

[188] I. A. Baburin et al. "From zeolite nets to sp^3 carbon allotropes: a topology-based multiscale theoretical study". In: *Physical Chemistry Chemical Physics* 17 (2015), p. 1332. DOI: 10.1039/c4cp04569f.

[189] T. Matsui et al. "Hypothetical ultralow-density ice polymorphs". In: *Journal of Chemical Physics* 147 (2017), p. 091101. DOI: 10.1063/1.4994757.

[190] C. Baerlocher and L.B. McCusker. *Database of Zeolite Structures*. http://www.iza-structure.org/databases/. Accessed: 28/06/2017. DOI: 10.1016/j.micromeso.2020.110000.

[191] E. A. Engel, B. Monserrat, and R. J. Needs. "Anharmonic nuclear motion and the relative stability of hexagonal and cubic ice". In: *Phys. Rev. X* 5 (2015), p. 021033. DOI: 10.1103/PhysRevX.5.021033.

[192] A. C. T. van Duin et al. "ReaxFF: A Reactive Force Field for Hydrocarbons". In: *Journal of Physical Chemistry A* 105 (2001), p. 9396. DOI: 10.1021/jp004368u.

[193] D. Raymand et al. "Water adsorption on stepped ZnO surfaces from MD simulation". In: *Surface Science* 604 (2010), p. 741. DOI: 10.1016/j.susc.2009.12.012.

[194] L. del Rosso et al. "Refined Structure of Metastable Ice XVII from Neutron Diffraction Measurements". In: *Journal of Physical Chemistry C* 120 (2016), p. 26955. DOI: 10.1021/acs.jpcc.6b10569.

[195] Mariana Rossi, Piero Gasparotto, and Michele Ceriotti. "Anharmonic and Quantum Fluctuations in Molecular Crystals: A First-Principles Study of the Stability of Paracetamol". In: *Phys. Rev. Lett.* 117 (2016), p. 115702. DOI: 10.1103/PhysRevLett.117.115702.

[196] W. Sun et al. "The thermodynamic scale of inorganic crystalline metastability". In: *Science Advances* 2 (2016), e1600225–e1600225. DOI: 10.1126/sciadv.1600225.

[197] Y. Liu and L. Ojamäe. "Clathrate ice sL: a new crystalline phase of ice with ultralow density predicted by first-principles phase diagram computations". In: *Physical Chemistry Chemical Physics* 20 (2018), p. 8333. DOI: 10.1039/C8CP00699G.

[198] B. Santra et al. "On the Accuracy of van der Waals Inclusive Density-Functional Theory Exchange-Correlation Functionals for Ice at Ambient and High Pressures". In: *Journal of Chemical Physics* 139 (2013), p. 154702. DOI: 10.1063/1.4824481.

[199] K. Lee et al. "Higher-accuracy van der Waals density functional". In: *Physical Review B* 82 (2010), 081101(R). DOI: 10.1103/PhysRevB.82.081101.

[200] J. Sun, A. Ruzsinszky, and J. P. Perdew. "Strongly Constrained and Appropriately Normed Semilocal Density Functional". In: *Physical Review Letters* 115 (2015), p. 036402. DOI: 10.1103/PhysRevLett.115.036402.

[201] J. Sun et al. "Accurate first-principles structures and energies of diversely bonded systems from an efficient density functional". In: *Nature Chemistry* 8 (2016), p. 831. DOI: 10.1038/nchem.2535.

[202] S. Grimme. "Semiempirical GGA-type density functional constructed with a long-range dispersion correction". In: *Journal of Computational Chemistry* 27 (2006), p. 1787. DOI: 10.1002/jcc.20495.

[203] J. P. K. Doye, D. J. Wales, and M. A. Miller. "Thermodynamics and the global optimization of Lennard-Jones clusters". In: *J. Chem. Phys.* 109 (1998), p. 8143. DOI: 10.1063/1.477477.

[204] J. P. K. Doye and C. P. Massen. "Characterizing the network topology of the energy landscapes of atomic clusters". In: *J. Chem. Phys.* 122 (2005), p. 084105. DOI: 10.1063/1.1850468.

[205] M. M. J. Treacy et al. "Enumeration of periodic tetrahedral frameworks". In: *Zeitschrift für Kristallographie* 212 (1997), p. 728. DOI: 10.1016/j.micromeso.2004.06.013.

[206] O. D. Friedrichs et al. "Systematic enumeration of crystalline networks". In: *Nature* 400 (1999), p. 644. DOI: 10.1038/nmat1090.

[207] Kumar et al. "The weighted histogram analysis method for free-energy calculations on biomolecules. I. The method". In: *J. Comp. Chem.* 13 (1992), p. 1011. DOI: 10.1002/jcc.540130812.

[208] A. Laio and M. Parrinello. "Escaping free energy minima". In: *Proc. Natl. Acad. Sci. USA* 99 (2002), p. 12562. DOI: 10.1073/pnas.202427399.

Andrea ANELLI

Personal Data

PLACE AND DATE OF BIRTH:	Rome, Italy | 01 September 1990
ADDRESS:	Avenue de l'Eglise-Anglaise 18, 1006 Lausanne, Switzerland
PHONE:	0041 78 6495351
EMAIL:	andrea.anelli@epfl.ch
GITHUB:	https://github.com/andreanelli

Work Experience

Apr 2015 - Oct 2015	Visiting graduate student in GROSSMAN GROUP, MA, USA *Massachusetts Institute of Technology* Molecular Dynamics simulations on water and ethanol system to investigate nanofiltration through a porous graphene membrane
Reference	Prof. Jeffrey C. GROSSMAN jcg@mit.edu
Sep 2014 - Feb 2015	Visiting graduate student in ENDOHLAB, Sendai, Japan *Tohoku University* Ab initio simulations on impurity placement effects on the electron transmission of a Vertical MOSFET
Reference	Prof. Tetsuo ENDOH tendoh@cir.tohoku.ac.jp

Education

2016 - 2020 PhD in Materials Science
École polytechnique fédérale de Lausanne, Lausanne, Switzerland
Thesis Characterising Structure and Stability
of Materials using Machine Learning
| Advisor: Prof. Michele CERIOTTI

2013-2015 Master of Science in NANOTECHNOLOGY APPLIED TO ICTS
Politecnico di Torino, Turin, Italy
Thesis Separation of ethanol and water using a nanoporous
graphene membrane under Reverse Osmosis
| Advisor: Prof. Giancarlo CICERO
FINAL MARK: 110 cum laude

2009-2013 Bachelor of Science in BIOMEDICAL ENGINEERING
Università Campus Biomedico, Rome, Italy
Thesis Optimal resource allocation under dynamic programming
| Advisor: Prof. Marco PAPI

Fellowships and Leadership

SEPT. 2014 JASSO scholarship for International researchers (Japan) (¥500,000)
SEPT. 2016 Best graduate student with UpToYou by Fondazione Agnelli (EUR 10,000)
MAY 2018 President of EDMX|Social organisation at EPFL

Languages

ENGLISH: Proficient - ITALIAN: Mothertongue

Skills

Atomistic Modelling : Quantum Espresso, Dalton, DFTB+, i-Pi, LAMMPS,Gromacs
Software Skills : FORTRAN, C++, Julia, Python, Bash, Perl, Go, git, OpenCL,CUDA
Data Science : scikit-learn, pyTorch, TensorFlow, R
Graphics Modeling : Adobe Suite, Blender, Cinema4D
Web Development : Javascript, HTML, CSS

Interests and Activities

Music production, Traveling, Cooking, Photography, Painting, Visual arts

Publications

- Automatic Selection of Atomic Fingerprints and Reference Configurations for Machine-Learning Potentials
  The Journal of Chemical Physics 148, 241730
  G Imbalzano, **A Anelli**, D Giofré, S Klees, J Behler, M Ceriotti
  DOI : https://doi.org/10.1063/1.5024611

- Mapping uncharted territory in ice from zeolite networks to ice structures
  Nature Communications volume 9, Article number: 2173 (2018)
  EA Engel, **A Anelli**, CJ Pickard, RJ Needs , M Ceriotti
  DOI : https://doi.org/10.1038/s41467-018-04618-6

- A Generalized Convex Hull Construction for Materials Discovery
  Physical Review Materials 2, 103804
  A Anelli, EA Engel, CJ Pickard, M Ceriotti
  DOI : https://doi.org/10.1103/PhysRevMaterials.2.103804

- A Bayesian approach to NMR crystal structure determination
  Phys. Chem. Chem. Phys., 2019,21, 23385-23400
  EA Engel, **A Anelli**, A Hofstetter, F Paruzzo, L Emsley and M Ceriotti
  DOI : https://doi.org/10.1039/C9CP04489B