Thèse n° 7997

A general and efficient framework for atomistic machine learning

Présentée le 30 avril 2021

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

Félix Benedito Clément MUSIL

Acceptée sur proposition du jury

Prof. K. Scrivener, présidente du jury
Prof. M. Ceriotti, directeur de thèse
Prof. C. Clementi, rapporteuse
Prof. G. Csanyi, rapporteur
Prof. L. Zdeborova, rapporteuse

École
polytechnique
fédérale
de Lausanne

2021

Puisque ces mystères nous dépassent, feignons d'en être l'organisateur…

—Jean Cocteau, *Les Mariés de la tour Eiffel*

Acknowledgements

I would like to first thank Prof. Michele Ceriotti, my thesis director, for supporting me up to the completion of this important chapter of my life. Under his supervision, Michele has shown me how to carry out impactful research projects, allowing me to develop the many skills required to tackle challenging scientific problems. Thanks to him, not only have I been able to contribute to an exciting research field, but I have also been given the opportunities to meet and learn from outstanding scientists by attending international schools, workshops, and conference. Moreover his enthusiasm for research has been a great motivation for pushing myself further throughout my graduate work.

In addition to Michele's mentorship, I benefitted greatly from my collaboration with a few exceptional researchers within the group, so I want to dedicate a special thank to Dr. Sandip De, Dr. Micheal Willat, and Dr. Max Veit. As a results of our close interactions, I have been able to transform ambitious project into exciting scientific achievements. Furthermore, their example provided me with the opportunity to develop some of the grit necessary to bear with the uncertainties and difficulties tied with these exploratory endeavour.

I feel very lucky to have found in the laboratory of computational science and modelling (COSMO) a warm and cheerful academic family. So I want to thank all of the former and current members of the group for building a friendly atmosphere sparking many enjoyable discussions.

An important part of my experiences in COSMO have been made remarkable by our numerous interactions within and outside of the 3^{rd} floor of the MXG building around a coffee and/or a drink. In particular, I would like to thank Venkat, Edgar, Max, Chiheb, and Alexander for transforming the unreliability of our coffee machines into an opportunity to both deepen my coffee making skills and share fun and relaxing break times. A special thanks to Natasha, Edoardo, Alexander, Chiheb, Jigyasa, Kevin, and Piero, for their advice, support, and interactions outside of the group.

I would like to express my deepest gratitude to Venkat, Giulio, and Andrea. I feel very fortunate that we could share most of our Ph.D. time together and I thank you for bearing with me so far.

Last but not least, I would like to thank my parents, family, and non-academic friends for their invaluable support and constant source of positivity and strength.

Merci encore à tous pour ces bons moment !

Lausanne, 30^{th} of March 2021 F. M.

Abstract

Over the last two decades, many technological and scientific discoveries, ranging from the development of materials for energy conversion and storage through the design of new drugs, have been accelerated by the use of preliminary *in silico* experiments, to steer and inform synthesis and characterization. This new computational paradigm has been particularly significant for simulations taking place at the atomic scale, which provide a predictive framework to determine the properties of condensed phases and molecular systems from first principles. Thanks to the steady improvement in accuracy and efficiency of *ab initio* methods, as well as to the increase in the performance (and reduction in the cost) of computational resources, once-prohibitive quantum mechanical calculations of atomic-scale properties have become affordable and ubiquitous. The rise of *ab initio* and high-throughput materials design and discovery, however, brings both challenges and opportunities.

Large repositories of atomistic data require complicated, time-consuming analyses to rationalize the relationship between the structure and the properties, and to determine the most promising candidates for a given application. Oftentimes - for instance when considering molecular dynamics simulations that sample the finite-temperature fluctuations of materials in realistic thermodynamic conditions - first-principle calculations contain large amounts of redundant data, for which a direct *ab initio* treatment is still prohibitively expensive. The availability of large amounts of data, and the fact that many applications require to sample repeatedly configurations that share considerable similarities, provide the ideal scenario to leverage statistical learning techniques. Machine-learning potentials (and more generally, atomistic property models) trained on a small number of reference quantum calculations accelerate by orders of magnitude the prediction of the stability and behavior of similar materials and molecules, while unsupervised (or semi-supervised) analyses automate the process of mining large computational databases for materials with improved performances, and for insights on the physical processes that determine their outstanding properties.

This thesis presents several methodological advances to the representation of condensed phase matter at the atomic scale to develop data-driven atomistic models. We present an atom density framework to build n-body representations encoding the chemical structure along with the fundamental symmetries of such systems and draw links between several popular frameworks. This formulation provides both a unifying picture of density-based representations and recipes to extract symmetry-adapted features from atomistic systems, one of the key factors for the successful application of – both supervised and unsupervised – machine learning algorithms. Building on this framework, we used a 3-body representation

to explore large databases of small peptides and molecular crystals using clustering and dimensionality reduction, unsupervised learning techniques, through maps of their structural correlations. These simple overviews of entire datasets allowed us to highlight structure-property relations and to check for their consistency and reliability. Thanks to the generality of this representation we also applied supervised learning to construct surrogate models of several quantum properties such as the chemical shifts in molecular materials and the stability of molecular materials, small molecules, and perovskites. We further improve the quality of these models by introducing property and system-specific knowledge into the representation to increase its correlation with the target properties. Such optimization of the representation helps reducing the error of model predictions, but being able to estimate the accuracy of these predictions is just as useful. To simplify computing uncertainty estimates for the predicted properties, we provided simple schemes to calibrate them and assess their accuracy thus increasing the reliability of data-driven models of materials.

The success of the supervised and unsupervised learning applications we presented within the atom density representation framework highlights the value for the atomic scale modeling toolbox in integrating machine learning algorithms to automate analyses and accelerate property predictions. This framework has already lead to several extensions – both on the representation and on the modeling strategy – and we expect it to be the cornerstone for the development of new knowledge-based computational materials methods.

Key words: Materials modelling, Machine learning, Atomistic Computer Simulation, Density-based representations, Kernel methods, New materials discovery, High-throughput screening, DFT, crystal structure prediction, Molecular Materials, Structure-property relationship

List of Publications directly related to this thesis

[1] Sandip De, Felix Musil, Teresa Ingram, Carsten Baldauf, and Michele Ceriotti. “Mapping and Classifying Molecules from a High-Throughput Structural Database”. In: *J. Cheminformatics* 9.1 (Dec. 2017), pp. 1–14. DOI: `10.1186/s13321-017-0192-4`.

[2] Félix Musil, Sandip De, Jack Yang, Joshua E. Campbell, Graeme M. Day, and Michele Ceriotti. “Machine Learning for the Structure-Energy-Property Landscapes of Molecular Crystals”. In: *Chem. Sci.* 9.5 (2018), pp. 1289–1300. DOI: `10.1039/c7sc04665k`.

[3] Federico M. Paruzzo, Albert Hofstetter, Félix Musil, Sandip De, Michele Ceriotti, and Lyndon Emsley. “Chemical shifts in molecular solids by machine learning”. In: *Nature Communications* 9.1 (Dec. 2018), p. 4501. DOI: `10.1038/s41467-018-06972-x`.

[4] Michael J. Willatt, Félix Musil, and Michele Ceriotti. “Feature Optimization for Atomistic Machine Learning Yields a Data-Driven Construction of the Periodic Table of the Elements”. In: *Phys. Chem. Chem. Phys.* 20.47 (2018), pp. 29661–29668. DOI: `10.1039/c8cp05921g`.

[5] Michael J. Willatt, Félix Musil, and Michele Ceriotti. “Atom-Density Representations for Machine Learning”. In: *J. Chem. Phys.* 150.15 (Apr. 2019), p. 154110. DOI: `10.1063/1.5090481`.

[6] Félix Musil, Michael J. Willatt, Mikhail A. Langovoy, and Michele Ceriotti. “Fast and Accurate Uncertainty Estimation in Chemical Machine Learning”. In: *Journal of Chemical Theory and Computation* 15.2 (Feb. 2019), pp. 906–915. DOI: `10.1021/acs.jctc.8b00959`.

[7] Félix Musil and Michele Ceriotti. “Machine learning at the atomic scale”. In: *Chimia* 73.12 (Dec. 2019), pp. 972–982. DOI: `10.2533/chimia.2019.972`.

[8] Tim Würger, Christian Feiler, Félix Musil, Gregor B.V. Feldbauer, Daniel Höche, Sviatlana V. Lamaka, Mikhail L. Zheludkevich, and Robert H. Meißner. “Data science based Mg corrosion engineering”. In: *Frontiers in Materials* 6 (Apr. 2019), pp. 1–9. DOI: `10.3389/fmats.2019.00053`.

Résumé

Au cours des deux dernières décennies, de nombreuses découvertes technologiques et scientifiques, allant du développement de matériaux pour la conversion et le stockage énergétique à la conception de nouveaux médicaments, ont été accélérées par l'utilisation d'expériences "numériques" pour orienter leur synthèse et leur caractérisation. Ce nouveau paradigme a été particulièrement important pour les simulations à l'échelle atomique, qui fournissent un cadre théorique solide pour prédire les propriétés de la matière condensée. Grâce à l'amélioration constante de la précision et de l'efficacité des méthodes *ab initio*, ainsi qu'à l'augmentation des performances des ordinateurs, les calculs précédemment prohibitifs des propriétés à l'échelle atomique sont devenus courants. L'essor de la conception et de la découverte de matériaux à haut débit utilisant des méthodes *ab initio*, cependant, apporte à la fois des défis et des opportunités.

Les grands dépôts de données atomiques nécessitent des analyses compliquées et longues afin de rationaliser la relation entre la structure et les propriétés, et de déterminer les candidats les plus prometteurs pour une application donnée. Souvent - par exemple les simulations de dynamique moléculaire qui échantillonnent les fluctuations à température finie des matériaux dans des conditions thermodynamiques réalistes - les interactions entre les atomes sont redondantes et les traiter directement avec une méthode *ab initio* sont encore prohibitif. La disponibilité de grandes quantités de données, et le fait que de nombreuses applications nécessitent d'échantillonner de façon répétée des configurations qui partagent des similitudes considérables, fournit le scénario idéal pour tirer parti des techniques d'apprentissage statistique. D'une part, les potentiels interatomiques (et plus généralement, les modèles de propriétés atomiques) entrainés avec des techniques d'apprentissage automatique sur un petit nombre de références *ab initio* accélèrent grandement la prédiction des propriétés de matériaux et de molécules similaires D'autre part, l'apprentissage non supervisé (ou semi-supervisé) permet d'automatiser l'exploration de grandes bases de données contenant des références *ab initio* afin de trouver des matériaux aux performances améliorées, et d'obtenir des informations sur les processus physiques qui déterminent leurs propriétés exceptionnelles.

Cette thèse présente un certain nombre d'avancées méthodologiques sur représentation de la matière condensée à l'échelle atomique dans le but de développer des modèles guidés par les données. Nous présentons un cadre théorique pour représenter les structures atomiques à travers les corrélations structurelles. De cette manière, la structure chimique ainsi que les symétries fondamentales de tels systèmes sont encodés succinctement, un des facteurs clés

pour l'application réussie des algorithmes d'apprentissage automatique. De plus, cette formulation fournit une image unifiée des représentations basées sur les densités atomiques. Sur la base de ce cadre, nous avons utilisé une représentation à trois corps pour explorer des bases de données de petits peptides et de cristaux moléculaires en utilisant des techniques de partitionnement et de réduction de la dimensionnalité, dit apprentissage non supervisé, en établissant des visualisations simples de leurs corrélations structurelles. Ces aperçus d'ensembles de ces bases de données nous ont permis de vérifier leur cohérence et leur fiabilité et de mettre en évidence des relations structures-propriétés pour ces systèmes. Grâce à la généralité de cette représentation, nous avons également appliqué des techniques d'apprentissage supervisé pour construire des modèles de substitution pour plusieurs propriétés quantiques telles que les déplacements chimiques dans des solides moléculaires et la stabilité des matériaux moléculaires, des petites molécules et des pérovskites. La qualité de ces modèles est améliorés davantage en introduisant des connaissances spécifiques des propriétés cible dans la représentation des structures atomiques afin d'augmenter leur corrélation réciproque. Une telle optimisation de la représentation permet de réduire l'erreur des prédictions du modèle, mais il est tout aussi utile de pouvoir estimer la précision de ces prédictions. Pour simplifier le calcul de l'estimation de l'incertitude associée avec les propriétés prédites, nous avons proposé des procédures simples pour la calibrer et évaluer sa précision, augmentant ainsi la fiabilité des modèles de matériaux guidés par les données.

Le succès des applications des algorithms d'apprentissage (non) supervisé que nous avons présentées dans le cadre de la représentation des systèmes atomiques par densité atomique souligne les possibilités d'amélioration qu'offrent les algorithmes d'apprentissage automatique aux méthodes classiques de modélisation des matériaux, que ce soit pour automatiser des analyses de leurs simulations ou pour accélérer la prédiction de leurs propriétés. Ce cadre théorique a déjà donné lieu à plusieurs extensions – tant sur le plan de la représentation que celui de la stratégie de modélisation – et nous pensons qu'il sera une pierre angulaire pour développement de nouvelles méthodes de modélisation des matériaux basées sur les données.

Contents

Acknowledgements i

Abstract (English/French) iii

1 Machine learning for atomic-scale modelling 1

2 Theory of atomic scale representations 7

2.1 A Dirac notation for atomic configurations and environments 8

2.1.1 Density-based structural representations 9

2.1.2 Symmetry-invariant representations 10

2.1.3 Tensor-product representations 11

2.1.4 Atom-centered representations 12

2.1.5 Rotationally-invariant representations 13

2.2 A unified picture of density-based representations 14

2.2.1 Plane waves 15

2.2.2 Many-body kernels and representations 15

2.2.3 Behler-Parrinello symmetry functions 17

2.2.4 Smooth Overlap of Atomic Positions 18

2.2.5 Tensorial Smooth Overlap of Atomic Positions (λ-SOAP) 19

2.2.6 Distributions vs sorted vectors 20

2.2.7 Use of density-based features in artificial neural networks 21

2.3 Generalized invariant density representations 23
2.3.1 Low-rank expansion of the $\hat{U}$ operator 24
2.3.2 Radially-scaled representations 24
2.3.3 Alchemical kernels 25
2.3.4 Multiple-kernel learning 27
2.3.5 Non-factorizable operators 27
2.3.6 Optimization of the density representation 29

3 Unsupervised ML **31**
3.1 REMatch Kernel and Distances 32
3.2 Sketch-map 33
3.3 Hierarchical clustering and HDBSCAN* 34
3.4 Mapping and classifying molecules from a high-throughput structural database 36
3.4.1 Finding the Dominant Features of a Structual Landscape 38
3.4.2 Understanding the Impact of Perturbations on conformational Space . . 43
3.4.3 Identifying Outliers and Checking for Consistency 44
3.5 The structure–energy–property landscapes of molecular crystals 46
3.5.1 A Benchmark Database of organic semiconductors 47
3.5.2 Results & Discussion 48

4 Accurate and reliable supervised ML **53**
4.1 Supervised Learning for atomistic modelling 54
4.1.1 GPR model 55
4.1.2 Approximate GPR model 56
4.1.3 Model ranking and optimization 56
4.2 Predictive models for molecular solids 57
4.2.1 Chemical shifts in molecular solids by machine learning 58

4.2.2 Property predictions for molecular crystals 73
4.3 Feature optimization for atomistic machine learning yields a data-driven construction of the periodic table of the elements 83
4.3.1 Benchmark Datasets 84
4.3.2 Reduced-dimensionality alchemical kernels 85
4.3.3 A data-driven periodic table of the elements 88
4.3.4 Radial scaling in the QM9 data set 89
4.4 Fast and Accurate Uncertainty Estimation in Chemical Machine Learning . . . 92
4.4.1 Resampling (RS) 92
4.4.2 Benchmark Datasets 97
4.4.3 Results and Discussion 98

5 Conclusions **107**

Appendices **111**

A Details on Spherical Invariants **113**
A.1 Angular basis 113
A.1.1 Spherical harmonics 113
A.1.2 Finite rotations 114
A.2 Spherical invariants of order $\nu = 1,2,3$ 115
A.3 Atom density expansions 117
A.3.1 Angular integration 117
A.3.2 Radial integration 118
A.4 Gradients of the atom density expansions 121
A.4.1 $\nabla_k Y_l^m(\hat{\mathbf{r}}_{ij})$ 122
A.4.2 $\nabla_k D_{nl}(r_{ij})$ 123

List of figures **125**

List of tables 135

Bibliography 137

1 Machine learning for atomic-scale modelling

The advent of new materials has been at the core of many deep societal changes over the last century, from widely available commercial flights to computers thanks to the use of super-alloys and silicon-based transistors. New materials have been typically discovered through labor-intensive searches. For instance, Edison tested about six thousand organic compounds to develop the filament of his long lasting light bulb while the formulation of a stable and cheap catalyst for the Haber-Bosch process required testing more than twenty thousands metallic candidates.

Following such protocol to find alternative materials to build solar cells or batteries to improve their efficiency would represent a prohibitive investment. As a result, computational materials methods have grown in the past two decades to be an essential guide for experimental searches and developments of new materials, providing atomistic insights to complex phenomena, pre-screening of hypothetical materials. The emergence of computer simulations to predict, *in silico*, the stability and the properties of hypothetical materials owes to the steady improvement of the modelling methods, computational resources, and in the efficiency of their implementations. Some recent achievements include highlighting the stabilization of high-temperature super-conductors by nuclear quantum effects[1] and the details of the deformation processes in tantalum metal with *ab initio* molecular dynamics,[2] finding promising 2D semiconductors for the next generation of transistors[3] and proposing alternative materials to extend the lifetime of Li-ion batteries.[4]

The most accurate description of atomic systems is provided by Schrödinger's equation but solving it numerically for all the electrons and nuclei system is impractical for more than a few atoms. Over the last fifty years a hierarchy of approximate quantum mechanical theories have been developed to predict their properties, e.g. formation energy, electronic band structure, NMR chemical shifts, and dipole-moment, from first principles.[5–9] These methods include by order of increasing accuracy Density Functional Theory (DFT), Møller-Plesset (MP) perturbation theory and Coupled-Clusters with singles and doubles (CCSD) methods and their typical scaling are respectively $\mathcal{O}(n^3)$, $\mathcal{O}(n^5)$ and $\mathcal{O}(n^6)$ where n is the number of electrons in the system. Therefore the length and time scales they can model, typically of the order of the

nm and ps with DFT, are often too small to directly study the effect of grain boundaries on the strength of alloys or the binding of a molecule to a protein, to mention only a few examples. Nevertheless, they provide the first step upon which multi-scale modelling techniques are built.[10] One of the most prominent examples of such procedure is the parametrization of empirical force fields (FF) using *ab initio* data where the electronic degrees of freedom are approximately incorporated into internal parameters of the model. This drastic simplification of the interactions within the atomic system coupled with fixed functional forms leaves the development of accurate and transferable reactive, multi-component FF as a major challenge.

Thermodynamic properties such as the phase diagram of a condensed phase system or the Raman spectra of a solvated molecule can be accurately estimated by sampling a statistical ensemble using Markov Chain Monte Carlo (MCMC) or Molecular Dynamics (MD) in the thermodynamic limit, i.e. for a large number of atoms, in conjunction with electronic structure methods.[11,12] Besides these generally applicable techniques, the atomic scale modelling toolbox also includes more targeted protocols. Among the most widely used design approaches figure Crystal Structure Prediction (CSP) to elucidate the polymorphism of molecular materials,[13,14] computer-aided drug design to accelerate drug discovery and development[15] and the combinatorial evaluation of materials properties,[16] e.g. optimize the composition of the perovskites structure for energy conversion.[17] Considering the large amount of reference data produced by such methods, several community efforts have emerged over the past few years[18–26] that aim at generating, and/or storing large amounts of simulation data in publicly available databases. The development of these repositories of structural data along with their associated materials properties (e.g. formation energy, band gap, polarizability, ...) and more generally the widespread availability of atomistic data enables the use of data-driven approaches to accelerate discoveries in the field of computational materials.[20,27,28]

In the past few years, machine learning (ML) models have become increasingly popular as a way to interpolate between *ab initio* calculations of both energy[29–34] and more complex properties[35–37] of atomistic structures (supervised learning), as well as automating time consuming analyses of atomistic simulations data (unsupervised learning).[38–43] The properties $\boldsymbol{y}$ of a physical system, A, obey a number of symmetries and conservation laws, and efforts to encode these at the core of atom-scale models have been shown to consistently improve the data efficiency of the regression scheme, making better use of the expensive electronic-structure calculations used for training. One option is to incorporate symmetries at the level of the model so that an appropriate representation can be learned from the data. For example, convolutional neural networks (CNN) architectures learn translational and scale-invariant features from images by design leading to a significant improvement of their performances compared to models based on features crafted by experts.[44] The main approach followed in the atomic scale modelling community this far has however been to develop representations of the atomic structure, defined by the positions and the species of its atoms $\{\boldsymbol{r}_i, a_i\}$ and the lattice vectors $\boldsymbol{h}_{1,2,3}$ for periodic structures, that extract features $\boldsymbol{x}$ that are equivariant with respect to these symmetries. Using these features as the input representation gives a ML model adapted to the desired symmetries.

Most of the current efforts have been geared towards the modelling of scalar properties, such as a system's energy, which are invariant with respect to permutations of the atoms label and rigid translations and rotations. Moreover, most physical observables are continuous functions of the atomic coordinates so an efficient representation would benefit from a certain level of smoothness. Different strategies have been proposed to incorporate these symmetries. Approaches based on internal coordinates (e.g. Coulomb matrices,[45,46] eigenvalues of overlap matrices[47] or bag of bonds[48]) are automatically invariant to rotations and translations but require an additional symmetrization over the permutation group. For low-dimensional problems this symmetrization can be performed exactly.[49–51] For larger systems, one can proceed by sorting the vector of interatomic distances or eigenvalues of a matrix that depends on interatomic distances.[47] However, both procedures introduce derivative discontinuities. Instead, many approaches to represent atomistic configurations rely more or less explicitly on atomic distributions, e.g radial distribution functions,[52] smooth overlap of atomic positions,[53] permutation invariant polynomials,[54] atom centered symmetry functions.[29] These *representations* need to be contrasted with the descriptors or fingerprints used in chemical and materials informatics. Instead of relying solely on atomic positions and types like *representations*, they incorporate heterogeneous information such as the degree of hybridization, atomic electronegativity, HOMO-LUMO energies...[28,55] with structural indicators such as backbone dihedral angles[56] and discrete secondary-structure categories[57,58] in proteins, graph representation of molecules[59] or histograms of coordination numbers[60] for clusters and condensed phase materials. Using domain-specific knowledge to model the relation between a material and its property can be very effective[35,37,61] but it also restricts the range of application of the method so this thesis will focus on *representations*.

The featurization of the atomic structure leads directly to the definition of distances over the chemical space, a key ingredient of *unsupervised learning* (UL). This family of techniques is centered around two main paradigms, clustering and dimensionality reduction, and aim at revealing patterns within a dataset of samples $\{\boldsymbol{x}_i\}_{i=1,2,\ldots,N}$ without labels. The clustering task tries to identify groups of samples within the data that are similar while reducing the dimensionality of the sample's features enables the identification of the key subspace in high-dimensional data.[62] Identifying groups of similar structures has found direct applications to improve crystal structure prediction methods,[63–65] where redundant crystals candidates are filtered, by providing more faithful metrics than the traditional root mean square displacement (RMSD).[47,66,67] Clustering techniques coupled with system dependant fingerprints have also been instrumental in the identification of metastable states in MD trajectories using Markov states models,[68,69] hierarchical clustering,[70] or gaussian mixture models.[40] Similarly, dimensionality reduction techniques can be used to infer the main slow structural transition path within a MD trajectory[41,71–73] or to provide a comprehensive visualization of entire databases of heterogeneous materials.[74,75] Low dimensional embeddings and visualizations of databases of materials are helpful tools to rationalize structure-property relations.[76]

On the other hand, *supervised learning* (SL) corresponds to a category of ML algorithms aiming to construct a model $f(\boldsymbol{x}) = y$ that can predict accurately the properties of a structure.[62] The internal parameters of the model are determined by optimizing the accuracy of prediction over a set of reference samples $\{(\boldsymbol{x}_i, y_i)\}_{i=1,2,\ldots,N}$ where $\boldsymbol{x}_i$ and y_i are respectively a set of features and target property associated with a training structure. The most appealing aspects of these methods lie in their systematic improvability with respect to the reference target by increasing the size of the training set.[77] One of the early applications of ML to the prediction of atomic-scale properties aimed at obtaining an accurate model of the potential energy surface (PES), which is crucial to assess the stability of a given configuration, and whose sampling underlies the evaluation of the thermodynamic properties of a system.[78] Contrary to traditional FFs, which assume physics-inspired functional forms for the interactions, and often use experimental observable as fitting targets, ML interatomic potentials (MLIPs) don't assume a fixed functional form and usually rely on electronic-structure calculations as a reference. In many cases, this more general, data-driven approach has been shown to result in more transferable and accurate models.[29,30,45,49] Besides the PES, ML models have also been successful at predicting other static lattice properties such as chemical shieldings,[79,80] density of states,[52] Hamiltonian matrix elements,[81] band gaps, electron affinities.[46]

In this thesis, we have aimed at improving the *description* of atomic systems to be used in conjunction with ML methods. We start from an abstract *representation* of structures and atomic environments to discuss atom-density-based approaches to chemical machine learning in Chapter 2.[82] We emphasize the basis-set independence of this representation by using the Dirac bra-ket notation. This framework provides a unifying picture of the field, in that several popular techniques can be seen as alternative representations of the same abstract feature vectors. In particular, we show that by representing these kets based on an expansion of atom-centered Gaussians in radial basis functions and spherical harmonics one recovers the smooth overlap of atomic positions (SOAP) representation.[53] From this formalism, we deduce general patterns to couple features of the representation. The resulting coupling weights are parameters that can be used to optimize and/or reduce the dimensionality of the original representation, recovering the flexibility of using different kinds of density-based representations within an elegant, unified framework.

Building on this foundation, we focus on the SOAP representation combined with (a) UL techniques to draw intuitions from repositories of hypothetical materials and (b) SL techniques accelerate reliable predictions of static lattice quantum properties. In Chapter 3, we expand from the REMatch metric,[75] non-linear dimensionality reduction and clustering techniques to address the challenges of navigating databases of molecular conformers[83] and molecular crystals,[84] checking their internal consistency and rationalizing structure-property relations. Even though we concentrate on particular databases of amino acid, dipeptide conformers, pentacene, and two azapentacene isomers obtained by an *ab initio* structure search,[85–87] many of the observations we infer are general and provide insight on the application of UL techniques to the analysis of structure-property relations in molecular and materials databases generated by high-throughput methodologies.

Finally, Chapter 4 demonstrate how Gaussian Process Regression (GPR)[88] with the SOAP representation is capable of predicting the relative energetics, chemical shifts, and transfer integrals that enter the evaluation of charge mobilities in molecular materials to high levels of accuracy.[84,89] To achieve state-of-the-art performances with machine-learned chemical shifts, we built a database of DFT calculated chemical shifts for structures taken from the Cambridge Structural Database (CSD),[90] chosen to be as structurally diverse as possible. Most significantly, even though no experimental shifts were used in training, we show that the model has sufficient accuracy to be used in a chemical shift driven NMR crystallography protocol to correctly determine, based on the match between experimentally-measured and ML-predicted shifts, the correct structure several pharmaceutical molecular crystals. To improve the performance of the regression further, we explore some optimizations of the input representation sketched in Chapter 2. We extend the SOAP representation by adapting the representation to the intrinsic length scales of atomic interactions, and by considering "alchemical" correlations between chemical species, which make it possible for instance to exploit the similar behavior of different elements to accelerate learning in very chemically heterogeneous data sets. Not only do these extensions improve significantly the performance of SOAP representations, but they do indeed offer insights into the chemistry of the system, for instance providing a data-driven representation of the similarity between elements that is reminiscent of the periodic table of the elements. Furthermore, to enhance the reliability of SL models, we compare uncertainty prediction estimators provided by sparse GPR and sub-sampling of the training set, by assessing their relative performances and propose a calibration procedure based on cross-validation to improve them. We demonstrate that the combination of sub-sampling with sparse GPR yields an inexpensive and reliable estimate of the uncertainty associated with the prediction of formation energies in the QM9,[25] Elpasolite crystal[33] and ^{1}H NMR chemical shieldings dataset.[89,90]

2 Theory of atomic scale representations[†]

The main reason underlying the development of representations of atomic structure is to build surrogate models of QM methods to predict static lattice properties by leveraging ML algorithms powered by collections of already computed references. These representations are therefore required to only incorporate inputs of QM methods, namely the set of atomic positions and species (with lattice vectors for periodic systems), and should act as a bridge between the atomistic and ML world. Moreover, the last decade of representation development has made it clear that ML models benefit greatly from including the physical knowledge of the target properties. QM observables are typically smooth functions of the atomic coordinates and they follow a certain set of symmetries. For instance, scalar quantities are invariant with respect to the permutation of atomic identities and rigid rotations and translations of the system. Including in the representation such invariances by design avoids the need to 'teach' them to the model, hence improving its data efficiency.[29,53,91,92] The effective locality of many QM interactions is another important feature to take advantage of when modelling extensive quantities, e.g. the total energy, since such methods are easier to transfer from small to larger systems compared to their global counterparts.

Following these guidelines, the main challenge is to devise representations of atomic systems that are at the same time complete and concise, so as to reduce the number of reference calculations that are needed to predict the properties of different types of materials reliably. This has led to a proliferation of alternative ways to convert an atomic structure into an input for a machine-learning model. We introduce an abstract definition of chemical environments that is based on a smoothed atomic density, using a bra-ket notation to emphasize basis set independence and to highlight the connections with some popular choices of representations for describing atomic systems. The correlations between the spatial distribution of atoms and their chemical identities are computed as inner products between these feature kets, which can be given an explicit representation in terms of the expansion of the atom density on orthogonal basis functions but also in real space, corresponding to n-body correlations

---

[†] This chapter has been adapted from the journal article [82] whose authors are Michael J. Willatt, **Félix Musil** and Michele Ceriotti. The author of this thesis worked on the links to other representations and the generalization of the invariant density representations of the article and contributed to the rest of the manuscript.

of the atom density. This formalism lays the foundations for a more systematic tuning of the behavior of the representations, by introducing operators that represent the correlations between structure, composition, and the target properties. It provides a unifying picture of recent developments in the field and indicates a way forward towards more effective and computationally affordable machine-learning schemes for molecules and materials.

This approach should be contrasted with the development of deep learning methods which integrate the definition of the representation into the learning procedure. The introduction of deep learning algorithms[44] has considerably improved the state-of-the-art performances in the fields of computer vision, speech recognition... which might lead to similar disruptive changes in atomistic ML. Nevertheless, QM observables are complex but well defined mathematical objects for which representations can take inspiration from more than half a century of theoretical modelling. Therefore it is not yet clear which of the two approaches is more effective or if such explicit assessment is even possible, and we focus on the development of explicit representations.

2.1 A Dirac notation for atomic configurations and environments

The Dirac (bra-ket) notation is often used to streamline the formulation of quantum-mechanical expressions, since it stresses basis-independence of quantum states, it helps avoiding manipulation errors and it is convenient to express linear transformations. For these reasons, we extend the use of this notation to the representation of atomic systems for ML (see also Appendix A where it simplifies greatly complex manipulations). Each atomic structure A that belongs to a dataset $\mathcal{D}$ is associated with a 'state' $|A; rep\rangle$ which gathers the elemental composition and geometric arrangement of atoms with the specification of the representation. When it is clear which representation is being used or that the discussion is focused on the representation $|A\rangle$ and $|rep\rangle$ will be used as shorthands. In the same spirit a property Y or the element of a basis b_n are casted into this bra-ket notation, such that $Y(A) := \langle Y|A\rangle$ is the property Y associated with A and $b_n := |n\rangle$ is the n^{th} basis function. Continuous basis such as the real space basis $\mathbf{x} := |\mathbf{x}\rangle$ are similarly transformed.

For the convenience of the reader, we summarize here a few shorthands that will be introduced later. Using the Dirac notation, basis set transformations are simply expressed as

$$\langle n|A\rangle = \sum_m \langle n|m\rangle \langle m|A\rangle , \tag{2.1}$$

$$\langle \mathbf{x}|A\rangle = \int \mathrm{d}\mathbf{r} \, \langle \mathbf{x}|\mathbf{r}\rangle \, \langle \mathbf{r}|A\rangle , \tag{2.2}$$

where $\langle A|m\rangle$ and $\langle A|\mathbf{r}\rangle$ are the coefficients for the change of basis. By abusing slightly this notation, we can express a linear model as

$$\langle Y|A\rangle \approx \sum_n \langle Y; rep|n\rangle \langle n|A; rep\rangle , \tag{2.3}$$

where $\langle Y|A\rangle$ is the prediction of the property Y of structure A, which is not a scalar product, and $\langle Y; rep|n\rangle$ are interpreted as the regression weights. By extension a kernel model can be written as

$$\langle Y|A\rangle \approx \sum_{T\in\mathcal{T}} \langle Y; rep|T; rep\rangle \langle T; rep|A; rep\rangle\,, \tag{2.4}$$

where T is an atomic configuration that belongs to the training set $\mathcal{T}$, $\langle Y; rep|T; rep\rangle$ are the model's weights and $\langle T; rep|A; rep\rangle$ is the kernel, e.g. inner product, between the training configurations and the structure A. The analogy between Eqs. (2.3) and (2.4) highlights the fact that in kernel models, predictions are performed using the training points as a basis. Supervised models are discussed in more details in Section 4.1 and an extensive discussion of linear and kernel methods can be found in Ref.[88]. Another useful construction is the tensor-product of representations

$$|(A; rep)\otimes(B; rep')\rangle = |A; rep\rangle \otimes |B; rep'\rangle\,, \tag{2.5}$$

which can be expressed as a Cartesian product of bases $|n\rangle\otimes|m\rangle$ or as a combined basis $|k\rangle$ with the shorthands

$$\langle n; m|A\otimes B\rangle = \langle n|A\rangle \langle m|B\rangle \rightarrow \langle k|A\otimes B\rangle\,. \tag{2.6}$$

Lastly, the symmetrization of a representation with respect to a symmetry group S of element $\hat{S}$ by Haar integration[93] is written as

$$|\langle A\otimes A\rangle_S\rangle = \int_{\hat{S}} \mathrm{d}\hat{S}\hat{S}|A\otimes A\rangle\,. \tag{2.7}$$

The symmetrized representation $|\langle A\otimes A\rangle_S\rangle$ will also be expressed in the more compact notation

$$|\langle A\otimes A\rangle_S\rangle \rightarrow |\overline{A\otimes A}\rangle \rightarrow |\overline{A^{\otimes 2}}\rangle\,, \tag{2.8}$$

where the group average and the tensor product are respectively highlighted by the overline and and the superscript.

2.1.1 Density-based structural representations

We represent the distribution of atoms in structure A as a density field $|A;\rho\rangle$ composed of smooth, real, positive and localized function $|g\rangle$, e.g. a Gaussian, centered on each atom and decorate them with orthonormal kets $|a\rangle$ to represent their elemental identities. Smoothness in the representation is beneficial as it leads to smooth kernels and better-behaved regression,[88] while the choice of a function that is clearly peaked at the atom positions encodes without ambiguity the full structural information. Moreover, the choice of a density field makes the representation automatically permutation invariant. Such an atomic configuration is written in position space as[‡]

$$\langle a\mathbf{x}|A;\rho\rangle = \sum_{i\in A}\delta_{aa_i}g(\mathbf{x}-\mathbf{r}_i), \tag{2.9}$$

where the sum is taken over all atoms in the configuration. This expansion could be generalized by using e.g. element-dependent widths in $g(\mathbf{x})$, i.e. $g(\mathbf{x}) \to g(s(a)\mathbf{x})$. Anisotropic functions may be used to represent entities with a preferential orientation, or some sort of internal structure. For a set of atoms, isotropy is a natural requirement for $g(\mathbf{x})$. Regardless of the particular form of $g(\mathbf{x})$ (provided that it is sufficiently localized), $|A\rangle$ provides a unique representation of the structure, but is variant with respect to fundamental physical symmetries, such as rigid translations $\hat{t}$ and rotations $\hat{R}$ of the constituent atoms $\{\mathbf{r}_i\} \to \{\hat{R}\hat{t}\mathbf{r}_i\}$.

2.1.2 Symmetry-invariant representations

To address the variance of Eq. (2.9) with respect to a symmetry operation $\hat{S}$, one can proceed by formally averaging the ket over the corresponding symmetry group (a procedure often referred to as Haar integration[93]):

$$|\langle\rho\rangle_S\rangle = \int_{\hat{S}} \mathrm{d}\hat{S}\,\hat{S}\,|\rho\rangle\,. \tag{2.10}$$

To see how this translates into symmetry-invariant representations, let us start by considering the relatively simple case of the integration over the translation group which simply corresponds to the integration over $\mathbb{R}^3$. Averaging directly over the position representation of $|\rho\rangle$ leads to a rather uninformative representation, which eliminates all structural information and only counts the number N_a of atoms belonging to each species,

$$\langle\mathbf{r}|\langle\rho\rangle_{\hat{t}}\rangle = \langle\mathbf{r}|\overline{\rho^{\otimes 1}}\rangle = \int_{\hat{t}}\mathrm{d}\hat{t}\,\langle\mathbf{r}|\,\hat{t}\,|\rho\rangle = \sum_i |a_i\rangle\int_{\mathbb{R}^3}\mathrm{d}\mathbf{t}\,g(\mathbf{t}+\mathbf{r}-\mathbf{r}_i) = \sum_a N_a\,|a\rangle\,, \tag{2.11}$$

where we have used the position representation of the translation operator. To avoid this information loss, one can perform the Haar integration over tensor products of $|\rho\rangle$, and define

$$|\overline{\rho^{\otimes\nu}}\rangle = \int_{\hat{t}}\mathrm{d}\hat{t}\,\underbrace{\hat{t}\,|\rho\rangle\otimes\hat{t}\,|\rho\rangle\ldots\hat{t}\,|\rho\rangle}_{\nu}. \tag{2.12}$$

[‡] We use the notation $\langle\mathbf{x}|A\rangle$ as shorthand for $\left[\,\langle\mathbf{x}|\otimes\hat{I}\right]|A\rangle$.

For $\nu = 2$, and assuming for simplicity that the same smooth density function is used for each atom, one gets

$$\begin{aligned}\langle \mathbf{r}\mathbf{r}'|\overline{\rho^{\otimes 2}}\rangle &= \int \mathrm{d}\hat{t} \sum_{ij} g(\hat{t}\mathbf{r}-\mathbf{r}_i)g(\hat{t}\mathbf{r}'-\mathbf{r}_j)\,|a_i a_j\rangle \\ &= \sum_{ij} |a_i a_j\rangle \int \mathrm{d}\mathbf{t}\, g(\mathbf{t}+\mathbf{r}-\mathbf{r}_i)g(\mathbf{t}+\mathbf{r}'-\mathbf{r}_j) \\ &= \sum_{ij} |a_i a_j\rangle\,(g\star g)(\mathbf{r}-\mathbf{r}'-\mathbf{r}_{ij}),\end{aligned} \tag{2.13}$$

where $\mathbf{r}_{ij} = \mathbf{r}_i - \mathbf{r}_j$, and $\star$ denotes the cross-correlation operation. We can simplify the notation for $|\overline{\rho^{\otimes 2}}\rangle$ in the position representation by (1) noting that the cross-correlation in Eq. (2.13) only depends on $\mathbf{r}-\mathbf{r}'$, so we can write the ket as a function of $\Delta\mathbf{r} = \mathbf{r}-\mathbf{r}'$ alone; (2) redefining the cross-correlation of two atom-density functions as $h = g\star g$:[§] [‖]

$$\langle \Delta\mathbf{r}|\overline{\rho^{\otimes 2}}\rangle = \sum_j |a_j\rangle \otimes \sum_i h(\Delta\mathbf{r}-\mathbf{r}_{ij})\,|a_i\rangle\,. \tag{2.14}$$

2.1.3 Tensor-product representations

Before proceeding further, let us comment briefly on the implications of the form of this ket for machine-learning of the properties associated with the structure $|A\rangle$, taking for simplicity a single-species compound so we can ignore the elemental kets. Learning a linear model is equivalent to the optimization of a linear mapping between the ket and the property, i.e.

$$\langle Y|A\rangle = \int_{\mathbb{R}^3} \mathrm{d}\mathbf{r}\,\langle Y;\overline{\rho^{\otimes 2}}|\mathbf{r}\rangle\,\langle \mathbf{r}|A;\overline{\rho^{\otimes 2}}\rangle\,, \tag{2.15}$$

where $\langle Y;\overline{\rho^{\otimes 2}}|\mathbf{r}\rangle$ represent the weights of the linear model. Taking the Dirac δ distribution limit of $g(\mathbf{r})$, one sees this is a (orientation-dependent) pair potential,

$$\langle Y|A\rangle = \sum_{ij} \langle Y;\overline{\delta^{\otimes 2}}|\mathbf{r}_{ij}\rangle\,, \tag{2.16}$$

and it is therefore easy to conceive properties that cannot be represented in this form. The feature vector itself, however, contains complete information about the structure, which can be recovered by taking tensor products of $|A\rangle$. For instance, if one takes the outer product of the translationally-symmetrized ket, learning amounts to the optimization of a function that

---

[§] For a generic basis function $g\star g$ might be a complicated function, but when g is a Gaussian, $g\star g$ is simply a Gaussian with double the variance

[‖] Like in Eq. (2.9), we use $\langle \Delta\mathbf{r}|\overline{\rho^{\otimes 2}}\rangle$ as shorthand for $\left[\langle \Delta\mathbf{r}|\otimes \hat{I}\otimes \hat{I}\right]|\overline{\rho^{\otimes 2}}\rangle$.

depends on two displacement vectors simultaneously,

$$\begin{aligned}\langle Y|A\rangle &= \int \mathrm{d}\mathbf{r}\mathrm{d}\mathbf{r}'\, \langle Y;\overline{\delta^{\otimes 2}}|\mathbf{r}\mathbf{r}'\rangle\, \langle \mathbf{r}'|A;\overline{\delta^{\otimes 2}}\rangle\, \langle \mathbf{r}|A;\overline{\delta^{\otimes 2}}\rangle \\ &= \sum_{iji'j'} \langle Y;\overline{\delta^{\otimes 2}}|\mathbf{r}_{ij}\mathbf{r}_{i'j'}\rangle\,,\end{aligned} \tag{2.17}$$

and so on. This simple example highlights how high-order correlations between atomic positions can be incorporated in the model by taking the tensor product of the structural ket before taking the Haar integral[53,94] (that is, choosing a high value of ν in Eq. (2.12)) or by taking a tensor product of the invariant ket,

$$\underbrace{|\langle\rho^{\otimes\nu}\rangle\rangle_{\hat{t}} \otimes |\langle\rho^{\otimes\nu}\rangle_{\hat{t}}\rangle \otimes \cdots \otimes |\langle\rho^{\otimes\nu}\rangle_{\hat{t}}\rangle}_{\zeta} \to |\langle\rho^{\otimes\nu}\rangle_{\hat{t}}\rangle^{\otimes\zeta}\,. \tag{2.18}$$

The latter choice corresponds to taking element-wise powers of the linear invariant kernel. In other terms, using a *unique* representation of a structure in a non-linear ML model can introduce higher body order correlations than those explicitly afforded by the feature vector itself.

2.1.4 Atom-centered representations

Having clarified how tensor-product kets can be used to incorporate higher-order correlations between the atoms, let us move on to discuss how the representation of a structure as a sum of atom-centered environments arises naturally, starting from a sum of atom-centered densities, as a by-product of symmetrization over the translation group. By grouping together the terms in the sum corresponding to displacement vectors involving atom j, the translationally-invariant second-order ket Eq. (2.14) decomposes into atom-centered contributions,

$$|A;\langle\rho_j\otimes\rho_j\rangle_{\hat{R}}\rangle = |A;\overline{\rho_j^{\otimes 2}}\rangle = \sum_j |\alpha_j\rangle\otimes|A;\rho_j\rangle\,, \tag{2.19}$$

where we have dropped the indication of the translational averaging from $|A;\rho_j\rangle$ to keep at bay the complexity of the notation. Note that Eq. (2.19) implies an additive definition of the relation between the representations of the entire structures, and those associated with atom-centered environments.

The position representation of the environmental atom-centered ket $|A;\rho_j\rangle$ is

$$\langle\mathbf{r}|\rho_j\rangle = \sum_{i\in j} f_c(r_{ij})h(\mathbf{r}-\mathbf{r}_{ij})\,|\alpha_i\rangle\,. \tag{2.20}$$

In this definition we have introduced a smooth cutoff function $f_c(r_{ij})$ so that each environment only depends on the position of the atoms in a spherical neighborhood centered on atom j. While one could in principle proceed with an atom-centered description that incorporates information from the entire structure, by making $f_c(r) = 1$, localisation is useful for computational reasons and is justified when studying atomic problems in light of the nearsightedness principle of electronic matter,[95] which underlies most linear-scaling electronic structure methods.[96–98] Note that when the ket is written in this form it might make sense to further generalize the definition of h, e.g. by making its width dependent on $r_{ij} = |\mathbf{r}_{ij}|$, $h(\mathbf{r}) \to h(s(r_{ij})\mathbf{r})$, or by choosing a form other than a Gaussian that is more flexible or computationally efficient. The notation can be further simplified by emphasizing the representations of structure and composition,

$$\langle a\mathbf{r}|\rho_j\rangle = \sum_{i\in j} \delta_{a_j a_i} f_c(r_{ij}) h(\mathbf{r}-\mathbf{r}_{ij}). \tag{2.21}$$

Writing the ket as a sum over all elements $a = \mathrm{H}, \mathrm{He}, \ldots$

$$\langle \mathbf{r}|\rho_j\rangle = \sum_a \langle a\mathbf{r}|\rho_j\rangle\, |a\rangle\,. \tag{2.22}$$

This translationally-invariant atom-centered environment representation can also be adapted by taking a linear transformation $\hat{U}\,|\rho_j\rangle \to |\rho_j\rangle$, where the linear operator $\hat{U}$ might act in the position space, the element space or both. As we will see, the freedom in choosing the form of $\hat{U}$ can be used to tune the behavior of the representation to describe in a more efficient way the relation between structure and properties.

2.1.5 Rotationally-invariant representations

In order to obtain a rotationally-invariant representation, one can formally average the ket $|\rho_j\rangle$ over the $SO(3)$ rotation group,

$$|\overline{\rho^{\otimes 1}}\rangle = \int_{SO(3)} d\hat{R}\,\hat{R}\,|\rho_j\rangle\,. \tag{2.23}$$

This ket can be readily computed in the position representation. Taking for simplicity the case where only one element is present, one gets

$$\langle \mathbf{r}|\overline{\rho^{\otimes 1}}\rangle = \int \mathrm{d}\hat{R}\,\langle \mathbf{r}|\hat{R}|\rho_j\rangle = \int \mathrm{d}\hat{R}\,\langle r\hat{R}\hat{\mathbf{e}}_z|\rho_j\rangle\,, \tag{2.24}$$

where we have used the fact that the integral is over all the rotation matrices, and so we can always rotate $\mathbf{r}$ to be aligned with the Cartesian z axis $\hat{\mathbf{e}}_z$ before taking the integral. The average can be written explicitly in terms of a suitable parameterization of the rotations, e.g. using Euler angles,

$$\frac{1}{8\pi^2}\int_0^{2\pi} \mathrm{d}\alpha \int_0^{\pi} \sin\beta \mathrm{d}\beta \int_0^{2\pi} \mathrm{d}\gamma\,\langle r\hat{R}(\alpha,\beta,\gamma)\hat{\mathbf{e}}_z|\rho_j\rangle\,. \tag{2.25}$$

One can then recognize that the γ angle does not affect $\hat{\mathbf{e}}_z$, so the integral can be written equivalently as an average over the unit sphere.[¶] We can then define

$$\langle r|\overline{\rho^{\otimes 1}}\rangle \propto r\int \mathrm{d}\hat{R}\, \langle r\hat{R}\hat{\mathbf{e}}_z|\rho_j\rangle = \frac{1}{4\pi} r \int \mathrm{d}\hat{\mathbf{r}}\, \langle r\hat{\mathbf{r}}|\rho_j\rangle\,, \tag{2.26}$$

where we have highlighted the fact that the position representation only depends on r, and we have explicitly included a factor of r so that

$$\int \mathrm{d}\mathbf{r}\, \langle\overline{\rho^{\otimes 1}}|\mathbf{r}\rangle\, \langle\mathbf{r}|\overline{\rho^{\otimes 1}}\rangle = \int \mathrm{d}r\, \langle\overline{\rho^{\otimes 1}}|r\rangle\, \langle r|\overline{\rho^{\otimes 1}}\rangle\,. \tag{2.27}$$

Much like in the case of translations, the average over rotations eliminates too much information, and $\overline{|\rho^{\otimes 1}\rangle}$ does not retain knowledge of the angular correlations of atoms around the center of the environment. A more general family of invariant kets can be obtained by starting from the tensor products of (possibly different) environmental kets, $\hat{U}_1\,|\rho_j^1\rangle \otimes \hat{U}_2\,|\rho_j^2\rangle \otimes \ldots$, and then symmetrizing over the rotation group,

$$\overline{|\rho_j^{\otimes\nu}\rangle} = \int \mathrm{d}\hat{R} \prod_{\aleph}^{\nu} \otimes\, \hat{R}\hat{U}_{\aleph}\,|\rho_j^{\aleph}\rangle\,. \tag{2.28}$$

As for the case of translational averages, one can use a linear map to build a machine-learning model of a property based on these symmetrized kets. Non-linear features correspond to tensor products of symmetrized kets such as

$$\underbrace{\overline{|\rho_j^{\otimes\nu}\rangle} \otimes \overline{|\rho_j^{\otimes\nu}\rangle} \otimes \ldots \otimes \overline{|\rho_j^{\otimes\nu}\rangle}}_{\zeta} \to \overline{|\rho_j^{\otimes\nu}\rangle}^{\otimes\zeta}\,, \tag{2.29}$$

and one could further generalize the construction by taking products of kets built from different $\hat{U}$ operators.

2.2 A unified picture of density-based representations

Equation (2.28) provides a very general – and abstract – definition of a density-based representation of an atomic structure that encodes translational, rotational and permutation symmetries. This level of abstraction provides a unifying picture of the field, in that many of the representations that have been used for machine-learning of atomic-scale properties can be seen as special cases of this form, or as the result of projection onto a particular choice of basis.

[¶] This is a consequence of the fact that $SO(3)$ is the product of $SO(2)$ and S^2

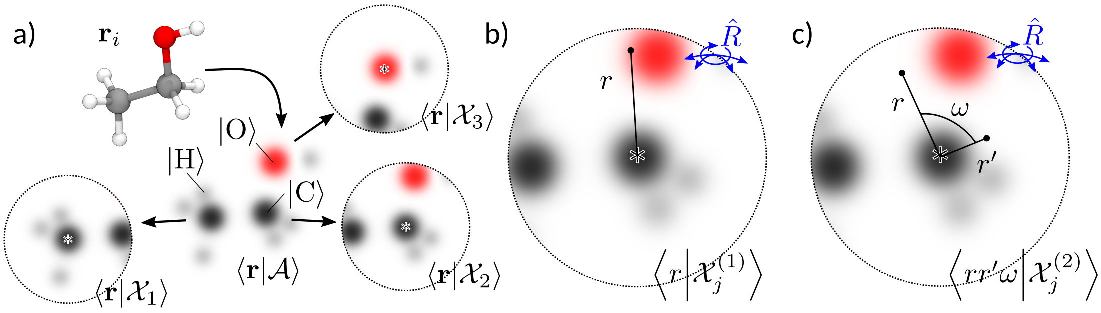

a)
$\mathbf{r}_i$
$|\mathrm{O}\rangle$
$|\mathrm{H}\rangle$
$|\mathrm{C}\rangle$
$\langle\mathbf{r}|\mathcal{X}_3\rangle$
$\langle\mathbf{r}|\mathcal{A}\rangle$
$\langle\mathbf{r}|\mathcal{X}_1\rangle$
$\langle\mathbf{r}|\mathcal{X}_2\rangle$
b)
r
$\hat{R}$
$\left\langle r \middle| \mathcal{X}_j^{(1)} \right\rangle$
c)
r
ω
r'
$\hat{R}$
$\left\langle r r' \omega \middle| \mathcal{X}_j^{(2)} \right\rangle$

Figure 2.1 – Atom-density-based structural representations, expressed in the real-space $\langle\mathbf{r}|$ basis. (a) A structure can be mapped onto a smooth atom density built as a superposition of smooth atom-centered functions. The overall density can be decomposed in atom-centered environments, and information on chemical compositions can be stored by decorating the functions with elemental kets. (b) The $\nu = 1$ invariant ket corresponds to spherical averaging of the environmental atom density. (c) The $\nu = 2$ invariant ket corresponds to three-body correlations, which are obtained by integrating over all rotations a stencil corresponding to two distances along two directions with a fixed angle $\arccos\omega$ between them.

2.2.1 Plane waves

Let us start by considering the translationally-invariant ket $|\langle\rho^{\otimes 2}\rangle_{\hat{t}}\rangle$, writing it in a plane-waves basis $\{|\mathbf{k}\rangle\}$, and taking for simplicity the $h \to \delta$ limit. One obtains a representation that is equivalent to the diffraction pattern generated by the structure, decomposed in multiple channels that correspond to the reciprocal-space correlations between different atomic species,

$$\langle\mathbf{k}|\langle\delta^{\otimes 2}\rangle_{\hat{t}}\rangle = \sum_{ij} |\alpha_i\alpha_j\rangle\, e^{\mathrm{i}\mathbf{k}\cdot\mathbf{r}_{ij}}. \tag{2.30}$$

When considering a periodic structure, and with an appropriate normalization, this representation is directly connected with the fingerprints that have been recently used to identify crystalline structures,[99] highlighting how different choices of basis may be best suited to different applications.

2.2.2 Many-body kernels and representations

Moving on to the case of rotationally-invariant kets, let us take for simplicity $\hat{U}_{\aleph} = 1$, and assume that all the environmental kets that are multiplied in Eq. (2.28) are the same. We will revisit later the possibility of introducing a linear operator to fine-tune the properties of the representation. Since we have started from a position representation for the environmental kets, it is natural to write Eq. (2.28) explicitly in a complete basis of position and element states, $|\prod_{\aleph}^{\nu}\alpha_{\aleph}\mathbf{r}_{\aleph}\rangle \equiv \prod_{\aleph}^{\nu}\otimes|\alpha_{\aleph}\mathbf{r}_{\aleph}\rangle$,

$$\langle\prod_{\aleph}^{\nu}\alpha_{\aleph}\mathbf{r}_{\aleph}|\overline{\rho_j^{\otimes\nu}}\rangle = \int d\hat{R}\prod_{\aleph}^{\nu}\langle\alpha_{\aleph}\hat{R}\mathbf{r}_{\aleph}|\rho_j\rangle\,. \tag{2.31}$$

One can see clearly that the kernels associated with Eq. (2.31) are in the form of the invariant n-body kernels discussed in Ref.[94] (more specifically, as we will see below, they correspond precisely to the SOAP kernels if h is a Gaussian). Considering the case with a single element,

$$\langle\overline{\rho_k^{\otimes\nu}}|\overline{\rho_j^{\otimes\nu}}\rangle = \int \mathrm{d}\hat{R}\mathrm{d}\hat{R}' \left[\int \mathrm{d}\mathbf{r}\,\langle\rho_k|\hat{R}'\mathbf{r}\rangle\,\langle\hat{R}\mathbf{r}|\rho_j\rangle\right]^{\nu}, \tag{2.32}$$

it is clear that one of the two Haar integrals is redundant and can be eliminated. Let us consider the effect of ν on the representation and the information that it captures. As discussed in the case of $\nu = 1$ following Eq. (2.24), one of the input vectors $\mathbf{r}$ can be aligned with a fixed reference axis, e.g. $\hat{\mathbf{e}}_z$. The fact that this axis is invariant under one of the Euler rotations makes it possible to align a second vector so that it lies in the xz plane. For $\nu = 1$ and $\nu = 2$ this analysis leads to

$$\begin{aligned}\langle ar|\overline{\rho_j^{\otimes 1}}\rangle &\propto r\int \mathrm{d}\hat{R}\,\langle ar\hat{R}\hat{\mathbf{e}}_z|\rho_j\rangle\\ \langle ara'r'\omega|\overline{\rho_j^{\otimes 2}}\rangle &\propto rr'\int \mathrm{d}\hat{R}\,\langle a;r\hat{R}\hat{\mathbf{e}}_z|\rho_j\rangle\,\langle a';r'\hat{R}(\omega\hat{\mathbf{e}}_z+\sqrt{1-\omega^2}\hat{\mathbf{e}}_x|\rho_j\rangle\,,\end{aligned} \tag{2.33}$$

where $\omega = \hat{\mathbf{r}}\cdot\hat{\mathbf{r}}'$ (see Fig. 2.1). After one has aligned the first two $\mathbf{r}_\aleph$'s, the position of all the other $\mathbf{r}_\aleph$'s cannot be manipulated, so in practice for $\nu > 2$ each further order brings in three degrees of freedom, that are expressed in the reference system in which the first two vectors are aligned along the z axis and lie in the xz plane. For $\nu = 3$,

$$\langle ara'r'\omega a''r''\hat{\mathbf{r}}''|\overline{\rho_j^{\otimes 3}}\rangle \propto rr'r''\int \mathrm{d}\hat{R}\,\langle a;r\hat{R}\hat{\mathbf{e}}_z|\rho_j\rangle\,\langle a';r'\hat{R}(\omega\hat{\mathbf{e}}_z+\sqrt{1-\omega^2}\hat{\mathbf{e}}_x|\rho_j\rangle\,\langle a'';r''\hat{R}\hat{\mathbf{r}}''|\rho_j\rangle\,. \tag{2.34}$$

Also note that we have incorporated the square root of the Jacobian in the definition of the representations so that the corresponding kernels can be computed straightforwardly as the inner product between two vectors without scaling.

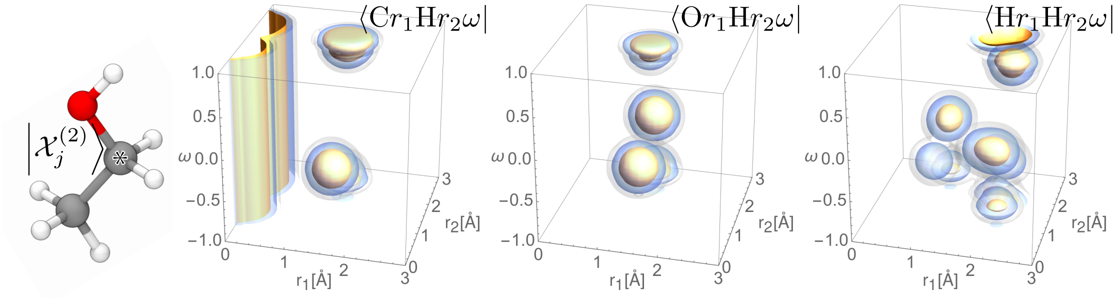


Figure 2.2 – Isocontours of the 3-body correlation functions associated with the environment centered on the tagged carbon atom of an ethanol molecule. From left to right, the figures correspond to $\langle \mathrm{C}r\mathrm{H}r'\omega|\overline{\rho_j^{\otimes 2}}\rangle/rr'$, $\langle \mathrm{O}r\mathrm{H}r'\omega|\overline{\rho_j^{\otimes 2}}\rangle/rr'$, $\langle \mathrm{O}r\mathrm{H}r'\omega|\overline{\rho_j^{\otimes 2}}\rangle/rr'$.

By expanding the densities as sums over atoms, it becomes clear that these kets are representations of the $(\nu+1)$-body order correlations between atoms within an environment[53,94] (Fig. 2.2). To start with, we return to the delta function limit of the atomic densities. In the limit in which each atomic density is represented by Dirac δ distributions, the position representations of the invariant vectors take very simple forms:

$$\begin{aligned}\langle ar|\overline{\delta_j^{\otimes 1}}\rangle &\propto r\sum_i \delta_{aa_i} f_c(r_{ij})\delta(r-r_{ij})\\ \langle ara'r'\omega|\overline{\delta_j^{\otimes 2}}\rangle &\propto rr'\sum_{ii'}\delta_{aa_i}\delta_{aa_{i'}}\delta(r-r_{ij})\delta(r'-r_{i'j})\delta(\omega-\hat{\mathbf{r}}_{ij}\cdot\hat{\mathbf{r}}_{i'j})f_c(r_{ij})f_c(r_{i'j}).\end{aligned} \tag{2.35}$$

The δ-distribution limit of the density kets is equivalent to the atomic cluster expansion (ACE) framework of Ref.[100,101], which has been independently developed to increase body order. Note that the symmetrized overlap between atomic densities defined in Eq. (2.32) does not vanish in the delta limit thanks to the smoothness of the radial basis functions used to expand the densities $|\delta_j\rangle$. Linear regression based on $|\overline{\rho_j^{\otimes\nu}}\rangle$ corresponds to $(\nu+1)$-body potentials[94] e.g. for the 3-body term,

$$\langle Y|A\rangle = \sum_j \int \mathrm{d}r\mathrm{d}r'\mathrm{d}\omega \, \langle Y;\overline{\rho_j^{\otimes 2}}|r,r',\omega\rangle \, \langle r,r',\omega|A;\overline{\rho_j^{\otimes 2}}\rangle \,. \tag{2.36}$$

There are however good reasons to use non-linear functions of the feature vector in an ML model. In the case of sufficiently sharp atom-centered density functions, the ket with $\nu = 1$ contains information on the list of all pair distances within an environment, which is not sufficient to reconstruct the structure of the environment unequivocally. The representation with $\nu = 2$, on the other hand, contains information on pair distances and angles between triplets of atoms. Contrary to the original understanding,[53] it has been recently shown that this information is not sufficient to represent arbitrarily complex invariant functions of the atomic coordinates.[102] Despite this limitation, the tensor products of the $|\overline{\rho_j^{\otimes 2}}\rangle$ ket seem good enough in practice to model complex scalar properties.

2.2.3 Behler-Parrinello symmetry functions

An expression of Eq. (2.28) in the position representation and in the $h \to \delta$ limit is an ideal starting point to investigate the relationship of $|\overline{\rho_j^{\otimes\nu}}\rangle$ with other density-based frameworks. These expressions reveal the connection between these invariant kets and several popular fingerprints designed to capture pair and 3-body interactions. The link between $\langle ar|\overline{\rho_j^{\otimes 1}}\rangle$ and the pair distribution function[103] is obvious. Behler-Parrinello symmetry functions, and similar weighed averages of n-body correlations, can be seen as projections of the $SO(3)$ invariant ket over suitable test functions G. For instance, for a 2-body symmetry function $G_2(r)$ one has

$$\langle aa'G_2|\overline{\delta^{\otimes 1}}\rangle = \langle a|a_j\rangle \int \mathrm{d}r \, G_2(r) r \, \langle a'r|\overline{\delta^{\otimes 1}}\rangle \,, \tag{2.37}$$

and an analogous expression can be written for a 3-body symmetry function $G_3(r,r',\omega)$. Expressions similar to Eq. (2.37) can be obtained by inserting into Eq. (2.33) Gaussians, or alternative basis functions. The relationship to other density-based representations, such as those discussed in Refs.[104,105] is less transparent, but several of the essential ingredients – such as scaling functions that modulate geometric and chemical correlations – can be introduced in terms of appropriate choices of the $\hat{U}$ operators, as we will discuss in the next section.

2.2.4 Smooth Overlap of Atomic Positions

We have left as a last example a discussion of the connection between the symmetrized ket and the Smooth Overlap of Atomic Positions (SOAP) power spectrum.[53,75] In fact, if we take as we did before $\hat{U}_{\aleph} = 1$ and $|\rho^1\rangle = |\rho^2\rangle = \ldots |\rho^\nu\rangle$ in Eq. (2.28), the SOAP power spectrum is nothing but an alternative representation of $|\overline{\rho_j^{\otimes 2}}\rangle$. To see how, one can start by expanding the translationally-invariant environmental ket Eq. (2.22) in a basis of orthonormal radial basis functions $R_n(r) = \langle r|n\rangle$ and spherical harmonics $Y_m^l(\hat{\mathbf{r}}) = \langle lm|\hat{\mathbf{r}}\rangle$,

$$\langle anlm|\rho_j\rangle = \int \mathrm{d}\mathbf{r}\, \langle n|x\rangle\, \langle lm|\hat{\mathbf{x}}\rangle\, \langle a\mathbf{x}|\rho_j\rangle\,. \tag{2.38}$$

Using a basis of spherical harmonics is extremely useful and practical because they block diagonalize the angular momentum operator (and thus the rotation operator), which allows for explicit integration over the rotation group in Eq. (2.31) (see Section A.3 for an explicit derivation of the $\langle anlm|\rho_j\rangle$ coefficients for some radial basis). or $\nu = 1$, this leads to the following feature vector,

$$\langle an|\overline{\rho_j^{\otimes 1}}\rangle \propto \langle an00|\rho_j\rangle\,. \tag{2.39}$$

For $\nu = 2$, the feature vector corresponds to the SOAP power spectrum,

$$\langle a_1n_1a_2n_2l|\overline{\rho_j^{\otimes 2}}\rangle \propto \frac{1}{\sqrt{2l+1}}\sum_m (-1)^m \langle a_1n_1lm|\rho_j\rangle\, \langle a_2n_2l\; -m|\rho_j\rangle\,. \tag{2.40}$$

For $\nu = 3$ the representation corresponds to the bispectrum,[53]

$$\langle a_1n_1l_1a_2n_2l_2a_3n_3l_3|\overline{\rho_j^{\otimes 3}}\rangle \propto \sum_{\substack{m_1m_2\\m_3}} \begin{pmatrix} l_1 & l_2 & l_3 \\ m_1 & m_2 & m_3 \end{pmatrix} \langle a_1n_1l_1m_1|\rho_j\rangle\, \langle a_2n_2l_2m_2|\rho_j\rangle\, \langle a_3n_3l_3m_3|\rho_j\rangle\,, \tag{2.41}$$

where the parentheses denote a Wigner 3j symbol. For a full derivation of Eqs. (2.39) to (2.41) refer to Section A.2. The bispectrum is used as a four-body feature vector in SOAP and in Spectral Neighbor Analysis Potentials (SNAP), where its high resolution is exploited to construct accurate interatomic potentials through linear regression.[106]

Seen in the light of the present formalism, the remarkable fact that the SOAP kernel (Eq. (2.32) with densities written as a sum of Gaussians) can be expressed as an explicit scalar product between vectors, representing a truncated expansion of the power spectrum, emerges as a

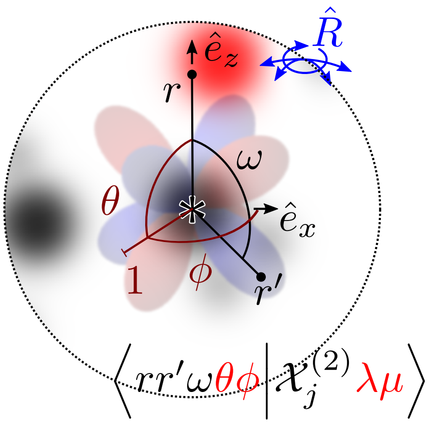

$\hat{R}$
$\hat{e}_z$
r
ω
θ
$\hat{e}_x$
1
ϕ
r'
$\left\langle rr'\omega\theta\phi\middle|\mathcal{X}_j^{(2)}\lambda\mu\right\rangle$

Figure 2.3 – Schematic representation of the construction of a real-space representation of a tensorial ket associated with a λ-SOAP kernel. The (smooth) atom density is evaluated at two points corresponding to a stencil (r,r',ω), and the spherical harmonic Y_μ^λ is evaluated at the angles (θ,ϕ), relative to the reference frame that is used to describe the stencil.

natural consequence of the definition of the kernel as the scalar product between invariant kets. It should also be noted that in practical applications of SOAP the kernels are often (but not always) normalized and raised to an integer power ζ, which corresponds to taking a tensor product of the kets and introduces a many-body character in the model built on such kernels.

2.2.5 Tensorial Smooth Overlap of Atomic Positions (λ-SOAP)

The feature vectors that appear in the tensorial extension of SOAP[92] are of the form in Eq. (2.28), with $\hat{U}_\aleph = \hat{I}$ for $\aleph = 1,2,\ldots,\nu+1$, $|\rho_j^\aleph\rangle = |\rho_j\rangle$ for $\aleph = 1,\ldots,\nu$ and $|\rho_j^{\otimes\nu+1}\rangle = |\lambda\mu\rangle$, where $|\lambda\mu\rangle$ is an angular momentum ket:

$$|\overline{\rho_j^{\otimes\nu};\lambda\mu}\rangle = \int \mathrm{d}\hat{R}\,\hat{R}|\lambda\mu\rangle \prod_{\aleph=1}^{\nu} \otimes \hat{R}|\rho_j\rangle\,. \tag{2.42}$$

The ket is rotationally invariant,

$$\left[\prod_{\aleph=1}^{\nu+1} \otimes \hat{R}\right] |\overline{\rho_j^{\otimes\nu};\lambda\mu}\rangle = |\overline{\rho_j^{\otimes\nu};\lambda\mu}\rangle\,, \tag{2.43}$$

but not in the subspace that describes the atomic environments,

$$\left[\hat{I}\otimes\prod_{\aleph=1}^{\nu} \otimes \hat{R}\right] |\overline{\rho_j^{\otimes\nu};\lambda\mu}\rangle \neq |\overline{\rho_j^{\otimes\nu};\lambda\mu}\rangle\,. \tag{2.44}$$

The inner product between two of these vectors is easily shown to be

$$\langle\overline{\rho_j^{\otimes\nu};\lambda\mu}|\overline{\rho_k^{\otimes\nu};\lambda'\mu'}\rangle = \delta_{\lambda\lambda'}\int d\hat{R}\,D_{\mu\mu'}^{\lambda}(\hat{R})\left[\langle\rho_j|\rho_k\rangle\right]^\nu, \tag{2.45}$$

which agrees with the usual definition of the λ-SOAP kernel,

$$\langle\overline{\rho_j^{\otimes\nu};\lambda\mu}|\overline{\rho_k^{\otimes\nu};\lambda'\mu'}\rangle = k^{\lambda}_{\mu\mu'}(\rho_j,\rho_k). \tag{2.46}$$

While $|\overline{\rho_j^{\otimes\nu};\lambda\mu}\rangle$ can be represented very effectively using a spherical-harmonics expansion of the atom density,[92] it is also possible to express it in terms of a real-space basis. Following arguments similar to those used to derive Eq. (2.34), one can see that in this form the tensorial ket corresponds to the evaluation of a three-body correlation function of the atom density, multiplied by a spherical harmonic of appropriate order computed in the reference frame of the (r,r',ω) stencil (see Fig. 2.3).

Taking tensor products of $|\overline{\rho_j^{\otimes\nu};\lambda\mu}\rangle$ with itself increases the order of body correlations that are explicitly included in the feature vector for which an efficient evaluation procedure can be found in Ref.[107]. Instead, one can take tensor products with $\lambda=0$ kets, which are rotationally invariant in the subspace that describes the atomic environments while preserving the desired symmetry of the representation, e.g.

$$|\overline{\rho_j^{\otimes\nu};\lambda\mu}\rangle\prod_{k=1}^{\zeta-1}\otimes|\overline{\rho_j^{\otimes\nu}}\rangle \rightarrow |\overline{\rho_j^{\otimes\nu};\lambda\mu}\rangle^{\otimes\zeta}. \tag{2.47}$$

This procedure has been found effective in practice for increasing the order of body correlations in tensorial SOAP.[108,109]

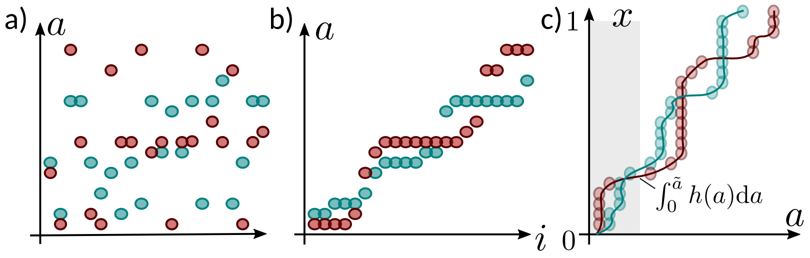


Figure 2.4 – (a) Permutation-variant structural descriptors can be stored in a vector to be used as an atomic-scale representation. (b) Sorting this vector makes it permutationally invariant. (c) It is easy to see how the sorted vector relates to the cumulative distribution function associated with the histogram of the values of the structural features.

2.2.6 Distributions vs sorted vectors

It is worth making some further considerations that extend somewhat the generality of this construction to include representations that are *not* based explicitly on atom densities. Many approaches in the literature rely on computing quantities that are not permutationally invariant *per se*, for instance the elements of the matrix of pair distances between atoms,[110] transformed elementwise by some function,[45] or the eigenvalues of such matrices.[47] In order to make these representations invariant to atom permutations, one often proceeds to sort these sets of items, and uses the Euclidean distance between the sorted vectors as the building block of kernels or other statistical learning frameworks.

In fact, it is easy to see that given a set of elements $\{e_i \in \mathbb{R}\}$, the sorted list contains the same amount of information as the histogram of the elements $h(e)$ (see Fig. 2.4). Scaling the index of the sorted items by the total number of items N, and considering the limit in which one can take $x = i/N$ as a continuous index, one sees that $x(\tilde{e})$ counts the fraction of entries that are smaller than $\tilde{e}$, that is

$$x(\tilde{e}) = \int_{-\infty}^{\tilde{e}} \mathrm{d}e\, h(e). \tag{2.48}$$

It follows that $e(x)$, which is a continuous representation of the vector of sorted distances, is just the inverse cumulative distribution function (iCDF) associated with $h(x)$. The Euclidean distance between two vectors of sorted elements is proportional to the $\mathcal{L}^2$ norm of the difference between the iCDF of the histograms associated with the two sets. Interestingly, if one considers the $\mathcal{L}^1$ norm, the distance between the sorted vectors corresponds to the earth mover's distance[111] between two distributions in one dimension.

The connection between different density-based representations is more direct than that which can be established between density-based and sorted-vector descriptions – also given that the relation between atom positions and the permutation-variant items might be far from trivial, e.g. when the representation involves the eigenvalues of an overlap matrix. However, the argument we present here highlights the fact that incorporating physical symmetries in the description of atomistic systems leads to representations that contain essentially the same information.

2.2.7 Use of density-based features in artificial neural networks

It should be stressed that the density-based representations presented so far could also be used as an input of an Artifical Neural Network (ANN). The simplest case is when one aims to predict a scalar property. Since the ket corresponding to a representation for learning scalars is invariant under permutations, translations and rotations, it follows that each of its components in a basis is an invariant, e.g.

$$\langle a r l a' r' l' \ldots | \hat{R} \overline{\rho_j^{\otimes \nu}} \rangle^{\otimes \zeta} = \langle a r l a' r' l' \ldots | \overline{\rho_j^{\otimes \nu}} \rangle^{\otimes \zeta}. \tag{2.49}$$

This means that one can construct a model of the form

$$\langle Y | A \rangle = \sum_j f \left[\left\{ \langle a r l a' r' l' \ldots | A; \overline{\rho_j^{\otimes \nu}} \rangle^{\otimes \zeta} \right\} \right], \tag{2.50}$$

where f is an arbitrary non-linear function of the components (with a nested structure in the case of an ANN), and the predicted property $\langle Y | A \rangle$ will necessarily fulfill the same invariances as the representation. Given the earlier demonstration of how Behler-Parrinello symmetry

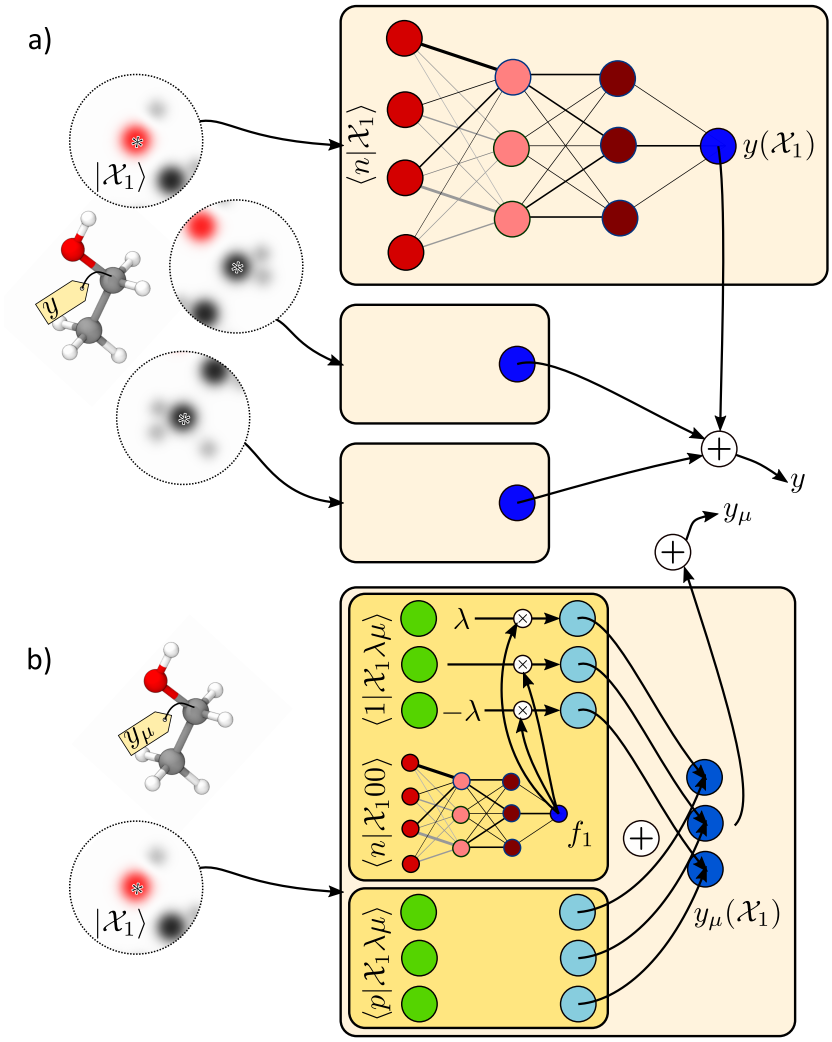

a)
$|\mathcal{X}_1\rangle$
y
$\langle n|\mathcal{X}_1\rangle$
$y(\mathcal{X}_1)$
y
y_μ
b)
y_μ
$\langle 1|\mathcal{X}_1\lambda\mu\rangle$
λ
$-\lambda$
$\langle n|\mathcal{X}_1 00\rangle$
f_1
$|\mathcal{X}_1\rangle$
$\langle p|\mathcal{X}_1\lambda\mu\rangle$
$y_\mu(\mathcal{X}_1)$

Figure 2.5 – (a) Schematic representation of an ANN in which the input layer corresponds to the elements of a density-based representation. The target property associated with the full structure is expressed as a sum of atomic contributions. (b) In the case of a tensorial property it is essential to preserve the covariant nature of the λ-SOAP ket. To do so, one can construct a NN using only scalar SOAP features, and use the output as a multiplier for the tensorial features. The output of several of these blocks must then be combined linearly and without mixing different $\lambda\mu$ components to obtain the environment's contribution to a tensorial property.

functions fit into the abstract framework described in this article, this observation should come as no surprise since Behler-Parrinello symmetry functions have enjoyed great success as ANN inputs.

One must be more careful with a representation engineered for use with a tensor model (e.g. λ-SOAP presented earlier). In the same way that taking tensor products of $|\overline{\rho_j^{\otimes\nu};\lambda\mu}\rangle$ with itself destroys the desired symmetry properties of the representation, feeding expansion coefficients of $|\overline{\rho_j^{\otimes\nu};\lambda\mu}\rangle$ into an arbitrary non-linear function (e.g. an ANN) will generally lead to predictions that do not satisfy the desired symmetry properties. One can, however, construct a generalisation of Eq. (2.50) in which expansion coefficients of $|\overline{\rho_j^{\otimes\nu};\lambda\mu}\rangle$ enter only at the last layer, i.e.

$$\langle Y|A;\lambda\mu\rangle = \sum_j f_{arla'r'l'\ldots}\left[\left\{\langle arla'r'l'\ldots|A;\overline{\rho_j^{\otimes\nu}}\rangle^{\otimes\zeta}\right\}\right]\langle arla'r'l'\ldots|A;\overline{\rho_j^{\otimes\nu};\lambda\mu}\rangle\,, \tag{2.51}$$

where each $f_{arla'r'l'\ldots}$ is an arbitrary non-linear function of the components (see Fig. 2.5). With such a structure, the predictions are guaranteed to satisfy the same symmetry properties as the representation $|\overline{\rho_j^{\otimes\nu};\lambda\mu}\rangle$, an idea that has been put in practice in the Cormorant architecture.[112,113]

2.3 Generalized invariant density representations

The formalism we have introduced in the previous section provides an elegant framework to construct a rotationally-invariant representation of the atomic density that can be used for machine-learning purposes. While the formalism provides a complete description of structural correlations of a given order within an atomic environment, the quality and the computational cost of the regression scheme can be improved substantially in practice by transforming the representation so that it incorporates some degree of chemical intuition. For instance, the combination of multiple kernels corresponding to different interatomic distances has been shown to improve the quality of the ML model.[114] Likewise, a scaling of the weights of different atomic distances within an environment has been shown to be beneficial when using ML to predict atomic-scale properties.[104,115]

We will discuss how many of these modifications can be incorporated through inclusion of a rotationally-invariant Hermitian operator $\hat{U} = \hat{U}_1 = \hat{U}_2 = \ldots$ (as introduced earlier) that leads to coupling of the geometric and elemental components of the translationally-invariant atom-centered ket $|\rho_j\rangle$. For concreteness, and to provide a formulation that can be directly applied to an existing framework, we discuss $\hat{U}$ written in the orthonormal basis of radial functions and spherical harmonics $\{|anlm\rangle\}$, that correspond to the SOAP power spectrum.

The requirement that $\hat{U}$ is rotationally-invariant (and thus commutes with an arbitrary rotation operator) means that it must have the following form

$$\langle anlm|\hat{U}|a'n'l'm'\rangle = \delta_{ll'}\delta_{mm'}\langle anl|\hat{U}|a'n'l'\rangle\,. \tag{2.52}$$

Equation (2.52) is the most general form compatible with $SO(3)$ symmetry, and can be seen as a way to introduce correlations between different radial and elemental components of the features, and to weight the contribution from different angular channels.

2.3.1 Low-rank expansion of the $\hat{U}$ operator

Since $\hat{U}$ is Hermitian, it can be diagonalized and expressed in the orthogonal basis of its eigenkets $\{|\bar{J}\rangle\}$,

$$\hat{U} = \sum_J |\bar{J}\rangle\, U_J \langle\bar{J}|\,. \tag{2.53}$$

Taking $U_J \langle\bar{J}| \to \langle J|$ allows us to express $\hat{U}$ as $\hat{U} = \sum_J |\bar{J}\rangle\, \langle J|$.

The transformed SO(3) vector components can be written in terms of the components of $|J\rangle$ in the chemical basis, $u_{Janl} = \langle J|anl\rangle$. This yields

$$\langle JJ'|\overline{\rho_j^{\otimes\nu}}\rangle = \sum_{aa'nn'l} u_{Janl} u_{J'a'n'l} \sum_m (-1)^m \langle anlm|\rho_j\rangle \langle a'n'l\text{-}m|\rho_j\rangle\,. \tag{2.54}$$

By choosing a low-rank expansion of $\hat{U}$ one can greatly reduce the dimensionality of the $SO(3)$ fingerprint vector, similarly to what was done in Ref.[116] applying standard sparse decomposition techniques to the $SO(3)$ fingerprints.

A possible approach is to determine this low-rank approximation based on the correlations found between environments that are part of the data set. For a given l, consider the spherically-symmetric covariance matrix between the features of the expanded atomic density,[††]

$$C^{(l)}_{ana'n'} = \frac{1}{N}\sum_j\sum_m (-1)^m \langle anlm|\rho_j\rangle \langle a'n'l\text{-}m|\rho_j\rangle = \frac{\sqrt{2l+1}}{N}\sum_j \langle ana'n'l|\overline{\rho_j^{\otimes\nu}}\rangle\,. \tag{2.55}$$

The eigenvectors of $\mathbf{C}^{(l)}$, $\boldsymbol{v}_J^{(l)}$, can then be used as u_{Janl} in Eq. (2.54). It is easy to see that this transformation identifies components of the data that are linearly independent within the training set, and have a spread that is equal to the corresponding eigenvalues $\lambda_J^{(l)}$. The feature space can then be compressed by only retaining a certain number of components n_J that could be determined using the magnitude of the associated eigenvalues.

2.3.2 Radially-scaled representations

In a system with relatively uniform atom density, the coefficients of the representation $|\overline{\rho_j^{\otimes\nu}}\rangle$ are dominated by the region farthest from the center. This could be regarded as rather un-

---

[††] Note that, apart from a l-dependent scaling, the covariance matrix is just the average of the SOAP power spectrum over the training set.

physical, since interactions between atoms decay with distance and the closest atoms should therefore give the most significant contribution to properties, which is reflected in the observation that multi-scale kernels tend to perform best when very low weights are assigned to the long-range kernels.[89,109,114] This effect can be counteracted by multiplying the atomic probability amplitude Eq. (2.20) with a radial scaling $u(r)$,

$$\langle a\mathbf{r}|\hat{U}|\rho_j\rangle = u(r)\,\langle a\mathbf{r}|\rho_j\rangle\,. \tag{2.56}$$

In the context of the SOAP power spectrum, this change can be represented in terms of a $\hat{U}$ operator that reads

$$\langle n|\hat{U}|n'\rangle = \int \mathrm{d}r\, r^2 R_n(r) R_{n'}(r)\, u(r), \tag{2.57}$$

since an operator that scales states in the position representation must be diagonal in it,

$$\langle r|\hat{U}|r'\rangle = \delta(r-r')\,u(r)/rr', \tag{2.58}$$

and its matrix elements in the basis of radial basis functions are

$$\langle n|\hat{U}|n'\rangle = \int dr \int dr' r^2 R_n(r) R_{n'}(r')\delta(r-r')\,u(r), \tag{2.59}$$

which reduces to Eq. (2.57).

Radial scaling in the form of Eq. (2.56) can be approximated, when using narrow atom-centered functions, with $\sum_i u(r_{ij}) f_c(r_{ij}) h(\mathbf{r}-\mathbf{r}_{ij})$, where we also consider for simplicity the case with a single species.[117] Besides the fact that it is simpler to implement this form of scaling in an existing code, this approximation also makes apparent the connection between the general density-based framework we introduce here and the descriptors of Ref.[104]. When h is taken to be a Gaussian function of width σ, the weight on the central atom is set to zero and one considers the two-body invariant representations, this ansatz is essentially equivalent to the two-body features in Ref.[104]:

$$\begin{aligned}\langle r|\hat{U}|\overline{\rho_j^{\otimes 1}}\rangle &= \sum_{i\neq j} u(r_{ij}) \frac{\sqrt{2\pi}}{\sigma r_{ij}} \left[e^{-(r-r_{ij})^2/2\sigma^2} - e^{-(r+r_{ij})^2/2\sigma^2} \right] \\ &\sim \sum_{i\neq j} u(r_{ij}) \frac{\sqrt{2\pi}}{r_{ij}} e^{-(r-r_{ij})^2/2\sigma^2}.\end{aligned} \tag{2.60}$$

2.3.3 Alchemical kernels

In the presence of multiple species, one could make the scaling element dependent, or devise a more complex operator that couples different channels of different species. As a first test of the generalization of SOAP in the presence of multiple elements, we consider an operator in the form

$$\langle anlm|\hat{U}|a'n'l'm'\rangle = \delta_{ll'}\delta_{mm'}\delta_{nn'}\,\langle a|\hat{U}|a'\rangle\,, \tag{2.61}$$

which ignores couplings between the structure of an environment and the elements within it. One can always write a low-rank expansion of the operator, $\hat{U} \approx \sum_{Ja}|\overline{J}\rangle\, u_{Ja}\,\langle a|$, which allows one to write

$$\hat{U}\otimes\hat{U}\,|\overline{\rho^{\otimes 2}}\rangle = \sum_{aa'}|\overline{JJ'}\rangle\, u_{Ja}u_{J'a'}\,\langle aa'|\overline{\rho^{\otimes 2}}\rangle\,. \tag{2.62}$$

In the context of SOAP, one can define the projections of the power spectrum in this “alchemical basis”,

$$\langle JnJ'n'l|\overline{\rho^{\otimes 2}}\rangle = \sum_{aa'} u_{Ja}u_{J'a'}\sum_{m}(-1)^{m}\,\langle anlm|\rho_j\rangle\,\langle a'n'l\,\text{-}m|\rho_j\rangle\,, \tag{2.63}$$

which was shown in Ref.[117] to yield a substantial improvement in the learning efficiency in the presence of many chemical elements, and to result in a low-dimensional representation of elemental space that shares some similarities with the grouping found in the periodic table of the elements. One can see the relationship between these “alchemical features” and previous attempts to incorporate cross-species correlations through the generalized SOAP environmental kernel,

$$\int \mathrm{d}\hat{R}\,\Big[\,\langle\rho_j|\,\hat{U}^{\dagger}\hat{U}\hat{R}\,|\rho_k\rangle\Big]^2 = \sum_{JnJ'n'l}\langle\overline{\rho_j^{\otimes 2}}|JnJ'n'l\rangle\,\langle JnJ'n'l|\overline{\rho_k^{\otimes 2}}\rangle\,. \tag{2.64}$$

By writing out explicitly this inner product in terms of the full power spectrum elements $\langle ana'n'l|\mathcal{X}_j^{(2)}\rangle\,\hat{R}$ one can see that the matrix elements $\langle a|\,\hat{U}^{\dagger}\hat{U}\,|a'\rangle$ are nothing but the elements of the alchemical kernel $\kappa_{aa'}$ that was introduced in Ref.[75], where it was shown that taking $\kappa_{aa'} \neq \delta_{aa'}$ can improve property predictions with kernel ridge regression.[75,114] Off-diagonal couplings between chemical elements have also been used in other representations, including those of Ref.[104].

The expression in terms of reduced features Eq. (2.63) is, however, more efficient to compute and clarifies how this approach enables the introduction of correlations between elements, as well as reduction of the space dimensionality. The full SOAP feature vector contains a number of components that is proportional to the square of the number of present species n_{sp}, while limiting to a number $d_J \ll n_{\mathrm{sp}}$ of basis kets reduces the dimensionality of the feature vector by a factor $(n_{\mathrm{sp}}/d_J)^2$. Note that one does not even need to compute all the elements in the $|anlm\rangle$ expansion of the density, since the alchemical projection can be brought down to the level of the atom density, which can be defined for d_J chemical “channels” rather than for each element separately,

$$\langle J\mathbf{r}|\rho_j\rangle = \sum_{a} u_{Ja}\,\langle a\mathbf{r}|\rho_j\rangle\,. \tag{2.65}$$

Density-based representations that assign a weight to each species have been explored as means to reduce the complexity of ML representations in cases where many elements are present simultaneously,[118–120] which correspond essentially to the case with $d_J = 1$. For instance, the *compositional descriptor* of Ref.[118] is equivalent to Eq. (2.37) computed on a single invariant density,

$$\langle r|\overline{\rho_j^{\otimes 1}}\rangle = \sum_i u_{a_i}\delta(r - r_{ij}) f_c(r_{ij}), \tag{2.66}$$

where the weights of different species are rather arbitrarily set to be $u_a = 0, \pm 1, \pm 2 \ldots$. The more general formulation in Eqs. (2.63)-(2.65) provides a way to alter the dimensionality of the representation, and to optimize the projections to obtain the most efficient features for a given regression problem.

2.3.4 Multiple-kernel learning

We have shown that by manipulating the form of the SOAP kernel, e.g. by including a radial scaling, by introducing correlations between elements, or by adjusting other hyperparameters, such as the cutoff radius or the shape of the atomic Gaussian functions, it is possible to obtain different perspectives of the structural correlations, and to tune them to give the best possible performance in a regression task. As done in Ref.[114], one can build a composite kernel out of a selection of different models, i.e.

$$K(A,B) = \sum_{\aleph} w_{\aleph} K_{\aleph}(A,B). \tag{2.67}$$

This multiple-kernel model makes it possible to find the best combination of different representations of the atomic environments, using short and long-range, 2 and 3-body, radially-scaled and alchemically-contracted terms. In a Gaussian Process Regression language, each model is meant to contribute $\sqrt{w_{\aleph}}$ to the variance of the target property. The weights can be set manually based on an intuitive understanding of how they contribute to a property, or – more simply – optimized by cross-validation (see Section 4.1.3). Note that such combined kernels can still be seen as an explicit inner product between representations. In other words, taking sums of multiple kernels can be interpreted equivalently as generalizations of kernels, or as generalizations of representations that take the form

$$|X\rangle = \sqrt{w_1}\,|X^1\rangle \oplus \sqrt{w_2}\,|X^2\rangle \oplus \ldots, \tag{2.68}$$

where $\oplus$ denotes concatenation.

2.3.5 Non-factorizable operators

In order to relate Eq. (2.28) to other density-based representations that involve more complicated scaling functions of the internal coordinates, it is necessary to introduce a further linear transformation $\hat{U}^{(\nu)}$ which does not factorize into components that act independently on each term in the ν-order tensor product. Such an operator must be chosen with care to ensure that it is rotationally-invariant, otherwise the rotational-invariance of the transformed ket will be lost. As far as the $\nu = 2$ rotationally-invariant kets are concerned, a generic operator is completely determined by its action on the basis vectors $\{|a\mathbf{r}a'\mathbf{r}'\rangle\}$. Rotationally-invariant operators must act on $|a\hat{R}\mathbf{r}a'\hat{R}\mathbf{r}'\rangle$ in the same was as on $|a\mathbf{r}a'\mathbf{r}'\rangle$, followed by the rotation $\hat{R}$. The upshot of this observation is

$$\langle a_1\mathbf{r}_1 a'_1\mathbf{r}'_1|\hat{U}^{\otimes 2}|a_2\mathbf{r}_2 a'_2\mathbf{r}'_2\rangle = \langle a_1 r_1 a'_1 r'_1\omega_1|\hat{U}^{\otimes 2}|a_2 r_2 a'_2 r'_2\omega_2\rangle\,, \tag{2.69}$$

i.e. any non-internal coordinate must be cyclic. If a distance and angle-based scaling is required, then the operator is diagonal,

$$\langle a_1\mathbf{r}_1 a'_1\mathbf{r}'_1|\hat{U}^{\otimes 2}|a_2\mathbf{r}_2 a'_2\mathbf{r}'_2\rangle = \delta_{a_1a_2}\delta_{a'_1a'_2}\delta(r_1-r_2)\delta(r'_1-r'_2)/r_1^2 r_1'^2\delta(\omega_1-\omega_2)u(a_1,r_1,a'_1,r'_1,\omega_1). \tag{2.70}$$

For example, the scaling function in the three-body descriptor in Ref.[104] corresponds to the following choice for $u(r_1,r_2,\omega)$,

$$u(r_1,r_2,\omega_1) = \frac{1-3\omega_1\omega_2\omega_3}{(r_1r_2r_3)^n}, \tag{2.71}$$

where $r_3^2 = r_1^2 + r_2^2 - 2r_1r_2\omega_1$, $\omega_2 = (r_1^2 - r_2^2 - r_3^2)/2r_2r_3$, $\omega_3 = (r_2^2 - r_1^2 - r_3^2)/2r_1r_3$ and n is an adjustable parameter. Faber *et al.* do not specify a scaling function for four-body and higher-body descriptors, but the analysis presented here clearly extends to any hypothetical scaling function that involves the internal coordinates of a collection of $\nu+1$ positions.

Starting from the SOAP power spectrum, one can exploit the fact that each component is separately symmetry invariant. It is then possible to introduce an arbitrary linear operator coupling the $|a_1na'_1n'l\rangle$ components, $\langle a_1n_1a'_1n'_1l_1|\hat{U}|a_2n_2a'_2n'_2l_2\rangle$. Being a linear operation, this transformation amounts to a change of regularization for the ridge regression problem, and is most useful if applied to reduce the dimensionality of the feature vectors. This can be done e.g. by finding the principal components of the covariance matrix of the SOAP power spectrum or – as done in Ref.[116] – by a sparse decomposition that singles out a subset of the components that suffice to obtain a thorough description of the relevant structures. This corresponds to the contracted representation

$$\langle J|\overline{\rho_j^{\otimes 2}}\rangle = \sum_{Jk} u_{Jk}\langle a_kn_ka'_kn'_kl_k|\overline{\rho_j^{\otimes 2}}\rangle\,, \tag{2.72}$$

where k runs over the set of selected components,[‡‡] which can be determined with different schemes, from a CUR decomposition[121] to farthest point sampling.[122,123] The coefficients u_{Jk}

[‡‡] The sparsification could be also represented explicitly by a $\hat{U}$ operator, that zeroes out all of the unnecessary components.

are the elements of a square matrix that ensures the contracted vectors in Eq. (2.72) generate a kernel that is as close as possible to the full kernel.

2.3.6 Optimization of the density representation

The optimization of the $\hat{U}$ operator in its more general form (see Eq. (2.52)) involves a large number of parameters, leading to a very concrete risk of overfitting. This is exacerbated by the fact that the feature vector is then used as the input for regression, and one has to balance the amount of data used to optimize the elements of $\hat{U}$ and that used for the training of the ridge regression model. The simplest approach to reduce the optimization of $\hat{U}$ to a small number of free parameters uses the compression method discussed in Section 2.3.1 to identify the most important combinations of $\langle anlm|\rho_j\rangle$ components that are linearly independent for the data at hand. We would like to be able to optimize based on the correlations found between environments that are part of the training set. The idea is that further optimization using target properties will be less likely to overfit after this dimensionality reduction.

Another possible use of the principal-component representation of $\langle anlm|\rho_j\rangle$ is to obtain a simpler ansatz to further optimize the $\hat{U}$ operator. For instance, one could combine linearly the different components using the $\hat{U}$ operator defined in Eq. (2.55)

$$\langle IJlm|\rho_j\rangle = \sum_{I'aJ'n} f^{(l)}_{II'JJ'} u^{(l)}_{I'aJ'n} \langle anlm|\rho_j\rangle\,, \tag{2.73}$$

where the scaling coefficients $f^{(l)}_{II'JJ'}$ are determined so as to make the representation better suited to build a regression model for the target property y. A systematic exploration of the different possibilities, as well as their benchmarking on different regression problems, is left for future work.

3 Unsupervised ML[†]

The large databases of structures and properties that result from computational searches, as well as the agglomeration of data of heterogeneous provenance, leads to considerable challenges when it comes to navigating the database, representing its structure at a glance, understanding structure-property relations, eliminating duplicates and identifying inconsistencies. In order to automate these tasks a number of different unsupervised learning (UL) algorithms have been developed, or adapted to the specific requirements of this field.[41,42,60,73,74,124–126] A fundamental ingredient in all of these approaches is the need to define distances over chemical space. As mentioned in Chapter 2 many options are available, with different levels of complexity and generality, starting from the commonly used Root Mean Square (RMS) distance. In order to deal with symmetry operations or condensed phase structures, several representation and "fingerprint" frameworks have been developed,[39,45,48,66,127–135] that assign a unique vector of order parameters to each molecular or crystalline configuration. Representations of the atomic structure are a solid foundation to build metrics comparing materials by taking some norm of the difference between feature vectors. The dissimilarity, i.e. distance, between the N atomic configurations in a database contains a large amount of information on the structural relations between the database items. However, this information is not readily interpretable, as it is encoded as a N^2 matrix of numbers. Any of these distances could be taken as the basis of the clustering and dimensionality reduction algorithms the main methods of UL that have been used on atomistic data. Both of these classes of techniques aim at discovering patterns or structures in an unlabeled dataset $\mathcal{D}$. In most cases, this can be thought of as learning a probabilistic model from $\mathcal{D}$. Dimensionality reduction involves building a low-dimensional "map", where each point corresponds to one of the structures in the database and where the (Euclidean) distances between points represent the information on the pairwise dissimilarity

---

[†] This chapter has been adapted from the journal article [83] [84] whose authors are respectively Sandip De, **Félix Musil**, Teresa Ingram, Carsten Baldauf, and Michele Ceriotti and **Félix Musil**, Sandip De, Jack Yang, Joshua E. Campbell, Graeme Matthew Day, and Michele Ceriotti. The author of this thesis built the unsupervised models, the dendrogram visualizations, made all the figures and contributed to the writing of [83] while he analyzed the data, built and benchmarked the supervised and unsupervised ML models and wrote the methods, results and discussions sections of [84]. Since Ref. [84] combines both supervised and unsupervised ML models, it has been split in Section 4.2.2 and here to follows the logic of this thesis. Please refer to Section 4.2.2 for the the details, results and discussion of the supervised learning side of this article.

matrix. Several methods have been proposed over the years to solve this dimensionality reduction problem, starting from principal component analysis[136] and the equivalent linear multi-dimensional scaling,[137] and proceeding to non-linear generalizations of the idea, such as ISOMAP,[138] diffusion maps,[139] kernel PCA.[140] An alternative approach to navigate a set of structures based on the dissimilarity matrix is to use clustering algorithms, that identify groups of objects having similar properties to hint at the presence of recurring motifs underlying the behavior of the system. A considerable number of clustering algorithms have been developed over the last few decades,[141–143] including connectivity models[144] (i.e. hierarchical clustering, centroid models[145–147] (i.e. k-means algorithm) and density based models.[42,148,149]

In Section 3.1, we present the regularized entropy matching (REMatch) kernel[75] based on SOAP features and the relation between kernels and distances. We introduce the sketch-map dimensionality reduction technique[60] and the HDBSCAN* clustering method[150] in Section 3.3 as tools to address the challenges of navigating a database of molecular conformers in Section 3.4, checking its internal consistency and rationalizing structure-property relations, and to develop a data-driven classification scheme that provides useful insight into the packing motifs in datasets of organic semiconductors in Section 3.5.

3.1 REMatch Kernel and Distances

A kernel function $K(\cdot,\cdot)$ between atomic configurations A and B measures their similarity. When this function is positive definite, it defines an inner product between vectors in a Hilbert space, i.e. $K(A,B) = \langle A|B\rangle$.[151] The power of kernel methods lays in their ability to transform a low-dimensional non-linear problem into a ‘more linear’ problem in a higher-dimensional space.[88,152] The ‘kernel trick’ allows transforming simple linear models into fully non-linear ones at the cost of defining this similarity measure.

In the case of local representations such as the SOAP power spectrum (see Section 2.2.4), the inner product between normalized feature vectors

$$k(A_i,B_j) = \langle A;\overline{\rho_i^{\otimes 2}}|B;\overline{\rho_j^{\otimes 2}}\rangle^{\zeta}\,, \tag{3.1}$$

defines an element $\boldsymbol{k}_{ij}(A,B)$ of the similarity matrix between local environments of the two structures. Contrary to the additive global similarity measure often associated with regression models (see Section 2.1.4), the REMatch kernel[75] measures the structural similarity by combining the local similarity measures to highlighting the pairs of local environments that exhibit the highest degree of structural similarity. For this purpose, the similarity between structure A and B is given by the weighted sum over the elements of $\mathbf{k}(A,B)$ where the weights are evaluated using a technique borrowed from optimal transport theory,[153]

$$\begin{aligned}&\hat{K}_\gamma(A,B) = \left[\operatorname{Tr}\mathbf{P}_\gamma\mathbf{k}(A,B)\right],\\&\mathbf{P}_\gamma = \underset{\mathbf{P}\in\mathcal{U}(N,N)}{\operatorname{argmin}}\sum_{ij} P_{ij}\left(1-k(A_i,B_j)+\gamma\ln P_{ij}\right).\end{aligned} \tag{3.2}$$

The optimal combination is obtained by searching the space of doubly stochastic matrices $\mathcal{U}(N, M)$ using the Sinkhorn algorithm[154] which minimizes the discrepancy between matching pairs of environments, regularized using the information entropy of the weight matrix $E(\mathbf{P}) = -\sum_{ij} P_{ij} \ln P_{ij}$; ζ affects the sensitivity of the kernel and γ enables switching between a strict and broad selection of best matching pairs of local environment (see Ref.[75] for more detail). Once a kernel between two configurations has been defined, it is then possible to introduce a kernel-induced distance[155]

$$D(A,B)^2 = \hat{K}_\gamma(A,A) + \hat{K}_\gamma(B,B) - 2\hat{K}_\gamma(A,B). \tag{3.3}$$

that can be used as the metric for clustering or dimensionality reduction.

3.2 Sketch-map

A widely adopted strategy used to perform dimensionality reduction is to find the low dimensional Cartesian projection that best reproduces the pairwise distances in the high dimensional space. Such an approach, called multi-dimensional scaling (MDS), is useful to represent high-dimensional data such as distance matrices. Sketch-map[60,123,126] is a particular non-linear MDS algorithm in which one iteratively optimizes the objective function

$$S^2 = \sum_{ij} \left[F\left[D(X_i, X_j)\right] - f\left[d(x_i, x_j)\right]\right]^2, \tag{3.4}$$

that measures the mismatch of the dissimilarity between atomic configurations $D(X_i, X_j)$ with the dissimilarity (typically just the Euclidean distance) between the corresponding low-dimensional projections $\{x_i\}$. The procedure is very similar to multi-dimensional scaling, except for the appearance of the transformations F and f, which are non-linear sigmoid functions of the form:

$$F(r) = 1 - (1 + (2^{a/b} - 1)(r/\sigma)^a)^{-b/a}. \tag{3.5}$$

The non-linear transformation focuses the optimization of Eq. (3.4) on the most significant distances (typically those of the order of σ), and disregards local distortions (e.g. induced by thermal fluctuations or by incomplete convergence of a geometry optimization) and the relation between completely unrelated portions of configuration landscape. The maps that we report in this work will be labeled synthetically using the notation σ`-A_B-a_b`, where A and B denote the exponents used for the high-dimensional function F, a and b denote the exponents for the low-dimensional function f, and σ the threshold for the switching function. The choice of these parameters of the sigmoid functions are discussed in detail elsewhere.[123] In practice, A, B, a, and b have a relatively small effect on the projection and can be optimized and kept fixed for systems belonging to the same family. Since the structures we consider here are minimum-energy configurations, and there are no thermal fluctuations that should be filtered out, we set $A = a = 1$ (so that at short range the algorithm will still try to represent distances faithfully) and set the long-range exponents to $B = b = 4$. The parameter σ is the one to which sketch-map is most sensitive, and needs to be tuned for each system separately. To automate the process of building sketch-maps of a large number of subsets of the database, we have used a simple heuristic procedure for determining the value of σ automatically. Following the prescriptions in Ref.[123], we first compute the histogram of distances in the dissimilarity matrix of each molecular set and detect the dissimilarity value (D_{max}) corresponding to the peak value of the histogram. We then set the value of σ to $0.8D_{max}$.

3.3 Hierarchical clustering and HDBSCAN*

Cluster analysis can be achieved through various strategies ranging from distribution models such as Gaussian mixture models to centroid models like k-means. Clustering models based on connectivity information such as hierarchical (or agglomerative) clustering[144] are particularly suited for the identification of recurring motifs underlying the behavior of the system. Starting from each configuration as its own cluster, the hierarchical clustering algorithm iteratively aggregates clusters together based on some assessment of their distance. The distance between two *clusters*, however, can be defined in many different ways. One possible choice is the RMS dissimilarity between the pair of members of the two clusters. The linkage distance Δ between two clusters $\mathbb{X}$ and $\mathbb{Y}$ is then defined as:

$$\Delta(\mathbb{X},\mathbb{Y}) = \sqrt{\frac{1}{N_{\mathbb{X}} N_{\mathbb{Y}}} \sum_{X\in\mathbb{X}, Y\in\mathbb{Y}} D^2(X,Y)}, \tag{3.6}$$

where $N_{\mathbb{X}}$ and $N_{\mathbb{Y}}$ are the total number of configurations within each cluster. The results of hierarchical clustering can be further processed to identify groups of structurally homogeneous atomic structures and unique configurations within the database as done in Section 3.4. The 'dendrogram' plot conveys visually the sequence of agglomerative clustering operations and the linkage distance at each step. The lowest level of the dendrogram is composed of single-structure clusters so that the x axis corresponds to individual configurations sorted according to the clustering procedure. Each merge operation is represented by a line joining the two underlying clusters, with the y position of the line representing the linkage distance for that pair, as defined by Eq. (3.6). In this kind of representation, at the bottom of the dendrogram, each structure can be thought of as an individual cluster containing only one item. Clusters are then merged iteratively, selecting at each step the pair of clusters that are closest to each other. This operation is repeated until all the clusters collapse into one single group that encompasses all the structures in the database, thus completing the dendrogram. To avoid overcrowding the bottom of the plot, one can hide the part that corresponds to very small linkage distances, while still graphically visualizing the size of the clusters by drawing bars that encompass the associated structures. Since the "leaves" of this dendrogram correspond to individual configurations, it is possible to complement the dendrogram with color-coded bar plots that represent the value of different properties of each structure, thereby giving a clear picture of the relation between structural clustering and the different properties.

In order to understand the basic motifs of a particular cluster $\mathbb{X}$, it is very useful to select one of its structures that is as representative as possible of the entire subset. In the case where stability estimates are available, such structure may be the lowest-energy structure in the cluster. For a definition that is based purely on conformational or configurational information, the most representative structure $RS(\mathbb{X})$ could be defined, as the item having the minimum mean square dissimilarity with respect to all other members of $\mathbb{X}$, i.e.

$$RS(\mathbb{X}) = \underset{X_1 \in \mathbb{X}}{\arg\min} \left[\frac{1}{N_{\mathbb{X}}} \sum_{X_2 \in \mathbb{X}} D^2(X_1, X_2) \right]. \tag{3.7}$$

Representative structures can be defined at each level of the hierarchy, and can therefore be very useful in navigating the database, and understanding what are its most crucial structural features. The spread of the cluster around $RS(\mathbb{X})$,

$$\sigma_D(\mathbb{X}) = \sqrt{\frac{1}{N_{\mathbb{X}}} \sum_{X \in \mathbb{X}} D^2(X, RS(\mathbb{X}))}, \tag{3.8}$$

can be used to quantify the range of structural landscape that is covered by the cluster. Combined with the dendrogram plot, such metric can help identify groups of structures characterized by similar structural patterns.

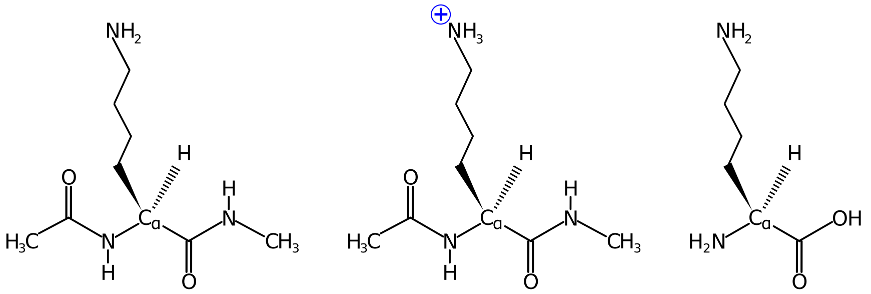
Another important aspect of database analysis is ‘outlier detection’.[156–161] An “outlier” configuration is defined as a configuration that is different from most of the samples in the database. Outlier configurations are very important as they are likely to have a unique structural motif in the whole database and are thus interesting for structure prediction applications. They also could represent chemical changes or indicate inconsistent configurations which are likely to be “errors” in the database. They can often be identified in the dendrogram plot as small clusters that are very different from their neighboring clusters.

Unfortunately, the analysis described above is quite time-consuming. A more systematic technique to identify automatically the main structural motifs is the HDBSCAN* algorithm[150] which has been applied in Section 3.5. This method introduces an adaptive density threshold, in a similar fashion as density-based clustering algorithms, to group together samples belonging to dense regions of the dataset. Indeed, HDBSCAN* has a single intuitive hyper-parameter, the minimal size of a cluster, which can be set to approximately 1% of the dataset size to discard configurations that belong to sparsely-populated regions that do not correspond to a recurring structural motif. Since such configurations are different from most of the samples in the database it automatically provides a list of “outlier” candidates.

3.4 Mapping and classifying molecules from a high-throughput structural database

This analysis focuses on a collection of *ab initio* datasets containing conformers of twenty proteinogenic amino acids and dipeptides, as well as their interactions with a series of divalent cations along with their calculated energy (Ca^{2+}, Ba^{2+}, Sr^{2+}, Cd^{2+}, Pb^{2+}, Hg^{2+}).[85] The potential-energy surfaces (PES) of 280 systems were explored using replica exchange molecular dynamics and selecting conformers up to 4 eV (390 kJ/mol), summing up to an overall of 45,892 stationary points on the respective potential-energy surfaces.[86] The underlying energetics were calculated by applying density-functional theory (DFT) in the generalized gradient approximation corrected for long-range van der Waals interactions (PBE+vdW).[162–164] A number of theory-theory and theory-experiment comparisons have shown the applicability of the method to amino acid and peptide systems.[86,165–170] The generation of this dataset involved significant manual intervention, and one would expect it to be an easy starting point for studying materials and molecules across chemical space.[171] Nevertheless, we will demonstrate that, even for such heavily curated data, automated techniques are needed to extract trends and to check for internal consistency.

In the following text we focus on the amino acid lysine (in short Lys) and investigate basic structural motifs of three forms, see Fig. 3.1. Furthermore, the UL techniques introduced in Sections 3.1 to 3.3 are used to detect the impact of perturbations (here Ca^{2+} cations) on the structural properties of the unperturbed systems. Finally, we demonstrate how the approach can also be applied to discover inconsistencies and outliers in the database.



(a) Lys dipeptide **(b)** LysH dipeptide **(c)** bare Lys

Figure 3.1 – The lysine building block was studied in three forms: (a) uncharged dipeptide, (b) protonated dipeptide, and (c) uncapped and uncharged amino acid.

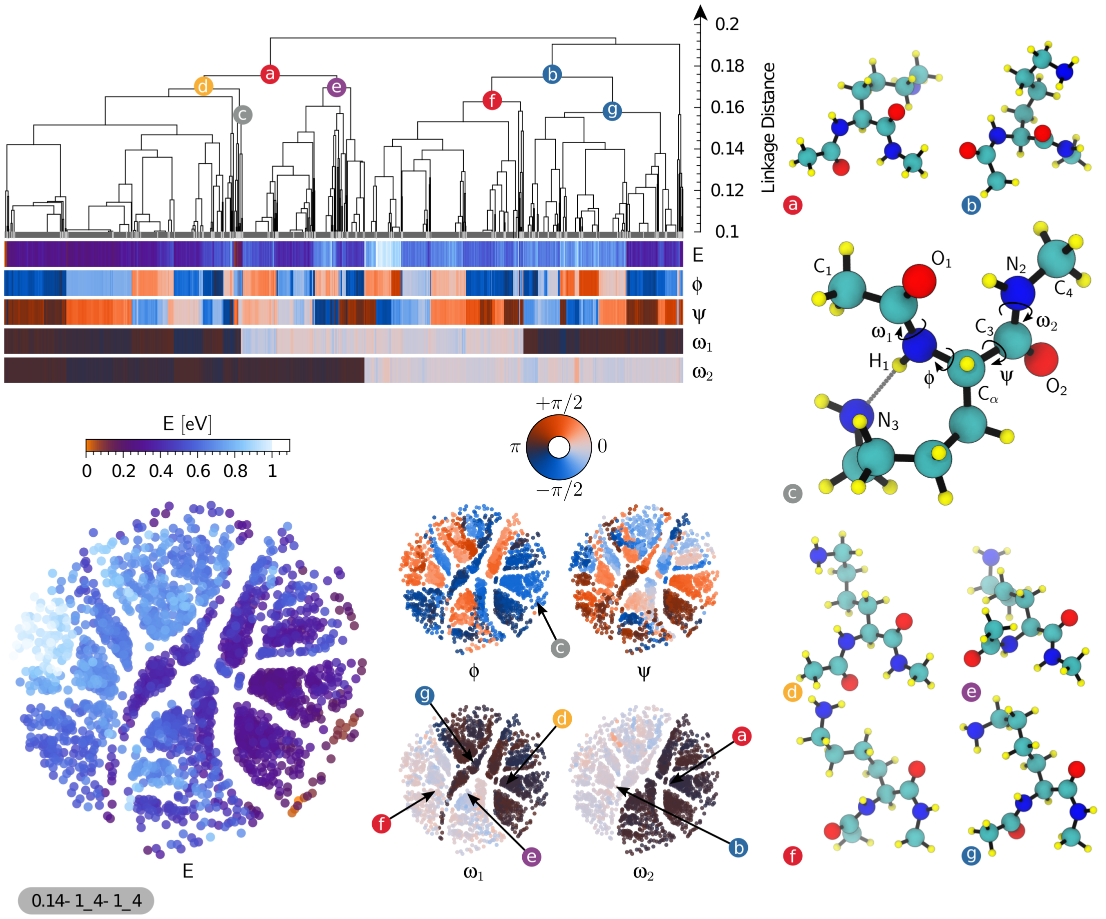


Figure 3.2 – Representation of the similarity matrix corresponding to the lysine dipeptide dataset using the agglomerative clustering algorithm (top) and the sketchmap algorithm (bottom, projection parameters shown following the scheme $\sigma\text{-}A\_B\text{-}a\_b$). A few representative structures (see Eq. (3.7)) of interesting clusters are shown (right) and their corresponding position on the sketch-maps and dendrogram representation is highlighted. The five sketch-maps are colored according to the conformational energy and the backbone dihedral angles ϕ, ψ, ω_1 and ω_2. The dendrogram shows the clustering hierarchy of the structures of the dataset. Each structure is vertically aligned with its properties shown using color bars below the dendrogram. The dendrogram is cut at a linkage distance of 0.1 since structural properties are very similar below this threshold, and the clusters that are merged at this level are shown as thick gray bars separated by light-gray lines. Clusters composed of only one structure are drawn as a black line reaching the bottom of the dendrogram. The main structural motifs of this set of structures are governed by the peptide bond dihedral angles ω_1 and ω_2. The two main clusters (a) and (b) are showing a global correlation with the angle ω_2 while the angle ω_1 splits them into two well correlated sub-clusters (d), (e) and (f), (g) respectively. The cluster (c) is highlighted as an example containing 'outlier' structures of low conformational energy.

3.4.1 Finding the Dominant Features of a Structual Landscape

Lysine Dipeptide

We take as our first example a subset of the database containing 2080 conformers of lysine dipeptide. we start by constructing the distance matrix using the SOAP-REMatch kernel. In Fig. 3.2 the dendrogram plot as well as sketch-maps have been shown along with five properties, energy and four dihedral angles, using the same color scales in both the sketchmap and dendrogram representations. In the sketchmap each circular 'disk' represents a conformer. Whereas in the case of the dendrogram plot, structures are represented by vertical lines at the bottom of the plot. The strong correlation between energy and conformational parameters on one side, and clustering and position on the map on the other, testifies how the the REMatch-SOAP kernel induces a meaningful classification of the structures in this dataset.

While both clustering and sketchmap show clearly that the dataset is composed of groups of structurally-related conformers, the agnostic nature of the underlying metric does not disclose immediately the structural features that most transparently differentiate between different clusters. Comparing the representative structures from the main clusters allowed us to quickly identify candidate structural motifs that could be used to rationalize the layout of the conformational landscape. By color-coding the dendrogram and the sketch-maps according to these indicators one can readily highlight the key correlations.

When considering existing literature on the stability of oligopeptides, the two structural parameters that are most often considered as the key coordinates to navigate the conformational landscape are the Ramachandran dihedral angles ϕ and ψ, that determine the structure of the backbone around the side chain bearing C_α atom[56] under the assumption of peptide bonds being solely in *trans* conformation. While fine-grained clusters are homogeneous with respect to the ϕ and ψ angles, it is clear that for the present systems the clear-cut branching at the top of the dendrogram is determined by some other order parameter. An analysis of the representative structures for the two main clusters (a) and (b) shows that the two molecules differ by the isomerization of the N-terminal peptide bond. Further splitting of these two clusters, i.e. (a) into clusters (d) and (e), and (b) into (f) and (g), depends on the isomerization of the C-terminal peptide bond. We can confirm this attribution of the main features of the dataset by color-coding the map and the dendrogram following the dihedral angles ω_1 and ω_2. The four main clusters are largely homogeneous with respect to peptide bond isomerization, and are then further subdivided based on ϕ and ψ. This observation deserves some further comment. Peptide bonds in naturally-occurring proteins are believed to almost exclusively exist in *trans* conformation with the exception of prolyl peptide bonds where a smaller energy difference to *trans* increases the chance for *cis* conformers.[172,173] This view is supported by the analysis of protein structures deposited in the protein databases where *cis* conformations are found for about 5% of the prolyl peptide bonds, but less than 0.1% for the others.[174] X-ray crystallographic structure represent however merely frozen snapshots of structural dynamics. The *ab initio* structure search protocol, instead, does consider the peptide bond torsions as variable and intentionally allowed simulations to overcome the isomerization barrier. Consequently, the dataset contains representatives of all four combinations of *cis* and *trans* conformers. Since these transitions are strongly bimodal, and reflect in significant changes of the favorable side chain conformations, they constitute the most significant feature to classify the conformers. As expected, the most stable conformers are largely in a *trans*-*trans* conformation. However, the large parts of conformational space of that is occupied by conformers with 1 or 2 *cis* peptide bonds suggests that *cis* isomers might play a role in the dynamics of peptides and proteins. Consequently, an analysis only focused on the Ramachandran dihedrals,ϕ and ψ, would have missed one of the main features of the structural landscape that is critical to characterize the relation between structure and energetics. One could then proceed further with the analysis, focusing for instance on small clusters containing low-energy structures such as that represented by the conformer (c). All the structure in this group are *trans*-*trans* isomers, that in addition have $\phi \approx -90$ degrees and $\psi \approx 90$ degrees, allowing for the formation of a H-bond between the side chain N_3 and H_1, and a favorable arrangement of the N_2 donating a H-bond to the carbonyl O_1 as shown in Fig. 3.2. Having access to the combined information on energetics, and on the grouping of structures with similar geometry makes it easier to rationalize the energy ordering of the structures, without having to separately juxtapose all the low-lying conformers but focusing on a few representative structures.

Protonated Lysine Dipeptide

As the second example we considered a dataset containing 897 conformers of gas phase protonated lysine dipeptide. We follow the same steps as described in the previous example in order to find the most basic structural motifs of this system. Figure 3.3 shows the dendrogram, the sketchmap and a few color coded properties of this system to show their correlation with the classification. The most prominent feature for this molecule, which is evident in both the dendrogram and the sketch maps, is the presence of a group of outliers, that are clearly separated from the bulk of the conformers. Inspection of the cluster centroid (g) clarifies the structural basis of this separation. Whereas in most of the structures the excess charge lies on the lysine side chain as a NH_3^+ group, conformers in this cluster experienced a proton transfer event, with the excess proton attached to one of the carbonyl oxygen O_1, stabilized by H-bonding to N_2. This is a result of the database generation where *ab initio* replica-exchange molecular dynamics including high T trajectories where used for structure sampling during which protons can eventually transfer.

Moving on to the main cluster of structures, we can see that similar to our previous example of the neutral dipeptide and again due to the unbiased sampling protocol and the high energy range the peptide bond angles are again more important than Ramachandran's dihedrals. Conformers (a) and (b) are the representative structure for groups having *cis* and *trans* ω_2 peptide bonds respectively. Group (a) is further split based on the *cis/trans* state of ω_1 into the clusters represented by structures (d) and (e).

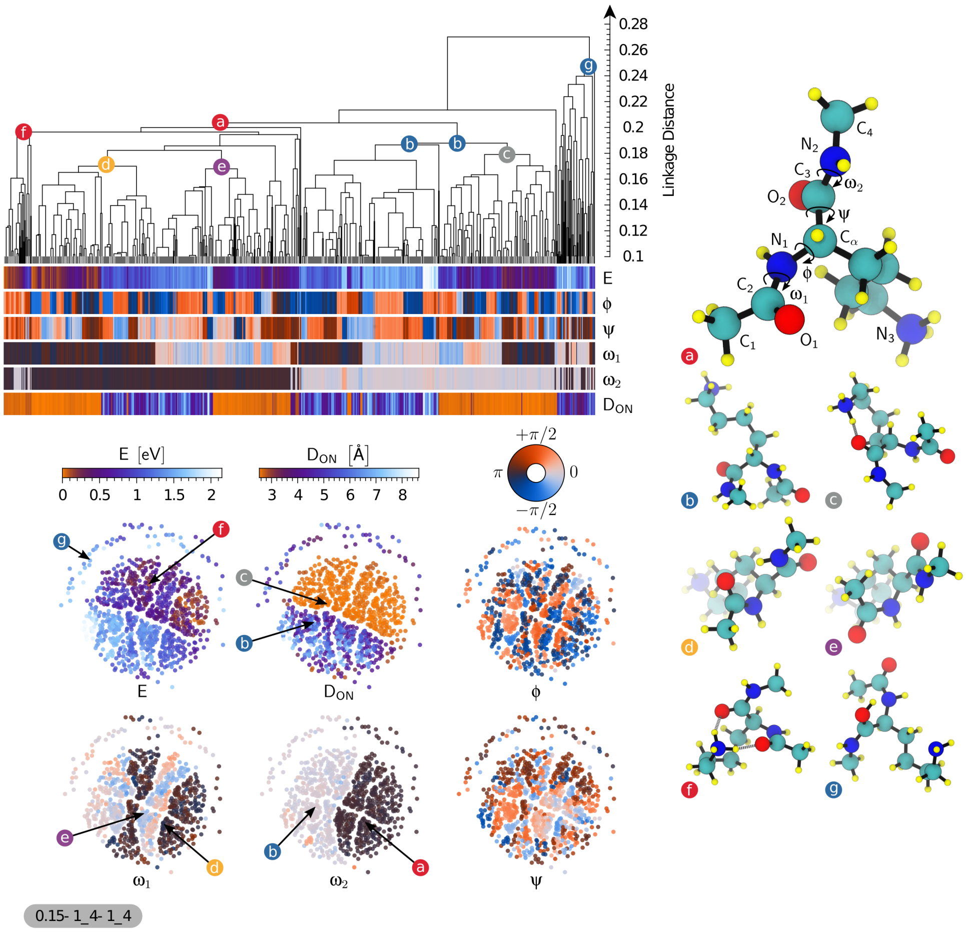


Figure 3.3 – Representation of the similarity matrix corresponding to the protonated lysine dipeptide dataset using the agglomerative clustering algorithm (top) and the sketchmap algorithm (bottom, projection parameters shown following the scheme σ-A_B-a_b). A few representative structures (see Eq. (3.7)) of interesting clusters are shown (right) and their corresponding position on the sketch-maps and dendrogram representation is highlighted. The six sketch-maps are colored according to the conformational energy, the minimal distance between O_1 or O_2 with N_3 called D_{ON}, and the backbone dihedral angles ϕ, ψ, ω_1 and ω_2. The dendrogram shows the clustering hierarchy of the structures of the dataset. Each structure is vertically aligned with its properties shown using color bars below the dendrogram. The dendrogram is cut at a linkage distance of 0.1 since structural properties are very similar below this threshold, and the clusters that are merged at this level are shown as thick gray bars separated by light-gray lines. Clusters composed of only one structure are drawn as a black line reaching the bottom of the dendrogram. The main structural motifs of this set of structures are governed by the dihedral angles ω_1 and ω_2 and the distance D_{ON}. The two main clusters (a) and (b) are showing a global correlation with the angle ω_2 while the angle ω_1 splits them into well correlated sub-clusters (e.g. sub-clusters (d) and (e)). The other important sub-clustering parameter is the distance D_{ON}, e.g. sub-clusters (c) and (b), which also correlates well with the separation between low and high conformational energy shown on the sketch-maps. Two sub-clusters are particular: (g) is a clear ‘outlier’ due to a chemical change and (f) features a H-bonding pattern with the side chain NH_3^+ pointing to both carboxy groups that sets this cluster apart from all others.

The presence of a charged side-chain leads to stronger H-bonds. As a consequence, peptide-bond isomerism plays a less crucial role in determining structural clustering than for the neutral dipeptide. An example of the importance of H-bonds is given for instance by the subcluster represented by conformer (f), in which the bent side chain leads to the formation of two H-bonds between NH_3^+ group and the carbonyl oxygens. H-bonds also dominate the partitioning of cluster (b), that is split into two groups – one of which is still best represented by the same conformer, and one that is epitomised by (c). Once again, inspection of these structural representatives reveals the organising principle behind the classification: (c)-like structures have an extended side chain, and are dominated by interactions among the peptide bond moieties, whereas (b)-like structures have a well-formed H-bond between the side chain and one of the two backbone O atoms. This structural pattern can be emphasized by color-coding conformers based on the parameter $\mathsf{D}_{\mathsf{ON}} = \min\left[d(\mathrm{O}_1,\mathrm{N}_3); d(\mathrm{O}_2,\mathrm{N}_3)\right]$: A small O-N distance indicates bending of the side chain and the formation of a H-bond between O and N. As it is clear from the sketchmap representation, there is a very strong correlation between the bending of the charged side chain and the energy of a conformer. All of the structures within 0.5 eV of the ground state feature this sidechain to backbone H-bonds.

It is worth noting that the importance of intramolecular H-bonds is a consequence of the gas-phase environment in which the structure search was performed. In a polar solvent like water, where intramolecular H-bonds that introduce strain compete with H-bonds with the surrounding water molecules, that do not require a bending of the side-chain, the energy balance might be different or less clear-cut. The analysis techniques we introduce in this work would be ideally suited to rationalize the changes in the (free) energetics of biological molecules when moving from the gas phase to (micro)solvated environments or to organic/inorganic interfaces.

Uncapped Lysine

Our third example is a dataset containing 733 conformers of the un-capped lysine molecule in the gas phase. We follow the same steps as described in the previous examples to construct the dendrogram shown in Fig. 3.4. The map has a simple structure, with few well-separated groups. Being a smaller molecule with fewer degrees of freedom, the Ramachandran angles are not defined. Still, the dihedral angles in the vicinity of the C_α atom display local structural correlation but once again they are not the main organizing factor that can rationalize the clustering. By juxtaposing representative conformers from the main clusters we could identify a better order parameter, that correlates strongly with H-bond patterns within the molecule. Namely, the distance (D_H) between the H atom in the OH group of the carboxyl function and the N atom in the backbone (N_1) discriminates well between structures based on H-bonding patterns[171] of *type I* between $N_1H \rightarrow O_2$ (e.g. conformer (b)) and of *type II* with a H-bond $O_1H \rightarrow N_1$ (e.g. conformer (a)). It can be seen from both the dendrogram and the sketch-maps that one could identify several subgroups based on particular values of D_H, representing specific orientations. Conformers (c) and (d) represent small groups of conformers having

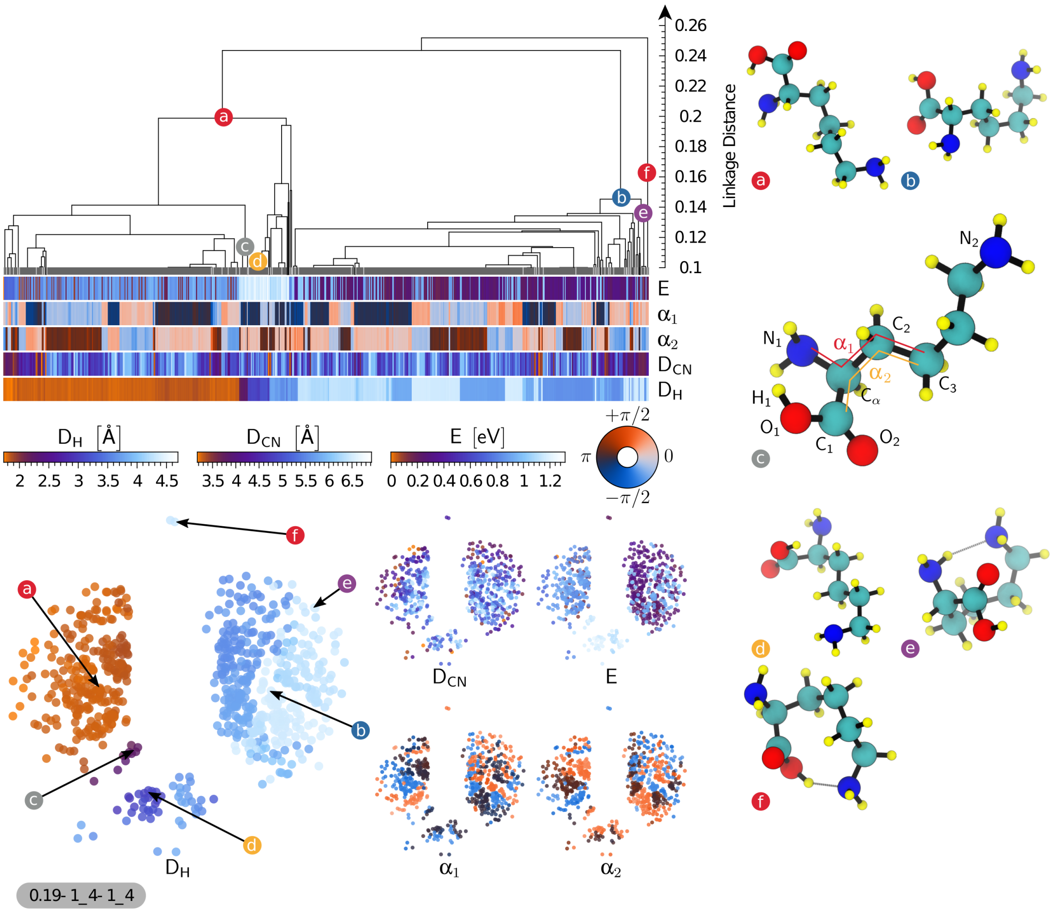


Figure 3.4 – Representation of the similarity matrix corresponding to the lysine uncapped dataset using the agglomerative clustering algorithm (top) and the sketchmap algorithm (bottom, projection parameters shown following the scheme σ-A_B-a_b). A few representative structures (see Eq. (3.7)) of interesting clusters are shown (right) and their corresponding position on the sketch-maps and dendrogram representation is highlighted. The five sketch-maps are colored according to the conformational energy, the distance between N_1 and the hydrogen in the carboxilic group H_1 (labelled $\mathsf{D_H}$), the distance between N_2 and C_α (labelled $\mathsf{D_{CN}}$), and the dihedral angles α_1 and α_2 which are respectively computed with the following atoms (N_1,C_α,C_2,C_3) and (C_1,C_α,C_2,C_3). The dendrogram shows the clustering hierarchy of the structures of the dataset. Each structure is vertically aligned with its properties shown using color bars below the dendrogram. The dendrogram is cut at a linkage distance of 0.1 since structural properties are very similar below this threshold, and the clusters that are merged at this level are shown as thick gray bars separated by light-gray lines. Clusters composed of only one structure are drawn as a black line reaching the bottom of the dendrogram. The main structural motifs of the database are governed by the distance $\mathsf{D_H}$. The two main clusters (a) and (b) are agglomerated according to the orientation of H_1 and the oxygen atom it is bonded to with respect to N_1 which is well described by the distance $\mathsf{D_H}$. The sub-cluster (e) is composed of ‘outlier’ structures showing an H-bond between N_2 and an hydrogen of N_1 resulting in a folded side chain structural motif. Finally, the outlier cluster (f) contains a H-bond between the carboxy H and the side-chain NH_2, that can be seen as a precursor to the zwitterionic form.

specific relative orientation between the OH and NH_2 groups. Conformer (e) is representative of a small outlier group with a well-defined bend of the side chain, leading to the formation of a further H-bond between the N_1 atom in the amino acid moiety and N_2, in the side chain. The lysine side chain is very flexible, and the distance between N and C_α only plays a role in defining the fine-grained structure of the dataset, but is minimally correlated with the most prominent features.

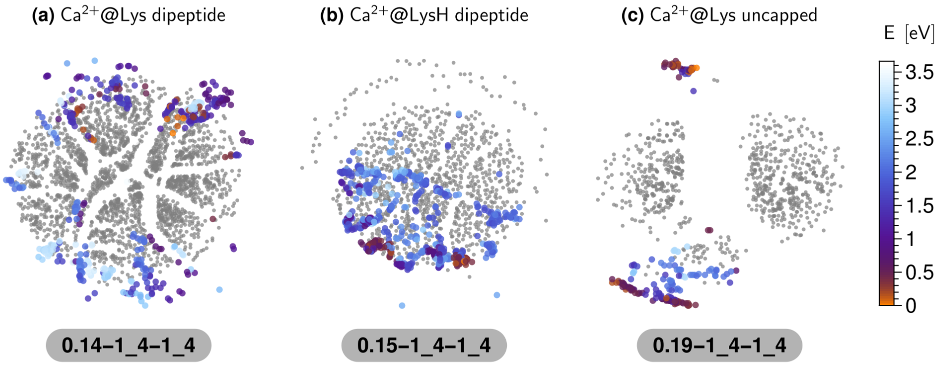


Figure 3.5 – The out-of-sample embedding of conformers with Ca^{2+} ion on the sketchmap of their pure counterpart, for the three systems we discussed in above: lysine dipeptide (a), protonated lysine dipeptide (b) and molecular lysine (c) systems. The projected conformers are colored with their energy where as the sketchmap on which they are projected are kept all in grey color. The location of the projected conformers allows us to understand how the conformational space of the pure conformers are affected due to presence of the Ca^{2+} ion.

While it appears that even in this case we could identify the basic structural motifs that characterize the conformational landscape of this system, the correlation with energy is very poor. There are several instances, in both the dendrogram and the sketchmap, where two conformers that are detected as structurally very similar display very different stability. Understanding whether this inconsistency signals a problem with our analysis brings us to the topic of outlier detection and consistency checks, that we will discuss in details in Section 3.4.3.

3.4.2 Understanding the Impact of Perturbations on conformational Space

Having elucidated the essential structural motifs that underlie the organization of a set of molecular conformers, one could also wonder how changes in the thermodynamic conditions, or other external perturbations such as solvation, the addition or subtraction of an electron[175] or that of an atom,[176–178] modify the conformations of the molecule and their stability. In addition to bare oligopeptides, the database[85,86] that we are using as an example contains sets of locally-stable conformers in the presence of cations of six different species, namely Ca^{2+}, Ba^{2+}, Sr^{2+}, Cd^{2+}, Pb^{2+} and Hg^{2+}. We consider the case of Ca^{2+} to describe how one can characterize its impact on the conformational space of the three molecular systems that we have discussed in our previous examples. We start by calculating the dissimilarity of all the conformers containing cations with their pure counterpart. In order to make the comparison on the same footings, we did not include the location of the cation in defining the SOAP kernels, so that the presence of Ca^{2+} only enters by distorting molecular geometries and/or altering their relative stability. Using this information, we then projected the cation-containing dataset on the top of the sketchmap of structures for the bare molecule. This is done using sketchmap out-of-sample embedding, and we refer our reader to see the relevant literature[60,123,126] for more details about the method. In Fig. 3.5 we show the resulting projection, colored according to the stability of the conformers, on top of the sketchmap of the pure molecule shown in grey color as a reference. A close proximity of projected conformers with a pure conformer signifies their structural similarity. Segregation of the projected conformers with the cation in some area of the reference sketchmap, represents the structural bias introduced by the strong electrostatic interaction with Ca^{2+}.

In the case of neutral lysine dipeptide (Fig. 3.5-a), the presence of the Ca^{2+} ion induces relatively small distortions of the stable conformers, that get pushed towards the outer region of the map but are still clearly related to the locally stable structures for the bare molecule. Energies are dramatically changed, with the most stable cluster in the original map being completely absent in the presence of the cation. These observations highlight the importance of sampling high-energy conformers during high-throughput structure searches, since the relative stability can be modulated strongly by external perturbations. In particular, *cis* conformers become energetically more competitive and are topologically closer to the global minima. In the case of protonated lysine dipeptide (Fig. 3.5-b), the same analysis shows an even clearer pattern. All the conformers with Ca^{2+} ions are projected in the lower part of the sketchmap, that correspond to conformers with an extended side chain (see Fig. 3.3). The Ca^{2+} ion preferably binds to the peptide O atoms, and the electrostatic repulsion with the protonated lysine residue strongly favors extended conformers, contrary to what we observed in the case of the bare molecule. Finally, one sees that for molecular lysine the addition of Ca^{2+} leads to conformers with very different structural motifs from those seen with the bare molecule, which is apparent in the sketchmap projection being concentrated far away from the unperturbed conformers (Fig. 3.5-c). In fact, inspection of the structures shows that Ca^{2+} often triggers the transition to the zwitterionic form, with the cation coupled to the carboxylate group, and the protonated side chain NH^3_+ extending as far as possible away from it. In analogy with what was observed for Lennard-Jones clusters[123] and solvated polypeptide segments,[179] sketch-maps proved to be a powerful tool to analyze the response of the system to external perturbations and changes in the boundary conditions, and – in this specific example – to draw connections between different subsets of a high-throughput molecular database.

3.4.3 Identifying Outliers and Checking for Consistency

The tools we introduced in this work are useful to address other important issues in data-driven science, such as outlier detection and consistency checks. We have already discussed the importance of detecting groups of “outlier” structures that are very different from the bulk of the dataset. These unusual items often signal the occurrence of unexpected effects that go beyond the original goal of the database construction effort. In the case of protonated lysine dipeptide, looking for outliers allowed us to reveal the presence of conformers with different chemical connectivity, or of strong H-bonds between the backbone and the charged side chain. Similar observations can also be made in the case of the bare lysine molecule (Fig. 3.4). Moreover, one can observe a branch at the topmost level of the dendrogram, containing only two conformers. These are the only two cases where a H-bond is formed between the N of the side chain and the H atom of the OH group in the backbone. In the sketchmap, these two conformers are projected on the top, clearly isolated from rest of the groups, and bear the most resemblance to the zwitterionic conformers that are stabilized in the presence of a divalent cation. Obviously, the definition of a group of "outliers" can be more nuanced, and refer to small groups of structures appearing at deeper levels in the hierarchy. Overall, the possibility of clustering together the structures from a large dataset and inspecting a few representative conformers, rather than hundreds or thousands, greatly facilitates the task of identifying trends and spotting interesting or unexpected structures.

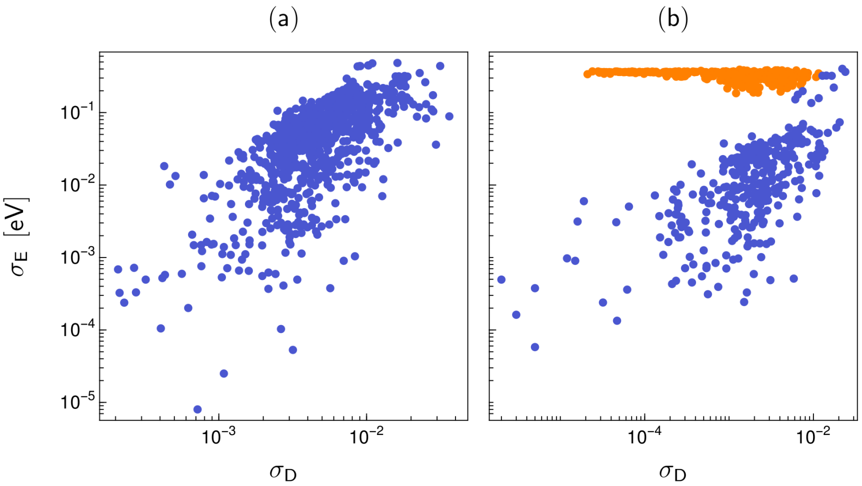


Figure 3.6 – This figure compares the homogeneity of clusters from the protonated lysine dipeptide (see a) and the bare lysine uncapped (see b) with respect to properties of their elements. The homogeneity of a cluster is probed using the standard deviation with respect to the distance between each cluster elements, σ_D, and the conformational energy, σ_E. The outliers of uncapped lysine (b) were manually highlighted in orange.

Outliers can signal interesting or important trends, but can also be a red flag for the presence of errors. The importance of database integrity has long been recognized by computer scientists,[180–183] and several tools are available to monitor and correct inconsistencies from the technical point of view, in terms of reliability of storing and retrieving data.[156–161] The issue is also crucial when it comes to the maintenance of automatically-generated databases, and to repositories that store data of heterogeneous provenance.[18,19,21–23] In these cases, problems have generally little to do with the integrity of the storage, but rather with the consistency of the simulation details of different sets of calculations. Rather, inconsistencies should manifest themselves in the presence of structures that are geometrically very similar, but are associated to very different values of particular properties.

For example the lysine molecule dataset shows signs of this kind of issues, with energies that vary wildly within clusters that are very homogeneous in structure. This problem can be seen from the maps, i.e. when comparing the energy-colored sketchmap in Fig. 3.4 to the respective maps for the other systems. However, a more robust and easy-to-automate approach to identify structure/property inconsistencies starts from the hierarchical clusters, and compares the structural variability within each cluster σ_D (Eq. (3.8)) with the variance of a given property, in this case energy, σ_E. Looking, for example, at a glassy energy landscape,[184] one can observe configurations that are very different from a structural point of view, but have similar energy, giving rise to clusters with large σ_D and small σ_E. The data points in Fig. 3.6 each represent individual clusters of lysine dipeptide and uncapped lysine, respectively, and illustrate their variation in energy and structure. In the case of lysine dipeptide (Fig. 3.6a) one sees a clear correlation between the structural and energetical variation of the clusters. The two quantities σ_D and σ_E are not necessarily strongly correlated, but in general clusters that contain very similar structures also have a low spread in energy. For uncapped lysine (Fig. 3.6b), however, one can identify a group of points (which we manually highlighted in orange for clarity) that has a distinctively different behavior, with σ_E converging to a constant value other than zero as σ_D decreases. This kind of feature indicates that the metric based on which structures were classified cannot detect one specific effect that has a dramatic impact on energetics, signaling either a failure of the metric or, as in this case, an inconsistency in the generated data. Further investigation of the lysine molecule dataset revealed that a subset of structures that had been generated at a lower level of theory in the initial stages of the structure-search procedure made their way by mistake into the final dataset. Using this measure of cluster homogeneity on all systems of the amino acid database revealed similar problems also for other molecules, for example Cys, Glu, and Arg. Thanks to this analysis we will be able to identify and rectify mistakes in all the affected datasets and subsequently update the on-line repository.[85]

3.5 The structure–energy–property landscapes of molecular crystals

Molecular crystals possess a diverse range of applications, including pharmaceutical,[185,186] electronics[187,188] and the food industry.[189] The directed assembly of molecules into crystalline materials with targeted properties is a central goal of the active research field of crystal engineering. However, material design guided by empirical rules of self-assembly often exhibit inconsistent success, particularly for the crystallization of molecular solids, because it is generally impossible to predict the outcome of self–assembly that is directed by many competing, weak non–covalent intermolecular interactions. A typical example is the phenomenon of polymorphism in molecular crystals,[13,190,191] whereby a given molecule can crystallize into different solid forms. Polymorphism is a central issue of the design of molecular materials since different stacking patterns have a direct effect the materials properties. To overcome this challenge, computational methods have been developed for crystal structure prediction (CSP) of organic molecules; over the past decade, CSP has been developed to the point where the experimentally-accessible polymorphs of small organic molecules can be predicted with reasonable success, as demonstrated by a series of CSP blind tests.[14] Recently, CSP has been combined with property prediction to produce energy-structure-function maps that describe the diversity of structures and properties available to a given molecule.[87,192] Hence, structure prediction methods are gaining increasing attention in the field of computer–guided materials design.[18,23,24] Nevertheless, in contrast to other fields of molecular science such as nano-clusters[76,184] and biomolecules,[60,73] little attention has been paid to the development of automatic analysis methods to rationalize the potential energy landscape and the structure-property relations in molecular crystals. Heuristic classifications of polymorphs based on the analysis of packing types[193] or of hydrogen bond (H-bond) patterns[194] are useful as they provide intuitive rules that can guide synthetic chemists in the design of crystallization protocols that yield the desired products. However, they lack transferability, and risk biasing the design of new materials based on outdated or partly irrelevant prior knowledge.

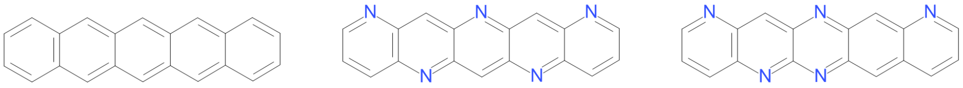
We discuss the application of the sketch-maps (see Section 3.2) and the HDBSCAN* (see Section 3.3) techniques to develop a data-driven classification scheme that provides useful insight into the packing motifs and structure-property relations. We use, as benchmark systems, pentacene (see Fig. 3.7a) and two azapentacene (see Figs. 3.7b and 3.7c) isomers, recently studied as possible organic semiconductors by CSP methods.[87] To best inform and automate the definition of this data-driven identification of structural patterns, we use the best performing SOAP-REMatch kernels obtained for modelling the lattice stability in Section 4.2.2. This classification scheme highlights families of structures on each CSP landscape and helps clarifying how introducing nitrogen substitutions in pentacene modifies the overall crystal packing landscape.

3.5.1 A Benchmark Database of organic semiconductors

We focus our present investigation on the lattice energies and charge mobility landscapes of three polyaromatic molecules: pentacene and two azapentacenes (5A and 5B), as depicted in Fig. 3.7. Pentacene is one of the most studied polyaromatic hydrocarbons, with promising electronic properties for organic semiconductor applications as a hole transporter. Without strong, directional intermolecular interactions, pentacene favours herringbone packing in crystalline phases, where molecules are arranged with a tilted edge–to–face arrangement in which neighboring molecules interact via C–H$\cdots\pi$ interactions. Generally, a co–facial π–stacking arrangement is preferable for crystalline organic semiconductors since it maximises the intermolecular charge transfer integrals.[195] Winkler and Houk,[196] suggested introducing a symmetric and complementary nitrogen substitution pattern along the long edges of the pentacene molecule to encourage hydrogen–bonding into a sheet–like packing in the crystal of the resulting azapentacene (molecule 5A, Fig. 3.7b), with the intention of increasing charge mobilities by promoting π–stackings. We have also studied molecule 5B (Fig. 3.7c) to further investigate if an irregular nitrogen substitution pattern would be less likely to promote sheet–like molecular arrangements in the crystal structure of this molecule.

Full details of the crystal structure and transport property predictions for these three molecules were presented in Ref.[87], and are summarized in Section 4.2.2. In brief, crystal structures were generated by quasi-random sampling[197] in a range of space groups, followed by lattice energy minimization with DMACRYS[198] using an empirically parameterized exp-6 force field model (W99[199]) combined with atomic multipolar electrostatics derived from a distributed multipole analysis (DMA).[200]

Besides this well-established semi-empirical model for predicting lattice energies, we also computed single-point energies of all the structures using density functional theory (DFT), with an expansion of Kohn-Sham orbitals in plane waves and a the generalized-gradient-approximation density functional PBE,[163] including Grimme's D2 dispersion corrections,[201] as implemented in Quantum ESPRESSO.[202] Further details of the DFT calculations are provided in Section 4.2.2.[128]



(a) Pentacene **(b)** 5A **(c)** 5B

Figure 3.7 – Molecules investigated in the present study.

The crystal packings of the predicted structures were classified into one of the categories typically used in describing polyaromatic hydrocarbon crystal packing:[193,203] herringbone, where all molecules adopt a tilted edge-to-face arrangement; sandwich-herringbone, in which pairs of coplanar molecules pack in a herringbone manner; γ, which features stacks of coplanar molecules; and sheet-like, where all molecules are coplanar. A fifth category, slipped–γ, was added in our previous publication[87] describing gamma structures in which the lateral offset between stacked molecules is so large that there is little $\pi - \pi$ contact along the stack of molecules. The classification was performed using an in–house algorithm based on a set of heuristic rules, by calculating the relative orientations of molecules in a sphere surrounding a central reference molecule in a given crystal, as described in Ref.[87].

3.5.2 Results & Discussion

While the "best" kernel for property prediction can be determined objectively based on the cross-validation error, it is more difficult to formulate objective criteria to optimize the parameters when a kernel is to be used for determining structural motifs, or generating low-dimensional maps of the crystal structure landscape. We found that by starting from the best parameters for energy prediction provided in Section 4.2.2, and modifying the cutoff radius to select different chemical features, e.g. H-bonds and CH$\cdots\pi$ interactions, we can change the representation of the structures in a predictable way. This turns out to be insightful, as we discuss below for the pentacene, 5A and 5B databases.

Pentacene

Figure 3.8 shows a sketch-map representation of the pentacene dataset color-coded according to the relative lattice energy (bottom right), a heuristic classification scheme developed in the previous publication on CSP of azapentancenes[87] (top right) and the clusters detected by HDBSCAN* based on the kernel-induced metric (left).

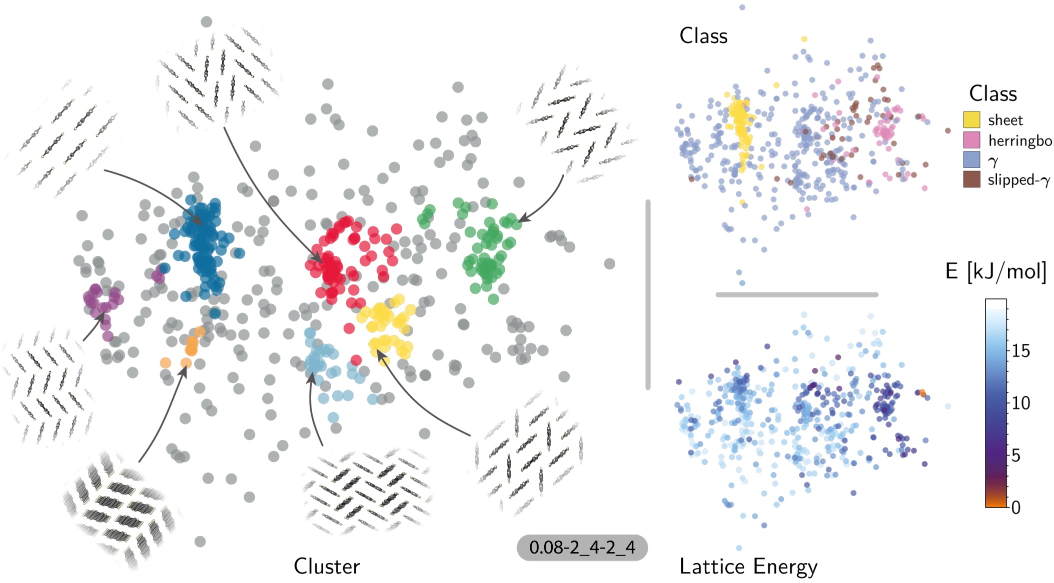


Figure 3.8 – Sketch-map representations of the pentacene crystal structure landscape's similarity matrix (projection parameters shown follow the scheme σ_{map}-A_B-a_b). The atomic configurations are color-coded according to their relative lattice energy (bottom right), class following the heuristic classification (top right) and cluster index (gray structures do not belong to a cluster) found using HDBSCAN* on the similarity matrix (left). The structural pattern of each cluster is illustrated from a view down the short edge of pentacene.

The 'islands' on the sketch-map indicate the presence of distinct structural motifs (Fig. 3.8). The HDBSCAN* technique identifies seven clusters among which two match clearly the herringbone and sheet heuristic classes. The correspondence between a classification based on unsupervised data analysis and one based on a well-established understanding o the behavior of π-stacked system provides a cross-validation of the two approaches. The combination of SOAP-REMatch kernels, sketch-map and clustering is capable of recognizing well-known stacking patterns, and vice versa these heuristic classes have a clear correspondence in the structure of the crystal structure landscape.

Cases in which the two classifications differ are similarly insightful. For example, γ packing is defined by a stacking column of molecules along their short axis, while neighboring columns could be tilted with respect to this reference stacking direction. The HDBSCAN* clustering shows that this broadly-defined grouping overlooks the existence of several well-defined clusters of 'mixed' character, that differ by the tilting pattern between neighboring molecules, making it possible to identify e.g. structures that are (i) closer to a sheet–like packing, e.g. the orange island shown in Fig. 3.8 where one nearest-neighbor column is parallel whereas another neighboring column is tilted with respect to each others, or (ii) further from a sheet–like packing, e.g. the purple island shown in Fig. 3.8 where all nearest-neighbor columns are tilted with respect to each other. The slipped-γ packing, on the other hand, does not correspond to a clear-cut group of structures, encompassing a sparse set of configurations that populate different portions of the map. Inspection of these structures, informed by the mapping and the automatic classifications, reveals that this heuristic class is not well-suited to rationalize packing in pentacene.

Clustering techniques like HDBSCAN*, which work in the high dimensional space, are also useful to complement non-linear projections based on the similarity matrix, making it possible to recognize the distortions brought about by the projection and develop a better understanding of the actual structure of the similarity matrix. For instance, small groups of structures such as the one on the lower right of the sketch-map might appear like a cluster because of the projection, while clusters such as the green and red ones might not seem fully homogeneous. Nevertheless, a careful inspection of these groups of structures confirms that clusters detected by HDBSCAN* are indeed structurally homogeneous while the group on the lower right corresponds to complex variations and distortions of the herringbone pattern which do not show an obvious common structural pattern.

The automatic classification based on kernels provides more fine-grained insights into the structural diversity in the lattice energy landscape compared to the heuristic classifications. To verify how these observations generalize to different classes of molecular crystals, we also considered the case of the two azapentacene isomers 5A and 5B.

Azapentacene 5A

The main difference between configurations, which is apparent by visual inspection, consists in the different arrangements of CH$\cdots$N H–bonds between molecules within each sheet. In order to focus our investigation on such patterns, without the confounding information associated with the relative arrangement of molecules in adjacent sheets, we use a kernel with a cutoff radius of 3 Å, which is sufficient to identify H–bonds but is insensitive to inter-sheet correlation, given that the typical distance between sheets is about ~3.5 Å. The outcome of this analysis is shown in Fig. 3.9. The HDBSCAN* automatic classification identifies nine main structural patterns, eight of which are sub-classes of the sheet motif. Representative structures for a few of these clusters (see Fig. 3.9) show that although a wide range of H-bond arrangements are possible within sheets, only a handful emerge as well-defined packing patterns. A single well-defined cluster that does not correspond to variations on the sheet

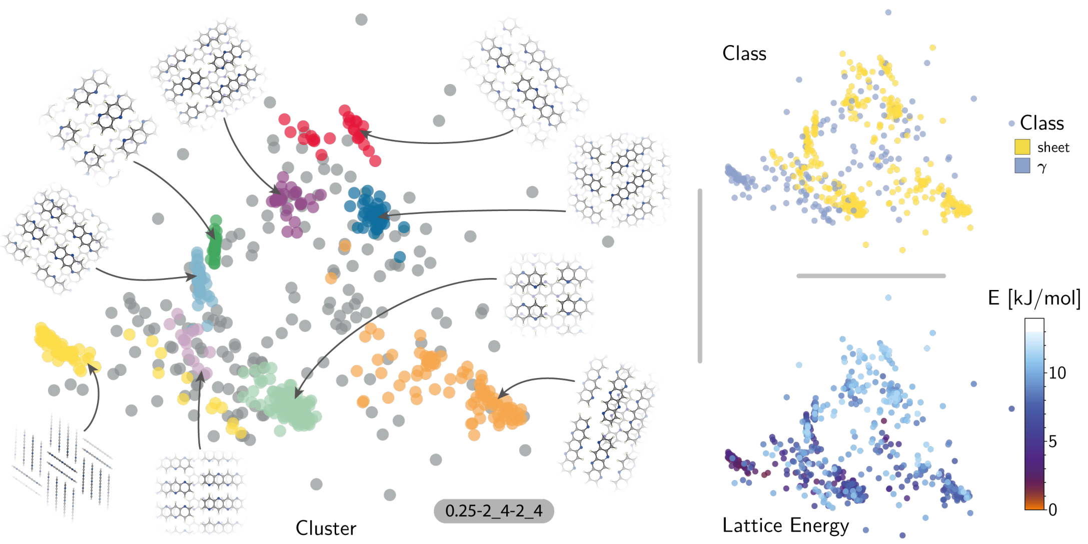


Figure 3.9 – Sketch-map representations of the 5A crystal structure landscape. The atomic configurations are color-coded according to their relative lattice energy (bottom right), class following the heuristic classification (top right) and cluster index (gray structure do not belong to a cluster) found using HDBSCAN* on the similarity matrix (left). The structural pattern of each cluster is illustrated with a top and long side (yellow cluster) view of the 5A polymorphs.

stacking is also present and identified, corresponding to the γ heuristic class, while other patterns are detected as background/outliers by HDBSCAN*.

The fact that the overwhelming majority of structures can be traced to a sheet motif, despite using a CSP protocol that is designed to sample as widely as possible the most–likely packing patterns for a given molecule, as demonstrated in the case of pentacene, underscores the fact that the nitrogen substitution favors the sheet stacking patterns and inhibits other kinds of structural motifs. However, we find relatively poor correlation between structural similarity and lattice energy (see Fig. 3.9, bottom right) when the kernel is tuned to disregard inter-layer correlations. This reflects the fact that in-sheet H-bonding is not the sole factor determining the stability of packing. This is an example of the insight that can be obtained by combining supervised and unsupervised ML analysis of the configurational landscape of molecular materials.

Azapentacene 5B

The structural basis of this greater complexity can be understood by performing an HDBSCAN* analysis and inspecting the sketch-map representation of the dataset. Even when using a 3Å cutoff for the kernel, the sketch-map representation of the similarity matrix does not show clear 'islands', i.e. recurring structural patterns (see Fig. 3.10), suggesting the presence of a glassy structural landscape in which many distinct patterns can be formed.[204] Indeed, even

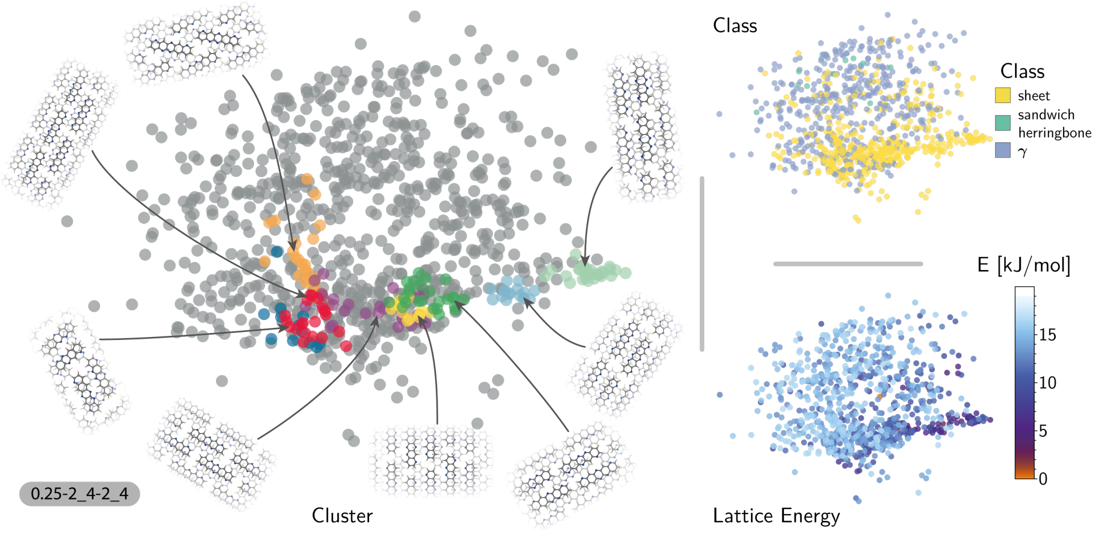


Figure 3.10 – Representation of the similarity matrix for 5B The atomic configurations, i.e. disks, on the three sketch-maps are color-coded according to their lattice energy (bottom right), class following the heuristic classification (top right) and cluster index (gray structure do not belong to a cluster) found using HDBSCAN* on the similarity matrix (left). The structural pattern of each cluster is illustrated with a top view of the 5B polymorphs.

though HDBSCAN* finds 8 clusters that can be described as sheet-like (see a few representative structures in Fig. 3.10), they correspond to less than 20% of the structures, and the majority of the database (760 samples) is too sparse to be partitioned into well-defined clusters.

This variety of complex and diverse stacking patterns that do not seem to fit into specific arrangements can be traced to the irregular substitutions of carbon atoms by nitrogen atoms, that determines a transition from a structure-seeker energy landscape to a glassy energy landscape.[204]

4 Accurate and reliable supervised ML

Regression algorithms aim to construct a model $F(A)$ that can predict accurately the target properties y associated with the input A.[62] The internal parameters of the model are determined by optimizing the accuracy of prediction over a set of training samples, $\mathcal{D}$, associated with their respective property, and their accuracy to that reference can be improved systematically by increasing the size of the training set.[77]

One of the early applications of ML to the prediction of atomic-scale properties aimed at obtaining an accurate model of the potential energy surface (PES), which is crucial to assess the stability of a given configuration, and whose sampling underlies the evaluation of the thermodynamic properties of a system.[78] Contrary to traditional FFs, which assume physics-inspired functional forms for the interactions, and often use experimental observable as fitting targets, ML interatomic potentials (MLIPs) have flexible functional forms and usually rely on electronic-structure calculations as a reference. In many cases, this more general, data-driven approach has been shown to result in more transferable and accurate models.[29,30,45,49] Besides the PES, ML models have also been successful at predicting other zero Kelvin properties such as chemical shieldings, band gaps, electron affinities, electron transfer integrals and static isotropic polarizabilities.[46,75,84,89,114,205–207] While considerable success has also been shown in using ML to predict complex properties that cannot be seen as arising from an individual atomic configuration (e.g. the free-energy of a state, the toxicity or pharmaceutical activity of a molecule, etc.), here we will focus entirely on the well-defined task of building a surrogate quantum model, which can sidestep the solution of the Schrödinger equation and predict the properties of a specific atomic configuration.

As mentioned in Chapter 2, we focus on using representations of the atomic structure to adapt any given ML model to these symmetries. Even though the regression techniques discussed in this chapter are compatible with most representations, all of the following models are based on the SOAP power spectrum representation (see Eq. (2.40)) and are trained to predict scalar properties. In Section 4.1, we start by discussing the supervised learning algorithms that are used to model lattice energy, chemical shieldings, and transfer integral in molecular crystals (see Sections 4.2.1 and 4.2.2). We then investigate how to improve these models by optimizing the representation on several benchmark datasets in Section 4.3. Finally, in Section 4.4, we present and benchmark a scheme to obtain an inexpensive and reliable estimate of the uncertainty associated with the predictions of a machine-learning model of atomic and molecular properties.

4.1 Supervised Learning for atomistic modelling†

A scalar property $y(\{\mathbf{r}_i, \alpha_i\})$ of a system A of N atoms of species α_i, located at positions $\mathbf{r}_i$, can be expressed formally as a function of a vector of features $|A\rangle$ that represents the structure,

$$y = F(|A\rangle). \tag{4.1}$$

The problem of modelling $F(|A\rangle)$ can therefore be decomposed into the problem of providing a concrete formulation of the feature vector (that we have discussed in detail in Chapter 2) and that of determining the functional form of the approximating model F. Irrespective of the regression technique used, most of the transferable property models that have been introduced in recent years decompose a property associated to a set of atoms A into atom-centered contributions, i.e.

$$F(|A\rangle) = \sum_{i\in A} f(|A_i\rangle), \tag{4.2}$$

where f is a trained ML model and A_i indicates the atomic environment centered on atom i of structure A. For the simplicity of the notation, we will use $|A\rangle$ and A interchangeably to refer to the representation of structure A. This choice can be motivated as a consequence of imposing the invariance of the property on the absolute position of the system (see Section 2.1.4), and – together with the limitation of the range of each environment to a region centered on the i-th atom – yields models of great transferability, since it allows breaking down the properties of large, complex configurations into a sum of contributions that only depends on the position of the neighboring atoms. In the cases in which this ansatz is not justified (e.g. for properties such as ligand binding affinity, or in the presence of significant long-range interactions) other strategies for combining local environments predictions like the REMatch kernel should be considered.[114]

Linear models based on permutation invariant polynomials (PIPs) have been very effective at reproducing accurate chemical reactions between small molecules[49,50,209] and to build efficient MLIPs with the many-body tensor (MTP) framework[54] that extends them to more complex systems.[210,211] Similarly linear models based on the n-body correlation function[100,106,212–214] have shown great promise. Fully non-linear models based on artificial neural networks (ANN) have however been the most popular this far. ANNs have been constructed based on the the expansion of the radial (and angular) distribution function on a basis

---

[†] This section has been adapted from section 3 of the journal article [208] whose authors are **Félix Musil** and Michele Ceriotti. The author of this thesis wrote this review article under the supervision of Prof. Michele Ceriotti.

such as the Behler-Parrinello symmetry functions,[29,105,215–220] Zernike polynomials,[221] Chebychev polynomials,[118] Gaussians,[205,222–224] and proved very successful at investigating the properties of complex systems.[207,225–229] Another class of models that have been both very popular and successful is based on Gaussian process regression (GPR),[88] that is formally equivalent to kernel ridge regression (KRR) and can be seen as a middle-ground solution that introduces non-linearity in the form of a kernel function $k(A,B)$ built on pairs of feature vectors, but effectively translates into a linear regression problem that uses (some of) the training set structures as the basis on which the structure-property relation is constructed. GPR has been used to predict the stability of molecules and solids[30,32,33,45,75,83,84,94,114,206,230] and build MLIPs for elemental solids,[127,231–233] nano clusters,[234] isolated molecules[235] and molecular liquids[236] as well as for the direct prediction of other quantum mechanical properties.[46,84,89,128,237–239]

4.1.1 GPR model

In the most straightforward form, a GPR model built on a kernel function k can be written based on a set $\mathcal{T}$ of N training structures, and the associated properties y. Assuming a Gaussian likelihood, and an additive, atom-centered property model, the prediction for a structure A becomes

$$F(A) = \sum_{T\in\mathcal{T}} x_T K(A,T), \tag{4.3}$$

where x_T is the weight associated with T, $K(A,T) = \sum_{i\in A}\sum_{j\in T} k(A_i, T_j)$ and the kernel function $k(\cdot,\cdot)$ quantifies the similarity between the local environments of T and A. The key ingredient of this model is the kernel function that - subject to a few conditions such as positive definiteness - defines an inner product between the inputs $k(A_i, T_j)$.

GPR is often preferred over the more sophisticated non-linear models because of its ease of use: it has a single interpretable hyperparameter λ, and the solution for the weights $\boldsymbol{x}$ has the closed form

$$\boldsymbol{x} = (\boldsymbol{K}_{\mathcal{T}\mathcal{T}} + \lambda^2 I)^{-1}\boldsymbol{y}, \tag{4.4}$$

where $\boldsymbol{K}_{\mathcal{T}\mathcal{T}}$ is the kernel matrix between the training inputs, I is the identity matrix and $\boldsymbol{y}$ are the property associated with the training set $\mathcal{T}$; λ corresponds to an expected Gaussian noise in the references $\boldsymbol{y}$ so it can account for small discrepancies in the convergence of the electronic structure method that are often found across a training set, and for errors caused by the local property ansatz. For simplicity we have used a single λ for the whole training sets, but this parameter might also take different values for each training sample as it is done in the GAP model.[30] In the language of kernel ridge regression, Eq. (4.4) can be obtained by minimizing the loss

$$L(\boldsymbol{x}) = \sum_{T\in\mathcal{T}} |F(T) - Y_T|^2 + \lambda^2 x_T^2. \tag{4.5}$$

It should be mentioned that GPR provides a simple approach to compute derivatives of the target properties with respect to atomic positions, e.g. the force consistent with the model, in which case F models the energy of a configuration. Derivatives can also be incorporated in the learning procedure,[88,94,235,239–241] by including the discrepancy between reference and predicted values in the loss Eq. (4.5). Finally, the probabilistic nature of GPR also allows one to estimate the uncertainty associated with the prediction

$$\sigma_Y^2(A) = \lambda^2 + K_{AA} - \boldsymbol{K}_{A\mathcal{T}}(\boldsymbol{K}_{\mathcal{T}\mathcal{T}} + \lambda^2 I)^{-1}\boldsymbol{K}_{\mathcal{T}A}. \tag{4.6}$$

4.1.2 Approximate GPR model

The drawback for such simplicity is the computational cost associated with the training phase - which scales cubically with the training set size - and the need to use the full training set as a basis to perform predictions. To address this issue, many approximations of the exact kernel matrix have been proposed,[242,243] among which the projected process (PP) approximation[242,244] has been shown to be quite practical to include force references[240,241] and effective from the point of view of the cost and accuracy of predictions.[233,245] The PP method introduces a set $\mathcal{S}$ of M sparse points with $M < N$ to approximate the GP prior which practically reduces the cost of training to the inversion of a $M \times M$ matrix, and ensures that predictions only require computing kernels between the new configurations and the M sparse points:

$$\begin{aligned} F^{\mathrm{PP}}(A) &= \boldsymbol{K}_{\mathcal{S}A}^T \tilde{\boldsymbol{K}}^{-1} \boldsymbol{K}_{\mathcal{S}\mathcal{T}} \boldsymbol{y}, \\ \sigma_Y^{\mathrm{PP}}(A)^2 &= \lambda^2 + K_{AA} - \boldsymbol{K}_{\mathcal{S}A}^T \boldsymbol{K}_{\mathcal{S}\mathcal{S}}^{-1} \boldsymbol{K}_{\mathcal{S}A} + \boldsymbol{K}_{\mathcal{S}A}^T \tilde{\boldsymbol{K}}^{-1} \boldsymbol{K}_{\mathcal{S}A}, \end{aligned} \tag{4.7}$$

where $\tilde{\boldsymbol{K}} = \boldsymbol{K}_{\mathcal{S}\mathcal{S}} + \lambda^{-2}\boldsymbol{K}_{\mathcal{T}\mathcal{S}}^T \boldsymbol{K}_{\mathcal{T}\mathcal{S}}$, $\boldsymbol{K}_{\mathcal{S}\mathcal{S}}$ indicates the kernel matrix between pseudo inputs, and $\boldsymbol{K}_{\mathcal{T}\mathcal{S}}$ the matrix between training points and pseudo inputs. For simplicity, the pseudo inputs (or active points) can be chosen directly from the training set and they represent a new basis in which the regression is performed. To maximize the cost reduction and the accuracy of the model, one needs to sample the active set carefully. Selecting randomly the active inputs is far from optimal so several approaches have been proposed[244,246–248] among which Farthest Point Sampling (FPS),[123] a greedy method that maximizes diversity, or a CUR decomposition[116,240] of the feature matrix associated with the training set, which minimizes the effect of the PP on the kernel matrix, have allowed significant reductions of the computational cost with minimal degradation of the accuracy.[233,245]

4.1.3 Model ranking and optimization

ML algorithms include recipes to train their parameters, e.g. Eq. (4.4), but they do not specify how to determine hyperparameters such as the regularization λ for GPR, the number of layers in an ANN, and the cutoff radius r_{cut} in the power spectrum representation, which can influence heavily the quality of the model. In the Bayesian context these hyperparameters can be interpreted as priors that should be inferred from our knowledge of the physical system,[240] or thought of as parameters that need to be optimized. In principle, the best parameters should allow for the lowest possible prediction error on all possible inputs. Given that one can only work on a finite-sized set of references, the problem becomes to find the parameters that best reproduce the available references and at the same time generalize well to unknown inputs. The performance of a model is measured by comparing the predicted values and the reference values with metrics such as the mean absolute error (MAE), the root mean square error (RMSE), the supremum error (SUP), the coefficient of determination (R^2) or the Spearman's rank-order correlation (COR). An effective technique to avoid overfitting these parameters, i.e. specialize the model for the training set which leads to poor generalization performances, is the so-called k-fold cross-validation where the performances are evaluated on several subsets of the training set (see Hansen et al. [230] and Refaeilzadeh, Tang, and Liu [249] for more details). Cross validated scores are more likely to match the generalization error which is a good basis to rank models and determine the optimal set of hyperparameters.[250] Learning curves are another standard diagnostic tool to characterize the performance of ML models. From statistical theory, the error of a given model decreases as a power-law with the size of the training set.[77] Figure 4.12 shows, on a logarithmic scale, three learning curves for models trained on datasets of molecular crystal polymorphs to reproduce their lattice energies. The GPR model performances vary with the considered training set because the learning rates (slopes of the curves) and off-sets are different. These curves are very useful because they help to differentiate between models that have a small offset and learning rates with models that have a larger off-set but also steeper slopes (see Fig. 4.16 for an example). Indeed, building a 'good' model with as few references as possible might be favored over a model that has a better learning power but poorer performances with few samples.

Even though learning curves and cross-validation procedures can benchmark quantitatively the ability of a model to perform well in production, demonstrating the performance of a model on practical test cases is typically more compelling.

4.2 Predictive models for molecular solids

Molecular solids are characterized by the combinatorial complexity and diversity of organic chemistry, the subtle dependence on conformations, and the long and short-range effects of crystal packing, which leads to great chemical diversity. They frequently crystallize in different polymorphs with substantially different physical properties.[190,191] To help guide the synthesis of candidate materials, atomic-scale modelling can be used to enumerate the stable polymorphs and to predict their properties. This is a critical issue, especially for the pharmaceutical industry, where properties of molecules, such as dissolution rate, must be strictly controlled because they can be significantly affected by the presence of different polymorphs.[13] Polymorphism also affects the opto–electronic performance of organic semiconductors, which are used in flexible electronic devices. To contribute to overcoming these challenges, we show that GPR with SOAP features is able to model the stability, the transfer integral (see Section 4.2.2), and chemical shifts (see Section 4.2.1) of molecular crystals.

4.2.1 Chemical shifts in molecular solids by machine learning[‡]

Solid-state nuclear magnetic resonance (NMR) spectroscopy is among the most powerful methods for determining the atomic-level structure and dynamics of powdered and amorphous solids. A revolution in solid-state NMR has occurred with the introduction of accurate methods to calculate chemical shifts,[251–253] in particular using plane wave DFT methods developed for periodic systems based on the PAW/GIPAW approach.[254–256] This has enabled very rapid development of chemical shift based NMR crystallography, which is now widely used to validate structures of molecular solids and identify known polymorphs,[257–275] or more recently in combination with crystal structure prediction (CSP) protocols, to determine de novo crystal structures from powders.[276–281] The power of the method arises from the fact that plane wave DFT with the GIPAW method is accurate enough to reproduce the exquisite sensitivity of chemical shifts to changes in local atomic environments. However, this approach also has severe limitations. The cubic scaling of the computational cost with system size prevents the application to larger and more complex crystals, or non-equilibrium structures. If one wanted to use more accurate *ab initio* calculations, the expense is prohibitive.

Data-driven prediction of chemical shifts for the specific case of proteins in solution using methods based on large experimental databases, using traditional[282–289] or machine learning approaches,[80,290,291] have met with considerable success in predicting shifts based on local sequence and structural motifs, and are widely used today. While there are some examples of machine learned experimental and *ab initio* chemical shifts of liquid and gas phase molecules,[79,292–295] to date there is only one example of machine learning being applied to calculations of chemical shifts in solids, which deals with the specific case of silicas.[296]

In the following we use GPR with the SOAP representation (see Sections 2.2.4 and 4.1.1) to predict chemical shifts in molecular solids. The protocol is schematically illustrated in Fig. 4.2. In the absence of a database of experimental shifts, and given that experiments alone do not provide a 1:1 mapping between chemical shifts and a single atomic configuration, we train the model on DFT calculated chemical shifts for structures taken from the Cambridge Structural Database (CSD),[90] chosen to be as diverse as possible, and then show that the method can predict chemical shifts in a test set with a R^2 coefficients between the chemical shifts calculated with DFT and with ML of 0.97 for ^{1}H, 0.99 for ^{13}C, 0.99 for ^{15}N, and 0.99 for ^{17}O, corresponding to root-mean-square-errors (RMSEs) of 0.49 ppm for ^{1}H, 4.3 ppm for ^{13}C, 13.3 ppm for ^{15}N, and 17.7 ppm for ^{17}O. Predicting the chemical shifts for a polymorph of cocaine, with 86 atoms in the unit-cell, using the ML method takes less than a minute of CPU time, thus reducing the computational time by a factor of between 5 to 10 thousand, without any significant loss in accuracy as compared to DFT. Most significantly, even though no experimental shifts were

---

[‡] This section has been adapted from the journal article [89] whose authors are Federico M. Paruzzo, Albert Hofstetter, **Félix Musil**, Sandip De, Michele Ceriotti, and Lyndon Emsley. The author of this thesis analyzed the data, built and benchmarked the predictive ML models and wrote the ML related sections of the article.

used in training, we show that the model has sufficient accuracy to be used in a chemical shift driven NMR crystallography protocol to correctly determine, based on the match between experimentally-measured and ML-predicted shifts, the correct structure of cocaine, and the drug 4-[4-(2-adamantylcarbamoyl)-5-tert-butylpyrazol-1-yl]benzoic acid (AZD8329). We also show that it is possible to calculate the NMR spectrum of very large molecular crystals. For this we calculate the chemical shifts of six structures from the CSD with between 768 and 1,584 atoms in the unit-cells.

Methods and Computational Details

Crystal Structure Datasets We describe here how the datasets used to build the ML models which are outlined in Fig. 4.2 have been built and the highlight the rational behind their construction. In the absence of an experimental database of shifts the model is developed by using a reference training set of structures for which chemical shifts are calculated with GIPAW DFT. Moreover, to obtain a model which is robust and general, the training set should be as large, as reliable, and as diverse as possible. All the crystal structures of CSD-61k and CSD-500 were obtained from the CSD.[90] A total of 88648 structures was downloaded from the CSD, using two different selection criteria: the maximum number and the type of atoms contained in the unit-cell. We selected only structures with a maximum of 200 atoms, containing either (i) only H and C or (ii) H, C and one hetero-atom between N and O or both. From this set we extracted a subset of 61,012 (CSD-61k) structures by removing structures with missing protons, and structures where the distance of at least one pair of atoms was smaller than the sum of their covalent radii minus 0.3 Å. In addition, structures containing partial occupancy were resolved by keeping only the first of the atoms with partial occupancy. If we were not able to resolve the disorder, the entire structure was not included. The disorder was assumed to be removed, if the number of atoms, for each atom type, was an integer multiple of the number of atoms given in the chemical formula. Note, that as we sorted through more that 60000 structures, the whole procedure was automatized and we didn't manually select the most stable structure for a given disorder. However, here we are not looking for ground state structures but instead only for physically reasonable structures to expand our data-set. Given that performing a GIPAW calculation for all of these structures would be prohibitively demanding, we then select a random subset of 500 structures (CSD-500) that are representative of the chemical diversity in the CSD, and we use it to test the accuracy of our model. For cross-validation and training, instead, we select 2000 structures (corresponding to about 185000 atomic environments) out of the CSD-61k using the FPS algorithm,[123] namely CSD-2k. This step ensures near-uniform sampling of the conformational space, improving the quality of the model when using a relatively small number of reference calculations.

DFT Calculations All the DFT calculations were carried out using the DFT program Quantum ESPRESSO.[202,297] For all structures in the CSD-2k and CSD-500 databases we first carried out geometry optimization using plane wave DFT. We used ultrasoft pseudopotentials with PAW[256] reconstruction, H.pbe-kjpaw_psl.0.1.UPF, C.pbe-n-kjpaw_psl.0.1.UPF, N.pbe-n-kjpaw_psl.0.1.UPF and O.pbe-n-kjpaw_psl.0.1.UPF from the USSP pseudopotential database.[298] The optimizations were done with the generalized-gradient-approximation (GGA) density functional PBE,[163] using a wave-function energy cutoff of 60 Ry, a charge density energy cutoff of 240 Ry, and without k-points. The Grimme van der Waals dispersion correction[299] was included in order to account for van der Waals interactions. The geometry optimization was done relaxing all atomic positions while keeping the lattice parameters fixed. A single point energy (scf) was then computed for the relaxed geometry, using higher wave-function and charge density energy cut-offs which were set to 100 Ry and 400 Ry respectively. For this calculation we also used a Monkhorst-Pack grid of k-points[300] corresponding to a maximum spacing of 0.06 $Å^{-1}$ in the reciprocal space. The k-points and energy cutoff values were optimized to ensure convergence of the electron density. Finally, we calculated the chemical shielding σ_{ref} using the GIPAW method, with the same parameters as used in the scf calculation. All the relaxed geometries, together with the GIPAW DFT calculated chemical shifts, are available from the SI of Ref.[89]. Note that using a convergence threshold of in the scf calculation of 10^{-8} Ry leads to a residual random error on the macroscopic contribution to the shifts of the order of 0.1 ppm. Fully converged results can be achieved with a threshold of $10^{-12} - 10^{-14}$ Ry.

Detection of Unusual Environments The quality of the training set is essential to ensure the optimal performance of a machine learning algorithm. However, the individual curation of the 2000 molecular crystals of the CSD-2k dataset would be very time consuming and cumbersome. Note, that the 2000 molecular crystals correspond to around 35000 symmetrically non-equivalent atomic environments for ^{1}H alone and the following detection procedure is applied directly to the individual atomic environments instead of the whole molecular crystals. We automate this detection procedure by assessing the ‘instability’ of the prediction of the shielding of a given local environment using the difference between the predictions of several GPR models and the reference DFT-shielding. We define this indicator as:

$$\epsilon(A_i) = \frac{1}{M}\sum_{m=1}^{M}(y_m(A_i) - y(A_i)), \tag{4.8}$$

where each of the M models is made using a 2-fold split of the shuffled training set that does not include the structure A. In total we generate $M = 40$ models, where each is generated using a different random shuffling of the data. Environments with a large value of $|\epsilon(A_i)|$ are not well-described by the rest of the training set within the SOAP-GPR framework. Note, that the error would cancel out in the case of random noise within the prediction, while a large value of $|\epsilon(A_i)|$ corresponds to a systematic error in the predicted chemical shielding, that could be associated to the limitations listed below. We define local environments to be unusual when $|\epsilon(A_i)|$ is larger than three times the standard deviation of $|\epsilon(A_i)|$ over the whole training set, and we then do not use them for training. We perform this elimination procedure on the CSD-2k dataset using a single kernel for each element ($r_{\mathrm{cut}} = 4.5$Å for ^{1}H, 4Å for ^{13}C, 4Å for ^{15}N and 3Å for ^{17}O). The hyperparameters of the single kernels used in the elimination procedure were determined using a grid search and 3-fold cross validation on the uncleaned CSD-2k training set. The ^{1}H environments excluded with this approach are shown in Fig. 4.1. It is interesting to see that in several cases we can trace the unusual behavior of the environment to subtle errors in the DFT calculations, or to physical phenomena that are ill described within our DFT model (metallic systems, zwitterions,...). However, note that we are not systematically removing such structures and that the training set still contains many structures with the listed features.

Most of the environments detected as 'unusual' are part of zwitterionic structures or charged structures such as VIWYEH, ZACSOO or EKUJIF (these six letters correspond to the molecular crystal identifier of the CSD). Others are metallic structures (E_{LUMO}–E_{HOMO} = 0), such as HAZQUV, QUICNA02, DMEBQU01 or AYUKIP, or have a partially empty unit cell (QAHVUQ). An intrinsic limit of this procedure is the fact that it might detect structures with uncommon functional groups as 'anomalies' (e.g. TIMCHX, which is an aziridine – a three membered heterocycle with one amine group, or FIGMAJ which has a cubane group), due to the fact that these structures are not well represented by the used training set. However, with increasing training size, we expect these structures to be better represented and they will not be detected as anomalies anymore.

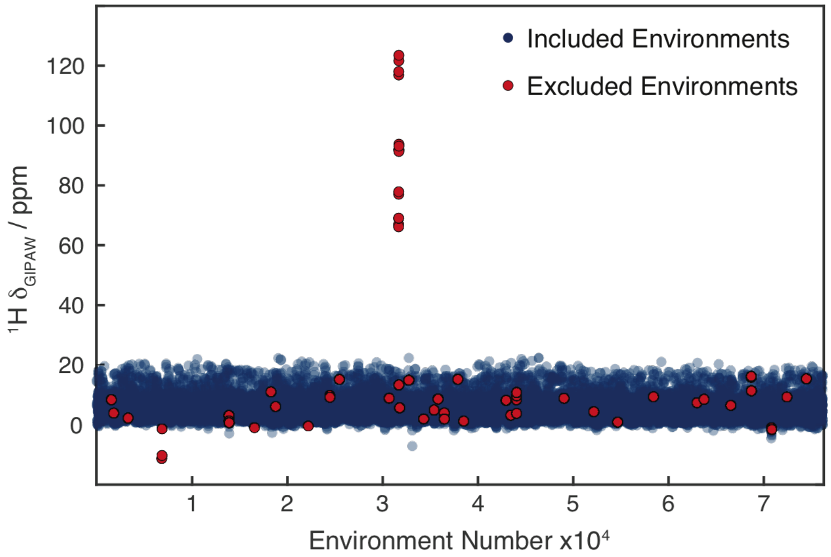


Figure 4.1 – ^{1}H chemical shifts of the 76214 environments in the CSD-2k set. The environments excluded using the unusual structures detection procedure described in Eq. (4.8) are shown in red.

ML models SOAP-based structural kernels contain several adjustable hyper-parameters, which we have not systematically explored. Instead we chose reasonable values of the parameters without extensive fine-tuning, based on previous experience[114] to select a small subset of parameters (see the SI of Ref.[89]) from which the optimal parameter sets were determined by cross-validation on the CSD-2k training set. We also combine kernels computed for different cutoff radii to capture the contributions to shifts from different length scales,[114] as is described in detail above. The parameters of the best single and multi-kernel models are summarized in Tables 4.1 and 4.2. The calculations of the local environment, the similarity kernel and the weighted correlations were done using the glosim2 package.[301]

Atom	r_{cut}	σ	$l_{\max}$	$n_{\max}$	λ	ζ
^{1}H	2	0.3	9	9	0.1	2
	3	0.3	9	9	0.1	2
	4	0.4	9	9	0.1	2
	5	0.4	9	9	0.1	2
	6	0.5	9	12	0.1	2
	7	0.5	9	12	0.1	2
^{13}C	2	0.3	9	9	0.01	2
	3	0.3	9	9	3.0	2
	4	0.4	9	9	5.0	2
	5	0.4	9	9	3.0	2
	6	0.5	9	12	1.0	2
	7	0.5	9	12	1.0	1
^{15}N	2	0.3	9	9	0.5	2
	3	0.3	9	9	1.0	2
	4	0.4	9	9	0.1	2
	5	0.4	9	9	0.1	2
	6	0.5	9	12	0.1	2
	7	0.5	9	12	0.05	2
^{17}O	2	0.3	9	9	0.5	2
	3	0.3	9	9	5.0	2
	4	0.4	9	9	5.0	2
	5	0.4	9	9	5.0	2
	6	0.5	9	12	1.0	2
	7	0.5	9	12	7.0	2

Table 4.1 – Kernel and GPR parameters. The GPR parameters (λ and ζ) are the ones used in single kernel predictions.

Crystal Structure Prediction Here we use a set of possible polymorphs predicted by CSP for cocaine and the drug 4-[4-(2-adamantylcarbamoyl)-5-tert -butylpyrazol-1-yl]-benzoic acid (also referred as AZD8329). General details on the CSP protocol can be found in Ref.[302]. In chemical shift based NMR crystallography, the CSP trial polymorphs are tested against experimental parameters (^{1}H chemical shifts) to determine the experimental crystal structure. We used 30 possible polymorph structures of cocaine and 14 trial structures of AZD8329 generated with CSP. The 30 structures of cocaine were obtained from the Electronic Supporting Information (ESI) of Ref.[265], and correspond to the most stable polymorphs obtained with

Atom	Multi-Scale Kernel Weights						λ	ζ
	$r_{\mathrm{cut}} = 2\,\text{Å}$	$r_{\mathrm{cut}} = 3\,\text{Å}$	$r_{\mathrm{cut}} = 4\,\text{Å}$	$r_{\mathrm{cut}} = 5\,\text{Å}$	$r_{\mathrm{cut}} = 6\,\text{Å}$	$r_{\mathrm{cut}} = 7\,\text{Å}$		
^{1}H	256	128	32	8	8	1	0.1	2
^{13}C	256	512	64	8	8	1	2.0	2
^{15}N	256	128	32	8	8	1	0.1	2
^{17}O	256	128	32	8	8	1	5.0	2

Table 4.2 – Kernel weights and GPR parameters used for multi-scale kernel prediction.

CSP. Crystal structures of AZD8329 were obtained from the ESI of Ref.[276], and correspond to the 14 most stable predicted polymorphs with the cis conformation of the amide bond. From the same sources we obtained chemical shifts for each structure calculated with GIPAW[254,255] using the DFT program CASTEP[303] and the experimental chemical shifts. Labels for the different polymorphs of each structure are based on their DFT calculated energy, with one being the most stable trial polymorph of a given molecule.

Results

Training and validation using DFT calculated shifts of known crystal structures By definition machine learning models must be trained on the property that is to be predicted which should ideally be the experimental chemical shifts. However, for molecular solids there are currently only around 100 compounds with reliable crystal structures and for which assigned ^{1}H or ^{13}C shifts have been published, despite the rapidly increasing activity of NMR in crystal structure determination. This is at least an order of magnitude too few structures to hope to determine a reliable prediction model. In this light, we note that today GIPAW chemical shift calculations can accurately reproduce experimental shifts.[262,304] Thus we propose to develop a machine learning model to predict chemical shifts by training the model on a database made up of GIPAW calculated shifts from a large and diverse set of reference crystal structures. If the model can then accurately predict GIPAW chemical shifts, we hypothesize that it should also be in good agreement with experimental shifts. We also note in this context that even if there was a database of experimental shifts, there would be a challenge to machine learning related to the fact that the experiment reports on structures that include dynamics or distributions, making the connection between shifts and environments ambiguous. Learning using GIPAW calculated shifts does not suffer from this problem. The approach we take to predicting chemical shifts in molecular solids is illustrated in Fig. 4.2. We use the GPR framework[88] described in Section 4.1 to predict the chemical shift of a new atomic configuration based on a statistical model that identifies the correlations between structure and shift for a reference set of training configurations, for which the chemical shifts have been determined by a GIPAW DFT calculation.

Figure 4.2 shows the general workflow of our methods from the construction of the reference dataset to the model validation and predictions. The training (CSD-2k) and test (CSD-500)

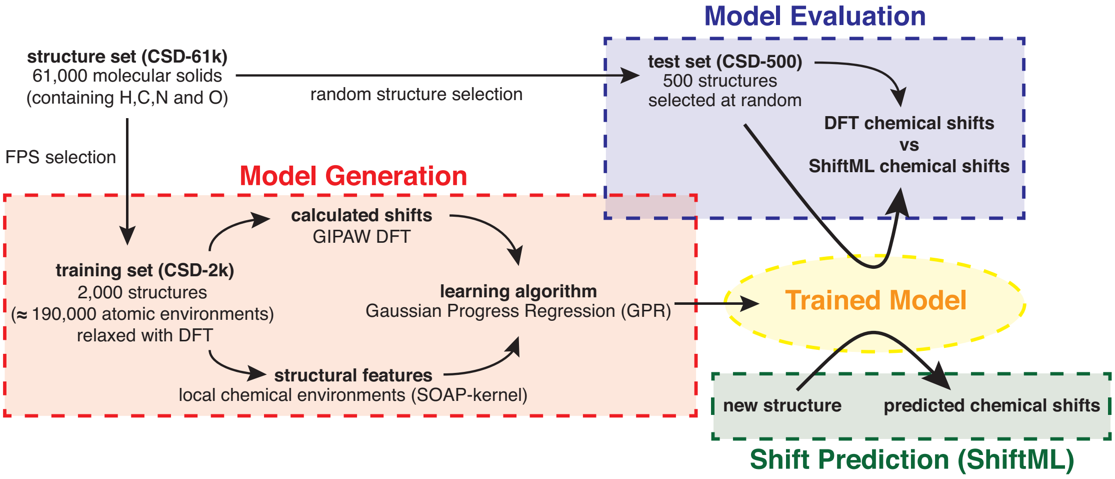


Figure 4.2 – Scheme of the machine learning model used for the chemical shift predictions.

sets are described above in Section 4.2.1. Furthermore, to avoid including spurious environments in the model, e.g. environments which might not be well described by DFT, we also automatically detect and discard from the training set atomic environments with values of the DFT calculated shifts that are anomalous based on a cross validation procedure described in Section 4.2.1. Note that using this unbiased statistical analysis we detected only a small fraction of environments as outliers (e.g. 211 out of 76214 for ^{1}H, or 0.3%).

We observe that the performance of the model degrades noticeably if one does not use this procedure. This pruning as well as the parameter optimization procedure, described below, were done exclusively using cross validation on the CSD-2k set (notably the test sets were not subject to any curation). In order to reduce the computational cost of the training and testing procedures we then finally remove from the training set all the symmetrically equivalent environments. In case of ^{1}H, this reduced the size of the training set from 70000 to about 35000 different atomic environments. The calculated chemical shieldings σ are converted to the corresponding chemical shifts δ through the relationship $\delta = \sigma_{\text{ref}} - \sigma$. Here, we used a σ_{ref} of 30.8 ppm (for ^{1}H) and 169.5 ppm (for ^{13}C), found through linear regression between the calculated and experimental chemical shifts for cocaine.

Figure 4.3 shows the chemical shift error between the DFT calculations and the ML predictions for the CSD-500 set, which is representative of the expected accuracy for the entire CSD-61k. The figure shows the overall prediction accuracy for ^{1}H chemical shifts as RMSE in ppm between the shifts calculated with DFT and with the protocol described above, which we refer to in the following as ShiftML, as a function of the cutoff radius (r_{cut}) and as a function of the number of training structures included from CSD-2k. The effect of the different cutoff radii is clearly visible. For example, for $r_{\text{cut}} = 2$Å the prediction error for a small training set (<10 structures or <100 atomic environments) can be smaller than for the larger radii, but does not improve significantly with increasing size of the training set. On the contrary, for $r_{\text{cut}} = 7$Å we observe a relatively large prediction error for a small training set, but even

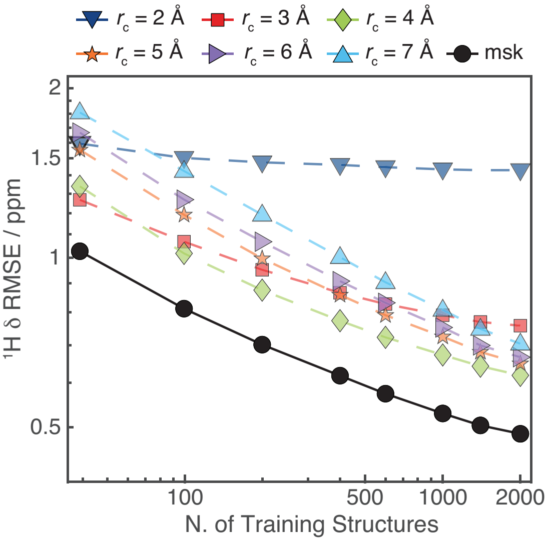


Figure 4.3 – ^{1}H chemical shift prediction error of the trained model for the CSD-500 set. The RMSE prediction error between chemical shifts calculated with ShiftML and GIPAW DFT is shown for different local environment cutoff radii, and for the multi-kernel (labelled as msk), as a function of the training set size.

with 2,000 structures (35000 environments), the prediction error is still decreasing. A similar behavior is observed in Fig. 4.4 for the prediction errors of the ^{13}C, ^{15}N and ^{17}O chemical shifts. The observed differences in the behavior of the prediction error with respect to r_{cut} clearly indicates the influence of the different extents of the local environment on the chemical shift. Short range interactions are sufficient to explain the rough order of magnitude of the shift, but long range interactions are required to learn about the higher order influences of next-nearest neighbors on shifts. However, for long range interactions, a much larger number of environments is needed in order to determine the correlation between environment and shift. We exploit these differences to generate a combined SOAP kernel consisting of a linear combination of the single local environment kernels,[114] with weightings of 256 ($r_{\text{cut}} = 2$Å), 128 ($r_{\text{cut}} = 3$Å), 32($r_{\text{cut}} = 4$Å), 8 ($r_{\text{cut}} = 5$Å and $r_{\text{cut}} = 6$Å) and 1 ($r_{\text{cut}} = 7$Å). This weighting was determined by rough optimization around values inspired by previous experience,[114] and by cross-validation on the CSD-2k training set (as described in Section 4.2.1). It is clear from Fig. 4.4 that learning with the combined kernel leads consistently to lower prediction errors than any of the single kernels, although the improvement in performance varies between nuclei. Figure 4.6a-d shows correlation plots between ^{1}H, ^{13}C, ^{15}N and ^{17}O chemical shifts calculated by DFT and by ShiftML for the CSD-500 set trained on the whole CSD-2k combined kernel. Using the combined kernel, we reach an error between ShiftML and DFT calculated chemical shifts of 0.49 ppm for ^{1}H (4.3 ppm for ^{13}C, 13.3 ppm for ^{15}N and 17.7 ppm for ^{17}O). This is very comparable with reported DFT chemical shift accuracy for ^{1}H of 0.33-0.43 ppm,13,57 while requiring a fraction of the computational time and cost: less than 1 CPU minute compared to 62-150 CPU hours for DFT chemical shift calculation on structures containing 86 atoms (around 350 valence electrons) as shown in Fig. 4.5. For the other nuclei,

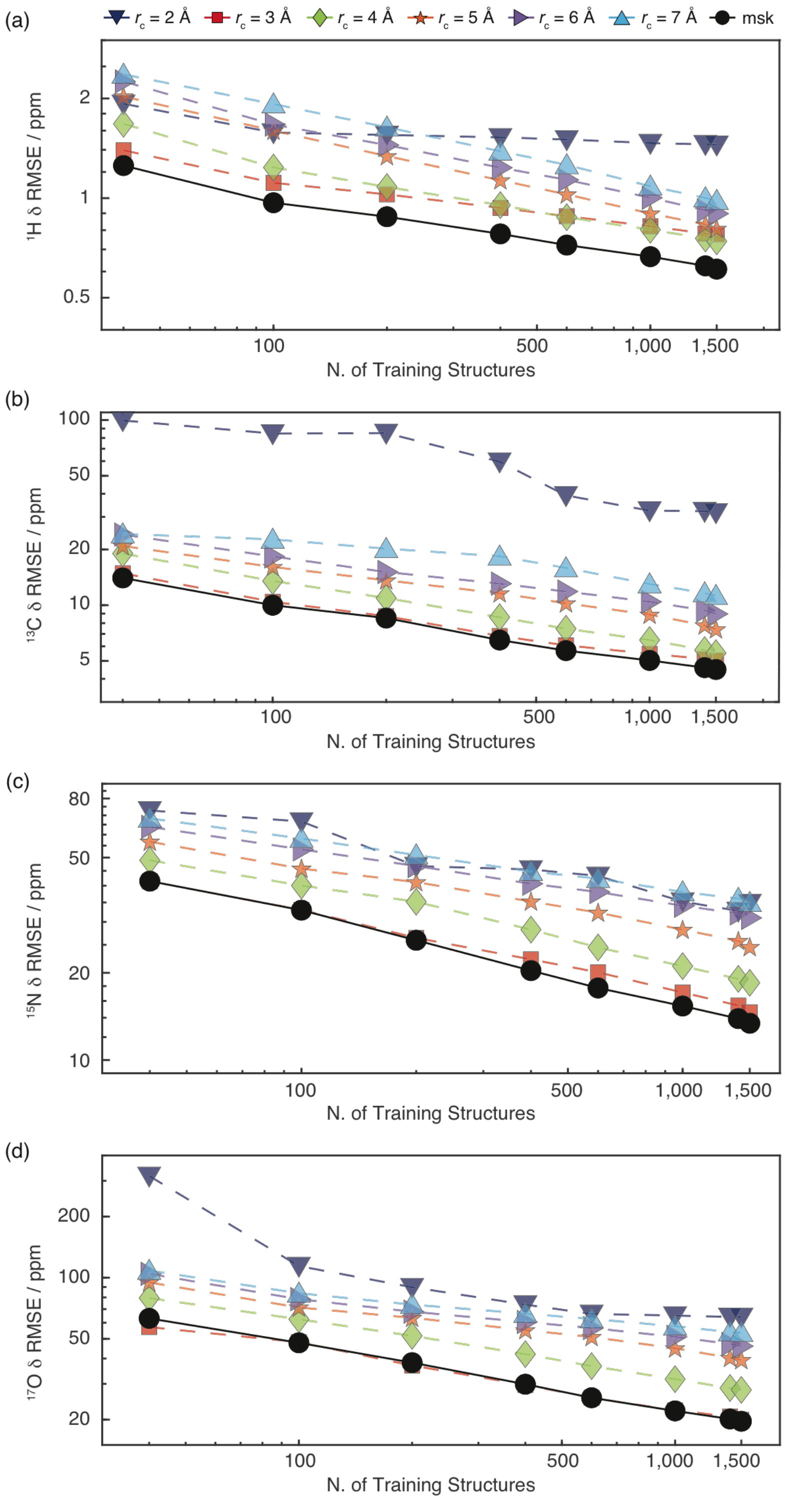


Figure 4.4 – RMSE learning curves showing the error between chemical shifts calculated with GIPAW DFT and ShiftML of the CSD-2k dataset for different local environment cutoff radii, and for the multi-kernel (labelled as msk), as a function of the training set size. The curves are for ^{1}H (a), ^{13}C (b), ^{15}N (c) and ^{17}O (d) chemical shieldings.

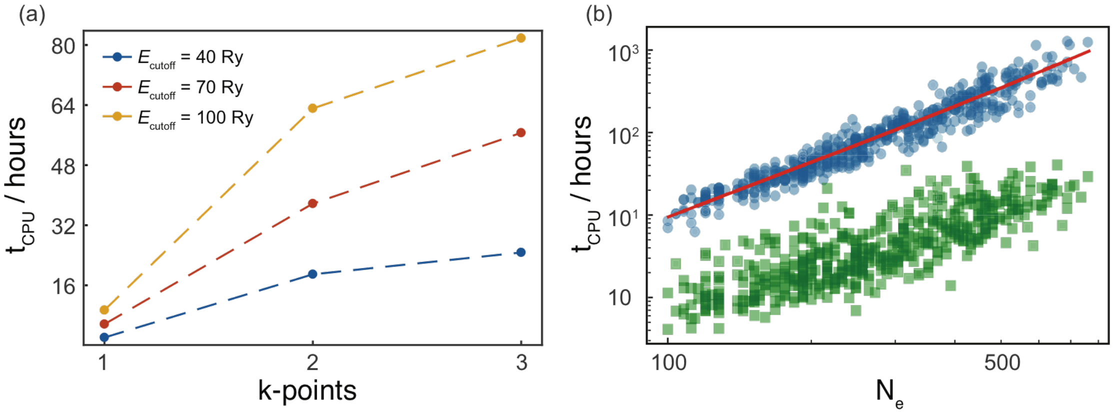


Figure 4.5 – CPU time for NMR chemical shift calculations using the GIPAW method. (a) The CPU time is shown as function of the DFT accuracy, determined by the plane-wave cutoff energy E_{cutoff} and the number of k-points in each dimension for polymorph 1 of cocaine. The charge density energy cut-offs were set to $E_\rho = 4E_{cutoff}$. (b) The CPU time is shown as function of increasing system size in CSD-2k. The green squares and blue dots show individual geometry optimization and GIPAW chemical shift DFT calculations, respectively. The red line shows the best fit between the number of valence electrons and the required CPU time as $t_{CPU} = aN_e^2 + bN_e^3$, with $a = 0.0162$ and $b = 5.91 \cdot 10^{-6}$.

the ML accuracy is slightly lower than reported values (1.9-2.2 ppm for ^{13}C, 5.4 ppm for ^{15}N and 7.2 ppm for ^{17}O),[262,305] which is not surprising as there are (currently) significantly less training environments for the heteronuclei than for ^{1}H. The R^2 coefficients between the chemical shifts calculated with DFT and with ShiftML are 0.97 for ^{1}H, 0.99 for ^{13}C, 0.99 for ^{15}N, and 0.99 for ^{17}O. Note that the CSD-500 set used for testing is selected randomly from CSD-61k and not curated. Indeed, we find that many of the atomic environments in the CSD-500 set with a relatively high prediction RMSE possess either unusual cavities inside their crystal structure, possibly indicating an organic cage surrounding non-crystalline solvent or other atoms, or exhibit strongly delocalised π-bonding networks. While there is no theoretical reason preventing the machine learning model from correctly describing such environments, they are rare and not well represented within the training set. CSD-500 thus constitutes a fairly demanding test set.

Predicting shifts for polymorphs Having evaluated the power of the trained model to predict the diverse CSD-500 set, we now look at the capacity to predict potentially subtler differences by looking at a set of polymorphs of a given structure. Figure 4.8a and b show the correlation between the ^{1}H shifts calculated by GIPAW DFT and by ShiftML for 30 polymorphs of cocaine and 14 polymorphs of AZD8329, all of which were previously generated with a CSP procedure.[265,276] The figure clearly shows that ShiftML is able to accurately predict the differences in ^{1}H chemical shift for different polymorphs.

We find a chemical shift prediction error (RMSE) between GIPAW DFT and ShiftML for ^{1}H for

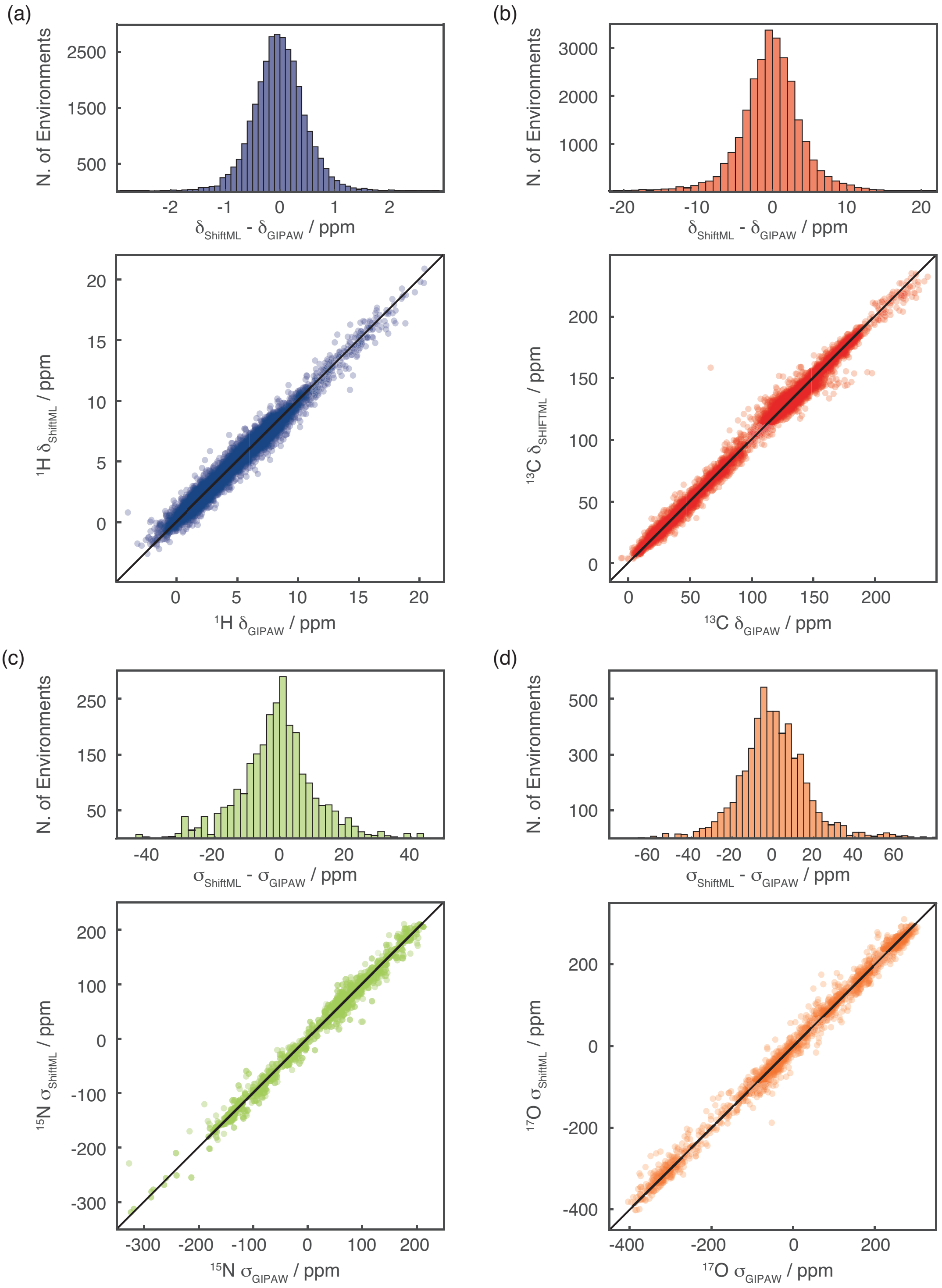


Figure 4.6 – Comparison of predictions from ShiftML and GIPAW DFT. Histograms and scatter-plots showing the correlation between ^{1}H (a), ^{13}C (b), ^{15}N (c) and ^{17}O (d) chemical shifts (shieldings) calculated with GIPAW and ShiftML. The black lines indicate a perfect correlation.

the cocaine polymorphs of 0.37 ppm and for AZD8329 of 0.46 ppm. Note that these values are slightly less than for the CSD-500 set, which might be expected when looking at these two fairly typical organic structures, and suggesting that the randomly selected CSD-500 indeed provides a good overall benchmark. Note that for these cases the DFT structure optimization and GIPAW chemical shift calculation were done with a different DFT program (CASTEP),[306] which suggests that ShiftML is robust with respect to small deviations from the fully optimized structures. For the heteronuclei we obtain an RMSE between GIPAW DFT and ShiftML for cocaine of 3.8 ppm for ^{13}C, 12.1 ppm for ^{15}N and 15.7 ppm for ^{17}O. For AZD8329 the ^{15}N and ^{17}O RMSEs are proportionally larger (17.7 and 54.7 ppm), and we attribute this to the fact that the molecule contains a rather unusual C-O$\cdots$H-N / C-O$\cdots$H-O H-bonded dimer structure, for which the learning is thus even sparser than for the heteronuclei in general. To illustrate the unusual nature of this motif, we note that the calculated ^{17}O shifts using DFT also change by up to 50 ppm for structures relaxed either by the CASTEP protocol used in Ref.[279], or the Quantum Espresso protocol used here (the RMSE between ML and DFT for the Quantum Espresso relaxed structures is reduced to 10.9 and 11.5 ppm for ^{15}N and ^{17}O). The RMSE of 4.0 ppm for ^{13}C for AZD8329 is in line with the other systems.

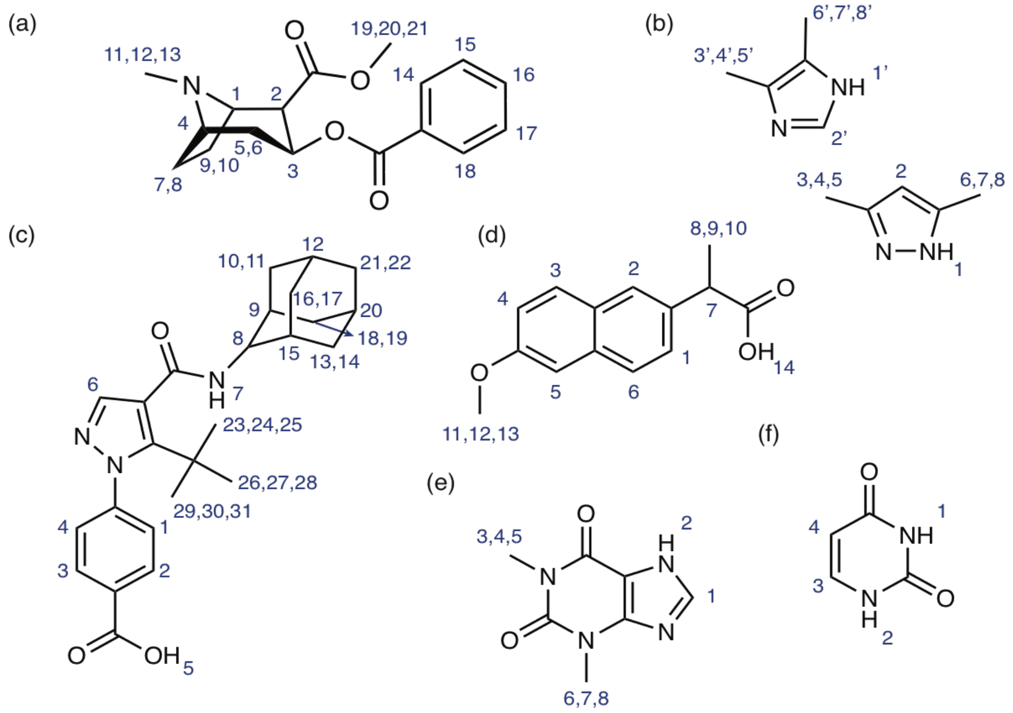


Figure 4.7 – Chemical structures of the compounds used for experimental comparison taken from Ref.[305]. In order, cocaine (a),[265] 3,5-dimethylimidazole and 4,5-dimethylimidazole (b),[307] AZD8329 (c),[276] naproxen (d),[308] theophylline (e)[265] and uracil (f),[309] and the labelling scheme used here.

Predicting experimental shifts and structure determination Further, the significance of the method is illustrated by comparison to experimentally measured shifts. This comparison is particularly important since the training protocol did not involve any experimentally measured chemical shifts. We find that the predicted shifts are accurate enough to allow crystal structure determination for both cocaine and AZD8329 from powder samples in a chemical shift driven NMR crystallography approach.

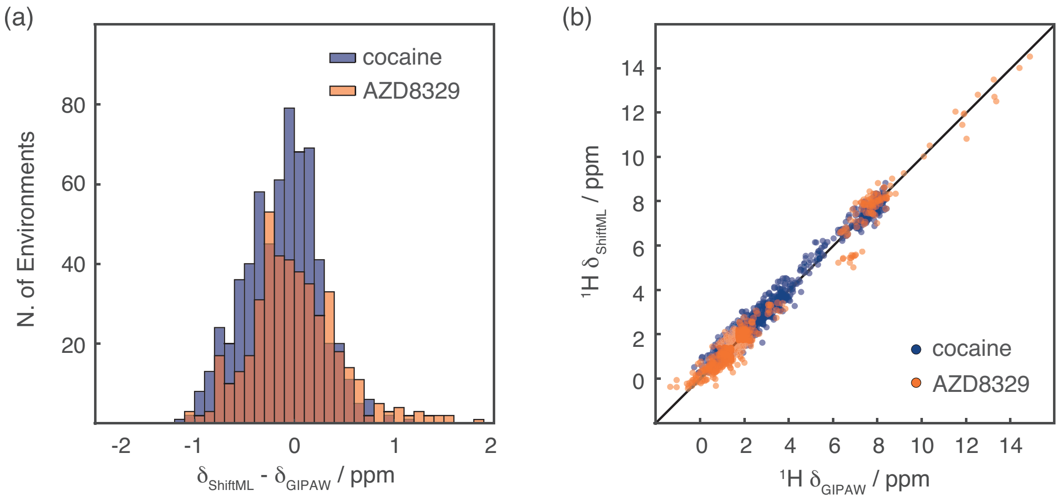


Figure 4.8 – Comparison of predictions from ShiftML and GIPAW DFT for polymorphs of cocaine and AZD8329. (a) Histogram showing the distribution of the differences between ^{1}H chemical shifts calculated with GIPAW and with ShiftML for the polymorphs of cocaine (blue), and the polymorphs of AZD8329 (orange). (b) Scatterplot showing the correlation between ^{1}H chemical shifts calculated with GIPAW and ShiftML for cocaine (blue) and AZD8329 (orange). The black line indicates a perfect correlation.

Figure 4.9a and b show the correlation between experimentally measured ^{1}H chemical shifts and the ^{1}H chemical shifts calculated by ShiftML for crystal structures and chemical shifts of the six molecules shown in Fig. 4.7. The comparison between experimental and calculated ^{1}H chemical shifts for all crystal structures (for a total of 68 shifts) gives an error (RMSE) of 0.39 ppm and a R^2 coefficient of 0.99. This compares very favorably to the equivalent agreement found between GIPAW DFT and experiment which for this set of structures is an RMSE of 0.38 ppm. Figure 4.9c and d show in blue the RMSE between DFT calculated and experimental ^{1}H chemical shifts for the 30 polymorphs predicted by CSP to have the lowest energy for cocaine and the 14 cis polymorphs of AZD8329. For both molecules the only structure in agreement with the GIPAW DFT calculations, to below a ^{1}H DFT chemical shift confidence interval of 0.49 ppm,[262] is the correct crystal structure. In the same plots we overlay the result where the experimental shifts are now compared to shifts predicted with ShiftML. Note that the RMSE between experiment and the predicted chemical shifts follows the same trends as for the DFT calculated shifts, and that here again the only structures below the confidence interval of 0.49 ppm are the two correct crystal structures. Note, that the cutoff of 0.49 ppm with respect to experiment has been evaluated for GIPAW DFT chemical shifts[262,304] and to rigorously

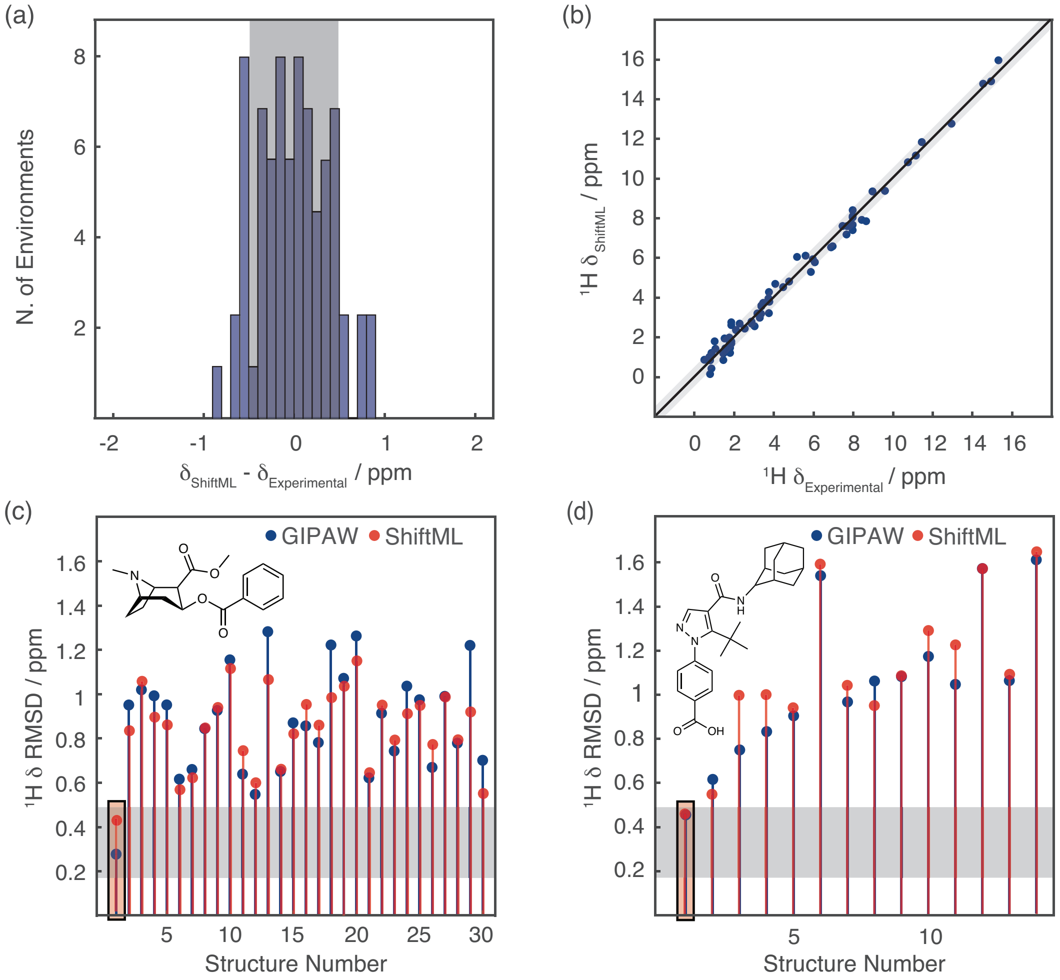


Figure 4.9 – Comparison of ShiftML to experimentally measured shifts. (a) Histogram showing the distribution of differences between experimentally measured ^{1}H chemical shifts and ^{1}H chemical shifts calculated with ShiftML for six different crystal structures (see Fig. 4.7 for the structures and numerical values of the shifts). (b) Scatter plot showing the correlation between these experimentally measured ^{1}H chemical shifts and shifts calculated with ShiftML. (c-d) Comparison between calculated and experimental ^{1}H chemical shifts for the most stable structures obtained with CSP for cocaine (c) and AZD8329 (d). For each candidate structure an aggregate RMSE is shown between experimentally measured shifts and shifts calculated using either GIPAW (blue) or ShiftML (red). The grey zones represent the confidence intervals of the GIPAW DFT ^{1}H chemical shift RMSD, as described in the text,[262] and candidates (in c and d) that have RMSEs within this range would be determined as correct crystal structures using a chemical shift driven solid-state NMR crystallography protocol.

repeat the CSP procedure for the ML method, the accuracy should be re-evaluated using more extensive benchmarking of ShiftML to experiment, which will be the subject of further work.

Predicting shifts for large structures Finally, we note that the accuracy of the method does not depend on the size of the structure, and that the prediction time is linear in the number of atoms. For the structures we calculate here the prediction time actually appears nearly constant, because it is dominated by the loading time of the reference SOAP vector (see Fig. 4.10a). We have used this method to calculate the NMR spectra (shown in Fig. 4.10b-g) for six structures from the CSD having among the largest numbers of atoms per unit cell (containing only H,C,N,O), with between 768 and 1,584 atoms per unit cell. Figure 4.10a shows the comparison between the GIPAW calculation time and the required ML prediction time. We estimate that the whole calculation would require around 16 CPU years by GIPAW. ShiftML requires less than 6 CPU minutes to calculate the shifts for all the compounds.

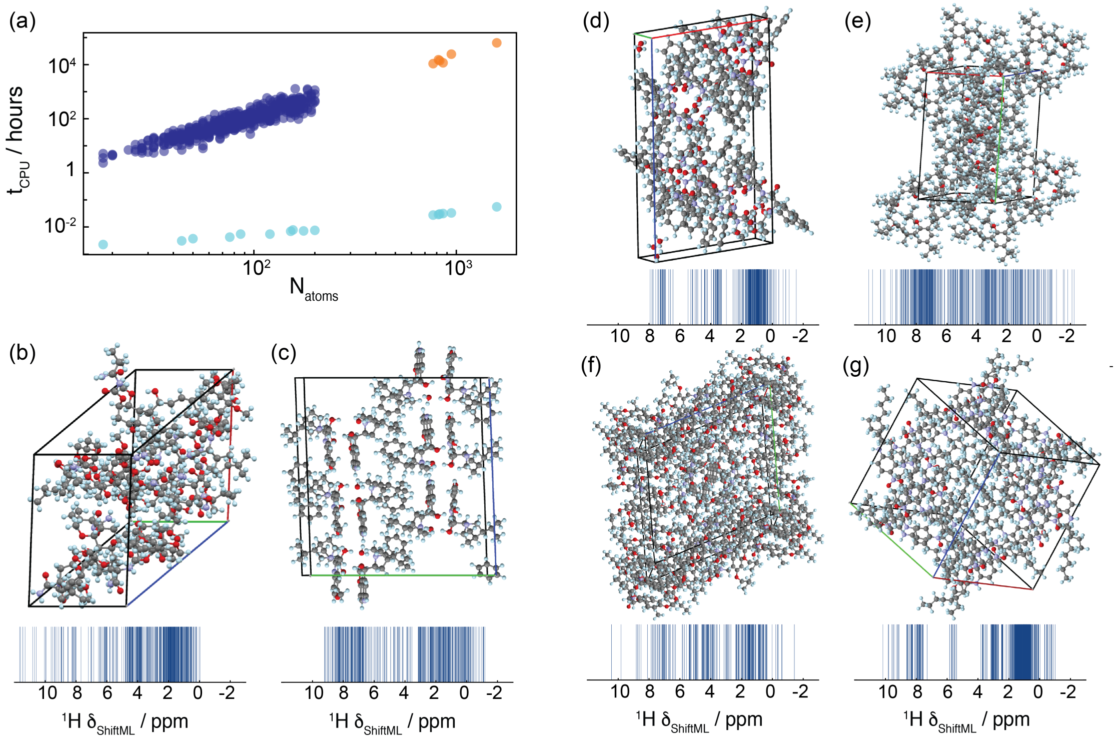


Figure 4.10 – Chemical shift calculation times and large structures. (a) DFT GIPAW calculation time (blue) and ShiftML prediction time (turquoise) for different system sizes. The GIPAW DFT calculation time for the six large structures (orange) is estimated from a cubic dependence on the number of valence electrons in the structure (see Fig. 4.5). (b-g) 3D-shemes and ^1H NMR spectra predicted with ShiftML, of the six large molecular crystals with CSD Refcodes: (b) CAJVUH,[310] $N_{atoms} = 828$, (c) RUKTOI,[311] $N_{atoms} = 768$, (d) EMEMUE,[312] $N_{atoms} = 860$, (e) GOKXOV,[313] $N_{atoms} = 945$, (f) HEJBUW,[314] $N_{atoms} = 816$, (g) RAYFEF,[315] $N_{atoms} = 1584$.

Discussion

We have presented a ML model based on local environments to predict chemical shifts of molecular solids containing HCNO to within current DFT accuracy. The R^2 coefficients between the chemical shifts calculated with DFT and with ShiftML are 0.97 for ^{1}H, 0.99 for ^{13}C, 0.99 for ^{15}N, and 0.99 for ^{17}O. The approach allows the calculation of chemical shifts for structures with 100 atoms in less than 1 minute, reducing the computational cost of chemical shift predictions in solids by a factor of between 5 to 10 thousand compared to current DFT chemical shift calculations, and thereby relieves a major bottleneck in the use of calculated chemical shifts for structure determination in solids. Far from being just a benchmark of a machine-learning scheme, the method is accurate enough to be used to determine structures by comparison to experimental shifts in chemical shift based NMR crystallography approaches to structure determination, as shown here for cocaine and AZD8329. The ML model only scales linearly with the number of atoms and, for the prediction of individual structures, is dominated by a constant I/O overhead. Here it allows the calculation of chemical shifts for a set of six structures with between 768 and 1584 atoms in their unit cells in less than six minutes (an acceleration of a factor 10^6 for the largest structure). The accuracy of the method is likely to increase further with the size of the training set, and subsequently with the future evolution of the accuracy of the method used to calculate the reference shifts used in training (here DFT), or by using experimental shifts if a large enough set were available. To simplify the dissemination of this model a web app has been developed and is publicly available at http://shiftml.epfl.ch The model used here can easily be extended to organic solids including halides or other nuclei, and to network materials such as oxides, and these will be the subject of further work.

4.2.2 Property predictions for molecular crystals[§]

The systematic design of molecular materials is a great challenge because of the competition of many weak interactions between their building blocks, i.e. constituent molecules. CSP methods have been developed to enumerate hypothetical polymorphs. Each putative crystal is ranked according to its likelihood to be observed experimentally and to its associated properties. The delicate balance between non-covalent interactions[316–318] and entropic and quantum fluctuations[319,320] call for a very precise description of the inter-molecular potential, in order to determine the cohesive energies of different polymorphs with predictive accuracy. The simplest stability indicator is the static lattice energy of the crystal which can be computed with empirical FFs or more expansive *ab initio* methods for a small pool of candidates. More-

---

[§] This section has been adapted from the journal article [84] whose authors are **Félix Musil**, Sandip De, Jack Yang, Joshua E. Campbell, Graeme Matthew Day and Michele Ceriotti. The author of this thesis analyzed the data, built and benchmarked the supervised and unsupervised ML models and wrote the methods, results and discussions sections of the article. Since this article combines both supervised and unsupervised ML models, it has been split in Section 3.5 and here to follows the logic of this thesis. Please refer to Section 3.5 for a general introduction to molecular materials and CSP, the details on the benchmark systems and the unsupervised learning side of this article.

over, the properties relevant to the design goals, e.g. electron/hole mobility for opto–electronic applications, are often modeled by expansive calculations.

In the following, we build GPR models based on the SOAP-REMatch kernel as described in Section 3.1 to reduce the cost associated with accurate lattice energy (see Section 4.2.2) and transfer integral (see Section 4.2.2) calculations in three benchmark systems[87] which are fully described in Section 3.5.1. We also propose a methodology to inform a data-driven classification of the patterns found in these datasets discussed in Section 3.5 through hyperparameters optimization with respect to the energy.

Lattice energy

CSP protocol and computational details CSP were performed with Global Lattice Energy Explorer (GLEE)[197] for possible crystal packings of a given molecules in the 23 most commonly adopted space groups for organic molecules in $Z' = 1$, and 12 common space groups for molecules that crystallize in $Z' = 2$.[321] This led to a total of 212,000 trial crystal structures, which were subsequently energy minimized in DMACRYS[198] using the W99 atom–atom intermolecular potentials[199,322–324], and multipolar electrostatics described by the distributed multipole model[200]. Duplicated crystal structures were removed using COMPACK[325] to consolidate a final list of structures for subsequent analysis.

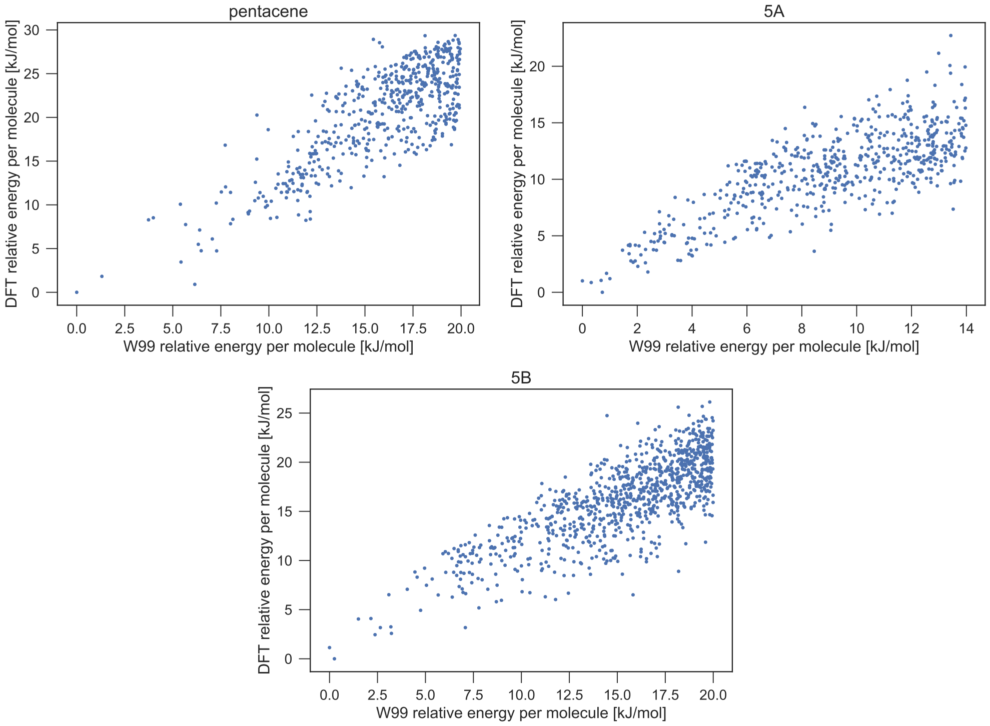


Figure 4.11 – The correlation between the W99 and DFT relative lattice energy of pentacene, 5A and 5B crystals, for W99-optimized geometries.

Single point energy calculations for the discussed set of molecular crystals have been carried out within Density functional theory (DFT) with quantum espresso code[202]. Plane wave basis set with wavefunction cutoff of 100Ry and charge density cutoff of 400Ry has been used, together with projector augmented wave (PAW) type pseudo potentials (non-linear core correction and scalar relativistic) and Perdew-Burke-Ernzerhof (PBE)[163] exchange correlation functional. To account for van der Waals interaction, Grimme's van der Waals dispersion correction[201] has been used with a cutoff radius of 80 bohr. The energy has been converged within an accuracy of 10^{-6} Hartree. The correlation between W99 and DFT energies are shown in Fig. 4.11

Hyperparameters and structural interpretation The form of the SOAP-REMatch kernels is general, and rather agnostic of the nature for the system. However, it contains many hyperparameters that can be tuned at will. The spread of the smooth Gaussians σ determines how important are small displacements of the atoms; the entropy regularization γ determines how much the combination of environments departs from a purely additive form.[114] The performance of the kernels are relatively insensitive to the value of most of these hyperparameters. The accuracy of cross-validated predictions provides an estimate of the generalization error of our models, i.e. the error for previously unseen data, which we used to optimize the performance of GPR for different systems. We found that a Gaussian width of $\sigma = 0.3$Å and a regularization $\gamma = 2$ provide the best performance for all the systems we considered.

The cutoff radius of the environment has the most significant influence on prediction performance and on the outcomes of the ML analysis. It also lends itself to a physical interpretation, since it determines the scale on which structural similarity is assessed. Although long-range electrostatics contribute significantly to the total lattice energies of crystalline structures, we found that a relatively short-range cutoff of $r_{\text{cut}} = 5$Å is sufficient to obtain remarkably accurate predictions of the reference lattice energies. This finding suggests that the most important *differences* in electrostatic interactions between competing crystal structures of a given molecule are those between nearest-neighbor molecules. It is important to note that the lattice energies were calculated using a pairwise additive force field, so the lattice energies lack contributions from polarization. Although we also observed excellent performance when predicting DFT energies, that contain full electrostatic responses, the slight degradation of the prediction accuracy suggests that a longer cutoff, or explicit treatment of the electrostatic terms, might be beneficial when learning energies that contain long-range many-body effects.

Pentacene Using the SOAP-REMatch kernel with the hyper-parameters $\gamma = 2$, $\sigma = 0.3$Å, $r_{\text{cut}} = 5$Å, the force field relative lattice energies of the pentacene crystals can be predicted with an accuracy of MAE $= 0.29 \pm 0.03$ kJ/mol and $R^2 = 0.979$ using 75% of the dataset (see Table 4.3). The learning curve for pentacene (see Fig. 4.12) shows a polynomial convergence of the error with respect to the training set size, indicating that the accuracy of the method can be improved systematically.

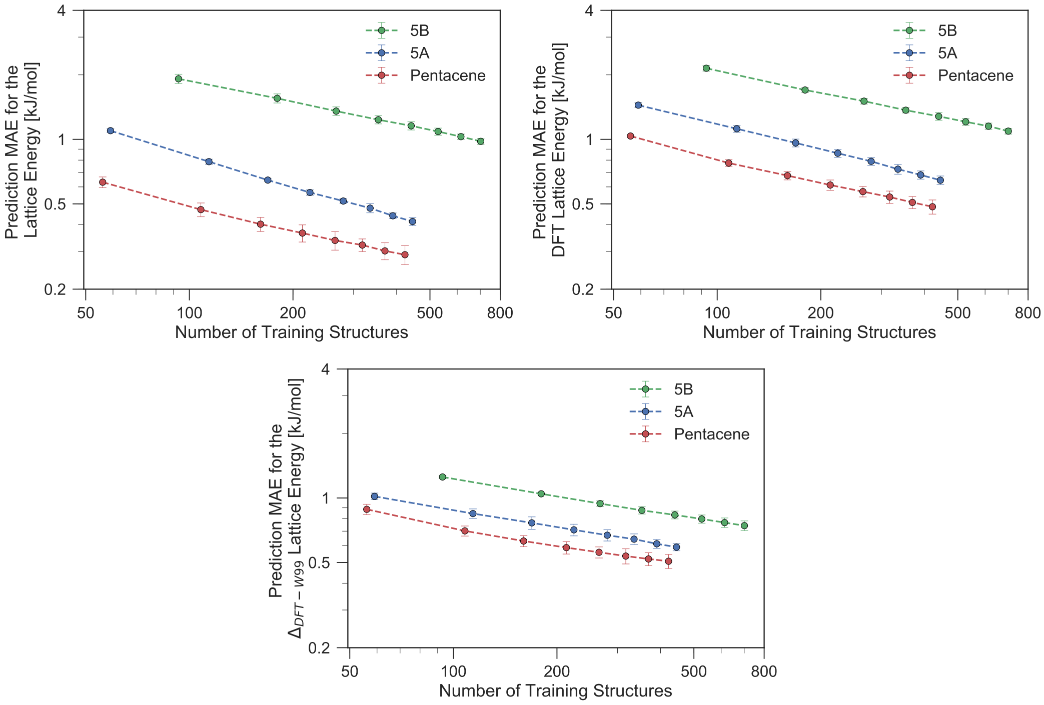


Figure 4.12 – Learning curves for the lattice energy predictions of pentacene, 5A and 5B datasets on a logarithmic scale. All hyper-parameters of our ML model are fixed except for the regularization parameter λ in the GPR model which is optimized on the fly at each training. We use 4-fold cross validation on the randomly shuffled dataset and randomly draw N times an increasing number of training samples from 75% of the dataset for each fold. The test MAE and error bars are, respectively, average and standard deviation over the folds. The left-hand panel corresponds to the prediction of W99 energies computed for W99-optimized geometries, the right-hand panel correspond to the prediction of DFT energies on such structures, and the bottom panel to the prediction of the difference between DFT and a W99 baseline.

Dataset	MAE [kJ/mol]	RMSE [kJ/mol]	R^2
Pentacene(W99)	0.29 ± 0.03	0.49 ± 0.08	0.979
Pentacene(DFT)	0.48 ± 0.04	0.68 ± 0.04	0.984
Pentacene(Δ)	0.51 ± 0.04	0.70 ± 0.06	0.96
5A(W99)	0.41 ± 0.02	0.59 ± 0.04	0.967
5A(DFT)	0.64 ± 0.03	0.91 ± 0.07	0.930
5A(Δ)	0.59 ± 0.03	0.85 ± 0.06	0.85
5B(W99)	0.98 ± 0.03	1.31 ± 0.03	0.877
5B(DFT)	1.09 ± 0.03	1.44 ± 0.04	0.870
5B(Δ)	0.74 ± 0.04	1.00 ± 0.05	0.83

Table 4.3 – Summary of the lattice energy prediction scores for pentacene, 5A and 5B (respectively 564, 594 and 936 structures). Our best accuracies on these datasets are estimated from average scores from a 4-fold cross validation (75% of the dataset is used for training). Δ-learning refers to the learning of the difference between W99 and DFT energies.

Errors in the absolute lattice energies calculated with the W99+DMA force field are, on average, about 15 kJ/mol when compared to benchmark experimental values,[326] which is 1.2 to 4 times larger than the error associated with dispersion–corrected DFT. However, these errors are largely systematic and so much of the error cancels in the evaluation of relative lattice energies. Thus, W99+DMA has been shown to be reliable in ranking the relative lattice energies in CSP studies on a large set of organic molecules[327] and was validated for this study by reproducing the known crystal structures of pentacene and an aza-substituted tetracene as global minima on their CSP landscapes.[87]

In the present study, using only a small fraction (5%) of the pentacene dataset for training, one can already very accurately reproduce the lattice energies calculated using the W99+DMA force field, with a MAE below 1 kJ/mol in the machine learned lattice energy predictions. The pentacene lattice energy landscape is dominated by the repulsion-dispersion contribution to intermolecular interactions and the above findings suggest that the predictions from the SOAP-REMatch kernel are robust in describing the relative thermodynamic stabilities of crystals of such non–polar molecules. The small fraction of structures required for training suggests that this approach could be used to reduce the cost of obtaining energy estimates at a higher level of theory, such as dispersion-corrected DFT, by performing training on a small number of high–level reference calculations. To verify this hypothesis we computed single-point dispersion-corrected DFT energies for each of the structures, which were then learned using the same kernel. As shown in Fig. 4.12, even though predictions are slightly less accurate, a ML model that uses just 50 training points can predict the DFT relative stability of different phases with a sub-kJ/mol error, opening the way to the use of more accurate energetics in large-scale CSP studies.

The quality of energy predictions based on SOAP-REMatch kernels for the predicted polymorphs of pentacene is remarkable, and the automatic classification based on kernels provides more fine-grained insights into the structural diversity in the lattice energy landscape compared to the heuristic classifications. To verify how these observations generalize to different classes of molecular crystals, we also considered the case of the two azapentacene isomers 5A and 5B.

Azapentacene 5A The quality of the lattice energy predictions for the 5A dataset is comparable to the pentacene dataset (see Table 4.3 and Fig. 4.12), showing similar accuracy estimations ($\mathrm{MAE} = 0.41 \pm 0.02\,\mathrm{kJ/mol}$ and $R^2 = 0.967$ for predicting W99 energies, and $\mathrm{MAE} = 0.64 \pm 0.03\,\mathrm{kJ/mol}$ and $R^2 = 0.930$ for DFT predictions) and trends in the learning curves. However, to reach 1 kJ/mol accuracy, we need at least twice as many training samples compared to pentacene. This can be rationalized by the introduction of stronger intermolecular electrostatic interactions involving the polar nitrogen atoms, which leads to the formation of $\mathrm{CH}\cdots\mathrm{N}$ H-bonds and the formation of molecular sheets. The presence of significant electrostatics as well as the dispersion interactions between arene rings results in a more complex lattice energy surface than that of pentacene, where dispersion interactions dominated and electrostatic contributions were small. The greater structural complexity of the landscape is reflected in the eigenvalue spectrum of the kernel matrices shown in Fig. 4.13, which decays more slowly than in the case of pentacene.

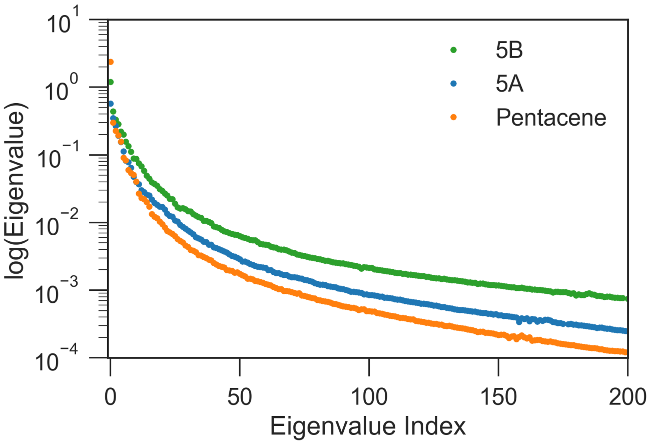


Figure 4.13 – First 200 largest eigenvalues corresponding to the centered kernel matrices[140] of pentacene, 5A and 5B datasets with cutoff radius of 5Å, gaussian width of 0.3Å and $\gamma = 2$.

Azapentacene 5B Our results on the learning of lattice energies of the 5B dataset are satisfactory, but not as good as those observed for pentacene and 5A datasets (Table 4.3 and Fig. 4.12); we reach an accuracy of about 2 kJ/mol with 100 training points and 1 kJ/mol accuracy with 75% of the dataset. Not only are the absolute errors larger, but also the slope of the learning curve is smaller, showing that it is difficult to improve the accuracy by simply including more structures in the training set.

The difficulty in learning can be traced to a higher inherent dimensionality of the dataset, as evidenced by the slow decay of the kernel eigenvalue spectrum (see Fig. 4.13). The relatively poor performance when learning lattice energies can then be understood in terms of the presence of a large number of distinct structural motifs that require a larger training set size in comparison to pentacene and 5A, which on the contrary are characterized by combinations of relatively few easy-to-rationalize and easy-to-learn stacking and H-bond patterns. Similar performance is observed when learning DFT energetics, with MAE and RMSE errors about 0.1 kJ/mol higher than learning the W99 lattice energies.

An alternative strategy for learning the DFT lattice energies is to use the W99 results as a baseline and to apply ML to predict the difference between the baseline and DFT. This approach was applied to all three molecules (Table 4.3 and Fig. 4.12). For pentacene and 5A and when using 75% of structures for training, the resulting errors are essentially the same as when learning the DFT lattice energies directly. For smaller train set sizes and for 5B, instead, this approach considerably improves the accuracy. This indicates that W99 baselining does reduce the intrinsic variance of the learning targets: given that W99 energies are an inevitable byproduct of the W99-based structure search, it is a good idea to use them as a starting point to compute more accurate lattice energies. It is however clear that the difference between W99 and DFT is a function that is as difficult to learn than the DFT or W99 energy itself, so the asymptotic accuracy is not improved much – contrary to what is observed e.g. when using a ML model to predict exact-exchange corrections to DFT, where the use of a baseline can improve the predictions by almost an order of magnitude.[32,114]

Mobility Prediction

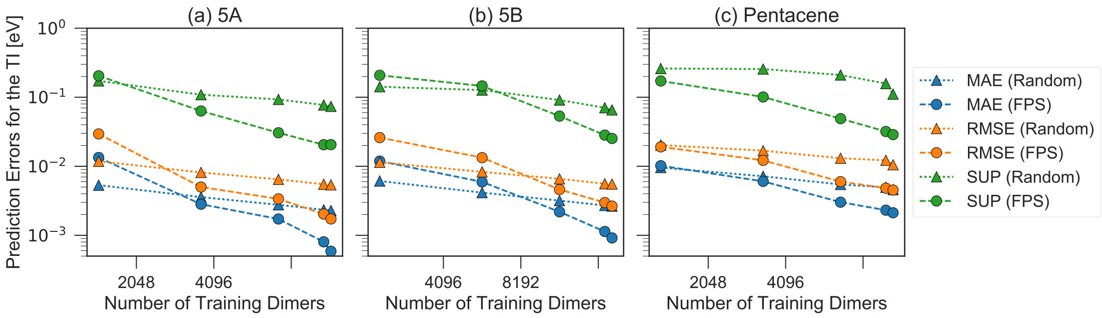


Figure 4.14 – Learning curves for the errors in predicting TI when selecting training dimers using a random or the FPS strategy. MAE, RMSE and SUP errors are defined in the text.

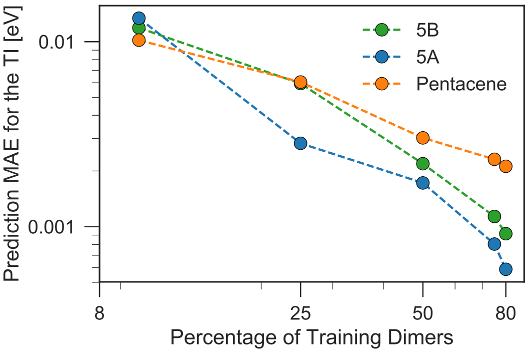


Figure 4.15 – Learning curves for the MAE in predicting TI when using the FPS selection of the training set. The three systems are compared as a function of the fraction of the total symmetry-independent dimer configurations.

Charge mobility is a key performance indicator for these set of molecular crystals considering their possible application to organic electronics. Therefore, being able to predict the hole (for pentacene) or electron (azapentacenes) mobility in putative crystal structures from CSP at a reasonable computational cost could accelerate property-driven design of functional organic semiconductors. However, contrary to the lattice energy for which bond-order expansions and additive energy models have been very successful, the charge mobility is commonly estimated through the computation of transfer integrals between pairs of molecules, each of which requires a rather demanding electronic structure calculation. The simulation protocol requires the collection from the crystal structures, up to 15 kJ/mol above the predicted global minimum for 5A, of all of the unique dimers within a specified distance cutoff, which are then used to calculate the corresponding TI values. Then instead of directly predicting the charge-carrier mobility of a given crystal structure, we thus decided to apply our ML framework to predict the value of TIs within dimers, which is the most computationally demanding part of the mobility calculation.

Computational details The charge mobility can be estimated using Einstein relationship:

$$\mu = \frac{e}{k_B T} D, \tag{4.9}$$

where e is the charge of electrons, k_B is the Boltzmann constant, T is the temperature and was set to 300 K. The electron diffusivity (D) is then evaluated as:

$$D = \frac{1}{2nM} \sum_{n=1}^{M} \sum_{j=1}^{N_i} r_{ij}^2 k_{ij} P_{ij}, \tag{4.10}$$

in which M is the total number of symmetrically independent molecules in a crystal that can be related to the Z' number for a crystal. For the i–th symmetrically independent molecule, N_i number of nearest–neighboring molecules will be extracted, which gives rise to a total of MN_i dimer pairs in the crystal structure. Symmetrically equivalent dimers based on an RMSD< 0.1Å criteria was filtered out from explicit transfer integral calculations with DFT to decrease the overall computational cost. For each dimer, r_{ij} denotes its inter–centroid distance, k_{ij} is the corresponding charge hopping rate, derived from Marcus theory:

$$k_{ij} = \frac{t_{ij}^2}{\hbar} \sqrt{\frac{\pi}{\lambda k_B T}} \exp\left[-\frac{\lambda}{4 k_B T}\right], \tag{4.11}$$

where t_{ij}, the transfer integral, describes the intermolecular electronic coupling which depends on the relative positions and orientations of the molecules in the crystal structure and λ is the intramolecular reorganization energy, and was calculated here using the conventional four–point models at B3LYP/6-311G** level of theory with GAUSSIAN09. P_{ij} is the probability for charge to hop between molecule i and j and it is related to the transfer integral as:

$$P_{ij} = \frac{k_{ij}}{\sum_{j=1}^{N_i} k_{ij}} = \frac{t_{ij}^2}{\sum_{j=1}^{N_i} t_{ij}^2}. \tag{4.12}$$

It should be clear from the above discussions that the key quantity that varies across crystal structures is t_{ij}, which is explicitly calculated with frozen–density embedding (FDE) DFT scheme. The calculations were performed at PW91/DZ level of theory with the non–additive kinetic energy modelled with PW91k functional. A threshold of $S < 10^{-2}$, below which the Penrose pseudo-inverse was applied in the final calculations of TI, was applied globally for all dimers considered, in order to avoid numerical instabilities when the orbital overlap between two monomers, S, is less than 10^{-2}. Hence our key effort here in accelerating mobility calculations will be focusing on direct prediction of t_{ij}'s for all dimers extracted from predicted crystal structures. Contrary to energies, the transfer integral is not an extensive observable because of the FDE scheme. FDE was built on the basis that the total electron densities of two interacting systems can be exactly partitioned into the sum of electron densities of two interacting systems as $\rho(\mathbf{r}) = \rho_I(\mathbf{r}) + \rho_{II}(\mathbf{r})$. In a Kohn–Sham scheme, where the total energy of the system is a functional of the total charge densities $E[\rho(\mathbf{r})]$, the same partition scheme for density does not apply for the total energy, in which a interacting non–additive component must be included as

$$E[\rho(\mathbf{r})] = E_I[\rho_I(\mathbf{r})] + E_{II}[\rho_{II}(\mathbf{r})] + E_{int}[\rho_I(\mathbf{r}), \rho_{II}(\mathbf{r})]. \tag{4.13}$$

In FDE, this is achieved by including a embedding potential $v_{emb}(\mathbf{r})$ in the Kohn–Sham equation, which takes into account contributions from non–additive kinetic and exchange–correlation energies. Furthermore, the embedding potential $v_{emb}^{I(II)}(\mathbf{r})$ acting on subsystem I (II) contains a Coulomb interaction between $\rho^I(\mathbf{r})$ and $\rho^{II}(\mathbf{r})$, and this was solved iteratively via 'freeze–and–thaw' cycles by updating the electron densities of one subsystem while keeping the other one frozen. For the evaluation of t_{ij}, one needs to introduce an additional electron/hole into the charge densities of the subsystems, thus $E_{int}[\rho_I(\mathbf{r}), \rho_{II}(\mathbf{r})]$ in Eq. (4.13) would also involve energetic contributions from polarized electron densities, which is also non–pairwise additive.

Discussion of the ML model of the TI Given that the molecules are rigid, and that the value of the TIs depends primarily on the relative intermolecular orientation, we use a simplified version of the SOAP similarity that does not require the computation of several overlap kernels for each dimer. We introduce a virtual atom situated at the center of mass of the dimers, which is used as the center of a single SOAP environment used to define the similarity between two dimers A and B. We set the environment cutoff to 10 Å, so that it encompasses the entirety of the two molecules, giving a complete information on the geometry of the dimer. We found that the accuracy of the resulting ML model obtained with this procedure is comparable to an optimized SOAP-REMatch model while being much faster to compute.

Given the total pool of dimer configurations for each system, one needs to question what is the most efficient strategy to obtain a given level of accuracy with the minimum computational effort. We considered two different strategies to determine the training structures (for which electronic structure calculations need to be performed) and the test structures (for which one would want to just use ML predictions). As the simplest possible method we considered a random selection of dimers as training references. As a second approach, we built a training set that simultaneously maximizes structural diversity while explicitly computing the value of the TI for unusual, outlier structures for which a ML prediction may fail. We do this by using FPS algorithm[122,123] which is detailed in Chapter 3.

We then used the similarity kernel of the training set to learn the TI values and perform predictions for the remaining dimers, within the GPR framework as described in Section 4.1.1, using the hyper-parameters $\zeta = 3$ and $\lambda = 5 \times 10^{-4}$ throughout. Fig. 4.14 shows the trend of the MAE, RMSE and SUP in prediction when the training set was increased systematically from 10-80% of the full set, while predicting on the remaining dimers. All systems show similar trends. The RMSE is consistently about a factor of 2 larger than the MAE, which indicates a heavy-tailed distribution of errors (for a Gaussian distribution RMSE/MAE=$\sqrt{\pi/2} \approx 1.2$).

There is a very substantial difference in the training curves between the random and the FPS selection of the training set. Similarly to what has been observed with isolated molecules,[114] a small training set size with random selection provides better MAE, since more training points are concentrated in the densely-populated portions of the structural landscape. The SUP error, however, shows that this improved MAE comes at the price of larger errors coming from the outlier structures. As the training set size is increased, the FPS learning curves decay much faster, and quickly outperform the random selection. On the one hand, this is due to the greater diversity of the training set which, for a given size, provides a relatively uniform coverage of the landscape. On the other hand, outlier configurations that may be hard to predict are computed explicitly, and so only "easy" configurations are left in the test set. Far from being an artifact of the FPS training set construction, this second element is a useful feature that can be used in a practical setting, since the selection can be performed based only on the structures. Being able to focus explicit simulations on "difficult" structures makes it possible to achieve the best overall accuracy for a given investment of computer time.

When discussing the absolute accuracy of predictions, one should keep in mind that the values of the TIs spread across several orders of magnitudes. Even when wavefunction–based methods, which were more accurate than the DFT–based method used here, were used to evaluate TIs, these could still lead to errors of the order of 5–10 meV compared to high–level reference values[328,329]; this indicates the intrinsic challenge in accurately predicting TIs. Here, it can be seen that this level of accuracy to predict DFT–derived TIs is easily achieved with about 10% of the dimer configurations, particularly if using a random selection. Using a FPS selection and increasing the training set size to about 25%, one can achieve more reliable predictions, with a MAE of about 3meV for 5A dimers, and about 7meV for 5B and pentacene (see Fig. 4.15). It is easy to see that the accuracy of predictions could be improved further. For instance, one could compute baseline values of the transfer integrals by a semi-empirical method,[32,114] or pre-select dimers with negligible TIs to reduce the computational expense. However, the present results already show that it is possible to use a straightforward ML protocol to reduce by a factor of 4-10 (depending on the desired level of accuracy) the cost of thoroughly screening all structures on a CSP landscape in terms of their charge carrier mobilities.

Conclusion

In Section 4.2.2, we have shown that sub-kJ/mol accuracy can also be obtained when predicting reference energies for the stability of different polymorphs of molecular crystals (relative lattice energies). Not only we can reproduce the energetics computed using an empirical atom-atom potential, but also predict accurately energies obtained at the dispersion-corrected DFT level. The possibility of interpolating between a few high-end reference calculations could improve the reliability of crystal structure prediction, while minimizing the added computational cost. Machine-learning models can also be used to predict properties other than polymorph stability. Given that the polyaromatic compounds studied here are relevant for molecular electronics, we chose as an example the calculation of charge mobility. In order to build a model that minimizes the investment of CPU time needed to achieve a quantitative prediction for the large numbers of crystal structures found on CSP landscapes, we focused on the bottleneck of the calculation, which is the evaluation of electronic transfer integrals between pairs of adjacent molecules. Because of their origin in the electronic structure of interacting molecules, there is no simple form for the relationship between the intermolecular arrangement and these transfer integrals. Despite the fact that transfer integrals vary over several orders of magnitude, we showed that our ML scheme could predict their value at a level of accuracy comparable to that of the electronic structure reference using only 10% of the dimer configurations – corresponding to a potential 90% reduction of the computational effort associated with the screening of crystal structures for their charge mobility.

4.3 Feature optimization for atomistic machine learning yields a data-driven construction of the periodic table of the elements[‖]

Machine-learning of atomic-scale properties amounts to extracting correlations between structure, composition, and the quantity that one wants to predict. Representing the input structure in a way that best reflects such correlations makes it possible to improve the accuracy of the model for a given amount of reference data. Indeed, after symmetries have been accounted for, there is still considerable freedom in how to define the details of an atomic-scale representation. Optimizing the input representation can improve substantially the performance of the regression, by adapting it to the specific structure-property relations associated with a given problem. What is more, in the process one can often recognize correlations that rely on intuitive information on such structure-property relations. We build on a generalization of the n-body invariant representation developed in Section 2.3 and GPR models to test adaptations of the SOAP representation to (a) the intrinsic length scales of atomic interactions, and to (b) "alchemical" correlations between chemical species, which make it possible for instance to exploit the similar behavior of different elements to accelerate learning in very chemically heterogeneous data sets. Not only do these extensions improve

---

[‖] This section has been adapted from the journal article [117] whose authors are Michael J. Willatt, **Félix Musil** and Michele Ceriotti. The author of this thesis built and benchmarked the QM9 models and wrote methods on the feature optimization and contributed to the rest of the manuscript.

significantly the performance of SOAP representations, but they do indeed offer insights into the chemistry of the system, for instance providing a data-driven representation of the similarity between elements that is reminiscent of the periodic table of the elements.

4.3.1 Benchmark Datasets

Having discussed different ways SOAP representations can be modified to represent in a more efficient way structure-property relations in complex data sets, we now verify what the practical implications of such modifications are. In order put these ideas to the test, we chose two data sets, one of which contains geometrically diverse, isolated organic molecules while the other contains elementally diverse inorganic crystals.

The QM9 data set is a collection of about 134k DFT-optimized (B3LYP/6-31G) structures of small organic molecules.[25] Each molecule contains up to nine heavy atoms (C, N, O and F) in addition to H. While the data set comprises only five atomic elements, it encompasses 621 distinct stoichiometries and is therefore very diverse geometrically. We followed Ref.[25] by removing all the 3,054 structures that failed the SMILES consistency test. The QM9 data set has been used in many pioneering studies of machine learning for molecules, notably for the demonstration of the predictive power of methods based on Coulomb matrices,[32,115] radial distribution functions[104] and SOAP.[114] It has also been used together with deep-learning schemes, such as Sch-Net[205] and HIP-NN.[223] QM9 is a very heterogeneous data set, with some stoichiometries being heavily represented, and some considerably less sampled (e.g. F-containing compounds). This, together with the fact that it has been thoroughly benchmarked with several different representations and regression strategies,[206] makes it an ideal benchmark to demonstrate the improved learning that is made possible by the scheme we introduce here.

The elpasolite data set comprises about 11k DFT-optimized quaternary structures with stoichiometry ABC_2D_6 (elpasolite $AlNaK_2F_6$ being the archetype). We have used the elpasolite data set of Faber *et al.*[33] in which the four elements constituting each structure were chosen from the 39 main group elements H to Bi. The DFT-relaxed geometries of each structure in the elpasolite crystals are almost identical which means that the data set is geometrically uniform but elementally diverse.

For each data set, we randomly selected two subsets: an optimization set (A) to be used to determine the hyperparameters of the model by cross-validation, and the other (B) to be used for training and testing. The optimizations discussed here were performed on the A set following the methods described in Sections 2.3.2 to 2.3.4, namely radial scaling, "alchemical" learning and multiple-kernel learning. Once each optimization was performed, we randomly shuffled and partitioned set B multiple times to produce training set and test set pairs. In order to account for the variability of the model accuracy with respect to the composition of the training and test sets, we averaged over the learning curves for each pair to create the figures presented here.

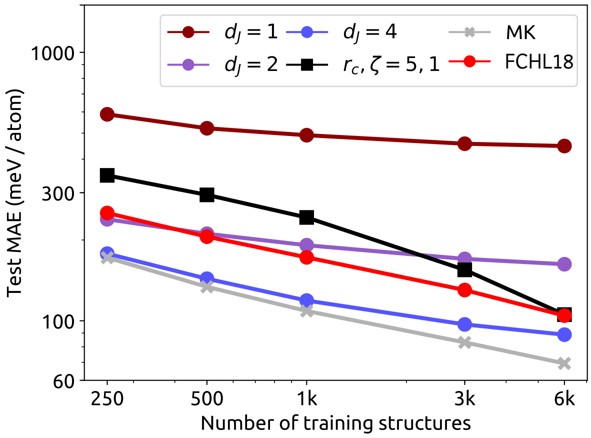


Figure 4.16 – Learning curves for the elpasolite crystals. The standard SOAP curve is shown in black, the best curve from Ref.[104] is shown in bright red and the optimized curves are shown in dark red $(d_J = 1)$, purple $(d_J = 2)$ and blue $(d_J = 4)$. For each of these models, the kernels were constructed with $r_{\text{cut}} = 5$Å and $\zeta = 1$. The multiple-kernel model (shown in grey) combines three standard SOAP kernels ($\zeta = 1$, $r_{\text{cut}} = 4$; $\zeta = 1$, $r_{\text{cut}} = 6$; $\zeta = 4$, $r_{\text{cut}} = 6$) and one optimized kernel $(d_J = 4,\ \zeta = 1,\ r_{\text{cut}} = 5)$ in the ratio $4:3:1:220$. All of the kernels were constructed with $\nu = 2$, $n_{\text{max}} = 12$ radial basis functions and $l_{\text{max}} = 9$ non-degenerate spherical harmonics. Error bars are omitted because they are as small as the data point markers.

4.3.2 Reduced-dimensionality alchemical kernels

For the elpasolite crystals, our optimization set contained 2k structures and the remainder were used to construct five training and test set pairs at random (6k and 2k structures respectively). Figure 4.16 shows the averaged learning curves. The reference curve (bright red line) was taken from Ref.[104] and corresponds to recently-proposed density-based representations.

The dark red, purple and blue curves show the result of optimizing the alchemical kernel, which we did by initializing low-dimensional u_{Ja} based on the d_J principal components of the alchemical kernel,

$$\kappa_{a_1 a_2} = e^{-(\epsilon_{a_1} - \epsilon_{a_2})^2/2\sigma_\epsilon^2 - (r_{a_1} - r_{a_2})^2/2\sigma_r^2}, \tag{4.14}$$

where ϵ_{a_1} and r_{a_1} correspond to Pauling atomic electronegativity and van der Waals radius for the element a_1. The values of u_{Ja} were then optimized with an iterative scheme working in the primal formulation of ridge regression for $\zeta = 1$. using the generic regularized loss function is defined by

$$L(\mathbf{u}, \mathbf{w}, \sigma_w; \{A\}) = \frac{1}{2} \sum_{N \in A} \left[y(N) - \langle w | N \rangle \right]^2 + \frac{1}{2} \sigma_w^2 \langle w | w \rangle, \tag{4.15}$$

where $\mathbf{w} = \{\langle n_1 J_1 n_2 J_2 l | w \rangle\}$ and $\mathbf{u} = \{u_{aJ}\}$, on which $|N\rangle$ is implicitly dependent. For k-fold cross validation, there are k optimal linear regression weights $\mathbf{w}_k$, which satisfy the vector equations

$$L_2(\mathbf{u}, \mathbf{w}_k, \sigma_w; \{A_k\}) = 0, \tag{4.16}$$

where the subscript denotes differentiation with respect to the second argument. Solving these equations, which are linear in $\mathbf{w}_k$, provides relations for $\mathbf{w}_k(\mathbf{u})$. Furthermore,

$$\mathbf{w}'_k(\mathbf{u}) = -L_{22}^{-1}(\mathbf{u}, \mathbf{w}_k(\mathbf{u}), \sigma_w; \{A_k\})\, L_{12}(\mathbf{u}, \mathbf{w}_k(\mathbf{u}), \sigma_w; \{A_k\}). \tag{4.17}$$

Having calculated these quantities, the total k-fold cross-validation error (the square of the total Root Mean Square Error),

$$L(\mathbf{u}) = \sum_k L(\mathbf{u}, \mathbf{w}_k(\mathbf{u}), \sigma_w = 0; \{A_k\}^c), \tag{4.18}$$

where the c superscript denotes the set complement, can be minimized (at least locally) by findings the roots of

$$L'(\mathbf{u}) = \sum_k L_1(\mathbf{u}, \mathbf{w}_k(\mathbf{u}), \sigma_w = 0; \{A_k\}^c) + L_2(\mathbf{u}, \mathbf{w}_k(\mathbf{u}), \sigma_w = 0; \{A_k\}^c)\mathbf{w}'_k(\mathbf{u}). \tag{4.19}$$

To optimize Eq. (4.19) with a 2-fold cross-validation error, we used the L-BFGS algorithm starting from the Cholesky factor u of the matrix κ defined in Eq. (4.14) and stopping after 500 iterations.

Note that the effective regularization parameter of the KRR model we use is given by $reg_{eff} = \sigma^2 \, \mathrm{Tr}\{K\}/var(y)N$, where σ is the reported regularization parameter, K is the kernel matrix for the training set, $var(y)$ is the variance of the training properties and N is the number of training samples.

Reducing the dimensionality of the SOAP representations by three orders of magnitude with $d_J = 1$ leads to a poor learning rate (dark red line). The learning behavior is much improved with $d_J = 2$ (purple line), which corresponds to a reduction in the dimensionality of the SOAP representations by a factor of 380. For fewer than 2k structures, the performance is better than standard SOAP (black line), but the learning rate gradually decreases (saturation) as the number of training structures increases. This suggests that the $d_J = 2$ representation is unable to represent diversity adequately in large sets of structures because of its low dimensionality, in much the same way as reducing ζ has been found to lead to saturation in SOAP models trained on the QM9 data set.[130]

By increasing d_J to 4 (blue line), which corresponds to a reduction in the dimensionality of the SOAP representations by 99%, the resulting model outperforms both the reference (bright red line) and standard SOAP models. There is still, however, a reduction in the learning rate as the number of training structures increases. Again, this is likely an indication that the low dimensionality of the representation is unable to represent diversity adequately in large sets

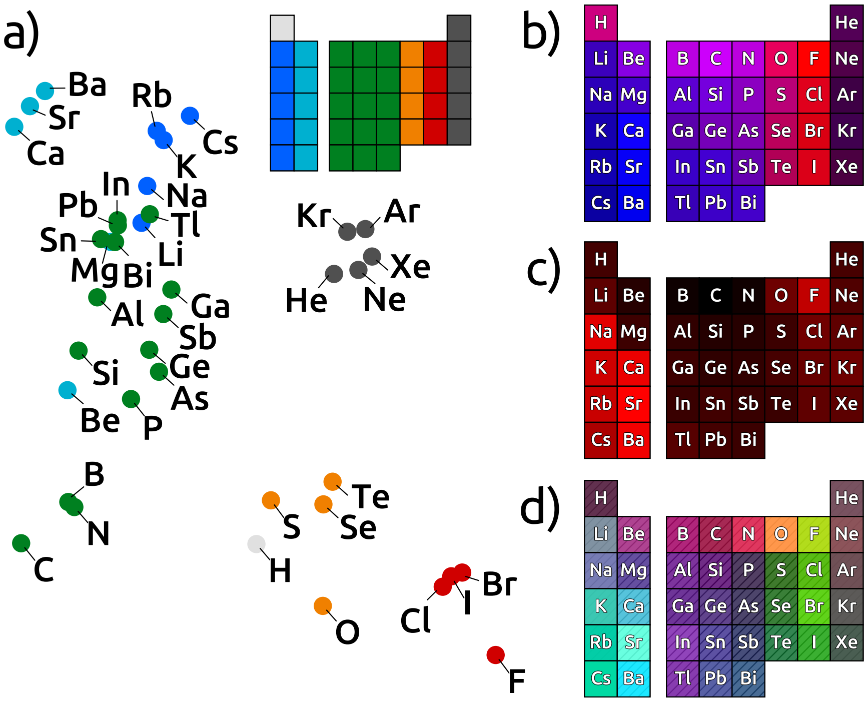


Figure 4.17 – Data-driven representations of the chemical space. (a) A 2D map of the elements contained in the elpasolite data set, with the coordinates corresponding to u_{1a} and u_{2a}, for the case $d_J = 2$. Points are colored according to the group. (b) A periodic table colored according to the coordinates in the 2D chemical space. u_{1a} corresponds to the red channel and u_{2a} to the blue channel. (c) A periodic table colored according to u_{1a} (red channel) for a 1D chemical space. (d) A periodic table colored according to 4D chemical coordinates (u_{1a}: red channel, u_{2a}: green channel, u_{3a}: blue channel, u_{4a}: hatches opacity)

of structures (in contrast to the higher-dimensional standard SOAP representation).

To test this idea, we combined multiple kernels in linear combination, including full dimensionality standard SOAP kernels for $r_{\text{cut}} = 4, 5, 6$ and $\zeta = 1, 2, 3, 4$, and the optimal alchemical kernels for $d_J = 1, 2, 4$. This multiple-kernel model (grey line) combines the optimized element correlations of the alchemical representation with the resistance to saturation of the standard SOAP representation, leading to an improvement in performance over standard SOAP and the state of the art by some 30% on the full training set. It is worth noting that our regression model also outperforms by a factor of two a recently-proposed scheme to determine similarities between elements based on artificial intelligence techniques.[330]

The performance of the model for different levels of compression of the chemical space reflect the tradeoff between the available data and the complexity of the representation. Training of the extended model entails non-linear optimization of $d_J \times n_{\text{elements}}$ weights, combined with KRR in a SOAP representation that contains d_J^2 "element channels". A low-dimensional model can extrapolate more reliably to combinations of elements that are not present in the train set, but may not have sufficient flexibility to maintain high learning rates when larger amounts of data are available. This tradeoff is evident when considering the apparent contradiction between the fact that we observed little improvement in model performance when increasing d_J beyond four, and the fact that a multi-kernel that includes full SOAP models does improve significantly the prediction accuracy. We attribute this to the fact that the number of free parameters grows steeply with d_J, which leads to failure of cross-validation scheme to extract meaningful information from the relatively small optimization set. Conversely, the multi-kernel model provides an approach to include full element information, with only a small number of hyperparameters defining how much weight this information should be given in comparison to more coarse-grained descriptions.

4.3.3 A data-driven periodic table of the elements

The eigenvectors of the alchemical kernel $\kappa_{aa'}$ lend themselves naturally to be interpreted as spanning a continuous alchemical space in which the element kets $|a\rangle$ are embedded. In other terms, they make it possible to obtain a low-dimensional representation of the elements, in which case elements that behave in a similar way with respect to the target property lie close to each other. Figure 4.17 (a) shows the optimized distribution of the elements u_{aJ} in the two-dimensional space spanned by $|1\rangle$ and $|2\rangle$ for $d_J = 2$. Elements within different groups of the periodic table are coloured differently. It is immediately apparent from this colouring scheme that optimization of the alchemical kernel leads to clustering of elements that is reminiscent of their position in the periodic table. The correlation between the data-driven element representations and the position in the periodic table is perhaps even more apparent in Fig. 4.17 (b), in which the periodic table is color-coded according to the values of u_{Ja}. This fascinating observation suggests that one could in principle construct a reasonable alchemical kernel using chemical intuition alone. However, there are two significant advantages to the approach presented here. First, the optimization is performed automatically on the data set under consideration. Second, the optimization can be performed just as well in a lower or higher-dimensional space (e.g. $d_J = 1$ or $d_J = 4$, Fig. 4.17 (c) and (d)), where intuition based on the (two-dimensional) periodic table is likely to hinder the performance of the model.

It should also be noted that the elpasolite data set consists of configurations that share the same structure, and span a space that is dominated by element correlations, making an optimization that ignores geometric correlations particularly effective. More structurally diverse data sets will imply stronger coupling between geometry and composition, making it advisable to consider more general extensions of the SOAP representations to extract comparable insight.

4.3.4 Radial scaling in the QM9 data set

Molecular databases such as the QM9 are less elementally diverse (containing only 5 elements), but contain a broad variety of structures. It has been shown that SOAP kernels can predict with great accuracy the stability of these molecules. However, reaching the best accuracy requires a combination of kernels, as in Eq. (2.67), with different cutoff radii. The combination of kernels with different length scales has been interpreted in terms of the need for encoding in the kernel the notion of multiple length scales in molecular interactions.[114] The same argument can be applied to the optimization of a radial scaling function $u(r)$ (see Section 2.3.2), so it should be possible to obtain similar accuracy to a multi-scale kernel by simply optimizing a suitable parameterization of such scaling.

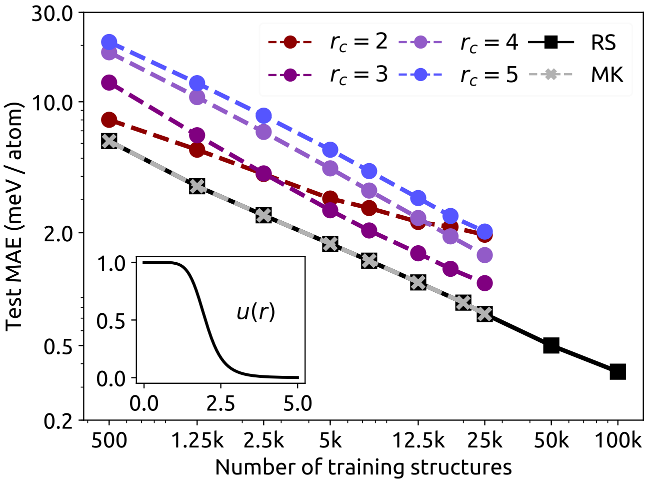


Figure 4.18 – Learning curves for the QM9 data set. Four of the lines show the MAE on the test set for various standard SOAP kernels ($\zeta = 2$) with different cutoff radii (dashed lines graduating from red to blue). The other lines show the MAE on the test set for the optimal radially-scaled (RS) and multiple-kernel (MK) SOAP models (black and grey lines respectively). In every model, the kernels were constructed with $\nu = 2$, $n_{\text{max}} = 12$ radial basis functions and $l_{\text{max}} = 9$ non-degenerate spherical harmonics. The inset shows the radial-scaling function $u(r)$ from $r = 0$Å to $r = 5$Å with the parameters that were found to minimize the ten-fold cross validation MAE on the optimization set through a grid search, $r_0 = 2$Å and $m = 7$. The multiple-kernel model combines the $r_{\text{cut}} = 2, 3, 4$ and RS kernels in the ratio $100{,}000 : 1 : 2 : 10{,}000$, and the learning curve agrees with the RS result to within graphical accuracy. Error bars are omitted because they are as small as the data point markers. Note that errors are expressed on a per-atom basis. Error per molecule expressed in kcal/mol can be obtained approximately by multiplying the scale by 0.4147, that is computed based on the average size of a molecule in the QM9 database.

Following Section 2.3.2, we consider a simple functional form with a long-range algebraic

σ	10^{-7}	10^{-6}	10^{-5}	10^{-4}	10^{-3}	0.01	0.1		
m	0	1	2	4	5	6	7	8	9
r_0	1	2	3	4					

Table 4.4 – Parameters used for the grid search of the optimal radial scaling on QM9. All the possible combinations of the three parameters were evaluated.

decay and smooth behavior at $r \to 0$,

$$u(r) = \begin{cases} 1/(r/r_0)^m & \text{if c=0,} \\ 1 & \text{if m=0,} \\ c/c+(r/r_0)^m & \text{else.} \end{cases} \tag{4.20}$$

We optimized r_0 and m and the regularization parameter (see Table 4.4 for the choice of values) using a grid search and 10-fold cross validation over an optimization set of 5,000 randomly-selected molecules with $c = 1$. The other parameters of the kernel were fixed to $r_{\mathrm{cut}} = 5$, $c = 1$, $u_0 = 1$, $\zeta = 2$, $\nu = 2$, $n_{max} = 12$, $l_{max} = 9$ and Gaussian width $g_w = 0.3$.

Figure 4.18 compares the learning curves of conventional SOAP for different cutoff radii with the best radial scaling determined on the A set. Radial scaling leads to a substantial (25%) improvement in the performance of the model.[¶] The learning rate does not decrease when the training is extended to larger fractions of the QM9. At the level of 100k reference configurations, the radially-scaled kernel achieves a MAE as low as 0.34 meV/atom, corresponding to 0.14 kcal/mol. When considering state-of-the-art results achieved in the past year using more generally-applicable representations, our optimized model achieves an improvement which is between 25 and 60%. Multi-kernel SOAP[331] yields 0.18 kcal/mol MAE, and two different neural network models reach 0.26[223] or 0.32[205] kcal/mol MAE. We also attempted to build a multi-kernel model including both conventional SOAP kernels and the best radially-scaled kernels. The improvement we could achieve is marginal, which reinforces the notion that an optimal radial scaling of the representation is essentially equivalent to an optimized combination of representations with different scales.

Although the QM9 data set exhibits a low degree of composition diversity, one can attempt to further improve the performance of the model by introducing correlations between chemical species. In this case it is necessary to use a $\zeta = 2$ exponent to incorporate many-body interactions in the regression, which makes the application of the primal-based optimization scheme we used for elpasolites impractical.[††] For this reason, and inspired by previous results based on a heuristic determination of $\kappa_{a_1 a_2}$ based on the Pauling electronegativity of the atoms,[114] we just used Eq. (4.14) and performed a grid search to find the optimal values of σ_ϵ and σ_r.

[¶] It is important to stress that the results we report here are about 20% better than those in Ref.[114], because we removed the 3,054 structures that failed the SMILES consistency test, as is done by other papers using this data set as benchmark, including Ref.[104].

[††] Note that the u_{Ja} optimized for the $\zeta = 1$ representations lead to a degradation of the accuracy when used for the $\zeta = 2$ case.

We considered the MAE averaged over 10-fold cross validation score for each combination on the range from 0.1 to 1 for σ_ϵ and 1 to 2 for σ_r with an increment of 0.1 and from 10^{-7} to 10^{-4} for σ with an increment of factors of 10. The other kernel parameters were fixed as follow: $r_{\text{cut}} = 5$, $\zeta = 2$, $\nu = 2$, $n_{max} = 12$, $l_{max} = 9$, $g_w = 0.3$, $u_0 = 2$, $c = 1$, $r_0 = 2$, $m = 7$. Note that the Pauling's atomic electronegativities and the van der Waals radii were standardized (mean is removed and scaled with the standard deviation) to simplify the determination of the search range for σ_ϵ and σ_r. Figure 4.19 shows that this simple ansatz improves by a further 10% the performance of a SOAP-based KRR model, and also combines with the optimized radial scaling to yield a model which is essentially equivalent in performance to the optimized representations of Ref.[206]. The success of the rather primitive form of this feature optimization protocol suggests that a more general strategy in which structural and chemical correlations are tuned simultaneously could improve even further beyond the state of the art.

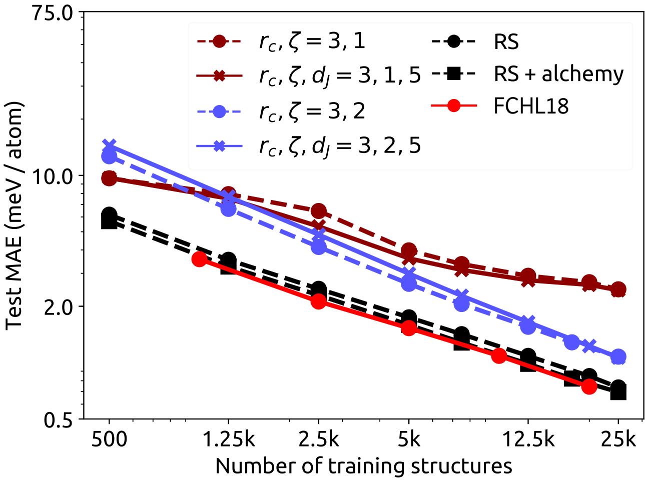


Figure 4.19 – Learning curves for the QM9 data set after inclusion of radially-scaled and alchemically-optimized SOAP kernels. Standard SOAP kernels with different cutoff radii are compared with the result of optimizing alchemical correlations using the scheme presented previously for the elpasolite crystal data set (blue and red lines). The learning curve of the optimized radially-scaled kernel (dashed black line with circles) is improved through inclusion of a Gaussian alchemical kernel (dashed black line with squares), which was optimized specifically for $\zeta = 2$ using a grid search. The combined optimization of the radial scaling and alchemical correlations leads to a model that matches the accuracy of the state of the art curve (dashed red line), which corresponds to the representations from Ref.[104], with the errors normalized by the average size of a molecules in the QM9 database. In every SOAP-based model, the kernels were constructed with $\nu = 2$, $n_{\text{max}} = 12$ radial basis functions and $l_{\text{max}} = 9$ non-degenerate spherical harmonics. Error bars are omitted because they are as small as the data point markers.

Thanks to their mathematically sound, unbiased constructions, SOAP representations are particularly well-suited to be extended, incorporating information on correlations between structure, composition and properties. We have given two examples of such extensions, representing the behavior of different chemical species as low-dimensional vectors, and modulating the information content of the representations with a radial scaling function. These optimizations improve significantly the performance of SOAP representations, matching or surpassing the state of the art on two very different data sets – a chemically diverse set of quaternary solid compounds, and a collection of small organic molecules. As we have demonstrated by re-discovering the periodic table of the elements, and extending it to one and four dimensions, they also makes it possible to extract useful insights from the inspection of the optimal combinations of features.

4.4 Fast and Accurate Uncertainty Estimation in Chemical Machine Learning[‡‡]

One of the possible approaches to quantify the accuracy of a machine learning model when presented with new inputs (the generalization error), involves measuring the residual error on a set of input-observation pairs (the test set) that are deliberately excluded from the training phase. These residual errors might be combined into a single score for the model, e.g. the Root Mean Square Error (RMSE) or Mean Absolute Error (MAE), which provides an estimate of the magnitude of the expected residual for an arbitrary input on average. Scores like the RMSE and MAE are undoubtedly useful guides, but one would often like an estimate of the error or uncertainty associated with a particular input.[233,332–336] Roughly speaking, one would like to know when the model is interpolating and when it is extrapolating (and thus likely to be less reliable). Not only can such a measure of prediction uncertainty allow the computational chemist to conclude more confidently from the model, but it can also direct the construction of the training set by highlighting important regions of the input space that are underrepresented and form the basis for an active learning strategy.[88,337]

In the following, we compare the GPR uncertainty estimator with another that is based on sub-sampling[338,339] of the training set, where multiple models are trained on different portions of the training set, and the distribution of predictions across the models is used to estimate the prediction uncertainty. We discuss ways to assess the relative performance of different uncertainty estimators and to improve the performance of an estimator by a calibration procedure based on cross-validation. We demonstrate this framework by assessing the accuracy of SOAP-GPR predictions of formation energies in the QM9[25] and Elpasolite crystal[33] datasets and ^{1}H NMR chemical shieldings in the CSD dataset.[89,90]

4.4.1 Resampling (RS)

Another approach to estimate the uncertainty associated with a prediction involves creating a family of models based on the same input data, which are representative of the statistical error

---

[‡‡] This section has been adapted from the journal article [245] whose authors are **Félix Musil**, Michael J. Willatt, Mikhail A. Langovoy and Michele Ceriotti. The author of this thesis computed the reference chemical shieldings, built the ML models for the molecular crystal and QM9 datasets and contributed to writing the manuscript.

associated with the finite amount of available training inputs.[333,338–340] Here we will discuss bootstrapping (BS) and sub-sampling (SS) techniques, which are applicable to any predictive statistical model.

Given the original training set $\mathcal{D}$ of size N, one creates N_R new datasets by drawing n input-observation pairs from $\mathcal{D}$ at random. In bootstrapping, the input-observation pairs are drawn from $\mathcal{D}$ with replacement and $n = N$, whereas in sub-sampling the selection is performed without replacement and $n < N$. Models trained independently on this ensemble of resampling datasets produce a fully non-parametric estimate of the distribution of the prediction for an input A, $P\left(y|A\right)$, whose moments can be calculated, e.g.

$$\begin{aligned} y_{RS}(A) &= \frac{1}{N_\mathrm{R}} \sum_i y^{(i)}(A) \\ \sigma^2_{RS}(A) &= \frac{1}{N_\mathrm{R}-1} \sum_i \left[y^{(i)}(A) - y_{RS}(A) \right]^2, \end{aligned} \tag{4.21}$$

where $y^{(i)}(A)$ is the prediction for the i^th resampling model. An advantage of this family of methods is that the ensemble of predictions $\left\{y^{(i)}(A)\right\}$ provides a full characterization of the error statistics which makes it possible to evaluate a non-parametric empirical model of $P\left(y|A\right)$. What is more, when considering multiple inputs $\mathcal{A}$, the ensemble of predictions relates to the fully correlated prediction distribution $P\left(\boldsymbol{y}|\mathcal{A}\right)$, making it trivial to estimate the uncertainty for any combination of the predictions.

In bootstrapping, the procedure is called the pairs resampling algorithm (as opposed to the bootstrap residuals resampling algorithm),[332,333] and it is commonly used in machine learning to construct committee models[341] and estimate prediction uncertainties. In the context of uncertainty estimation, the bootstrap variance $\sigma^2_{RS}(A)$ is sometimes used to estimate uncertainties in predictions from neural networks, where it has been found to be more reliable than alternative estimators because the variability of the predictions with respect to the initialization parameters of the neural network is incorporated automatically.[333,342]

Bootstrapping generates random samples of the right size N from the wrong distribution. On the other hand, sub-sampling generates random samples of the wrong size $n < N$ from the right distribution, provided $n \ll N$.[339] There are a variety of approaches to correct for this shortcoming of the sub-sampling approach. The most common is to assume a power law relationship between the statistic one is interested in and the size of the sub-sample n. Linear regression for a variety of sub-sample sizes then allows one to infer the exponent in the power law and extrapolate to the $n \to N$ limit.[339] Instead, to extrapolate to the $n \to N$ limit, we apply a scaling to the statistic that is optimized using a validation set, as discussed in detail later. As shown below, we have found that the criteria we use to assess uncertainty estimates suggest bootstrapping offers no advantage over sub-sampling for uncertainty estimation with the GPR PP models presented here (see Tables 4.5 to 4.7). Since sub-sampling is simpler and computationally cheaper, we focus on sub-sampling in the remainder of this article.

A practical concern regarding resampling algorithms is that they require N_R models to be trained, which increases the computational cost N_R-fold.[333,334] However, this added computational cost is associated with the training phase, whereas calculating the GPR variance is expensive in the testing phase. In most situations where one would like an uncertainty estimate for each prediction, an extended training phase is often preferable over an increased cost of making predictions. Moreover, if one exploits the model compression scheme (GPR PP) outlined in Section 4.1.2, where the training set is partitioned into active and passive components, then the computational expense of training and testing can be reduced significantly for both the resampling and GPR approaches to uncertainty estimation. Furthermore, if one uses the same representative set for all models then the cost of predicting the uncertainty for a resampling estimator becomes effectively zero. In fact, one only needs to compute once the kernel between the new input and the representative set (which is typically the expensive step), and evaluating multiple models requires only the calculation of N_R scalar products. This is in stark contrast to similar approaches based on neural networks[334,343] in which the evaluation of multiple models entails a substantial overhead. The PP approximation is, however, known to be detrimental to the quality of the GPR variance and, to a lesser extent, the prediction.

Log-likelihood assessment of uncertainty estimates

Assessing the prediction accuracy of a machine learning model is straightforward, as it suffices to compute some average of the prediction errors $y_A - y(A)$ for an appropriate validation/test set of points. However, how should one assess the quality of a model that provides an estimate of the uncertainty $\sigma(A)$ as well as of the value of the property $y(A)$? Here, we use for this purpose a log-likelihood estimator that has also been adopted for the same purpose in some classical works on statistical regression.[335,344,345]

In a nutshell, we assume that the true values of the properties $\boldsymbol{y}$ associated with the test structures $\mathcal{T}$ are uncorrelated and follow a Gaussian probability distribution,

$$P\left(\boldsymbol{y}\middle|\mathcal{T}\right) = \prod_{A\in\mathcal{T}} \frac{1}{\sqrt{2\pi\sigma^2(A)}} \exp\left(-\frac{(y_A - y(A))^2}{2\sigma^2(A)}\right), \tag{4.22}$$

whose means $y(A)$ are the GPR predictions for a reference model based on the full training set, and whose variances $\sigma^2(A)$ are determined with a statistical model – a Gaussian process or a committee of Gaussian processes in the present work (see Eqs. (4.7) and (4.21) respectively). The match between the predictions and the actual values of y can be quantified by summing the logarithms of $P\left(y\middle|A\right)$ over an appropriate test set – corresponding to the logarithm of the probability that the true targets are a realization of the model,

$$LL = \frac{1}{N_\mathrm{test}} \sum_{A\in\mathcal{T}} \log P\left(y_A\middle|A\right). \tag{4.23}$$

When using the same test set to compare two models that only differ by the uncertainty estimate, the best model will yield the highest value of LL. Note that a more general discussion of the likelihood for Gaussian probability distributions can be found in Ref.[62].

Maximum likelihood estimation for scaling uncertainty estimates

Since sub-sampling models are trained with $n < N$ input-observation pairs, each term in the joint distribution of predictions about the references over a set of sub-sampling models (see Eq. (4.22)) is likely to be too broad or narrow in general. If we assume that distributions for different inputs are broadened or narrowed by roughly the same amount, then this distortion can be corrected by scaling each term in the product by the same constant. The same approach can of course be applied to the GPR predictive distribution, to correct for the detrimental effect of a small representative set, and also to bootstrapping to correct for artificial correlations between the resampled models.

To make this notion concrete, we suppose the following form for the predictive distribution,

$$P\left(\mathbf{y}\middle|\mathcal{T}\right) = \prod_{A\in\mathcal{T}} \frac{1}{\sqrt{2\pi\alpha^2\sigma^{\gamma+2}(A)}} \exp\left(-\frac{(y_A - y(A))^2}{2\alpha^2\sigma^{\gamma+2}(A)}\right), \tag{4.24}$$

where α and γ are optimizable parameters, and Eq. (4.22) is recovered with $\alpha = 1$ and $\gamma = 0$. The calibrated uncertainty estimate for input A is then given by the standard deviations of the corresponding marginal distribution,

$$\overline{\sigma}(A) = \alpha\sigma^{\gamma/2+1}(A). \tag{4.25}$$

For $\gamma = 0$ (linear scaling) the value of α that maximizes the log likelihood over a validation set $\mathcal{V}$ of size N_{val} is simply,

$$\alpha_0^2 = \frac{1}{N_{\text{val}}} \sum_{A\in\mathcal{V}} \frac{z_A^2}{\sigma^2(A)}, \tag{4.26}$$

where z_A is the residual error for input A. For $\gamma \neq 0$ (non-linear scaling) the optimal values of α and γ can be determined easily by numerical optimization. Note that this a biased estimator α_0 for which a correction can be found in Ref.[346].

As is often the case, one should be wary of overfitting. While we mitigate this risk by a cross-validation procedure, in the context of maximum likelihood estimation it is customary to introduce a penalty for the complexity of the model. Commonly used techniques, such as the Bayesian Information Criterion,[347] or the Akaike Information Criterion,[348] also allow for an information-theoretic interpretation that justifies a comparison between models of different complexity. For the Gaussian process case, the scaling factor α_0 is the scalar product norm of the score vector for the sub-problem with N_{val} observations. The covariance matrix is the Gram matrix of the score vector. Therefore, in view of the Cauchy-Schwarz inequality for scalar products in Hilbert spaces, α_0 serves as a normalizing factor for the LL. This implies that LL corresponding to the modified distribution (Eq. (4.24)) with $\alpha = \alpha_0$ and $\gamma = 0$ can be also used as a likelihood-based test statistic for testing the above assumption of uncorrelated normally distributed reference predictions.[349] In the case of non-linear scaling with $\gamma \neq 0$, the LL for (Eq. (4.24)) is statistically equivalent to a penalized log-likelihood with the penalty depending on γ and α. In order to prevent overfitting, it might still be necessary to introduce an additional penalty for the complexity of the model. The norm of LL in this case can be used for hypothesis testing, but one has to be aware that the null distribution of the corresponding test statistic is typically more complex than in the case of linear scaling.

In a demanding, real application, removing input-observation pairs from the training set might be overly wasteful. In Ref.[342], Heskes points out that in a randomly-resampled dataset $\mathcal{D}_i$, many of the input-observation pairs in $\mathcal{D}$ will be absent and can thus be used for validation. An attractive way of optimizing α and γ without explicitly constructing a validation set is therefore the following internal validation scheme,

$$y_{\mathrm{INT}}(A) = \frac{1}{N_{\mathrm{R}}(A)} \sum_{\substack{i \\ A \notin \mathcal{D}_i}} y^{(i)}(A) \tag{4.27}$$

$$\sigma^2_{\mathrm{INT}}(A) = \frac{1}{N_{\mathrm{R}}(A) - 1} \sum_{\substack{i \\ A \notin \mathcal{D}_i}} \left[y^{(i)}(A) - y_{\mathrm{INT}}(A) \right]^2, \tag{4.28}$$

where $N_{\mathrm{R}}(A)$ is the number of resampling models that do not contain A in the training set. By using

$$P\left(\mathbf{y}|\mathcal{D}\right) = \prod_{A \in \mathcal{D}} \frac{1}{\sqrt{2\pi\alpha^2\sigma^{\gamma+2}_{\mathrm{INT}}(A)}} \exp\left(-\frac{(y_A - y(A))^2}{2\alpha^2\sigma^{\gamma+2}_{\mathrm{INT}}(A)} \right), \tag{4.29}$$

where $y(A)$ is the leave-one-out[350] prediction for input A, the log likelihood corresponding to this distribution can be maximised with respect to α and γ over the set of training inputs $\mathcal{D}$ that are absent from at least a few of the resampled models (so that Eq. (4.28) is finite and well converged for each A). In this work we only used training inputs that were absent from at least five of the resampled models. It is straightforward to show that, as the size N of the training set grows, the fraction of absent inputs for a random bootstrap sample tends to e^{-1}, while for sub-sampling it is always $1 - n/N$. This means for example that if one takes $N_{\mathrm{R}} = 10$ and $n = N/2$, slightly more than 50% of the training inputs are expected to be absent from at least five of the resampling models, and the size of the effective validation set for the procedure described above is therefore roughly half of the full training set.

By default, SOAP kernels are dimensionless. For the GPR interpretation of the kernel as a covariance to make sense, they must assume the squared units of the property one wants to predict. This is easily achieved by taking

$$K \frac{\mathrm{Var}[\mathbf{y}]}{\mathrm{Tr}\,[K]} \to K, \tag{4.30}$$

and the regularization parameter λ^2 must then be scaled by the same amount since it adopts the same units. This procedure has absolutely no effect on the prediction $y(A)$ but is essential to make the uncertainty estimate $\sigma^2_{\mathrm{GPR}}(A)$ dimensionally correct and therefore meaningful. Note that a linear scaling of the kernel with a hyperparameter ($\alpha K \rightarrow K$) is sometimes exploited to improve GPR model performance.[88] It has the same effect on the GPR variance as the linear scaling introduced earlier in Eq. (4.25) with $\gamma = 0$, provided the kernel and λ^2 are scaled by the same amount. If the regularisation parameter is not simultaneously scaled then the predictive mean $y(A)$ also changes, which is an undesirable effect in the present context since our aim is to calibrate the variation of predictive distribution about its fixed mean. If one wishes to relax this constraint by optimizing the log likelihood with respect to both λ^2 and the scale of the kernel, then the two approaches are identical. We have avoided the latter approach in the present work since it adds an extra degree of freedom to the log likelihood, which could exacerbate overfitting, and one often has a good reason for determining λ^2 in advance, especially if the variance of the noise contaminating the data points is known.

4.4.2 Benchmark Datasets

Molecular crystals dataset To generate a benchmark database for this work, that includes a more diverse set of off-equilibrium environments, the structures from the CSD-500 dataset, described in Section 4.2.1, were randomly perturbed away from their optimal geometries, so that the Root Mean Squared Deviation (RMSD) between the positions of an optimized structure and its rattled counterparts is 0.25Å and 0.5Å. We call this dataset, that contains 890 structures, the CSD-890-R. The ^{1}H chemical shieldings were calculated using the Quantum Espresso package.[202,297,351] We used PBE[163] ultrasoft pseudopotentials with GIPAW[254,255] reconstruction, H.Perdew1996-kjpaw_psl.0.1.UPF, C.Perdew1996-n-kjpaw_psl.0.1.UPF, N.Perdew1996-n-kjpaw_psl.0.1.UPF and O.Perdew1996-n-kjpaw_psl.0.1.UPF from the PSlibrary 0.3.1.[352] The wave-function and charge density energy cut-offs were set to 100 Ry and 400 Ry respectively, the convergence threshold of the self consistent cycle is set to 10^{-12} Ry and a Monkhorst-Pack grid of k-points[300] corresponding to a maximum spacing of 0.06Å in the reciprocal space. The scalar chemical shieldings are obtained from the average of the diagonal of the chemical shielding tensor using a linear response wave-vector of 0.02 bohrradius^{-1} and a convergence threshold of 10^{-14} Ry^2 for the Green's function solver.

We randomly partitioned 30k of the H environments into 20k, 5k and 5k sets. One of the 5k sets was used for validation and the other was used for testing. We sorted the 20k training environments using FPS and used the 10k most diverse environments as the representative set for the PP approach. For sub-sampling we selected 64 random sub-samples of the training set for each sub-sample size.

QM9 dataset We randomly partitioned 30k of the QM9 structures (see Section 4.3.1for more details) into 20k, 5k and 5k sets. One of the 5k sets was used for validation and the other was used for testing. We sorted the 361k environments present in the 20k training structures with

FPS and used the 5k most distant environments as the representative set in the PP approach. For sub-sampling we selected 64 random sub-samples of the training set for each sub-sample size.

Elpasolite crystal dataset We randomly partitioned the elpasolite dataset (see Section 4.3.1 for more details) into 8k and 1k sets. The 1k set was used for validation and the rest of the structures were used for testing. We sorted the 8k randomly selected structures using FPS and used the 4k most distant ones as the representative set in the PP approach. For sub-sampling we selected 64 random sub-samples of the training set for each sub-sample size (as for the molecular crystals and QM9 datasets).

4.4.3 Results and Discussion

Prediction of CSD ^{1}H NMR chemical shieldings

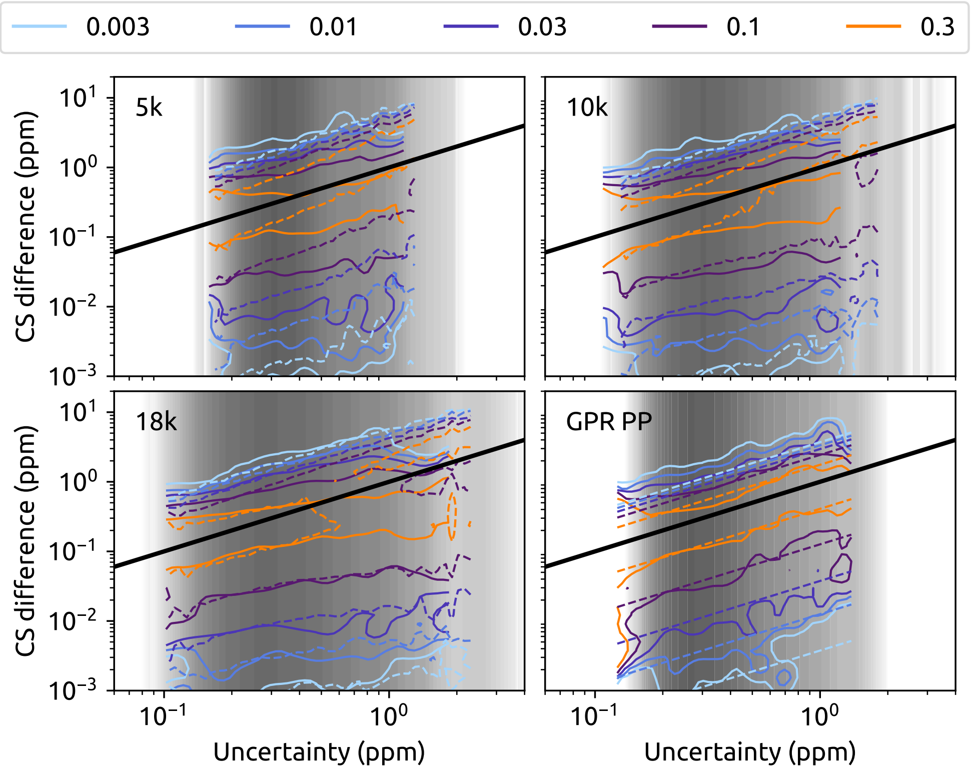


Figure 4.20 – Distribution of ^{1}H chemical shielding predictions. The colored solid lines show contours of $P(\ln\epsilon_t|\ln\sigma)$, while the colored dashed lines show contours of $P(\ln\epsilon_m|\ln\sigma)$ (see Eq. (4.31) and the corresponding contour levels are shown in the legend. The grayscale density plot and the solid black line respectively correspond to the marginal distribution of the predicted uncertainty $P(\ln\sigma)$ and to $y = x$.

Table 4.5 shows the log likelihood on the test set (Eq. (4.23)) for different sub-sample sizes, before scaling with the maximum likelihood estimation scheme and after scaling, using either a validation set or internally with the training set as described earlier. It also shows the GPR log likelihoods before and after scaling with the validation set. Note that scaling the GPR variances using the training set for internal validation is impossible, hence the corresponding cells are empty. We remark here that LL (Raw) likelihoods in the first column of Table 4.5 can be directly compared to each other, when our goal is to evaluate relative efficiency of different sub-sample sizes. This is due to the fact that such comparison is equivalent to likelihood-based model selection with an Akaike information criterion-type (AIC) penalty.[348] Indeed, AIC penalties depend only on the problem's dimensionality and not on the sub-sample size, so comparing penalized likelihoods in the AIC framework coincides with comparing LL likelihoods in the first column of the Table.

This is a convenient feature of our approach, as, in general, values of log-likelihoods for different sub-problems cannot be directly compared to each other.[349,353] Strictly speaking, the powerful machinery of maximum likelihood only guarantees good properties of the best solution, but does not always directly induce a quality scale to rank other solutions via the use of the likelihood function. Before scaling, the log likelihood shows considerable variability between resampling estimators obtained with different sample size. It appears that the estimator based on sub-samples of 5k environments (i.e. one quarter of the training set) strikes the best balance between resampling from the correct distribution but with the wrong size (small sub-sample sizes), and resampling from the wrong distribution but with the right size (large sub-sample sizes). Rescaling the uncertainty estimator leads to a substantially more stable, standardized version of the log-likelihood, and reduces greatly the impact of the sample size for RS estimators. Additionally, we observed that after rescaling, the GPR-PP estimator leads to noticeably higher log likelihood.

The log likelihood provides a measure of the accuracy of the uncertainty estimation that is quantitative but hardly intuitive. To provide a more straightforward representation of the accuracy of an uncertainty estimator, we observe that in an ideal scenario, the distribution of actual errors relative to the reference should match the distribution of the predictions of RS models around their mean. The equality of the distributions should be true for an arbitrarily-selected subset of the test set.

Based on this observation, we computed the distribution of actual errors $\epsilon_t(A) = \left|y(A) - y_A\right|$ conditioned on the value of the predicted uncertainty $\overline{\sigma}(A)$ (calibrated with a linear scaling), and the distribution of model errors $\epsilon_m(A) = \left|y^{(i)}(A) - y(A)\right|$. Given that the predicted (and actual) errors can span a broad range, we computed the conditional on a log scale, e.g.

$$P\left(\ln\epsilon|\ln\sigma\right) = P\left(\ln\epsilon, \ln\sigma\right)/P\left(\ln\sigma\right). \tag{4.31}$$

The plots comparing the predicted and actual error distributions from the linearly scaled estimators are shown in Fig. 4.20. One sees in all cases there is a good qualitative agreement between the distribution of the model (which is Gaussian by construction for the GPR model,

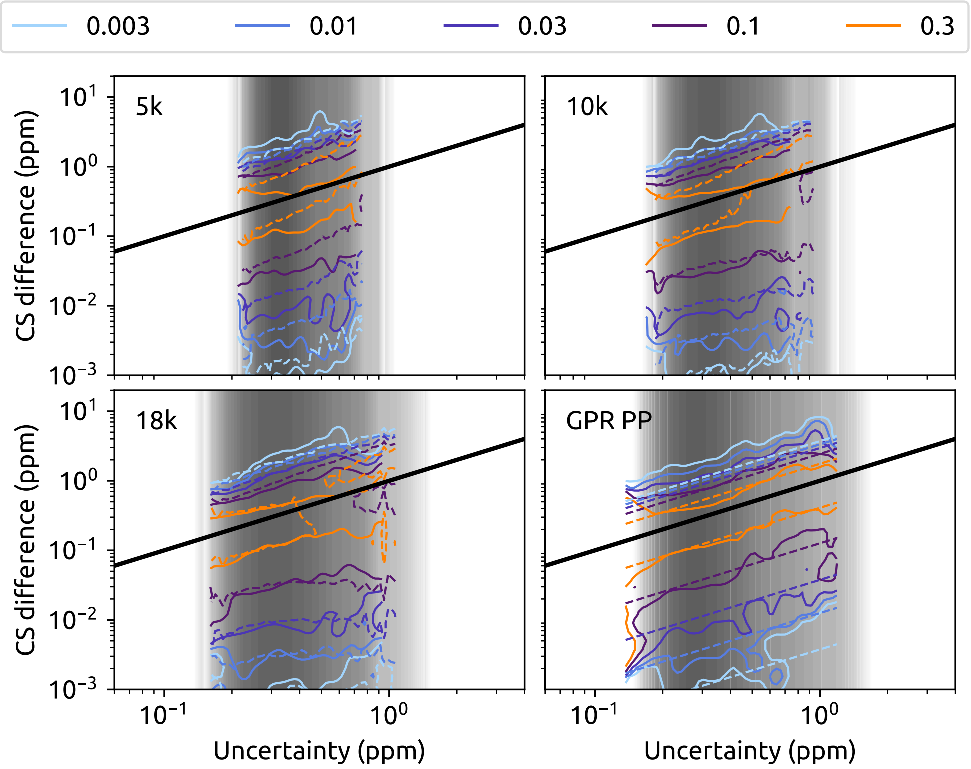


Figure 4.21 – Distribution of ^{1}H chemical shielding predictions. The solid lines show contours of $P(\ln\epsilon_t|\ln\sigma)$, while the dashed lines show contours of $P(\ln\epsilon_m|\ln\sigma)$ (see Eq. (4.31)), including a non-linear scaling of the uncertainty corresponding to Eq. (4.25) with $\gamma \neq 0$, and the corresponding contour levels are shown in the legend. The grayscale density plot and the solid black line respectively correspond to the marginal distribution of the predicted uncertainty $P(\ln\sigma)$ and to $y = x$.

and in some cases strongly non-Gaussian for RS estimators), with essentially none of the samples with large true errors being associated with a small $\overline{\sigma}(A)$.

Method	Active/Train Size	LL (Raw)	LL (Val.)	LL (Val./N-L)	LL (Int.)	LL (Int./N-L)
SS	5k/1k	1.989	2.178	2.203	2.177	2.202
	5k/5k	2.188	2.19	2.222	2.190	2.221
	5k/10k	1.846	2.21	2.243	2.210	2.243
	5k/15k	-0.534	2.214	2.259	2.203	2.257
	5k/18k	-10.546	2.203	2.261	2.048	2.242
BS	5k/20k	1.409	2.211	2.249	2.211	2.249
GPR PP	5k/20k	-3.054	2.301	2.302	N/A	N/A

Table 4.5 – Log likelihood (LL) of chemical shielding predictions on the test set for different sub-sample sizes. After scaling the variances through maximum likelihood estimation – internally (Int.) or on the validation set (Val.) – the final log likelihood is insensitive to the sub-sample size. A non-linear scaling of the uncertainty (N-L), i.e. $\gamma \neq 0$, further improves the quality of the uncertainty estimates. To normalize the results, the log likelihood (-2.560) of a model with a constant mean and variance corresponding to the empirical chemical shielding mean and variance of the full training set has been subtracted from each value.

On the other hand, there are also substantial differences between the various estimators. An obvious difference is the range spanned by the predicted $\overline{\sigma}(A)$. One could argue that - for a given LL - the model that spans the broader range of values is the most useful, as it yields better resolution between more or less trustworthy predictions. From this point of view, large-sample-size RS estimators appear to be superior, spanning almost two orders of magnitude in the value of $\overline{\sigma}(A)$. The GPR PP model, however, clearly displays the best agreement between predicted and actual error distributions, which is consistent with the higher LL. Looking more carefully at the distributions for the RS models, one can see that the actual errors tend to increase monotonically as a function of $\overline{\sigma}(A)$, even though they do not follow the trend predicted by the sample distribution. This suggests that the performance of the estimator can be improved by optimizing away from $\gamma = 0$ in Eq. (4.25). As shown in Fig. 4.21 this procedure improves the agreement between the RS and the actual error distribution, and brings the LL close to the level of the GPR PP estimator (see also Table 4.5). The observed improvement in fit hints at the possibility that the data are a better match to a light heavy-tailed process[354] rather than to a standard Gaussian process. This is entirely plausible due to the high complexity of molecular models and high dimensionality of the associated data.

Prediction of QM9 formation energies

Method	Active/Train Size	LL (Raw)	LL (Val.)	LL (Val./N-L)	LL (Int.)	LL (Int./N-L)
SS	5k/1k	4.145	4.465	4.465	4.466	4.466
	5k/5k	4.395	4.485	4.486	4.485	4.484
	5k/10k	3.462	4.491	4.492	4.492	4.491
	5k/15k	0.030	4.495	4.496	4.492	4.488
	5k/18k	-10.503	4.492	4.492	4.450	4.452
BS	5k/20k	2.997	4.496	4.496	4.497	4.493
GPR PP	5k/20k	4.151	4.178	4.22	N/A	N/A
	10k/20k	3.981	4.366	4.428	N/A	N/A

Table 4.6 – Log likelihood (LL) of formation energy predictions (QM9 dataset) on the test set for different sub-sample sizes. After scaling the variances through maximum likelihood estimation – internally (Int.) or on the validation set (Val.) – the final log likelihood is insensitive to the sub-sample size. To normalize the results, the log likelihood (0.207) of a model with a constant mean and variance corresponding to the empirical formation energy mean and variance of the full training set has been subtracted from each value.

Table 4.6 is the analogue of Table 4.5 for the QM9 dataset. The same trend is observed as for the CSD dataset with the 5k sub-sampling estimator as the most reliable before scaling through maximum likelihood estimation. Despite the differences between predicting chemical shieldings (a local property) and formation energies (a global property), we see again that the (non-linear) scaling procedure with a validation set makes the GPR and the sub-sample uncertainty estimators more or less equally reliable.

Figure 4.22 is the QM9 analogue of Fig. 4.21, reporting a more analytical representation of the correspondence between predicted and actual errors for the non-linearly scaled models. Again, the fundamental assumption of sub-sampling – that the sub-samples are to the reference what the reference is to the target – appears to be reliable. As for the CSD chemical shielding results, we found the quality of this agreement to be roughly the same, regardless of the sub-sample size. In this case, even after non-linear scaling, the GPR estimator yields a narrow uncertainty distribution, while the RS models accurately predict the uncertainty over a span of two orders of magnitude.

The fact that in the QM9 dataset about 3k molecular configurations are tagged as 'unreliable' (as their SMILES strings after geometry optimization differ from those of the corresponding starting configurations) provides an interesting benchmark for the uncertainty estimation. In Fig. 4.22 we also show the predicted variance - actual error pairs for 100 randomly selected SMILES-inconsistent compounds. Note that no SMILES-inconsistent structures were used in the training, validation and testing of the models reported in the table; the randomly-selected 100 were only added to the test set when making the figure. While most of the structures

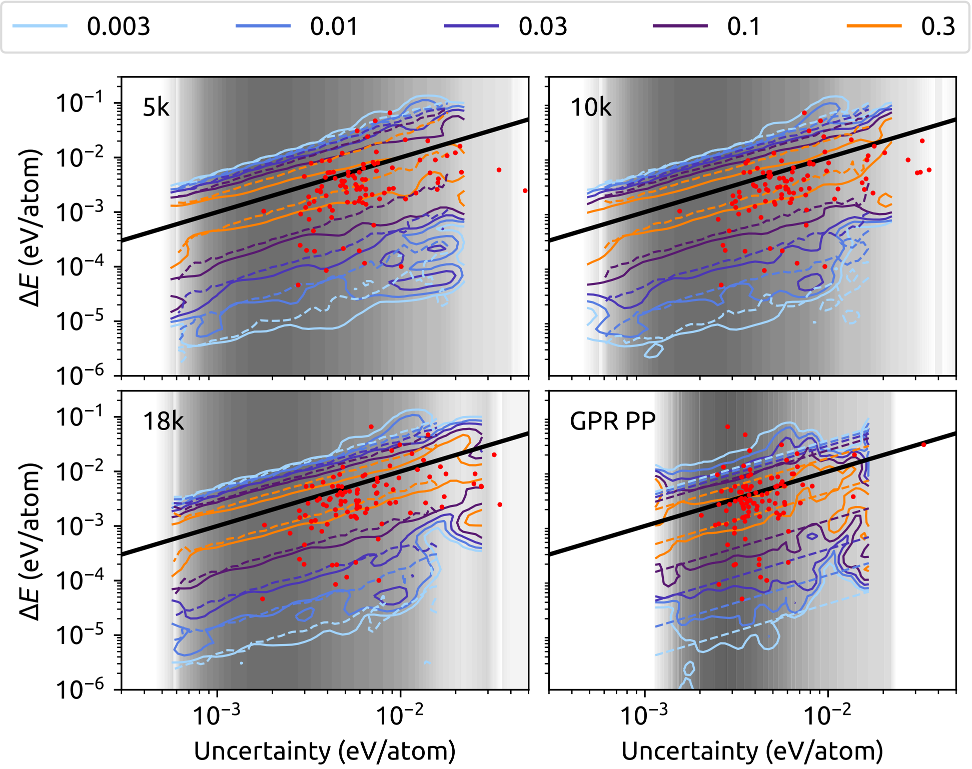


Figure 4.22 – Distribution of formation energy differences for the QM9 dataset. The colored solid lines show contours of $P(\ln\epsilon_t|\ln\sigma)$, while the colored dashed lines show the contours of $P(\ln\epsilon_m|\ln\sigma)$ (see Eq. (4.31)), including a non-linear scaling of the uncertainty corresponding to Eq. (4.25) with $\gamma \neq 0$, and the corresponding contour levels are shown in the legend. The grayscale density plot and the solid black line respectively correspond to the marginal distribution of the predicted uncertainty $P(\ln\sigma)$ and to $y = x$. The red dots show the distribution of actual errors and predicted uncertainties for 100 randomly selected structures that failed the SMILES consistency test when the QM9 dataset was constructed.

lie in the high-predicted-variance range of the data, reflecting the fact they have somewhat unusual structure, they span an order of magnitude range of $\overline{\sigma}(A)$, and there are several SMILES-consistent structures having larger predicted (and actual) errors. This observation underscores the fact that predictive uncertainty estimates can be a better guide than heuristic arguments to detect outliers and to identify structures that are needed to enlarge a training set.

Prediction of Elpasolite crystal formation energies

Method	Active/Train Size	LL (Raw)	LL (Val.)	LL (Val./N-L)	LL (Int.)	LL (Int./N-L)
SS	4k/1k	1.771	1.794	1.802	1.665	1.665
	4k/2k	1.772	1.783	1.798	1.711	1.711
	4k/4k	1.716	1.723	1.794	1.717	1.745
	4k/6k	1.103	1.695	1.800	1.691	1.761
BS	4k/8k	1.646	1.697	1.798	1.696	1.759
GPR PP	4k/8k	1.660	1.779	1.791	N/A	N/A

Table 4.7 – Log likelihood (LL) of formation energy predictions (Elpasolite crystal dataset) on the test set for different sub-sample sizes. After scaling the variances through maximum likelihood estimation – internally (Int.) or on the validation set (Val.) – the final log likelihood is insensitive to the sub-sample size. A non-linear scaling of the uncertainty (N-L) further improves the uncertainty model. To normalize the results, the log likelihood (-1.353) of a model with a constant mean and variance corresponding to the empirical formation energy mean and variance of the full training set has been subtracted from each value.

Table 4.7 shows the log likelihood of formation energy predictions on the Elpasolite crystal dataset, and Fig. 4.23 shows the distributions of actual and predicted errors after non-linear scaling. While there is a large variation in the log likelihoods before scaling, all the uncertainty estimators appear to become more or less equally effective after optimizing the log likelihood with respect to α and γ, with the GPR PP uncertainties faring very slightly worse. Interestingly, the effect of introducing a non-linear scaling offers no advantage over a linear scaling according to the log likelihood values. Also, using the training set for internal validation leads to significantly worse resampling uncertainty estimators, in contrast to the CSD and QM9 results.

We speculate that these minor differences in behavior can be traced to the discrete structure of the Elpasolite dataset; whereas CSD and QM9 are chemically homogeneous and contain a broad variety of atomic structures, Elpasolite crystals possess essentially a fixed geometry, and differ primarily by the chemical composition, that spans quaternary combinations of 39 elements, suggesting that the uncertainty estimation framework works even when the dataset reflect discrete chemical differences rather than smooth position variables.

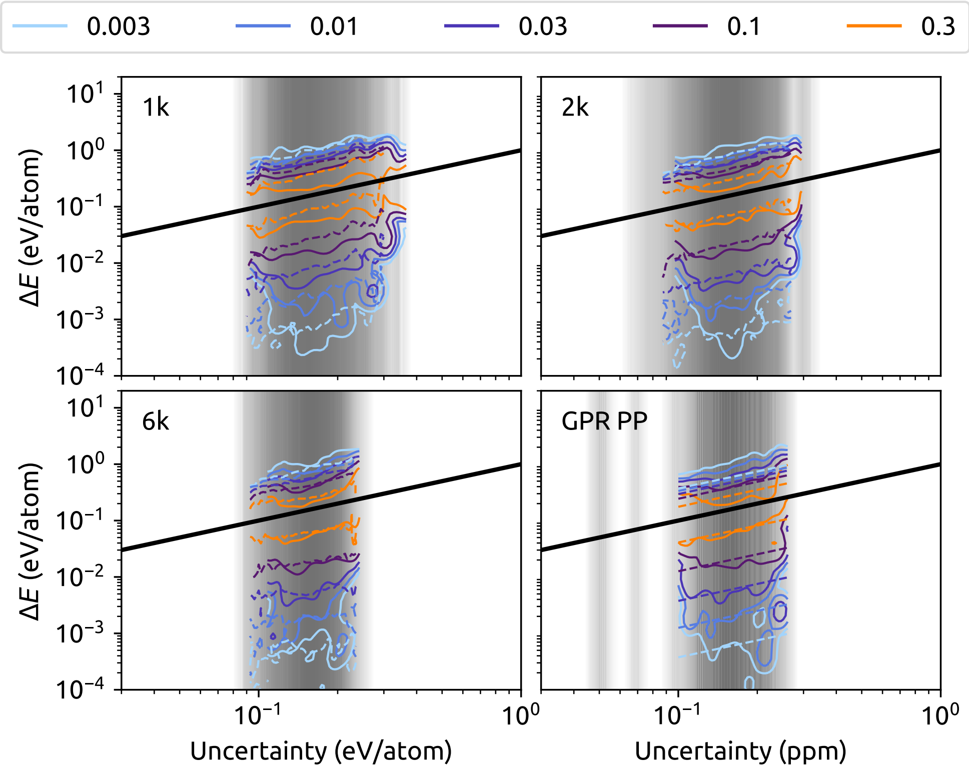


Figure 4.23 – Distribution of formation energy differences for the Elpasolite crystal dataset. The colored solid lines show contours of $P(\ln\epsilon_t|\ln\sigma)$, while the colored dashed lines show contours of $P(\ln\epsilon_m|\ln\sigma)$ (see Eq. (4.31)), including a non-linear scaling of the uncertainty corresponding to Eq. (4.25) with $\gamma \neq 0$, and the corresponding contour levels are shown in the legend. The grayscale density plot and the solid black line respectively correspond to the marginal distribution of the predicted uncertainty $P(\ln\sigma)$ and to $y = x$.

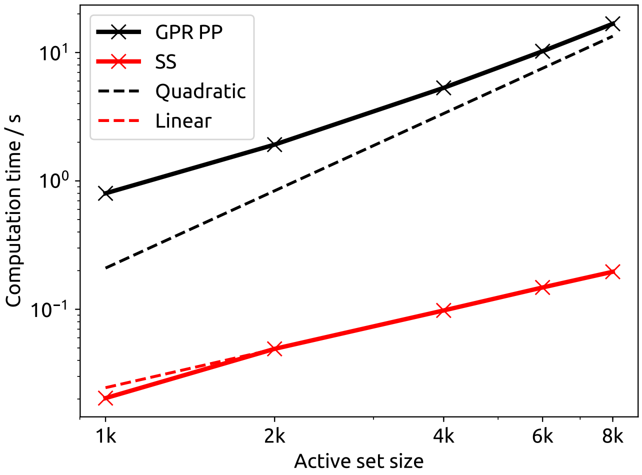


Figure 4.24 – Time taken to compute the GPR PP and sub-sampling (SS) uncertainty estimates (Eqs. (4.7) and (4.21) respectively) as a function of the active set size for 2.5k Elpasolite crystal structures as a test set. 64 training set sub-samples were used to produce the sub-sampling result (i.e. 64 linear regression weights). The time taken to generate the kernels between active and test points is excluded from the reported computation times. The simulations were performed using the NumPy module in Python for the necessary linear algebra (one Intel(R) Xeon(R) CPU E5-4627 v2 @ 3.30GHz core).

We use this dataset to also provide a benchmark of the overhead associated with uncertainty estimation. Figure 4.24 gives a comparison of the computation times of evaluating the GPR PP and sub-sampling uncertainty estimates for the Elpasolite crystal dataset. The time taken to compute kernels between active and test points is excluded, since those quantities are necessary for the property estimation, and are therefore already available to compute the uncertainty. We also do not consider the cost of the training phase: as stressed earlier, the main computational advantage of sub-sampling over GPR PP for uncertainty estimation is expected in the testing phase. The figure shows clearly that the sub-sampling approach to uncertainty estimation is significantly cheaper than the GPR PP approach in the testing phase. Sub-sampling is expected to scale linearly with the active set size M because the computationally demanding step is the evaluation of vector dot products of length M. On the other hand, GPR PP is expected to scale quadratically with M because, before taking vector dot products of length M, one of the vectors must be multiplied by an $M \times M$ matrix. The timings shown in the figure reflect these asymptotic considerations.

5 Conclusions

In this thesis, we have presented an array of methodological improvements accelerating several aspects of atomic-scale modelling by using ML algorithms. We have introduced a general formulation of the problem of representing atomic structures in terms of a (smooth) atom density, which is independent of the basis that is used to expand it. Starting from a representation of a 3D structure in terms of a superposition of atom-centered functions decorated with elemental kets, we introduce symmetries by formally averaging the feature vectors over the continuous translation and rotation groups. The averaging removes information, but a complete, unique description can be retained by taking tensor products of the ket before computing the integral. Different representations, capturing varying amounts of inter-atomic correlations, can be obtained depending on the combination of tensor products and symmetrized averages. This formulation provides a unified picture of density-based representations for machine learning of atomic-scale properties, with several popular frameworks emerging by taking different limits, or using specific basis sets to represent the abstract invariant kets. In particular, using a basis of radial functions and spherical harmonics shows clearly the 1:1 mapping between the symmetrized kets and different flavors of the SOAP representation. Even alternative schemes that start from rotationally and translationally-invariant internal coordinates and proceed to ensure permutation invariance appear to contain comparable information.

We discussed how several modifications and optimizations can be introduced in terms of operators that couple and scale different channels of the representation, focusing in particular on the SOAP power spectrum representation. We have given two examples of such extensions, representing the behavior of different chemical species as low-dimensional vectors, and modulating the information content of the representations with a radial scaling function. These optimizations improve significantly the performance of SOAP representations, matching or surpassing the state of the art on two very different data sets – a chemically diverse set of quaternary solid compounds, and a collection of small organic molecules. The framework we use to simplify the description of atomic species can reduce dramatically the complexity and computational costs of machine-learning models for multi-component systems, and could also be applied to coarse-grained models, in which beads correspond to functional groups, and a reduced-dimensionality description could identify features such as polarity or hydrophobicity. Moreover by re-discovering a data-driven version of the periodic table of the elements, and extending it to one and four dimensions, it also makes it possible to extract useful insights from the inspection of the optimal combinations of features.

We have demonstrated how a set of unsupervised learning techniques such as hierarchical clustering and sketch-map coupled with the SOAP power spectrum and the REMatch kernel can be used to navigate databases of molecules and molecular materials. Rather than simply reflecting preconceived notions of what would be the key structural parameters to differentiate conformers and polymorphs, automatic clustering identifies motifs that can be easily related to heuristic structural classifications, while capturing finer details and being fully data-driven. In the particular case of oligopeptide structures in the gas phase, such analysis reveals the importance of peptide bond isomerization in describing the high-energy portion of conformational space of oligopeptides, the possibility of changes in chemical connectivity in the course of the *ab initio* structural search, and the interplay between hydrogen-bonding, backbone dihedrals, and electrostatic interactions. Moreover, a similar study applied to pentacene molecular crystals and nitrogen substituted pentacenes 5A and 5B confirmed that a regular substitution leads to regular H-bond patterns within the molecular planes, while an asymmetric substitution leads to less robust H-bonding patterns and a generally glassy potential energy landscape. At the same time, comparing energy predictions and structural classification showed clearly that H-bonding alone is not sufficient to characterize the lattice energies of 5A and 5B, but inter-sheet arrangements also need to be properly accounted for. Lastly, we also show the importance of automated analysis techniques in assessing the integrity and the internal consistency of a database, by successfully identifying a subset of structures associated with ill-converged energetics. By simplifying the analysis and the interpretation of computational datasets containing thousands or millions of hypothetical compounds, these methods will be crucial to unleash the full potential of computational materials design.

We have applied the Gaussian process regression technique using the SOAP power spectrum representation to model several important ground state properties of molecular materials. Here we have shown that sub-kJ/mol accuracy can be obtained when predicting reference energies for the stability of different polymorphs of molecular crystals. Moreover, we have built accurate models for the prediction of chemical shifts and electron/hole mobility in molecular crystals, demonstrating how supervised learning can reduce the cost associated with crystal structure prediction and determination. Indeed, the chemical shielding models are accurate enough to be used to determine structures by comparison to experimental shifts in chemical shift based NMR crystallography approaches to structure determination, as shown here for cocaine and AZD8329. The ML model only scales linearly with the number of atoms and, for the prediction of individual structures, is dominated by a constant I/O overhead. Here it allows the calculation of chemical shifts for a set of six structures with between 768 and 1584 atoms in their unit cells in less than six minutes (an acceleration of factor 106 for the largest structure). In order to build a model of the charge mobility that minimizes the investment of CPU time needed to achieve a quantitative prediction for the large numbers of crystal structures found on CSP landscapes, we focused on the bottleneck of the calculation, which is the evaluation of electronic transfer integrals between pairs of adjacent molecules. Because of their origin in the electronic structure of interacting molecules, there is no simple form for the relationship between the intermolecular arrangement and these transfer integrals. Even though transfer integrals vary over several orders of magnitude, we showed that our ML scheme could predict their value at a level of accuracy comparable to that of the electronic structure reference using only 10% of the dimer configurations – corresponding to a potential 90% reduction of the computational effort associated with the screening of crystal structures for their charge mobility.

Finally, we have presented a scheme to obtain an inexpensive and reliable estimate of the uncertainty associated with the predictions of a machine-learning model of atomic and molecular properties. The scheme is based on sub-sampling and sparse Gaussian Process Regression. We have investigated the reliability of this approach for two applications: the prediction of ^{1}H NMR chemical shieldings in organic crystals and the prediction of formation energies of small organic molecules and inorganic crystals. In every case, we found the sub-sampling estimator to be reliable based on log-likelihood results and the good agreement between the true and predicted distribution of errors on a test set. Besides the computational savings, the fact that the sub-sampling models generate an ensemble of predictions makes it trivial to predict uncertainties in derived properties, that are obtained by a linear or non-linear combination of multiple predictions such as thermal averages.

To conclude, ML techniques have demonstrated their utility in the context of atomistic simulations over the last decade, by automating the post-processing of large amounts of data, e.g. molecular dynamics trajectories, and by improving the efficiency and/or the accuracy of the prediction of atomic-scale properties. Over the course of this thesis, we have developed a toolbox, based on a general framework for representing atomic structures and ML algorithms, to accelerate the analysis of structure-property relations and the reliable property prediction in atomistic systems. Far from exhausting all possible aspects of the this research field, we believe this thesis exposes a few questions regarding the incorporation of physical priors into the representation and/or ML model and the need for high body order representations along with their efficient implementation. Moreover, an accurate and inexpensive uncertainty estimation might help to streamline more atomistic ML by developing active-learning strategies to efficiently generate training datasets. More generally, several aspects of atomistic modelling such as enhanced sampling and coarse-graining have yet to benefit fully from the integration of ML techniques despite clear potential benefits. In light of this thesis, data-driven approaches applied to atomistic simulation are about to pass the stage of proof concepts to enrich well established frameworks.

Appendices

A Details on Spherical Invariants

This appendix gathers the various observations and notes that improved my understanding and familiarity with the SOAP power spectrum and more generally the n-body spherical invariant representations. I will be using the notation introduced in Chapter 2 for this purpose. A particular emphasis is put on their explicit derivations and features associated with gradients evaluation with explicit references to the *NIST Digital Library of Mathematical Functions* [355] or other online libraries are provided as footnote. I start discussing some important properties of the angular basis, or spherical harmonics, in Section A.1. Then I show in Section A.2 how this basis is convenient to derive explicit expressions for invariant features based on the density expansion coefficients. Formulas for the evaluation of atom density expansion coefficients and their derivatives w.r.t. atomic positions are derived in Section A.3 and Section A.4 respectively.

A.1 Angular basis

In this section I summarize a few results on the complete and orthonormal spherical ket basis $|lm\rangle$ for which full discussions can be found in Ref.[356] and Ref.[357].

A.1.1 Spherical harmonics

In real space the angular basis is composed of spherical harmonics (SPHs) and we use the following convention

$$\langle lm|\hat{\mathbf{r}}\rangle = Y_l^m(\hat{\mathbf{r}}) = Y_l^m\left(\theta,\phi\right) = A_l^m e^{im\phi} P_l^m(\cos\theta)\,, \tag{A.1}$$

where $A_l^m = \sqrt{(l-m)!(2l+1)/4\pi(l+m)!}$, the associated Legendre Polynomials (ALPs) are given by

$$P_l^m(x) = (-1)^m (1-x^2)^{m/2} \frac{\mathrm{d}^m}{\mathrm{d}x^m} P_l(x) \tag{A.2}$$

and $\hat{\mathbf{r}}$ is the direction vector defined by the spherical coordinates θ and ϕ orientated from $\hat{\boldsymbol{e}}_z$.

The complex SPHs of Eq. (A.1) are useful for analytical derivations but the real SPHs defined as

$$\left.\begin{aligned}\bar{Y}_{lm}(\hat{\mathbf{r}}_{ij}) &= \cos\big(m\phi\big)\bar{P}_l^m(\cos\theta)\\ \bar{Y}_{l,-m}(\hat{\mathbf{r}}_{ij}) &= \sin\big(m\phi\big)\bar{P}_l^m(\cos\theta)\end{aligned}\right\}\text{ for } m>0 \qquad \text{(A.3a)}$$
$$\bar{Y}_{l,0}(\hat{\mathbf{r}}_{ij}) = \frac{1}{\sqrt{2}}\bar{P}_l^0(\cos\theta)$$

where

$$\bar{P}_l^m(\cos\theta) = \sqrt{\frac{2l+1}{2\pi}\frac{(l-m)!}{(l+m)!}}P_l^m(\cos\theta), \qquad \text{(A.3b)}$$

are computationally more efficient.[358] Indeed the effect of conjugation

$$\langle\hat{\mathbf{r}}|l-m\rangle = (-1)^m\langle lm|\hat{\mathbf{r}}\rangle \qquad \text{(A.4)}$$

allow to retrieve the terms missing from Eq. (A.3) trivially.

A.1.2 Finite rotations

The rotation matrix $\hat{R} := \hat{R}(\alpha,\beta,\gamma)$ where α,β,γ are the Euler angles as defined in the $z-y-z$ convention is an element of the $\mathcal{SO}(3)$ group. The irreducible representation of this rotation group in the angular basis is given by the Wigner-D matrix elements

$$\langle lm|\hat{R}|\lambda\mu\rangle = \delta_{l\lambda}D^l_{m\mu}\big(\hat{R}\big), \qquad \text{(A.5)}$$

which follow the orthogonality relation

$$\int_0^{2\pi}\mathrm{d}\alpha\int_0^{\pi}\sin\beta\mathrm{d}\beta\int_0^{2\pi}\mathrm{d}\gamma D^l_{m\mu}\big(\hat{R}\big)^* D^{l'}_{m'\mu'}\big(\hat{R}\big) = \frac{8\pi^2}{2l+1}\delta_{ll'}\delta_{mm'}\delta_{\mu\mu'}, \qquad \text{(A.6)}$$

where their complex conjugate are given by

$$D^l_{m\mu}\big(\hat{R}\big) = (-1)^{m-\mu}D^l_{-m,-\mu}\big(\hat{R}\big)^*. \qquad \text{(A.7)}$$

Products of Wigner-D matrix elements can be reduced using

$$D^l_{m\mu}\big(\hat{R}\big)D^{l'}_{m'\mu'}\big(\hat{R}\big) = \sum_{L=|l-l'|}^{l+l'}\langle lml'm'|L(m+m')\rangle\langle l\mu l'\mu'|L(\mu+\mu')\rangle D^L_{(m+m'),(\mu+\mu')}\big(\hat{R}\big), \qquad \text{(A.8)}$$

where $\langle l_1m_1l_2m_2|l_3m_3\rangle$ is a Clebsch-Gordan (CG) coefficient. The relation between Wigner-D matrices and SPHs is given by

$$\mathrm{D}^l_{m0}\big(\alpha,\beta,\gamma\big) = \sqrt{\frac{4\pi}{2l+1}}Y_l^m\big(\beta,\alpha\big)^*, \qquad \text{(A.9)}$$

and the rotation of a spherical harmonic is expressed in terms of Wigner-D matrices

$$Y_l^m\left(\hat{\mathrm{R}}\hat{\mathbf{r}}\right) = \sum_{m'=-l}^{l} \mathrm{D}^l_{mm'}\left(\hat{R}\right)^* Y_l^{m'}\left(\hat{\mathbf{r}}\right). \tag{A.10}$$

A.2 Spherical invariants of order $\nu = 1,2,3$

For an atom density expressed on the angular basis and a radial basis $\langle nlm|$, the 2-body invariant representation ($\nu = 1$) is

$$\begin{aligned}
\langle nlm|\overline{\rho_j^{\otimes 1}}\rangle &= \int_{\mathcal{SO}(3)} \mathrm{d}\hat{R}\,\langle nlm|\hat{R}\rho_j\rangle\,,\\
&= \sum_{\mu\mu'} \langle nl\mu|\rho_j\rangle\,\langle n'l'\mu'|\rho_j\rangle \int_{\mathcal{SO}(3)} \mathrm{d}\hat{R}\,\langle lm|\hat{R}|l\mu\rangle\,,\\
&= \sum_{\mu\mu'} \langle nl\mu|\rho_j\rangle\,\langle n'l'\mu'|\rho_j\rangle \int_{\mathcal{SO}(3)} \mathrm{d}\hat{R} D^l_{m\mu}\left(\hat{R}\right) D^0_{00}\left(\hat{R}\right),\\
&= \frac{8\pi}{2l+1}\langle nlm|\rho_j\rangle\,\delta_{l0}\delta_{m0},\\
\Rightarrow \langle n|\overline{\rho_j^{\otimes 1}}\rangle &= 8\pi\,\langle n00|\rho_j\rangle\,,
\end{aligned} \tag{A.11}$$

where we have used Eqs. (A.5) to (A.7) and $D^0_{00}\left(\hat{R}\right) = 1\ \forall \hat{R} \in \mathcal{SO}(3)$. The 3-body invariant representation ($\nu = 2$) is given by

$$\begin{aligned}
\langle n_1 l_1 m_1; n_2 l_2 m_2|\overline{\rho_j^{\otimes 2}}\rangle &= \int_{\mathcal{SO}(3)} \mathrm{d}\hat{R}\,\langle n_1 l_1 m_1|\hat{R}\rho_j\rangle\,\langle n_2 l_2 m_2|\hat{R}\rho_j\rangle\,,\\
&= \sum_{m_1' m_2'} \langle n_1 l_1 m_1'|\rho_j\rangle\,\langle n_2 l_2 m_2'|\rho_j\rangle\\
&\qquad \int_{\mathcal{SO}(3)} \mathrm{d}\hat{R}\,\langle l_1 m_1|\hat{R}|l_1 m_1'\rangle\,\langle l_2 m_2|\hat{R}|l_2 m_2'\rangle\,,\\
&= \sum_{m_1' m_2'} \langle n_1 l_1 m_1'|\rho_j\rangle\,\langle n_2 l_2 m_2'|\rho_j\rangle\\
&\qquad \int_{\mathcal{SO}(3)} \mathrm{d}\hat{R} D^{l_1}_{m_1 m_1'}\left(\hat{R}\right) D^{l_2}_{m_2 m_2'}\left(\hat{R}\right),\\
&= \frac{8\pi}{2l+1}(-1)^{-m_1}\delta_{ll'}\delta_{m_1,-m_2}\\
&\qquad \sum_m (-1)^m \langle n_1 l_1 m|\rho_j\rangle\,\langle n_2 l_2 -m|\rho_j\rangle\,,
\end{aligned} \tag{A.12}$$

where we have used the Eqs. (A.5) to (A.7). Inspecting Eq. (A.12) shows that the angular basis $\langle l_1 m_1; l_2 m_2|$ has a large null space since the orders m_i do not affect the representation. More generally the $\nu+1$-order invariant representation can be simplified by coupling the angular basis $\langle l_1 m_1; \ldots; l_\nu m_\nu|$ (see Ref.[357] for a complete discussion of the procedure). The coupled

3-body invariant representation ($\nu = 2$) is then given by

$$
\begin{aligned}
\langle n_1 l_1; n_1 l_2; LM | \overline{\rho_j^{\otimes 2}} \rangle = &\sum_{m_1 m_2} \langle n_1 l_1; n_1 l_2; LM | n_1 l_1 m_1; n_2 l_2 m_2 \rangle \\
&\qquad\qquad \langle n_1 l_1 m_1; n_2 l_2 m_2 | \overline{\rho_j^{\otimes 2}} \rangle , \\
=&\delta_{L0}\delta_{M0}(-1)^{l_1 - m_1}\sqrt{2l_1+1}\, \langle n_1 l_1 m_1; n_2 l_2 m_2 | \overline{\rho_j^{\otimes 2}} \rangle ,
\end{aligned}
$$

where we have used the orthonormality of the basis $|l_1 m_1; l_2 m_2\rangle$ along with the contraction of CG coefficients

$$
\sum_m (-1)^{j-m} \langle j m j - m | J0 \rangle = \sqrt{2J+1}\sqrt{2j+1}\delta_{J0}. \tag{A.13}
$$

Equation (A.12) simplifies then into

$$
\langle n_1 n_2 l | \overline{\rho_j^{\otimes 2}} \rangle = \frac{8\pi^2 (-1)^l}{\sqrt{2l+1}} \sum_m (-1)^m \langle n_1 l m | \rho_j \rangle \langle n_2 l (-m) | \rho_j \rangle , \tag{A.14}
$$

which corresponds to the SOAP powerspectrum up to factors that do not change the SOAP kernel.[53] Following the same procedure as for Eq. (A.12) with Eq. (A.8) the uncoupled 4-body spherical invariant is

$$
\begin{aligned}
\langle n_1 l_1 m_1; n_2 l_2 m_2; a_3 n_3 l_3 m_3 | \overline{\rho_j^{\otimes 3}} \rangle = &\frac{8\pi^2}{2l_1+1} (-1)^{-m_1} \langle l_2 m_2 l_3 m_3 | l_1 - m_1 \rangle \\
\sum_{\nu_1 \nu_2 \nu_3} (-1)^{-\nu_1} \langle l_2 \nu_2; l_3 \nu_3 | l_1 (-\nu_1) \rangle & \langle n_1 l_1 \nu_1 | \rho_j \rangle \langle n_2 l_2 \nu_2 | \rho_j \rangle \langle n_3 l_3 \nu_3 | \rho_j \rangle ,
\end{aligned} \tag{A.15}
$$

which simplifies into

$$
\begin{aligned}
\langle n_1 l_1; n_2 l_2; n_3 l_3 | \overline{\rho_j^{\otimes 3}} \rangle = \frac{8\pi^2 (-1)^{l_1}}{\sqrt{2l_1+1}} &\sum_{m_1 m_2 m_3} (-1)^{m_1} \langle l_2 m_2; l_3 m_3 | l_1 (-m_1) \rangle \langle n_1 l_1 m_1 | \rho_j \rangle \\
&\langle n_2 l_2 m_2 | \rho_j \rangle \langle n_3 l_3 m_3 | \rho_j \rangle
\end{aligned} \tag{A.16}
$$

by reducing the coupling

$$
\langle l_1 l_2 l_3 l_{23}; LM | \overline{\rho_j^{\otimes 3}} \rangle = \sum_{\substack{m_1 m_2 \\ m_3 m_{23}}} \langle l_1 l_{23}; LM | l_1 m_1 \rangle \langle l_2 l_3; l_{23} m_{23} | l_2 m_2 l_3 m_3 \rangle \langle l_1 m_1 l_2 m_2 l_3 m_3 | \overline{\rho_j^{\otimes 3}} \rangle , \tag{A.17}
$$

using the exclusion rules of the Clebsch-Gordan coefficients, the orthonormality of the angular basis and Eq. (A.13). Equation (A.16) is the SOAP bispectrum up to factors that do not change the SOAP kernel.[53]

A.3 Atom density expansions

Section A.2 shows how invariant features can be computed from contractions of the atom density expanded on a basis. In this section we derive expressions for the coefficients of the density expansion and their derivative with respect to the atomic coordinates using several basis sets. In real space the atom density is given by

$$\langle a\mathbf{r}|\rho_i\rangle = \sum_{j\in i}\delta_{aa_j}\exp\left[-c\left(\mathbf{r}-\mathbf{r}_{ij}\right)^2\right] f_c(r_{ij}), \tag{A.18}$$

where $c = 1/2\sigma^2$, σ is the width of the Gaussian, $f_c(r_{ij})$ is a cutoff function with cutoff radii r_{cut} and a_j is the atomic species of atom j, a neighbor of atom i. Then the density coefficients become

$$\langle anlm|\rho_i\rangle = \int_{\mathbb{R}^3}\mathrm{d}\mathbf{r}\,\langle n|r\rangle\,\langle lm|\hat{\mathbf{r}}\rangle\,\langle a\mathbf{r}|\rho_i\rangle = \sum_{j\in i}\delta_{aa_j} f_c(r_{ij}) C^{ij}_{nlm}, \tag{A.19}$$

where $\langle n|r\rangle = R_n(r)$ is a radial basis and $\langle lm|\hat{\mathbf{r}}\rangle = Y_l^m(\hat{\mathbf{r}})$ is a spherical harmonic.

A.3.1 Angular integration

We use the orthonormality of the basis set to compute the expression for the density coefficients and express the resulting integral over $\mathbb{R}^3$ in spherical coordinates using the law of cosines $\|\mathbf{r}-\mathbf{r}_{ij}\|^2 = \|\mathbf{r}\|^2 + \|\mathbf{r}_{ij}\|^2 - 2\|\mathbf{r}\|\,\|\mathbf{r}_{ij}\|\cos\theta$:

$$\begin{aligned}\langle anlm|\rho_i\rangle = &\sum_{j\in i}\delta_{aa_j} f_c(r_{ij})\int_0^\infty \mathrm{d}r\; r^2\exp\left[-c\left(r^2+r_{ij}^2\right)\right] R_n(r)\int_{-1}^{1}\mathrm{d}\left(\cos\theta\right)\\ &\int_0^{2\pi}\mathrm{d}\phi\exp\left[2crr_{ij}\cos\theta\right] Y_l^m\left(\hat{\mathrm{R}}\hat{\boldsymbol{e}}_z\right),\end{aligned} \tag{A.20}$$

where $\hat{\mathrm{R}} = \hat{\mathrm{R}}_{ZYZ}\left(\alpha_{ij},\beta_{ij},0\right)$ is the ZYZ-Euler matrix that rotate $\hat{\boldsymbol{e}}_z$ onto $\hat{\boldsymbol{r}}_{ij}$ to match θ with the angle of the integral. Note that global normalization constants are omitted because of a normalization at the end. The integration over the angular part yields

$$\begin{aligned}\int_{-1}^{1}\mathrm{d}\left(\cos\theta\right)\exp\left[crr_{ij}\cos\theta\right]\int_0^{2\pi}\mathrm{d}\phi Y_l^m\left(\hat{\mathrm{R}}\left(\alpha_{ij},\beta_{ij},0\right)\hat{\boldsymbol{r}}\right) =& 4\pi Y_l^m\left(\beta_{ij},\alpha_{ij}\right)\mathrm{i}_l\left(arr_{ij}\right),\\ =& Y_l^m\left(\hat{\mathbf{r}}_{ij}\right)\mathrm{i}_l\left(arr_{ij}\right),\end{aligned} \tag{A.21}$$

where the intermediate steps are detailed in the following paragraphs.

Integration over ϕ The integration over the polar angle cancels out all orders of m from the SPH

$$\int_0^{2\pi}\mathrm{d}\phi Y_l^m\left(\theta,\phi\right) = \sqrt{\pi\left(2l+1\right)}\mathrm{P}_l^m\left(\cos\theta\right)\delta_{m0}, \tag{A.22}$$

since

$$\int_0^{2\pi} \mathrm{d}\phi \exp\left[im\phi\right] = 2\pi\delta_{0m}. \tag{A.23}$$

Using Eqs. (A.9) and (A.10), the polar integral over the rotated SPH simplifies into

$$\begin{aligned}\int_0^{2\pi} \mathrm{d}\phi\, Y_l^m\left(\hat{\mathrm{R}}\hat{\mathbf{r}}\right) = &\sum_{m'=-l}^{l} \mathrm{D}^l_{mm'}\left(\alpha_{ij},\beta_{ij},0\right)\sqrt{\pi(2l+1)}P_l^{m'}\left(\cos\theta\right)\delta_{m'0}\\ =&2\pi Y_l^m\left(\beta_{ij},\alpha_{ij}\right)P_l^0\left(\cos\theta\right).\end{aligned} \tag{A.24}$$

Integration over θ The modified spherical Bessel function of the first kind (MSBF) admit the following integral representation

$$\mathsf{i}_n\left(z\right) = \frac{1}{2}\int_{-1}^{1} \mathrm{d}x \exp\left(zx\right)\mathrm{P}_n^0\left(x\right), \tag{A.25}$$

which can be shown using the reference relations Eqs. (A.26) to (A.28)

$$\mathsf{j}_n\left(z\right) = \frac{(-i)^n}{2}\int_{-1}^{1} \mathrm{d}x \exp\left[izx\right]P_n^0\left(x\right), ^{\dagger} \tag{A.26}$$

$$\mathsf{i}_n\left(z\right) = (-i)^n \mathsf{j}_n\left(iz\right), ^{\ddagger} \tag{A.27}$$

$$\mathsf{i}_n\left(z\right) = (-1)^n \mathsf{i}_n\left(-z\right), ^{\S} \tag{A.28}$$

j_n is the spherical Bessel function of the first kind. The integral over the polar angle is then given by

$$\int_{-1}^{1} \mathrm{d}\left(\cos\theta\right)\exp\left[2crr_{ij}\cos\theta\right]P_l^0\left(\cos\theta\right) = 2\mathsf{i}_l(2crr_{ij}). \tag{A.29}$$

A.3.2 Radial integration

Summing up the results from the previous section

$$C_{nlm}^{ij} = 4\pi Y_l^m(\hat{\mathbf{r}}_{ij})\exp\left[-cr_{ij}^2\right]\underbrace{\int_0^{\infty} \mathrm{d}r\, r^2 R_n(r)e^{-cr^2}\mathsf{i}_l\left(2crr_{ij}\right)}_{=\mathrm{I}_{nl}^{ij}}, \tag{A.30}$$

[†] http://dlmf.nist.gov/10.54.E2

[‡] http://dlmf.nist.gov/10.47.E12

[§] http://dlmf.nist.gov/10.47.E16

we identify the radial integral I_{nl}^{ij} for which an explicit choice of a radial basis function has to be done. Inspecting Eq. (A.30) shows that a swap of the ij indices in the pair coefficient only affects the spherical harmonics and it corresponds to reflecting the orientation of the $\hat{\boldsymbol{r}}_{ij}$. The parity operator applied on spherical harmonics[357] yields the following relation for the pair coefficients:

$$C_{nlm}^{ija_j} = (-1)^l C_{nlm}^{jia_i}, \tag{A.31}$$

where a_j is the type of atom j and a_i is the type of atom i.

In the following we provide explicit results for I_{nl}^{ij} using several radial basis.

Gaussian Type Orbital (GTO) like radial basis

The GTO radial basis is defined as

$$R_n^{GTO}(r) = \mathcal{N}_n\, r^n \exp\left[-b_n r^2\right], \tag{A.32}$$

where $b_n = 1/2\sigma_n^2$, $\sigma_n = r_{\mathrm{cut}} \max(\sqrt{n}, 1)/n_{\max}$ and the normalization factor is given by

$$\mathcal{N}_n^2 = \frac{2}{\sigma_n^{2n+3}\Gamma(n+3/2)}, \tag{A.33}$$

where $\Gamma(x)$ is the Gamma function. The GTO radial basis is interesting because it allows for an analytical integration of the radial integral (see the next paragraph for the detailed derivation steps)

$$I_{nl}^{ij\,\mathrm{GTO}} = \mathcal{N}_n \frac{\sqrt{\pi}}{4} \frac{\Gamma\left((n+l+3)/2\right)}{\Gamma\left(l+3/2\right)} c^l r_{ij}^l (c+b_n)^{-(n+l+3)/2} {}_1F_1\left(\frac{n+l+3}{2}, l+\frac{3}{2}, \frac{c^2 r_{ij}^2}{c+b_n}\right), \tag{A.34}$$

where ${}_1F_1$ is the confluent hypergeometric function of the first kind and the neighbor contribution becomes

$$C_{nlm}^{ij\,\mathrm{GTO}} = (\pi)^{3/2} \mathcal{N}_n \frac{\Gamma\left((n+l+3)/2\right)}{\Gamma\left(l+3/2\right)} (c+b_n)^{-(n+l+3)/2} \tag{A.35}$$

$$Y_l^m(\hat{\mathbf{r}}_{ij}) \exp\left[-c r_{ij}^2\right] (c r_{ij})^l {}_1F_1\left(\frac{n+l+3}{2}, l+\frac{3}{2}, \frac{c^2 r_{ij}^2}{c+b_n}\right). \tag{A.36}$$

The GTO radial basis is not orthonormal so the expansion coefficients are given by

$$\langle anlm|\rho_i;\mathrm{GTO}\rangle = \sum_{n'} S_{nn'}^{-1/2} \sum_{j\in i} \delta_{aa_j} f_c(r_{ij}) C_{n'lm}^{ij\,\mathrm{GTO}} \tag{A.37}$$

where the overlap matrix is

$$S_{nn'} = \int_0^\infty R_n^{GTO}(r) R_{n'}^{GTO}(r) r^2 \,\mathrm{d}r = \frac{1}{2}\mathcal{N}_n \mathcal{N}_{n'} \left(b_n + b_{n'}\right)^{-0.5(3+n+n')} \Gamma(\frac{3+n+n'}{2}).$$

Analytic radial integral We write an integral representation of the confluent hypergeometric function ${}_1F_1\left(a,b,z\right)$ (CHF) in terms of MSBF:

$${}_1F_1\left(a, l+\frac{3}{2}, x\right) = \frac{2x^{-l/2}}{\sqrt{\pi}} \frac{\Gamma\left(l+3/2\right)}{\Gamma\left(a\right)} \int_0^\infty e^{-t} t^{a-1-l/2} \mathrm{i}_l(2\sqrt{xt}) \,\mathrm{d}t, \tag{A.38}$$

using these relations

$$\begin{aligned}
{}_1F_1\left(a,b,z\right) =& \frac{1}{\Gamma\left(a\right)} \int_0^\infty e^{-t} t^{a-1} {}_0F_1\left(b, zt\right) \mathrm{d}t,^{\P\P} && \text{(A.39)}\\
I_l(z) =& \frac{(z/2)^l}{\Gamma\left(l+1\right)} {}_0F_1\left(l+1, \frac{z^2}{4}\right),^{\dagger\dagger\dagger} && \text{(A.40)}\\
\mathrm{i}_l(z) =& \sqrt{\frac{\pi}{2z}} I_{l+1/2}(z),^{\ddagger\ddagger\ddagger} && \text{(A.41)}\\
\mathrm{i}_l(z) =& \sqrt{\frac{\pi}{4}} \frac{(z/2)^l}{\Gamma\left(l+3/2\right)} {}_0F_1\left(l+\frac{3}{2}, \frac{z^2}{4}\right), && \text{(A.42)}\\
{}_0F_1\left(l+\frac{3}{2}, xt\right) =& \sqrt{\frac{4}{\pi}} \Gamma\left(l+\frac{3}{2}\right) x^{-l/2} t^{-l/2} \mathrm{i}_l(2\sqrt{xt}), && \text{(A.43)}
\end{aligned}$$

where I_l is the modified Bessel function and ${}_0F_1\left(b,z\right)$ is the limit conflent hypergeometric function.

The radial integral with GTO radial basis function is

$$I_{nl}^{ij\,\mathrm{GTO}} = \int_0^\infty \mathrm{d}r\, r^2 R_n^{\mathrm{GTO}}(r) e^{-r^2/2\sigma^2} \mathrm{i}_l\left(r r_{ij}/\sigma^2\right) = \mathcal{N}_n \int_0^\infty \mathrm{d}r r^{2+n} e^{-r^2(c+b_n)} \mathrm{i}_l\left(2c r r_{ij}\right). \tag{A.44}$$

We partially identify the terms between Eq. (A.38) and Eq. (A.44):

$$\begin{aligned}
t =& r^2(c+b_n), && \text{(A.45)}\\
\mathrm{d}t =& 2r\,\mathrm{d}r\,(c+b_n), && \text{(A.46)}\\
x =& \frac{c^2 r_{ij}^2}{c+b_n}, && \text{(A.47)}
\end{aligned}$$

[¶¶] http://functions.wolfram.com/HypergeometricFunctions/Hypergeometric1F1/07/01/01/0002/ or http://dlmf.nist.gov/16.5.E3

[†††] https://en.wikipedia.org/wiki/Generalized_hypergeometric_function#The_series_0F1

[‡‡‡] http://mathworld.wolfram.com/ModifiedSphericalBesselFunctionoftheFirstKind.html

to change the integrand of the radial integral

$$I_{nl}^{ij\,\mathrm{GTO}} = \mathcal{N}_n \int_0^\infty \frac{\mathrm{d}t}{2(c+b_n)} (c+b_n)^{-(n+1)/2} t^{(n+1)/2} e^{-t} \mathrm{i}_l\left(2\sqrt{xt}\right), \tag{A.48}$$

and identify the last term to $(n+l+3)/2$

Descrete Variable Representation (DVR) radial basis

Alternatively, the radial integral can be solved numerically

$$\mathrm{I}_{nl}^{ij} = \sum_{k=1}^{K} \omega_k r_k^2 R_n(r_k) e^{-cr_k^2} \mathrm{i}_l\left(2cr_k r_{ij}\right), \tag{A.49}$$

where the ω_k are the quadrature weights evaluated at the quadrature nodes r_k. Depending on the quadrature rule, the following shifting formula is useful,

$$\int_a^b f(x)\,\mathrm{d}x \approx \frac{b-a}{2} \sum_{i=1}^{n} w_i f\left(\frac{b-a}{2} x_i + \frac{a+b}{2}\right).$$

The cost associated with the K function evaluations can be mitigated by choosing the DVR radial basis and the Gauss-Legendre quadrature rule (see Ref.[359] for more details)

$$\mathrm{I}_{nl}^{ij\,\mathrm{DVR}} = x_n \sqrt{\omega_n} e^{-cx_n^2} \mathrm{i}_l\left(2cx_n r_{ij}\right), \tag{A.50}$$

where x_n and ω_n have been shifted to the proper range. Note we have x_n instead of x_n^2 because when starting from the kernel for the power spectrum in real space and approximating the radial integrals with Gauss quadrature, the r^2 gets split into the two spherical expansions of the power spectrum.

A.4 Gradients of the atom density expansions

From Section A.3, it is clear that irrespective of the radial basis the atom density expansion can be written as

$$\langle anlm|\rho_i\rangle = \sum_{j\in i} \delta_{aa_j} Y_l^m(\hat{\mathbf{r}}_{ij}) D_{nl}(r_{ij}), \tag{A.51}$$

where j is a neighbor of atom i of species a_j and $D_{nl}(r_{ij}) = f_c(r_{ij})\exp\left[-cr_{ij}^2\right] I_{nl}^{ij}$. Hence the derivative w.r.t. the atomic positions instead

$$\begin{aligned}\nabla_k \langle anlm|\rho_i\rangle =& \sum_{j\in i} \delta_{aa_j} \nabla_k C^{ij}_{anlm} \\ =& \sum_{j\in i} \delta_{aa_j} \left[\nabla_k Y_l^m(\hat{\mathbf{r}}_{ij}) D_{nl}(r_{ij}) + Y_l^m(\hat{\mathbf{r}}_{ij}) \nabla_k D_{nl}(r_{ij})\right],\end{aligned} \tag{A.52}$$

where k can be i or j. From this expression we can derive a few properties of $\nabla_k \langle anlm|\rho_i\rangle$ that are helpfull to its efficient implementation:

$$\nabla_i \langle anlm|\rho_i\rangle = \sum_{j\in i} \delta_{aa_j} \nabla_i C^{ij}_{a_jnlm} \tag{A.53}$$

$$\nabla_j \langle anlm|\rho_i\rangle = \delta_{aa_j} \nabla_j C^{ij}_{a_jnlm}. \tag{A.54}$$

Inspection of Eq. (A.52) shows that

$$\nabla_i C^{ij}_{a_jnlm} = -\nabla_j C^{ij}_{a_jnlm}, \tag{A.55}$$

which allow to simply write Eq. (A.54) in terms of Eq. (A.53)

$$\nabla_j \langle anlm|\rho_i\rangle = -\delta_{aa_j} \nabla_i C^{ij}_{a_jnlm}. \tag{A.56}$$

In a similar fashion, the missing ji pairs when $i < j$ using the half neighbor list can be recovered. The terms $\nabla_i \langle anlm|\rho_j\rangle$ are not explicitly present but we can get them using:

$$\nabla_i \langle anlm|\rho_j\rangle = \nabla_i C^{ji}_{anlm} = (-1)^l \nabla_i C^{ij}_{a_jnlm} = (-1)^{l+1} \nabla_j C^{ij}_{a_jnlm} = (-1)^{l+1} \nabla_j \langle anlm|\rho_i\rangle. \tag{A.57}$$

A.4.1 $\nabla_k Y_l^m(\hat{\mathbf{r}}_{ij})$

The derivative of the SPH can be expressed in a few different ways. The real SPHs definition is used in the following derivation (see Eq. (A.3)). Derivative w.r.t. the z coordinate:

$$\begin{aligned}\frac{\partial \bar{Y}_l^m}{\partial z_i} =& \frac{-\sin\theta}{2r_{ij}} \cos(m\phi) \Big(\sqrt{(l+m)(l-m+1)}\bar{P}_l^{m-1}(\cos\theta) \\ & -\sqrt{(l-m)(l+m+1)}\bar{P}_l^{m+1}(\cos\theta)\Big)\end{aligned} \tag{A.58}$$

$$\begin{aligned}\frac{\partial \bar{Y}_l^{-m}}{\partial z_i} =& \frac{-\sin\theta}{2r_{ij}} \sin(m\phi) \Big(\sqrt{(l+m)(l-m+1)}\bar{P}_l^{m-1}(\cos\theta) \\ & -\sqrt{(l-m)(l+m+1)}\bar{P}_l^{m+1}(\cos\theta)\Big)\end{aligned} \tag{A.59}$$

$$\frac{\partial \bar{Y}_l^0}{\partial z_i} = \frac{\sin\theta}{r_{ij}} \sqrt{\frac{l(l+1)}{2}} \bar{P}_l^1(\cos\theta)) \tag{A.60}$$

The x component is:

$$\begin{aligned}\frac{\partial \bar{Y}_l^m}{\partial x_i} &= \frac{-m\sin\phi}{\sqrt{x_{ij}^2+y_{ij}^2}}\bar{Y}_l^{-m} + \frac{\cos\phi\cos\theta}{2r_{ij}}\cos\big(m\phi\big)\Big(\sqrt{(l+m)(l-m+1)}\bar{P}_l^{m-1}(\cos\theta) \\ &\qquad - \sqrt{(l-m)(l+m+1)}\bar{P}_l^{m+1}(\cos\theta)\Big)\end{aligned} \tag{A.61}$$

$$\begin{aligned}\frac{\partial \bar{Y}_l^{-m}}{\partial x_i} &= \frac{m\sin\phi}{\sqrt{x_{ij}^2+y_{ij}^2}}\bar{Y}_l^{m} + \frac{\cos\phi\cos\theta}{2r_{ij}}\sin\big(m\phi\big)\Big(\sqrt{(l+m)(l-m+1)}\bar{P}_l^{m-1}(\cos\theta) \\ &\qquad - \sqrt{(l-m)(l+m+1)}\bar{P}_l^{m+1}(\cos\theta)\Big)\end{aligned} \tag{A.62}$$

$$\frac{\partial \bar{Y}_l^0}{\partial x_i} = \frac{-\cos\phi\cos\theta}{r_{ij}}\sqrt{\frac{l(l+1)}{2}}\bar{P}_l^1(\cos\theta) \tag{A.63}$$

and for the y component, similarly:

$$\begin{aligned}\frac{\partial \bar{Y}_l^m}{\partial y_i} &= \frac{m\cos\phi}{\sqrt{x_{ij}^2+y_{ij}^2}}\bar{Y}_l^{-m} + \frac{\sin\phi\cos\theta}{2r_{ij}}\cos\big(m\phi\big)\Big(\sqrt{(l+m)(l-m+1)}\bar{P}_l^{m-1}(\cos\theta) \\ &\qquad - \sqrt{(l-m)(l+m+1)}\bar{P}_l^{m+1}(\cos\theta)\Big)\end{aligned} \tag{A.64}$$

$$\begin{aligned}\frac{\partial \bar{Y}_l^{-m}}{\partial y_i} &= \frac{-m\cos\phi}{\sqrt{x_{ij}^2+y_{ij}^2}}\bar{Y}_l^{m} + \frac{\sin\phi\cos\theta}{2r_{ij}}\sin\big(m\phi\big)\Big(\sqrt{(l+m)(l-m+1)}\bar{P}_l^{m-1}(\cos\theta) \\ &\qquad - \sqrt{(l-m)(l+m+1)}\bar{P}_l^{m+1}(\cos\theta)\Big)\end{aligned} \tag{A.65}$$

$$\frac{\partial \bar{Y}_l^0}{\partial y_i} = \frac{-\sin\phi\cos\theta}{r_{ij}}\sqrt{\frac{l(l+1)}{2}}\bar{P}_l^1(\cos\theta) \tag{A.66}$$

The formulæ above have a singularity at the poles for $m \neq 0$, so the following identity:

$$\begin{aligned}\frac{m}{\sqrt{x_{ij}^2+y_{ij}^2}}\begin{pmatrix}\bar{Y}_{l,-m}(\hat{r}_{ij}) \\ \bar{Y}_{l,m}(\hat{r}_{ij})\end{pmatrix} = \frac{-1}{2z_{ij}}\begin{pmatrix}\sin\big(m\phi\big) \\ \cos\big(m\phi\big)\end{pmatrix}&\Big(\sqrt{(l+m)(l-m+1)}\bar{P}_l^{m-1}(\cos\theta) \\ &+ \sqrt{(l-m)(l+m+1)}\bar{P}_l^{m+1}(\cos\theta)\Big)\end{aligned} \tag{A.67}$$

can be used to shift the singularity to the equator ($z = 0$).

A.4.2 $\nabla_k D_{nl}(r_{ij})$

$$\nabla_k D_{nl}(r_{ij}) = \left[\frac{\mathrm{d}f_c(r_{ij})}{\mathrm{d}r_{ij}}I_{nl}^{ij} - 2cr_{ij}f_c(r_{ij})I_{nl}^{ij} + f_c(r_{ij})\frac{\mathrm{d}I_{nl}^{ij}}{\mathrm{d}r_{ij}}\right]\exp\Big[-cr_{ij}^2\Big]\nabla_k r_{ij} \tag{A.68}$$

where $\nabla_{i,j} r_{ij} = \mp\mathbf{r}_{ij}/r_{ij}$ and $\mathbf{r}_{ij} = \mathbf{r}_j - \mathbf{r}_i$.

Derivative of the radial integral

GTO radial basis Using the recurrence relation of the ${}_1F_1$[†]:

$$\frac{\mathrm{d}}{\mathrm{d}z}{}_1F_1\left(a,b,z\right) = \frac{a}{b}{}_1F_1\left(a+1,b+1,z\right), \tag{A.69}$$

the gradient of the GTO radial integral becomes:

$$\frac{\mathrm{d}I_{nl}^{ij\,GTO}}{\mathrm{d}r_{ij}} = \frac{\mathcal{N}_n}{2} c^{l+2} r_{ij}^{l+1} (c+b_n)^{-n+l+5/2} \frac{\Gamma(n+l+5/2)}{\Gamma(l+5/2)} {}_1F_1\left(\frac{n+l+5}{2}, l+\frac{5}{2}, \frac{c^2 r_{ij}^2}{c+b_n}\right) + \frac{l}{r_{ij}} I_{nl}^{ij\,GTO} \tag{A.70}$$

DVR radial basis Using the recurrence relation of the MSBF[‡]:

$$\frac{\mathrm{d}\mathrm{i}_l(x)}{\mathrm{d}x} = \frac{1}{2l+1}[l\mathrm{i}_{l-1}(x) + (l+1)\mathrm{i}_{l+1}(x)], \tag{A.71}$$

the gradient of the DVR radial integral becomes:

$$\frac{\mathrm{d}I_{nl}^{ij\,DVR}}{\mathrm{d}r_{ij}} = \frac{2c\sqrt{\omega_n}}{2l+1}\frac{r_{\mathrm{cut}}}{2} x_n^3 e^{-cx_n^2}[l\mathrm{i}_{l-1}(2cx_n r_{ij}) + (l+1)\mathrm{i}_{l+1}(2cx_n r_{ij})], \tag{A.72}$$

where $x_n = r_{\mathrm{cut}}/2r_n + r_{\mathrm{cut}}/2$ and r_n are the Gauss-Legendre quadrature points.

[†] `http://dlmf.nist.gov/13.3.E15`

[‡] `http://mathworld.wolfram.com/ModifiedSphericalBesselFunctionoftheFirstKind.html`

List of Figures

2.1 Atom-density-based structural representations, expressed in the real-space $\langle \mathbf{r}|$ basis. (a) A structure can be mapped onto a smooth atom density built as a superposition of smooth atom-centered functions. The overall density can be decomposed in atom-centered environments, and information on chemical compositions can be stored by decorating the functions with elemental kets. (b) The $\nu = 1$ invariant ket corresponds to spherical averaging of the environmental atom density. (c) The $\nu = 2$ invariant ket corresponds to three-body correlations, which are obtained by integrating over all rotations a stencil corresponding to two distances along two directions with a fixed angle $\arccos\omega$ between them. . 15

2.2 Isocontours of the 3-body correlation functions associated with the environment centered on the tagged carbon atom of an ethanol molecule. From left to right, the figures correspond to $\langle \mathrm{C}r\mathrm{H}r'\omega|\overline{\rho_j^{\otimes 2}}\rangle\,/rr'$, $\langle \mathrm{O}r\mathrm{H}r'\omega|\overline{\rho_j^{\otimes 2}}\rangle\,/rr'$, $\langle \mathrm{O}r\mathrm{H}r'\omega|\overline{\rho_j^{\otimes 2}}\rangle\,/rr'$. 16

2.3 Schematic representation of the construction of a real-space representation of a tensorial ket associated with a λ-SOAP kernel. The (smooth) atom density is evaluated at two points corresponding to a stencil (r, r', ω), and the spherical harmonic Y_μ^λ is evaluated at the angles (θ, ϕ), relative to the reference frame that is used to describe the stencil. 19

2.4 (a) Permutation-variant structural descriptors can be stored in a vector to be used as an atomic-scale representation. (b) Sorting this vector makes it permutationally invariant. (c) It is easy to see how the sorted vector relates to the cumulative distribution function associated with the histogram of the values of the structural features. 20

2.5 (a) Schematic representation of an ANN in which the input layer corresponds to the elements of a density-based representation. The target property associated with the full structure is expressed as a sum of atomic contributions. (b) In the case of a tensorial property it is essential to preserve the covariant nature of the λ-SOAP ket. To do so, one can construct a NN using only scalar SOAP features, and use the output as a multiplier for the tensorial features. The output of several of these blocks must then be combined linearly and without mixing different $\lambda\mu$ components to obtain the environment's contribution to a tensorial property. 22

3.1 The lysine building block was studied in three forms: (a) uncharged dipeptide, (b) protonated dipeptide, and (c) uncapped and uncharged amino acid. 36

3.2 Representation of the similarity matrix corresponding to the lysine dipeptide dataset using the agglomerative clustering algorithm (top) and the sketchmap algorithm (bottom, projection parameters shown following the scheme σ-A_B-a_b). A few representative structures (see Eq. (3.7)) of interesting clusters are shown (right) and their corresponding position on the sketch-maps and dendrogram representation is highlighted. The five sketch-maps are colored according to the conformational energy and the backbone dihedral angles ϕ, ψ, ω_1 and ω_2. The dendrogram shows the clustering hierarchy of the structures of the dataset. Each structure is vertically aligned with its properties shown using color bars below the dendrogram. The dendrogram is cut at a linkage distance of 0.1 since structural properties are very similar below this threshold, and the clusters that are merged at this level are shown as thick gray bars separated by light-gray lines. Clusters composed of only one structure are drawn as a black line reaching the bottom of the dendrogram. The main structural motifs of this set of structures are governed by the peptide bond dihedral angles ω_1 and ω_2. The two main clusters (a) and (b) are showing a global correlation with the angle ω_2 while the angle ω_1 splits them into two well correlated sub-clusters (d), (e) and (f), (g) respectively. The cluster (c) is highlighted as an example containing 'outlier' structures of low conformational energy. 37

3.3 Representation of the similarity matrix corresponding to the protonated lysine dipeptide dataset using the agglomerative clustering algorithm (top) and the sketchmap algorithm (bottom, projection parameters shown following the scheme σ-A_B-a_b). A few representative structures (see Eq. (3.7)) of interesting clusters are shown (right) and their corresponding position on the sketch-maps and dendrogram representation is highlighted. The six sketch-maps are colored according to the conformational energy, the minimal distance between O_1 or O_2 with N_3 called D_{ON}, and the backbone dihedral angles ϕ, ψ, ω_1 and ω_2. The dendrogram shows the clustering hierarchy of the structures of the dataset. Each structure is vertically aligned with its properties shown using color bars below the dendrogram. The dendrogram is cut at a linkage distance of 0.1 since structural properties are very similar below this threshold, and the clusters that are merged at this level are shown as thick gray bars separated by light-gray lines. Clusters composed of only one structure are drawn as a black line reaching the bottom of the dendrogram. The main structural motifs of this set of structures are governed by the dihedral angles ω_1 and ω_2 and the distance D_{ON}. The two main clusters (a) and (b) are showing a global correlation with the angle ω_2 while the angle ω_1 splits them into well correlated sub-clusters (e.g. sub-clusters (d) and (e)). The other important sub-clustering parameter is the distance D_{ON}, e.g. sub-clusters (c) and (b), which also correlates well with the separation between low and high conformational energy shown on the sketch-maps. Two sub-clusters are particular: (g) is a clear ‘outlier’ due to a chemical change and (f) features a H-bonding pattern with the side chain NH_3^+ pointing to both carboxy groups that sets this cluster apart from all others. 40

3.4 Representation of the similarity matrix corresponding to the lysine uncapped dataset using the agglomerative clustering algorithm (top) and the sketchmap algorithm (bottom, projection parameters shown following the scheme σ-A_B-a_b). A few representative structures (see Eq. (3.7)) of interesting clusters are shown (right) and their corresponding position on the sketch-maps and dendrogram representation is highlighted. The five sketch-maps are colored according to the conformational energy, the distance between N_1 and the hydrogen in the carboxilic group H_1 (labelled D_H), the distance between N_2 and C_α (labelled D_{CN}), and the dihedral angles α_1 and α_2 which are respectively computed with the following atoms (N_1,C_α,C_2,C_3) and (C_1,C_α,C_2,C_3). The dendrogram shows the clustering hierarchy of the structures of the dataset. Each structure is vertically aligned with its properties shown using color bars below the dendrogram. The dendrogram is cut at a linkage distance of 0.1 since structural properties are very similar below this threshold, and the clusters that are merged at this level are shown as thick gray bars separated by light-gray lines. Clusters composed of only one structure are drawn as a black line reaching the bottom of the dendrogram. The main structural motifs of the database are governed by the distance D_H. The two main clusters (a) and (b) are agglomerated according to the orientation of H_1 and the oxygen atom it is bonded to with respect to N_1 which is well described by the distance D_H. The sub-cluster (e) is composed of 'outlier' structures showing an H-bond between N_2 and an hydrogen of N_1 resulting in a folded side chain structural motif. Finally, the outlier cluster (f) contains a H-bond between the carboxy H and the side-chain NH_2, that can be seen as a precursor to the zwitterionic form. 42

3.5 The out-of-sample embedding of conformers with Ca^{2+} ion on the sketchmap of their pure counterpart, for the three systems we discussed in above: lysine dipeptide (a), protonated lysine dipeptide (b) and molecular lysine (c) systems. The projected conformers are colored with their energy where as the sketchmap on which they are projected are kept all in grey color. The location of the projected conformers allows us to understand how the conformational space of the pure conformers are affected due to presence of the Ca^{2+} ion. 43

3.6 This figure compares the homogeneity of clusters from the protonated lysine dipeptide (see a) and the bare lysine uncapped (see b) with respect to properties of their elements. The homogeneity of a cluster is probed using the standard deviation with respect to the distance between each cluster elements, σ_D, and the conformational energy, σ_E. The outliers of uncapped lysine (b) were manually highlighted in orange. 45

3.7 Molecules investigated in the present study. 48

3.8 Sketch-map representations of the pentacene crystal structure landscape's similarity matrix (projection parameters shown follow the scheme σ_{map}-A_B-a_b). The atomic configurations are color-coded according to their relative lattice energy (bottom right), class following the heuristic classification (top right) and cluster index (gray structures do not belong to a cluster) found using HDBSCAN* on the similarity matrix (left). The structural pattern of each cluster is illustrated from a view down the short edge of pentacene. 49

3.9 Sketch-map representations of the 5A crystal structure landscape. The atomic configurations are color-coded according to their relative lattice energy (bottom right), class following the heuristic classification (top right) and cluster index (gray structure do not belong to a cluster) found using HDBSCAN* on the similarity matrix (left). The structural pattern of each cluster is illustrated with a top and long side (yellow cluster) view of the 5A polymorphs. 51

3.10 Representation of the similarity matrix for 5B The atomic configurations, i.e. disks, on the three sketch-maps are color-coded according to their lattice energy (bottom right), class following the heuristic classification (top right) and cluster index (gray structure do not belong to a cluster) found using HDBSCAN* on the similarity matrix (left). The structural pattern of each cluster is illustrated with a top view of the 5B polymorphs. 52

4.1 ^{1}H chemical shifts of the 76214 environments in the CSD-2k set. The environments excluded using the unusual structures detection procedure described in Eq. (4.8) are shown in red. 61

4.2 Scheme of the machine learning model used for the chemical shift predictions. 64

4.3 ^{1}H chemical shift prediction error of the trained model for the CSD-500 set. The RMSE prediction error between chemical shifts calculated with ShiftML and GIPAW DFT is shown for different local environment cutoff radii, and for the multi-kernel (labelled as msk), as a function of the training set size. 65

4.4 RMSE learning curves showing the error between chemical shifts calculated with GIPAW DFT and ShiftML of the CSD-2k dataset for different local environment cutoff radii, and for the multi-kernel (labelled as msk), as a function of the training set size. The curves are for ^{1}H (a), ^{13}C (b), ^{15}N (c) and ^{17}O (d) chemical shieldings. 66

4.5 CPU time for NMR chemical shift calculations using the GIPAW method. (a) The CPU time is shown as function of the DFT accuracy, determined by the plane-wave cutoff energy E_{cutoff} and the number of k-points in each dimension for polymorph 1 of cocaine. The charge density energy cut-offs were set to $E_\rho = 4E_{cutoff}$. (b) The CPU time is shown as function of increasing system size in CSD-2k. The green squares and blue dots show individual geometry optimization and GIPAW chemical shift DFT calculations, respectively. The red line shows the best fit between the number of valence electrons and the required CPU time as $t_{CPU} = aN_e^2 + bN_e^3$, with $a = 0.0162$ and $b = 5.91 \cdot 10^{-6}$. 67

4.6 Comparison of predictions from ShiftML and GIPAW DFT. Histograms and scatter-plots showing the correlation between ^{1}H (a), ^{13}C (b), ^{15}N (c) and ^{17}O (d) chemical shifts (shieldings) calculated with GIPAW and ShiftML. The black lines indicate a perfect correlation. 68

4.7 Chemical structures of the compounds used for experimental comparison taken from Ref.[305]. In order, cocaine (a),[265] 3,5-dimethylimidazole and 4,5-dimethylimidazole (b),[307] AZD8329 (c),[276] naproxen (d),[308] theophylline (e)[265] and uracil (f),[309] and the labelling scheme used here. 69

4.8 Comparison of predictions from ShiftML and GIPAW DFT for polymorphs of cocaine and AZD8329. (a) Histogram showing the distribution of the differences between ^{1}H chemical shifts calculated with GIPAW and with ShiftML for the polymorphs of cocaine (blue), and the polymorphs of AZD8329 (orange). (b) Scatterplot showing the correlation between ^{1}H chemical shifts calculated with GIPAW and ShiftML for cocaine (blue) and AZD8329 (orange). The black line indicates a perfect correlation. 70

4.9 Comparison of ShiftML to experimentally measured shifts. (a) Histogram showing the distribution of differences between experimentally measured ^{1}H chemical shifts and ^{1}H chemical shifts calculated with ShiftML for six different crystal structures (see Fig. 4.7 for the structures and numerical values of the shifts). (b) Scatter plot showing the correlation between these experimentally measured ^{1}H chemical shifts and shifts calculated with ShiftML. (c-d) Comparison between calculated and experimental ^{1}H chemical shifts for the most stable structures obtained with CSP for cocaine (c) and AZD8329 (d). For each candidate structure an aggregate RMSE is shown between experimentally measured shifts and shifts calculated using either GIPAW (blue) or ShiftML (red). The grey zones represent the confidence intervals of the GIPAW DFT ^{1}H chemical shift RMSD, as described in the text,[262] and candidates (in c and d) that have RMSEs within this range would be determined as correct crystal structures using a chemical shift driven solid-state NMR crystallography protocol. 71

4.10 Chemical shift calculation times and large structures. (a) DFT GIPAW calculation time (blue) and ShiftML prediction time (turquoise) for different system sizes. The GIPAW DFT calculation time for the six large structures (orange) is estimated from a cubic dependence on the number of valence electrons in the structure (see Fig. 4.5). (b-g) 3D-shemes and ^{1}H NMR spectra predicted with ShiftML, of the six large molecular crystals with CSD Refcodes: (b) CAJVUH,[310] $N_{atoms} = 828$, (c) RUKTOI,[311] $N_{atoms} = 768$, (d) EMEMUE,[312] $N_{atoms} = 860$, (e) GOKXOV,[313] $N_{atoms} = 945$, (f) HEJBUW,[314] $N_{atoms} = 816$, (g) RAYFEF,[315] $N_{atoms} = 1584$. . . . 72

4.11 The correlation between the W99 and DFT relative lattice energy of pentacene, 5A and 5B crystals, for W99-optimized geometries. 74

4.12 Learning curves for the lattice energy predictions of pentacene, 5A and 5B datasets on a logarithmic scale. All hyper-parameters of our ML model are fixed except for the regularization parameter λ in the GPR model which is optimized on the fly at each training. We use 4-fold cross validation on the randomly shuffled dataset and randomly draw N times an increasing number of training samples from 75% of the dataset for each fold. The test MAE and error bars are, respectively, average and standard deviation over the folds. The left-hand panel corresponds to the prediction of W99 energies computed for W99-optimized geometries, the right-hand panel correspond to the prediction of DFT energies on such structures, and the bottom panel to the prediction of the difference between DFT and a W99 baseline. 76

4.13 First 200 largest eigenvalues corresponding to the centered kernel matrices[140] of pentacene, 5A and 5B datasets with cutoff radius of 5Å, gaussian width of 0.3Å and $\gamma = 2$. 78

4.14 Learning curves for the errors in predicting TI when selecting training dimers using a random or the FPS strategy. MAE, RMSE and SUP errors are defined in the text. 79

4.15 Learning curves for the MAE in predicting TI when using the FPS selection of the training set. The three systems are compared as a function of the fraction of the total symmetry-independent dimer configurations. 79

4.16 Learning curves for the elpasolite crystals. The standard SOAP curve is shown in black, the best curve from Ref.[104] is shown in bright red and the optimized curves are shown in dark red ($d_J = 1$), purple ($d_J = 2$) and blue ($d_J = 4$). For each of these models, the kernels were constructed with $r_{\text{cut}} = 5$Å and $\zeta = 1$. The multiple-kernel model (shown in grey) combines three standard SOAP kernels ($\zeta = 1$, $r_{\text{cut}} = 4$; $\zeta = 1$, $r_{\text{cut}} = 6$; $\zeta = 4$, $r_{\text{cut}} = 6$) and one optimized kernel ($d_J = 4$, $\zeta = 1$, $r_{\text{cut}} = 5$) in the ratio $4:3:1:220$. All of the kernels were constructed with $\nu = 2$, $n_{\text{max}} = 12$ radial basis functions and $l_{\text{max}} = 9$ non-degenerate spherical harmonics. Error bars are omitted because they are as small as the data point markers. 85

4.17 Data-driven representations of the chemical space. (a) A 2D map of the elements contained in the elpasolite data set, with the coordinates corresponding to u_{1a} and u_{2a}, for the case $d_J = 2$. Points are colored according to the group. (b) A periodic table colored according to the coordinates in the 2D chemical space. u_{1a} corresponds to the red channel and u_{2a} to the blue channel. (c) A periodic table colored according to u_{1a} (red channel) for a 1D chemical space. (d) A periodic table colored according to 4D chemical coordinates (u_{1a}: red channel, u_{2a}: green channel, u_{3a}: blue channel, u_{4a}: hatches opacity) 87

4.18 Learning curves for the QM9 data set. Four of the lines show the MAE on the test set for various standard SOAP kernels ($\zeta = 2$) with different cutoff radii (dashed lines graduating from red to blue). The other lines show the MAE on the test set for the optimal radially-scaled (RS) and multiple-kernel (MK) SOAP models (black and grey lines respectively). In every model, the kernels were constructed with $\nu = 2$, $n_{\text{max}} = 12$ radial basis functions and $l_{\text{max}} = 9$ non-degenerate spherical harmonics. The inset shows the radial-scaling function $u(r)$ from $r = 0$Å to $r = 5$Å with the parameters that were found to minimize the ten-fold cross validation MAE on the optimization set through a grid search, $r_0 = 2$Å and $m = 7$. The multiple-kernel model combines the $r_{\text{cut}} = 2, 3, 4$ and RS kernels in the ratio $100,000:1:2:10,000$, and the learning curve agrees with the RS result to within graphical accuracy. Error bars are omitted because they are as small as the data point markers. Note that errors are expressed on a per-atom basis. Error per molecule expressed in kcal/mol can be obtained approximately by multiplying the scale by 0.4147, that is computed based on the average size of a molecule in the QM9 database. 89

4.19 Learning curves for the QM9 data set after inclusion of radially-scaled and alchemically-optimized SOAP kernels. Standard SOAP kernels with different cut-off radii are compared with the result of optimizing alchemical correlations using the scheme presented previously for the elpasolite crystal data set (blue and red lines). The learning curve of the optimized radially-scaled kernel (dashed black line with circles) is improved through inclusion of a Gaussian alchemical kernel (dashed black line with squares), which was optimized specifically for $\zeta = 2$ using a grid search. The combined optimization of the radial scaling and alchemical correlations leads to a model that matches the accuracy of the state of the art curve (dashed red line), which corresponds to the representations from Ref.[104], with the errors normalized by the average size of a molecules in the QM9 database. In every SOAP-based model, the kernels were constructed with $\nu = 2$, $n_{\max} = 12$ radial basis functions and $l_{\max} = 9$ non-degenerate spherical harmonics. Error bars are omitted because they are as small as the data point markers. 91

4.20 Distribution of ^{1}H chemical shielding predictions. The colored solid lines show contours of $P(\ln\epsilon_t|\ln\sigma)$, while the colored dashed lines show contours of $P(\ln\epsilon_m|\ln\sigma)$ (see Eq. (4.31) and the corresponding contour levels are shown in the legend. The grayscale density plot and the solid black line respectively correspond to the marginal distribution of the predicted uncertainty $P(\ln\sigma)$ and to $y = x$. 98

4.21 Distribution of ^{1}H chemical shielding predictions. The solid lines show contours of $P(\ln\epsilon_t|\ln\sigma)$, while the dashed lines show contours of $P(\ln\epsilon_m|\ln\sigma)$ (see Eq. (4.31)), including a non-linear scaling of the uncertainty corresponding to Eq. (4.25) with $\gamma \neq 0$, and the corresponding contour levels are shown in the legend. The grayscale density plot and the solid black line respectively correspond to the marginal distribution of the predicted uncertainty $P(\ln\sigma)$ and to $y = x$. . 100

4.22 Distribution of formation energy differences for the QM9 dataset. The colored solid lines show contours of $P(\ln\epsilon_t|\ln\sigma)$, while the colored dashed lines show the contours of $P(\ln\epsilon_m|\ln\sigma)$ (see Eq. (4.31)), including a non-linear scaling of the uncertainty corresponding to Eq. (4.25) with $\gamma \neq 0$, and the corresponding contour levels are shown in the legend. The grayscale density plot and the solid black line respectively correspond to the marginal distribution of the predicted uncertainty $P(\ln\sigma)$ and to $y = x$. The red dots show the distribution of actual errors and predicted uncertainties for 100 randomly selected structures that failed the SMILES consistency test when the QM9 dataset was constructed. . . 103

4.23 Distribution of formation energy differences for the Elpasolite crystal dataset. The colored solid lines show contours of $P(\ln\epsilon_t|\ln\sigma)$, while the colored dashed lines show contours of $P(\ln\epsilon_m|\ln\sigma)$ (see Eq. (4.31)), including a non-linear scaling of the uncertainty corresponding to Eq. (4.25) with $\gamma \neq 0$, and the corresponding contour levels are shown in the legend. The grayscale density plot and the solid black line respectively correspond to the marginal distribution of the predicted uncertainty $P(\ln\sigma)$ and to $y = x$. 105

4.24 Time taken to compute the GPR PP and sub-sampling (SS) uncertainty estimates (Eqs. (4.7) and (4.21) respectively) as a function of the active set size for 2.5k Elpasolite crystal structures as a test set. 64 training set sub-samples were used to produce the sub-sampling result (i.e. 64 linear regression weights). The time taken to generate the kernels between active and test points is excluded from the reported computation times. The simulations were performed using the NumPy module in Python for the necessary linear algebra (one Intel(R) Xeon(R) CPU E5-4627 v2 @ 3.30GHz core). 106

List of Tables

4.1 Kernel and GPR parameters. The GPR parameters (λ and ζ) are the ones used in single kernel predictions. 62

4.2 Kernel weights and GPR parameters used for multi-scale kernel prediction. . . 63

4.3 Summary of the lattice energy prediction scores for pentacene, 5A and 5B (respectively 564, 594 and 936 structures). Our best accuracies on these datasets are estimated from average scores from a 4-fold cross validation (75% of the dataset is used for training). Δ-learning refers to the learning of the difference between W99 and DFT energies. 76

4.4 Parameters used for the grid search of the optimal radial scaling on QM9. All the possible combinations of the three parameters were evaluated. 90

4.5 Log likelihood (LL) of chemical shielding predictions on the test set for different sub-sample sizes. After scaling the variances through maximum likelihood estimation – internally (Int.) or on the validation set (Val.) – the final log likelihood is insensitive to the sub-sample size. A non-linear scaling of the uncertainty (N-L), i.e. $\gamma \neq 0$, further improves the quality of the uncertainty estimates. To normalize the results, the log likelihood (-2.560) of a model with a constant mean and variance corresponding to the empirical chemical shielding mean and variance of the full training set has been subtracted from each value. 101

4.6 Log likelihood (LL) of formation energy predictions (QM9 dataset) on the test set for different sub-sample sizes. After scaling the variances through maximum likelihood estimation – internally (Int.) or on the validation set (Val.) – the final log likelihood is insensitive to the sub-sample size. To normalize the results, the log likelihood (0.207) of a model with a constant mean and variance corresponding to the empirical formation energy mean and variance of the full training set has been subtracted from each value. 102

4.7 Log likelihood (LL) of formation energy predictions (Elpasolite crystal dataset) on the test set for different sub-sample sizes. After scaling the variances through maximum likelihood estimation – internally (Int.) or on the validation set (Val.) – the final log likelihood is insensitive to the sub-sample size. A non-linear scaling of the uncertainty (N-L) further improves the uncertainty model. To normalize the results, the log likelihood (-1.353) of a model with a constant mean and variance corresponding to the empirical formation energy mean and variance of the full training set has been subtracted from each value. 104

Bibliography

[1] Ion Errea, Francesco Belli, Lorenzo Monacelli, Antonio Sanna, Takashi Koretsune, Terumasa Tadano, Raffaello Bianco, Matteo Calandra, Ryotaro Arita, Francesco Mauri, and José A. Flores-Livas. "Quantum crystal structure in the 250-kelvin superconducting lanthanum hydride". In: *Nature* 578.7793 (Feb. 2020), pp. 66–69. DOI: `10.1038/s41586-020-1955-z`.

[2] Luis A. Zepeda-Ruiz, Alexander Stukowski, Tomas Oppelstrup, and Vasily V. Bulatov. "Probing the limits of metal plasticity with molecular dynamics simulations". In: *Nature* 550.7677 (Oct. 2017), pp. 492–495. DOI: `10.1038/nature23472`.

[3] Cedric Klinkert, Áron Szabó, Christian Stieger, Davide Campi, Nicola Marzari, and Mathieu Luisier. "2-D Materials for Ultrascaled Field-Effect Transistors: One Hundred Candidates under the <i>Ab Initio</i> Microscope". In: *ACS Nano* 14.7 (July 2020), pp. 8605–8615. DOI: `10.1021/acsnano.0c02983`.

[4] Muratahan Aykol, Soo Kim, Vinay I. Hegde, David Snydacker, Zhi Lu, Shiqiang Hao, Scott Kirklin, Dane Morgan, and C. Wolverton. "High-throughput computational design of cathode coatings for Li-ion batteries". In: *Nature Communications* 7.1 (Dec. 2016), p. 13779. DOI: `10.1038/ncomms13779`.

[5] Richard M. Martin. *Electronic Structure*. Cambridge: Cambridge University Press, 2004. DOI: `10.1017/CBO9780511805769`.

[6] Christopher J. Cramer. *Essentials of Computational Chemistry Theories and Models*. Vol. 42. 2. Wiley, 2004, pp. 334–342. DOI: `10.1021/ci010445m`.

[7] Frank Neese, Mihail Atanasov, Giovanni Bistoni, Dimitrios Maganas, and Shengfa Ye. *Chemistry and Quantum Mechanics in 2019: Give Us Insight and Numbers*. Feb. 2019. DOI: `10.1021/jacs.8b13313`.

[8] Brian M. Austin, Dmitry Yu. Zubarev, and William A. Lester. *Quantum monte carlo and related approaches*. Jan. 2012. DOI: `10.1021/cr2001564`.

[9] E.K.U. Gross and R.M. Dreizler. *Density Functional Theory: An Approximation to the Quantum Many-Body Problems*. Springer Berlin Heidelberg, 1990, p. 302.

[10] J A Elliott. *Novel approaches to multiscale modelling in materials science*. July 2011. DOI: `10.1179/1743280410Y.0000000002`.

[11] Daan Frenkel and Berend Smit. *Understanding molecular simulation: From algorithms to applications.* 1996, p. 638. DOI: 10.1016/B978-0-12-267351-1.X5000-7.

[12] Dominik Marx and Jürg Hutter. *Ab initio molecular dynamics: Basic theory and advanced methods.* Vol. 9780521898. Cambridge University Press, 2009, pp. 1–567. DOI: 10.1017/CBO9780511609633.

[13] John Bauer, Stephen Spanton, Rodger Henry, John Quick, Walter Dziki, William Porter, and John Morris. "Ritonavir: an extraordinary example of conformational polymorphism". In: *Pharm. Res* 18.6 (2001), p. 859.

[14] Anthony M. Reilly, Richard I. Cooper, Claire S. Adjiman, Saswata Bhattacharya, A. Daniel Boese, Jan Gerit Brandenburg, Peter J. Bygrave, Rita Bylsma, Josh E. Campbell, Roberto Car, David H. Case, Renu Chadha, Jason C. Cole, Katherine Cosburn, Herma M. Cuppen, Farren Curtis, Graeme M. Day, Robert A. DiStasio, Alexander Dzyabchenko, Bouke P. Van Eijck, Dennis M. Elking, Joost A. Van Den Ende, Julio C. Facelli, Marta B. Ferraro, Laszlo Fusti-Molnar, Christina Anna Gatsiou, Thomas S. Gee, René De Gelder, Luca M. Ghiringhelli, Hitoshi Goto, Stefan Grimme, Rui Guo, Detlef W.M. Hofmann, Johannes Hoja, Rebecca K. Hylton, Luca Iuzzolino, Wojciech Jankiewicz, Daniël T. De Jong, John Kendrick, Niek J.J. De Klerk, Hsin Yu Ko, Liudmila N. Kuleshova, Xiayue Li, Sanjaya Lohani, Frank J.J. Leusen, Albert M. Lund, Jian Lv, Yanming Ma, Noa Marom, Artëm E. Masunov, Patrick McCabe, David P. McMahon, Hugo Meekes, Michael P. Metz, Alston J. Misquitta, Sharmarke Mohamed, Bartomeu Monserrat, Richard J. Needs, Marcus A. Neumann, Jonas Nyman, Shigeaki Obata, Harald Oberhofer, Artem R. Oganov, Anita M. Orendt, Gabriel I. Pagola, Constantinos C. Pantelides, Chris J. Pickard, Rafal Podeszwa, Louise S. Price, Sarah L. Price, Angeles Pulido, Murray G. Read, Karsten Reuter, Elia Schneider, Christoph Schober, Gregory P. Shields, Pawanpreet Singh, Isaac J. Sugden, Krzysztof Szalewicz, Christopher R. Taylor, Alexandre Tkatchenko, Mark E. Tuckerman, Francesca Vacarro, Manolis Vasileiadis, Alvaro Vazquez-Mayagoitia, Leslie Vogt, Yanchao Wang, Rona E. Watson, Gilles A. De Wijs, Jack Yang, Qiang Zhu, and Colin R. Groom. "Report on the sixth blind test of organic crystal structure prediction methods". In: *Acta Crystallographica Section B: Structural Science, Crystal Engineering and Materials* 72.4 (Aug. 2016), pp. 439–459. DOI: 10.1107/S2052520616007447.

[15] Si-sheng Ou-Yang, Jun-yan Lu, Xiang-qian Kong, Zhong-jie Liang, Cheng Luo, and Hualiang Jiang. "Computational drug discovery". In: *Acta Pharmacologica Sinica* 33.9 (Sept. 2012), pp. 1131–1140. DOI: 10.1038/aps.2012.109.

[16] Alfred Ludwig. "Discovery of new materials using combinatorial synthesis and high-throughput characterization of thin-film materials libraries combined with computational methods". In: *npj Computational Materials* 5.1 (Dec. 2019), p. 70. DOI: 10.1038/s41524-019-0205-0.

[17] Ivano E. Castelli, Thomas Olsen, Soumendu Datta, David D. Landis, Søren Dahl, Kristian S. Thygesen, and Karsten W. Jacobsen. "Computational screening of perovskite metal oxides for optimal solar light capture". In: *Energy and Environmental Science* 5.2 (Jan. 2012), pp. 5814–5819. DOI: 10.1039/c1ee02717d.

[18] Giovanni Pizzi, Andrea Cepellotti, Riccardo Sabatini, Nicola Marzari, and Boris Kozinsky. “AiiDA: automated interactive infrastructure and database for computational science”. In: *Computational Materials Science* 111 (2016), pp. 218–230. DOI: `http://dx.doi.org/10.1016/j.commatsci.2015.09.013`.

[19] Johannes Hachmann, Roberto Olivares-Amaya, Sule Atahan-Evrenk, Carlos Amador-Bedolla, Roel S. Sánchez-Carrera, Aryeh Gold-Parker, Leslie Vogt, Anna M. Brockway, and Alán Aspuru-Guzik. “The Harvard Clean Energy Project: Large-Scale Computational Screening and Design of Organic Photovoltaics on the World Community Grid”. In: *The Journal of Physical Chemistry Letters* 2.17 (2011), pp. 2241–2251. DOI: `10.1021/jz200866s`.

[20] C. Ortiz, O. Eriksson, and M. Klintenberg. “Data mining and accelerated electronic structure theory as a tool in the search for new functional materials”. In: *Computational Materials Science* 44.4 (Feb. 2009), pp. 1042–1049. DOI: `10.1016/j.commatsci.2008.07.016`.

[21] James E. Saal, Scott Kirklin, Muratahan Aykol, Bryce Meredig, and C. Wolverton. “Materials Design and Discovery with High-Throughput Density Functional Theory: The Open Quantum Materials Database (OQMD)”. In: *JOM* 65.11 (2013), pp. 1501–1509. DOI: `10.1007/s11837-013-0755-4`.

[22] P. Villars, M. Berndt, K. Brandenburg, K. Cenzual, J. Daams, F. Hulliger, T. Massalski, H. Okamoto, K. Osaki, A. Prince, H. Putz, and S. Iwata. “The Pauling File, Binaries Edition”. In: *Journal of Alloys and Compounds* 367.1–2 (2004), pp. 293–297. DOI: `http://dx.doi.org/10.1016/j.jallcom.2003.08.058`.

[23] Anubhav Jain, Shyue Ping Ong, Geoffroy Hautier, Wei Chen, William Davidson Richards, Stephen Dacek, Shreyas Cholia, Dan Gunter, David Skinner, Gerbrand Ceder, and Kristin A. Persson. “Commentary: The Materials Project: A materials genome approach to accelerating materials innovation”. In: *APL Mater.* 1.1 (2013), p. 011002. DOI: `http://dx.doi.org/10.1063/1.4812323`.

[24] Ashley White. “The Materials Genome Initiative: One year on”. In: *MRS Bulletin* 37.8 (Aug. 2012), pp. 715–716. DOI: `10.1557/mrs.2012.194`.

[25] Raghunathan Ramakrishnan, Pavlo O. Dral, Matthias Rupp, and O. Anatole von Lilienfeld. “Quantum chemistry structures and properties of 134 kilo molecules”. In: *Sci. Data* 1 (Aug. 2014), pp. 1–7. DOI: `10.1038/sdata.2014.22`.

[26] Stefano Curtarolo, Wahyu Setyawan, Gus L.W. Hart, Michal Jahnatek, Roman V. Chepulskii, Richard H. Taylor, Shidong Wang, Junkai Xue, Kesong Yang, Ohad Levy, Michael J. Mehl, Harold T. Stokes, Denis O. Demchenko, and Dane Morgan. “AFLOW: An automatic framework for high-throughput materials discovery”. In: *Computational Materials Science* 58 (June 2012), pp. 218–226. DOI: `10.1016/j.commatsci.2012.02.005`.

[27] Kristin M. Tolle, D. Stewart W. Tansley, and Anthony J.G. Hey. “The fourth Paradigm: Data-intensive scientific discovery”. In: *Proceedings of the IEEE*. Vol. 99. 8. Oct. 2011, pp. 1334–1337. DOI: `10.1109/JPROC.2011.2155130`.

[28] Stefano Curtarolo, Gus L. W. Hart, Marco Buongiorno Nardelli, Natalio Mingo, Stefano Sanvito, and Ohad Levy. “The high-throughput highway to computational materials design”. In: *Nature Materials* 12.3 (Feb. 2013), pp. 191–201. DOI: `10.1038/nmat3568`.

[29] Jörg Behler and Michele Parrinello. “Generalized Neural-Network Representation of High-Dimensional Potential-Energy Surfaces”. In: *Phys. Rev. Lett.* 98.14 (Apr. 2007), p. 146401. DOI: `10.1103/PhysRevLett.98.146401`.

[30] Albert P. Bartók, Mike C. Payne, Risi Kondor, and Gábor Csányi. “Gaussian Approximation Potentials: The Accuracy of Quantum Mechanics, without the Electrons”. In: *Phys. Rev. Lett.* 104.13 (Apr. 2010), p. 136403. DOI: `10.1103/PhysRevLett.104.136403`.

[31] Dipti Jasrasaria, Edward O. Pyzer-Knapp, Dmitrij Rappoport, and Alan Aspuru-Guzik. “Space-Filling Curves as a Novel Crystal Structure Representation for Machine Learning Models”. In: *http://arxiv.org/abs/1608.05747* (Aug. 2016).

[32] Raghunathan Ramakrishnan, Pavlo O. Dral, Matthias Rupp, and O. Anatole Von Lilienfeld. “Big data meets quantum chemistry approximations: The Δ-machine learning approach”. In: *Journal of Chemical Theory and Computation* 11.5 (May 2015), pp. 2087–2096. DOI: `10.1021/acs.jctc.5b00099`.

[33] Felix A. Faber, Alexander Lindmaa, O. Anatole Von Lilienfeld, and Rickard Armiento. “Machine Learning Energies of 2 Million Elpasolite (ABC2D6) Crystals”. In: *Phys. Rev. Lett.* 117.13 (Sept. 2016), p. 135502. DOI: `10.1103/PhysRevLett.117.135502`.

[34] Atsuto Seko, Hiroyuki Hayashi, Keita Nakayama, Akira Takahashi, and Isao Tanaka. “Representation of compounds for machine-learning prediction of physical properties”. In: *Physical Review B* 95.14 (Apr. 2017), pp. 1–10. DOI: `10.1103/PhysRevB.95.144110`.

[35] Maarten de Jong, Wei Chen, Randy Notestine, Kristin Persson, Gerbrand Ceder, Anubhav Jain, Mark Asta, and Anthony Gamst. “A Statistical Learning Framework for Materials Science: Application to Elastic Moduli of k-nary Inorganic Polycrystalline Compounds”. In: *Sci. Rep.* 6.1 (Dec. 2016), p. 34256. DOI: `10.1038/srep34256`.

[36] Edward O. Pyzer-Knapp, Kewei Li, and Alan Aspuru-Guzik. “Learning from the Harvard Clean Energy Project: The Use of Neural Networks to Accelerate Materials Discovery”. In: *Adv. Func. Mater.* 25.41 (Nov. 2015), pp. 6495–6502. DOI: `10.1002/adfm.201501919`.

[37] Jesús Carrete, Wu Li, Natalio Mingo, Shidong Wang, and Stefano Curtarolo. “Finding unprecedentedly low-thermal-conductivity half-heusler semiconductors via high-throughput materials modeling”. In: *Physical Review X* 4.1 (Feb. 2014), p. 011019. DOI: `10.1103/PhysRevX.4.011019`.

[38] Ruth Nussinov and Haim J Wolfson. “Efficient detection of three-dimensional structural motifs in biological macromolecules by computer vision techniques”. In: *Proceedings of the National Academy of Sciences* 88.23 (1991), pp. 10495–10499.

[39] Fabio Pietrucci and Wanda Andreoni. “Graph Theory Meets Ab Initio Molecular Dynamics: Atomic Structures and Transformations at the Nanoscale”. In: *Phys. Rev. Lett.* 107 (Aug. 2011), p. 085504. DOI: `10.1103/PhysRevLett.107.085504`.

[40] Piero Gasparotto and Michele Ceriotti. “Recognizing Molecular Patterns by Machine Learning: An Agnostic Structural Definition of the Hydrogen Bond”. In: *J. Chem. Phys.* 141.17 (Nov. 2014), p. 174110. DOI: `10.1063/1.4900655`.

[41] Mary A Rohrdanz, Wenwei Zheng, and Cecilia Clementi. “Discovering Mountain Passes via Torchlight: Methods for the Definition of Reaction Coordinates and Pathways in Complex Macromolecular Reactions.” In: *Annu. Rev. Phys. Chem.* 64 (Jan. 2013), pp. 295–316. DOI: `10.1146/annurev-physchem-040412-110006`.

[42] Alex Rodriguez and Alessandro Laio. “Clustering by Fast Search and Find of Density Peaks”. In: *Science* 344.6191 (2014), pp. 1492–1496. DOI: `10.1126/science.1242072`.

[43] Nicolas Blöchliger, Andreas Vitalis, and Amedeo Caflisch. “High-Resolution visualisation of the states and pathways sampled in molecular dynamics simulations”. In: *Scientific Reports* 4.1 (May 2014), p. 6264. DOI: `10.1038/srep06264`.

[44] Yann Lecun, Yoshua Bengio, and Geoffrey Hinton. “Deep learning”. In: *Nature* 521.7553 (May 2015), pp. 436–444. DOI: `10.1038/nature14539`.

[45] Matthias Rupp, Alexandre Tkatchenko, Klaus-Robert Müller, and O. Anatole von Lilienfeld. “Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning”. In: *Phys. Rev. Lett.* 108.5 (Jan. 2012), p. 058301. DOI: `10.1103/PhysRevLett.108.058301`.

[46] Grégoire Montavon, Matthias Rupp, Vivekanand Gobre, Alvaro Vazquez-Mayagoitia, Katja Hansen, Alexandre Tkatchenko, Klaus Robert Müller, and O. Anatole Von Lilienfeld. “Machine learning of molecular electronic properties in chemical compound space”. In: *New Journal of Physics* 15.9 (Sept. 2013), p. 095003. DOI: `10.1088/1367-2630/15/9/095003`.

[47] Ali Sadeghi, S. Alireza Ghasemi, Bastian Schaefer, Stephan Mohr, Markus A. Lill, and Stefan Goedecker. “Metrics for Measuring Distances in Configuration Spaces”. In: *J. Chem. Phys.* 139.18 (Nov. 2013), p. 184118. DOI: `10.1063/1.4828704`.

[48] Katja Hansen, Franziska Biegler, Raghunathan Ramakrishnan, Wiktor Pronobis, O. Anatole Von Lilienfeld, Klaus Robert Müller, and Alexandre Tkatchenko. “Machine Learning Predictions of Molecular Properties: Accurate Many-Body Potentials and Nonlocality in Chemical Space”. In: *J. Phys. Chem. Lett.* 6.12 (June 2015), pp. 2326–2331. DOI: `10.1021/acs.jpclett.5b00831`.

[49] Bastiaan J. Braams and Joel M. Bowman. “Permutationally Invariant Potential Energy Surfaces in High Dimensionality”. In: *Int. Rev. Phys. Chem.* 28.4 (Oct. 2009), pp. 577–606. DOI: `10.1080/01442350903234923`.

[50] Zhen Xie and Joel M. Bowman. "Permutationally invariant polynomial basis for molecular energy surface fitting via monomial symmetrization". In: *Journal of Chemical Theory and Computation* 6.1 (Jan. 2010), pp. 26–34. DOI: 10.1021/ct9004917.

[51] Bin Jiang and Hua Guo. "Permutation invariant polynomial neural network approach to fitting potential energy surfaces Permutation invariant polynomial neural network approach to fitting". In: *The Journal of Chemical Physics* 054112.5 (Aug. 2013), pp. 0–5. DOI: 10.1063/1.4817187.

[52] K. T. Schütt, H. Glawe, F. Brockherde, A. Sanna, K. R. Müller, and E. K. U. Gross. "How to represent crystal structures for machine learning: Towards fast prediction of electronic properties". In: *Phys. Rev. B* 89.20 (May 2014), p. 205118. DOI: 10.1103/PhysRevB.89.205118.

[53] 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.

[54] A. Shapeev. "Moment Tensor Potentials: A Class of Systematically Improvable Interatomic Potentials". In: *Multiscale Model. Sim.* 14.3 (2016), pp. 1153–1173.

[55] John B. O. Mitchell. "Machine learning methods in chemoinformatics". In: *Wiley Interdisciplinary Reviews: Computational Molecular Science* 4.5 (Sept. 2014), pp. 468–481. DOI: 10.1002/wcms.1183.

[56] G.N. Ramachandran, C. Ramakrishnan, and V. Sasisekharan. "Stereochemistry of Polypeptide Chain Configurations". In: *Journal of Molecular Biology* 7.1 (July 1963), pp. 95–99. DOI: 10.1016/S0022-2836(63)80023-6.

[57] Dmitrij Frishman and Patrick Argos. "Incorporation of Non-Local Interactions in Protein Secondary Structure Prediction from the Amino Acid Sequence". In: *Protein Eng Des Sel* 9.2 (1996), pp. 133–142. DOI: 10.1093/protein/9.2.133.

[58] Wolfgang Kabsch and Christian Sander. "Dictionary of Protein Secondary Structure: Pattern Recognition of Hydrogen-Bonded and Geometrical Features". In: *Biopolymers* 22.12 (Dec. 1983), pp. 2577–2637. DOI: 10.1002/bip.360221211.

[59] David Weininger. "SMILES, a Chemical Language and Information System: 1: Introduction to Methodology and Encoding Rules". In: *Journal of Chemical Information and Computer Sciences* 28.1 (Feb. 1988), pp. 31–36. DOI: 10.1021/ci00057a005.

[60] Michele Ceriotti, Gareth A. Tribello, and Michele Parrinello. "Simplifying the Representation of Complex Free-Energy Landscapes Using Sketch-Map". In: *Proc. Natl. Acad. Sci. U. S. A.* 108.32 (Aug. 2011), pp. 13023–13028. DOI: 10.1073/pnas.1108486108.

[61] Logan Ward, Ankit Agrawal, Alok Choudhary, and Christopher Wolverton. "A general-purpose machine learning framework for predicting properties of inorganic materials". In: *Computational Materials* 2.1 (Nov. 2016), p. 16028. DOI: 10.1038/npjcompumats.2016.28.

[62] Christopher M. Bishop et al. *Pattern recognition and machine learning*. 4. Springer, 2006, p. 12. DOI: 10.1117/1.2819119.

[63] "Crystal structure prediction using the minima hopping method". In: *The Journal of Chemical Physics* 133.22 (Dec. 2010), p. 224104. DOI: 10.1063/1.3512900.

[64] Chris J Pickard and R J Needs. *Ab initio random structure searching*. Feb. 2011. DOI: 10.1088/0953-8984/23/5/053201.

[65] Artem R. Oganov and Colin W. Glass. "Crystal structure prediction using <i>ab initio</i> evolutionary techniques: Principles and applications". In: *The Journal of Chemical Physics* 124.24 (June 2006), p. 244704. DOI: 10.1063/1.2210932.

[66] Li Zhu, Maximilian Amsler, Tobias Fuhrer, Bastian Schaefer, Somayeh Faraji, Samare Rostami, S. Alireza Ghasemi, Ali Sadeghi, Migle Grauzinyte, Chris Wolverton, and Stefan Goedecker. "A fingerprint based metric for measuring similarities of crystalline structures". In: *The Journal of Chemical Physics* 144.3 (Jan. 2016), p. 034203. DOI: 10.1063/1.4940026.

[67] Grégoire Ferré, Jean Bernard Maillet, and Gabriel Stoltz. "Permutation-invariant distance between atomic configurations". In: *Journal of Chemical Physics* 143.10 (Sept. 2015), p. 104114. DOI: 10.1063/1.4930541.

[68] Gregory R. Bowman, Kyle A. Beauchamp, George Boxer, and Vijay S. Pande. "Progress and challenges in the automated construction of Markov state models for full protein systems". In: *Journal of Chemical Physics* 131.12 (Sept. 2009), p. 124101. DOI: 10.1063/1.3216567.

[69] Jan Hendrik Prinz, Hao Wu, Marco Sarich, Bettina Keller, Martin Senne, Martin Held, John D. Chodera, Christof Schtte, and Frank Noé. "Markov models of molecular kinetics: Generation and validation". In: *Journal of Chemical Physics* 134.17 (May 2011), p. 174105. DOI: 10.1063/1.3565032.

[70] Andreas Vitalis and Amedeo Caflisch. "Efficient construction of mesostate networks from molecular dynamics trajectories". In: *Journal of Chemical Theory and Computation* 8.3 (Mar. 2012), pp. 1108–1120. DOI: 10.1021/ct200801b.

[71] Guillermo Pérez-Hernández, Fabian Paul, Toni Giorgino, Gianni De Fabritiis, and Frank Noé. "Identification of slow molecular order parameters for Markov model construction". In: *Journal of Chemical Physics* 139.1 (July 2013), p. 015102. DOI: 10.1063/1.4811489.

[72] Mary A. Rohrdanz, Wenwei Zheng, Mauro Maggioni, and Cecilia Clementi. "Determination of reaction coordinates via locally scaled diffusion map". In: *The Journal of Chemical Physics* 134.12 (Mar. 2011), p. 124116. DOI: 10.1063/1.3569857.

[73] Andrew L Ferguson, Athanassios Z Panagiotopoulos, Pablo G Debenedetti, and Ioannis G Kevrekidis. "Systematic Determination of Order Parameters for Chain Dynamics Using Diffusion Maps." In: *Proc. Natl. Acad. Sci. U. S. A.* 107.31 (Aug. 2010), pp. 13597–602. DOI: 10.1073/pnas.1003293107.

[74] Olexandr Isayev, Denis Fourches, Eugene N. Muratov, Corey Oses, Kevin Rasch, Alexander Tropsha, and Stefano Curtarolo. “Materials Cartography: Representing and Mining Materials Space Using Structural and Electronic Fingerprints”. In: *Chemistry of Materials* 27.3 (Feb. 2015), pp. 735–743. DOI: `10.1021/cm503507h`.

[75] Sandip De, Albert P. Bartók, Gábor Csányi, and Michele Ceriotti. “Comparing Molecules and Solids across Structural and Alchemical Space”. In: *Phys. Chem. Chem. Phys.* 18.20 (2016), pp. 13754–13769. DOI: `10.1039/c6cp00415f`.

[76] David Wales. *Energy Landscapes: Applications to Clusters, Biomolecules and Glasses.* Cambridge University Press, 2003. DOI: `10.1017/cbo9780511721724`.

[77] Shun-ichi Amari and Noboru Murata. “Statistical Theory of Learning Curves under Entropic Loss Criterion”. In: *Neural Computation* 5.1 (Jan. 1993), pp. 140–153. DOI: `10.1162/neco.1993.5.1.140`.

[78] Tuckerman Mark. *Statistical Mechanics: Theory and Molecular Simulation.* Oxford University Press, 1972, p. 696.

[79] Matthias Rupp, Raghunathan Ramakrishnan, and O. Anatole Von Lilienfeld. “Machine Learning for Quantum Mechanical Properties of Atoms in Molecules”. In: *Journal of Physical Chemistry Letters* 6.16 (Aug. 2015), pp. 3309–3313. DOI: `10.1021/acs.jpclett.5b01456`.

[80] Beomsoo Han, Yifeng Liu, Simon W. Ginzinger, and David S. Wishart. “SHIFTX2: significantly improved protein chemical shift prediction”. In: *Journal of Biomolecular NMR* 50.1 (Mar. 2011), p. 43. DOI: `10.1007/s10858-011-9478-4`.

[81] Ganesh Hegde and R Chris Bowen. “Machine-learned approximations to Density Functional Theory Hamiltonians”. In: *Scientific reports* 7 (Feb. 2016), p. 42669. DOI: `10.1038/srep42669`.

[82] Michael J. Willatt, Félix Musil, and Michele Ceriotti. “Atom-Density Representations for Machine Learning”. In: *J. Chem. Phys.* 150.15 (Apr. 2019), p. 154110. DOI: `10.1063/1.5090481`.

[83] Sandip De, Felix Musil, Teresa Ingram, Carsten Baldauf, and Michele Ceriotti. “Mapping and Classifying Molecules from a High-Throughput Structural Database”. In: *J. Cheminformatics* 9.1 (Dec. 2017), pp. 1–14. DOI: `10.1186/s13321-017-0192-4`.

[84] Félix Musil, Sandip De, Jack Yang, Joshua E. Campbell, Graeme M. Day, and Michele Ceriotti. “Machine Learning for the Structure-Energy-Property Landscapes of Molecular Crystals”. In: *Chem. Sci.* 9.5 (2018), pp. 1289–1300. DOI: `10.1039/c7sc04665k`.

[85] M. Ropo, C. Baldauf, and V. Blum. “Energy/structure database of all proteinogenic amino acids and dipeptides without and with divalent cations”. In: *ArXiv e-prints* (Apr. 2015).

[86] Matti Ropo, Markus Schneider, Carsten Baldauf, and Volker Blum. “First-principles data set of 45,892 isolated and cation-coordinated conformers of 20 proteinogenic amino acids”. In: *Scientific Data* 3 (2016), p. 160009. DOI: `10.1038/sdata.2016.9`.

[87] Josh E. Campbell, Jack Yang, and Graeme M. Day. “Predicted energy–structure–function maps for the evaluation of small molecule organic semiconductors”. In: *J. Mater. Chem. C* 5.30 (Aug. 2017), pp. 7574–7584. DOI: `10.1039/C7TC02553J`.

[88] Carl E. Rasmussen and Christopher K. I. Williams. *Gaussian processes for machine learning.* 2. World Scientific Publishing Company, Apr. 2006, pp. 69–106. DOI: `10.1142/S0129065704001899`.

[89] Federico M. Paruzzo, Albert Hofstetter, Félix Musil, Sandip De, Michele Ceriotti, and Lyndon Emsley. “Chemical shifts in molecular solids by machine learning”. In: *Nature Communications* 9.1 (Dec. 2018), p. 4501. DOI: `10.1038/s41467-018-06972-x`.

[90] C. R. Groom, I. J. Bruno, M. P. Lightfoot, and S. C. Ward. “The Cambridge Structural Database”. In: *Acta Crystallographica Section B: Structural Science, Crystal Engineering and Materials* 72.2 (Apr. 2016), pp. 171–179. DOI: `10.1107/S2052520616003954`.

[91] Aldo Glielmo, Peter Sollich, and Alessandro De Vita. “Accurate Interatomic Force Fields via Machine Learning with Covariant Kernels”. In: *Phys. Rev. B* 95.21 (June 2017), p. 214302. DOI: `10.1103/PhysRevB.95.214302`.

[92] Andrea Grisafi, David M. Wilkins, Gábor Csányi, and Michele Ceriotti. “Symmetry-Adapted Machine Learning for Tensorial Properties of Atomistic Systems”. In: *Phys. Rev. Lett.* 120.3 (Jan. 2018), p. 036002. DOI: `10.1103/PhysRevLett.120.036002`.

[93] L Nachbin. *The Haar integral.* R. E. Krieger Pub. Co., 1976.

[94] Aldo Glielmo, Claudio Zeni, and Alessandro De Vita. “Efficient Nonparametric n -Body Force Fields from Machine Learning”. In: *Phys. Rev. B* 97.18 (May 2018), p. 184307. DOI: `10.1103/PhysRevB.97.184307`.

[95] E Prodan and W Kohn. “Nearsightedness of Electronic Matter”. In: *Proc. Natl. Acad. Sci.* 102.33 (Aug. 2005), pp. 11635–11638. DOI: `10.1073/pnas.0505436102`.

[96] S Goedecker. “Linear Scaling Electronic Structure Methods”. In: *Rev. Mod. Phys.* 71.4 (1999), pp. 1085–1123.

[97] Manthos G. Papadopoulos, Robert Zalesny, and Paul G. Mezey. *Linear-Scaling Techniques in Computational Chemistry and Physics.* Dordrecht: Springer Netherlands, 2011, p. 536. DOI: `10.1007/978-90-481-2853-2`.

[98] D. R. Bowler and T. Miyazaki. “O(N) methods in electronic structure calculations”. In: *Reports Prog. Phys.* 75.3 (2012), p. 1. DOI: `10.1088/0034-4885/75/3/036503`.

[99] Angelo Ziletti, Devinder Kumar, Matthias Scheffler, and Luca M. Ghiringhelli. “Insightful Classification of Crystal Structures Using Deep Learning”. In: *Nat. Commun.* 9.1 (Dec. 2018), p. 2775. DOI: `10.1038/s41467-018-05169-6`.

[100] Ralf Drautz. “Atomic Cluster Expansion for Accurate and Transferable Interatomic Potentials”. In: *Phys. Rev. B* 99.1 (Jan. 2019), p. 014104. DOI: `10.1103/PhysRevB.99.014104`.

[101] Markus Bachmayr, Gabor Csanyi, Ralf Drautz, Genevieve Dusson, Simon Etter, Cas van der Oord, and Christoph Ortner. "Atomic Cluster Expansion: Completeness, Efficiency and Stability". In: (Nov. 2019).

[102] Sergey N. Pozdnyakov, Michael J. Willatt, Albert P. Bartók, Christoph Ortner, Gábor Csányi, and Michele Ceriotti. "Incompleteness of Atomic Structure Representations". In: *Physical Review Letters* 125.16 (Oct. 2020), p. 166001. DOI: `10.1103/PhysRevLett.125.166001`.

[103] O. Anatole von Lilienfeld, Raghunathan Ramakrishnan, Matthias Rupp, and Aaron Knoll. "Fourier Series of Atomic Radial Distribution Functions: A Molecular Fingerprint for Machine Learning Models of Quantum Chemical Properties". In: *Int. J. Quantum Chem.* 115.16 (Aug. 2015), pp. 1084–1093. DOI: `10.1002/qua.24912`.

[104] Felix A. Faber, Anders S. Christensen, Bing Huang, and O. Anatole Von Lilienfeld. "Alchemical and Structural Distribution Based Representation for Universal Quantum Machine Learning". In: *J. Chem. Phys.* 148.24 (June 2018), p. 241717. DOI: `10.1063/1.5020710`.

[105] Linfeng Zhang, Jiequn Han, Han Wang, Roberto Car, and Weinan E. "Deep Potential Molecular Dynamics: A Scalable Model with the Accuracy of Quantum Mechanics". In: *Phys. Rev. Lett.* 120 (2018), p. 143001.

[106] A. P. Thompson, L. P. Swiler, C. R. Trott, S. M. Foiles, and G. J. Tucker. "Spectral neighbor analysis method for automated generation of quantum-accurate interatomic potentials". In: *Journal of Computational Physics* 285 (Mar. 2015), pp. 316–330. DOI: `10.1016/j.jcp.2014.12.018`.

[107] Jigyasa Nigam, Sergey Pozdnyakov, and Michele Ceriotti. "Recursive evaluation and iterative contraction of N-body equivariant features". In: *Journal of Chemical Physics* 153.12 (Sept. 2020), p. 121101. DOI: `10.1063/5.0021116`.

[108] Andrea Grisafi, Alberto Fabrizio, Benjamin Meyer, David M. Wilkins, Clemence Corminboeuf, and Michele Ceriotti. "Transferable Machine-Learning Model of the Electron Density". In: *ACS Central Science* 5.1 (Jan. 2019), pp. 57–64. DOI: `10.1021/acscentsci.8b00551`.

[109] David M. Wilkins, Andrea Grisafi, Yang Yang, Ka Un Lao, Robert A. DiStasio, and Michele Ceriotti. "Accurate Molecular Polarizabilities with Coupled Cluster Theory and Machine Learning". In: *Proc. Natl. Acad. Sci. U. S. A.* 116.9 (Feb. 2019), pp. 3401–3406. DOI: `10.1073/pnas.1816132116`.

[110] Fabio Pietrucci and Roman Martoňák. "Systematic Comparison of Crystalline and Amorphous Phases: Charting the Landscape of Water Structures and Transformations". In: *J. Chem. Phys.* 142.10 (2015), p. 104704. DOI: `10.1063/1.4914138`.

[111] Victor M. Panaretos and Yoav Zemel. "Statistical Aspects of Wasserstein Distances". In: *Annual Review of Statistics and Its Application* 6.1 (Mar. 2019). DOI: `10.1146/annurev-statistics-030718-104938`.

[112] Brandon Anderson, Truong Son Hy, and Risi Kondor. “Cormorant: CovariantMolecular neural networks”. In: *arXiv* (June 2019).

[113] Risi Kondor, Zhen Lin, and Shubhendu Trivedi. “Clebsch-Gordan Nets: a Fully Fourier Space Spherical Convolutional Neural Network”. In: *ArXiv e-prints* (June 2018).

[114] Albert P. Bartók, Sandip De, Carl Poelking, Noam Bernstein, James R. Kermode, Gábor Csányi, and Michele Ceriotti. “Machine Learning Unifies the Modeling of Materials and Molecules”. In: *Sci. Adv.* 3.12 (Dec. 2017), e1701816. DOI: `10.1126/sciadv.1701816`.

[115] Bing Huang and O. Anatole von Lilienfeld. “Communication: Understanding molecular representations in machine learning: The role of uniqueness and target similarity”. In: *J. Chem. Phys.* 145.16 (Oct. 2016), p. 161102. DOI: `10.1063/1.4964627`.

[116] Giulio Imbalzano, Andrea Anelli, Daniele Giofré, Sinja Klees, Jörg Behler, and Michele Ceriotti. “Automatic selection of atomic fingerprints and reference configurations for machine-learning potentials”. In: *Journal of Chemical Physics* 148.24 (June 2018), p. 241730. DOI: `10.1063/1.5024611`.

[117] Michael J. Willatt, Félix Musil, and Michele Ceriotti. “Feature Optimization for Atomistic Machine Learning Yields a Data-Driven Construction of the Periodic Table of the Elements”. In: *Phys. Chem. Chem. Phys.* 20.47 (2018), pp. 29661–29668. DOI: `10.1039/c8cp05921g`.

[118] Nongnuch Artrith, Alexander Urban, and Gerbrand Ceder. “Efficient and accurate machine-learning interpolation of atomic energies in compositions with many species”. In: *Physical Review B* 96.1 (July 2017), p. 014112. DOI: `10.1103/PhysRevB.96.014112`.

[119] M. Gastegger, L. Schwiedrzik, M. Bittermann, F. Berzsenyi, and P. Marquetand. “wACSF—Weighted Atom-Centered Symmetry Functions as Descriptors in Machine Learning Potentials”. In: *J. Chem. Phys.* 148.24 (June 2018), p. 241709. DOI: `10.1063/1.5019667`.

[120] Haoyan Huo and Matthias Rupp. “Unified Representation for Machine Learning of Molecules and Crystals”. In: *ArXiv e-prints* (Apr. 2017).

[121] Michael W Mahoney and Petros Drineas. “CUR matrix decompositions for improved data analysis”. In: *Proceedings of the National Academy of Sciences* 106.3 (Jan. 2009), pp. 697–702. DOI: `10.1073/pnas.0803205106`.

[122] Daniel J. Rosenkrantz, Richard E. Stearns, and Philip M. Lewis II. “An Analysis of Several Heuristics for the Traveling Salesman Problem”. In: *SIAM J. Comput.* 6.3 (Sept. 1977), pp. 563–581. DOI: `10.1137/0206041`.

[123] Michele Ceriotti, Gareth A. Tribello, and Michele Parrinello. “Demonstrating the Transferability and the Descriptive Power of Sketch-Map”. In: *J. Chem. Theory Comput.* 9.3 (Mar. 2013), pp. 1521–1532. DOI: `10.1021/ct3010563`.

[124] Gang Yu, Jingzhong Chen, and Li Zhu. “Data Mining Techniques for Materials Informatics: Datasets Preparing and Applications”. In: *Knowledge Acquisition and Modeling, 2009. KAM ’09. Second International Symposium on.* Vol. 2. Nov. 2009, pp. 189–192. DOI: `10.1109/KAM.2009.98`.

[125] Prasanna V. Balachandran, James Theiler, James M. Rondinelli, and Turab Lookman. "Materials Prediction via Classification Learning". In: *Scientific Reports* 5 (Aug. 2015), 13285 EP -.

[126] Gareth A. Tribello, Michele Ceriotti, and Michele Parrinello. "Using Sketch-Map Coordinates to Analyze and Bias Molecular Dynamics Simulations". In: *Proc. Natl. Acad. Sci. U. S. A.* 109.14 (Apr. 2012), pp. 5196–5201. DOI: `10.1073/pnas.1201152109`.

[127] Wojciech J. Szlachta, Albert P. Bartók, and Gábor Csányi. "Accuracy and Transferability of Gaussian Approximation Potential Models for Tungsten". In: *Phys. Rev. B* 90.10 (Sept. 2014), p. 104108. DOI: `10.1103/PhysRevB.90.104108`.

[128] Alejandro Lopez-Bezanilla and O. Anatole von Lilienfeld. "Modeling electronic quantum transport with machine learning". In: *Phys. Rev. B* 89 (June 2014), p. 235411. DOI: `10.1103/PhysRevB.89.235411`.

[129] Ghanshyam Pilania, Chenchen Wang, Xun Jiang, Sanguthevar Rajasekaran, and Ramamurthy Ramprasad. "Accelerating materials property predictions using machine learning". In: *Scientific Reports* 3 (Sept. 2013), 2810 EP -.

[130] Albert P. Bartók, Michael J. Gillan, Frederick R. Manby, and Gábor Csányi. "Machine-Learning Approach for One- and Two-Body Corrections to Density Functional Theory: Applications to Molecular and Condensed Water". In: *Phys. Rev. B* 88.5 (Aug. 2013), p. 054104. DOI: `10.1103/PhysRevB.88.054104`.

[131] Matthias Rupp, Ewgenij Proschak, and Gisbert Schneider. "Kernel Approach to Molecular Similarity Based on Iterative Graph Similarity". In: *J. Chem. Inf. Model.* 47.6 (Nov. 2007), pp. 2280–2286. DOI: `10.1021/ci700274r`.

[132] Matthew Hirn, Nicolas Poilvert, and Stephane Mallat. "Quantum Energy Regression Using Scattering Transforms". In: *ArXiv Prepr. ArXiv150202077* (2015).

[133] John C. Snyder, Matthias Rupp, Katja Hansen, Klaus-Robert Müller, and Kieron Burke. "Finding Density Functionals with Machine Learning". In: *Phys. Rev. Lett.* 108 (June 2012), p. 253002. DOI: `10.1103/PhysRevLett.108.253002`.

[134] S. Alireza Ghasemi, Albert Hofstetter, Santanu Saha, and Stefan Goedecker. "Interatomic potentials for ionic systems with density functional accuracy based on charge densities obtained by a neural network". In: *Phys. Rev. B* 92 (July 2015), p. 045131. DOI: `10.1103/PhysRevB.92.045131`.

[135] O. Anatole von Lilienfeld. "First principles view on chemical compound space: Gaining rigorous atomistic control of molecular properties". In: *International Journal of Quantum Chemistry* 113.12 (June 2013), pp. 1676–1689. DOI: `10.1002/qua.24375`.

[136] Svante Wold, Kim Esbensen, and Paul Geladi. "Principal component analysis". In: *Chemometrics and Intelligent Laboratory Systems* 2.1 (1987), pp. 37–52. DOI: `http://dx.doi.org/10.1016/0169-7439(87)80084-9`.

[137] J. B. Kruskal. "Nonmetric multidimensional scaling: A numerical method". In: *Psychometrika* 29.2 (1964), pp. 115–129. DOI: `10.1007/BF02289694`.

[138] J B Tenenbaum, V de Silva, and J C Langford. “A Global Geometric Framework for Nonlinear Dimensionality Reduction.” In: *Science* 290.5500 (Dec. 2000), pp. 2319–2323. DOI: `10.1126/science.290.5500.2319`.

[139] R. R. Coifman, S. Lafon, A. B. Lee, M. Maggioni, B. Nadler, F. Warner, and S. W. Zucker. “Geometric Diffusions as a Tool for Harmonic Analysis and Structure Definition of Data: Diffusion Maps”. In: *Proc. Natl. Acad. Sci. U. S. A.* 102.21 (2005), pp. 7426–7431. DOI: `10.1073/pnas.0500334102`.

[140] Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Müller. “Nonlinear Component Analysis as a Kernel Eigenvalue Problem”. In: *Neural Computation* 10.5 (July 1998), pp. 1299–1319. DOI: `10.1162/089976698300017467`.

[141] A. K. Jain, M. N. Murty, and P. J. Flynn. “Data Clustering: A Review”. In: *ACM Comput. Surv.* 31.3 (Sept. 1999), pp. 264–323. DOI: `10.1145/331499.331504`.

[142] Rui Xu and II Wunsch D. “Survey of clustering algorithms”. In: *Neural Networks, IEEE Transactions on* 16.3 (May 2005), pp. 645–678. DOI: `10.1109/TNN.2005.845141`.

[143] Charu C Aggarwal and Chandan K Reddy. *Data clustering: algorithms and applications.* CRC Press, 2013.

[144] Fionn Murtagh and Pedro Contreras. “Algorithms for hierarchical clustering: an overview”. In: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 2.1 (2012), pp. 86–97. DOI: `10.1002/widm.53`.

[145] Zhexue Huang. “Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values”. In: *Data Mining and Knowledge Discovery* 2.3 (1998), pp. 283–304. DOI: `10.1023/A:1009769707641`.

[146] L. Jing, M. K. Ng, and J. Z. Huang. “An Entropy Weighting k-Means Algorithm for Subspace Clustering of High-Dimensional Sparse Data”. In: *IEEE Transactions on Knowledge and Data Engineering* 19.8 (Aug. 2007), pp. 1026–1041. DOI: `10.1109/TKDE.2007.1048`.

[147] Mu-Chun Su and Chien-Hsing Chou. “A Modified Version of the K-Means Algorithm with a Distance Based on Cluster Symmetry”. In: *IEEE Trans. Pattern Anal. Mach. Intell.* 23.6 (June 2001), pp. 674–680. DOI: `10.1109/34.927466`.

[148] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. “A density-based algorithm for discovering clusters in large spatial databases with noise”. In: AAAI Press, 1996, pp. 226–231.

[149] Mihael Ankerst, Markus M. Breunig, Hans-peter Kriegel, and Jörg Sander. “OPTICS: Ordering Points To Identify the Clustering Structure”. In: ACM Press, 1999, pp. 49–60.

[150] Ricardo J. G. B. Campello, Davoud Moulavi, Arthur Zimek, and Jörg Sander. “Hierarchical Density Estimates for Data Clustering, Visualization, and Outlier Detection”. In: *ACM Transactions on Knowledge Discovery from Data* 10.1 (July 2015), pp. 1–51. DOI: `10.1145/2733381`.

[151] Bernhard Schölkopf, Ralf Herbrich, and Alex J. Smola. “A Generalized Representer Theorem”. In: *Computational Learning Theory*. Ed. by David Helmbold and Bob Williamson. Berlin, Heidelberg: Springer Berlin Heidelberg, 2001, pp. 416–426. DOI: 10.1007/3-540-44581-1_27.

[152] Marco Cuturi. “Positive Definite Kernels in Machine Learning”. In: *arXiv* 0911 (Nov. 2009), p. 5367.

[153] Marco Cuturi. “Sinkhorn Distances: Lightspeed Computation of Optimal Transport”. In: *Advances in Neural Information Processing Systems 26*. Ed. by C J C Burges, L Bottou, M Welling, Z Ghahramani, and K Q Weinberger. Curran Associates, Inc., 2013, pp. 2292–2300.

[154] Richard Sinkhorn. “Diagonal Equivalence to Matrices with Prescribed Row and Column Sums”. In: *The American Mathematical Monthly* 74.4 (1967), pp. 402–405.

[155] Christian Berg, JPR Christensen, and Paul Ressel. “General Results on Positive and Negative Definite Matrices and Kernels”. In: *Harmonic Analysis on Semigroups* (1984), pp. 86–143. DOI: 10.1007/978-1-4612-1128-0_4.

[156] Xingwang Zhao, Jiye Liang, and Fuyuan Cao. “A simple and effective outlier detection algorithm for categorical data”. In: *International Journal of Machine Learning and Cybernetics* 5.3 (2014), pp. 469–477. DOI: 10.1007/s13042-013-0202-4.

[157] Kenji Yamanishi, Jun-ichi Takeuchi, Graham Williams, and Peter Milne. “On-Line Unsupervised Outlier Detection Using Finite Mixtures with Discounting Learning Algorithms”. In: *Data Mining and Knowledge Discovery* 8.3 (2004), pp. 275–300. DOI: 10.1023/B:DAMI.0000023676.72185.7c.

[158] M. I. Petrovskiy. “Outlier Detection Algorithms in Data Mining Systems”. In: *Programming and Computer Software* 29.4 (2003), pp. 228–237. DOI: 10.1023/A:1024974810270.

[159] Fabrizio Angiulli and Clara Pizzuti. “Fast Outlier Detection in High Dimensional Spaces”. In: *Principles of Data Mining and Knowledge Discovery: 6th European Conference, PKDD 2002 Helsinki, Finland, August 19–23, 2002 Proceedings*. Ed. by Tapio Elomaa, Heikki Mannila, and Hannu Toivonen. Berlin, Heidelberg: Springer Berlin Heidelberg, 2002, pp. 15–27. DOI: 10.1007/3-540-45681-3_2.

[160] Markus M. Breunig, Hans-Peter Kriegel, Raymond T. Ng, and Jörg Sander. “LOF: Identifying Density-based Local Outliers”. In: *SIGMOD Rec.* 29.2 (May 2000), pp. 93–104. DOI: 10.1145/335191.335388.

[161] Charu C. Aggarwal and Philip S. Yu. “Outlier Detection for High Dimensional Data”. In: *SIGMOD Rec.* 30.2 (May 2001), pp. 37–46. DOI: 10.1145/376284.375668.

[162] Volker Blum, Ralf Gehrke, Felix Hanke, Paula Havu, Ville Havu, Xinguo Ren, Karsten Reuter, and Matthias Scheffler. “Ab initio molecular simulations with numeric atom-centered orbitals”. In: *Computer Physics Communications* 180.11 (2009), pp. 2175–2196. DOI: 10.1016/j.cpc.2009.06.022.

[163] John P. Perdew, Kieron Burke, and Matthias Ernzerhof. “Generalized Gradient Approximation Made Simple”. In: *Phys. Rev. Lett.* 77.18 (1996), pp. 3865–3868.

[164] Alexandre Tkatchenko and Matthias Scheffler. “Accurate Molecular Van Der Waals Interactions from Ground-State Electron Density and Free-Atom Reference Data”. In: *Physical Review Letters* 102.7 (Feb. 2009), p. 073005. DOI: `10.1103/PhysRevLett.102.073005`.

[165] Alexandre Tkatchenko, Mariana Rossi, Volker Blum, Joel Ireta, and Matthias Scheffler. “Unraveling the stability of polypeptide helices: Critical role of van der Waals interactions”. In: *Physical Review Letters* 106.11 (2011), p. 118102. DOI: `10.1103/PhysRevLett.106.118102`.

[166] Carsten Baldauf, Kevin Pagel, Stephan Warnke, Gert Von Helden, Beate Koksch, Volker Blum, and Matthias Scheffler. “How cations change peptide structure”. In: *Chemistry - A European Journal* 19.34 (2013), pp. 11224–11234. DOI: `10.1002/chem.201204554`.

[167] Franziska Schubert, Mariana Rossi, Carsten Baldauf, Kevin Pagel, Stephan Warnke, Gert von Helden, Frank Filsinger, Peter Kupser, Gerard Meijer, Mario Salwiczek, Beate Koksch, Matthias Scheffler, and Volker Blum. “Exploring the conformational preferences of 20-residue peptides in isolation: Ac-Ala$_{19}$-Lys + H$^+$ *vs.* Ac-Lys-Ala$_{19}$ + H$^+$ and the current reach of DFT”. In: *Phys. Chem. Chem. Phys.* 17 (2015), pp. 7373–7385.

[168] Franziska Schubert, Kevin Pagel, Mariana Rossi, Stephan Warnke, Mario Salwiczek, Beate Koksch, Gert von Helden, Volker Blum, Carsten Baldauf, and Matthias Scheffler. “Native like helices in a specially designed β peptide in the gas phase”. In: *Phys. Chem. Chem. Phys.* 17 (2015), pp. 5376–5385.

[169] Mariana Rossi, Sucismita Chutia, Matthias Scheffler, and Volker Blum. “Validation Challenge of Density-Functional Theory for Peptides-Example of Ac-Phe-Ala5-LysH(+).” In: *The journal of physical chemistry. A* 118.35 (2014), pp. 7349–59. DOI: `10.1021/jp412055r`.

[170] Carsten Baldauf and Mariana Rossi. “Going clean: structure and dynamics of peptides in the gas phase and paths to solvation.” In: *Journal of physics. Condensed matter : an Institute of Physics journal* 27.49 (2015), p. 493002. DOI: `10.1088/0953-8984/27/49/493002`.

[171] Matti Ropo, Volker Blum, and Carsten Baldauf. “Trends for isolated amino acids and dipeptides: Conformation, divalent ion binding, and remarkable similarity of binding to calcium and lead”. In: *arXiv:1606.02151 [physics, q-bio]* (2016).

[172] Gunter Fischer. “Chemical aspects of peptide bond isomerisation”. In: *Chemical Society Reviews* 29.2 (2000), pp. 119–127. DOI: `10.1039/a803742f`.

[173] Christophe Dugave and Luc Demange. “Cis-trans isomerization of organic molecules and biomolecules: Implications and applications”. In: *Chemical Reviews* 103.7 (2003), pp. 2475–2532. DOI: `10.1021/cr0104375`.

[174] M S Weiss, A Jabs, and R Hilgenfeld. “Peptide bonds revisited”. In: *Nature structural biology* 5.8 (1998), p. 676. DOI: `10.1038/1368`.

[175] Sandip De, S. Alireza Ghasemi, Alexander Willand, Luigi Genovese, Dilip Kanhere, and Stefan Goedecker. “The effect of ionization on the global minima of small and medium sized silicon and magnesium clusters”. In: *The Journal of Chemical Physics* 134.12 (2011). DOI: `http://dx.doi.org/10.1063/1.3569564`.

[176] Ideh Heidari, Sandip De, S. M. Ghazi, Stefan Goedecker, and D. G. Kanhere. “Growth and Structural Properties of MgN (N = 10–56) Clusters: Density Functional Theory Study”. In: *The Journal of Physical Chemistry A* 115.44 (2011), pp. 12307–12314. DOI: `10.1021/jp204442e`.

[177] Seyed Mohammad Ghazi, Sandip De, D G Kanhere, and Stefan Goedecker. “Density functional investigations on structural and electronic properties of anionic and neutral sodium clusters Na N (N = 40–147): comparison with the experimental photoelectron spectra”. In: *Journal of Physics: Condensed Matter* 23.40 (2011), p. 405303.

[178] Pascal Pochet, Luigi Genovese, Sandip De, Stefan Goedecker, Damien Caliste, S. Alireza Ghasemi, Kuo Bao, and Thierry Deutsch. “Low-energy boron fullerenes: Role of disorder and potential synthesis pathways”. In: *Phys. Rev. B* 83 (Feb. 2011), p. 081403. DOI: `10.1103/PhysRevB.83.081403`.

[179] Albert Ardevol, Gareth A. Tribello, Michele Ceriotti, and Michele Parrinello. “Probing the Unfolded Configurations of a β-Hairpin Using Sketch-Map”. In: *J. Chem. Theory Comput.* 11.3 (Mar. 2015), pp. 1086–1093. DOI: `10.1021/ct500950z`.

[180] Saša Baškarada and Andy Koronios. “A Critical Success Factor Framework for Information Quality Management”. In: *Information Systems Management* 31.4 (2014), pp. 276–295. DOI: `10.1080/10580530.2014.958023`.

[181] Jan Van den Broeck, Solveig Argeseanu Cunningham, Roger Eeckels, and Kobus Herbst. “Data Cleaning: Detecting, Diagnosing, and Editing Data Abnormalities”. In: *PLoS Med* 2.10 (Sept. 2005). DOI: `10.1371/journal.pmed.0020267`.

[182] Albert Gevorgyan, Mark G. Poolman, and David A. Fell. “Detection of stoichiometric inconsistencies in biomolecular models”. In: *Bioinformatics* 24.19 (2008), pp. 2245–2251. DOI: `10.1093/bioinformatics/btn425`.

[183] L. Ferretti, M. Colajanni, and M. Marchetti. “Distributed, Concurrent, and Independent Access to Encrypted Cloud Databases”. In: *IEEE Transactions on Parallel and Distributed Systems* 25.2 (Feb. 2014), pp. 437–446. DOI: `10.1109/TPDS.2013.154`.

[184] Sandip De, Alexander Willand, Maximilian Amsler, Pascal Pochet, Luigi Genovese, and Stefan Goedecker. “Energy Landscape of Fullerene Materials: A Comparison of Boron to Boron Nitride and Carbon”. In: *Physical Review Letters* 106.22 (June 2011), p. 225502. DOI: `10.1103/PhysRevLett.106.225502`.

[185] Peddy Vishweshwar, Jennifer A McMahon, Matthew L Peterson, Magali B Hickey, Tanise R Shattock, and Michael J Zaworotko. “Crystal engineering of pharmaceutical co-crystals from polymorphic active pharmaceutical ingredients”. In: *Chem. Commun.* 36 (2005), pp. 4601–4603.

[186] Naga K Duggirala, Miranda L Perry, Örn Almarsson, and Michael J Zaworotko. “Pharmaceutical cocrystals: along the path to improved medicines”. In: *Chem. Commun.* 52.4 (2016), pp. 640–655.

[187] Stephen R. Forrest. “The path to ubiquitous and low-cost organic electronic appliances on plastic.” In: *Nature* 428.6986 (2004), p. 911. DOI: `10.1038/nature02498`.

[188] Michele Muccini. “A bright future for organic field-effect transistors.” In: *Nature Mater.* 5.8 (2006), p. 605. DOI: `10.1038/nmat1699`.

[189] Dorothy Crowfoot Hodgkin, Jenny Pickworth, John H Robertson, KENNETH N Trueblood, RICHARD J Prosen, JOHN G White, et al. “The crystal structure of the hexacarboxylic acid derived from B12 and the molecular structure of the vitamin.” In: *Nature* 176 (1955), pp. 325–328.

[190] J. Bernstein. *Polymorphism in Molecular Crystals.* IUCr monographs on crystallography. Clarendon Press, 2002.

[191] Lian Yu. “Polymorphism in molecular solids: an extraordinary system of red, orange, and yellow crystals”. In: *Acc. Chem. Res.* 43.9 (2010), p. 1257.

[192] Angeles Pulido, Linjiang Chen, Tomasz Kaczorowski, Daniel Holden, Marc A. Little, Samantha Y. Chong, Benjamin J. Slater, David P. Mcmahon, Baltasar Bonillo, Chloe J. Stackhouse, Andrew Stephenson, Christopher M. Kane, Rob Clowes, Tom Hasell, Andrew I. Cooper, and Graeme M. Day. “Functional materials discovery using energy–structure–function maps”. In: *Nature* (Jan. 2017), pp. 1–34.

[193] Gautam R Desiraju and A Gavezzotti. “Crystal structures of polynuclear aromatic hydrocarbons. Classification, rationalization and prediction from molecular structure”. In: *Acta Cryst. B* 45.5 (1989), p. 473.

[194] Margaret C Etter, John C MacDonald, and Joel Bernstein. “Graph-set analysis of hydrogen-bond patterns in organic crystals”. In: *Acta Cryst. B* 46.2 (1990), p. 256.

[195] Edward F. Valeev, Veaceslav Coropceanu, Demetrio A. da Silva Filho, Seyhan Salman, and Jean-Luc Bredas. “Effect of Electronic Polarization on Charge-Transport Parameters in Molecular Organic Semiconductors”. In: *J. Am. Chem. Soc.* 128.30 (2006), pp. 9882–9886. DOI: `10.1021/ja061827h`.

[196] Michael Winkler and K. Houk. “Nitrogen-rich oligoacenes: candidates for n-channel organic semiconductors”. In: *J. Amer. Chem. Soc* 129.6 (2007), p. 1805.

[197] David H Case, Josh E Campbell, Peter J Bygrave, and Graeme M Day. “Convergence properties of crystal structure prediction by quasi-random sampling”. In: *J. Chem. Theory Comput.* (2015).

[198] Sarah L. Price, Maurice Leslie, Gareth W. A. Welch, Matthew Habgood, Louise S. Price, Panagiotis G. Karamertzanis, and Graeme M. Day. “Modelling organic crystal structures using distributed multipole and polarizability-based model intermolecular potentials”. In: *Phys. Chem. Chem. Phys.* 12.30 (2010), pp. 8478–8490.

[199] Donald E. Williams. “Improved intermolecular force field for molecules containing H, C, N, and O atoms, with application to nucleoside and peptide crystals”. In: *J. Comp. Chem.* 22.11 (2001), pp. 1154–1166. DOI: `10.1002/jcc.1074`.

[200] AJ Stone and M Alderton. “Distributed multipole analysis methods and applications”. In: *Molecular Physics* 100.1 (2002), pp. 221–233.

[201] Stefan Grimme. “Semiempirical GGA-type density functional constructed with a long-range dispersion correction”. In: *Journal of Computational Chemistry* 27.15 (2006), pp. 1787–1799. DOI: `10.1002/jcc.20495`.

[202] Paolo Giannozzi, Stefano Baroni, Nicola Bonini, Matteo Calandra, Roberto Car, Carlo Cavazzoni, Davide Ceresoli, Guido L Chiarotti, Matteo Cococcioni, Ismaila Dabo, Andrea Dal Corso, Stefano De Gironcoli, Stefano Fabris, Guido Fratesi, Ralph Gebauer, Uwe Gerstmann, Christos Gougoussis, Anton Kokalj, Michele Lazzeri, Layla Martin-Samos, Nicola Marzari, Francesco Mauri, Riccardo Mazzarello, Stefano Paolini, Alfredo Pasquarello, Lorenzo Paulatto, Carlo Sbraccia, Sandro Scandolo, Gabriele Sclauzero, Ari P Seitsonen, Alexander Smogunov, Paolo Umari, and Renata M Wentzcovitch. “QUANTUM ESPRESSO: A modular and open-source software project for quantum simulations of materials”. In: *Journal of Physics Condensed Matter* 21.39 (Sept. 2009), p. 395502. DOI: `10.1088/0953-8984/21/39/395502`.

[203] Leigh Loots and Leonard J. Barbour. “A simple and robust method for the identification of [small pi]-[small pi] packing motifs of aromatic compounds”. In: *CrystEngComm* 14 (2012), pp. 300–304. DOI: `10.1039/C1CE05763D`.

[204] Sandip De, Bastian Schaefer, Ali Sadeghi, Michael Sicher, D. G. Kanhere, and Stefan Goedecker. “Relation between the Dynamics of Glassy Clusters and Characteristic Features of their Energy Landscape”. In: *Physical Review Letters* 112.8 (Feb. 2014), p. 083401. DOI: `10.1103/PhysRevLett.112.083401`.

[205] K. T. Schütt, H. E. Sauceda, P. J. Kindermans, A. Tkatchenko, and K. R. Müller. “SchNet - A deep learning architecture for molecules and materials”. In: *Journal of Chemical Physics* 148.24 (June 2018), p. 241722. DOI: `10.1063/1.5019779`.

[206] Felix A. Faber, Luke Hutchison, Bing Huang, Justin Gilmer, Samuel S. Schoenholz, George E. Dahl, Oriol Vinyals, Steven Kearnes, Patrick F. Riley, and O. Anatole von Lilienfeld. “Prediction Errors of Molecular Machine Learning Models Lower than Hybrid DFT Error”. In: *J. Chem. Theory Comput.* 13.11 (Nov. 2017), pp. 5255–5264. DOI: `10.1021/acs.jctc.7b00577`.

[207] Eric R. Homer, Derek M. Hensley, Conrad W. Rosenbrock, Andrew H. Nguyen, and Gus L. W. Hart. “Machine-Learning Informed Representations for Grain Boundary Structures”. In: *Frontiers in Materials* 6 (July 2019), p. 168. DOI: `10.3389/fmats.2019.00168`.

[208] Félix Musil and Michele Ceriotti. “Machine learning at the atomic scale”. In: *Chimia* 73.12 (Dec. 2019), pp. 972–982. DOI: `10.2533/chimia.2019.972`.

[209] Chen Qu, Qi Yu, and Joel M. Bowman. “Permutationally Invariant Potential Energy Surfaces”. In: *Annual Review of Physical Chemistry* 69.1 (Apr. 2018), pp. 151–175. DOI: `10.1146/annurev-physchem-050317-021139`.

[210] Mehdi Jafary-Zadeh, Khoong Hong Khoo, Robert Laskowski, Paulo S. Branicio, and Alexander V. Shapeev. “Applying a machine learning interatomic potential to unravel the effects of local lattice distortion on the elastic properties of multi-principal element alloys”. In: *Journal of Alloys and Compounds* 803 (Sept. 2019), pp. 1054–1062. DOI: `10.1016/j.jallcom.2019.06.318`.

[211] I. I. Novoselov, A. V. Yanilkin, A. V. Shapeev, and E. V. Podryabinkin. “Moment tensor potentials as a promising tool to study diffusion processes”. In: *Computational Materials Science* 164 (June 2019), pp. 46–56. DOI: `10.1016/j.commatsci.2019.03.049`.

[212] Mitchell A. Wood and Aidan P. Thompson. “Extending the accuracy of the SNAP interatomic potential form”. In: *Journal of Chemical Physics* 148.24 (June 2018), p. 241721. DOI: `10.1063/1.5017641`.

[213] Atsuto Seko, Atsushi Togo, and Isao Tanaka. “Group-theoretical high-order rotational invariants for structural representations: Application to linearized machine learning interatomic potential”. In: *Physical Review B* 99.21 (June 2019), p. 214108. DOI: `10.1103/PhysRevB.99.214108`.

[214] Zhi Deng, Chi Chen, Xiang Guo Li, and Shyue Ping Ong. “An electrostatic spectral neighbor analysis potential for lithium nitride”. In: *npj Computational Materials* 5.1 (Dec. 2019), p. 75. DOI: `10.1038/s41524-019-0212-1`.

[215] Jörg Behler. “Atom-Centered Symmetry Functions for Constructing High-Dimensional Neural Network Potentials”. In: *J. Chem. Phys.* 134.7 (2011). DOI: `10.1063/1.3553717`.

[216] Nongnuch Artrith and Alexander Urban. “An implementation of artificial neural-network potentials for atomistic materials simulations: Performance for TiO2”. In: *Computational Materials Science* 114 (Mar. 2016), pp. 135–150. DOI: `10.1016/j.commatsci.2015.11.047`.

[217] J. S. Smith, O. Isayev, and A. E. Roitberg. “ANI-1: an extensible neural network potential with DFT accuracy at force field computational cost”. In: *Chemical Science* 8.4 (2017), pp. 3192–3203. DOI: `10.1039/C6SC05720A`.

[218] Kun Yao, John E. Herr, David W. Toth, Ryker McKintyre, and John Parkhill. “The TensorMol-0.1 model chemistry: A neural network augmented with long-range physics”. In: *Chemical Science* 9.8 (Feb. 2018), pp. 2261–2269. DOI: `10.1039/c7sc04934j`.

[219] Kyuhyun Lee, Dongsun Yoo, Wonseok Jeong, and Seungwu Han. “SIMPLE-NN: An efficient package for training and executing neural-network interatomic potentials”. In: *Computer Physics Communications* 242 (Sept. 2019), pp. 95–103. DOI: 10.1016/j.cpc.2019.04.014.

[220] John E. Herr, Kevin Koh, Kun Yao, and John Parkhill. “Compressing physics with an autoencoder: Creating an atomic species representation to improve machine learning models in the chemical sciences”. In: *The Journal of Chemical Physics* 151.8 (Aug. 2019), p. 084103. DOI: 10.1063/1.5108803.

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

[222] Kristof T. Schütt, Farhad Arbabzadah, Stefan Chmiela, Klaus R. Müller, and Alexandre Tkatchenko. “Quantum-chemical insights from deep tensor neural networks”. In: *Nature Communications* 8 (Jan. 2017), pp. 6–13. DOI: 10.1038/ncomms13890.

[223] Nicholas Lubbers, Justin S. Smith, and Kipton Barros. “Hierarchical modeling of molecular energies using a deep neural network”. In: *Journal of Chemical Physics* 148.24 (June 2018), p. 241715. DOI: 10.1063/1.5011181.

[224] Oliver T. Unke and Markus Meuwly. “PhysNet: A Neural Network for Predicting Energies, Forces, Dipole Moments, and Partial Charges”. In: *Journal of Chemical Theory and Computation* 15.6 (Feb. 2019), pp. 3678–3693. DOI: 10.1021/acs.jctc.9b00181.

[225] Suresh Kondati Natarajan and Jörg Behler. “Neural network molecular dynamics simulations of solid–liquid interfaces: water at low-index copper surfaces”. In: *Phys. Chem. Chem. Phys.* 18.41 (Oct. 2016), pp. 28704–28725. DOI: 10.1039/C6CP05711J.

[226] Michael Gastegger, Jörg Behler, and Philipp Marquetand. “Machine Learning Molecular Dynamics for the Simulation of Infrared Spectra”. In: *Chem. Sci.* 8.10 (Sept. 2017), pp. 6924–6935. DOI: 10.1039/C7SC02267K.

[227] Venkat Kapil, Jörg Behler, and Michele Ceriotti. “High order path integrals made easy”. In: *Journal of Chemical Physics* 145.23 (Dec. 2016), p. 234103. DOI: 10.1063/1.4971438.

[228] Si Da Huang, Cheng Shang, Pei Lin Kang, and Zhi Pan Liu. “Atomic structure of boron resolved using machine learning and global sampling”. In: *Chemical Science* 9.46 (Nov. 2018), pp. 8644–8655. DOI: 10.1039/c8sc03427c.

[229] Marco Eckhoff and Jörg Behler. “From Molecular Fragments to the Bulk: Development of a Neural Network Potential for MOF-5”. In: *Journal of Chemical Theory and Computation* 15.6 (June 2019), pp. 3793–3809. DOI: 10.1021/acs.jctc.8b01288.

[230] Katja Hansen, Grégoire Montavon, Franziska Biegler, Siamac Fazli, Matthias Rupp, Matthias Scheffler, O. Anatole Von Lilienfeld, Alexandre Tkatchenko, and Klaus Robert Müller. “Assessment and validation of machine learning methods for predicting molecular atomization energies”. In: *Journal of Chemical Theory and Computation* 9.8 (Aug. 2013), pp. 3404–3419. DOI: `10.1021/ct400195d`.

[231] Volker L. Deringer and Gábor Csányi. “Machine Learning Based Interatomic Potential for Amorphous Carbon”. In: *Phys. Rev. B* 95.9 (Mar. 2017), p. 094203. DOI: `10.1103/PhysRevB.95.094203`.

[232] Volker L. Deringer, Noam Bernstein, Albert P. Bartók, Matthew J. Cliffe, Rachel N. Kerber, Lauren E. Marbella, Clare P. Grey, Stephen R. Elliott, and Gábor Csányi. “Realistic Atomistic Structure of Amorphous Silicon from Machine-Learning-Driven Molecular Dynamics”. In: *Journal of Physical Chemistry Letters* 9.11 (June 2018), pp. 2879–2885. DOI: `10.1021/acs.jpclett.8b00902`.

[233] Albert P. Bartók, James Kermode, Noam Bernstein, and Gábor Csányi. “Machine Learning a General-Purpose Interatomic Potential for Silicon”. In: *Physical Review X* 8.4 (May 2018).

[234] Claudio Zeni, Kevin Rossi, Aldo Glielmo, Ádám Fekete, Nicola Gaston, Francesca Baletto, and Alessandro De Vita. “Building machine learning force fields for nanoclusters”. In: *Journal of Chemical Physics* 148.24 (Feb. 2018), p. 241739.

[235] Stefan Chmiela, Alexandre Tkatchenko, Huziel E. Sauceda, Igor Poltavsky, Kristof T. Schütt, and Klaus Robert Müller. “Machine learning of accurate energy-conserving molecular force fields”. In: *Science Advances* 3.5 (May 2017), e1603015. DOI: `10.1126/sciadv.1603015`.

[236] Max Veit, Sandeep Kumar Jain, Satyanarayana Bonakala, Indranil Rudra, Detlef Hohl, and Gábor Csányi. “Equation of State of Fluid Methane from First Principles with Machine Learning Potentials”. In: *Journal of Chemical Theory and Computation* 15.4 (Oct. 2019), pp. 2574–2586. DOI: `10.1021/acs.jctc.8b01242`.

[237] Nathaniel Raimbault, Andrea Grisafi, Michele Ceriotti, and Mariana Rossi. “Using Gaussian Process Regression to Simulate the Vibrational Raman Spectra of Molecular Crystals”. In: *ArXiv e-prints* (June 2019).

[238] Raghunathan Ramakrishnan, Mia Hartmann, Enrico Tapavicza, and O. Anatole Von Lilienfeld. “Electronic spectra from TDDFT and machine learning in chemical space”. In: *Journal of Chemical Physics* 143.8 (Aug. 2015), p. 084111. DOI: `10.1063/1.4928757`.

[239] Anders S. Christensen, Felix A. Faber, and O. Anatole von Lilienfeld. “Operators in quantum machine learning: Response properties in chemical space”. In: *J. Chem. Phys.* 150.6 (2019), p. 064105.

[240] A. P. Bartók and G. Csányi. In: *International Journal of Quantum Chemistry* 115 (2015), pp. 1051–1057.

[241] Michele Ceriotti, Michael J. Willatt, and Gábor Csányi. "Machine Learning of Atomic-Scale Properties Based on Physical Principles". In: *Handbook of Materials Modeling*. Ed. by Wanda Andreoni and Sidney Yip. Cham: Springer International Publishing, 2018, pp. 1–27. DOI: 10.1007/978-3-319-42913-7_68-1.

[242] Joaquin Quiñonero-Candela and Carl Edward Rasmussen. "A Unifying View of Sparse Approximate Gaussian Process Regression". In: *Journal of Machine Learning Research* 6.Dec (2005), pp. 1939–1959.

[243] Haitao Liu, Yew-Soon Ong, Xiaobo Shen, and Jianfei Cai. "When Gaussian Process Meets Big Data: A Review of Scalable GPs". In: *ArXiv e-prints* (July 2018).

[244] Matthias Seeger, Christopher K I Williams, and Neil D Lawrence. "Fast Forward Selection to Speed Up Sparse Gaussian Process Regression". In: *Artificial Intelligence and Statistics 9*. 2003.

[245] Félix Musil, Michael J. Willatt, Mikhail A. Langovoy, and Michele Ceriotti. "Fast and Accurate Uncertainty Estimation in Chemical Machine Learning". In: *Journal of Chemical Theory and Computation* 15.2 (Feb. 2019), pp. 906–915. DOI: 10.1021/acs.jctc.8b00959.

[246] Alex J. Smola and Peter Bartlett. "Sparse Greedy Gaussian Process Regression". In: *Advances in Neural Information Processing Systems 13* 13 (2001), pp. 619–625.

[247] S. Sathiya Keerthi and Wei Chu. "A matching pursuit approach to sparse Gaussian process regression". In: *Advances in Neural Information Processing Systems*. 2005, pp. 643–650.

[248] Jens Schreiter, Duy Nguyen-Tuong, and Marc Toussaint. "Efficient sparsification for Gaussian process regression". In: *Neurocomputing* 192 (June 2016), pp. 29–37. DOI: 10.1016/j.neucom.2016.02.032.

[249] Payam Refaeilzadeh, Lei Tang, and Huan Liu. "Cross-Validation". In: *Encyclopedia of Database Systems*. Ed. by LING LIU and M. TAMER ÖZSU. Boston, MA: Springer US, 2009, pp. 532–538. DOI: 10.1007/978-0-387-39940-9_565.

[250] O. Anatole von Lilienfeld. "Quantum Machine Learning in Chemical Compound Space". In: *Angewandte Chemie - International Edition* 57.16 (Apr. 2018), pp. 4164–4169. DOI: 10.1002/anie.201709686.

[251] AC de Dios, J. G. Pearson, and E. Oldfield. "Secondary and tertiary structural effects on protein NMR chemical shifts: an ab initio approach". In: *Science* 260.5113 (June 1993), pp. 1491–1496. DOI: 10.1126/science.8502992.

[252] Julio C. Facelli and David M. Grant. "Determination of molecular symmetry in crystalline naphthalene using solid-state NMR". In: *Nature* 365.6444 (Sept. 1993), pp. 325–327. DOI: 10.1038/365325a0.

[253] Daniel Sebastiani and Michele Parrinello. "A New ab-Initio Approach for NMR Chemical Shifts in Periodic Systems". In: *The Journal of Physical Chemistry A* 105.10 (Mar. 2001), pp. 1951–1958. DOI: 10.1021/jp002807j.

[254] Chris J. Pickard and Francesco Mauri. “All-electron magnetic response with pseudopotentials: NMR chemical shifts”. In: *Physical Review B* 63.24 (May 2001), pp. 2451011–2451013. DOI: 10.1103/PhysRevB.63.245101.

[255] Jonathan R. Yates, Chris J. Pickard, and Francesco Mauri. “Calculation of NMR chemical shifts for extended systems using ultrasoft pseudopotentials”. In: *Physical Review B* 76.2 (July 2007), p. 024401. DOI: 10.1103/PhysRevB.76.024401.

[256] P. E. Blöchl. “Projector augmented-wave method”. In: *Physical Review B* 50.24 (Dec. 1994), pp. 17953–17979. DOI: 10.1103/PhysRevB.50.17953.

[257] C. Ochsenfeld, S. P. Brown, I. Schnell, J. Gauss, and H. W. Spiess. “Structure assignment in the solid state by the coupling of quantum chemical calculations with NMR experiments: a columnar hexabenzocoronene derivative”. In: *Journal of the American Chemical Society* 123.11 (Mar. 2001), pp. 2597–2606. DOI: 10.1021/ja0021823.

[258] James K. Harper and David M. Grant. “Enhancing Crystal-Structure Prediction with NMR Tensor Data”. In: *Crystal Growth & Design* 6.10 (Oct. 2006), pp. 2315–2321. DOI: 10.1021/cg060244g.

[259] Robin K. Harris. “NMR crystallography: the use of chemical shifts”. In: *Solid State Sciences* 6.10 (Oct. 2004), pp. 1025–1037. DOI: 10.1016/j.solidstatesciences.2004.03.040.

[260] Abdullah Othman, John S. O. Evans, Ivana Radosavljevic Evans, Robin K. Harris, and Paul Hodgkinson. “Structural Study of Polymorphs and Solvates of Finasteride”. In: *Journal of Pharmaceutical Sciences* 96.5 (May 2007), pp. 1380–1397. DOI: 10.1002/jps.20940.

[261] Elodie Salager, Robin S. Stein, Chris J. Pickard, Bénédicte Elena, and Lyndon Emsley. “Powder NMR crystallography of thymol”. In: *Physical Chemistry Chemical Physics* 11.15 (Mar. 2009), pp. 2610–2621. DOI: 10.1039/B821018G.

[262] Elodie Salager, Graeme M. Day, Robin S. Stein, Chris J. Pickard, Bénédicte Elena, and Lyndon Emsley. “Powder crystallography by combined crystal structure prediction and high-resolution 1 H solid-state NMR spectroscopy”. In: *Journal of the American Chemical Society* 132.8 (Mar. 2010), pp. 2564–2566. DOI: 10.1021/ja909449k.

[263] Amy L. Webber, Lyndon Emsley, Rosa M. Claramunt, and Steven P. Brown. “NMR Crystallography of Campho[2,3-c]pyrazole (Z’ = 6): Combining High-Resolution 1H-13C Solid-State MAS NMR Spectroscopy and GIPAW Chemical-Shift Calculations”. In: *The Journal of Physical Chemistry A* 114.38 (Sept. 2010), pp. 10435–10442. DOI: 10.1021/jp104901j.

[264] Dmytro Dudenko, Adam Kiersnowski, Jie Shu, Wojciech Pisula, Daniel Sebastiani, Hans Wolfgang Spiess, and Michael Ryan Hansen. “A Strategy for Revealing the Packing in Semicrystalline π-Conjugated Polymers: Crystal Structure of Bulk Poly-3-hexylthiophene (P3HT)”. In: *Angewandte Chemie International Edition* 51.44 (2012), pp. 11068–11072. DOI: 10.1002/anie.201205075.

[265] Maria Baias, Cory M. Widdifield, Jean-Nicolas Dumez, Hugh P. G. Thompson, Timothy G. Cooper, Elodie Salager, Sirena Bassil, Robin S. Stein, Anne Lesage, Graeme M. Day, and Lyndon Emsley. "Powder crystallography of pharmaceutical materials by combined crystal structure prediction and solid-state 1H NMR spectroscopy". In: *Physical Chemistry Chemical Physics* 15.21 (May 2013), pp. 8069–8080. DOI: `10.1039/C3CP41095A`.

[266] Tomasz Pawlak, Magdalena Jaworska, and Marek J. Potrzebowski. "NMR crystallography of α-poly(L-lactide)". In: *Physical Chemistry Chemical Physics* 15.9 (Feb. 2013), pp. 3137–3145. DOI: `10.1039/C2CP43174B`.

[267] Sérgio M. Santos, João Rocha, and Luís Mafra. "NMR Crystallography: Toward Chemical Shift-Driven Crystal Structure Determination of the β-Lactam Antibiotic Amoxicillin Trihydrate". In: *Crystal Growth & Design* 13.6 (June 2013), pp. 2390–2395. DOI: `10.1021/cg4002785`.

[268] David Lüdeker and Gunther Brunklaus. "NMR crystallography of ezetimibe co-crystals". In: *Solid State Nuclear Magnetic Resonance* 65 (Feb. 2015), pp. 29–40. DOI: `10.1016/j.ssnmr.2014.11.002`.

[269] Piotr Paluch, Tomasz Pawlak, Marcin Oszajca, Wieslaw Lasocha, and Marek J. Potrzebowski. "Fine refinement of solid state structure of racemic form of phospho-tyrosine employing NMR Crystallography approach". In: *Solid State Nuclear Magnetic Resonance* 65 (Feb. 2015), pp. 2–11. DOI: `10.1016/j.ssnmr.2014.08.002`.

[270] Abigail E. Watts, Keisuke Maruyoshi, Colan E. Hughes, Steven P. Brown, and Kenneth D. M. Harris. "Combining the Advantages of Powder X-ray Diffraction and NMR Crystallography in Structure Determination of the Pharmaceutical Material Cimetidine Hydrochloride". In: *Crystal Growth & Design* 16.4 (Apr. 2016), pp. 1798–1804. DOI: `10.1021/acs.cgd.6b00016`.

[271] Cory M. Widdifield, Harry Robson, and Paul Hodgkinson. "Furosemide's one little hydrogen atom: NMR crystallography structure verification of powdered molecular organics". In: *Chemical Communications* 52.40 (May 2016), pp. 6685–6688. DOI: `10.1039/C6CC02171A`.

[272] G. Mali. "Ab initio crystal structure prediction of magnesium (poly)sulfides and calculation of their NMR parameters". In: *Acta Crystallographica Section C: Structural Chemistry* 73.3 (Mar. 2017), pp. 229–233. DOI: `10.1107/S2053229617000687`.

[273] Robin K. Harris, Siân A. Joyce, Chris J. Pickard, Sylvian Cadars, and Lyndon Emsley. "Assigning carbon-13 NMR spectra to crystal structures by the INADEQUATE pulse sequence and first principles computation: a case study of two forms of testosterone". In: *Physical Chemistry Chemical Physics* 8.1 (Dec. 2006), pp. 137–143. DOI: `10.1039/B513392K`.

[274] Nicolas Mifsud, Bénédicte Elena, Chris J. Pickard, Anne Lesage, and Lyndon Emsley. “Assigning powders to crystal structures by high-resolution 1H–1H double quantum and 1H–13C J-INEPT solid-state NMR spectroscopy and first principles computation. A case study of penicillin G”. In: *Physical Chemistry Chemical Physics* 8.29 (July 2006), pp. 3418–3422. DOI: `10.1039/B605227D`.

[275] Elizabeth M. Heider, James K. Harper, and David M. Grant. “Structural characterization of an anhydrous polymorph of paclitaxel by solid-state NMR”. In: *Physical Chemistry Chemical Physics* 9.46 (Nov. 2007), pp. 6083–6097. DOI: `10.1039/B711027H`.

[276] Maria Baias, Jean-Nicolas Dumez, Per H. Svensson, Staffan Schantz, Graeme M. Day, and Lyndon Emsley. “De Novo Determination of the Crystal Structure of a Large Drug Molecule by Crystal Structure Prediction-Based Powder NMR Crystallography”. In: *Journal of the American Chemical Society* 135.46 (Nov. 2013), pp. 17501–17507. DOI: `10.1021/ja4088874`.

[277] José A. Fernandes, Mariana Sardo, Luís Mafra, Duane Choquesillo-Lazarte, and Norberto Masciocchi. “X-ray and NMR Crystallography Studies of Novel Theophylline Cocrystals Prepared by Liquid Assisted Grinding”. In: *Crystal Growth & Design* 15.8 (Aug. 2015), pp. 3674–3683. DOI: `10.1021/acs.cgd.5b00279`.

[278] Julien Leclaire, Guillaume Poisson, Fabio Ziarelli, Gerard Pepe, Frédéric Fotiadu, Federico M. Paruzzo, Aaron J. Rossini, Jean-Nicolas Dumez, Bénédicte Elena-Herrmann, and Lyndon Emsley. “Structure elucidation of a complex CO2-based organic framework material by NMR crystallography”. In: *Chemical Science* 7.7 (June 2016), pp. 4379–4390. DOI: `10.1039/C5SC03810C`.

[279] Marcin Selent, Jonas Nyman, Juho Roukala, Marek Ilczyszyn, Raija Oilunkaniemi, Peter J. Bygrave, Risto Laitinen, Jukka Jokisaari, Graeme M. Day, and Perttu Lantto. “Clathrate Structure Determination by Combining Crystal Structure Prediction with Computational and Experimental 129Xe NMR Spectroscopy”. In: *Chemistry – A European Journal* 23.22 (Apr. 2017), pp. 5258–5269. DOI: `10.1002/chem.201604797`.

[280] Cory M. Widdifield, Sten O. Nilsson Lill, Anders Broo, Maria Lindkvist, Anna Pettersen, Anna Svensk Ankarberg, Peter Aldred, Staffan Schantz, and Lyndon Emsley. “Does Z’ equal 1 or 2? Enhanced powder NMR crystallography verification of a disordered room temperature crystal structure of a p38 inhibitor for chronic obstructive pulmonary disease”. In: *Physical Chemistry Chemical Physics* 19.25 (June 2017), pp. 16650–16661. DOI: `10.1039/C7CP02349A`.

[281] Sten O. Nilsson Lill, Cory M. Widdifield, Anna Pettersen, Anna Svensk Ankarberg, Maria Lindkvist, Peter Aldred, Sandra Gracin, Norman Shankland, Kenneth Shankland, Staffan Schantz, and Lyndon Emsley. “Elucidating an Amorphous Form Stabilization Mechanism for Tenapanor Hydrochloride: Crystal Structure Analysis Using X-ray Diffraction, NMR Crystallography, and Molecular Modeling”. In: *Molecular Pharmaceutics* 15.4 (Apr. 2018), pp. 1476–1487. DOI: `10.1021/acs.molpharmaceut.7b01047`.

[282] Yang Shen and Ad Bax. “Protein backbone chemical shifts predicted from searching a database for torsion angle and sequence homology”. In: *Journal of Biomolecular NMR* 38.4 (Aug. 2007), pp. 289–302. DOI: `10.1007/s10858-007-9166-6`.

[283] Stephen Neal, Alex M. Nip, Haiyan Zhang, and David S. Wishart. “Rapid and accurate calculation of protein 1H, 13C and 15N chemical shifts”. In: *Journal of Biomolecular NMR* 26.3 (July 2003), pp. 215–240. DOI: `10.1023/A:1023812930288`.

[284] David S. Wishart, M. Scott Watson, Robert F. Boyko, and Brian D. Sykes. “Automated 1H and 13C chemical shift prediction using the BioMagResBank”. In: *Journal of Biomolecular NMR* 10.4 (Dec. 1997), pp. 329–336. DOI: `10.1023/A:1018373822088`.

[285] Mitsuo Iwadate, Tetsuo Asakura, and Michael P. Williamson. “Cα and Cβ Carbon-13 Chemical Shifts in Proteins From an Empirical Database”. In: *Journal of Biomolecular NMR* 13.3 (Mar. 1999), pp. 199–211. DOI: `10.1023/A:1008376710086`.

[286] Xiao-Ping Xu and David A. Case. “Automated prediction of 15N, 13Cα, 13Cβ and 13C’ chemical shifts in proteins using a density functional database”. In: *Journal of Biomolecular NMR* 21.4 (Dec. 2001), pp. 321–333. DOI: `10.1023/A:1013324104681`.

[287] Seongho Moon and David A. Case. “A new model for chemical shifts of amide hydrogens in proteins”. In: *Journal of Biomolecular NMR* 38.2 (Apr. 2007), p. 139. DOI: `10.1007/s10858-007-9156-8`.

[288] Jorge A. Vila, Yelena A. Arnautova, Osvaldo A. Martin, and Harold A. Scheraga. “Quantum-mechanics-derived 13Cα chemical shift server (CheShift) for protein structure validation”. In: *Proceedings of the National Academy of Sciences* 106.40 (Oct. 2009), pp. 16972–16977. DOI: `10.1073/pnas.0908833106`.

[289] Kai J. Kohlhoff, Paul Robustelli, Andrea Cavalli, Xavier Salvatella, and Michele Vendruscolo. “Fast and Accurate Predictions of Protein NMR Chemical Shifts from Interatomic Distances”. In: *Journal of the American Chemical Society* 131.39 (Oct. 2009), pp. 13894–13895. DOI: `10.1021/ja903772t`.

[290] Jens Meiler. “PROSHIFT: Protein chemical shift prediction using artificial neural networks”. In: *Journal of Biomolecular NMR* 26.1 (May 2003), pp. 25–37. DOI: `10.1023/A:1023060720156`.

[291] Yang Shen and Ad Bax. “SPARTA+: a modest improvement in empirical NMR chemical shift prediction by means of an artificial neural network”. In: *Journal of Biomolecular NMR* 48.1 (Sept. 2010), pp. 13–22. DOI: `10.1007/s10858-010-9433-9`.

[292] K. A. Blinov, Y. D. Smurnyy, M. E. Elyashberg, T. S. Churanova, M. Kvasha, C. Steinbeck, B. A. Lefebvre, and A. J. Williams. “Performance Validation of Neural Network Based 13C NMR Prediction Using a Publicly Available Data Source”. In: *Journal of Chemical Information and Modeling* 48.3 (Mar. 2008), pp. 550–555. DOI: `10.1021/ci700363r`.

[293] Yegor D. Smurnyy, Kirill A. Blinov, Tatiana S. Churanova, Mikhail E. Elyashberg, and Antony J. Williams. “Toward More Reliable 13C and 1H Chemical Shift Prediction: A Systematic Comparison of Neural-Network and Least-Squares Regression Based Approaches”. In: *Journal of Chemical Information and Modeling* 48.1 (Jan. 2008), pp. 128–134. DOI: `10.1021/ci700256n`.

[294] João Aires-de-Sousa, Markus C. Hemmer, and Johann Gasteiger. “Prediction of 1H NMR Chemical Shifts Using Neural Networks”. In: *Analytical Chemistry* 74.1 (Jan. 2002), pp. 80–90. DOI: `10.1021/ac010737m`.

[295] Stefan Kuhn, Björn Egert, Steffen Neumann, and Christoph Steinbeck. “Building blocks for automated elucidation of metabolites: Machine learning methods for NMR prediction”. In: *BMC Bioinformatics* 9.1 (Sept. 2008), p. 400. DOI: `10.1186/1471-2105-9-400`.

[296] Jérôme Cuny, Yu Xie, Chris J. Pickard, and Ali A. Hassanali. “Ab Initio Quality NMR Parameters in Solid-State Materials Using a High-Dimensional Neural-Network Representation”. In: *Journal of Chemical Theory and Computation* 12.2 (Feb. 2016), pp. 765–773. DOI: `10.1021/acs.jctc.5b01006`.

[297] P Giannozzi, O Andreussi, T Brumme, O Bunau, M Buongiorno Nardelli, M Calandra, R Car, C Cavazzoni, D Ceresoli, M Cococcioni, N Colonna, I Carnimeo, A Dal Corso, S. De Gironcoli, P Delugas, R. A. Distasio, A Ferretti, A Floris, G Fratesi, G Fugallo, R Gebauer, U Gerstmann, F Giustino, T Gorni, J Jia, M Kawamura, H. Y. Ko, A Kokalj, E. Kücükbenli, M Lazzeri, M Marsili, N Marzari, F Mauri, N. L. Nguyen, H. V. Nguyen, A. Otero-De-La-Roza, L Paulatto, S Poncé, D Rocca, R Sabatini, B Santra, M Schlipf, A P Seitsonen, A Smogunov, I Timrov, T Thonhauser, P Umari, N Vast, X Wu, and S Baroni. “Advanced capabilities for materials modelling with Quantum ESPRESSO”. In: *Journal of Physics Condensed Matter* 29.46 (Nov. 2017), p. 465901. DOI: `10.1088/1361-648X/aa8f79`.

[298] Kurt Lejaeghere, Gustav Bihlmayer, Torbjörn Björkman, Peter Blaha, Stefan Blügel, Volker Blum, Damien Caliste, Ivano E Castelli, Stewart J Clark, Andrea Dal Corso, Stefano De Gironcoli, Thierry Deutsch, John Kay Dewhurst, Igor Di Marco, Claudia Draxl, Marcin Dułak, Olle Eriksson, José A Flores-Livas, Kevin F Garrity, Luigi Genovese, Paolo Giannozzi, Matteo Giantomassi, Stefan Goedecker, Xavier Gonze, Oscar Grånäs, E. K.U. Gross, Andris Gulans, François Gygi, D R Hamann, Phil J Hasnip, N. A.W. Holzwarth, Diana Iuşan, Dominik B Jochym, François Jollet, Daniel Jones, Georg Kresse, Klaus Koepernik, Emine Küçükbenli, Yaroslav O Kvashnin, Inka L.M. Locht, Sven Lubeck, Martijn Marsman, Nicola Marzari, Ulrike Nitzsche, Lars Nordström, Taisuke Ozaki, Lorenzo Paulatto, Chris J Pickard, Ward Poelmans, Matt I.J. Probert, Keith Refson, Manuel Richter, Gian Marco Rignanese, Santanu Saha, Matthias Scheffler, Martin Schlipf, Karlheinz Schwarz, Sangeeta Sharma, Francesca Tavazza, Patrik Thunström, Alexandre Tkatchenko, Marc Torrent, David Vanderbilt, Michiel J. Van Setten, Veronique Van Speybroeck, John M Wills, Jonathan R Yates, Guo Xu Zhang, and Stefaan Cottenier. “Reproducibility in density functional theory calculations of solids”. In: *Science* 351.6280 (Mar. 2016), aad3000. DOI: `10.1126/science.aad3000`.

[299] Stefan Grimme. “Semiempirical GGA-type density functional constructed with a long-range dispersion correction”. In: *Journal of Computational Chemistry* 27.15 (Nov. 2006), pp. 1787–1799. DOI: 10.1002/jcc.20495.

[300] H Monkhorst and J Pack. “Special points for Brillouin zone integrations”. In: *Physical Review B* 13.12 (June 1976), pp. 5188–5192. DOI: 10.1103/PhysRevB.13.5188.

[301] Felix Musil, Sandip De, and Michele Ceriotti. *Glosim2*. original-date: 2017-06-25T13:35:25Z. Mar. 2020.

[302] G. M. Day, W. D.S. Motherwell, and W. Jones. “A strategy for predicting the crystal structures of flexible molecules: The polymorphism of phenobarbital”. In: *Physical Chemistry Chemical Physics* 9.14 (Mar. 2007), pp. 1693–1704. DOI: 10.1039/b612190j.

[303] Stewart J. Clark, Matthew D. Segall, Chris J. Pickard, Phil J. Hasnip, Matt I. J. Probert, Keith Refson, and title = "First principles methods using CASTEP Mike C. Payne". In: *Zeitschrift für Kristallographie - Crystalline Materials* 220.5-6 (2005), pp. 567–570. DOI: 10.1524/zkri.220.5.567.65075.

[304] Joshua D. Hartman, Ryan A. Kudla, Graeme M. Day, Leonard J. Mueller, and Gregory J. O. Beran. “Benchmark fragment-based 1H, 13C, 15N and 17O chemical shift predictions in molecular crystals”. In: *Physical Chemistry Chemical Physics* 18.31 (Aug. 2016), pp. 21686–21709. DOI: 10.1039/C6CP01831A.

[305] Joshua D. Hartman, Ryan A. Kudla, Graeme M. Day, Leonard J. Mueller, and Gregory J. O. Beran. “Benchmark fragment-based 1H, 13C, 15N and 17O chemical shift predictions in molecular crystals”. In: *Phys. Chem. Chem. Phys.* 18 (31 2016), pp. 21686–21709. DOI: 10.1039/C6CP01831A.

[306] Stewart J. Clark, Matthew D. Segall, Chris J. Pickard, Phil J. Hasnip, Matt I. J. Probert, Keith Refson, and Mike C. Payne. “First principles methods using CASTEP”. In: *Zeitschrift für Kristallographie - Crystalline Materials* 220.5/6 (Jan. 2005). DOI: 10.1524/zkri.220.5.567.65075.

[307] Mariana Sardo, Sérgio M. Santos, Artem A. Babaryk, Concepción López, Ibon Alkorta, José Elguero, Rosa M. Claramunt, and Luís Mafra. “Diazole-based powdered cocrystal featuring a helical hydrogen-bonded network: Structure determination from PXRD, solid-state NMR and computer modeling”. In: *Solid State Nuclear Magnetic Resonance* 65 (2015). NMR Crystallography, pp. 49–63. DOI: https://doi.org/10.1016/j.ssnmr.2014.12.005.

[308] Elisa Carignani, Silvia Borsacchi, Jonathan P. Bradley, Steven P. Brown, and Marco Geppi. “Strong Intermolecular Ring Current Influence on 1H Chemical Shifts in Two Crystalline Forms of Naproxen: a Combined Solid-State NMR and DFT Study”. In: *The Journal of Physical Chemistry C* 117.34 (Aug. 2013), pp. 17731–17740. DOI: 10.1021/jp4044946.

[309] Anne-Christine Uldry, John M. Griffin, Jonathan R. Yates, Marta Perez-Torralba, M. Dolores Santa Maria, Amy L. Webber, Maximus L. L. Beaumont, Ago Samoson, Rosa Maria Claramunt, Chris J. Pickard, and Steven P. Brown. "Quantifying Weak Hydrogen Bonding in Uracil and 4-Cyano-4-ethynylbiphenyl: A Combined Computational and Experimental Investigation of NMR Chemical Shifts in the Solid State". In: *Journal of the American Chemical Society* 130.3 (Jan. 2008), pp. 945–954. DOI: `10.1021/ja075892i`.

[310] Christopher C. Arico-Muendel, Heather Blanchette, Dennis R. Benjamin, Teresa M. Caiazzo, Paolo A. Centrella, Jennifer DeLorey, Elisabeth G. Doyle, Steven R. Johnson, Matthew T. Labenski, Barry A. Morgan, Gary O'Donovan, Amy A. Sarjeant, Steven Skinner, Charles D. Thompson, Sarah T. Griffin, William Westlin, and Kerry F. White. "Orally Active Fumagillin Analogues: Transformations of a Reactive Warhead in the Gastric Environment". In: *ACS Medicinal Chemistry Letters* 4.4 (Apr. 2013), pp. 381–386. DOI: `10.1021/ml3003633`.

[311] Hai T. Dao, Chao Li, Quentin Michaudel, Brad D. Maxwell, and Phil S. Baran. "Hydromethylation of Unactivated Olefins". In: *Journal of the American Chemical Society* 137.25 (July 2015), pp. 8046–8049. DOI: `10.1021/jacs.5b05144`.

[312] Takeharu Haino, Youko Matsumoto, and Yoshimasa Fukazawa. "Supramolecular Nano Networks Formed by Molecular-Recognition-Directed Self-Assembly of Ditopic Calix[5]arene and Dumbbell [60]Fullerene". In: *Journal of the American Chemical Society* 127.25 (June 2005), pp. 8936–8937. DOI: `10.1021/ja0524088`.

[313] J.W. Bats. *CCDC 802297: Experimental Crystal Structure Determination.* 2014.

[314] Guo-Bao Huang, Wei-Er Liu, Arto Valkonen, Huan Yao, Kari Rissanen, and Wei Jiang. "Selective recognition of aromatic hydrocarbons by endo-functionalized molecular tubes via C/N-H$\cdots\pi$ interactions". In: *Chinese Chemical Letters* 29.1 (Jan. 2018), pp. 91–94. DOI: `10.1016/j.cclet.2017.07.005`.

[315] M. John Plater, William T.A. Harrison, Laura M. Machado de los Toyos, and Lewis Hendry. "The Consistent Hexameric Paddle-Wheel Crystallisation Motif of a Family of 2,4-Bis(n-Alkylamino)Nitrobenzenes: Alkyl = Pentyl, Hexyl, Heptyl and Octyl". In: *Journal of Chemical Research* 41.4 (Apr. 2017), pp. 235–238. DOI: `10.3184/174751917X14902201357356`.

[316] Anthony M. Reilly and Alexandre Tkatchenko. "Role of Dispersion Interactions in the Polymorphism and Entropic Stabilization of the Aspirin Crystal". In: *Phys. Rev. Lett.* 113.5 (July 2014), p. 055701. DOI: `10.1103/PhysRevLett.113.055701`.

[317] Sarah (Sally) L. Price. "Quantifying intermolecular interactions and their use in computational crystal structure prediction". In: *CrystEngComm* 6 (2004), pp. 344–353. DOI: `10.1039/B406598K`.

[318] Farren Curtis, Xiaopeng Wang, and Noa Marom. "Effect of packing motifs on the energy ranking and electronic properties of putative crystal structures of tricyano-1,4-dithiino[c]-isothiazole". In: *Acta Crystallographica Section B: Structural Science, Crystal Engineering and Materials* 72.4 (Aug. 2016), pp. 562–570. DOI: `10.1107/S2052520616009227`.

[319] Jonas Nyman and Graeme M. Day. "Static and lattice vibrational energy differences between polymorphs". In: *CrystEngComm* 17 (2015), pp. 5154–5165. DOI: 10.1039/C5CE00045A.

[320] 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.* 117.11 (Sept. 2016), p. 115702. DOI: 10.1103/PhysRevLett.117.115702.

[321] Carolyn Pratt Brock and Jack D. Dunitz. "Towards a Grammar of Crystal Packing". In: *Chem. Mater.* 6.8 (1994), pp. 1118–1127.

[322] Donald E Williams. "Improved intermolecular force field for crystalline hydrocarbons containing four-or three-coordinated carbon". In: *J. Mol. Struct.* 485-486 (1999), pp. 321–347.

[323] Donald E. Williams. "Improved intermolecular force field for crystalline oxohydrocarbons including OHO hydrogen bonding". In: *J. Comp. Chem.* 22.1 (2001), pp. 1–20. DOI: 10.1002/1096-987X(20010115)22:1<1::AID-JCC2>3.0.CO;2-6.

[324] Edward O Pyzer-Knapp, Hugh PG Thompson, and Graeme M Day. "An optimized intermolecular force field for hydrogen-bonded organic molecular crystals using atomic multipole electrostatics". In: *Acta Cryst. B* 72.4 (2016), p. 477.

[325] James Alexander Chisholm and Sam Motherwell. "COMPACK : a program for identifying crystal structure similarity using distances". In: *J. Appl. Cryst* 38 (2005), pp. 228–231.

[326] Jonas Nyman, Orla Sheehan Pundyke, and Graeme M Day. "Accurate force fields and methods for modelling organic molecular crystals at finite temperatures". In: *Phys. Chem. Chem. Phys.* 18.23 (2016), p. 15828.

[327] Graeme M. Day, W. D. Sam Motherwell, and William Jones. "Beyond the Isotropic Atom Model in Crystal Structure Prediction of Rigid Molecules:â€‰ Atomic Multipoles versus Point Charges". In: *Crystal Growth & Design* 5.3 (2005), pp. 1023–1033. DOI: 10.1021/cg049651n.

[328] Anton Pershin and Pater G. Szalay. "Improving the Accuracy of the Charge Transfer Integrals Obtained by Coupled Cluster Theory, MBPT(2), and TDDFT". In: *Journal of Chemical Theory and Computation* 11.12 (2015), pp. 5705–5711. DOI: 10.1021/acs.jctc.5b00837.

[329] Adam Kubas, Felix Hoffmann, Alexander Heck, Harald Oberhofer, Marcus Elstner, and Jochen Blumberger. "Electronic couplings for molecular charge transfer: Benchmarking CDFT, FODFT, and FODFTB against high-level ab initio calculations". In: *J. Chem. Phys.* 140.10 (2014), p. 104105. DOI: 10.1063/1.4867077.

[330] Quan Zhou, Peizhe Tang, Shenxiu Liu, Jinbo Pan, Qimin Yan, and Shou-Cheng Zhang. "Learning atoms for materials discovery". In: *Proceedings of the National Academy of Sciences* 115.28 (July 2018), E6411–E6417. DOI: 10.1073/pnas.1801181115.

[331] James Barker, Johannes Bulin, Jan Hamaekers, and Sonja Mathias. "LC-GAP: Localized coulomb descriptors for the gaussian approximation potential". In: *Scientific Computing and Algorithms in Industrial Simulations: Projects and Products of Fraunhofer SCAI.* Nov. 2017, pp. 25–42. DOI: `10.1007/978-3-319-62458-7_2`.

[332] John Fox. *Applied regression analysis and generalized linear models.* Sage Publications, 2015.

[333] Robert Tibshirani. "A Comparison of Some Error Estimates for Neural Network Models". In: *Neural Computation* 8.1 (Jan. 1996), pp. 152–163. DOI: `10.1162/neco.1996.8.1.152`.

[334] Andrew A. Peterson, Rune Christensen, and Alireza Khorshidi. "Addressing Uncertainty in Atomistic Machine Learning". In: *Phys. Chem. Chem. Phys.* 19.18 (2017), pp. 10978–10985. DOI: `10.1039/c7cp00375g`.

[335] Christopher M. Bishop and Cazhaow S. Qazaz. "Regression with input-dependet noise: A bayesian treatment". In: *Advances in Neural Information Processing Systems.* 1997, pp. 347–353.

[336] D.A. Nix and A.S. Weigend. "Estimating the mean and variance of the target probability distribution". In: *Proc. 1994 IEEE Int. Conf. Neural Networks* (1994), 55–60 vol.1. DOI: `10.1109/ICNN.1994.374138`.

[337] M Titsias. "Variational learning of inducing variables in sparse Gaussian processes". In: *Proc. Twelth Int. Conf. Artif. Intell. Stat.* 5 (2009), pp. 567–574.

[338] Dimitris N. Politis and Joseph P. Romano. "Large Sample Confidence Regions Based on Subsamples under Minimal Assumptions". In: *The Annals of Statistics* 22.4 (Dec. 1994), pp. 2031–2050. DOI: `10.1214/aos/1176325770`.

[339] Dimitris Politis, Joseph P. Romano, and Michael Wolf. "Weak convergence of dependent empirical measures with application to subsampling in function spaces". In: *Journal of Statistical Planning and Inference* 79.2 (July 1999), pp. 179–190. DOI: `10.1016/S0378-3758(98)00174-8`.

[340] B. Efron. "Bootstrap Methods: Another Look at the Jackknife". In: *The Annals of Statistics* 7.1 (Jan. 1979), pp. 1–26. DOI: `10.1214/aos/1176344552`.

[341] Leo Breiman. "Bagging predictors". In: *Machine Learning* 24.2 (1996), pp. 123–140. DOI: `10.1023/A:1018054314350`.

[342] Tom Heskes. "Practical confidence and prediction intervals". In: *Adv. Neural Inf. Process. Syst.* i (1997), pp. 176–182.

[343] Jörg Behler. "Representing potential energy surfaces by high-dimensional neural network potentials". In: *J. Phys. Condens. Matter* 26.18 (2014). DOI: `10.1088/0953-8984/26/18/183001`.

[344] I.A. Ibragimov and I.U.A. Rozanov. *Gaussian random processes.* Applications of mathematics. Springer-Verlag, 1978.

[345] Joaquin Quiñonero-Candela, Carl Edward Rasmussen, Fabian Sinz, Olivier Bousquet, and Bernhard Schölkopf. "Evaluating Predictive Uncertainty Challenge". In: Springer, Berlin, Heidelberg, 2006, pp. 1–27. DOI: `10.1007/11736790_1`.

[346] Giulio Imbalzano, Yongbin Zhuang, Venkat Kapil, Kevin Rossi, Edgar A. Engel, Federico Grasselli, and Michele Ceriotti. *Uncertainty estimation by committee models for molecular dynamics and thermodynamic averages.* Nov. 2020.

[347] Gideon Schwarz. "Estimating the Dimension of a Model". In: *Ann. Statist.* 6.2 (Mar. 1978), pp. 461–464. DOI: `10.1214/aos/1176344136`.

[348] H. Akaike. "A new look at the statistical model identification". In: *IEEE Transactions on Automatic Control* 19.6 (Dec. 1974), pp. 716–723. DOI: `10.1109/TAC.1974.1100705`.

[349] M. Langovoy. *Data-driven goodness-of-fit tests.* Ph.D. thesis. Göttingen: University of Göttingen, 2007, pp. ix+89.

[350] Gavin C. Cawley, Nicola L. C. Talbot, and Olivier Chapelle. "Estimating Predictive Variances with Kernel Ridge Regression". In: *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics).* Vol. 3944 LNAI. 2006, pp. 56–77. DOI: `10.1007/11736790_5`.

[351] Nicola Varini, Davide Ceresoli, Layla Martin-Samos, Ivan Girotto, and Carlo Cavazzoni. "Enhancement of DFT-calculations at petascale: Nuclear Magnetic Resonance, Hybrid Density Functional Theory and Car-Parrinello calculations". In: *Computer Physics Communications* 184.8 (Aug. 2013), pp. 1827–1833. DOI: `10.1016/j.cpc.2013.03.003`.

[352] E. Kucukbenli, M. Monni, B. I. Adetunji, X. Ge, G. A. Adebayo, N. Marzari, S. de Gironcoli, and A. Dal Corso. "Projector augmented-wave and all-electron calculations across the periodic table: a comparison of structural and energetic properties". In: *arxiv:1404.3015* (Apr. 2014).

[353] L. Birgé and P. Massart. "Gaussian model selection". In: *J. Eur. Math. Soc. (JEMS)* 3.3 (2001), pp. 203–268.

[354] B.V. Gnedenko and A.N. Kolmogorov. *Limit distributions for sums of independent random variables.* Addison-Wesley series in statistics. Addison-Wesley, 1968.

[355] *NIST Digital Library of Mathematical Functions.* http://dlmf.nist.gov/, Release 1.0.28 of 2020-09-15. F. W. J. Olver, A. B. Olde Daalhuis, D. W. Lozier, B. I. Schneider, R. F. Boisvert, C. W. Clark, B. R. Miller, B. V. Saunders, H. S. Cohl, and M. A. McClain, eds.

[356] William J. (William Jackson) Thompson. *Angular momentum : an illustrated guide to rotational symmetries for physical systems.* Wiley-VCH, 2004, p. 461.

[357] Jouni Suhonen. *From Nucleons to Nucleus.* Springer Berlin Heidelberg, 2007. DOI: `10.1007/978-3-540-48861-3`.

[358] Taweetham Limpanuparb and Josh Milthorpe. "Associated Legendre Polynomials and Spherical Harmonics Computation for Chemistry Applications". In: (Oct. 2014).

[359] John C. Light and Tucker Carrington. “Discrete-Variable Representations and their Utilization”. In: John Wiley & Sons, Ltd, Mar. 2007, pp. 263–310. DOI: 10.1002/9780470141731.ch4.

Félix Musil

Curriculum Vitae

18, avenue de l'Église Anglaise
1006 Lausanne
☎ *+41 78 822 58 75*
✉ *musil.felix@gmail.com*
felixmusil
29 years old (03.02.1991)
French

Education

	EPFL, École Polytechnique Fédérale de Lausanne, Switzerland
2016 - 2021	**PhD in Material Science**, *Laboratory of computational science and modelling*, advisor: Prof. Michele Ceriotti. A general and efficient framework for atomistic machine learning
2009 - 2015	**Bachelor and Master in Engineering Physics**.
2012 - 2013	**3$^{\text{rd}}$ year of Bachelor**, *KTH*, Royal Institute of Technology, Sweden. Erasmus exchange

Work Experiences

Feb. 2015 to Jul. 2015	**Master thesis**, *Swiss Plasma Center*, advisor: Prof. Ricci and Dr. Halpern. The impact of the Boussinesq approximation on the simulation of scrape-off layer plasma turbulence
Sep. 2014 to Jan. 2015	**EMS**, *Electro-Medical Systems*, Nyon, Switzerland. Experimental investigation of tribo-electric effects in powder handling
Jun. 2013 to Aug. 2013	**ONERA**, *Office Nationale d'Études et de Recherche Aérospatiales*, Châtillon, France. Numerical modelling to optimize the elaboration process of an anti-oxidation coating
2011 to 2019	**EPFL**, *Teaching assistant*. Tutor of a ten students group during the exercise sessions of Physics course (app. 4h./week)

Features

C++ Project	Main developer of an open source library to compute representations for atomic scale learning `https://github.com/cosmo-epfl/librascal`
Web App	ShiftML : An interactive web-app that calculates chemical shifts for molecular crystals using machine learning `http://shiftml.org/`
Outreach	Opinion article featuring data-driven periodic tables on chemistryworld
Award	Best poster prize "Developing High-Dimensional Potential Energy Surfaces – From the Gas Phase to Materials" 24.-26. April 2019, Göttingen

Computer Skills

Programming	Python, C++14, Fortran, LaTex
Atomistic Modelling	Quantum Espresso, DFTB+, LAMMPS, i-Pi
Data Science	scikit-learn, pyTorch, TensorFlow
Software	Adobe Suite, Blender, Microsoft Office

Languages

English	**C2 level,** written and spoken
Spanish	**B1 level,** written
French	**Mother tongue**

Interest and Hobbies

Culture	Classical concert, opera Travel, painting and sculpture
Sport	Badminton, hiking

Presentations

Poster	"Theory and Practice of Atom-Density Representations for Machine Learning", Developing High-Dimensional Potential Energy Surfaces – From the Gas Phase to Materials, 2019, Göttingen
Oral presentation	"Machine Learning for Molecular Materials: Stability, Properties and Experimental Obeservables", International Workshop on Machine Learning for Materials Science, 2018, Aalto University, Helsinki
Oral presentation	"Machine learning the structure-energy-property landscapes of molecular crystals", DPG 2018, TU Berlin
Poster	"Mapping and Classifying Molecules from a High-Throughput Structural Database", International Workshop on Machine Learning for Materials Science, 2017, Aalto University, Helsinki

Publications

[1] **Musil, F.**, Ceriotti, M., "Machine learning at the atomic scale". In: *Chimia* 73.12 (Dec. 2019), pp. 972–982. DOI: `10.2533/chimia.2019.972`.

[2] **Musil, F.**, Willatt, M. J., Langovoy, M. A., Ceriotti, M., "Fast and Accurate Uncertainty Estimation in Chemical Machine Learning". In: *Journal of Chemical Theory and Computation* 15.2 (Feb. 2019), pp. 906–915. DOI: `10.1021/acs.jctc.8b00959`.

[3] Willatt, M. J., **Musil, F.**, Ceriotti, M., "Atom-density representations for machine learning". In: *Journal of Chemical Physics* 150.15 (Apr. 2019), p. 154110. DOI: `10.1063/1.5090481`.

[4] **Musil, F.**, De, S., Yang, J., Campbell, J. E., Day, G. M., Ceriotti, M., "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`.

[5] Paruzzo, F. M., Hofstetter, A., **Musil, F.**, De, S., Ceriotti, M., Emsley, L., "Chemical shifts in molecular solids by machine learning". In: *Nature Communications* 9.1 (Dec. 2018), p. 4501. DOI: `10.1038/s41467-018-06972-x`.

[6] Willatt, M. J., **Musil, F.**, Ceriotti, M., "Feature optimization for atomistic machine learning yields a data-driven construction of the periodic table of the elements". In: *Nature Communications* 20.47 (Dec. 2018), pp. 29661–29668. DOI: `10.1039/c8cp05921g`.

[7] De, S., **Musil, F.**, Ingram, T., Baldauf, C., Ceriotti, M., "Mapping and Classifying Molecules from a High-Throughput Structural Database". In: *Journal of Cheminformatics* 9.1 (Dec. 2016), p. 6. DOI: `10.1186/s13321-017-0192-4`.