Title

Applied Lattice Models for the Study of Configuration Thermodynamics of Multicomponent Crystalline Materials

Author

Barroso-Luque, Luis

Publication Date

2022

Peer reviewed|Thesis/dissertation

Applied Lattice Models for the Study of Configuration Thermodynamics of Multicomponent Crystalline Materials

by

Luis Barroso-Luque

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Materials Science and Engineering

and the Designated Emphasis

in

Computational and Data Science and Engineering

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Gerbrand Ceder, Chair
Professor Daryl Chrzan
Professor Ahmad Omar

Fall 2022

Applied Lattice Models for the Study of Configuration Thermodynamics of Multicomponent Crystalline Materials

Abstract

Applied Lattice Models for the Study of Configuration Thermodynamics of Multicomponent Crystalline Materials

by

Luis Barroso-Luque

Doctor of Philosophy in Materials Science and Engineering

and the Designated Emphasis in

Computational and Data Science and Engineering

University of California, Berkeley

Professor Gerbrand Ceder, Chair

Computational methods that enable calculations of thermodynamic properties emergent from the specific arrangements of distinct atomic species in materials have become indispensable in the advent of flourishing research into materials with increasingly large numbers of components. Calculations involving lattice models fitted to the energies of a set of representative atomic configurations of a material—predominantly by way of the cluster expansion method—are now standard and commonly used by researchers. As researchers explore materials with growing numbers of components, continued development of cluster expansion-based methodology has ensued. Although substantial progress has been made, the vast majority of developments have focused solely on statistical regression methodology to fit expansions using the original underlying mathematical formalism largely unchanged.

In this thesis, we revisit the mathematical framework underlying the cluster expansion method and re-establish it in a more general form as a representation for generalized lattice Hamiltonians of atomic configuration. In doing so, we present two categories of representation that are found to be direct generalizations of the Ising and Potts models respectively. We rigorously define Fourier cluster expansions—those used in the original formalism of the cluster expansion method—and present some of their useful mathematical properties. We then show how, regardless of the particular choice of basis, Fourier cluster expansions are essentially expressions of a unique *cluster decomposition*. The intimate relation between the cluster decomposition and well-established function decompositions used in statistics establishes an avenue to a formal interpretation of expansion terms as the mean of statistically independent atomic interactions. The second representation, which we have named the *gen-*

eralized Potts frame, involves a redundant representation by way of a mathematical frame. By constructing a representation that is *over-complete* (more functions than dimensions) additional robustness and expressiveness when estimating coefficients are obtained. We illustrate the capability of the Potts frame representation to fit the most accurate Hamiltonian for a system, which to the best of our knowledge, represents the largest configuration space attempted to date. We also describe general and practical ways to implement the aforementioned representations of lattice Hamiltonian and methods to carry out calculations with improved time complexity relative to other available cluster expansion implementations.

The formal structure of Fourier cluster expansions and Potts frame expansions are then used to motivate and develop novel structured-sparsity-based linear regression methods that allow robust parametrization of generalized lattice models from first principles electronic structure calculations. The methods developed rely on establishing structural priors on the expansion coefficients—some of which have previously been based on heuristics—which we motivate and justify with more rigorous mathematical and statistical arguments. The regression methods were developed with the goal of enabling accurate estimation of expansion coefficients in high dimensional configuration spaces using relatively small samples of training structures. We describe a series of practical implementations and auxiliary methods necessary for the practical implementation and learning of applied lattice models of complex multi-component materials. Finally, we demonstrate the successful application of the methodology developed to learn lattice models of several Li transition metal oxides and medium entropy alloys that have garnered considerable attention from researchers due to their remarkable and technologically relevant properties.

The thesis is concluded by suggesting avenues for continued development of lattice-based methods geared toward studying order and partial disordered in inorganic multi-component materials. A general commentary on the suite of lattice-based methodology in the context of the rapidly growing development of machine learning potentials is given.

To Mr. Baco
a loyal friend and an unexpected teacher

Contents

List of Figures

List of Tables

Acknowledgments

First and foremost, I want to thank my family for their unconditional support throughout my life. Without the fortune of a privileged socioeconomic and psycho-emotional upbringing that my parents, Catalina and Luis, so determinedly provided me with, none of this would have been possible. Likewise, my close-knit relationship with my parents and siblings, Catalina and Roberto, has been a steadfast source of relief that consistently made me feel well accompanied even in the deepest periods of solitude brought on by the nature of graduate school—on occasion also aggravated by my own proclivities.

I would also like to express my sincere gratitude to my advisor, Professor Gerd Ceder. Gerd has a distinctive ability to offer deep insight into research problems pushing the furthermost borders of his core expertise. Gerd provided me the freedom and guidance to channel my curiosity and efforts into pursuing the research (this work) that I found most exciting.[1]

Special thanks to my colleague, collaborator, and friend, Julia Yang. She has been an invaluable companion in the entirety of this journey. I met Julia before even committing to attend Berkeley; and since then, she helped me navigate a majority of the personal, academic, intellectual, and bureaucratic obstacles that inevitably emerged during graduate school. Thank you Julia for your support, advice, countless hours of CE discussions, and perhaps even more hours of venting frustrations.

The work presented in this thesis would have fallen much shorter without the contributions from the members of the *cluster expansion* subgroup.[2] Witnessing how many of the tools I developed were used and extended in ways I never considered, has been truly rewarding well beyond simply satisfying my own intellectual curiosities. Thanks to Tina Chen, Fengyu Xie, Zinab Jadidi, Peichen Zhong, and Ronald Kam—working with this group of incredibly talented people has been incredibly humbling.

I also want to thank the Materials Science & Engineering department staff and in particular Ariana Castro. Ariana patiently guided me through countless expected and unexpected official procedures necessary to complete this degree. Her kind, timely and informative responses to my emails (which I am sure were only but a few on top of a much larger pile) were immensely helpful in the otherwise baffling experience of navigating the bureaucracy of UC Berkeley at large.

Lastly, I am incredibly grateful for the friendships that I have made during my time at UC Berkeley. I have never appreciated the value of a supportive community more than I did during the challenging final months of completing this dissertation. To all my Berkeley friends, thank you for your kindness, your care, your support, and for making life in Berkeley and the rest of California so memorable and encompassing much more than graduate school.

[1] Albeit not after a struggle to get his attention.

[2] Now renamed the *statistical learning and simulation* subgroup—because we had to jump on the machine learning bandwagon, but were not ready to drink the entirety of the Kool-Aid.

Chapter 1

Introduction

> Science grows through accretion but becomes potent through distillation.
>
> - Professor James P. Sethna [200]

The work in this thesis concerns the formal development of mathematical representations of generalized Lattice models with discrete configurations. In addition, practical methodology to fit lattice models of real multi-component materials is developed and benchmark applications are reported. These methods subsequently allow efficient thermodynamic calculations of properties that depend on atomic ordering. We first briefly motivate this work by highlighting both the scientific and technological value in deepening our understanding of the roles of (partial) disorder of atomic configuration in the thermodynamic stability and emergent properties of multi-component inorganic crystalline materials. Subsequently, we provide a basic exposition of statistical mechanics and thermodynamics driven by atomic configuration. Finally, we introduce and motivate the role of lattice models as a compelling mathematical framework that enables practical statistical thermodynamic calculations in complex multi-component crystalline materials of recent scientific and technological interest.

1.1 Multi-principal element materials

The mixing of different components—in the present case chemical species—with the goal of tailoring physical properties of materials is at the very foundation of materials science and engineering research. One might expect that a mixture of different components would simply yield materials with properties that are averages of the constituent properties. Notwithstanding, the intricacy of atomic interactions between components results in a true materials *gestalt*, where quite different and often remarkably enhanced properties can emerge.

In the last two decades the mixing of several elemental species in almost equal proportions, in what are now referred to as high entropy or multi-principal element (MPE) materials, has become predominant in materials research and novel materials design [143]. The combinatorial growth of composition spaces under the MPE paradigm has allowed the

discovery and precise tuning of materials with properties of significant technological value [76, 78, 163]. Among many applications, the MPE paradigm has led to the discovery of several classes of materials with remarkable structural and functional properties. For example, high entropy metallic alloys, such as the Cantor (CrMnFeCoNi) and related alloys can exhibit both high strength and high ductility [78]. MPE CrMoNbVZr based nitrides have been explored as supercapacitors due to their notably high capacitance [110]. CuS based semiconductor MPEs have been found to have exceptionally high thermoelectric figures of merit (ZT) [257]. MPE materials involving LiMnOF have been also found to have remarkable electrochemical properties when used as high-capacity battery electrodes [45]. Furthermore, a myriad of other MPE materials have been studied for several other technological applications including hydrogen storage, piezo-electrics, electronic transformers, soft magnets, spintronics, and thermal insulators [76, 163].

The viability and promising properties of MPE materials are strongly dependent on the nature of atomic interactions and the resulting energy landscape in terms of atomic configurations. A multitude of relevant properties including stability and resulting phase diagrams [3, 48, 118, 139, 160, 164, 182, 227], short-range order [41, 46, 67, 108, 136, 159, 204], ionic percolation [165, 224], among many others [143] are strongly driven by the thermodynamics of atomic configuration. As a result, the development of accurate computational thermodynamic methods is indispensable to deepen our understanding and further explore the nature of these atomic interactions and the effects of different levels of atomic disorder in determining the thermodynamic stability and emergent physical properties of MPE materials.

1.2 A statistical thermodynamics primer

We begin by giving a brief distillation of the basic concepts of equilibrium statistical thermodynamics that are necessary for the study of atomic configuration and disorder in materials science. Although it is common practice to separate the subject into statistical mechanics (the microscopic details of matter) and thermodynamics (the emerging macroscopic states of matter), we attempt to give our exposition of some of the basic concepts in the thermal physics of materials in a cohesive manner. This hopefully makes our exposition briefer and helps to emphasize the practical value of (microscopic) first principles-based computational methods in extending our understanding of technologically relevant (macroscopic) properties of materials. In short, the essence of equilibrium statistical thermodynamics is to describe the macroscopic states of matter, their properties, and their responses to changes in their environment, all of which emanate from the physical interactions of its microscopic constituents and the fluctuations of said interactions at finite temperatures.

Macroscopic and microscopic states

A macroscopic state of matter is specified by a small and finite set of *extensive variables*.[1] Relevant extensive variables include the internal energy E, the number of chemical species present N_i, total volume V, electric polarization $\boldsymbol{P}$ and magnetization $\boldsymbol{M}$ [71]. We universally denote an extensive variable with the scalar X and a set of several extensive variables with the vector $\boldsymbol{X}$. A macroscopic state of matter is fully specified by the values of all relevant extensive variables. In addition, a conjugate intensive variable[2]—which we denote with Y and $\boldsymbol{Y}$—is associated with each extensive variable. Intensive variables include chemical potentials μ_i, pressure P, electric fields $\boldsymbol{E}$, and magnetic fields $\boldsymbol{H}$. Conjugate pairs of extensive and intensive variables represent means for matter to exchange energy with its surroundings [26, 71]. Changes in the macroscopic states of matter and the accompanying energy exchanges can be fully specified by the initial and final set of extensive variables X and the values of their conjugate intensive variables Y. The set of pairs of conjugate variables X, Y is collectively referred to as *thermodynamic coordinates*. Among the set of extensive variables, the internal energy U plays a central role and must always be present to fully specify any macroscopic state of matter [26, 71]. Changes in internal energy at a given temperature constitute precisely what we call thermal interactions—that is energy transfer that occurs as heat—which are essential to give rise to the emergent thermodynamic properties of materials at finite temperatures. The internal energy U, along with its conjugate intensive variable, inverse temperature $1/T$, is thus imperative to obtain a complete description of the equilibrium properties of macroscopic states interacting with their surroundings.

In contrast, microscopic states or microstates must be specified by a very large number of mechanical variables. The specific mechanical variables depend on the physical principles used in the description at hand. For example, in a classical system, the positions $\boldsymbol{r}$ and momenta $\boldsymbol{p}$ of all particles specify a microscopic state. For a quantum mechanical system, a microstate is specified by the precise components of an infinite dimensional state vector $|\Psi\rangle$ in a suitable basis [112]. A theoretically formal treatment of matter requires a quantum mechanical treatment, however, the framework of statistical thermodynamics is independent of the specific mechanical description of microscopic constituents [71, 173]. In fact, many practical approximations can be obtained from classical descriptions or from a mixture of quantum and classical (semi-classical) descriptions. A semi-classical description of materials is the foundation of the methods developed in this thesis. Specifically, we will label a general microstate as $\boldsymbol{s}$—with the implication that fully specifying it requires knowing a large, sometimes infinite set of *coordinates*. The space of all possible microstates $\boldsymbol{s}$ of a particular system is referred to as its *phase space*. Furthermore, any mechanical description of a microstate in statistical thermodynamics must include a *Hamiltonian* function $\mathcal{H}$ that

[1] That is variables that are first-order homogeneous functions of the size of a system (commonly specified by the number of particles).

[2] These are formally zeroth order homogeneous functions of the system size; meaning their values are independent of system size.

maps any possible microstate $\boldsymbol{s}$ to its total energy.[3]

Functions $\widetilde{X}(\boldsymbol{s})$ defined over phase space and which vary smoothly $\boldsymbol{s}$ with are referred to as *observables* [173]. The connection between macroscopic and microscopic states lies in the statistical relationship between extensive variables X and observables $\widetilde{X}(\boldsymbol{s})$. Specifically, the values of macroscopic extensive variables X are the expectation of thermodynamic *observables*,

$$X = \langle \widetilde{X}(\boldsymbol{s}) \rangle \tag{1.1}$$

Notably, the internal energy E for a thermally interacting system is the expectation value of its Hamiltonian, $E = \langle H \rangle$.

The fundamental postulate

Having described the formal connection between microscopic and macroscopic states as the expectation of observables, we are still left with specifying the details of the probability distribution over which the expectation in Equation 1.1 is taken. In reality, observables measured in experimental measurements, are time averages over the dynamic evolution of microstates.[4] However, to do so computationally, we would need to specify the time evolution of each of the large number of variables specifying the microstate of the system—an effort that is intractable for all but the simplest systems. The situation is elegantly resolved in statistical thermodynamics by replacing time averages with the expectation with respect to appropriate probability distributions—referred to as thermodynamic ensembles—over accessible volumes of phase space.[5] The prescription to specify these *thermodynamic ensembles* rests on a fundamental postulate, here named *the fundamental postulate*, that establishes a deeper connection between macroscopic states and the consistent volume of phase space (i.e. sets of microscopic states).

The fundamental postulate of statistical thermodynamics presupposes the existence of the concept of entropy as a function of the total number of accessible microstates, or in other words, the accessible volume of phase space [173]. Furthermore, the set of accessible states is determined by the values of macroscopic variables set by the environment, which are referred to as the *thermodynamic boundary* conditions. In light of this, the entropy can also be considered an extensive function of the prescribed set of thermodynamic coordinates which define the resulting macroscopic state.

A general expression for entropy is given by the Gibbs entropy formula [14, 112],

$$S(\Gamma(\boldsymbol{s})) = -k_B \sum_{\boldsymbol{s}} \mathbb{P}(\boldsymbol{s}) \ln \mathbb{P}(\boldsymbol{s}) \tag{1.2}$$

[3] In classical and semi-classical descriptions $\mathcal{H}$ is usually a function. In quantum description it is an *operator* [112].

[4] As governed by their Hamiltonian and associated equations of motion [112].

[5] Rigorous justification for the validity of doing so requires the phase space and dynamics of microstates to satisfy certain conditions, such as ergodicity [113]. Fortunately, these conditions are favorably met in almost all physical systems addressed in practice.

where $\Gamma(\boldsymbol{s})$ denotes the accessible phase space volume; k_B is Boltzmann's constant; and $\mathbb{P}(\boldsymbol{s})$ is the thermodynamic ensemble distribution—that with respect to which the expectation values to compute observables are taken with.

The maximization of entropy determines equilibrium macroscopic states and the distribution of the underlying thermodynamic ensemble of microstates as stated in the fundamental postulate of statistical thermodynamics:

> The values of the extensive parameters $(E, \boldsymbol{X})$ which specify a macroscopic state in thermodynamic equilibrium are those that maximize the entropy over all internal degrees of freedom—microscopic states—consistent with the thermodynamic boundaries [26, 173]. Equivalently, the equilibrium thermodynamic ensemble is the probability distribution over accessible microstates which maximizes the entropy subject to constraints compatible with the thermodynamic boundaries.

When entropy is expressed as a function of extensive variables $S = S(E, \boldsymbol{X})$, it is referred to as the fundamental equation of thermodynamics since it contains a *complete* thermodynamic description of the system considered [26, 173]. Apart from determining equilibrium and stability conditions, all conjugate variables Y and thermodynamic material properties[6] γ (i.e. heat capacities, compressibility, susceptibilities) can be obtained simply by taking derivatives of the fundamental equation,

$$Y = \left(\frac{\partial S}{\partial X}\right)_{\boldsymbol{X}\setminus X}$$
$$\gamma = \left(\frac{\partial^2 S}{\partial X^2}\right)_{\boldsymbol{X}\setminus X}$$

where $\boldsymbol{X} \setminus X$ means all extensive variables of S except the variable X being differentiated with respect to. We have lumped the energy E into the generic set of extensive variables $\boldsymbol{X}$. Explicitly, the derivative of entropy with respect to internal energy is the inverse temperature $1/T = (\partial S/\partial X)_{\boldsymbol{X}}$, which indeed can be shown to correspond to our intuitive concept of temperature [26].

A final requirement to complete the basis of statistical thermodynamics is the principle of *equal apriori probabilities* [162]. The principle states that lacking further information, the accessible microstates in an isolated system—all values of extensive variables fixed—in equilibrium occur with equal probability [37, 173]. This implies that for isolated systems—in the *micro-canonical ensemble*—the entropy and resulting thermodynamic ensemble distribution are given by,

$$\begin{aligned} S(\Gamma(E)) &= k_B \ln \Gamma(E) \\ \mathbb{P}(\boldsymbol{s}) &= \exp\left(\frac{S}{k_B}\right) = \frac{1}{\Gamma(E)} \end{aligned} \tag{1.3}$$

[6] Materials properties are also known as response functions, since they can be measured experimentally as a response to changes in thermodynamic coordinates [112].

where $\Gamma(E)$ represents the phase space volume or total number of micro-states with energy equal to the fixed value of energy E: $H(\boldsymbol{s}) = E$.

It is easy to check that the uniform probability distribution in Equation 1.3 gives the solution to the maximization of entropy under no external constraints[7], which implies fixed extensive thermodynamic coordinates as boundary conditions. However, in many situations, matter does not exist as an isolated system. Specifically, for a vast majority of cases, materials on earth exist in an environment (i.e. the atmosphere) that sets, at a minimum, the values of temperature and pressure as boundary conditions.

Thermodynamic potentials and ensembles

Situations where matter is not isolated, meaning some thermodynamic boundary conditions involve fixed values of intensive quantities, are described in statistical thermodynamics by appealing to different *thermodynamic potentials* and their associated thermodynamic ensembles. Conceptually, this requires treating the extensive variables conjugate to the intensive quantities being fixed explicitly as observables $X = \langle \widetilde{X}(\boldsymbol{s}) \rangle_T$. In order to find the respective distribution one now needs to maximize the entropy subject to constraints that will ensure that observables correctly match the value of their corresponding extensive variable, $X = \langle \widetilde{X}(\boldsymbol{s}) \rangle_T$. This prescription is achieved by a constrained maximization of the entropy over all possible distributions $\mathbb{P}(\boldsymbol{s})$ as follows,

$$\underset{\mathbb{P}(\boldsymbol{s})}{\operatorname{argmax}} \; -k_B \sum_{\boldsymbol{s}} \mathbb{P}(\boldsymbol{s}) \ln \mathbb{P}(\boldsymbol{s}) - \alpha \left(\sum_{\boldsymbol{s}} \mathbb{P}(\boldsymbol{s}) - 1 \right) - \boldsymbol{\lambda}^\top \left(\sum_{\boldsymbol{s}} \mathbb{P}(\boldsymbol{s}) \widetilde{\boldsymbol{X}}(\boldsymbol{s}) - \boldsymbol{X} \right) \tag{1.4}$$

where α and $\boldsymbol{\lambda}$ are Lagrange multipliers for the normalization of the distribution and thermodynamic consistency constraints respectively.

The solutions of the above maximization are of the following form [112],

$$\mathbb{P}(\boldsymbol{s}) = e^{-\boldsymbol{\lambda}^\top \widetilde{\boldsymbol{X}}(\boldsymbol{s}) - \Psi / k_B} \tag{1.5}$$

Where the Lagrange multipliers $\boldsymbol{\lambda}$ are in fact the respective negative conjugate extensive quantities[8] normalized by the thermal unit of energy $k_B T$, $\boldsymbol{\lambda} = -\boldsymbol{Y}/k_B T$ [14, 173]. The remaining factor in the exponential, $e^{-\Psi/k_B}$, represents a normalization factor that depends on the function Ψ which is determined based on the fluctuating extensive variables $\boldsymbol{X}$.

Moreover, to account for thermal interactions, energy must be allowed to fluctuate ($\langle H(\boldsymbol{s}) \rangle_T = E$) such that writing it out explicitly in Equation 1.5 gives a more familiar expression known as a *Boltzmann distribution*,

$$\mathbb{P}(\boldsymbol{s}) = \frac{1}{Z} \exp\left(-\beta H(\boldsymbol{s}) + \beta \boldsymbol{Y}^\top \widetilde{\boldsymbol{X}}(\boldsymbol{s}) \right) \tag{1.6}$$

[7] A normalization constraint is always necessary to ensure the solution is a valid probability distribution.

[8] Formally they correspond to conjugate extensive quantities in the thermodynamic representation of entropy [26].

where $\beta = 1/k_B T$ is the inverse temperature in units of energy. Z is a re-writing of the normalization factor $e^{-\Psi/k_B}$ as what is known as the *partition function*. Explicitly, since Z is used to ensure that $\mathbb{P}(\boldsymbol{s})$ is properly normalized, it is given by,

$$Z(\beta, \beta\boldsymbol{Y}) = \sum_{\boldsymbol{s}} \exp\left(-\beta H(\boldsymbol{s}) + \beta \boldsymbol{Y}^\top \widetilde{\boldsymbol{X}}(\boldsymbol{s})\right) \tag{1.7}$$

It is precisely a Boltzmann probability given in Equation 1.6 over which the thermodynamic expectation values to compute observables $\langle \widetilde{X}(\boldsymbol{s}) \rangle_T$ are taken.

Using Equation 1.7 for the partition function, we can obtain the following expression for the function Ψ,

$$\Psi(\beta, \beta\boldsymbol{Y}) = k_B \ln Z(\beta, \beta\boldsymbol{Y}) \tag{1.8}$$

And based on the solution of the maximization of Equation 1.4 and additional thermodynamic relations [26, 112], one can show that the function Ψ can also be expressed as,

$$\Psi(T, \boldsymbol{Y}, \boldsymbol{X}') = S(E, \boldsymbol{X}, \boldsymbol{X}') - \frac{1}{T}E + \frac{1}{T}\boldsymbol{Y}^\top \boldsymbol{X} \tag{1.9}$$

Where $\boldsymbol{X}'$ refers to the set of all extensive variables that are set as environmental boundary conditions (i.e. not allowed to fluctuate). We recognize Equation 1.9 as a Legendre transformation of the entropy $S(E, \boldsymbol{X}, \boldsymbol{X}')$ with respect to the intensive variables $\boldsymbol{Y}$.

A rearrangement of Equation 1.9 results in what is known as the *free energy* or thermodynamic potentials for the given thermodynamic system under the prescribed set of thermodynamic boundary conditions,

$$F(T, \boldsymbol{Y}, \boldsymbol{X}') = -T\Psi(T, \boldsymbol{Y}, \boldsymbol{X}') = E(S, \boldsymbol{X}, \boldsymbol{X}') - TS - \boldsymbol{Y}^\top \boldsymbol{X} \tag{1.10}$$

where now the equation represents a Legendre transformation of the internal energy $E(S, \boldsymbol{X}, \boldsymbol{X}')$ with respect to the intensive variables $\boldsymbol{Y}$ [26, 37, 173].

Lastly, using Equations 1.8 and 1.10 we can also obtain the free energy in terms of the partition function,

$$F(\beta, \beta\boldsymbol{Y}, \boldsymbol{X}') = -k_B T \ln Z(\beta, \beta\boldsymbol{Y}) \tag{1.11}$$

The expression for free energy in terms of the partition function given in Equation 1.11 is essentially what allows us to carry out thermodynamic calculations based on a microscopic description of matter. In fact, an expression for free energy constitutes a fundamental thermodynamic equation in the sense that it also contains a complete description of a system's thermodynamics, by which fluctuating thermodynamic coordinates and response functions can be computed by taking the appropriate derivatives.

The methodology developed in this thesis allows direct statistical thermodynamic calculations in the following two thermodynamic ensembles, which are commonly used thermodynamic ensembles in materials research,

- The canonical ensemble where only energy is allowed to fluctuate. The corresponding free energy $F(T) = A(T) = E - TS$ is called the Helmholtz free energy.

- The (semi) grand canonical ensemble where energy and species compositions are allowed to fluctuate. The free energy $F(T, \boldsymbol{\mu}) = \Phi(T, \boldsymbol{\mu}) = E - TS - N \sum \mu_i x_i$ (where $\boldsymbol{\mu}$ are relative chemical potentials for the species whose compositions $\boldsymbol{x}$ are allowed to fluctuate) is known as the Grand potential.

Specifically, the methods we develop enable statistical thermodynamic calculations for micro-states of *atomic configuration* in multi-component crystalline materials. By configurations, we mean the precise arrangement of species in space. For inorganic crystalline materials, this means the precise arrangement of chemical species over crystallographic sites; or in other words the possible *colorings* or *decorations* of the sites in a crystal structure with species chosen from prescribed sets of allowed species.

1.3 Generalized lattice models of atomic configuration

In order to make use of Equations 1.6 and 1.11 for statistical thermodynamic calculations, we need a precise way to construct a Hamiltonian $H(\boldsymbol{s})$ where the microstates $\boldsymbol{s}$ correspond to atomic configurations. Such an endeavor, and precisely the one taken up in this work, can be suitably achieved using *lattice models*. Lattice models have always played an essential role in statistical physics since the very inception of the field. Their use has been crucial to understanding and discovering a variety of complex physical phenomena, such as phase transitions, critical phenomena, and universality classes [14, 200].

Methods based on generalized lattice models are also some of the most widely used techniques in the computational study of configuration thermodynamics in materials. Specifically, the cluster expansion method [192] coupled with Monte Carlo sampling has become an established tool in the computational study of first principles thermodynamics and statistical mechanics of multi-component metal alloys and ionic materials [212, 231]. The mathematical framework of the cluster expansion method has been used to further establish additional methods that have themselves become invaluable in the computational study of multi-component materials. For example, special quasi-random structures (SQS) [263], special quasi-ordered structures (SQoS) [135, 184], and small sets of ordered structures (SSOS) [109], enable generation of representative structures for different levels of configurational disorder within manageable supercell sizes, such that they are amenable to calculations with highly accurate first-principles electronic structure methods. Additionally, cluster expansion-based methods allow practical—and in some cases rigorously provable—generation of ground state and near ground state configurations[100, 103, 128].

However, the development and applications of the aforementioned methods has been mostly limited to binary and ternary metallic alloys and transition metal oxides. Recent interest in applying such methods to higher dimensional and more complex configuration spaces has required development of additional methodology to address many practical challenges [9, 150, 153, 253]. Despite several developments in parameter estimation and training data sampling for building cluster expansions, scant attention has been given recently to the

formal mathematical framework on which the vast majority of these methods are based. An essential part of the present work is focused precisely on revisiting and extending the underlying formal mathematical framework for representing and fitting generalized Lattice models. Before doing so, let us briefly contextualize and motivate the endeavor by introducing three of the most prominent classical lattice models.

The Ising, lattice gas and Potts models

We briefly introduce three of the simplest yet most studied classical lattice models, the Ising model, the lattice gas model, and the Potts model, because they are deeply connected to the generalized representations that we develop later on.

The Ising model is one of the simplest but most heavily studied lattice models that exhibits non-trivial statistical thermodynamic behavior. In its original form, which involves only nearest neighbor interactions, the Ising model exhibits a phase transition at finite temperature for spatial dimensions > 2 [22]. Such is the versatility of the Ising model, that it can be used as a simple model for magnetism, ordering in binary alloys, and liquid-gas transitions.

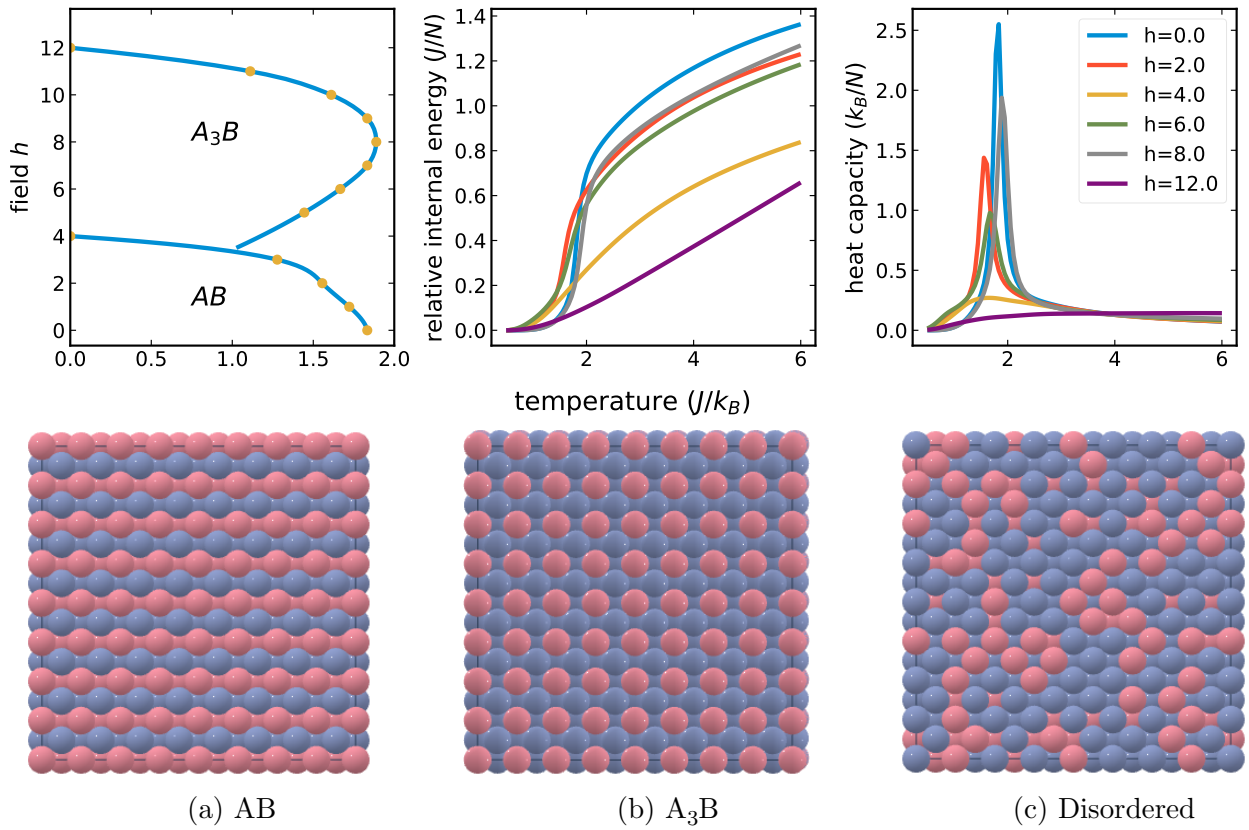
The Ising model Hamiltonian involves a collection of two-state *spin* variables $s_i = \pm 1$ located at sites on a specified lattice. The Ising Hamiltonian with nearest neighbor pair interactions and a magnetic field is given as,

$$H(\boldsymbol{s}) = -h \sum_i s_i - J \sum_{<ij>} s_i s_j \tag{1.12}$$

where $\boldsymbol{s} = s_1, s_2, \ldots, s_N$ is a particular configuration of N *spins*. The sum over $< ij >$ is done over all nearest neighbor sites i, j. h is the value of an externally applied field. And J is a pair interaction parameter.

A variety of rich thermodynamic behavior of the Ising model results from considering distinct domains specified by lattices with different symmetries [15, 16, 44, 126]. Yet even when considering the same lattice, the relative magnitudes of h and J can also drastically change the resulting thermodynamic behavior. As an example exhibiting complex thermodynamic behavior, Figure 1.1 shows a computed phase diagram and sampled configurations from Monte Carlo calculations of a face-centered cubic (FCC) anti-ferromagnetic (AF) ($J = -1$) Ising model. The simple FCC-AF Ising model exhibits first order ($h = 0$) and second order ($h > 0$) phase transitions and two different states order. When used as a model for a binary AB alloy (A = 1, B = -1, h $\geq$ 0) the model exhibits different ordering transitions based on the magnitude of the field h (which in this case h can be thought of as a relative chemical potential). Indeed, if the chemical pair interactions for a binary alloy are E_{AA}, E_{BB} and E_{AB}, then the interaction parameter is given by $J = \frac{1}{4}(E_{AA} + E_{BB} - 2E_{AB})$ and the field as $h = \frac{z}{4}(E_{AA} - E_{BB})$ (z is the number of nearest neighbors per site).[9]

[9] We are assuming positive values of the parameters, that is $E_{\circ\circ} > 0$ represent attractive interactions (i.e. they lower the energy) based on the form of Equation 1.12.



(a) AB (b) A_3B (c) Disordered

Figure 1.1: Monte Carlo results of a face-centered cubic anti-ferromagnetic Ising model. The top panel shows the phase diagram, relative internal energy, and heat capacities for different values of magnetic field (relative chemical potentials) from Wang-Landau sampling in a 256-site supercell. The bottom panel shows examples of ordered and disordered configurations from Metropolis-Hastings simulated annealing of a 2048 site supercell.

The Ising model can also be used to model liquid-gas transitions simply by choosing a spin value, say $s = 1$ to represent an occupied site and the other $s = -1$ to represent a vacant site. However, it is much more explicit to instead use the change of variables $n = (s + 1)/2$, so that the resulting *occupation variable* $n = \{0, 1\}$ is equal to 1 if the site is occupied and zero otherwise. With this change of variables, one obtains the *lattice gas model*, and its Hamiltonian is given by,

$$H(\boldsymbol{n}) = -\mu \sum_i n_i - \varepsilon \sum_{<ij>} n_i n_j \tag{1.13}$$

where the parameters are related to the Ising model as follows, the chemical potential is given by $\mu = 2(h - qJ)$ and the interaction energy by $\varepsilon = 4J$.

In fact, a lattice gas model can also be used to model a binary alloy. If we simply re-interpret the occupation variables n to represent the presence of one of the species of the

alloy, and the absence of said species implies the other allowed species is present. For example $n_i = 1$ means that species A sits at the i-th site, and $n_i = 0$ that species A is not at the i-th, implying that B is. In such a case the relationship between lattice gas interactions and species interactions E_{AA}, E_{BB} and E_{AB} are given by, $\mu = zE_{AB}$ and $\varepsilon = (E_{AA}+E_{BB}-2E_{AB})$.

The equivalence between the Ising model and the lattice gas model can be made more evident by replacing the spin and occupation variables with a general site function $\phi : \Omega \to [-1, 1]$; where $\Omega = \{A, B\}$ is a set with two elements that represent the states each site can have. In doing so, the Ising and the lattice gas Hamiltonian can be expressed in the following general form,

$$H(\boldsymbol{\sigma}) = -J_1 \sum_i \phi(\sigma_i) - J_2 \sum_{<ij>} \phi(\sigma_i)\phi(\sigma_j) \tag{1.14}$$

Where one can now think of the variables $\sigma \in \Omega$ abstractly as states—explicitly as species for the binary alloy scenario $\sigma \in \{A, B\}$. For the case of the Ising model the site function can be explicitly expressed as $\phi(\sigma) = 2\mathbf{1}_A(\sigma) - 1$. In the Lattice gas, the site function is simply $\phi(\sigma) = \mathbf{1}_A(\sigma)$. Where $\mathbf{1}_A(\sigma)$ is an singleton indicator function for state A (i.e. $\mathbf{1}_A(\sigma) = 1$ if $\sigma = A$ and $\mathbf{1}_A(\sigma) = 0$ if $\sigma \neq A$).

Using the above abstraction, we can trivially extend the model to allow more than 2 states at each site, $\Omega = \{A, B, C, \ldots\}$. For example, we can express a Hamiltonian representing a simplified ternary alloy with nearest neighbor interactions as follows,

$$H(\boldsymbol{\sigma}) = -J \sum_{<ij>} \mathbf{1}_A(\sigma_i)\mathbf{1}_A(\sigma_j) + \mathbf{1}_B(\sigma_i)\mathbf{1}_B(\sigma_j) + \mathbf{1}_C(\sigma)\mathbf{1}_C(\sigma) \tag{1.15}$$

$$= -J \sum_{<ij>} \delta_{\sigma_i\sigma_j} \tag{1.16}$$

The above is clearly still a simplified model. We have set all interactions between like species to a single value parameter J and all interactions between different species to zero. This simplified model of a ternary alloy corresponds to the 3-state *standard Potts model*. The simple extension to q-states using the same Hamiltonian in Equation 1.16 is known as the standard q-state Potts model [250].

The standard q-state Potts model (and a related version the planar Potts model), although deceivingly simple, shows a variety of complex behavior based on the number of states q and the lattice symmetry that is still an active research area in statistical physics [13, 62, 125, 241]. Further extensions of the Hamiltonian in Equation 1.14 that allow longer range, multi-body interactions, and random coupling constants are the foundation of the modern field of *spin glasses* and disordered systems [19, 142, 173].

Generalized lattice models as coarse-grained models of atomic configuration

We have introduced three simple yet quintessential examples of classical lattice models, the Ising model, the lattice gas, and the Potts model; all of which can be used as models of

atomic configuration. In these models, and more generally in the study of statistical physics of lattice models, a specific form of the Hamiltonian is defined *a-priori*, and the focus is on rigorously studying the rich behavior of these predetermined lattice models. This has led to groundbreaking results and insights, particularly for the mathematical treatment of phase transitions, which have had far-reaching impacts beyond statistical physics [142]. However, the focus of the work presented here is to instead construct lattice models that approximate a Hamiltonian of atomic configuration for a real material as faithfully as possible. Doing so by only introducing constraints to the functional form of the Hamiltonian that can be formally justified (i.e. those required by physical symmetry). The final objective is to be able to use the fitted Hamiltonian to effectively compute thermodynamic properties of real materials.

We start by adopting a generalized form for a classical lattice model and generalize the geometry from a simple lattice to any crystallographic structure.[10] We also generalize the functional form of the Hamiltonian beyond just pair interactions and allow the Hamiltonian to be any possible function of configuration.[11] For this general case, we can intuitively write the Hamiltonian as an expansion in terms of multiple body interactions between species residing on the crystallographic sites,

$$H(\boldsymbol{\sigma}) = \sum_{S \in [N]} H_S(\boldsymbol{\sigma}_S) \tag{1.17}$$

where $\boldsymbol{\sigma}$ is a particular configuration specified by the species occupying each site. The sum runs over subsets of sites S, which we will call clusters. And $\boldsymbol{\sigma}_S = \{\sigma_i \ : \ \forall i \in S\}$ are the configuration variables for each of the sites in S. The interaction terms H_S can be any possible function of the configurations of cluster S. The most general form for a Hamiltonian as expressed in Equation 1.17 will include an interaction term for every possible cluster of sites,[12] however, for the majority of physical systems, the general locality of physical interactions usually requires only a subset of clusters with a small number of physically compact sites from all the possible clusters.

Lattice models as coarse-grained models for real materials

Using generalized lattice models to study properties of a real material constitutes a *coarse-graining* of the mechanical variables present in the full representation of the true physical Hamiltonian. The general Hamiltonian for a crystalline solid can be written as a function of the canonical coordinates of all valence electrons $\{\boldsymbol{p}_i, \boldsymbol{r}_i\}$ and ion cores $\{\boldsymbol{P}_i, \boldsymbol{R}_i\}$. Additionally, for multi-component systems, the chemical nature of each core $\{\sigma_i\}$ must be specified.

[10] That is to say we generalize to the case of a lattice with a basis.

[11] We will have more to say about the structure of both the space of configurations and the space of functions over them in Chapter 2.

[12] Formally the sum is over the elements of the *powerset* of the set of N sites. Which is a total of 2^N terms.

The Hamiltonian can be generally expressed in the following general form,

$$\mathcal{H} = \mathcal{H}(\{\boldsymbol{p}_i, \boldsymbol{r}_i\}, \{\boldsymbol{P}_i, \boldsymbol{R}_i, \sigma_i\}) \tag{1.18}$$

The possible system excitations for the Hamiltonian above can be divided into two categories based on their characteristic time scales [70]. Those involving canonical coordinates, which we label as *dynamic* excitations—these include electronic, magnetic, vibrational excitations—and those involving the possible configurations of the chemical nature of the ion cores which we call *static* excitations. Since the *static* excitations are considered to be much slower than any of the *dynamic* excitations, we can re-write the Hamiltonian above parametrized by configuration to a good approximation as follows,

$$\mathcal{H}(\{\boldsymbol{p}_i, \boldsymbol{r}_i\}, \{\boldsymbol{P}_i, \boldsymbol{R}_i, \sigma_i\}) = \mathcal{H}_{\boldsymbol{\sigma}}(\{\boldsymbol{p}_i, \boldsymbol{r}_i\}, \{\boldsymbol{P}_i, \boldsymbol{R}_i\}) \tag{1.19}$$

where the atomic configuration is given by the string $\boldsymbol{\sigma} = (\sigma_1, \sigma_2, \ldots, \sigma_i, \ldots)$ which specifies the chemical nature of the i-th ion core. We note that Equation 1.20 does not represent a *decoupling* of *dynamic* and *static* degrees of freedom, since the possible *dynamic* excitations still depend on the given configuration [70]. The form of the Hamiltonian given in Equation 1.19 is precisely what most standard *first-principles* methods are able to approximate to a very high degree of accuracy.

We will not discuss the effects of the ion core momenta $\{\boldsymbol{P}_i\}$[13], which requires an additional step involving phonon and/or force constant calculations. Although these vibrational effects can play critical roles in the thermodynamic properties of many materials [166, 167, 231], their treatment are beyond the scope of this dissertation. We will simply ignore the effects of the ion core momenta $\{\boldsymbol{R}_i\}$ by considering them to be static, or that they can be effectively parameterized for each configuration, such that Equation 1.19 simplifies to,

$$\mathcal{H}(\{\boldsymbol{p}_i, \boldsymbol{r}_i\}, \{\boldsymbol{P}_i, \boldsymbol{R}_i, \sigma_i\}) = \mathcal{H}(\{\boldsymbol{p}_i, \boldsymbol{r}_i\}, \{\boldsymbol{R}_i\}; \boldsymbol{\sigma}) \tag{1.20}$$

There is an important distinction between how the positions of the ion cores $\{\boldsymbol{R}_i\}$ are handled in the coarse-graining of the Hamiltonian in Equation 1.20. We can coarse grain the Hamiltonian by minimizing the electronic degrees of freedom while keeping the ion cores fixed (i.e. keeping the structure fixed),

$$H_{\boldsymbol{R}}(\boldsymbol{\sigma}) = \min_{\{\boldsymbol{r}_i, \boldsymbol{p}_i\}} \mathcal{H}(\{\boldsymbol{p}_i, \boldsymbol{r}_i\}; \boldsymbol{\sigma}) \tag{1.21}$$

where we have used the notation $H_{\boldsymbol{R}}$ to represent the coarse-grained Hamiltonian of atomic configuration for fixed positions of the ion cores $\{\boldsymbol{R}_i\}$, or in other words for a rigid lattice.

Alternatively, we can obtain a full coarse-graining by minimizing over the positions $\{\boldsymbol{R}_i\}$ of the ion cores as well (i.e. including structural relaxations),

$$H(\boldsymbol{\sigma}) = \min_{\{\boldsymbol{r}_i, \boldsymbol{p}_i\}, \{\boldsymbol{R}_i\}} \mathcal{H}(\{\boldsymbol{p}_i, \boldsymbol{r}_i\}, \{\boldsymbol{R}_i\}; \boldsymbol{\sigma}) \tag{1.22}$$

[13] These are themselves coarse-grained in many electronic structure calculations following the Born-Oppenheimer approximation [47, 203].

Both Equation 1.21 and Equation 1.22 represent a perfectly valid starting point for fitting applied lattice models to subsequently approximate thermodynamic properties that depend on the atomic configuration of real materials. However, both approaches come with very different fundamental and practical implications.

1. Using Equation 1.21 one obtains a formal separation of the structural (the positions of the ion cores $\{\boldsymbol{R}_i\}$) and the configurational (the atomic species at each site $\boldsymbol{\sigma}$) degrees of freedom. This separation results in expansions based on Equation 1.17 that are of the same order as the underlying full Hamiltonian 1.20[14]. As a result, obtaining *well converged*[15] lattice models usually require fewer expansion terms and less training data. However, since the final thermodynamic properties can be strongly influenced by structural relaxations, one would need to incorporate these into the model, which at least in a brute force manner, would require several lattice models fitted to the configurational energy of different sets of structural parameters.

2. Using Equation 1.22 breaks the formal separation of structural and configurational degrees of freedom [189], which may require expansions with terms of degree beyond the order of the real Hamiltonian 1.20 to fully capture the influence of structural relaxations on the configurational energy landscape. However, doing so allows one to construct a lattice model that represents the ground state energy landscape as a function of configuration. The end result is that computing final thermodynamic properties of configuration can require only a single lattice model.

In the remainder of this work we will focus mostly on the second approach because it is now the predominant approach for such calculations in materials research as it provides a much more direct route to computing final thermodynamic properties, albeit at the cost of having to put in a little more effort to construct sufficiently accurate models [212, 231].

1.4 Thermodynamic ensembles of generalized lattice models

Finally, we briefly describe the link between the statistical thermodynamics presented and the framework we explore for representing generalized lattice models. In order to carry out statistical thermodynamic calculations of systems with configurational degrees of freedom using a generalized lattice model, it is particularly useful to work with a *semi-grand canonical* (SGC) ensemble. The SGC ensemble proves to be computationally simpler to implement compared to other ensembles [72]. For example, Monte Carlo steps have fewer constraints in the SGC ensemble compared to a canonical ensemble. Additionally, there is no need to deal

[14] A proof of this is given in Appendix B.1.

[15] This broadly means high predictive accuracy with a general trend of expansion coefficients decaying with respect to the physical distance between sites and the number of interacting sites involved.

with the intricacies of particle addition and removal when compared to a grand canonical formalism. And unlike the grand canonical ensemble, an isobaric version of the SGC ensemble formally exists [119]. Finally, and most importantly, many of the mathematical properties of the representations of lattice Hamiltonians that we develop in this work only rigorously hold in an SGC ensemble [187, 189].

To derive the formal expression of the pertinent SGC ensemble, we write the Helmholtz free energy differential of a multi-component system with N total species and L substructures each with n_l allowed species at each site as follows [119],[16]

$$dA = -SdT + \sum_{l=1}^{L}\sum_{i=1}^{n_l} \mu_i^{(l)} dN_i^{(l)} \tag{1.23}$$

where for now we broadly define a substructure as the set of all sites that have the same set of allowed species. By choosing a reference species for each substructure, Equation 1.23 can be written as,

$$dA = -SdT + \sum_{l=1}^{L}\left(\mu_r^{(l)} + \sum_{i=1}^{n_l} \bar{\mu}_i^{(l)} dN_i^{(l)}\right) \tag{1.24}$$

where the chemical potential differences are given as $\bar{\mu}_i = \mu_i - \mu_r$. The total number of sites in each substructure is fixed, $N^{(l)} = \sum_i^{n_l} N_i^{(l)}$, and so is the total number of sites in the structure, $N = \sum_l^L N^{(l)}$.

We obtain the semi-grand potential as a Legendre transformation of the free energy in Equation 1.24,[17]

$$Y = A - \sum_{l=1}^{L}\sum_{i=1}^{n_l} \bar{\mu}_i^{(l)} N_i^{(l)} = \sum_{l=1}^{L} \mu_r^{(l)} N^{(l)} \tag{1.25}$$

The SGC partition function can be written explicitly by summing over all possible atomic configurations $\boldsymbol{\sigma}$,

$$Z(\beta, \bar{\boldsymbol{\mu}}) = \sum_{\boldsymbol{\sigma}} \exp\left(\beta \sum_{l=1}^{L}\sum_{i=1}^{n_l} \bar{\mu}_i^{(l)} N_i^{(l)}(\boldsymbol{\sigma})\right) Z_{\boldsymbol{\sigma}} \tag{1.26}$$

where the sum over configurations represents N total sums, one for each lattice site. The number of species i in the l-th substructure can be represented as a function of the configuration, $N_i^{(l)}(\boldsymbol{\sigma})$. The canonical partition function for a given configuration is $Z_{\boldsymbol{\sigma}} = \mathrm{tr} e^{-\beta\mathcal{H}}$, where the trace operation is taken over all *dynamic* degrees of freedom, i.e. all canonical coordinates as done in Equation 1.22.

[16] Where we are ignoring any nuances regarding PV pressure/volume variables, which for cases considered for 1 atmosphere of pressure can be practically justified by the negligible contribution that changes in volume will have to the total energy of crystalline solids [55].

[17] One must be careful when the same chemical species is allowed in more than one substructure. In those cases, equilibrium requires that the absolute chemical potentials for the same species in the different substructures be equal.

Finally, we introduce the fugacities, $z_i = e^{\beta\mu_i}$ and the fugacity fractions, $\xi_i = z_i / \sum_i^n z_i$ as a direct link to the *a-priori* probability measure that we will use in our development of function spaces over configuration spaces.[18] We re-write the SGC partition function in terms of fugacity fractions accordingly,

$$\begin{aligned} Z(\beta, \bar{\boldsymbol{\mu}}) &= \prod_l^L \left(\xi_r^{(l)}\right)^{-N^{(l)}} \sum_{\boldsymbol{\sigma}} \prod_l^L \prod_i^{n_l} \left(\xi_i^{(l)}\right)^{N_i^{(l)}(\boldsymbol{\sigma})} Z_{\boldsymbol{\sigma}} \\ &= \sum_{\boldsymbol{\sigma}} \prod_l^L \prod_i^{n_l} \rho_i^{(l)}(\boldsymbol{\sigma}) Z_{\boldsymbol{\sigma}} \qquad (1.27) \\ &= \sum_{\boldsymbol{\sigma}} \boldsymbol{\rho}(\boldsymbol{\sigma}) Z_{\boldsymbol{\sigma}} \qquad (1.28) \end{aligned}$$

where we have dropped the constant term involving a product over reference fugacity fractions, and have introduced the compact notation $\boldsymbol{\rho}(\boldsymbol{\sigma})$ to represent the product of all involved fugacity fractions.

Furthermore, to tractably compute thermodynamic properties that depend on configurational degrees of freedom, we can approximate the canonical partition function using the maximum term method[14] as follows,

$$Z_{\boldsymbol{\sigma}} \approx e^{-\beta H(\boldsymbol{\sigma})} \qquad (1.29)$$

where the Hamiltonian, $H_{\boldsymbol{\sigma}}$ is the minimum of Equation 1.24 over canonical coordinates for the corresponding configuration $\boldsymbol{\sigma}$. With this approximation, which corresponds to the coarse-grained Hamiltonian given in Equation 1.22, the SGC partition function in Equation 1.28 requires only a sum over all configurations,

$$Z(\beta, \boldsymbol{\rho}) = \sum_{\boldsymbol{\sigma}} \boldsymbol{\rho}(\boldsymbol{\sigma}) e^{-\beta H(\boldsymbol{\sigma})} \qquad (1.30)$$

And the associated Boltzmann distribution based on the SGC partition function above is given by,

$$\mathbb{P}(\boldsymbol{\sigma}) = \frac{\boldsymbol{\rho}(\boldsymbol{\sigma}) e^{-\beta H(\boldsymbol{\sigma})}}{\sum_{\boldsymbol{\sigma}} \boldsymbol{\rho}(\boldsymbol{\sigma}) e^{-\beta H(\boldsymbol{\sigma})}} \qquad (1.31)$$

It is within the formalism of the SGC ensemble distribution given in Equation 1.31 that the development of the mathematical framework to represent generalized lattice models will be carried out in this work. That isn't to say that once a Hamiltonian $H(\boldsymbol{\sigma})$ is obtained, one can only calculate statistical thermodynamic properties in the SGC ensemble. In practice, it is perfectly valid, even trivial, to use $H(\boldsymbol{\sigma})$ to define a canonical ensemble. However many mathematical properties and resulting interpretations of the structure of $H(\boldsymbol{\sigma})$ only rigorously hold in an SGC ensemble.

[18] Additionally, it is practically useful to work with fugacity fractions since their values are bounded, $\xi_i \in [0, 1]$, are equal to zero when the corresponding species is absent from the substructure and equal to one when the corresponding species is the only species present on the substructure.

1.5 Thesis overview

We have so far presented the technological and scientific motivation to develop methods that enable the computational study of atomic disorder in multi-component or multi-principal element materials. In addition, we gave a brief exposition of the basic fundamental concepts required to carry out statistical thermodynamic calculations for systems with configurational degrees of freedom. Finally, we motivated lattice models as a compelling, natural, and effective framework to carry out such calculations in practice.

The rest of this thesis deals with developing a mathematical framework to construct generalized Lattice models that can be used to represent the energy in terms of atomic configurations by including interactions of any range and between any number of species. Formally the framework involves defining mathematical representations for any multi-variate function of discrete states that is invariant under the pertinent crystallographic symmetries. The specific representations developed can be seen as direct generalizations of the Ising and Potts models. We subsequently use the mathematical structure of these representations to motivate, develop and justify linear regression estimation methods that allow efficient parametrization of lattice models using first-principles electronic structure calculations. A set of practical training data sampling and preparation methods are also presented, along with select examples of fitted Hamiltonians for technologically relevant multi-component ionic materials and medium entropy alloys. Finally, we conclude this thesis by presenting open areas in which the work presented here can be used in novel ways to study complex multi-component materials.

The work is organized into the following Chapters:

- Chapter 2 develops the formal representation of generalized Hamiltonians as symmetrically invariant functions of a multicomponent crystal's atomic configuration. Two different representations are developed: Fourier cluster expansions and Potts frame expansions. Fourier cluster expansions rely on orthonormal basis sets (as was originally proposed in the cluster expansion method [192]), and can be seen as a generalization of the Ising model. Fourier cluster expansions are then re-cast in a unique basis-agnostic expansion that we call the *cluster decomposition*, which allows formal interpretation of expansion terms. A Potts frame expansion is a direct generalization of the Potts model to arbitrary interactions. Potts frame representations constitute a redundant expansion which results in particularly useful properties for robust and accurate reconstruction from limited data.

- Chapter 3 presents practical ways to numerically implement the mathematical framework developed in Chapter 2. In addition, procedures to convert any representation to a Fourier cluster expansion to enable model interpretation, as well as procedures to directly compute cluster occupation averages and probabilities from expansion function values are derived. All methodology described has been implemented and is openly available [11].

- In Chapter 4 novel parameter estimation models are formulated to fit generalized lattice Hamiltonians using data first principles calculated data. Structural priors of expansion coefficients are derived and justified based on the representations established in Chapter 2. Novel regularized linear regression models that result in parameters that satisfy these structural priors are presented. Finally, procedures to appropriately construct generalized Lattice models mixed with empirical pair potentials, are described and motivated for handling long-range electrostatic interactions in ionic materials.

- Chapter 5 describes training data sampling and data preparation methods necessary to parametrize lattice Hamiltonians of complex materials with large numbers of allowed species. In addition, applications of the methods developed are presented for lattice models of several Li-transition metal-OF and NiCoCr alloys.

- Chapter 6 concludes this thesis. A handful of open areas to extend and improve methods and applications are suggested. Specifically, new lattice model learning paradigms, improved special structure generation, ground state identification and methods for optimization, and machine learning-based thermodynamic inference are suggested.

Chapter 2

Representation of generalized lattice models

In this chapter, we give a detailed exposition of the mathematical *representation* of lattice models. By *representation* we mean the framework necessary to precisely specify a class of object. The objects we aim to specify are symmetrically invariant functions of configuration. In order to do so, we will first precisely define the *configuration spaces* which compromise the domain of said functions. Afterward, we will formally describe the *function space* in which the Hamiltonian functions we seek are found. Finally, in order to represent a Hamiltonian for real applications, we construct *spanning sets* of functions that allow us to expand any such Hamiltonian. Specifically, we construct a particular class of *basis* sets, in addition to a particular *frame* that span the aforementioned function space.

Succinctly, the subject of this chapter is then to formulate representations for a Hamiltonian function H with domain $\mathbf{\Omega}$ (a configuration space) and codomain $\mathbb{R}$,

$$H : \mathbf{\Omega} \to \mathbb{R} \tag{2.1}$$

The endeavor to specify representations of Hamiltonians of configuration lies in the realm of functional analysis[5, 107, 243]. More specifically, for the subject matter here, it is the realm of discrete harmonic analysis [30], and finite frame theory [43, 235]. Functional and harmonic analysis and frame theory are rich and extensive fields, but for our purposes, we only use a handful of basic concepts and results.

In this chapter, we focus on carefully defining what is meant by a configuration space $\mathbf{\Omega}$ in terms of a probability product space of individual elementary probability spaces associated with degrees of freedom of each crystallographic site, i.e. the allowed species and an associated *a priori* probability. We also briefly discuss configuration spaces with composition constraints, which are necessary to work with ionic structures where only charge-neutral configurations are of interest. We then move to formally describe the function space that contains Hamiltonians of configuration H. Finally, we construct two general forms to represent Hamiltonians of configuration and describe some of their mathematical properties.

Specifically, we treat the following representations:

- Fourier cluster expansions, which are constructed using an orthonormal basis and thus allow several useful properties to compute norms, means, and variances.
  - The cluster decomposition, which is essentially a re-expression of Fourier cluster expansions that yields a basis agnostic and unique representation that in turn allows further insight and interpretation of expansion terms.
- The generalized Potts frame, which is a redundant representation as an intuitive generalization of the Potts model, and allows both practical and robust learning even for cases of insufficiently sampled training data.

Preliminary notation conventions and auxiliary definitions that we use in this and following Chapters are detailed in Appendix A.

2.1 Configuration spaces

The *configurations* we are concerned with here are all the possible decorations or colorings of the sites in a given crystal structure constructed by choosing a specific species from a set of allowed species at each site. The set of all the possible decorations of the sites in a crystal structure makes up a *configuration space*. For example, in the Ising model, the configuration space is relatively simple since there are only two choices of allowed species (or spins) for any site. The total number of configurations for a lattice with N sites is, therefore, 2^N[1]. The number of configurations for a Potts model is not much different, we simply need to change the base 2 to the number n of allowed species, meaning there are n^N configurations.

In both the Ising and Potts models, all sites share the same set of allowed species. In more complex crystal structures, it is not uncommon to have sites with distinct sets of allowed species[2]. In order to accurately describe these situations, we will formalize (and generalize) the sets of allowed species to take the form of a *discrete probability space*[3], which we call a *site space* Ω.

Definition 2.1.1 (Site Space)**.** *A site space is a discrete probability space where the sample space is the set of the species allowed at a given site* $\vec{p}$*,*[4]

$$(\Omega, \rho)_{\vec{p}} = (\{\sigma_j, \rho_j : \ \textit{for } j = 1, \ldots, n\},)_{\vec{p}}$$

[1] Although many of these configurations can be symmetrically equivalent based on the geometry/lattice used.

[2] An example arising in practice, involves interstitial defects in a metal where the set of allowed species at interstitial sites is different than that at regular sites. Another simple example is cation and anion sites in ionic materials.

[3] For our purposes we always assume that the event space includes only elementary events.

[4] A (crystallographic) site can be specified by a point $\vec{p} \in \mathbb{R}^3$ where one of a set of allowed chemical species may reside. When dealing with materials of other dimensions sites are points in the corresponding space $\mathbb{R}^n$—i.e. $\vec{p} \in \mathbb{R}^2$ for 2D materials.

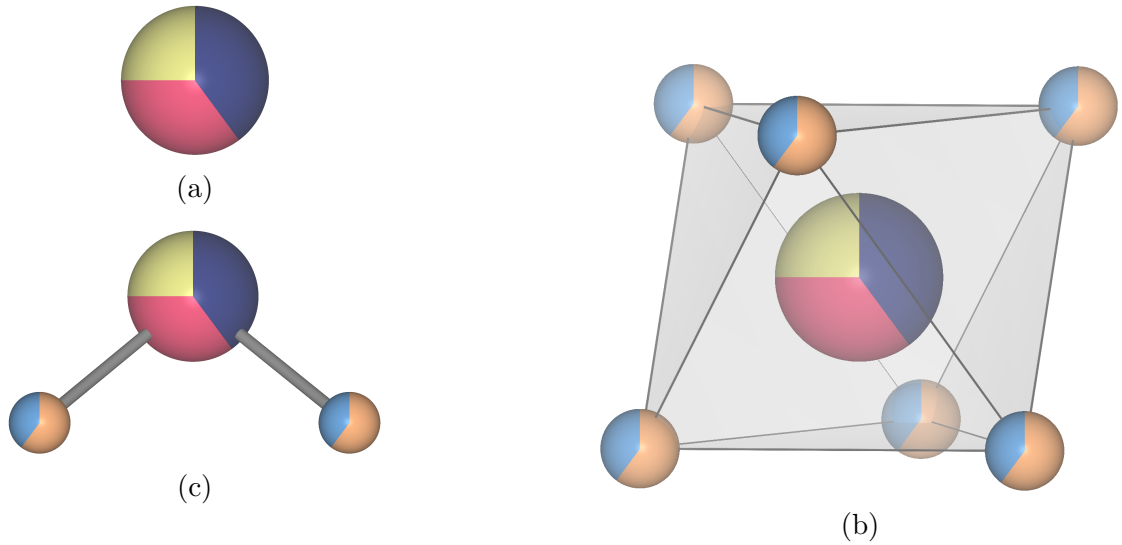


Figure 2.1: (a) Visual representation of a ternary *site space*. The colors represent the different species and the relative fractions colored represent the values of the *a-priori* measure (species concentrations). (b) An example of a rocksalt *disordered crystal structure* with a basis including two different site spaces. (c) An example of a triplet *site space cluster*.

Each occupation variable σ_j is a specific chemical species. This is most often an elemental species or ion, but in the general case can be polyatomic. $\rho(\sigma_j) = \rho_j$ is an apriori measure for the probabilities corresponding to each species. The size of a site space is the total number of allowed species $n = |\Omega|$.[5]

The *a priori* measure ρ describes the state over a particular site for a non-interacting system, or in other words, in the *random limit*. This is to say, we can think of ρ as specifying the concentrations of species, and the probability of finding a particular species at the i-th site at very high temperatures $T \to \infty$. Figure 2.1a shows a graphical depiction of a ternary site space. The different colors represent the different allowed species, and their relative proportions represent the *a-priori* measure. For the most part we will assume the measure ρ, of a site space is uniform[6]; and simply state that a site space is $\Omega_{\vec{p}} = \{\sigma_j, j = 1, \dots |\Omega|\}$.

Remark 2.1.1 (Categorical occupation variables)**.** *In the present work, the occupation variables σ are considered categorical variables as opposed to numerical variables. Meaning they directly represent a chemical species or ions, e.g. $\sigma = Li^+$, in contrast to numerical* spin *or* occupation *variables, as used in the original Ising model $\sigma = \pm 1$ or lattice gas models $\sigma = 0, 1$.*[7]

[5] We call site space with only one species, i.e. $n = |\Omega| = 1$, *inactive*, since it does not have compositional degrees of freedom.

[6] $\rho_j = 1/n$ for all j

[7] This may seem esoteric, and more so considering that when implementing the method we in fact need

By assigning site spaces instead of specific species to the crystallographic sites in a structure, we construct what we call a *disordered crystal structure*.

Definition 2.1.2 (Disordered Crystal Structure)**.** *A disordered crystal structure is a crystal structure defined by a* Bravais *lattice and a crystallographic basis specifying the location of all* site spaces *within a crystallographic unit cell.*

Figure 2.1b shows the unit cell of a rocksalt disordered crystal structure. In the example shown, the crystallographic basis includes two different site spaces.

Using the definition of a disordered crystal structure, we can define *site space clusters*, which will be a necessary concept in the coming sections when we construct spanning functions whose domains are essentially the configuration spaces of site space clusters only.

Definition 2.1.3 (Site Space Cluster)**.** *A site space cluster* $A = \{(\Omega, \rho)_{\vec{p}_i} \ : \ i = 1, \ldots, N_A\}$ *is a set of site spaces. The size of the cluster is the number of sites spaces in the cluster,* $N_A = |A|$.[8]

Figure 2.1c shows an example of a triplet site space cluster. In the case shown, there are three site spaces, two of which are equivalent in the sense that they have the same set of species and concentrations but are associated with different sites. In the remainder of this work, we may use *site cluster* interchangeably to refer to a *site space cluster*, making it implicit that the sites have associated site spaces.

Finally, a *configuration space* is a probability product space, constructed from the product of all site spaces in a disordered crystal structure.

Definition 2.1.4 (Configuration space)**.** *A configuration space* $(\boldsymbol{\Omega}, \boldsymbol{\rho})$ *is a probability product space of* N *site spaces. Where the set* $\boldsymbol{\Omega}$ *is the Cartesian product of each* Ω_i *and* $\boldsymbol{\rho}$ *is a corresponding product measure.*

$$\boldsymbol{\Omega} = \bigtimes_{i=1}^{N} \Omega_i \tag{2.2}$$

$$\boldsymbol{\rho}(\boldsymbol{\sigma}) = \bigtimes_{i=1}^{N} \rho_i(\sigma_i) \tag{2.3}$$

In the notation above we have dropped the explicit reference to the position of the sites by making it implicit that the i*-th site has an associated position* $\vec{p}_i$*. For compactness, we will continue to use this notation relying only on the site indices* i*.*

to use numbers to *encode* the actual species. However, the choice of encoding depends on the details of the implementation, so treating σ as a number in our analysis only obfuscates its real meaning in representing a particular chemical species.

[8]The empty set with $N_A = 0$ and the set of all sites $N_A \to \infty$ are well-defined clusters.

The size of a configuration space $(\boldsymbol{\Omega}, \boldsymbol{\rho})$ is equal the total number of possible configurations, $|\boldsymbol{\Omega}| = \prod_i^N |\Omega_i|$, and the dimension of the configuration spaces is equal to the number of sites N considered. In theory, for bulk structures, we take $N \to \infty$, but for all practical purposes, we can think of N as large *enough* and use periodic boundary conditions.

A configuration space $\boldsymbol{\Omega}$ can be represented as a *hypergrid*. Figure 2.2 shows the configuration space for a hypothetical finite three site structure with two ternary sites (A_1, A_2) and one binary site (B). The configuration space is depicted as the representative *disordered structure*, and the corresponding three-dimensional configuration grid. Increasing the number of allowed species at a site translates to adding vertices to the hypergrid along the dimension corresponding to said site. Adding more sites to the structure increases the dimension of the hypergrid since the dimension of the configuration space is equal to the number of sites in the given system. For a structure with N sites, the configuration space corresponds to a hypergrid in N dimensions.

Although a hypergrid is the formal mathematical representation of a configuration space, it is also useful to simply represent a configuration space using the underlying disordered crystal structure only. Specific configurations (vertices in the hypergrid) are all the possible *decorations* of the N sites with a specific species chosen from their corresponding site spaces. Furthermore, to simplify matters further, only a small set of distinct site spaces are needed to define a configuration space for applications materials science. Such that the product

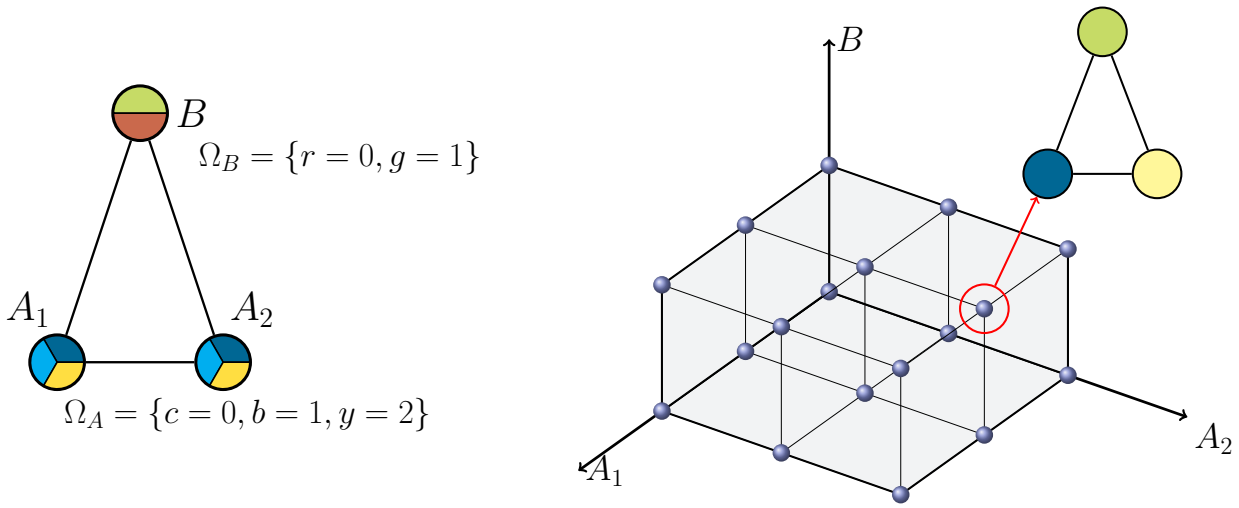
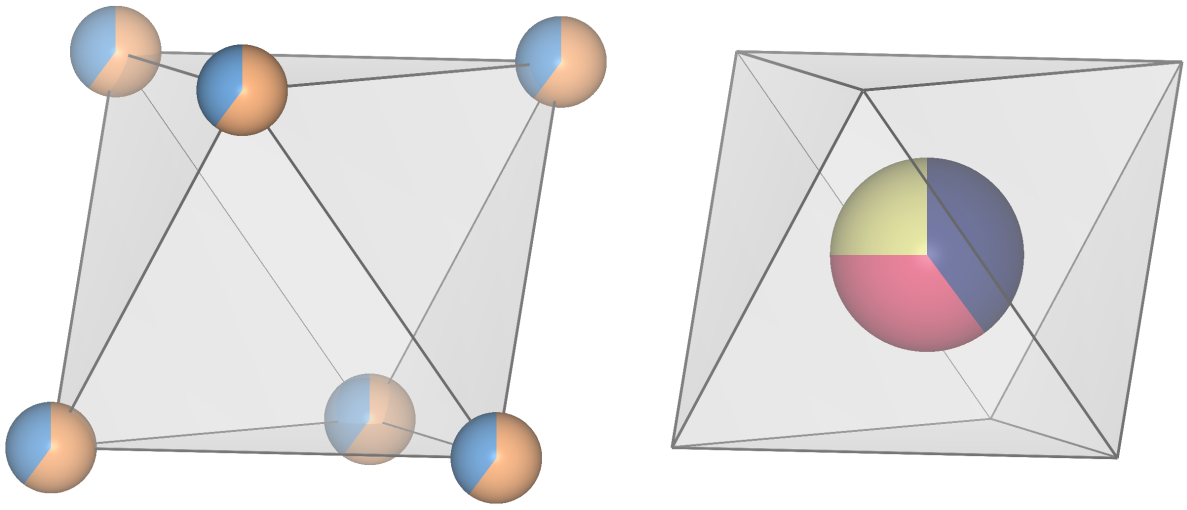


Figure 2.2: Illustration of the configuration space as a hypergrid for a disordered structure illustrated by the triangular figure shown on the left. The structure has two ternary sites A_1, A_2 where the allowed species are represented by the colors green (g) and red (r); and one binary site B with allowed species, cyan (t), blue (b), and yellow (y). The vertex of the hypergrid corresponding to specific configuration, $\boldsymbol{\sigma}$ = (blue b=1, yellow y=2, green g=1), is pointed out as an example of a point in the (A_1, A_2, B) configuration space.



(a) (b)

Figure 2.3: Primitive cells of two *disordered substructures* in the disordered structure shown in Figure 2.1b.

space in Equation 2.2 can be more compactly written as products over distinct site spaces only,

$$\mathbf{\Omega} = \bigtimes_{l=1}^{L} \Omega_l^{N_l} \tag{2.4}$$

where L is the total number of distinct site spaces, and N_l is the number of sites with site space Ω_l.

The sets of sites with the same associated site spaces taken together can be used to define *substructures* of the disordered structure.

Definition 2.1.5 (Disordered substructure)**.** *Given a disordered structure with $L > 1$ distinct site spaces, a disordered substructure is a crystallographic orbit generated by sites associated with a subset L_s of the L distinct site spaces such that $1 \leq L_s < L$.*

Figure 2.3 shows the two distinct disordered substructures for the disordered substructure shown in Figure 2.1b.

In the majority of the cluster expansion literature the *disordered structure* is referred to as the *lattice*, and *disordered substructures* are referred to as *sublattices*. However, the use of both words in the context of more complex crystallographic structures (i.e. those with a two or more atom basis) does not follow their definition from crystallography. For this reason, we will use *disordered structure* and *substructure* as previously defined to adhere to their crystallographic definitions [154].

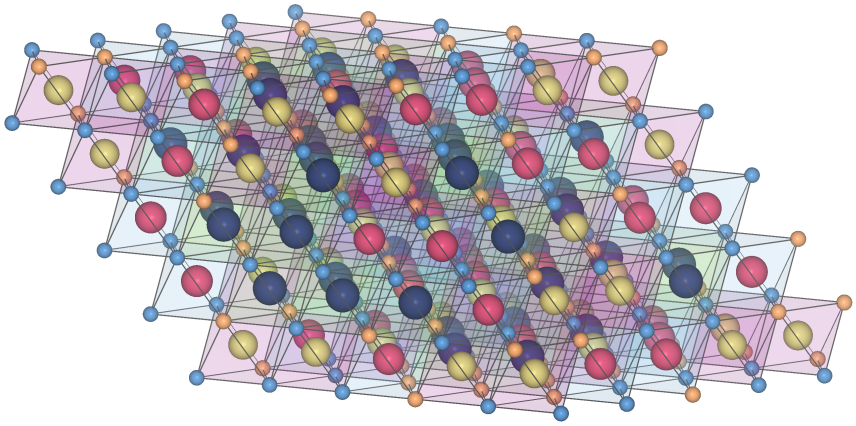


Figure 2.4: An example *ordered structure* corresponding to a specific configuration $\boldsymbol{\sigma} \in \boldsymbol{\Omega}$ for the disordered structure shown in Figure 2.1b.

Ionic materials are straightforward examples of structures that involve multiple disordered substructures (i.e., those corresponding to distinct anion and cation substructures). For example, in an ionic system with one anion and one cation disordered substructure, the configuration space can be expressed as,

$$\begin{aligned} \boldsymbol{\Omega} &= \left(\bigtimes_{i \in A} \Omega_i \right) \times \left(\bigtimes_{i \in C} \Omega_i \right) \\ &= \Omega_A^{N_A} \times \Omega_C^{N_C} = \boldsymbol{\Omega}_A \times \boldsymbol{\Omega}_C \end{aligned} \tag{2.5}$$

where A and C denote the sites in the anion and cation substructure respectively, and N_A and N_C refer to the number of anion and cation sites in their respective substructures.

Finally, a specific configuration is represented by an *occupancy string*, which is an element of a configuration space $\boldsymbol{\sigma} \in \boldsymbol{\Omega}$. The configuration $\boldsymbol{\sigma}$ represents an *ordered structure*, i.e. a crystal structure with a specific species residing at each site. Meaning, for each site i in the structure, we pick a specific species from its associated site spaces Ω_i.

Definition 2.1.6 (Occupancy string)**.** *An occupancy string is an element of a configuration space $\boldsymbol{\sigma} \in \boldsymbol{\Omega}$. More explicitly it is a string of occupation variables where each occupation variable is an element of the corresponding site space $\sigma_i \in \Omega_i$,*

$$\boldsymbol{\sigma} = (\sigma_1, \sigma_2, \ldots, \sigma_N \mid \sigma_i \in \Omega_i \text{ for } i = 1, \ldots, N) \tag{2.6}$$

Figure 2.4 shows the ordered structure for a given configuration $\boldsymbol{\sigma}$ with $N = 64$ sites. In the figure, we assume periodic boundary conditions, so that the structure is representative of a bulk material.

Composition constraints

The configuration space for ionic materials, in particular, requires introducing the concept of *composition constraints* on a configuration space. We will specifically focus on composition constraints necessary to guarantee charge-neutral configurations of ionic materials. Because the species in ionic materials are charged ions, the configurations that are considered for the vast majority of practical applications are *charge neutral* configurations only. Charge neutrality constraints on composition can be expressed as a sum of the oxidation states associated with the species in a configuration string in the following manner,

$$\sum_{\sigma_i \in \boldsymbol{\sigma}} z(\sigma_i) = 0 \tag{2.7}$$

where $z(\sigma_i)$ represents a mapping from the species represented by the occupation variable σ_i to its corresponding oxidation state.

Although Equation 2.7 is straightforward, it has important repercussions in both the formalism for representing functions of configuration and in its practical application to ionic material systems with heterovalent ions. Composition constraints formally change the domain of valid configurations, such that the function space over the configurations of a heterovalent ionic material system is not the same as the product configuration space $\boldsymbol{\Omega}$ introduced in Equation 2.2. More specifically, the configuration space $\hat{\boldsymbol{\Omega}}$ of an ionic material system is a set of *slices* of the product configuration space—or equivalently a set of *slices of the configuration hypergrid*, given as,

$$\hat{\boldsymbol{\Omega}} = \left\{ \boldsymbol{\sigma} \in \boldsymbol{\Omega} : \sum_{\sigma_i \in \boldsymbol{\sigma}} z(\sigma_i) = 0 \right\} \tag{2.8}$$

By construction $|\hat{\boldsymbol{\Omega}}| \leq |\boldsymbol{\Omega}|$. The constrained space $\hat{\boldsymbol{\Omega}}$ is equal to the unconstrained space only for cases where all species associated with each substructure in the system are iso-valent, and thus every point in the full configuration space is charge neutral.

Figure 2.5 shows an example of the configuration space with composition constraints for the previously introduced three-site system from Figure 2.2. The composition constraints considered in the example reduce the total number of configurations in the unconstrained configuration space from 18 configurations to only 8 charge-neutral configurations. Additionally, since the configuration can be expressed by explicitly specifying the occupation of 2 out of the 3 sites, the configuration grid can be represented in a lower dimension as shown. These observations extend to higher dimensional configuration spaces such that when charge neutrality constraints are considered, the dimensionality of the constrained space is effectively reduced, and the total number of configurations is in most cases substantially reduced.

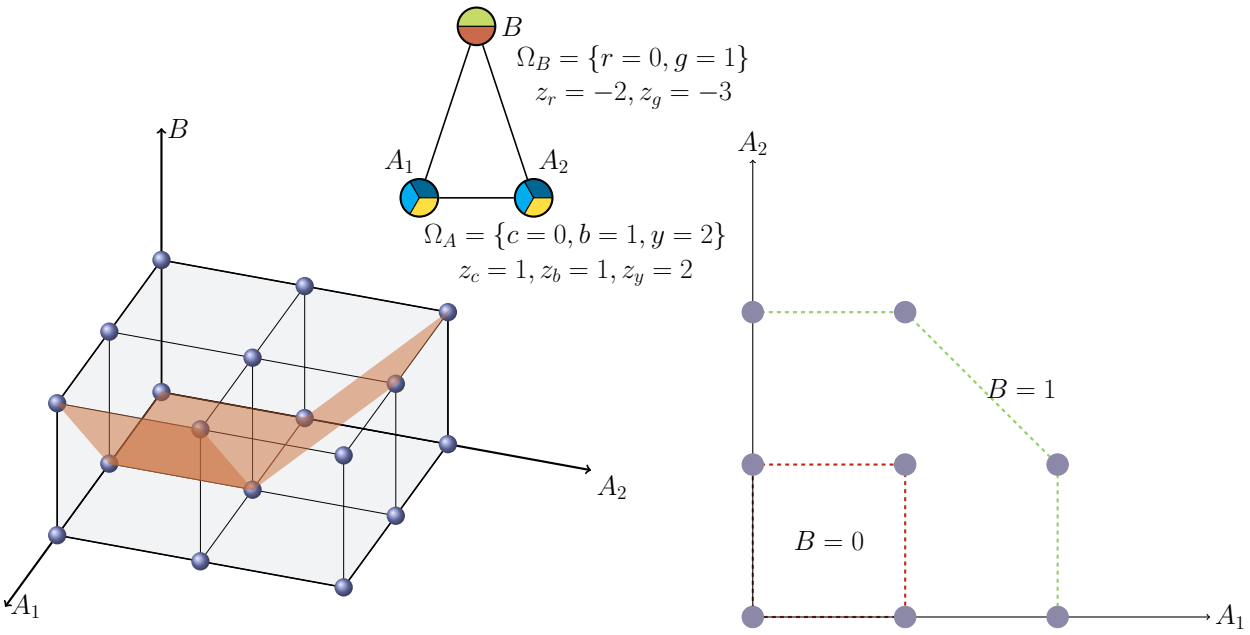


Figure 2.5: Illustration of the configuration space as a slice of the hypergrid for the triangular figure shown on the top. The figure has two ternary sites A_1, A_2 and one binary site B, and the labeled ternary sites and the binary site have positive and negative oxidation states respectively. The charge neutral slice of the original unconstrained 3D grid is shown as the grid points intersected by the orange planes. The two-dimensional figures depict the constrained space where the occupation of the binary site is implicit given the occupations of the two ternary sites based on charge neutrality constraints.

2.2 Function spaces over configuration spaces

Having carefully defined what a *configuration space* is and how it relates to a *disordered crystal structure*, we now move on to define and subsequently develop representations for the *function spaces* over configurations. Following the definition of a configuration space, the function spaces we will define are essentially function spaces over *probability product spaces* [6, 157, 214], and so the product structure of configuration spaces extends in an analogous fashion to a product structure of the function spaces over them. Following a similar recipe as before we start by defining a function space over a single site space.

Definition 2.2.1 (Function space over a site space)**.** *The function space $L^2(\Omega, \rho)$ over a site space (Ω, ρ) is the space of all functions $f : \Omega \to \mathbb{R}$, such that $\langle f^2 \rangle_\rho < \infty$. Where the inner product $\langle f, g \rangle_\rho$ is defined as follows,*

$$\langle f, g \rangle_\rho = \sum_{\sigma \in \Omega} \rho(\sigma) f(\sigma) g(\sigma) \tag{2.9}$$

The function space $L^2(\Omega, \rho)$ is a Hilbert space of dimension: $\dim L^2(\Omega, \rho) = |\Omega|$ [157, 243]. The inner product in Equation 2.9 has the following probabilistic interpretation when f and g are considered random variables[9],

$$\langle f, g \rangle_\rho = \mathbb{E}_\rho [fg] \tag{2.10}$$

In other words, the inner product in Equation 2.9 can be thought of as an expectation over the probability distribution with probability mass function $\rho(\sigma_i)$. With this interpretation, we recognize that the inner product with the constant function 1, represents the expectation value of a function: $\langle f, 1 \rangle_0 = \mathbb{E}[f]$. We will continue to use the bracket notation, with the underlying understanding that it represents an expectation under the *a-priori* distribution ρ. Furthermore for the majority of the work presented we take ρ to be the uniform distribution, such that the inner product in Equation 2.9 becomes,

$$\langle f, g \rangle_\rho = \frac{1}{|\Omega|} \sum_{\sigma_i \in \Omega} f(\sigma_i) g(\sigma_i) \tag{2.11}$$

Using Definition 2.2.1, we can formally define the function space over configuration space as the tensor product of function spaces over site spaces.

Definition 2.2.2 (Function space over configuration space)**.** *The function space over a configuration space, $L^2(\boldsymbol{\Omega}, \boldsymbol{\rho})$ is the tensor product space taken over all function spaces $L^2(\Omega, \rho)_i$ for $i \in [N]$ that make up the configuration space $\boldsymbol{\Omega}$,*

$$L^2(\boldsymbol{\Omega}, \boldsymbol{\rho}) = \bigotimes_i^N L^2(\Omega, \rho)_i \tag{2.12}$$

Now, $L^2(\boldsymbol{\Omega}, \boldsymbol{\rho})$ is itself a Hilbert space of dimension[10] $\dim L^2(\boldsymbol{\Omega}, \boldsymbol{\rho}) = |\boldsymbol{\Omega}|$. An equivalent way to define $L^2(\boldsymbol{\Omega}, \boldsymbol{\rho})$ such that its nature as a Hilbert space is more explicit is as follows,

$$L^2(\boldsymbol{\Omega}, \boldsymbol{\rho}) = \{F : \boldsymbol{\Omega} \to \mathbb{R}, \langle F^2 \rangle_{\boldsymbol{\rho}} < \infty\} \tag{2.13}$$

Where the inner product is defined as,

$$\langle F, G \rangle_{\boldsymbol{\rho}} = \sum_{\boldsymbol{\sigma} \in \boldsymbol{\Omega}} \boldsymbol{\rho}(\boldsymbol{\sigma}) F(\boldsymbol{\sigma}) G(\boldsymbol{\sigma}) \tag{2.14}$$

And for the frequent case where we take the distribution $\boldsymbol{\rho}$ to be uniform, the inner product becomes,

$$\langle F, G \rangle_{\boldsymbol{\rho}} = \frac{1}{|\boldsymbol{\Omega}|} \sum_{\boldsymbol{\sigma} \in \boldsymbol{\Omega}} F(\boldsymbol{\sigma}) G(\boldsymbol{\sigma}) \tag{2.15}$$

[9] Actually they are functions of the random variable s.

[10] Basically the dimension of $L^2(\boldsymbol{\Omega}, \boldsymbol{\rho})$ is equal to the total number of configurations in $\boldsymbol{\Omega}$.

Similarly, we can extend the previous probabilistic interpretation to the inner product in Equation 2.14,

$$\langle F, G \rangle_{\boldsymbol{\rho}} = \mathbb{E}_{\boldsymbol{\rho}} [FG] \tag{2.16}$$

With this, we can interpret the inner product of a function of configuration with the constant function 1, as the expectation value of that function over the *a priori* product distribution $\boldsymbol{\rho}$: $\langle F, 1 \rangle_{\boldsymbol{\rho}} = \mathbb{E}_{\boldsymbol{\rho}} [F]$. Furthermore, considering that $\boldsymbol{\rho}$ represents the probability in the non-interacting limit, we can further interpret $\langle F, 1 \rangle_{\boldsymbol{\rho}}$ as the expectation value of F in this limit. This interpretation is then in direct accordance with the high-temperature limit of the Boltzmann probability for the semi-grand canonical ensemble given in Equation 1.31 which was derived in Chapter 1.

The inner product in Equation 2.15 is a generalization of the inner product used in the original development of the cluster expansion method [192]. The Equation 2.14 is a generalization of the *concentration* dependent inner product that was later introduced [187, 188]. And the generalization of configuration spaces with multiple substructures is a formalization of configuration space used in the development of the so-called *coupled* cluster expansion [215].

Finally, Hamiltonians of configuration $H(\boldsymbol{\sigma})$ should be invariant to operations in the symmetry group $\mathcal{G}$ of the underlying disordered crystal structure. Symmetry invariance for a function of configuration means that all permutations T_π $\forall \pi \in \mathcal{G}$ of the occupation variables in $\boldsymbol{\sigma}$ should leave the value of $H(\boldsymbol{\sigma})$ unchanged,

$$T_\pi(H(\boldsymbol{\sigma})) = H(\boldsymbol{\sigma}^\pi) = H(\boldsymbol{\sigma}) \;\; \forall \pi \in \mathcal{G} \tag{2.17}$$

where $\boldsymbol{\sigma}^\pi = (\sigma_{\pi(1)}, \ldots, \sigma_{\pi(N)})$ is a permutation of $\boldsymbol{\sigma}$.

Based on our requirement of symmetry invariance, the lattice model Hamiltonians we seek to represent are actually elements of symmetrically invariant subspaces of $L^2(\boldsymbol{\Omega}, \boldsymbol{\rho})$, which we will denote $L^2(\boldsymbol{\Omega}, \boldsymbol{\rho})^{\mathcal{G}}$.

Definition 2.2.3 ($\mathcal{G}$-Invariant subspace of $L^2(\boldsymbol{\Omega}, \boldsymbol{\rho})$)**.** *A $\mathcal{G}$-invariant subspace of $L^2(\boldsymbol{\Omega}, \boldsymbol{\rho})$ denoted by $L^2(\boldsymbol{\Omega}, \boldsymbol{\rho})^{\mathcal{G}}$ is the set of all permutation invariant functions for all permutation operations in the symmetry group $\mathcal{G}$ in $L^2(\boldsymbol{\Omega}, \boldsymbol{\rho})$,*

$$L^2(\boldsymbol{\Omega}, \boldsymbol{\rho})^{\mathcal{G}} = \{F \in L^2(\boldsymbol{\Omega}, \boldsymbol{\rho}) \;:\; T_\pi(F(\boldsymbol{\sigma})) = F(\boldsymbol{\sigma}) \; \forall \pi \in \mathcal{G}\} \tag{2.18}$$

Having explicitly defined the function space $L^2(\boldsymbol{\Omega}, \boldsymbol{\rho})$ and a symmetrically invariant subspace $L^2(\boldsymbol{\Omega}, \boldsymbol{\rho})^{\mathcal{G}}$, we have left to develop the actual *representation* of any function $F \in L^2(\boldsymbol{\Omega}, \boldsymbol{\rho})$, and more specifically of symmetrically invariant functions $H \in L^2(\boldsymbol{\Omega}, \boldsymbol{\rho})^{\mathcal{G}}$ for the symmetry group $\mathcal{G}$ of a given disordered crystal structure. We will do so in two different ways,

- The first is by way of a symmetrized Fourier product basis [30].[11] We will also extend this formalism by showing how any expansion in any Fourier product basis corresponds to a unique representation in what we call the *cluster decomposition*.
- The second form entails a *redundant* representation using a mathematical frame, which we call the *generalized Potts frame* due to its evident connection to the original Potts model introduced in Chapter 1.3.

2.3 Fourier cluster expansions

The recipe we will use to construct basis functions for $L^2(\mathbf{\Omega}, \boldsymbol{\rho})$ takes advantage of its tensor product structure given in Equation 2.12. Precisely, it has been shown in the development of the cluster expansion method [192], and is generally known from discrete harmonic analysis [30, 157], that a basis for the function space $L^2(\mathbf{\Omega}, \boldsymbol{\rho})$ can be obtained by taking the tensor product of basis functions over the included single site function spaces $L^2(\Omega, \rho)_i$.

Standard site basis sets

Accordingly, we begin by constructing basis sets for functions over a single site space (Ω, ρ). Any linearly independent set of functions $\{\phi_0, \ldots, \phi_{n-1}\}$ of size equal to the dimension of the corresponding space, $n = |\Omega_i|$, constitutes a basis for the space of functions $L^2(\Omega, \rho)$ [30]. This implies that any site function $f : \Omega \to \mathbb{R}$ can be expanded as follows,

$$f(\sigma) = \sum_{j=0}^{n-1} a_j \phi_j(\sigma) \tag{2.19}$$

where a_j are scalar expansion coefficients.

Although a variety of site basis generating schemes[12] have been proposed in the cluster expansion literature [192, 230, 258], we will instead focus on broad class of equivalent basis sets that we call a *standard site basis*.

Definition 2.3.1 (Standard site basis)**.** *A standard site basis for $L^2(\Omega, \rho)$, is a basis $\{\phi_0, \ldots, \phi_{n-1}\}$, that satisfies the following two properties,*

1. $\phi_0 \coloneqq 1$
2. $\langle \phi_j, \phi_k \rangle_\rho = \delta_{jk}$ *for all* $j, k \in \{0, \ldots, n_i - 1\}$

That is, the basis includes the constant function $\phi_0 \coloneqq 1$ and it is orthonormal.

[11] This is precisely the formalism underlying the cluster expansion method [188, 192].

[12] Expressions for such basis generating schemes are detailed in Appendix A.2.

Under the probabilistic interpretation for the inner product in Equation 2.9, and the expansion of a function given in Equation 2.19 using a standard basis, we obtain the following properties [156, 157],

1. $\mathbb{E}_\rho[f] = \langle f \rangle_\rho = a_0$
2. $||f||_2^2 = \langle f^2 \rangle_\rho = \sum_{i=0}^{n-1} a_i^2$
3. $\mathrm{Var}_\rho[f] = \langle f^2 \rangle_\rho - \langle f \rangle_\rho^2 = \sum_{i=1}^{n-1} a_i^2$
4. $\mathrm{Cov}_\rho[f, g] = \langle f, g \rangle_\rho - \langle f \rangle_\rho \langle g \rangle_\rho = \sum_{i=1}^{n-1} a_i b_i$

A standard basis can be obtained from any basis (i.e. using any of the listed basis set generating schemes in Appendix A.2), by ensuring that the constant function is included, and carrying out a Graham-Schmidt orthonormalization procedure. As an example, lets consider a binary site space $|\Omega| = 2$. If we start with a basis $\{\psi_0, \psi_1\}$ which is not a standard basis, we can obtain a standard basis $\{\phi_0, \phi_1\}$ as follows,

1. Define,
$$\phi_0 \coloneqq 1$$
‘
2. And set
$$\phi_1(\sigma) = \frac{\psi_1(\sigma) - \mathbb{E}_\rho[\psi_1(\sigma)]}{\sqrt{\mathrm{Var}_\rho[\psi_1(\sigma)]}}$$

A detailed derivation of this is given in Appendix A.2.

In the above procedure we could have also used ψ_0 instead of ψ_1 (if $\psi_0 \neq 1$) in the Graham-Schmidt process in step 2[13]. Furthermore, if we use $\sigma = \pm 1$, set the probability $\rho(+1) = c$, and start from the basis set $\{\psi_0 = 1, \psi_1 = \sigma\}$, we obtain the following standard basis,

$$\phi_0 = 1$$
$$\phi_1 = \frac{\sigma - \mu}{\sqrt{1 - \mu^2}}$$

where $\mu = 2c - 1$ is the expectation of $\mathbb{E}_\rho[\phi_1] = \mathbb{E}_\rho[\sigma]$.

The expression above is exactly the expression proposed in the formulation of the cluster expansion with *concentration* dependent basis functions [4, 187, 188]. However, to keep to our precept of treating occupation variables σ as categorical, we note that one can always remove the reliance on a specific choice of numerical encoding for occupation variables by simply constructing an initial basis by taking $\phi_0 = 1$, extending the set using indicator

[13]Which goes to say that we need only a set of $|\Omega| - 1$ linearly independent functions to construct a standard basis, since we will include the constant function $\phi_0 = 1$.

functions for each of $|\Omega| - 1$ allowed species, and finally carrying out a Graham-Schmidt process to obtain a standard basis.

Extending the example now to a ternary site space $|\Omega| = 3$, we obtain a standard basis $\{\phi_0, \phi_1, \phi_2\}$ for $L^2(\Omega, \rho)$ starting from the general basis $\{1, \psi_1, \psi_2\}$ as follows,

$$\begin{aligned}
\phi_0 &= 1 \\
\phi_1 &= \frac{\psi_1(\sigma) - \mathbb{E}_\rho\left[\psi_1(\sigma)\right]}{\sqrt{\mathrm{Var}_\rho[\psi_1(\sigma)]}} \\
\phi_2 &= \frac{\psi_2 - \mathbb{E}_\rho\left[\psi_2\right] - \frac{\mathrm{Cov}_\rho[\psi_2\psi_1]}{\mathrm{Var}_\rho[\psi_1]}(\psi_1 - \mathbb{E}_\rho\left[\psi_1\right])}{\sqrt{\mathrm{Var}_\rho[\psi_2] - \frac{\mathrm{Cov}_\rho[\psi_2\psi_1]^2}{\mathrm{Var}_\rho[\psi_1]}}}
\end{aligned}$$

From which tt is clear that under the statistical interpretation of the inner product as expectation values for the *a priori* probability ρ, constructing a standard basis amounts to obtaining basis functions that are *whitened* in the statistical sense. In other words, the non-constant basis functions $\{\phi_i, i \geq 1\}$ are centered (have mean zero), uncorrelated, and have unit variance. This is in fact what gives rise to the properties of standard site basis sets listed above.

Fourier product basis sets

We now proceed to construct a *tensor product* basis for $L^2(\boldsymbol{\Omega}, \boldsymbol{\rho})$ following Equation 2.12 and using *standard site basis functions*. In following, discrete harmonic analysis, we call the resulting basis a Fourier product basis [30, 157].

Definition 2.3.2 (Fourier product basis)**.** *A Fourier product basis* $\{\Phi_{\boldsymbol{\alpha}} : \forall \boldsymbol{\alpha} \in \mathbb{N}^N_{<\boldsymbol{n}}\}$ *is a tensor product basis for* $L^2(\boldsymbol{\Omega}, \boldsymbol{\rho})$ *constructed from standard site basis sets for all* $L^2(\Omega, \rho)_i$, $i \in [N]$ *in a configuration space of dimension* $|\boldsymbol{\Omega}| = N$*. The multi-indices* $\boldsymbol{\alpha} \in \mathbb{N}^N_{<\boldsymbol{n}}$ *for* $\boldsymbol{n} = (|\Omega_i| \mid \forall i \in [N])$ *label each of the total* $|\boldsymbol{\Omega}|$ *functions.*

Since the domain of functions in $L^2(\boldsymbol{\Omega}, \boldsymbol{\rho})$ is discrete, the *tensor product* in Equation 2.12 simplifies to an N-fold product. As a result, a Fourier product basis includes all possible N-fold products among site basis functions for each of the N site spaces in $\boldsymbol{\Omega}$. We can therefore write the basis functions $\Phi_{\boldsymbol{\alpha}}$ explicitly as,

$$\Phi_{\boldsymbol{\alpha}}(\boldsymbol{\sigma}) = \prod_{i=0}^{N-1} \phi^{(i)}_{\alpha_i}(\sigma_i) \tag{2.20}$$

Figure 2.6 shows a schematic illustrating different N-fold products for constructing a product basis for the function space over configurations of the three site system shown in Figure 2.2. A few N-fold products are illustrated with colored arrows connecting the three terms involved in the product. The total number of possible products in the example shown is 18.

A Fourier product basis has the following properties,

1. The basis includes the constant function $\Phi_{\mathbf{0}} = 1$
2. The basis is orthonormal,
$$\langle \Phi_{\boldsymbol{\alpha}}, \Phi_{\boldsymbol{\eta}} \rangle_{\boldsymbol{\rho}} = \delta_{\boldsymbol{\alpha}\boldsymbol{\eta}}$$
3. Basis functions have a *cluster* framework. Meaning they can be thought of as functions only of the occupation variables in the support of their multi-index $\boldsymbol{\sigma}_{\mathrm{supp}(\boldsymbol{\alpha})} = (\sigma_i \mid \forall i \in \mathrm{supp}(\boldsymbol{\alpha}))$,[14]
$$\Phi_{\boldsymbol{\alpha}}(\boldsymbol{\sigma}) = \prod_{i \in \mathrm{supp}(\boldsymbol{\alpha})} \phi^{(i)}_{\alpha_i}(\sigma_i)$$

Proofs for the properties are straightforward and given in Appendix B.2. Property (3) is the tenet of the original *cluster expansion* method [192], in the sense that a basis function $\Phi_{\boldsymbol{\alpha}}$ can be thought of as a function over the configuration of the *site space cluster* whose sites are specified by the support of the function's multi-index $\mathrm{supp}(\boldsymbol{\alpha})$.

Furthermore, it will be useful in coming sections to consider just the non-zero values of the multi-index $\boldsymbol{\alpha}$, as a contracted multi-index $\mathrm{ctr}(\boldsymbol{\alpha}) = (\alpha_i \in \boldsymbol{\alpha} \mid \alpha_i \neq 0)$.

Definition 2.3.3 (Contracted multi-index)**.** *A contracted multi-index* $\mathrm{ctr}(\boldsymbol{\alpha})$ *is a multi-index of non-zero index elements in a given multi-index* $\boldsymbol{\alpha}$*,*

$$\mathrm{ctr}(\boldsymbol{\alpha}) = \boldsymbol{\alpha}_{\mathrm{supp}(\boldsymbol{\alpha})} = (\alpha_i \in \boldsymbol{\alpha} \mid \alpha_i \neq 0) \tag{2.21}$$

[14] All other site basis functions not in the support are constant, i.e. $\phi_0 = 1$

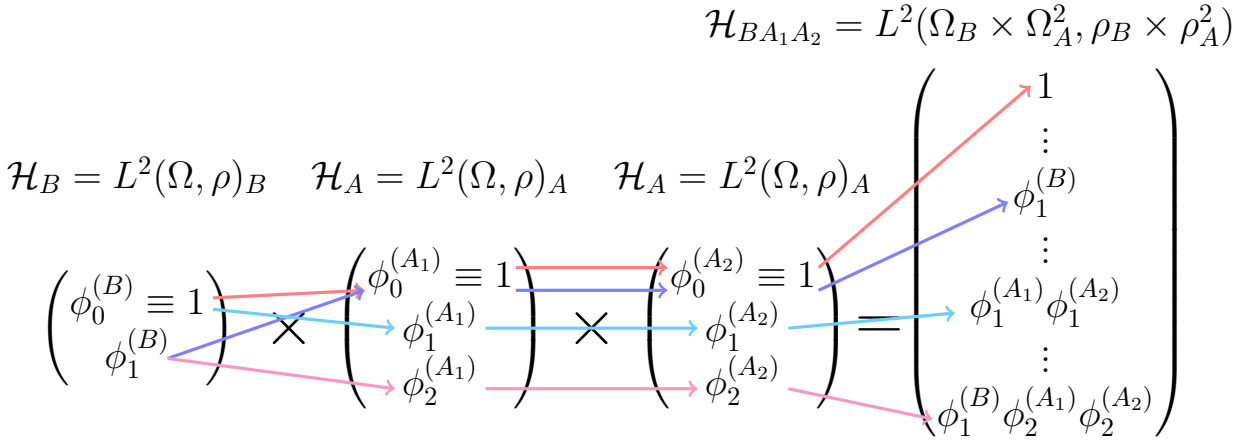


Figure 2.6: Schematic illustrating the construction of a product basis from a set of site basis functions for functions over the configuration space illustrated in Figure 2.2. The construction of subsets of product basis functions from the corresponding site basis functions is depicted with colored arrows. The site spaces $\mathcal{H}_A, \mathcal{H}_A$ and the product space $\mathcal{H}_{BA_1A_2}$ are L^2 Hilbert spaces over their respective domains.

On occasion, we will use a hat to more compactly write a contracted multi-index $\hat{\boldsymbol{\alpha}} \equiv \mathrm{ctr}(\boldsymbol{\alpha})$.

Using a Fourier product basis, any function of configuration $F \in L^2(\boldsymbol{\Omega}, \boldsymbol{\rho})$ can be expanded as follows,

$$F(\boldsymbol{\sigma}) = \sum_{\boldsymbol{\alpha} \in \mathbb{N}^N_{<\boldsymbol{n}}} \hat{F}_{\boldsymbol{\alpha}} \Phi_{\boldsymbol{\alpha}}(\boldsymbol{\sigma}) \tag{2.22}$$

Where the Fourier coefficients, $\hat{F}_{\boldsymbol{\alpha}}$ are given by their projections of the function F onto the respective basis function.

$$\hat{F}_{\boldsymbol{\alpha}} = \langle F(\boldsymbol{\sigma}), \Phi_{\boldsymbol{\alpha}}(\boldsymbol{\sigma}) \rangle_{\boldsymbol{\rho}} \tag{2.23}$$

In analogous fashion to the the properties of an expansion using site basis functions, expansions based on Fourier product basis as given in Equation 2.22 will have the following Fourier formulas [30, 157],

1. $\mathbb{E}_{\boldsymbol{\rho}}[F] = \langle F(\boldsymbol{\sigma}) \rangle_{\boldsymbol{\rho}} = \hat{F}_{\mathbf{0}}$
2. $\mathbb{E}_{\boldsymbol{\rho}}[F^2] = \langle F(\boldsymbol{\sigma})^2 \rangle_{\boldsymbol{\rho}} = \sum_{\boldsymbol{\alpha}} \hat{F}^2_{\boldsymbol{\alpha}}$
3. $\mathrm{Var}_{\boldsymbol{\rho}}[F] = \langle F(\boldsymbol{\sigma})^2 \rangle_{\boldsymbol{\rho}} - \langle F(\boldsymbol{\sigma}) \rangle^2_{\boldsymbol{\rho}} = \sum_{\boldsymbol{\alpha} \neq \mathbf{0}} \hat{F}^2_{\boldsymbol{\alpha}}$
4. $\mathrm{Cov}_{\boldsymbol{\rho}}[FG] = \langle F(\boldsymbol{\sigma}) G(\boldsymbol{\sigma}) \rangle_{\boldsymbol{\rho}} - \langle F(\boldsymbol{\sigma}) \rangle_{\boldsymbol{\rho}} \langle G(\boldsymbol{\sigma}) \rangle_{\boldsymbol{\rho}} = \sum_{\boldsymbol{\alpha} \neq \mathbf{0}} \hat{F}_{\boldsymbol{\alpha}} \hat{G}_{\boldsymbol{\alpha}}$

When using a Fourier product basis to represent or fit a Hamiltonian of configuration, the above Fourier formulas give us direct access to statistical properties of the Hamiltonian under the *a-priori* distribution. In applications this can be interpreted as having direct access to thermodynamic properties in the random or equivalently the high-temperature limit.

Fourier correlation basis sets

A Fourier product basis can be used to represent *any* function of configuration, including functions that are invariant to any set permutations as expressed in Equation 2.17. However, since we will always require a Hamiltonian to be symmetrically invariant, it is advantageous to construct a basis for $L^2(\boldsymbol{\Omega}, \boldsymbol{\rho})^{\mathcal{G}}$ as well.

The trick to obtaining a basis for $L^2(\boldsymbol{\Omega}, \boldsymbol{\rho})^{\mathcal{G}}$ is to use what is known as the Reynolds [74, 218] or averaging operator $\mathcal{R}$,

$$\mathcal{R}(\Phi_{\boldsymbol{\alpha}}) = \frac{1}{|\mathcal{G}|} \sum_{\pi \in \mathcal{G}} \Phi_{\boldsymbol{\alpha}^\pi}(\boldsymbol{\sigma}) = \frac{1}{|\mathcal{G}|} \sum_{\pi \in \mathcal{G}} \prod_{i=1}^{N} \phi^{(i)}_{\pi(\alpha_i)}(\sigma_i) \tag{2.24}$$

However, the sum in Equation 2.24 will usually involve the same Fourier product basis function $\Phi_{\boldsymbol{\alpha}}$ more than once from cases where a multi-index is mapped back to itself $\boldsymbol{\alpha}^\pi = \boldsymbol{\alpha}$. An equivalent expression can be obtained by averaging only over the unique terms, in other words, those with multi-indices in the same orbit.

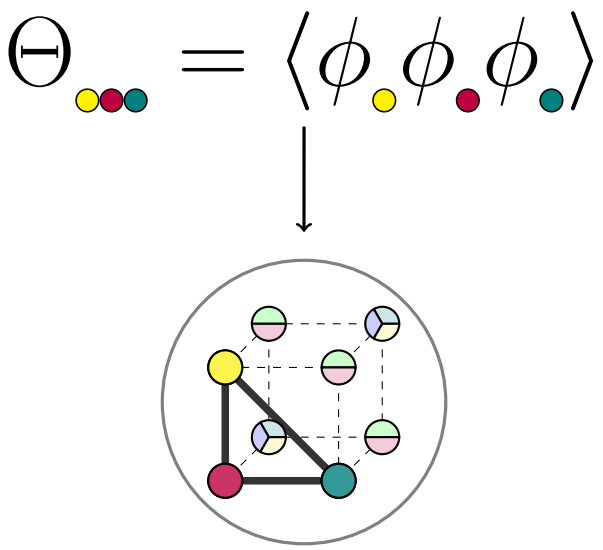


Figure 2.7: Schematic illustrations of a correlation function and its associated function orbit β for a template disordered rocksalt structure. The site coloring in the images represent non-constant site functions. In the illustration, there are two types of site spaces, one with 4 allowed species (3 non-constant site functions); and another with 3 allowed species (2 non-constant site functions).

Definition 2.3.4 (Multi-index orbit)**.** *A multi-index orbit β is a set of multi-indices $\boldsymbol{\alpha}$ generated by permutations of its elements using the permutation operations in a given symmetry group $\mathcal{G}$.*

$$\beta = \{\boldsymbol{\alpha}^{\pi} : \forall \pi \in \mathcal{G}\} \tag{2.25}$$

Definition 2.3.5 (Function orbit)**.** *A function orbit represented, which we also denote by β, is a set of product functions generated by the application of permutation operations of a symmetry group $\mathcal{G}$ to the site functions in a product function $\Phi_{\boldsymbol{\alpha}}$,*

$$\{\Phi_{\boldsymbol{\alpha}^{\pi}} : \forall \pi \in \mathcal{G}\}_{\boldsymbol{\alpha} \in \beta} \tag{2.26}$$

We use multi-index orbits and function orbits almost interchangeably (unless explicitly specified) since a multi-index orbit β is precisely how we specify the product functions that are part of the corresponding function orbit.

We can simplify the action of the Reynolds operator using only averages over *basis function orbits*, and construct a basis for $L^2(\boldsymbol{\Omega}, \boldsymbol{\rho})^{\mathcal{G}}$, which we call a *Fourier correlation basis*, from the set of all symmetrically averaged Fourier basis functions.

Definition 2.3.6 (Fourier correlation basis)**.** *A Fourier correlation basis $\{\Theta_{\beta}\}_{\beta \in \mathcal{G}(\mathbb{N}^N_{\leq \boldsymbol{n}})}$ is the set of correlation functions generated from all possible function orbits β of a given Fourier*

product basis $\{\Phi_{\boldsymbol{\alpha}}\}_{\boldsymbol{\alpha}\in\mathbb{N}^N_{\leq \boldsymbol{n}}}$ *generated by way of the permutations of a symmetry group* $\mathcal{G}$*. Correlation functions* Θ_β *are given by,*

$$\Theta_\beta(\boldsymbol{\sigma}) = \frac{1}{|\beta|}\sum_{\boldsymbol{\alpha}\in\beta}\Phi_{\boldsymbol{\alpha}}(\boldsymbol{\sigma}) \tag{2.27}$$

A schematic illustration of a triplet correlation function over a particular function orbit β is shown in Figure 2.7, where the values of the contracted multi-index $\mathrm{ctr}(\boldsymbol{\alpha})$ are depicted by different colors. The function orbit is depicted as highlighted sites in a representative unit cell colored by the values of their corresponding entry in $\mathrm{ctr}(\boldsymbol{\alpha})$. The unit cell is a representation of the underlying disordered crystal structure.

Intuitively, the procedure outlined above implies or equivalently can be also be derived starting from the requirement that the expansion coefficients $\langle H(\boldsymbol{\sigma}), \Phi_{\boldsymbol{\alpha}}(\boldsymbol{\sigma})\rangle_{\boldsymbol{\rho}}$ of a given Hamiltonian H be invariant to all operations T_π $\forall\pi\in\mathcal{G}$ [187, 188, 192].

Furthermore, since the main application will be bulk crystals (periodic systems), it is customary to write the expression for a correlation basis function in Equation 2.27 as follows,

$$\Theta_\beta(\boldsymbol{\sigma}) = \frac{1}{Nm_\beta}\sum_{\boldsymbol{\alpha}\in\beta}\Phi_{\boldsymbol{\alpha}}(\boldsymbol{\sigma}) \tag{2.28}$$

where the *multiplicity* m_β is defined as a density of the size of function orbit β, $m_\beta = |\beta|/N$. This expression allows practical calculations for periodic structures with different numbers of sites N on equivalent grounds.

Figure 2.8 shows the two possible standard basis sets for a binary site space, and the two possible resulting Fourier correlation basis sets for functions over the three distinct configurations of the symmetric diatomic molecule pictured above.

The expansion of a symmetrically invariant function $H(\boldsymbol{\sigma})$ of configuration using a Fourier correlation basis is called a *Fourier cluster expansion.*

Definition 2.3.7 (Fourier cluster expansion)**.** *A Fourier cluster expansion is a representation of a function* $H\in L^2(\boldsymbol{\Omega},\boldsymbol{\rho})^{\mathcal{G}}$ *using a Fourier correlation basis* $\{\Theta_\beta\}_{\beta\in\mathcal{G}(\mathbb{N}^N_{\leq \boldsymbol{n}})}$*. We can express the Fourier cluster expansion of* H *as follows,*

$$H(\boldsymbol{\sigma}) = \sum_{\beta\in\mathcal{G}(\mathbb{N}^N_{\leq \boldsymbol{n}})} Nm_\beta J_\beta\Theta_\beta(\boldsymbol{\sigma}) \tag{2.29}$$

where the sum is carried out over all multi-index orbits β *generated by elements in the symmetry group* $\mathcal{G}$ *of the underlying disordered crystal structure. The expansion coefficients* J_β *are known in the literature as effective cluster interactions*[15]

[15]The use of the name effective cluster interactions is somewhat misleading as their magnitudes and sign depend on the specific choice of basis; and as we will see in our development of the cluster decomposition, the actual mean cluster interactions (which are invariant to the choice of basis) and represent contributions of site clusters to the energy are not simply the expansion coefficients J_β.

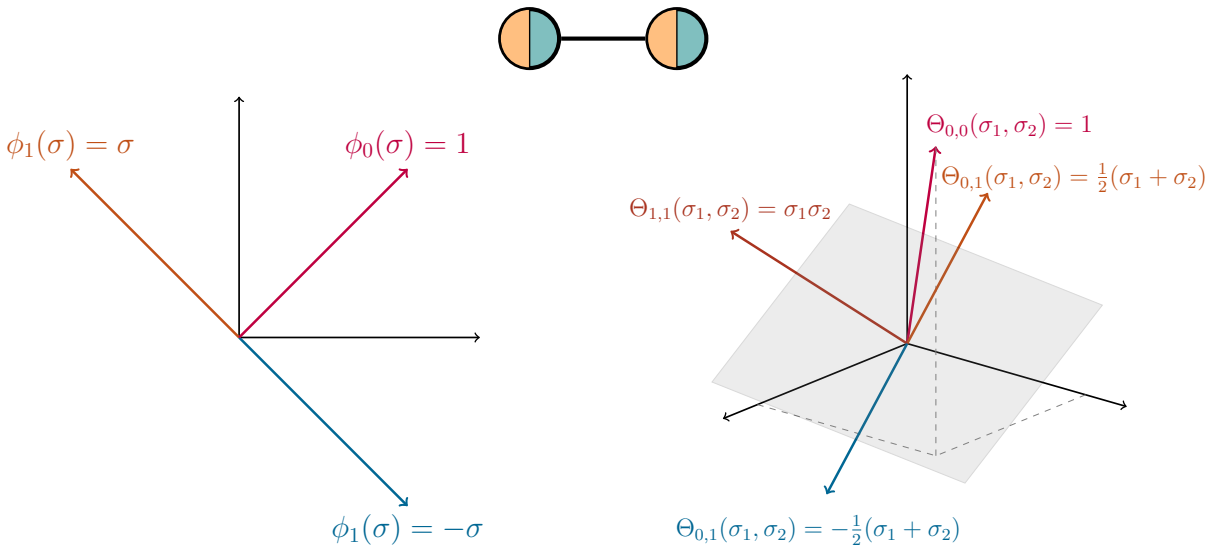


Figure 2.8: Function space representations over the configurations of a single binary site and a symmetric binary diatomic molecule. (a) Function space over a single binary site space. The two different choices for standard site bases are colored blue and orange. Each set also includes the purple $\phi_0 \equiv 1$ function. (b) Function space over symmetrically distinct configurations of the molecule. The two possible Fourier correlation basis sets include the constant $\Phi_{0,0}$ (red) and the $\Phi_{1,1}$ pair correlation function (brown) and either the blue colored or the orange colored point function $\Phi_{0,1}$.

The *effective cluster interactions* J_β are given by the projection of the function H onto the respective correlation function,

$$J_\beta = \langle H(\boldsymbol{\sigma}), \Theta_\beta(\boldsymbol{\sigma}) \rangle_{\boldsymbol{\rho}} \tag{2.30}$$

Equivalently, it follows from the symmetry invariance of the Hamiltonian H that the effective cluster interactions are also the projection onto any Fourier basis function $\Phi_{\boldsymbol{\alpha}}$ for $\boldsymbol{\alpha} \in \beta$.

The procedure outlined in constructing Fourier product basis functions from standard site basis functions, and subsequently averaging Fourier product functions to obtain a Fourier correlation basis, can be carried out starting from any arbitrary set of site basis functions. However, if the set of site basis functions does not include the constant $\phi_0 = 1$ function, then the resulting product basis set will not have a *cluster* framework and thus are not of much practical use since all basis functions will depend on all occupation variables rather than just those associated with site space clusters only. It is precisely the cluster framework that allows one to obtain a useful correlation basis with which one can seek to express Hamiltonians using only a small subset of the large (possibly infinite) set of all basis functions.

If however, the constant function is included in the site basis sets used, but the set is not orthonormal, then one can still obtain practically useful product and correlation basis

sets. Indeed two of the most commonly used site basis sets are not orthonormal [230, 258].[16] In these cases, one does obtain a cluster framework as we have described, but the resulting basis set will not result in the analytic/statistical properties outlined above. We will call expansions in a correlation basis constructed from any type of site basis that includes $\phi_0 = 1$ simply as *cluster expansions* to emphasize the distinction with *Fourier cluster expansions* which must be constructed using *standard site basis* sets.

Fourier cluster expansions over constrained spaces

We make a small digression to discuss some implications of using Fourier correlation basis sets to represent functions over configuration spaces with charge neutrality constraints $\hat{\mathbf{\Omega}}$ as detailed in Equation 2.7. Since the total number of functions in a product basis over $\mathbf{\Omega}$ is precisely $|\mathbf{\Omega}|$, a set of product basis functions is *over-complete* for the function space over the constrained configuration space $\hat{\mathbf{\Omega}}$ of a heterovalent ionic system. Consequentially, orthonormal/orthogonal product basis sets have no such orthogonality properties when restricted to $\hat{\mathbf{\Omega}}$.

The *overcompleteness* of the correlation basis sets for ionic materials does not formally prevent their use, since the set still spans the functions over charge neutral configurations. However, since it is over-complete, the set of all correlation functions is not linearly independent. The result is that the use of an over-complete set to express functions over a configuration space with composition constraints $\hat{\mathbf{\Omega}}$ introduces linear dependencies between correlation functions. This implies that—in contrast to systems without any active composition constraints such as metallic alloys—material properties of ionic materials as a function of their configuration do not have a unique Fourier cluster expansion for any given set of correlation functions.

The simplest case can be illustrated by considering the most trivial linear relation that arises among the constant and single site cluster functions only,

$$\sum_k \rho_k \Theta_k(\boldsymbol{\sigma}) + \rho_\emptyset = 0 \tag{2.31}$$

where the sum runs over all orbits k with single site clusters only, and the constants ρ_k can be obtained from the composition constraints in Equation 2.7, the respective multiplicities associated with each point function, and the particular choice of site basis set used.

To further illustrate these linear dependencies, consider the constant and set of single site correlation functions based on site indicator functions (see details in Appendix A.2) for the system in Figure 2.5. The resulting linear constraint is,

$$\langle \mathbf{1}_r(\boldsymbol{\sigma}) \rangle - \langle \mathbf{1}_t(\boldsymbol{\sigma}) \rangle - \langle \mathbf{1}_b(\boldsymbol{\sigma}) \rangle - 1 = 0 \tag{2.32}$$

[16] The trigonometric basis [230] is orthonormal for binary site spaces, since it is exactly one of the 2 choices for a standard basis. However, for $|\Omega| > 3$ the resulting site basis is orthogonal but not normalized.

The example constraint in Equation 2.32 results in linear dependencies that give rise to infinitely many expressions for the same CE. This implies that, once a set of ECI coefficients are obtained for a particular CE, the coefficients can be transformed accordingly,

$$\begin{aligned} J'_\emptyset &= J_\emptyset - x \\ J'_r &= J_r + x \\ J'_t &= J_t - x \\ J'_b &= J_b - x \end{aligned}$$

for any scalar x. The cluster expansion with transformed coefficients represents exactly the same function as the one with the original set of coefficients.

Additional linear dependencies exist among higher-order correlation functions as well. Referring again to the finite example in Figure 2.5, the following linear relationships exist among pair functions,

$$\begin{aligned} \langle \mathbf{1}_r(\boldsymbol{\sigma})\mathbf{1}_t(\boldsymbol{\sigma})\rangle - \langle \mathbf{1}_t(\boldsymbol{\sigma})\mathbf{1}_t(\boldsymbol{\sigma})\rangle - \frac{1}{2}\langle \mathbf{1}_t(\boldsymbol{\sigma})\mathbf{1}_b(\boldsymbol{\sigma})\rangle = 0 \\ \langle \mathbf{1}_r(\boldsymbol{\sigma})\mathbf{1}_b(\boldsymbol{\sigma})\rangle - \langle \mathbf{1}_b(\boldsymbol{\sigma})\mathbf{1}_b(\boldsymbol{\sigma})\rangle - \frac{1}{2}\langle \mathbf{1}_t(\boldsymbol{\sigma})\mathbf{1}_b(\boldsymbol{\sigma})\rangle = 0 \end{aligned}$$

In order to remove the resulting linear dependencies, one could in theory remove $|\boldsymbol{\Omega}| - |\hat{\boldsymbol{\Omega}}|$ functions to obtain a linearly independent set. However, for real bulk ionic systems, obtaining analytical expressions for the linear relationships among higher-order correlation functions may be too lengthy of a task, let alone constructing an orthogonal basis set which is far from trivial—as for example has been done for slices of the boolean hypercube [68]. Furthermore, it is not clear that removing all linear dependencies is even necessary. The cluster basis still spans the constrained configuration space, and correlation basis sets have been successfully used as is to fit properties of ionic materials [34, 129, 130, 165, 182, 198, 199, 236, 245, 247]. Indeed, the deleterious effects of linear dependencies on expansion coefficients can be managed with appropriate sampling strategies and choices of regularization during fitting, which we will describe in Chapter 4.

2.4 The cluster decomposition

So far we have presented the formalism to construct Fourier or correlation basis sets as a flexible representation of any lattice model Hamiltonian. This formalism so far corresponds to that of the original cluster expansion method [187, 188, 192]. However, infinitely many Fourier product basis sets that span the same function space $L^2(\boldsymbol{\Omega}, \boldsymbol{\rho})$ can be constructed from different *standard site basis* choices. This in itself is not a practical problem, any Hamiltonian representation is equally valid and will give the same results regardless of the choice of standard basis.[17] Yet different values of the expansion coefficients (or ECI) are

[17] In fact the site basis does not even need to be *standard*. Any site basis will do!

needed to represent the same Hamiltonian expansion with different bases. Further, since the choice of site basis is arbitrary, then so are the associated expansion coefficients for a particular Hamiltonian; and as a result, any effort to interpret expansion coefficients is going to be tenuous at best. With a deeper consideration of the structure of Fourier correlation basis sets, a Fourier cluster expansion can be re-written in a form that is both *unique*[18] and *interpretable*.

The basic idea to obtain a *unique* representation from a Fourier cluster expansion simply requires one to group the correlation functions by the site space cluster orbits $B = \text{orb}(A)$ for $A \subseteq [N]$ over which they operate. The resulting Hamiltonian expansion will then essentially be expressed in the form given in Equation 1.17 in our introduction for any general Hamiltonian up to arbitrary multiple body interactions. When this re-writing procedure is carried out with a Fourier cluster expansion, it results in what we call a *cluster decomposition*.

Definition 2.4.1 (Cluster decomposition)**.** *The* ***cluster decomposition*** *of a lattice Hamiltonian $H(\boldsymbol{\sigma}) \in L^2(\boldsymbol{\Omega}, \boldsymbol{\rho})^{\mathcal{G}}$ is an expansion of the following form,*

$$H(\boldsymbol{\sigma}) = N \sum_{B \in \mathcal{G}\,\mathrm{P}([N])} m_B \sum_{\beta \in L(B)} \hat{m}_\beta J_\beta \Theta_\beta(\boldsymbol{\sigma}) \tag{2.33}$$

where $L(B) = \{\beta : \text{supp}(\boldsymbol{\alpha}) \in B \ \forall \boldsymbol{\alpha} \in \beta\}$ are sets of function cluster orbits β containing multi-indices $\boldsymbol{\alpha}$ with symmetrically equivalent supports, i.e. with support that belongs to the same orbit B of site space clusters. The constant m_B is the multiplicity of the site-space cluster orbit B per unit cell, and $\hat{m}_\beta = |\hat{\beta}|$ is the permutation multiplicity for the orbit $\hat{\beta}$ of contracted multi-indices $\hat{\boldsymbol{\alpha}}$, i.e. $\hat{\beta} = \{\text{ctr}(\boldsymbol{\alpha}), \forall \boldsymbol{\alpha} \in \beta\}$[19].

A schematic illustrating the sets $L(B)$ following the same conventions from Figure 2.7 is shown in Figure 2.9. The set $L(B)$ includes all symmetrically distinct function clusters β that act over the orbit of site space clusters B. $L(B)$ is depicted as highlighted sites that are colored with all possible values that the entries of an associated contracted multi-index $\hat{\boldsymbol{\alpha}}$ (for any $\boldsymbol{\alpha} \in \beta$) can take.[20]

Mean cluster interactions

Each inner sum in the cluster decomposition in Equation 2.33 corresponds to a term that acts over all of the clusters S in each orbit B included in the Hamiltonian. Each such term constitutes a particular multi-body interaction term. Accordingly, we will refer to these terms as *mean cluster interactions*.

[18] That is agnostic to the choice of standard site bases.

[19] More straightforward, this multiplicity is simply the number of symmetrically equivalent permutations of labels over the sites of a fixed cluster $\text{supp}(\boldsymbol{\alpha})$ with non-constant basis functions.

[20] In fact though the decoration, in this case, is with "non-constant" basis functions, instead of site spaces, both cases will have the same symmetry group $\mathcal{G}$. So we will be a bit sloppy and use the two concepts interchangeably.

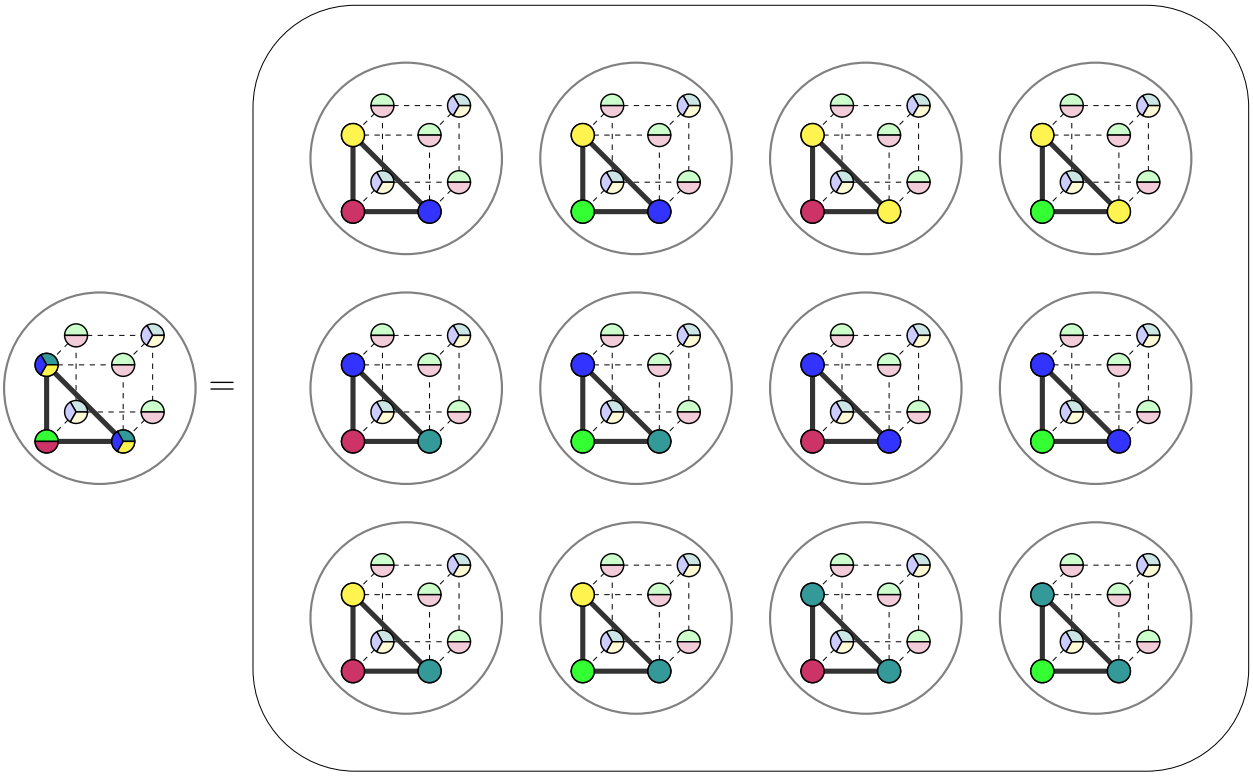


Figure 2.9: Schematic depicting a set $L(B)$ for a triplet interaction. The colored sites correspond to a site space cluster S, and the orbit of all symmetrically equivalent site clusters is $B = \text{orb}(S)$. The figures with different combinations of colors over the sites S represent all the function clusters β that operate over the orbit of site space clusters B, i.e. β such that $\text{supp}(\boldsymbol{\alpha}) \in B \ \forall \boldsymbol{\alpha} \in \beta$. There are two types of site spaces in the illustration: one with 4 allowed species (3 non-constant site functions); and another with 3 allowed species (2 non-constant site functions).

Definition 2.4.2 (Mean cluster interaction)**.** *A **mean cluster interaction** is a term in a **cluster decomposition** that acts only over the clusters of sites in a particular orbit B, and is given by,*

$$H_B(\boldsymbol{\sigma}) = \sum_{\beta \in L(B)} \hat{m}_\beta J_\beta \Theta_\beta(\boldsymbol{\sigma}) \tag{2.34}$$

The cluster decomposition can be conveniently expressed in terms of mean cluster interactions, making its resemblance to the general Hamiltonian in equation 1.17 much more explicit[21].

$$H(\boldsymbol{\sigma}) = N \sum_{B \in \mathcal{G}\,\mathrm{P}([N])} m_B H_B(\boldsymbol{\sigma}) \tag{2.35}$$

[21]It is simply a re-grouping of the terms in Equation 1.17 that are symmetrically equivalent

Each mean cluster interaction has an associated *effective cluster weight* determined by the precise values of the ECI or Fourier expansion coefficients included.

Definition 2.4.3 (Effective cluster weight)**.** *The* ***effective cluster weights****,* $\mathrm{W}[H_B]$ *are weighted norms of the mean cluster interactions. The value of an effective cluster weight* $\mathrm{W}[H_B]$ *is,*

$$\mathrm{W}[H_B] = m_B N ||H_B||_2^2 \tag{2.36}$$

We note that despite the explicit factor of N, $\mathrm{W}[H_B]$ in fact does not scale with the system size N; this is because H_B is itself normalized by N, i.e. the correlation functions that make up H_B are themselves normalized per Equation 2.27.

A simple, and quite useful, expression for the effective cluster weights follows from the orthogonality of Fourier correlation functions,

$$\mathrm{W}[H_B] = \sum_{\beta \in L(B)} \hat{m}_\beta J_\beta^2 \tag{2.37}$$

We pause here to note that mean cluster interactions and their effective cluster weights as given in Equations 2.34 and 2.37 are defined based on a particular Hamiltonian. In practice, constructing mean cluster interactions requires both the expansion coefficients J_β (which are model specific) and the Fourier correlation functions. However, in what follows, we will show that the actual resulting mean cluster interactions (the functions) and the associated effective cluster weights (the function norms) are indeed independent of the particular choice of standard basis.

The value of re-writing a cluster expansion as a cluster decomposition arises from two important properties of the mean cluster interactions. Specifically, the mean cluster interactions H_B in the cluster decomposition have the following two properties,[22]

1. They are orthogonal,
$$\langle H_B, H_D \rangle = \begin{cases} ||H_B||_2^2 \text{ if } B = D \\ 0 \text{ if } B \neq D \end{cases}$$
2. They are *unique*,
$$\sum_{\beta \in L(B)} \hat{m}_\beta J_\beta^{(1)} \Theta_\beta^{(1)}(\boldsymbol{\sigma}) = \sum_{\beta \in L(B)} \hat{m}_\beta J_\beta^{(2)} \Theta_\beta^{(2)}(\boldsymbol{\sigma})$$
For any choice of Fourier correlation basis $\{\Theta_\beta^{(1)}\}$ and $\{\Theta_\beta^{(2)}\}$.

The two properties above are noteworthy since they give way to a formal analysis and interpretation of the terms in Fourier cluster expansion of any applied lattice model. As a start, simply by the orthogonality of any mean cluster interaction H_B with the constant

[22] Proofs for these properties are given in Appendix B.2.

interaction $H_\emptyset$, we can identify effective cluster weights as a measure of the variance of a mean cluster interaction,

$$\mathrm{Var}_{\boldsymbol{\rho}}[H_B] = \langle H_B, H_B \rangle_{\boldsymbol{\rho}} - \langle H_B, 1 \rangle_{\boldsymbol{\rho}}^2 = \frac{\mathrm{W}[H_B]}{m_B N} \tag{2.38}$$

In fact, we will show that the mean cluster interactions and effective cluster weights represent a unique decomposition of the variance of a Hamiltonian. Based on these results, we will formalize and outline ways to effectively *interpret* lattice models by way of the cluster decomposition in Section 2.4.

Site basis rotations and invariance of effective cluster weights

Before discussing the question of interpretation, we make a digression to show how effective cluster weights are *invariant* to the specific choice of Fourier correlation basis. To do so we will first derive a change of basis matrix between two different standard Fourier product basis sets. We can do this by first considering the properties of standard site basis functions. Since all standard site basis sets must include the constant function ($\phi_0 \equiv 1$) and must be orthonormal, it follows that different standard site basis sets are related by rotations orthogonal to the constant function[23]. Figure 2.10 shows the case of two site basis sets over a ternary site space. In this case, the rotations are simply those about the orthogonal plane [111].

For simplicity, let's consider a simple lattice system,[24] i.e. only one site space per lattice point. We start with a set of Fourier product basis functions constructed from a standard site basis $\{\phi_i,\ i = 0, n-1\}$, written out as follows,

$$\Phi_\alpha(\boldsymbol{\sigma}) = \prod_i^N \phi_{\alpha_i}$$

Any another standard site basis $\{\psi_i\}$, must be related to $\{\phi_i\}$ by some rotation R orthogonal to ϕ_0, i.e. $\psi_i = \mathrm{R}\,\phi_i$, as depicted in Figure 2.10. The resulting product basis functions can then be expressed as follows,

$$\begin{aligned}\Psi_\alpha(\boldsymbol{\sigma}) &= \prod_i^N \psi_{\alpha_i} \\ &= \prod_i^N R\phi_{\alpha_i}\end{aligned}$$

[23] Explicitly these are rotations in a hyperplane of rotation orthogonal to the constant vector.

[24] Extending to the general case with different site spaces is straightforward.

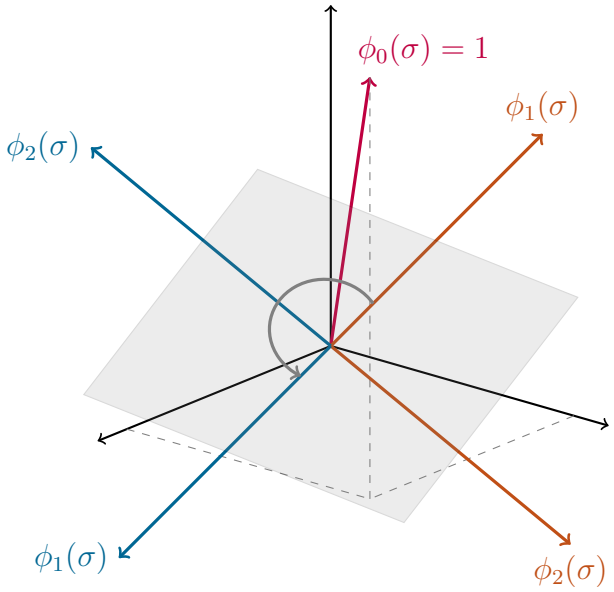


Figure 2.10: Two different choices of standard site basis sets for functions of the configuration for a ternary site space, and the rotation R relating them. Both basis sets by definition include the constant $\phi_0 = 1$ colored in red. Any arbitrary rotation about ϕ_0 results in a standard site basis.

Now we can construct the change of basis matrix from $\Psi \to \Phi$ is $U_{\gamma\alpha} = \langle \Psi_\gamma, \Phi_\alpha \rangle$, starting from the following expression for the relation between the basis functions,

$$\Phi_\alpha(\boldsymbol{\sigma}) = \sum_{\boldsymbol{\gamma}} \langle \Psi_{\boldsymbol{\gamma}}, \Phi_{\boldsymbol{\alpha}} \rangle \Psi_{\boldsymbol{\gamma}}(\boldsymbol{\sigma})$$

By expressing the un-rotated basis Φ_α in terms of the rotated basis Ψ_α we obtain the following expression for the change of basis matrix elements,

$$\begin{aligned} \langle \Psi_\gamma, \Phi_\alpha \rangle &= \langle \prod_i^N R\phi_{\gamma_i}, \prod_i^N \phi_{\alpha_i} \rangle \\ &= \prod_i^N \langle R\phi_{\gamma_i}, \phi_{\alpha_i} \rangle \\ &= \left(\prod_i^N R_{\alpha_i,\gamma_i} \right) \delta_{\mathrm{supp}(\boldsymbol{\gamma})\,\mathrm{supp}(\boldsymbol{\alpha})} \end{aligned} \tag{2.39}$$

where we used the fact that $\langle R\phi_{\gamma_i}, \phi_0 \rangle = \delta_{\gamma_i 0}$, since by definition all non-constant functions must be orthogonal to ϕ_0. We observe that the change of basis matrix is simply the product

of elements of the site rotations matrix expressed in the $\{\phi_i\}$ basis for elements corresponding to product functions that act over the same cluster of sites S.

Since it is a change of basis matrix, $U_{\gamma,\alpha}$ must be orthogonal. But more importantly, the change of basis matrix $U_{\gamma\alpha}$, is block diagonal; where the blocks correspond to the product functions acting over the same set of site space clusters S identified by the support of their multi-indices ($\text{supp}(\boldsymbol{\alpha})$). Furthermore, since $U_{\gamma,\alpha}$ is orthogonal, it follows that the blocks themselves are orthogonal.[25] The blocks being orthogonal implies that for a given Hamiltonian H the norm of all the expansion terms in a given block is left unchanged from a change of Fourier correlation basis,

$$\left\langle \left(\sum_{\boldsymbol{\gamma} \,:\, \text{supp}(\boldsymbol{\gamma})=S} J'_{\boldsymbol{\gamma}} \Psi_{\boldsymbol{\gamma}} \right)^2 \right\rangle_{\boldsymbol{\rho}} = \left\langle \left(\sum_{\boldsymbol{\alpha} \,:\, \text{supp}(\boldsymbol{\alpha})=S} J_{\boldsymbol{\alpha}} \Phi_{\boldsymbol{\alpha}} \right)^2 \right\rangle_{\boldsymbol{\rho}}$$
$$\sum_{\boldsymbol{\gamma} \,:\, \text{supp}(\boldsymbol{\gamma})=S} J'^2_{\boldsymbol{\gamma}} = \sum_{\boldsymbol{\alpha} \,:\, \text{supp}(\boldsymbol{\alpha})=S} J^2_{\boldsymbol{\alpha}} \tag{2.40}$$

The expression Equation 2.40 above applies to any function of configuration, however, when dealing with a symmetrically invariant Hamiltonian, we can group the sums by site space clusters B and obtain the following invariance relation,

$$\sum_{\eta \in L(B)} \hat{m}_\eta J'^2_\eta = \sum_{\beta \in L(B)} \hat{m}_\beta J^2_\beta \tag{2.41}$$

which is simply an expression that the cluster weights $\text{W}[H_B]$ are invariant to the choice of basis. This should not be a surprise considering that we have already stated that the mean cluster interactions are unique.

As an example, Figure 2.11a shows vertical stems for the values of coefficients for cluster correlation functions that act over three different orbits of clusters of three sites (triplets) and three orbits of clusters of four sites (quads) in a faced-centered cubic ternary disordered structure. Two standard site basis sets related by a rotation of $2\pi/3$ radians are used to construct two different sets of Fourier correlation functions, as shown in Figure 2.11b. A shaded area with height corresponding to the corresponding effective cluster weight is overlaid over each set of correlation functions. The values of the correlation functions, along with the corresponding value of the mean cluster interaction for a randomly chosen configuration $\boldsymbol{\sigma}$, are also plotted. The change of basis matrix for the correlation functions shown is visualized in Figure 2.11c.

As we have already mentioned, this *invariace* among mean cluster interactions is due to the block diagonal and orthogonal nature of the change of basis matrix relating two Fourier correlation basis sets as shown in Figure 2.11. It is evident from Figure 2.11, how the effective cluster weights and the particular values of the mean cluster interactions remain

[25] Proofs are given in Appendix B.2.

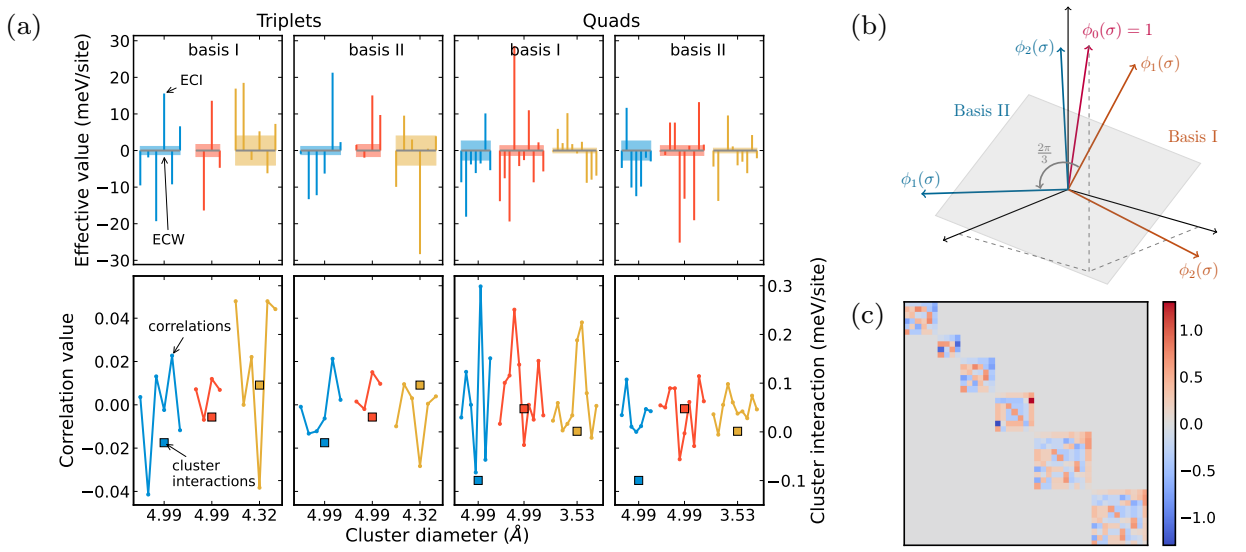


Figure 2.11: (a) Correlation function coefficients, effective cluster weights; and correlation function and mean cluster interaction values for a randomly chosen configuration $\boldsymbol{\sigma}$ of a ternary face-centered cubic disordered structure for functions acting over three (triplet and four (quad) site cluster orbits. (b) The two standard site basis sets used to compute the values plotted in (a). (c) Change of basis matrix relating the Fourier correlation basis functions shown.

the same irrespective of the choice of site basis, and in effect represent the contribution to the total energy from said interactions. In what follows, we reveal further meaning and possible interpretations by establishing connections with well-known statistical decompositions of functions of (discrete) random variables.[26]

Interpretation of mean cluster interactions

The cluster decomposition is, at its essence, an expansion of a Hamiltonian where the expansion terms are the energy contribution of all possible site space clusters. If symmetry were ignored there are 2^N site spaces clusters in a structure with N sites—essentially the power set $\mathrm{P}([N])$ of the N sites. When symmetry is taken into consideration, we simply sum all interactions associated with clusters in the same orbit (by symmetry these terms represent the same function over their respective sites) and obtain Expression 2.35. Furthermore, since the mean cluster interactions are orthogonal and the cluster decomposition is unique, it follows that the cluster decomposition is a *symmetrized* Sobol decomposition [206]. The

[26] These statistical decompositions are not actually limited to discrete variables, in fact, much of the original development of such decompositions has been carried out for continuous random variables [96, 206].

Sobol decomposition is also known as the functional ANOVA (f-ANOVA) decomposition[27]

We can thus leverage the fact that the cluster decomposition is a *symmetrized* form of the Sobol decomposition to obtain a deeper understanding and enable interpretation of the expansion terms of a generalized lattice Hamiltonian. As the name ANOVA—which stands for *analysis of variance* [65, 96]—suggests, this connection provides deeper insight into the variance structure of a Hamiltonian beyond the total variance Fourier formula given in Section 2.3. Additionally, we can also leverage tools and results from the field of variance based Sensitivity Analysis [65, 101, 186] to obtain further insight into the structure of a lattice Hamiltonians by permitting us to formally rank the importance the included cluster interactions.

We first list the formulation of the Sobol decomposition explicitly. The Sobol decomposition was originally derived for multivariate functions over the interior of the unit hypercube in N dimensions, $f : [0,1]^N \to \mathbb{R}$. However, the decomposition can be used for any multivariate function in a Hilbert space (in the present case $L^2(\boldsymbol{\Omega}, \boldsymbol{\rho})$). Any such function f has a Sobol decomposition expressed as follows [96, 206],

$$f(\boldsymbol{x}) = \sum_{S \in [N]} f_S(\boldsymbol{x}) \tag{2.42}$$

The expansion is a f-ANOVA representation (and is unique) if the terms f_S are mean zero and orthogonal [96, 206], that is,

$$\int_{[0,1]^N} f_S(\boldsymbol{x}) \rho(\boldsymbol{x}) d\boldsymbol{x} = \mathbb{E}_\rho \left[f_S(\boldsymbol{x}) \right] = 0 \quad \forall S \neq \emptyset \tag{2.43}$$

$$\int_{[0,1]^N} f_S(\boldsymbol{x}) f_T(\boldsymbol{x}) \rho(\boldsymbol{x}) d\boldsymbol{x} = \mathbb{E}_\rho \left[f_S(\boldsymbol{x}) f_T(\boldsymbol{x}) \right] = 0 \quad \forall S \neq T \tag{2.44}$$

where the expectation is taken using a product probability density, thus implying that the elements of $\boldsymbol{x}$ are independently distributed.

Under the above properties, it follows that the terms f_S can be constructed with the following procedure [96, 206],

$$f_\emptyset = \mathbb{E}_\rho \left[f(\boldsymbol{x}) \right] \tag{2.45}$$

$$f_{\{i\}}(\boldsymbol{x}) = \mathbb{E}_\rho \left[f(\boldsymbol{x}) - f_\emptyset | \boldsymbol{x}_i \right] \tag{2.46}$$

$$f_S = \mathbb{E}_\rho \left[f(\boldsymbol{x}) - \sum_{T \subset U} f_T(\boldsymbol{x}) \middle| \boldsymbol{x}_S \right] \tag{2.47}$$

where the conditional expectations are taken over all elements of $\boldsymbol{x}$ except those listed, and $\boldsymbol{x}_S = (\boldsymbol{x}_i \mid \forall i \in S)$.

[27] The Sobol or f-ANOVA decomposition is used in a variety of fields and applications where it known with many different names, including Hoeffding, Efron, and Stein decomposition [96, 97, 157, 208].

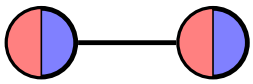


Figure 2.12: A hypothetical symmetric diatomic molecule with binary degrees of freedom.

The first term $f_\emptyset$ given by Equation 2.45 is called the *grand mean*. The terms $f_{\{i\}}$ over a single variable are called *main effects*. All other terms f_S for $|S| > 1$ are called *interactions* [65, 96]. From the prescription above, we see that the expansion terms f_S represent the portion of the function f arising only from the interactions of variables $\boldsymbol{x}_S$ (resultant from averaging out over all other elements), and that is not captured by any lower order terms f_T for $T \subset S$.

Moreover, an important consequence of the orthogonality of the terms f_S and the statistical independence of each element of $\boldsymbol{x}$ is a guaranteed *variance decomposition* of the following form,

$$\mathrm{Var}_\rho[f(\boldsymbol{x})] = \sum_{S \in [N]} \mathrm{Var}_\rho[f_S(\boldsymbol{x})] \tag{2.48}$$

Where the variance of each term $\mathrm{Var}_\rho[f_S(\boldsymbol{x})]$ can be obtained using Equation 2.47 as follows,

$$\mathrm{Var}_\rho[f_S(\boldsymbol{x})] = \mathrm{Var}_\rho[\mathbb{E}_\rho\left[f(\boldsymbol{x})|\boldsymbol{x}_S\right]] - \sum_{T \subset S} \mathrm{Var}_\rho[\mathbb{E}_\rho\left[f(\boldsymbol{x})|\boldsymbol{x}_T\right]] \tag{2.49}$$

Similarly, we see that the variance of each term f_S represents the variance from only those variables in $\boldsymbol{x}_S$ after averaging out all other variables, and so removing the variance arising from all subsets of variables $\boldsymbol{x}_T$ for $T \subset S$. Since they are disjoint, these variances can be used to measure the importance of the interaction of the variables $\boldsymbol{x}_S$ captured by the term f_S in the fANOVA expansion of $f(\boldsymbol{x})$. In other words, the variance of f_S can be used to measure how *sensitive* the function f is to interactions of input variables $\boldsymbol{x}_S$. More precisely, this is done by using *global sensitivity indices*, also known as *Sobol indices* τ_S; which are given by,

$$\tau_S = \frac{\mathrm{Var}_\rho[f_S(\boldsymbol{x})]}{\mathrm{Var}_\rho[f(\boldsymbol{x})]} = \frac{\mathrm{Var}_\rho[f_S(\boldsymbol{x})]}{\sum_{T \in [N]} \mathrm{Var}_\rho[f_T(\boldsymbol{x})]} \tag{2.50}$$

Sobol indices make up an important tool in the field of variance based Sensitivity Analysis [65, 101, 186], where they are used to rank the importance of input variables and their interactions. Having established that the cluster decomposition is a Sobol decomposition, means that we can directly leverage results from Sensitivity Analysis to obtain better insights into the mean cluster interactions of a given lattice Hamiltonian.

ANOVA of a diatomic molecule

Before discussing the general application of the Sobol/f-ANOVA decomposition and Sobol indices to lattice Hamiltonians, let us provide further intuition and motivation by way of a simple example. Consider the configurations of a symmetric diatomic molecule with binary degrees of freedom $\Omega = \{1, 2\}$ as depicted in Figure 2.12. We can express the Hamiltonian, i.e. the energy of each possible configuration, in a 2×2 matrix,[28]

$$H(\sigma_1, \sigma_2) = \begin{bmatrix} \epsilon_{11} & \epsilon_{12} \\ \epsilon_{21} & \epsilon_{22} \end{bmatrix}$$

where the ϵ_{ij} represents the energy of a configuration where the first site is in state $\sigma_1 = i$ and the second in state $\sigma_2 = j$. By our assumption that the molecule is symmetric $\epsilon_{12} = \epsilon_{21}$.

The ANOVA decomposition of the molecule is expressed as follows,

$$H(\boldsymbol{\sigma}) = H_0 + H_1(\sigma_1) + H_2(\sigma_2) + H_{12}(\sigma_1, \sigma_2) \tag{2.51}$$

where each term is computed as,

$$\begin{aligned}
\textbf{mean:} \quad & H_0 = \mathbb{E}\left[H(\boldsymbol{\sigma})\right] = \frac{1}{4}(\epsilon_{11} + \epsilon_{21} + \epsilon_{12} + \epsilon_{22}) \\
\textbf{main effect 1:} \quad & H_1(\sigma_1) = \mathbb{E}\left[H(\boldsymbol{\sigma})|\sigma_1\right] - H_0 \\
\textbf{main effect 2:} \quad & H_2(\sigma_2) = \mathbb{E}\left[H(\boldsymbol{\sigma})|\sigma_2\right] - H_0 \\
\textbf{interaction \{1,2\}:} \quad & H_{12}(\sigma_1, \sigma_2) = H(\sigma_1, \sigma_2) - H_1(\sigma_1) - H_2(\sigma_2) - H_0 \\
& \qquad = H(\sigma_1, \sigma_2) - \mathbb{E}\left[H(\boldsymbol{\sigma})|\sigma_1\right] - \mathbb{E}\left[H(\boldsymbol{\sigma})|\sigma_2\right] + \mathbb{E}\left[H(\boldsymbol{\sigma})\right]
\end{aligned}$$

Finally, computing the above values for the Hamiltonian matrix in Equation 2.4 gives the following explicit expressions for each term,

$$\begin{aligned}
\textbf{mean:} \quad & H_0 = \begin{bmatrix} \frac{1}{4}(\epsilon_{11} + 2\epsilon_{12} + \epsilon_{22}) & \frac{1}{4}(\epsilon_{11} + 2\epsilon_{12} + \epsilon_{22}) \\ \frac{1}{4}(\epsilon_{11} + 2\epsilon_{12} + \epsilon_{22}) & \frac{1}{4}(\epsilon_{11} + 2\epsilon_{12} + \epsilon_{22}) \end{bmatrix} \\
\textbf{main effect 1:} \quad & H_1(\sigma_1) = \begin{bmatrix} \frac{1}{4}(\epsilon_{11} - \epsilon_{22}) & \frac{1}{4}(\epsilon_{11} - \epsilon_{22}) \\ \frac{1}{4}(\epsilon_{22} - \epsilon_{11}) & \frac{1}{4}(\epsilon_{22} - \epsilon_{11}) \end{bmatrix} \\
\textbf{main effect 2:} \quad & H_2(\sigma_1) = \begin{bmatrix} \frac{1}{4}(\epsilon_{11} - \epsilon_{22}) & \frac{1}{4}(\epsilon_{22} - \epsilon_{11}) \\ \frac{1}{4}(\epsilon_{11} - \epsilon_{22}) & \frac{1}{4}(\epsilon_{22} - \epsilon_{11}) \end{bmatrix} \\
\textbf{interaction \{1,2\}:} \quad & H_{12}(\sigma_1) = \begin{bmatrix} \frac{1}{4}(\epsilon_{11} + \epsilon_{22} - 2\epsilon_{12}) & \frac{1}{4}(2\epsilon_{12} - \epsilon_{11} - \epsilon_{22}) \\ \frac{1}{4}(2\epsilon_{12} - \epsilon_{11} - \epsilon_{22}) & \frac{1}{4}(\epsilon_{11} + \epsilon_{22} - 2\epsilon_{12}) \end{bmatrix}
\end{aligned}$$

[28] We do this to follow the convention used in the *tabular* ANOVA.

Now if we use the encoding $(\sigma_i, \sigma_j) \in \{\pm 1\}^2$ for the configuration string under a uniform probability $\rho = 1/2$, the ANOVA expression can be expressed as follows,

$$\begin{aligned} H(\boldsymbol{\sigma}) &= \frac{1}{4}(\epsilon_{11} + 2\epsilon_{12} + \epsilon_{22}) + \frac{1}{4}(\epsilon_{22} - \epsilon_{11})(\sigma_1 + \sigma_2) + \frac{1}{4}(\epsilon_{11} + \epsilon_{22} - 2\epsilon_{12})\sigma_1\sigma_2 \\ &= J_0 + 2J_\circ \frac{\sigma_1 + \sigma_2}{2} + J_{\circ\circ}\sigma_1\sigma_2 \\ &= J_0 + 2J_\circ \Theta_\circ(\boldsymbol{\sigma}) + J_{\circ\circ}\Theta_{\circ\circ}(\boldsymbol{\sigma}) \\ &= J_0 + 2H_\circ(\boldsymbol{\sigma}) + H_{\circ\circ}(\boldsymbol{\sigma}) \end{aligned}$$

where we can immediately recognize that the second-to-final expression above is the Fourier cluster expansion of $H(\boldsymbol{\sigma})$; which in the binary case is trivially re-written as the cluster decomposition in the final line. In addition, by relying on the Fourier expressions for variance, we can identify the meaning of the expansion coefficients as independent variances,

$$\mathrm{Var}_\rho[H(\boldsymbol{\sigma})] = \mathrm{Var}_\rho[H_1(\boldsymbol{\sigma})] + \mathrm{Var}_\rho[H_2(\boldsymbol{\sigma})] + \mathrm{Var}_\rho[H_{12}(\boldsymbol{\sigma})] = 2J_\circ^2 + J_{\circ\circ}^2$$

The resulting Sobol indices then follow directly from their definition in Equation 2.50: $\tau_1 = \tau_2 = \frac{J_\circ^2}{J_\circ^2 + J_{\circ\circ}^2}$ and $\tau_{12} = \frac{J_{\circ\circ}^2}{J_\circ^2 + J_{\circ\circ}^2}$. This means that our expansion coefficients—more precisely the effective cluster weights $\mathrm{W}[H_B]$ for the general case beyond a binary system—directly give us the terms of a variance decomposition of the Hamiltonian.

Mean cluster interactions as independent energy contributions

Now that we have motivated the nature of the cluster decomposition as a Sobol decomposition and foreshadowed the meaning of expansion terms and coefficients, we now proceed to describe them generally and illustrate how these concepts can be used to gain insight and to interpret expansion terms of lattice Hamiltonians.

Based on the construction of f-ANOVA terms as conditional expectation values, we can obtain a more specific understanding of the meaning of mean cluster interactions in a given generalized lattice Hamiltonian. First, let us resolve the structure of *main effects*—those terms that depend on a single occupation variable only. We can re-write the mean cluster interactions H_{B_1} for an orbit B_1 of singleton site space clusters as follows,

$$H_{B_1}(\boldsymbol{\sigma}) = \frac{1}{m_{B_1} N} \sum_{i \in B_1} \mathbb{E}_{\boldsymbol{\rho}}\left[H(\boldsymbol{\sigma}) | \sigma_i\right] = \mathbb{E}_{\boldsymbol{\rho}}\left[H(\boldsymbol{\sigma}) | \sigma_{i \in B_1}\right] \tag{2.52}$$

where we have used the fact that all conditional expectations are equal by symmetry, and there are a total of $|B_1| = m_{B_1} N$ sites in the orbit B_1.

We can see from Equation 2.52 that the *main effects* in the cluster decomposition represent the conditional expectation of the Hamiltonian conditioned on the occupancy of a single site. In other words, the main effects are the contribution that a specific occupancy σ_i on the i-th site has on the total energy (normalized per unit cell). Taken altogether, all

main effects or point terms of the cluster decomposition H_{B_1} represent the portion of the Hamiltonian $H(\boldsymbol{\sigma})$ that depends on composition only.

Similarly, we can write out the f-ANOVA form of a higher order mean cluster interaction H_B as follows,

$$H_B(\boldsymbol{\sigma}) = \frac{1}{m_B N} \sum_{S \in B} \left(\mathbb{E}_{\boldsymbol{\rho}} \left[H(\boldsymbol{\sigma}) | \boldsymbol{\sigma}_S \right] \right) - \sum_{D \sqsubset B} H_D(\boldsymbol{\sigma}) = \mathbb{E}_{\boldsymbol{\rho}} \left[H(\boldsymbol{\sigma}) | \boldsymbol{\sigma}_{S \in B} \right] - \sum_{D \sqsubset B} H_D(\boldsymbol{\sigma}) \quad (2.53)$$

where we subtract all of the mean cluster interactions for orbits D of subclusters of the clusters in B.

Equation 2.53 clarifies the meaning of a mean cluster interaction H_B as the contribution to the total energy coming solely from a single cluster $S \in B$ and none of its subclusters. Thus we see that the terms in the cluster decomposition generally represent the *energetic interactions* between species for all the configurations of sites in a cluster. These energetic interacations are composed of chemical interactions and possibly elastic interactions when structural relaxations are included during the expansion parameter estimation process.

In addition to allowing us to determine the energy contribution of each cluster, the cluster decomposition allows us to formally rank the importance of each contribution by the same prescription of Sobol's indices [206], which we call *cluster sensitivity indices*.

Definition 2.4.4 (Total cluster sensitivity index)**.** *For a given Hamiltonian $H \in L^2(\boldsymbol{\Omega}, \boldsymbol{\rho})^{\mathcal{G}}$, the cluster sensitivity index τ_B is the fraction of the total variance of H contributed by the mean cluster interaction H_B multiplied by its multiplicity m_B,*

$$\tau_B = \frac{m_B N \operatorname{Var}_\rho[H_B(\boldsymbol{\sigma})]}{\operatorname{Var}_\rho[H(\boldsymbol{\sigma})]} \quad (2.54)$$

The total cluster sensitivity index τ_B represents the variance contributed by the interactions of all clusters in the orbit B per unit cell. We can also define the *effective* cluster sensitivity index which represents the portion of the variance carried by only a single cluster in the given orbit normalized per unit cell.

Definition 2.4.5 (Effective cluster sensitivity index)**.** *For a given Hamiltonian $H \in L^2(\boldsymbol{\Omega}, \boldsymbol{\rho})^{\mathcal{G}}$, the effective cluster sensitivity index $\bar{\tau}_B$ is the fraction of the total variance of H carried by only one cluster $S \in B$ from the mean cluster interaction H_B.*

$$\bar{\tau}_B = \frac{\tau_B}{m_B} = \frac{N \operatorname{Var}_\rho[H_B(\boldsymbol{\sigma})]}{\operatorname{Var}_\rho[H(\boldsymbol{\sigma})]} \quad (2.55)$$

More so, using Equation 2.38 cluster sensitivities can be calculated directly from a cluster decomposition as the ratios of effective cluster weights,

$$\tau_B = \frac{m_B \operatorname{W}[H_B]}{\sum_B m_B \operatorname{W}[H_B]} \quad (2.56)$$

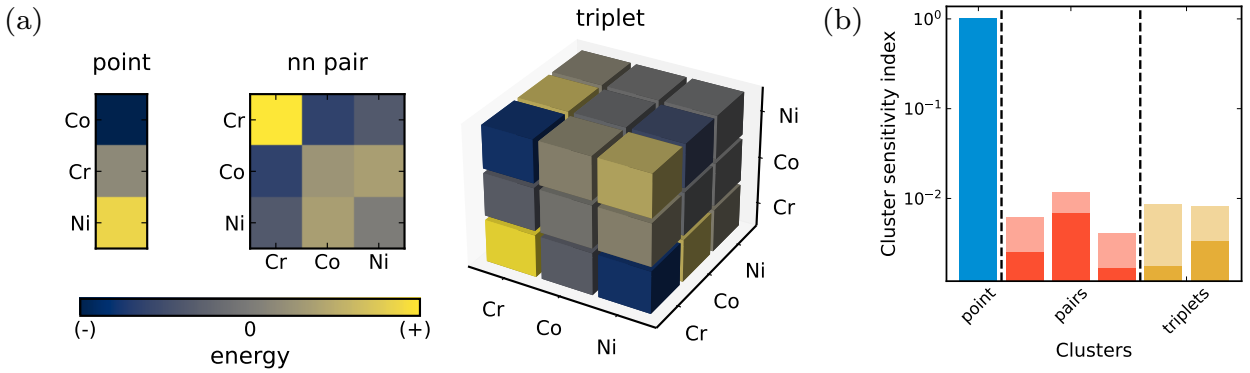


Figure 2.13: (a) Main species effect, nearest neighbor pair, and triplet mean cluster interaction tensors for a face-centered cubic ternary alloy (Ni−Co−Cr) cluster decomposition. The contribution of each cluster configuration is shown by a symmetric color map centered at 0 (neutral/no contribution). The magnitudes of the interactions are of different orders of magnitude; the main effect point term contributions are of eV magnitude, and higher degree interactions are of meV magnitude. (b) Cluster sensitivity indices for the face-centered cubic ternary alloy cluster decomposition. The point term main effects represent almost the entirety of the energy variance. The most important higher degree terms are the second pair, the triplet interactions.

As an example of interpreting a particular Hamiltonian using the terms of its cluster decomposition, Figure 2.13a shows a visualization of the point, nearest neighbor pair, and a triplet mean cluster interactions for a fitted cluster decomposition of a NiCoCr ternary alloy. The relative magnitudes of the chemical interaction among species can be read directly from the colors used. Values towards the blue (yellow) end of the spectrum give negative (positive) energies and thus favorable (unfavorable) energetic interactions. We can observe the relative importance of each interaction by the corresponding cluster sensitivity indices plotted in Figure 2.13b, where the effective cluster sensitivities $\bar{\tau}_B$ are plotted with solid colors on top of the total cluster sensitivities τ_B shown with transparency. In this example, the point interactions carry are by far the most important, suggesting the biggest contribution to the energy landscape for the NiCoCr ternary alloy is simply composition[29]. This can be further corroborated by inspecting the ternary phase diagram shown in Appendix C.3 from energies calculated with density functional theory and which were used to fit this Hamiltonian. The most important contribution from higher order mean cluster interactions is from the second pair shown, both by a single cluster ($\bar{\tau}_B$) and by the total number of clusters per unit cell (τ_B).

Lastly, having shown how the uniqueness, orthogonality, and the connection of the cluster decomposition to the Sobol decomposition allows formal interpretation of expansion terms,

[29] As is most often the case in any physical system.

we remark that when using a representation with a non-Fourier basis or a mathematical frame we do not obtain these properties, as thus a rigorous interpretation of expansion terms is not directly available. In practice, however, one can always re-write any expansion in the form given in Equation 2.33 and thus have terms that operate over the clusters in given orbits B as follows,

$$H(\boldsymbol{\sigma}) = \sum_{B \in \mathcal{G}\,\mathrm{P}([N])} m_B \widetilde{H}_B(\boldsymbol{\sigma}) \tag{2.57}$$

where each term $\widetilde{H}_B(\boldsymbol{\sigma})v$ is computed according to Equation 2.34 but with functions that are not strictly Fourier correlations.

We will refer to such a representation as an *pseudo cluster decomposition* and call the terms mean *pseudo cluster interactions*. Nevertheless, the values of such terms for non-Fourier basis expansions are basis dependent and will not have ANOVA properties, such that making model interpretations would be insubstantial. However, as we will show in Chapter 3 one can always transform any general cluster expansion into a Fourier cluster expansion in order to allow formal interpretation of the energy contributions of site space clusters.

Expected cluster energies at finite temperatures

We can obtain further insight from the decomposition of expectations and variances at finite temperatures.[30] For the expectation value of the energy (the internal energy) we still have the same decomposition in terms of *finite* temperature mean cluster interactions,

$$\langle H(\boldsymbol{\sigma})\rangle_T = N \sum_{B \in \mathcal{G}\,\mathrm{P}([N])} m_B \langle H_B(\boldsymbol{\sigma})\rangle_T \tag{2.58}$$

In contrast, we do not get a complete decomposition for the variance of the energy at finite temperatures—which is proportional to the heat capacity—because we now need to account for the covariances between mean cluster interactions. The finite temperature variance of a Hamiltonian is explicitly given as follows,

$$\mathrm{Var}_T[H(\boldsymbol{\sigma})] = N^2 \sum_{B \in \mathcal{G}\,\mathrm{P}([N])} \mathrm{Var}_T[m_B H_B(\boldsymbol{\sigma})] + N^2 \sum_{B \neq D} \mathrm{Cov}_T[m_B H_B(\boldsymbol{\sigma}), m_D H_D(\boldsymbol{\sigma})] \tag{2.59}$$

Although a decomposition of variance in independent terms is not obtained at finite temperatures, the expression in Equation 2.59 still allows a useful breakdown of the finite temperature variance. Using such Equation 2.59 we can still gain insight into which cluster interactions contribute most to the total energy at a particular temperature. Figure 2.14

[30] We are now using ensemble or thermodynamic averages with respect to the Boltzmann distribution at finite temperature T.

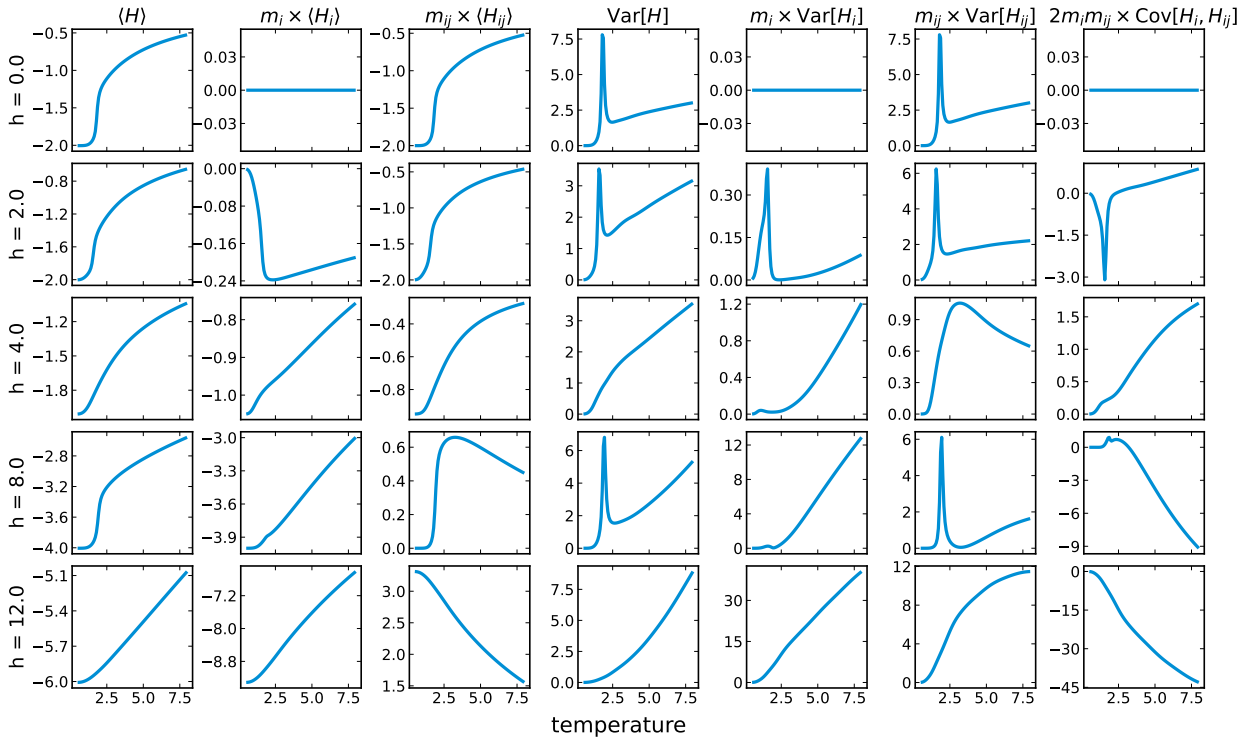


Figure 2.14: Internal energy $\langle H \rangle$, mean cluster decomposition of the internal energy into point and pair interactions $m_i \langle H_i \rangle$ and $m_{ij} \langle H_{ij} \rangle$, total energy variance $\mathrm{Var}[H]$, variance decomposition $m_i \mathrm{Var}[H_i]$ and $m_{ij} \mathrm{Var}[H_{ij}]$, and covariance $2m_i m_{ij} \mathrm{Cov}[H_i, H_{ij}]$ decomposition at finite temperatures for different values of an external field h in an antiferromagnetic face-centered cubic Ising model. For lower fields the majority of the variance is carried by the magnetic pair interactions

shows the internal energy and variance decomposition at finite temperatures for the FCC antiferromagnetic Ising model introduced in Chapter 1.3 for select values of the field h. We observe that for low values of h the energy and variance associated with pair interactions make up the majority of the total values and the covariances remain relatively small. For larger values of h the portion associated with the magnetization or single site energies becomes more important until they fully dominate at the critical point $h = 12$.

We can qualitatively understand this behavior by considering the cluster indices for the two expansion terms. Figure 2.15 shows the phase diagrams for BCC and FCC antiferromagnetic Ising models as well as the Sobol indices for the magnetization energy and the pair interaction energy as a function of the field h. The importance of the effects of each term can be understood, at least qualitatively, from the decay (growth) of the pair sensitivity index (point sensitivity index).

Although our last example using the Ising model may be overly simple, the formal con-

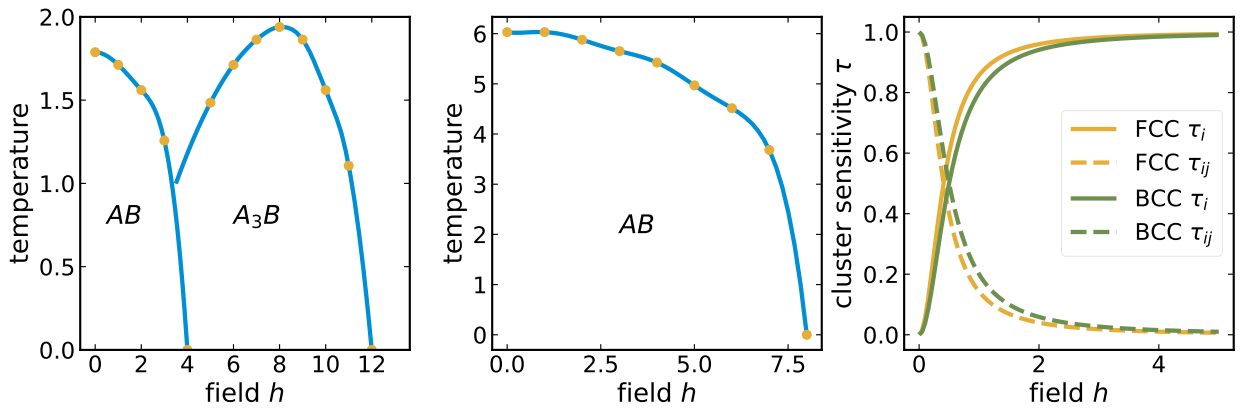


Figure 2.15: Computed phase diagrams and cluster sensitivity indices τ for BCC and FCC antiferromagnetic Ising models

nection here between the cluster decomposition, the Sobol/ANOVA decomposition, and the use of Sobol indices to gain further insight and interpretability of generalized lattice Hamiltonians should open up a realm of possible applications and new methodology development by leveraging the many tools, results and developments of the various related fields, such as Sensitivity Analysis, ANOVA and other statistical diagnostics [65, 84, 96, 97, 101, 206, 208].

2.5 Potts frame expansions

We have detailed a procedure to construct Fourier basis sets for the space of functions over atomic configurations $L^2(\mathbf{\Omega}, \boldsymbol{\rho})$. We have further shown how Fourier basis sets have useful formulas that allow us to immediately obtain statistical properties, such as the mean, variance, and covariances between functions under an *a-priori* distribution and a finite temperature semigrand canonical Boltzmann distribution. Additionally, we have shown that all the possible Fourier basis sets represent the same decomposition for any given Hamiltonian, which we have used to reveal how to obtain even more detailed statistical properties and interpretability in applied lattice models.

Although the analytical properties and resulting interpretations permitted by Fourier basis sets and the cluster decomposition are profoundly useful, using such representations to fit a Hamiltonian in practice, can sometimes be difficult and inefficient; and in some cases, can also limit the predictive accuracy of the fitted Hamiltonian. In broad terms, such effects show up in practice precisely due to the restrictive conditions that provide useful analytic properties. Specifically, the requirements of orthonormality and linear independence can be over-restrictive in certain practical applications. For example, since correlation functions are orthogonal/uncorrelated, the robustness of a fitted Hamiltonian can be severely compromised if the coefficient for a single but important correlation function is not accurately recovered

or missed altogether from a finite and imperfectly sampled training data set.

Many of the practical limitations of using basis sets can be addressed simply by dropping the requirements of linear independence and orthonormality when constructing a spanning set of functions, and instead using what is known as a mathematical *frame*. We can think of a frame simply as a basis set with additional functions included, such that the set will always span the original space, but allows some level of *redundancy* (i.e. *more* functions than needed). Including redundancy in the representation of applied lattice Models can provide additional robustness and in some cases less restrictive training data requirements.

Definition 2.5.1 (Frame)**.** *A countable sequence* $\{\Phi_\gamma\}_{\gamma\in I}$ *is said to be a frame for a Hilbert space* $\mathcal{H}$ *if there exist frame bounds* $A, B > 0$ *such that,*

$$A||F||^2 \leq \sum_{\gamma\in I} |\langle F, \Phi_\gamma\rangle|^2 \leq B||F||^2 \quad \forall F \in \mathcal{H}$$

The definition above implies that the frame $\{\Phi_\gamma\}_{\gamma\in I}$ spans $\mathcal{H}$. [43, 235]. For our purposes, we will take the Hilbert space $\mathcal{H}$ to be $L^2(\boldsymbol{\Omega}, \boldsymbol{\rho})$. Since a frame by definition spans the space of functions over configurations $L^2(\boldsymbol{\Omega}, \boldsymbol{\rho})$, any function can be expanded in terms of the functions making up the frame,

$$F(\boldsymbol{\sigma}) = \sum_{\gamma\in I} J_\gamma \Phi_\gamma(\boldsymbol{\sigma}) \tag{2.60}$$

In contrast to an expansion in an orthonormal basis, such as a Fourier expansion, the expansion coefficients J_γ are not necessarily projections of the function F onto the expansion functions Φ_γ. Furthermore, there are infinitely many sets of expansion coefficients that can be used to represent the same function F with the same frame $\{\Phi_\gamma\}_{\gamma\in I}$. Actually, a basis is itself a frame where the number of functions included is exactly the dimension of the space being spanned, (in our case $|I| = \dim(L^2(\boldsymbol{\Omega}, \boldsymbol{\rho}))$). However, in order to benefit from redundancy, we will construct a frame that includes many more functions than the dimension of the space, i.e. $|I| > \dim(L^2(\boldsymbol{\Omega}, \boldsymbol{\rho}))$.

To get a better understanding and motivate the rest of this section, consider the simple case of representing vectors in $\mathbb{R}^2$. We are well aware that we can *uniquely* represent any vector $\vec{v} \in \mathbb{R}^2$ using the canonical orthonormal basis $\{\hat{\imath}, \hat{\jmath}\}$, as shown on the left side in Figure 2.16. But we can also use a set of 3 non-collinear vectors to represent the same vector $\vec{v}$. In particular we can use what is called the *Mercedes-Benz* (MB) frame [43, 235], which is shown on the right in Figure 2.16. We only show one out of the infinitely many possible sets of coefficients representing $\vec{v}$ with the MB frame. In Figure 2.16 an orange dot is also shown on the $\hat{\imath}$ basis vector and on the $\hat{\epsilon}_2$ vector. The orange dot represents the best approximation of $\vec{v}$ if only that single vector was used; say for example if we were unable to recover the coefficients of the other vectors properly. The approximation error in both cases is equal to the length of the dotted line from the tip of $\vec{v}$ to the orange dot. In the example, the approximation using the MB frame is substantially more robust since it has

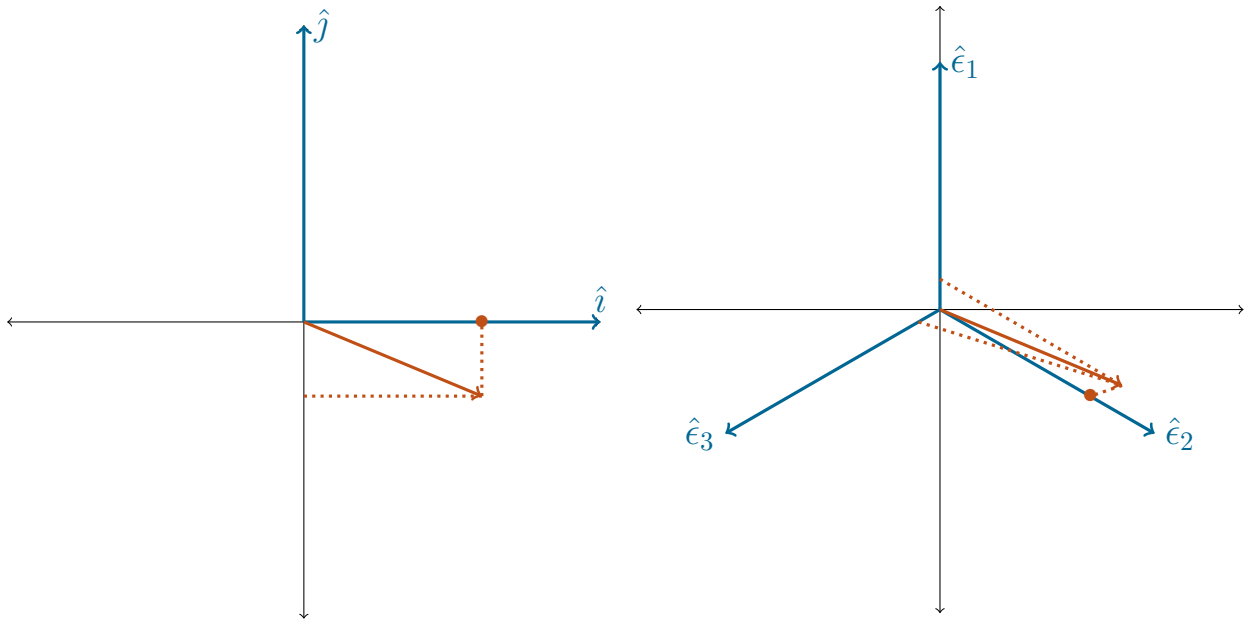


Figure 2.16: Canonical basis for $\mathbb{R}^2$ (left) and the *Mercedes-Benz* frame spanning $\mathbb{R}^2$ (right). The same vector $\vec{v}$ is shown in orange, and a representation of $\vec{v}$ is depicted by the intersection of the dotted lines from the tip of $\vec{v}$ to each of the spanning vectors. For the canonical basis, the representation of $\vec{v}$ has two unique coefficients. In contrast, only one infinitely many sets of coefficients that can be used to represent $\vec{v}$ are shown. An orange dot is shown for both representations, as the best approximation to $\vec{v}$ if only the single vector depicted was used.

a much lower error. We can also interpret this example as an illustration of how a vector can often have a much better *sparse* approximation when using a frame over a basis—in this example, by using only one vector in two dimensions. Although this is only a single example[31], the general intuition is that *more often than not* a carefully chosen frame can result in more robust and better sparse approximations of vectors, or in our present case lattice Hamiltonians of atomic configuration.

In the remainder of this chapter, we construct a particular frame that spans the space functions over configurations $L^2(\boldsymbol{\Omega}, \boldsymbol{\rho})$, and that closely resembles the Potts frame introduced in Chapter 1.3. Using this redundant representation to fit Hamiltonians provides numerous practical advantages, as we will discuss in Chapter 4.6, and show examples in Chapter 5.3. However, since the *redundancy* implies a lack of *uniqueness* are forced to abandon the useful analytic properties that a Fourier basis representation has. Nevertheless, we can always use the Frame representation to fit a particular Hamiltonian when it is convenient and afterward convert the Hamiltonian to its corresponding *unique* cluster decomposition when analytical

[31] We could certainly also come up with many examples where the representation using the basis comes out as more robust/sparse

expressions for statistical properties and interpretations are sought. We describe practical procedures to convert any representation into a Fourier cluster expansion in Chapter 3.

Cluster expansions with site indicator basis functions

Before introducing the generalized Potts frame, we give a brief exposition of the process of constructing a product basis using a particular choice of *non-standard* site basis. We focus on one commonly used basis set composed of site *indicator* functions,[32] [258].

Definition 2.5.2 (Site indicator function)**.** *A site indicator function* $\mathbf{1}_{\sigma_j} \in L^2(\Omega, \rho)$ *is a function that* indicates *whether a given species occupies a site or not,*

$$\mathbf{1}_{\sigma_j}(\sigma_i) = \begin{cases} 1 & \text{if } \sigma_i = \sigma_j \\ 0 & \text{otherwise} \end{cases} \tag{2.61}$$

Since the set of basis functions must include the constant function in order to yield basis functions that follow a cluster framework, the constant function needs to replace an indicator function for one of the species at each site. As a result, one of the total n allowed species at each site will not have an associated indicator function [258]. The sets of site basis functions $\{\phi_j; j = 0, \ldots, n-1\}$ in terms of site indicator functions are then given by,

$$\phi_j(\sigma_i) = \begin{cases} 1 & \text{if } j = 0 \\ \mathbf{1}_{\sigma_j}(\sigma_i) & \text{if } j = 1, \ldots, n-1 \end{cases} \tag{2.62}$$

Following the procedure described in Section 2.3, the correlation functions are constructed from symmetry adapted averages of the following N-fold products of site indicator functions,

$$\Phi_{\boldsymbol{\alpha}}(\boldsymbol{\sigma}) = \prod_{i \text{ s.t. } \alpha_i \neq 0} \mathbf{1}_{\sigma_{\alpha_i}}(\sigma_i), \tag{2.63}$$

where we notice that by construction, the functions in Equation 2.63 will indicate whether a specific occupancy of a given cluster represented by the nonzero elements of the multi-index $\text{supp}(\boldsymbol{\alpha})$ is present in a structure. We can more briefly write Equation 2.63 as a *cluster* indicator function.

Definition 2.5.3 (Cluster indicator function)**.** *A cluster indicator function* $\mathbf{1}_{\boldsymbol{\alpha}}(\boldsymbol{\sigma}) \in L^2(\boldsymbol{\alpha}, \boldsymbol{\rho})$*, is an N-fold product function of site indicator functions,*

$$\mathbf{1}_{\boldsymbol{\alpha}}(\boldsymbol{\sigma}) = \begin{cases} 1 & \text{if } \boldsymbol{\sigma}_{\text{supp}(\boldsymbol{\alpha})} = \text{ctr}(\boldsymbol{\alpha}) \\ 0 & \text{otherwise} \end{cases} \tag{2.64}$$

[32] Site indicator functions are also referred to as site *occupancy* functions.

Furthermore, since the final correlation functions are constructed from averages of functions given by Equation 2.64 over symmetrically equivalent clusters, the correlation functions represent (up to multiplicity constants) the *concentration* of a specific occupancy of clusters for the different crystallographic orbits.

Definition 2.5.4 (Cluster indicator correlation function)**.** *A cluster indicator correlation function* $I_\beta \in L^2(\boldsymbol{\Omega}, \boldsymbol{\rho})$ *is a symmetrically invariant average of cluster indicator functions* $\mathbf{1}_{\boldsymbol{\alpha}}$*,*

$$I_\beta(\boldsymbol{\sigma}) = \frac{1}{m_\beta N_{\boldsymbol{\sigma}}} \sum_{\boldsymbol{\alpha} \in \beta} \mathbf{1}_{\boldsymbol{\alpha}}(\boldsymbol{\sigma}) \tag{2.65}$$

where β *represents an orbit of site indicator function clusters (i.e. with each site labeled with a particular indicator function), and* $m_\beta N_\sigma$ *is the total number of labeled clusters in the orbit* β*. The sum is carried over all labeled clusters* $\boldsymbol{\alpha}$ *that are part of the orbit* β*.*

We make two notable observations regarding correlation functions with site indicator basis functions. The first is that the basis sets in Equation 2.62 are not orthogonal and as a result, the resulting correlation functions in Equation 2.65 are not orthogonal either. This lack of orthogonality can further complicate constructing highly incoherent measurement matrices compared to using orthogonal/orthonormal basis sets.

The second observation is concerned with the set of clusters/orbits that are selected by the correlation functions in Equation 2.65. The set of correlation functions in Equation 2.65 indeed represents a basis for the function space over all possible configurations $\boldsymbol{\sigma}$ and thus the set is linearly independent [258]. As a consequence, however, the included functions do not give the concentrations of all possible occupied clusters. Namely, the correlation functions in Equation 2.65 never include any occupied clusters that involve the species that do not have an associated indicator function in the corresponding site basis functions. Additionally, the number of clusters not indicated for in the cluster indicator basis grows quickly with the number of components. In fact, the number grows as $O\left(n^{N_{\boldsymbol{\alpha}}} - (n-1)^{N_{\boldsymbol{\alpha}}}\right)$, where n is the number of allowed species at a site and $N_{\boldsymbol{\alpha}}$ is the number of sites in a cluster.

Notwithstanding these observations, cluster expansions using site indicator functions formally constitute a basis for function spaces over crystalline configurations, and have been successfully used in the study of configurational thermodynamics for some time [165, 181, 258]. Such an expansion corresponds to direct generalizations of the lattice gas model. However, as alluded to before, the lack of orthogonality complicates obtaining incoherent measurements to maximize accurate ECI estimation from classical CS. Additionally, the choice of species left out in site basis sets is mathematically meaningless, but can often lead to precarious interpretations of fitted coefficients. This begs the question of whether we can do away with using cluster indicator basis sets and seeking maximally incoherent measurements, yet still obtain suitably sparse and accurate expansions by instead relying on redundancy by including functions labeled for all possible occupations over the included clusters.

The generalized Potts frame

We introduce a specific redundant set of functions that spans the same function $L^2(\mathbf{\Omega}, \boldsymbol{\rho})$. The redundant set of functions can be obtained simply by including cluster indicator functions of the form in Equation 2.63 for all possible occupations. More formally we build all the N-fold product functions from redundant site function sets that include $\phi_0 \equiv 1$ and site indicator functions for *all* allowed species at each site, such that for each site with n allowed species, we associate a redundant set of functions given by,

$$\phi_j(\sigma_i) = \begin{cases} 1 & \text{if } j = 0 \\ \mathbf{1}_{\sigma_j}(\sigma_i) & \text{if } j = 1, \ldots, n \end{cases} \tag{2.66}$$

This results in $n+1$ total functions for a space of dimension n, where the redundancy is the trivial linear relation,

$$\phi_0(\sigma) - \sum_{i=1}^{n} \mathbf{1}_{\sigma_i}(\sigma) = 0 \tag{2.67}$$

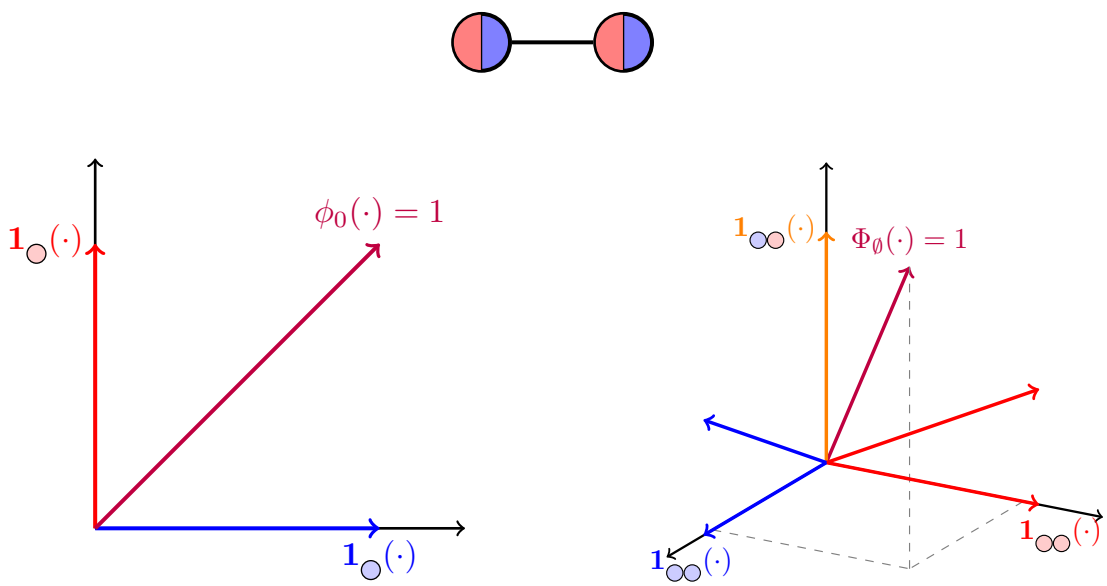
Carrying out the same operation of taking all possible N-fold products of functions from each redundant set and taking symmetry adapted averages over crystallographic orbits results in the set of symmetrized product functions which give *concentrations* for any possible cluster occupancy. This resulting set is highly redundant since the total number of functions obtained are of order $O\left((n+1)^N\right)$ and the dimension of the function space is of order $O(n^N)$. In other words the combinatorics [87] for identifying symmetrically equivalent clusters involve n possible *labels* of basis functions for each site for all clusters in distinct orbits compared with the $n-1$ *labels* involved in a cluster indicator basis. Nonetheless, the resulting set of functions spans the same function space $L^2(\mathbf{\Omega}, \boldsymbol{\rho})$ and therefore formally the set constitutes a *frame* [43, 235]. We provide a derivation for a set of frame bounds in Appendix B.3, but make no effort to optimize them. The resulting frame has a strong connection to the well-known Potts model [176]. Indeed it is a (normalized) generalization in both spatial extent and interaction size of the original nearest neighbor pair Potts model introduced in Chapter 1.3 [250]. Hence we refer to the proposed frame as the *generalized Potts frame*.

Definition 2.5.5 (Generalized Potts frame)**.** *The generalized Potts frame for $L^2(\mathbf{\Omega}, \boldsymbol{\rho})$ is the sequence of all possible cluster indicator functions $\{\mathbf{1}_{\boldsymbol{\alpha}} \mid \forall \boldsymbol{\alpha} \in \mathbb{N}^N_{\leq \boldsymbol{n}}\}$ constructed from products of local site frames which include a constant function and site indicator functions for each of the allowed species in the corresponding site space.*

Definition 2.5.6 (Symmetrized Potts frame)**.** *A symmetrized Potts frame for a symmetrically invariant subspace of $L^2(\mathbf{\Omega}, \boldsymbol{\rho})$ to operations in a given symmetry group $\mathcal{G}$ is the sequence of all possible cluster indicator correlation functions $\{I_\beta \mid \forall \beta \in \mathcal{G}(\mathbb{N}^N_{\leq \boldsymbol{n}})\}$.*

Any lattice Hamiltonian $H \in L^2(\mathbf{\Omega}, \boldsymbol{\rho})^{\mathcal{G}}$ can thus be expressed in terms of a symmetrized Potts frame as,

$$H(\boldsymbol{\sigma}) = N \sum_{\beta \in \mathcal{G}(\mathbb{N}^N_{\leq \boldsymbol{n}})} m_\beta J_\beta I_\beta(\boldsymbol{\sigma}) \tag{2.68}$$



$$H(\sigma) = J_{\emptyset} + J_{\circ}\mathbf{1}_{\circ}(\sigma) + J_{\circ\circ}\mathbf{1}_{\circ\circ}(\sigma)$$
$$H(\sigma) = J_{\emptyset} + J_{\circ}\mathbf{1}_{\circ}(\sigma) + J_{\circ}\mathbf{1}_{\circ}(\sigma) + J_{\circ\circ}\mathbf{1}_{\circ\circ}(\sigma) + J_{\circ\circ}\mathbf{1}_{\circ\circ}(\sigma) + J_{\circ\circ}\mathbf{1}_{\circ\circ}(\sigma)$$

Figure 2.17: Function space representations over the configurations of a single binary site and a symmetric binary diatomic molecule. (a) Function space over a single binary site space. The two different choices for site bases to construct a cluster expansion are colored red and blue, and each set also includes the purple $\phi_0 \equiv 1$ function. (b) Function space over symmetrically distinct configurations of the molecule. A cluster expansion basis includes either the blue-colored or the red-colored functions. The generalized Potts frame includes all colored functions (blue/red/yellow). All function sets also include the magenta-colored constant function.

Where it is clear that the total number of expansion terms I_β far surpasses the dimensional of $|\mathcal{G}(\mathbb{N}^N_{\leq \boldsymbol{n}})| > \dim(L^2(\boldsymbol{\Omega}, \boldsymbol{\rho})) = |\mathcal{G}(\mathbb{N}^N_{<\boldsymbol{n}})|$.

As a simple illustration of the geometry of a symmetrized Potts frame, consider a symmetric binary diatomic system. Figure 2.17a shows the corresponding site spaces and site indicator functions along with the constant function ϕ_0. Additionally, the resulting symmetrized product bases are also shown in Figure 2.17b. The basis functions in both cases are colored to represent the possible cluster indicator function bases. The union of all basis functions from these cluster indicator function bases corresponds to the symmetrized generalized Potts frame for this simple diatomic system. Again we highlight also the inclusion of products of *mixed* site basis sets, which corresponds to the orange vector built from the

product of a red and blue basis function associated with each site respectively.

The generalized Potts frame can also be seen as the union of every possible cluster indicator basis. All possible cluster indicator bases can be generated by cycling over the species that is not *indicated* for in its corresponding site basis and building the corresponding cluster indicator correlation basis for every possible combination of site basis sets. We make a special note that apart from the normal cluster indicator basis sets, this includes correlations of *mixed* products where symmetrically equivalent sites in the underlying random structure can have distinct basis sets, i.e. a different subset of species with associated indicator functions. Symbolically, the generalized Potts frame includes all the cluster indicator functions in the following set,

$$\bigcup_{\sigma} \left\{ \mathbf{1}_{\boldsymbol{\alpha}};\ \forall \boldsymbol{\alpha} \ \text{ s.t. } \ \text{ctr}(\boldsymbol{\alpha}) \neq \boldsymbol{\sigma}_{\text{supp}(\boldsymbol{\alpha})} \right\} \tag{2.69}$$

As alluded to previously, by using the generalized Potts Frame to represent Hamiltonians of configuration we lose the useful properties of a Fourier cluster expansion, however, we will see in Chapter 4.6 and the results given in Chapter 5.3 that using the generalized Potts Frame in the process of learning a Hamiltonian for a real system can result in more accurate, more robust and sparser approximations than can be obtained using a Fourier cluster expansion. Additionally, we also show in Chapter 3.2 how one can convert a fitted Hamiltonian using the generalized Potts Frame to its corresponding Fourier cluster expansion in order to allow a formal analysis of the Hamiltonian using the methods we have developed in Chapter 2.4.

Chapter 3

Practical numerical implementations

Before delving into the details of fitting generalized lattice models, we describe practical and efficient ways to implement the mathematical framework and formalism developed in Chapter 2. We propose implementations to numerically represent Hamiltonians of configuration in applied lattice models, to efficiently convert between representations, and to obtain cluster configuration statistics directly from expansion functions. The methodology and implementations discussed here represent only one of many ways to do so. However, we have found that these implementations generalize, simplify and improve the space and time complexity compared to most other available implementations [2, 38, 132, 177, 226, 230]. All of the methodology described in this chapter has been implemented and is openly available [11].

3.1 Reduced correlation tensors and cluster interaction tensors

When implementing the different representations (cluster or frame expansions) of functions in $L^2(\mathbf{\Omega}, \boldsymbol{\rho})$, it is advantageous to leverage the *cluster-like* framework and treat the product functions introduced in Equations 2.20 and 2.64 explicitly as functions only over the sites in their corresponding support $\text{supp}(\boldsymbol{\alpha})$, i.e. as functions over the configurations of a single site space cluster $S = \text{supp}(\boldsymbol{\alpha})$. Furthermore, by symmetry we will have that product functions with multi-indices in the same orbit β, are for practical purposes the exact same function—they simply operate over the sites of another (symmetrically equivalent) site space cluster. The practical implications of this formality allows one to only deal with a small set of distinct functions—one for each multi-index orbit included in the expansion—of a small number of variables only, regardless of the domain (supercell) size being used in applications. In order to make this notion explicit, we refer to a correlation function of only the sites in a cluster S as a *reduced correlation function*. The domain of a reduced correlation function is the configuration space of a single site space cluster, whose set of site indices is equal to the support of a multi-index $S = \text{supp}(\boldsymbol{\alpha})$ in a given multi-index orbit $\boldsymbol{\alpha} \in \beta$. Such a

configuration space is explicitly expressed as follows,

$$\mathbf{\Omega}_{\text{supp}(\boldsymbol{\alpha})} = \bigtimes_{i:\alpha_i \neq 0} \Omega_i \tag{3.1}$$

where in the above equation we omitted the associated *a-priori* measure, however, it is simply made up of the product of the corresponding site space measures ρ_i in the product.

Definition 3.1.1 (Reduced correlation function)**.** *A reduced correlation function $\hat{\Theta}_\beta$ is a correlation function over the configuration space of a single site space cluster $S = \text{supp}(\boldsymbol{\alpha})$ for any multi-index $\boldsymbol{\alpha} \in \beta$,*

$$\hat{\Theta}_\beta : \mathbf{\Omega}_{\text{supp}(\boldsymbol{\alpha})} \to \mathbb{R} \tag{3.2}$$

Equivalently, a reduced correlation function is an element of the function space over configurations of a single site space cluster, $\hat{\Theta}_\beta \in L^2(\mathbf{\Omega}_{\text{supp}(\boldsymbol{\alpha})})$

The corresponding correlation function for a complete structure, can be computed simply by averaging the reduced correlation functions over all of the site clusters in the associated orbit $B = \{\text{supp}(\alpha) : \forall \boldsymbol{\alpha} \in \beta\}$,

$$\Theta_\beta(\boldsymbol{\sigma}) = \frac{1}{|B|} \sum_{S \in B} \hat{\Theta}_\beta(\boldsymbol{\sigma}_S) \tag{3.3}$$

Reduced correlation functions are practically useful because their input is only a small set of variables, and so they are fast to compute and memory efficient. More importantly, reduced correlation functions can be built efficiently and stored in memory such that their evaluation is reduced to a simple data access operation. This can be made more precise by considering the formal definition of the domain $\mathbf{\Omega}_{\text{supp}(\alpha)}$ for a given set of reduced correlation functions, and re-invoking the construction of basis functions for any tensor product space in Chapter 2. For the case of reduced correlation functions, the product space is relatively manageable since it consists only of a few site spaces Ω_i. As a result, it is straightforward to compute the product functions $\Phi_{\text{ctr}(\boldsymbol{\alpha})}$ directly as tensor products of single site functions as follows,

$$\Phi_{\text{ctr}(\boldsymbol{\alpha})} = \bigotimes_{\alpha_i \in \text{ctr}(\boldsymbol{\alpha})} \phi^{(i)}_{\alpha_i} \tag{3.4}$$

where the product is now only over the elements of the reduced multi-index $\text{ctr}(\boldsymbol{\alpha})$. Furthermore, since we are dealing with discrete domains, each site function is isomorphic to a real vector $\phi^{(i)}_{\alpha_i} \in \mathbb{R}^{|\Omega_i|}$, where the dimension is the size of the site space $|\Omega_i|$. Accordingly, the tensor products in Equation 3.4 can be computed efficiently using real space vectors. The resulting product function is simply a Cartesian tensor $\Phi_{\text{ctr}(\boldsymbol{\alpha})} \in \mathbb{R}^{\times_{i \in \text{supp}(\boldsymbol{\alpha})} |\Omega_i|}$. Equivalently, we can also think of the product function as a vector itself, $\Phi_{\text{ctr}(\boldsymbol{\alpha})} \in \mathbb{R}^{|\mathbf{\Omega}_{\text{supp}(\boldsymbol{\alpha})}|}$ by simply arranging all elements in a consistent manner—for example in lexicographical order.[1]

[1] In fact, Fourier product basis functions, correlation functions, cluster interactions, and full Hamiltonian can be thought of as high dimensional real space vectors, $H \in \mathbb{R}^{|\mathbf{\Omega}|}$, where the dimension is the size of the total configuration space $\mathbf{\Omega}$. Or equivalently high order Cartesian tensors in $\mathbb{R}^{\times_i^N |\Omega_i|}$.

Following Chapter 2, a reduced correlation function is obtained by averaging over product functions given by Equation 3.4 over symmetrically equivalent reduced multi-indices, i.e. those belonging to the same orbit $\hat{\beta} = \{\text{ctr}(\boldsymbol{\alpha}) \; \forall \boldsymbol{\alpha} \in \beta\}$.[2]

$$\hat{\Theta}_\beta = \frac{1}{|\hat{\beta}|} \sum_{\hat{\boldsymbol{\alpha}} \in \hat{\beta}} \Phi_{\hat{\boldsymbol{\alpha}}} \tag{3.5}$$

Therefore, we can also think of reduced correlation functions as either tensors or vectors, since they are explicitly constructed as linear combinations of tensors/vectors of the form given in Equation 3.4. To make the nature of treating reduced correlation functions as tensors more apparent, we will write them out between square brackets $\left[\hat{\Theta}_\beta\right]$, and refer to them as *correlation tensors*.

It is now evident that when using pre-computed *correlation tensors* evaluating a correlation function reduces to a data access operation, specifically an array access operation. For example, evaluating a reduced correlation function over a cluster of only three sites simplifies to accessing an element of a rank three tensor (a three-dimensional array),

$$\hat{\Theta}_\beta(\sigma_i, \sigma_j, \sigma_k) = \left[\hat{\Theta}_\beta\right]_{\sigma_i \sigma_j \sigma_k} \tag{3.6}$$

And correspondingly, to compute the value of a correlation function for an arbitrary configuration $\boldsymbol{\sigma}$ of a full structure, we simply average the reduced tensor entry for each site cluster $S \in B$, where the orbit $B = \{\text{supp}(\boldsymbol{\alpha}), \; \forall \boldsymbol{\alpha} \in \beta\}$. For example, the value of a triplet cluster correlation function is computed from its reduced correlation tensor as follows,

$$\begin{aligned} \Theta_\beta(\boldsymbol{\sigma}) &= \frac{1}{|B|} \sum_{\{i,j,k\} \in B} \left[\hat{\Theta}_\beta\right]_{\sigma_i \sigma_j \sigma_k} \\ &= \frac{1}{|B|} \sum_{S \in B} \left[\hat{\Theta}_\beta\right]_{\boldsymbol{\sigma}_S} \end{aligned} \tag{3.7}$$

This practical implementation of correlation functions can be suitably extended to express cluster interactions in a similarly useful manner. In fact, we have already relied on the concept of the interaction of single clusters in establishing the interpretation of cluster interactions in Chapter 2.4. We can thus directly represent a single *cluster interaction* in terms of reduced correlation functions.

Definition 3.1.2 (Cluster interaction)**.** *A cluster interaction $\hat{H}_B$ represents the energy of a single site space cluster $S \in B$ for a given lattice Hamiltonian H. A reduced cluster*

[2] The seemingly sloppy use of β vs $\hat{\beta}$ as indices is intentional to highlight the fact that a reduced correlation functions are independent of the precise site cluster S such that $S = \text{supp}(\boldsymbol{\alpha})$ for any $\boldsymbol{\alpha} \in \beta$. Thus when used as indices the precise elements of the set β and $\hat{\beta}$ are superfluous. However, we must use them carefully in most other cases, such as sums over their respective elements since they are not the same set $\beta \neq \hat{\beta}$.

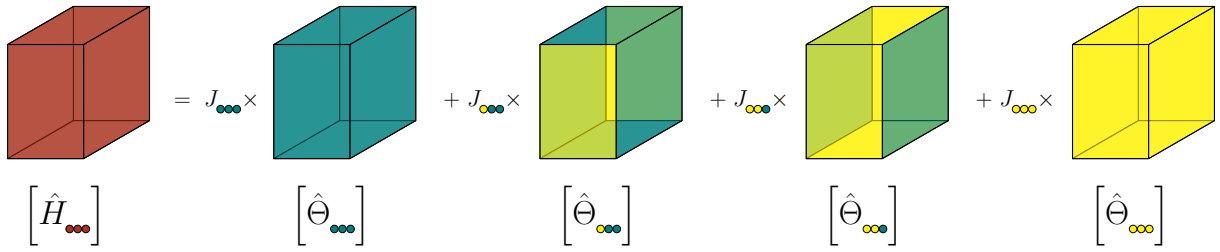


Figure 3.1: Schematic depiction of an arbitrary triplet cluster interaction tensor as a linear combination of reduced correlation tensors. In the depiction we have assumed that permutation multiplicities have been absorbed into the expansion coefficients J.

interaction can be expressed as a sum of reduced correlation functions as follows,

$$\hat{H}_B(\boldsymbol{\sigma}_S) = \sum_{\hat{\beta}\in\hat{L}(S)} \hat{m}_\beta J_\beta \hat{\Theta}_\beta(\boldsymbol{\sigma}_S) \tag{3.8}$$

Where the set $\hat{L}(S) = \{\hat{\beta} : \text{supp}(\boldsymbol{\alpha}) = S\ \forall \boldsymbol{\alpha} \in \beta\}$ *is the set of orbits of symmetrically distinct contracted multi-indices* $\text{ctr}(\boldsymbol{\alpha})$*—i.e. sets of symmetrically distinct permutations of the possible site basis function choices for each site in* S*.*

Correspondingly, *cluster interaction tensors* $\left[\hat{H}_B\right]$, can be directly computed as a linear combination of correlation tensors,

$$\left[\hat{H}_B\right] = \sum_{\hat{\beta}\in L(S)} \hat{m}_\beta J_\beta \left[\hat{\Theta}_\beta\right] \tag{3.9}$$

Figure 3.1 shows a schematic depiction of the cluster interaction tensor for a triplet cluster as a linear combination of the corresponding reduced correlation tensors.

Similarly, the value of a mean cluster interaction H_B for any configuration $\boldsymbol{\sigma}$ can be computed by averaging the values of the cluster interactions for each site cluster $S \in B$,

$$H_B(\boldsymbol{\sigma}) = \frac{1}{|B|} \sum_{S\in B} \left[\hat{H}_B\right]_{\boldsymbol{\sigma}_S} \tag{3.10}$$

Pre-computing all cluster interactions for the cluster decomposition of a given lattice Hamiltonian using Equation 3.10, is particularly useful for applications that require many evaluations of the lattice Hamiltonian, such as Monte Carlo sampling. By computing values or differences in values of a lattice Hamiltonian using cluster interaction tensors directly, the time complexity of the calculation becomes independent of the number of components or species allowed per site as it will only depend on the number of orbits of site clusters S present in the expansion. In other words, evaluating a lattice Hamiltonian for any arbitrary

number of components is just as fast as doing so for a binary system with the same underlying random crystal structure.

Further, the benefits of using cluster interactions are not limited only to expansions that rigorously constitute a cluster decomposition. Tensors of the form given in equation 3.9 can certainly be constructed using non-orthonormal basis representations or the Potts frame representation, which we have called *pseudo cluster interactions*, and will denote with a tilde $\widetilde{H}_B$. Using pseudo cluster interaction tensors in turn also reduces the computation complexity to be independent of the number of allowed components for any non-standard cluster or Potts frame expansion.

Maximal cluster representations

Additional computation run-time improvements can be obtained by re-writing any expansion of a lattice Hamiltonian in an expansion that includes only pseudo cluster interactions that operate over *maximal clusters*. That is an expansion that does not include any term that operates over sub-clusters of clusters associated with another term. This will result in a representation that further reduces the number of terms that have to be evaluated to compute properties or their changes from local configuration changes.

Formally, the procedure above represents another *representation* for a generalized lattice Hamiltonian. We call such a representation a *maximal cluster representation* since it contains functions over a set of maximal clusters only; such that an expansion function that acts over sub-clusters of the clusters that an included function already acts on is never included in a maximal cluster representation.

Definition 3.1.3 (Maximal cluster representation)**.** *A maximal cluster representation is an expansion of a lattice Hamiltonian $H \in L^2(\boldsymbol{\Omega}, \boldsymbol{\rho})$ in terms of pseudo cluster interactions $\widetilde{H}_B$ such that $B \not\sqsubset D$ for any terms $\widetilde{H}_B$ and $\widetilde{H}_D$ included in the expansion.*

In order to appropriately account for terms $\widetilde{H}_D$ that do not act over maximal clusters, we can modify the calculation of mean (pseudo) cluster interactions $\widetilde{H}_D$ based on Equation 3.10 by instead summing over an orbit of maximal clusters B, such that $D \sqsubset B$, as follows,

$$\begin{aligned} \widetilde{H}_D(\boldsymbol{\sigma}) &= \frac{1}{|D|} \sum_{S \in D} \left[\hat{H}_D\right]_{\boldsymbol{\sigma}_S} \\ &= \frac{c_{DB}}{|B|} \sum_{S \in B} \sum_{T \leftarrow S} \left[\hat{H}_D\right]_{\boldsymbol{\sigma}_T} \end{aligned} \tag{3.11}$$

where the notation $T \leftarrow S$, refers to clusters T in orbit D that are sub-clusters of cluster $S \in B$, that is $T \in \{T \subset S \text{ for } T \in D\}$. The *counting factors* c_{DB} are given by,

$$c_{DB} = \frac{m_D}{N_{DB} m_B} \tag{3.12}$$

where $N_{DB} = |\{T \ : \ \forall T \leftarrow S\}|$ are the number of subclusters T contained in cluster S.

Although there exist infinitely many ways of re-writing an expansion in a maximal cluster representation, the procedure to do so is simple to carry out in terms of (pseudo) correlation tensors. In concept, one needs to appropriately broadcast the contributions of lower degree terms into effective pseudo cluster interactions acting over maximal clusters only. This involves identifying the set of maximal cluster orbits $\mathcal{M}$ and for each orbit $B \in \mathcal{M}$ identifying all other orbits D such that $D \sqsubset B$. Subsequently each maximal pseudo interaction $\widetilde{H}_B$ can be computed as follows,

$$\widetilde{H}_B = \frac{1}{|B|} \sum_{S \in B} \sum_{D \sqsubseteq B} \frac{c_{DB}}{n_D} \sum_{T \leftarrow S} \left[\hat{H}_D \right]_{\boldsymbol{\sigma}_T} \tag{3.13}$$

where $n_D = |\{B \in \mathcal{M} \ : \ D \sqsubset B\}|$ is the number of orbits of maximal clusters $B \in \mathcal{M}$ that contain superclusters of the clusters in D. In other words how many maximal clusters will share the contribution of H_D.[3]

Finally, the inner sum in equation 3.13 can be used to define *maximal pseudo cluster interaction* tensors for a practical implementation of a maximal cluster representation,

$$\left[\widetilde{H}_B \right]_{\boldsymbol{\sigma}_S} = \sum_{D \sqsubseteq B} \frac{c_{DB}}{n_D} \sum_{T \leftarrow S} \left[\hat{H}_D \right]_{\boldsymbol{\sigma}_T} \tag{3.14}$$

Results showing the improved evaluation performance when using the tensor-based formulations described earlier are plotted in Figure 3.2. Figure 3.2 shows the run-time for evaluating the total energy for a given configuration in a ternary body-centered cubic system using four different methods: (1) direct evaluation of correlation functions, (2) using pre-computed correlation tensors, (3) using pre-computed cluster interactions, and (4) using pre-computed maximal cluster interactions. The run-time for calculating the change in energy from changing the occupancy of a single site using the four methods is also shown.

Run-time scaling plots are shown for increasing the super-cell size (increasing the number of sites) for an expansion with 1306 total correlation function terms. The results show that the time complexity with respect to system size is the same (linear complexity for full energy evaluations and constant time complexity for energy changes) for all evaluation methods. However, the scaling pre-factor is substantially different to make all the tensor-based and in particular the cluster interaction tensor method more than an order of magnitude faster.

In addition, run-time scaling with model size (the number of terms in the expansion) shows that for both full energy evaluations and energy change evaluations using a cluster interaction-based method noticeably improves the time complexity. This is no surprise since cluster interaction-based methods scale with respect to the number of site space cluster orbits, and not the total number of correlation functions. Furthermore, model evaluation using cluster interactions is of considerable value for systems with a large number of allowed

[3]The equal partitioning of contributions of a non-maximal cluster interaction H_D to each maximal cluster is only one straightforward way of the infinitely many choices to partition such contributions.

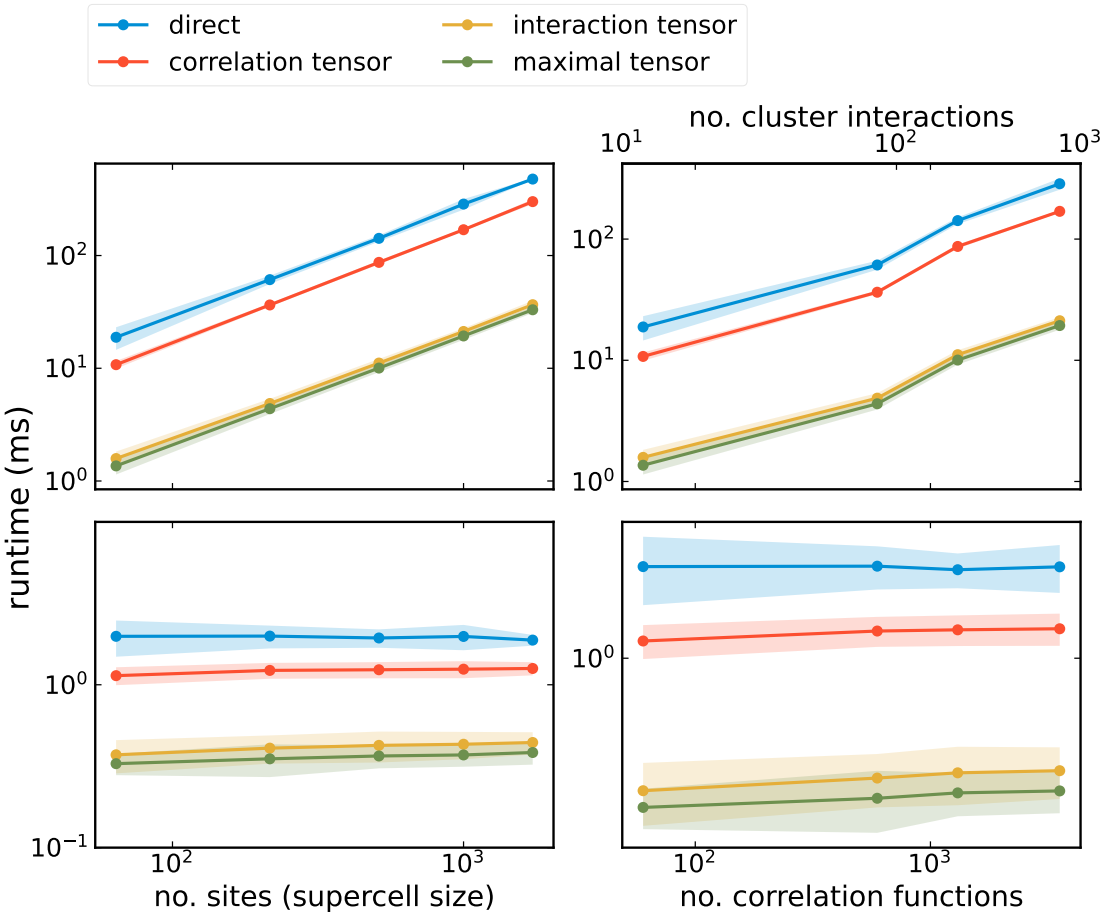


Figure 3.2: Run-time scaling curves for (top) full energy evaluations and (bottom) evaluation of energy differences from a change of the occupancy at a single site. (Left) Run-time scaling with respect to the number of sites (supercell size) for an expansion with 1306 terms (corresponding to 154 cluster interactions). (Right) Run-time scaling curves with respect to the number of correlation functions (model size). In all curves, the values are the average run time of 2000 evaluations. The shaded regions denote one standard deviation. All runs were done using an Intel Core i7-7700HQ 2.80 GHz processor.

species considering that such evaluation only depends on the number of orbits and is thus independent of the number of components.[4]

3.2 Converting expansions to Fourier representations

In Chapter 2 we developed two useful representation schemes for lattice Hamiltonians. Additionally, several related basis function representations—such as those listed in Appendix A.2—have been developed in the cluster expansion literature [188, 230, 258]. The different representations have different properties and as a result different practical advantages and limitations. Based on the different properties of lattice Hamiltonian representations, it becomes of practical importance to be able to convert fitted expansions between different representations. Doing so allows practitioners to leverage the strengths of each representation based on the application at hand.

Notably, when fitting a lattice Hamiltonian a Potts frame or non-orthonormal representation using imperfectly sampled training data, may result in models with improved prediction accuracy than those fitted using a Fourier cluster expansion [9, 258]—such as the example given in Chapter 5.3 using a Potts frame reconstruct Hamiltonians over very high dimensional configuration space. On the other hand, any serious attempt to characterize or interpret interactions between species directly from a fitted expansion should be only carried out using a Fourier cluster expansion and the resulting cluster decomposition. Then, for example, one could fit a lattice Hamiltonian using the Potts frame and subsequently convert it to a Fourier cluster expansion to permit interpretation of cluster interactions. As a solution, we present an efficient method to convert any Hamiltonian representation to a Fourier cluster expansion.

In theory, any transformation between two basis representations can be obtained by the appropriate change of basis matrix. That is, by expressing the old correlation basis functions Θ'_β in terms of the new basis correlation functions Θ_β,

$$\Theta'_\beta = \sum_{\gamma \sqsubseteq \beta} M_{\gamma\beta} \Theta_\beta \tag{3.15}$$

It follows that expansion coefficients for a Hamiltonian can be computed from those expressed in the old basis as follows,

$$J_\gamma = \sum_{\beta \sqsupseteq \gamma} M_{\gamma\beta} J'_\beta \tag{3.16}$$

Specifically, when the new basis is a Fourier correlation basis, the entries of the change of basis matrix are simply proportional to the projections of the old basis functions onto the Fourier correlation basis functions,

$$M_{\gamma\beta} = m_\gamma \langle \Theta'_\beta, \Theta_\gamma \rangle_{\boldsymbol{\rho}} \tag{3.17}$$

[4] In other words this allows evaluation of the expansion of any multi-component material at the same time-complexity of a binary system.

However, in practice it is much more efficient, to use a pseudo cluster decomposition to represent the Hamiltonian in the old basis, and obtain the Fourier coefficients by carrying out projections of the pseudo cluster interactions $\widetilde{H}_B$ onto the Fourier correlation basis. Thus, the Fourier coefficients are computed as follows,[5]

$$J_\gamma = N \sum_{B \supseteq D(\gamma)} m_B \langle \widetilde{H}_B, \Theta_\gamma \rangle_{\boldsymbol{\rho}} \tag{3.18}$$

where the notation $D(\gamma) = \mathrm{Orb}_{\mathcal{G}}(\mathrm{supp}(\gamma))$ represents the site space clusters D over which the function Θ_γ acts on.

In practice, one can compute the inner products in equation 3.18 by computing only the inner product between pseudo interaction tensors and Fourier correlation tensors. The correlation tensors must be suitably expanded to match the dimensions of the interaction tensor being projected onto; similar to what was done to broadcast lower interactions when constructing the maximal cluster representation. Since the correlation functions involved in Equation 3.18 always act over sub-clusters of the clusters that $\tilde{H}_B$ acts on, the *broadcasted* reduced correlation functions $\hat{\Theta}_{\gamma B}(\boldsymbol{\sigma}_S)$ are represented as follows,

$$\hat{\Theta}_{\gamma B}(\boldsymbol{\sigma}_S) = \sum_{T \leftarrow S} \hat{\Theta}_\gamma(\boldsymbol{\sigma}_T) \quad \text{for any } S \in B \tag{3.19}$$

The inner product can be efficiently computed in terms of the (pseudo) cluster interaction tensor $\left[\hat{H}_B\right]$ and the broadcasted reduced correlation tensor $\left[\hat{\Theta}_{\gamma B}\right]$, such that the new expansion coefficients—those appearing in Equation 3.18)—can be calculated as follows,[6]

$$J_\gamma = \sum_{B \supseteq D(\gamma)} \frac{m_B}{m_{D(\gamma)}|S|} \left[\hat{H}_B\right] \cdot \left[\hat{\Theta}_{\gamma B}\right] \tag{3.20}$$

Using any site space cluster $S \in B$, and where the tensor dot product above is the sum of all element wise-products.

Using pseudo cluster interactions allows converting any basis expansion as well as Potts frame expansion coefficients to Fourier cluster expansions with the same procedure outlined above. As an illustration, Figure 3.3 shows the expansion coefficients for the Hamiltonian of a ternary alloy fitted with a using a Potts frame and the expansion coefficients for the same Hamiltonian converted to a Fourier correlation basis representation.

An important observation to make is that the corresponding Fourier coefficients will always have more weight on lower degree terms since all cluster indicator correlations have nonzero projections onto lower degree Fourier correlation functions. In practice, fits using a non-orthogonal basis will usually result in coefficients of larger magnitude and possibly

[5] We are essentially expanding each $\widetilde{H}_B$ in its own Fourier cluster expansion and summing the coefficients from each $\widetilde{H}_B$ corresponding to the same correlation function.

[6] A derivation is given in Appendix B.4.

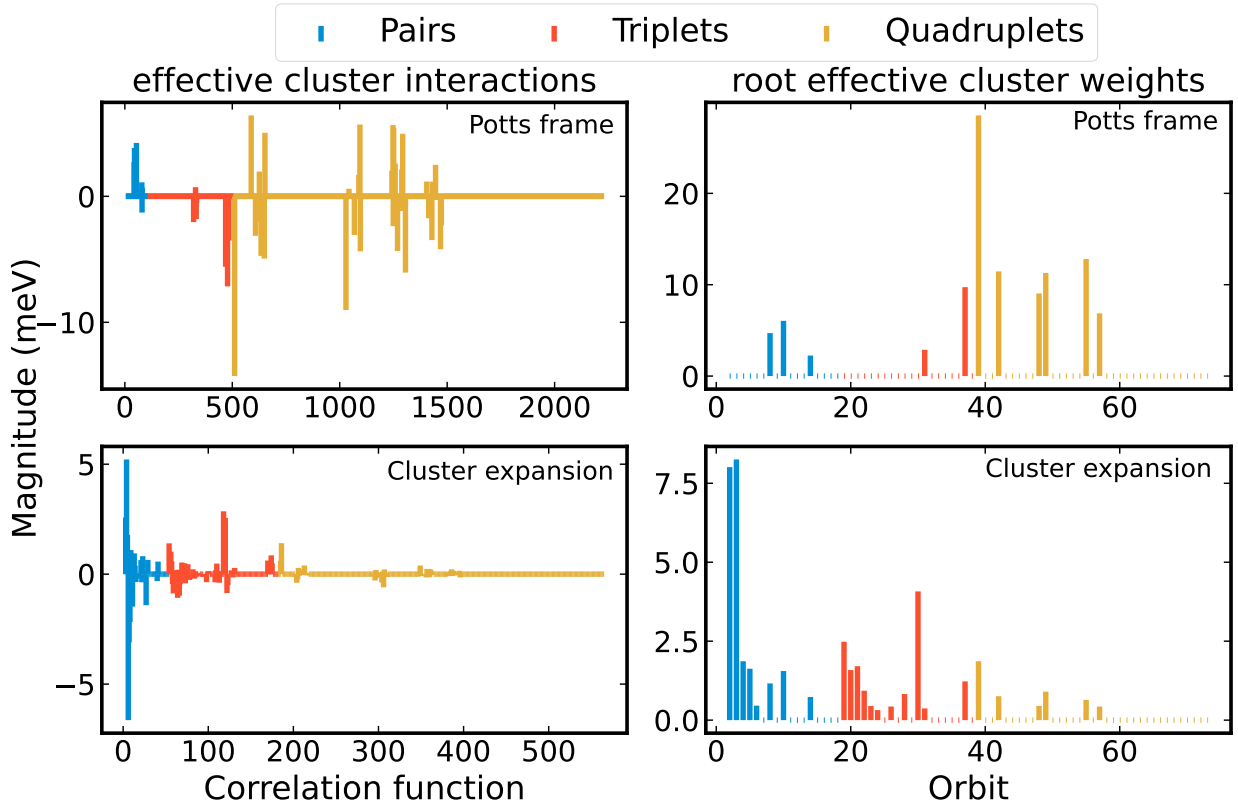


Figure 3.3: Effective cluster interactions and root cluster weights for a Potts frame fit of a CrCoNi alloy system, and the corresponding expansion coefficients converted to a Fourier cluster expansion according to the procedure outlined using Equation 3.18

exhibit a weaker trend in the decay of coefficient magnitude with decay size as can be observed in Figure 3.3. A handful of additional examples of this behavior are given in Chapter 5.2. However, practitioners should understand that the corresponding Fourier cluster expansion coefficients will most often reflect a much stronger decay versus the degree of expansion coefficients. Based on this, practitioners should be wary of using arguments that rely on heuristics that have been established based on Fourier cluster expansions when analyzing trends in coefficients of a non-orthogonal cluster expansion.

Finally, we contend that there never really is a need to transform coefficients in the reverse direction, i.e. from a Fourier cluster expansion to a non-Fourier representation. If a sufficiently accurate fit has been obtained as a Fourier cluster expansion then, to the best of our knowledge, there are few theoretical or practical benefits obtained by converting a Fourier expansion to another representation. For the few cases where Boolean operations may be important, such as determining ground-state configurations [100], any expansion expressed as a (pseudo) cluster decomposition (including the maximal cluster representation), can

trivially be expressed as a Potts frame simply by setting each term $J_\beta I_\beta(\boldsymbol{\sigma})$ as follows

$$J_\beta I_\beta(\boldsymbol{\sigma}) = \frac{1}{m_\beta N} \sum_{\boldsymbol{\alpha} \in \beta} \hat{H}_{B(\beta)}(\boldsymbol{\sigma}_{\boldsymbol{\alpha}}) \mathbf{1}_{\boldsymbol{\alpha}}(\boldsymbol{\sigma}) \tag{3.21}$$

where $\boldsymbol{\sigma}_{\boldsymbol{\alpha}}$ refers to the precise configuration the cluster indicator function $\mathbf{1}_{\boldsymbol{\alpha}}$ indicates for. That is the expansion coefficients J_β are just the values of the corresponding reduced correlation function $\hat{H}_{B(\beta)}$ evaluated at the configuration being indicated for.

3.3 Computing cluster occupation probabilities and averages

In addition to predicting formation energies, which can, in turn, be used in calculations of Free energies and thermodynamic properties of atomic configuration, the ability to resolve atomic ordering statistics at finite temperatures permits a much deeper understanding of the properties and behavior of materials exhibiting partial or full atomic disordering [36, 81, 159, 204]. As detailed in Chapter 1 short-range order parameters at finite temperatures have an important role in the study of disordered materials [35, 50, 54, 81].

Short range order (SRO) parameters are computed in terms of *cluster probabilities* [36, 50, 54]. More explicitly, SRO parameters are computed using the marginal probabilities that the site space clusters for a particular orbit B have a specific atomic configuration.

Definition 3.3.1 (Cluster probability)**.** *A cluster probability $\mathbb{P}_S(\boldsymbol{\sigma}_S)$ is the marginal of the Boltzmann probability which gives the probability for the configuration $\boldsymbol{\sigma}_S$ of a given cluster of sites S,*

$$\mathbb{P}_S(\boldsymbol{\sigma}_S \mid T) = \sum_{\boldsymbol{\sigma} \setminus \boldsymbol{\sigma}_S} \mathbb{P}(\boldsymbol{\sigma} \mid T) \tag{3.22}$$

where the sum is carried over all possible configurations of the sites not in S; and $\mathbb{P}(\boldsymbol{\sigma} \mid T)$ is a generalized Boltzmann distribution as introduced in Chapter 1.4.

Many methods exist for calculating approximations to cluster probabilities [67, 114, 144, 192, 216]. A subset of these methods leverages the particular representation of the lattice Hamiltonian [67, 192, 193] in terms of a Fourier cluster expansion.[7] The main idea underlying these methods is to represent the cluster probabilities themselves using a Fourier cluster expansion. The notion is straightforward since cluster probabilities at a given temperature $\mathbb{P}_S(\boldsymbol{\sigma}_S \mid T) \in L^2(\boldsymbol{\Omega}_S)$ are themselves functions over configuration space.[8] Moreover, since

[7] As a matter of fact, approximating cluster probabilities was the motivation to formalize the representation of lattice Hamiltonians using Fourier cluster expansions [192].

[8] And usually a very manageable configuration space since usually only small clusters S are important!

the cluster probabilities for any cluster in the same orbit $S \in B$ are equivalent by symmetry, the cluster probability $\mathbb{P}_B$ for all clusters $S \in B$ can be expanded as follows,

$$\mathbb{P}_B(\boldsymbol{\sigma}_{S\in B} \mid T) = \sum_{\beta\in\mathcal{G}(\mathbb{N}^N_{<\boldsymbol{n}})} m_\beta \xi_\beta(T)\Theta_\beta(\boldsymbol{\sigma}) \tag{3.23}$$

where the expansion coefficients $\xi_\beta(T)$ depend on the thermodynamic temperature T. When Equation 3.23 is expressed in terms of a Fourier cluster expansion, then the expansion coefficients can be shown to correspond to the expected values of each correlation function as follows [193],

$$\xi_\beta(T) = \langle \mathbb{P}_B(\boldsymbol{\sigma}_{S\in B} \mid T), \Theta_\beta\rangle_{\boldsymbol{\rho}} = \langle \Theta_\beta\rangle_T \tag{3.24}$$

Therefore, cluster probabilities can be directly computed from the thermodynamic expectation values of Fourier correlation functions. For many practical calculations—in particular, those based on Monte Carlo sampling—the above approach can be generalized by expressing cluster probabilities as the sum of expectations over the indicator functions for each possible configuration of a site space cluster S,

$$\mathbb{P}_S(\boldsymbol{\sigma}_S \mid T) = \sum_{\boldsymbol{\sigma}'_S\in\boldsymbol{\Omega}_S} \langle \mathbf{1}_{\boldsymbol{\sigma}'_S}\rangle_T \mathbf{1}_{\boldsymbol{\sigma}'_S}(\boldsymbol{\sigma}_S) \tag{3.25}$$

And the probability for all clusters in an orbit $S \in B$ can be obtained as the mean of all probabilities for each cluster,

$$\begin{aligned}
\mathbb{P}_B(\boldsymbol{\sigma}_{S\in B} \mid T) &= \frac{1}{|B|}\sum_{S\in B} \mathbb{P}_S(\boldsymbol{\sigma}_S \mid T) \\
&= \frac{1}{|B|}\sum_{S\in B}\sum_{\boldsymbol{\sigma}'_S\in\boldsymbol{\Omega}_S} \langle \mathbf{1}_{\boldsymbol{\sigma}'_S}\rangle_T \mathbf{1}_{\boldsymbol{\sigma}'_S}(\boldsymbol{\sigma}_S) \\
&= \frac{1}{|B|}\sum_{S\in B}\sum_{\hat{\beta}\in\mathcal{G}(\boldsymbol{\Omega}_S)} \langle \hat{m}_\beta I_\beta\rangle_T \hat{m}_\beta \hat{I}_\beta(\boldsymbol{\sigma}_S)
\end{aligned} \tag{3.26}$$

where $\mathcal{G}(\boldsymbol{\Omega}_S)$ is the set of sets of symmetrically equivalent occupancy $\boldsymbol{\sigma}_S$ of a cluster $S \in B$. We explicitly include the multiplicities $\hat{m}_\beta$, so that reduced cluster indicator correlations give values of 1—opposed to $1/\hat{m}_\beta$—for any of the symmetrically equivalent occupancies $\boldsymbol{\sigma}_S$ included in β (i.e. they are symmetrized indicator functions).

Equation 3.26 is simply a re-expression of the canonical expansion of $\mathbb{P}_B$ in a symmetrized Potts frame written as follows,[9]

$$\mathbb{P}_B(\boldsymbol{\sigma}_{S\in B} \mid T) = \sum_{\beta\in L(B)} \langle \hat{m}_\beta I_\beta\rangle_T m_\beta I_\beta(\boldsymbol{\sigma}) \tag{3.27}$$

[9]From a more careful inspection, we can also see this as the mean pseudo cluster interaction of cluster probabilities!

where the set $L(B) = \{\beta : \text{supp}(\boldsymbol{\alpha}) \in B \; \forall \boldsymbol{\alpha} \in \beta\}$ is constructed in the same way as done for the cluster decomposition given in Equation 2.33. However, the multi-indices $\boldsymbol{\alpha}$ are obtained from the *overcomplete* set $\boldsymbol{\alpha} \in \mathbb{N}^N_{\leq \boldsymbol{n}}$ used in the Potts frame construction (in contrast to the multi-index set $\boldsymbol{\alpha} \in \mathbb{N}^N_{<\boldsymbol{n}}$ used in Fourier cluster expansions.

For practical purposes, the cluster indicator correlation functions in Equation 3.25 can be suitably expressed in terms of reduced cluster correlation functions as follows,

$$\hat{m}_\beta \hat{I}_\beta(\boldsymbol{\sigma}_S) = \sum_{\gamma \in \mathcal{G}(\mathbb{N}^{|S|}_{<\boldsymbol{n}})} I_{\gamma\beta} \hat{\Theta}_{\gamma B}(\boldsymbol{\sigma}_S) \tag{3.28}$$

where $I_{\gamma\beta}$ are expansion coefficients. Note that the left hand side is a Fourier cluster expansion, hence the functions are indexed using multi-indices from the set $\mathcal{G}(\mathbb{N}^{|S|}_{<\boldsymbol{n}})$; and $\hat{\Theta}_{\gamma B}$ are reduced correlation functions broadcasted to site space clusters $S \in B(\beta)$ following Equation 3.19.

The set of equations as given in Equation 3.26 for all symmetrically distinct configurations represented by $\hat{\beta} \in \mathcal{G}(\boldsymbol{\Omega}_S)$ can be compactly expressed as the following matrix equation,[10]

$$\mathbb{I}_{SS} = \mathbb{V}_S \boldsymbol{I}_{\gamma\beta} \tag{3.29}$$

where the columns of the matrix $\mathbb{V}_S \in \mathbb{R}^{|\mathcal{G}(\boldsymbol{\Omega}_S)| \times |\mathcal{G}(\mathbb{N}^{|S|}_{<\boldsymbol{n}})|}$ are the corresponding reduced correlation functions expressed as vectors in $\mathbb{R}^{|\boldsymbol{\Omega}_S|}$.[11] The matrix $\boldsymbol{I}_{\gamma\beta} \in \mathbb{R}^{|\mathcal{G}(\mathbb{N}^{|S|}_{<\boldsymbol{n}})| \times |\mathcal{G}(\boldsymbol{\Omega}_S)|}$ is made up of columns with the expansion coefficients $I_{\gamma\beta}$ for each reduced indicator correlation function $\hat{I}_\beta(\boldsymbol{\sigma}_S)$. And $\mathbb{I}_{SS}$ is a $|\mathcal{G}(\boldsymbol{\Omega}_S)| \times |\mathcal{G}(\boldsymbol{\Omega}_S)|$ identity matrix.

Before continuing, we make an important remark regarding Equation 3.29. Even though we are considering symmetrically distinct configurations, this may not necessarily result in a square $\mathbb{V}$ matrix for any given cluster S. For example any cluster S of symmetrically equivalent sites under the space group $\mathcal{G}$ of a disordered crystal structure that are not symmetrically equivalent with respect to its point group $\mathcal{P}$ will result in $|\mathcal{G}(\mathbb{N}^{|S|}_{<\boldsymbol{n}})| < |\mathcal{G}(\boldsymbol{\Omega}_S)|$. Meaning that the set of reduced correlation functions constructed with the symmetry $\mathcal{G}$ of the underlying lattice spans a lower dimensional space than the set of indicator functions for all symmetrically distinct configurations. Nevertheless, this does not prevent the use of Equation 3.29 to obtain a practical expression for the probability $\mathbb{P}_B$ of clusters in an orbit using Equation 3.26. Indeed, this is just a statement that the probability of the different configurations of clusters of an orbit in Equation 3.26 must be symmetrically invariant under operations of the symmetry group $\mathcal{G}$.

Furthermore, the fact that $\mathbb{V}_S$ may not be square, does not prevent the inversion of Equation 3.29, since it will always be either square or overdetermined, and will always be

[10] This formulation is equivalent to the original V-matrix construction used in the solution of the CVM by way of Fourier cluster expansions [33, 192, 193].

[11] In this case correlation functions over orbits $\gamma \sqsubset \beta$ must be accordingly expanded to the dimension $|\mathcal{G}(\boldsymbol{\Omega}_S)|$ by an appropriate ordering of the corresponding site clusters.

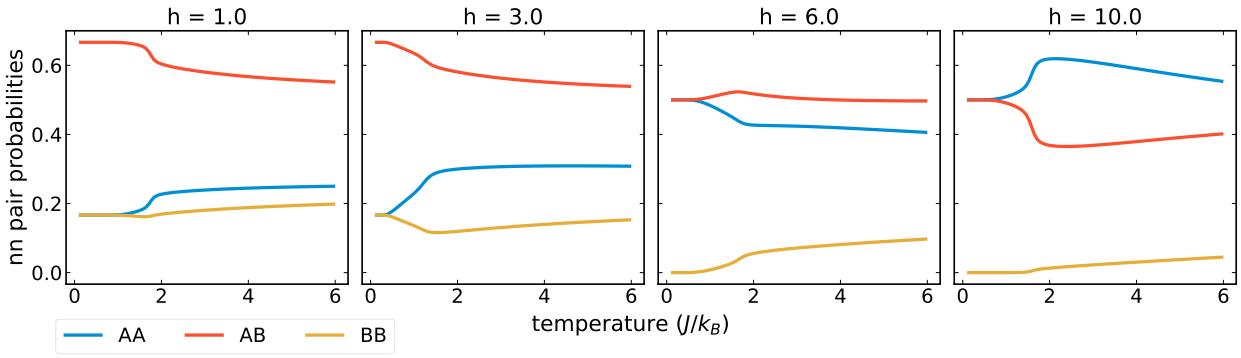


Figure 3.4: Nearest neighbor pair probabilities for a face-centered cubic antiferromagnetic Ising model under different applied fields h calculated with Wang-Landau sampling using a supercell with 256 sites.

full rank.[12] Thus, the expansion coefficients for the reduced correlation indicator functions are given by the right (pseudo)-inverse of $\mathbb{V}_S$,

$$\boldsymbol{I}_{\gamma\beta} = (\mathbb{V}_S^\top \mathbb{V}_S)^{-1} \mathbb{V}_S^\top = \mathbb{V}_S^+ \tag{3.30}$$

Consequently, the pseudo-inverse can be used to compute the occupation probabilities of clusters in an orbit $\mathbb{P}_B$ in terms of correlation functions using the following matrix-vector equation,[13]

$$\mathbb{P}_B = M_{DB} \mathbb{V}_S^+ \langle \boldsymbol{\Pi}_B \rangle_T \tag{3.31}$$

where the elements of probability vector $\mathbb{P}_B$ are the thermodynamic probability of symmetrically distinct configuration $\boldsymbol{\sigma}_S$ over all symmetrically equivalent clusters $S \in B$. The elements of the vector $\langle \boldsymbol{\Pi}_B \rangle_T \in \mathbb{R}^{|\mathcal{G}(\mathbb{N}^{|S|}_{<\boldsymbol{n}})|}$ are given by the expectation value of each correlation function $\beta \in \mathcal{G}(\mathbb{N}^{|S|}_{<\boldsymbol{n}})$, i.e. $(\langle \boldsymbol{\Pi}_B \rangle_T)_\beta = \langle \Theta_\beta \rangle_T$. The matrix $M_{DB} \in \mathbb{R}^{|\mathcal{G}(\mathbb{N}^{|S|}_{<\boldsymbol{n}})| \times |\mathcal{G}(\mathbb{N}^{|S|}_{<\boldsymbol{n}})|}$ accounts for the *over-counting* factors between the broadcasted correlation functions used to compute $\mathbb{V}$ based on Equations 3.28 and 3.29 and the full correlation functions in $\langle \boldsymbol{\Pi}_B \rangle_T$.

Equation 3.31 is quite useful for Monte Carlo sampling because it allows the thermodynamic averages of correlation functions to be easily approximated without any additional computing—computing the energy of each configuration already requires computing the correlation function values for that configuration [216]. Figure 3.4 shows an example of the nearest neighbor pair cluster probabilities as a function of temperature for an FCC antiferromagnetic Ising models calculated from Wang-Landau [237] Monte Carlo sampling using

[12] In fact, this was one of the main motivations to use such construction to find solutions to the cluster variation method (CVM). Apart from guaranteeing that cluster probabilities satisfy marginality constraints, this prescription guarantees that the probability functions are symmetrically invariant under the symmetry of the disordered structure, thereby transforming the CVM problem into an unconstrained one over the expansion coefficients $\xi_\beta(T)$ [192, 193].

[13] A derivation is given in Appendix B.5.

Equation 3.31 for several values of non-critical fields. An example illustrating the computation of short-range ordering for a real material using this method is given in Chapter 5.4.

Following the same reasoning, the *cluster occupation averages* for a specific configuration $\boldsymbol{\sigma}$ can be suitably computed as,

$$\bar{S}_B(\boldsymbol{\sigma}) = M_{DB}\mathbb{V}_S^{+}\boldsymbol{\Pi}_B(\boldsymbol{\sigma}) \tag{3.32}$$

Where now $\bar{S}_B(\boldsymbol{\sigma})$ is made up of cluster occupation averages for the configuration $\boldsymbol{\sigma}$, and the elements of the vector $\boldsymbol{\Pi}_B(\boldsymbol{\sigma})$ are now simply the correlation functions evaluated for the configuration $\boldsymbol{\sigma}$, i.e. $(\boldsymbol{\Pi}_B(\boldsymbol{\sigma}))_\beta = \Theta_\beta(\boldsymbol{\sigma})$.

Chapter 4

Learning applied lattice models

We have motivated their use and presented the formal representation and interpretation of generalized lattice models for the study of thermodynamic properties of atomic configurations in crystalline materials. In this chapter, we discuss several aspects that enable fitting a generalized lattice model to a particular material system using training data computed from *first-principles calculations*. We focus particularly on using training data calculated with *density functional theory* (DFT) [120], since this is by far the most prevalent method used in computational materials science research [23, 89]. There have been a handful of different learning algorithms developed to parametrize applied lattice models since the cluster expansion method was proposed. After only a brief description of some of these methods, we will focus solely on regularized linear regression-based learning and in particular regularization that results in *structured sparsity*. The majority methodology developed in this chapter is based on our recently published work [9, 10, 253, 259].

4.1 Overview of learning algorithms

Learning or *fitting* an applied lattice model is the process by which to estimate expansion coefficients for a lattice Hamiltonian that can most accurately predict a particular materials property. Expansion coefficients are predominantly obtained by statistical estimation, such that the predicted energy matches the energy calculated for a set of configurations by first-principles calculations. However, we can also take a *parametric* approach [55], where different sets of expansion coefficients for a prescribed set of correlation functions are used to calculate thermodynamic properties. Particular Hamiltonians that match observed thermodynamic properties for the application at hand can subsequently be used to further probe relevant phenomena. Additionally, some form of bench-marking with observed thermodynamic properties should always be carried out when using calculations from an applied lattice model obtained with either a *fitting* or a *parametric* approach to ensure a faithful representation of the Hamiltonian has been obtained. Furthermore, there are a variety of ways both approaches can be used together. Nonetheless, in this chapter, we will only focus on the

learning approach since it is the approach used predominantly in practice—though the former still comprises a worthwhile path to develop a powerful methodology for the study of disordered materials.

A large number of different algorithms and methods for learning cluster expansions have been proposed and bench-marked in literature. It would be overly ambitious to properly cover them all in a single chapter. Nevertheless, we provide a high-level overview, by grouping algorithms into general three categories based on the underlying conceptual approach. These learning categories can be roughly summarized as follows [55],

1. **Mean medium perturbation methods**, such as the Generalized Perturbation method [61], the concentration wave method [85] and the embedded cluster method [79], were the first set of methods proposed. Broadly, in these methods, local concentration perturbations are introduced into an effective medium representing the non-interacting (or fully random) electronic energy in a periodic potential (on a lattice). The expansion coefficients are calculated directly from a perturbation treatment of an effective medium theory, such as the coherent potential approximation [80]. Although these methods can be regarded as the most formal, in the sense of their direct inclusion of the electronic structure of a disordered solid in parameter estimation, they have mostly fallen out of favor in practical calculations, in particular, because they lack structural relaxations; so we will not discuss them further.

2. **Direct configuration averaging** involves directly estimating expansion coefficients as empirical averages using the statistical interpretation given in Equation 2.14 [58] in Chapter 2. This method has also fallen out of favor due to the large number of training structures required to obtain accurate estimates of coefficients. However, we will briefly discuss it since it may possibly have renewed interest in light of improvements in computation power and recently developed methods to calculate training data. We give some brief suggestions on how to revisit this method as an outlook in our conclusions in Chapter 6.

3. **Structure inversion methods** are those that involve linear regression estimation of expansion parameters [49]. These methods have become the de-facto method for fitting Hamiltonians of configuration when constructing an applied lattice model. As such, we will discuss them extensively, and focus on novel techniques and methodology that we have found greatly improves the accuracy and robustness of the resulting Hamiltonians.

Direct configuration averaging

The method of direct configuration averaging (DCA) to fit a lattice Hamiltonian of configuration involves directly approximating the expansion parameters using the inner-product given in Equation 2.14 [55, 58] under its statistical interpretation as an expectation. The

concept is to simply approximate the *expected* value as an arithmetic mean,

$$\begin{aligned} J_\beta &= \mathbb{E}_\rho \left[H(\boldsymbol{\sigma}) \Theta_\beta(\boldsymbol{\sigma}) \right] \\ &= \sum_{\boldsymbol{\sigma} \in \boldsymbol{\Omega}} \boldsymbol{\rho}(\boldsymbol{\sigma}) H(\boldsymbol{\sigma}) \Theta_\beta(\boldsymbol{\sigma}) \\ &= \sum_{\boldsymbol{\sigma} \in \boldsymbol{\Omega}} \boldsymbol{\rho}(\boldsymbol{\sigma}) H(\boldsymbol{\sigma}) \Phi_{\boldsymbol{\alpha}}(\boldsymbol{\sigma}) \quad \text{for} \quad \boldsymbol{\alpha} \in \beta \qquad (4.1) \\ &\approx \frac{1}{|\boldsymbol{\Omega}_T|} \sum_{\boldsymbol{\sigma} \in \boldsymbol{\Omega}_T} E_{\boldsymbol{\sigma}} \Phi_{\boldsymbol{\alpha}}(\boldsymbol{\sigma}) \quad \text{for} \quad \boldsymbol{\alpha} \in \beta \qquad (4.2) \end{aligned}$$

where we have used the fact that all projections of $H(\boldsymbol{\sigma})$ onto any function in the orbit of product functions $\{\Phi_\alpha\}_\beta$ are equal by symmetry. The approximation is made by using only a set of configurations $\boldsymbol{\Omega}_T$ corresponding to a set of training structures for which the corresponding energies $E_{\boldsymbol{\sigma}}$ have been computed.

The DCA method was originally developed and proposed only estimation under a uniform apriori distribution $\rho(\boldsymbol{\sigma}) = 1/|\boldsymbol{\Omega}|$, and not for any general product distribution as stated in Equation 4.1.[1] Yet generally, in order for the estimate given in Equation 4.2 to be accurate, the set $\boldsymbol{\Omega}_T$ must be sampled according to the apriori distribution being targeted.[2]

DCA offers a mathematically rigorous method to estimate expansion coefficients. Further, estimating a coefficient using DCA is done completely independently of all other coefficients and expansion terms. However, as already mentioned, using DCA fell out of practice due to the large number of structures required to obtain accurate expansions compared to the *structure inversion method* [55, 246]. However, in light of the momentous advancements in computing and methods to approximate the energy of crystalline materials, revisiting the DCA may prove a worthwhile endeavor. We provide a few suggestions and remarks on doing so in the outlook given in Chapter 6.

The structure inversion method

The structure inversion method (SIM) as originally proposed, involved solving for expansion coefficients of a binary Fourier cluster expansion by inverting a set of linear equations in the expansion coefficients [49]. In its original development, the SIM involved constructing a system with the number of independent equations equal to the number of unknown coefficients. This constituted an invertible square linear system, expressed as follows,

$$\boldsymbol{\Pi} \boldsymbol{J} = \boldsymbol{E} \qquad (4.3)$$

where $\boldsymbol{\Pi} \in \mathbb{R}^{m \times m}$ is a square matrix, where each row $\boldsymbol{\Pi}_{\boldsymbol{\sigma}} \in \mathbb{R}^m$ is a vector—referred to as an *correlation vector*—made up of the values of a predefined set of Fourier correlation

[1] Actually, the DCA was used to some extent for other distributions by considering configurations $\boldsymbol{\sigma}$ at a fixed concentration.

[2] The estimates can always be improved by including weights proportional to $\boldsymbol{\rho}(\boldsymbol{\sigma})$.

basis functions evaluated for a given configuration $\boldsymbol{\sigma}$. $\boldsymbol{J} \in \mathbb{R}^m$ is a vector of the unknown expansion coefficients, and $\boldsymbol{E} \in \mathbb{R}^m$ is a vector of the energy $E_{\boldsymbol{\sigma}}$ of each configuration $\boldsymbol{\sigma}$ computed by some appropriate method, e.g. DFT or a related first-principles electronic structure method.

The estimated expansion coefficients can be determined by solving the linear system, i.e. inverting the matrix $\boldsymbol{\Pi}$,

$$\boldsymbol{J} = \boldsymbol{\Pi}^{-1}\boldsymbol{E} \tag{4.4}$$

The structure inversion method quickly became popular for its simplicity and its accuracy in predicting energies of binary alloys that often surpassed the accuracy of other learning methods [246]. However, the original structure inversion is limited in two important aspects. First, the expansion coefficients depend on the selected structures used to construct the square linear system. Second, a pre-selected set of expansion functions needs to be selected beforehand.

The first issue was promptly addressed by using Ordinary Least Squares (OLS) regression to estimate the coefficients instead of the direct inversion in Equation 4.4 [55]. Using OLS regression, the process to obtain expansion coefficients becomes a statistical estimation problem rather than an exact solution. The OLS regression optimization problem for estimated expansion coefficients is written as follows,

$$\boldsymbol{J}^* = \underset{\boldsymbol{J}}{\operatorname{argmin}} \, ||\boldsymbol{\Pi}\boldsymbol{J} - \boldsymbol{E}||_2^2 \tag{4.5}$$

where $\boldsymbol{\Pi} \in \mathbb{R}^{m \times d}$ is similarly a *correlation matrix* (also referred to as feature matrix) where the rows are truncated correlation vectors $\boldsymbol{\Pi}_{\boldsymbol{\sigma}} \in \mathbb{R}^d$ for m training structures. Notice now that the number of training structures m does not need to equal the number of expansion terms p. $\boldsymbol{J}, \boldsymbol{J}^* \in \mathbb{R}^d$ are vectors of the expansion coefficients. In the vector of expansion coefficients, $\boldsymbol{J}, \boldsymbol{J}^* \in \mathbb{R}^d$, the multiplicities for of the corresponding correlation function are usually treated implicitly as $\boldsymbol{J} = (J_0, m_{\beta_1} J_{\beta_1}, \ldots, m_{\beta_{d-1}} J_{\beta_{d-1}})$. However, the expansion coefficients can be fitted directly by accounting for the multiplicities in the feature matrix as well.

The OLS estimation of expansion coefficients is still often referred to as the SIM. However compared to the direct inversion original proposed, OLS allows estimation of coefficients for both *overdetermined* cases—when there are more training structures than correlation functions ($m > d$)—and for *underdetermined* scenarios—where there are more correlation functions than training structures ($m < d$). To some extent, OLS also addresses the dependence of expansion parameters on a particular set of training structures. However, it still requires using a predefined set of expansion terms. Further, it suffers from all of the complexities and limitations that can arise in OLS estimation, such as solutions that are highly sensitive to changes in the training data, and a tendency to over-fit to training data [66, 90, 93].

These remaining obstacles have been addressed by resorting to *regularized linear regression* and/or including linear constraints in the regression optimization problem. A variety of different regression models have been proposed and tested in the literature [2, 131, 148, 153,

227]. In the rest of this chapter, we will focus on the basics of regularized regression. We will then discuss how physically and mathematically motivated priors on the structure of the expansion coefficients in lattice Hamiltonians can be introduced. Finally, we give examples of how these methods can be used to fit accurate and robust Hamiltonians of configuration for complex multi-component crystalline materials with an eye toward the growing number of components in materials of technological interest.

4.2 Regularized linear regression

Regularized linear regression seeks to estimate coefficients by including a penalization or regularization term to the OLS regression objective. A general regularized linear regression estimation can be expressed as follows,

$$\boldsymbol{J}^* = \underset{\boldsymbol{J}}{\operatorname{argmin}} \; ||\boldsymbol{\Pi}\boldsymbol{J} - \boldsymbol{E}||_2^2 + ||\boldsymbol{J}||$$

where $||\boldsymbol{J}||$ is the regularization term, which usually involves the power of a norm or combination of norms of the coefficients $\boldsymbol{J}$. The coefficient for the data offset, or formally the empty cluster coefficient, is commonly not penalized in the regularization [90, 148]. Additional constraints can be added to the optimization problem in Equation 4.6, such as cluster hierarchy constraints [99, 131] or constraints to preserve certain configurations as ground states [99].

In current research into MPE or complex ionic materials, some form of regularized regression is almost always used and necessary in practice when estimating expansion coefficients. Furthermore, regularized regression models can be derived and/or interpreted under a Bayesian framework, such that the choice of regularization function is based on the assumed prior distribution of coefficients [90, 148]. We will not discuss the details of a Bayesian motivation or interpretation, but make note that such an interpretation in terms of a prior distribution—that corresponds to the specific regularization used—always exists [90, 93]. Instead, we focus on the practical and numerical benefits resulting from the use of regularization.

Effects of regularization

In order to better understand the effects of including a regularization term in linear regression, it is important to first understand the basic properties and geometry of solutions in OLS problems. The matrix-vector multiplication between the correlation matrix and the expansion coefficient to be estimated can be understood as the expansion of a finite vector in terms of the columns/features made up of the sampled values of the correlation functions included,

$$\boldsymbol{\Pi}\boldsymbol{J} = \sum_{i=1}^{d} \boldsymbol{J}_i \boldsymbol{\Pi}^{(i)} \tag{4.6}$$

where $\mathbf{\Pi}^{(i)}$ is a column or *feature vector* of $\mathbf{\Pi} \in \mathbb{R}^{n \times d}$ made up of the values of a corresponding correlation function $\Theta_{\beta^{(i)}}$ evaluated for each of the configurations $\boldsymbol{\sigma}^{(i)}$ for $i = 1, \ldots, n$.

Equation 4.6 makes it explicit, that the approximation done when minimizing the OLS objective (without regularization) in Equation 4.5 amounts to finding a vector that is closest to $\boldsymbol{E}$ (in terms of the Euclidean distance in $\mathbb{R}^n$) as a linear combination of feature vectors $\mathbf{\Pi}^{(i)}$, i.e. in the range of the correlation matrix $\mathbf{\Pi}$, $\hat{\boldsymbol{E}} = \mathbf{\Pi}\boldsymbol{J}^* \in \mathcal{R}(\mathbf{\Pi})$ [25, 180]. Further, the dimension of the range of the correlation matrix $\dim \mathcal{R}(\mathbf{\Pi})$, is equal to the number of *linearly independent* features, and is bounded by $\dim \mathcal{R}(\mathbf{\Pi}) \leq \min(m, d)$. Whenever $\dim \mathcal{R}(\mathbf{\Pi}) = \min(m, d)$, we say that $\mathbf{\Pi}$ is *full rank*.

In addition to the fact that any $\mathbf{\Pi}\boldsymbol{J} \in \mathbb{R}(\mathbf{\Pi})$, any set of coefficients J^* that are a solution to Equation 4.5, satisfy the well-known *normal equations* [25, 66, 90, 180],

$$\mathbf{\Pi}^\top \mathbf{\Pi} \boldsymbol{J}^* = \mathbf{\Pi}^\top \boldsymbol{E} \tag{4.7}$$

Whenever $\mathbf{\Pi}$ is full rank, a solution to the normal equations 4.7 can be obtained in terms of the *psuedo-inverse* $\mathbf{\Pi}^+$ [25],

$$\boldsymbol{J}^* = \mathbf{\Pi}^+ \boldsymbol{E} \tag{4.8}$$

The pseudo-inverse can be expressed explicitly as follows:

- For an overdetermined system, $\text{rank}(\mathbf{\Pi}) = d < m$,

$$\mathbf{\Pi}^+ = (\mathbf{\Pi}^\top \mathbf{\Pi})^{-1} \mathbf{\Pi}^\top \tag{4.9}$$

- For an underdetermined system, $\text{rank}(\mathbf{\Pi}) = m < d$,

$$\mathbf{\Pi}^+ = \mathbf{\Pi}^\top (\mathbf{\Pi}\mathbf{\Pi}^\top)^{-1} \tag{4.10}$$

When the system is overdetermined, the solution given by Equation 4.8 is the sole unique solution to the OLS problem. In contrast, when the system is underdetermined, the solution corresponds to the solution with minimum norm ($||\boldsymbol{J}^*||_2$), taken from the set of infinitely many solutions [25].

Furthermore, by using the pseudo-inverse and considering $\mathbf{\Pi}\boldsymbol{J}^*$ as a linear combination of the feature vectors, the OLS problem 4.5 can be seen as obtaining the projection of the vector of calculated energies $\boldsymbol{E}$ onto the range of the correlation matrix $\mathcal{R}(\mathbf{\Pi})$ [25], such that the distance or *residual* r^* or equivalently the square of the residual minimum,

$$\begin{aligned} r^2 &= ||\boldsymbol{E} - \mathbf{\Pi}\boldsymbol{J}^*||_2^2 \\ &= ||\boldsymbol{E} - \mathbf{\Pi}\mathbf{\Pi}^+ \boldsymbol{E}||_2^2 \\ &= ||\boldsymbol{E} - \mathcal{P}_{\mathcal{R}(\mathbf{\Pi})}(\boldsymbol{E})||_2^2 \end{aligned} \tag{4.11}$$

where the projection onto $\mathcal{R}(\mathbf{\Pi})$, denoted $\mathcal{P}_{\mathcal{R}(\mathbf{\Pi})}$, is obtained using the projection matrix $H = \mathbf{\Pi}\mathbf{\Pi}^+$.

The fact that OLS solutions must satisfy the normal equations 4.7 and the corresponding in-sample prediction vector $\boldsymbol{\Pi}\boldsymbol{J}^*$ is the projection of the target vector $\boldsymbol{E}$ onto $\mathcal{R}(\boldsymbol{\Pi})$, helps to motivate and better understand the use of regularization in linear regression and particularly for learning Hamiltonians of configuration in applied lattice models.

The first reason to use *regularization*, as the name suggests, is for improving the stability of the solution to small disturbances of the sampled correlation functions. We can obtain a measure of the sensibility of a solution to disturbances in the calculated energies $\boldsymbol{E}$ by considering a bound on the norm of the corresponding change in the obtained coefficients $\boldsymbol{J}^*$. Such a bound can be obtained directly from the fact $\boldsymbol{J}^*$ satisfies the normal equations, and is given by the *condition number* of the matrix $\boldsymbol{\Pi}^\top\boldsymbol{\Pi}$ [25],

$$\kappa(\boldsymbol{\Pi}^\top\boldsymbol{\Pi}) = \frac{s_{\max}}{s_{\min}} = \kappa^2(\boldsymbol{\Pi}) \tag{4.12}$$

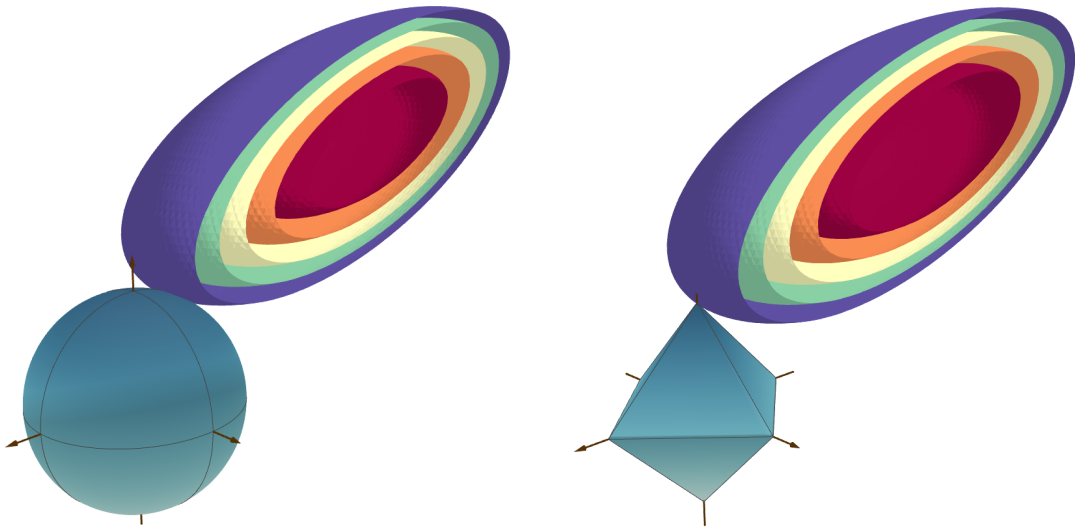
where $s_{\max}$ and $s_{\min}$ are the singular values of the matrix $\boldsymbol{\Pi}^\top\boldsymbol{\Pi}$, and are each equal to the square of the corresponding maximum and minimum singular values of $\boldsymbol{\Pi}$.

When the smallest singular value of the matrix $\boldsymbol{\Pi}$ is small relative to the largest, the condition number will be large; and when the matrix is singular (i.e. $s_{\min} = 0$), its condition number diverges, $\kappa(\boldsymbol{\Pi}) = \infty$. Regularization prevents the linear system from being close to singular. Due to sampling complications which are discussed in detail in Chapter 5.1, correlation matrices can often be poorly conditioned. Regularization directly improves the condition number by modifying the matrix in the normal equations to one where all the original singular values are shifted by a positive value proportional to the regularization. By improving the condition number of the regression problem, more numerically stable solutions can be obtained [25, 90].

The second purpose of regularization is that of *shrinkage*, which entails forcing solutions to have a small norm. The motivation for seeking regularization and shrinkage can be succinctly summed up in the *bias-variance trade-off*; where introducing a regularization term will increase the model bias—or its flexibility to represent the training data—but lower the model variance and as a result yield more stable model coefficients [90]. The *bias-variance trade-off* usually leads to better out-of-sample prediction accuracy at the cost of lowering in-sample prediction accuracy, or in other words, preventing over-fitting.

Another way to understand *shrinkage* of regression solutions with regularization norms, is to consider the geometry of solutions in terms of the elliptical OLS *level sets* (isosurfaces),[3] and the level sets of the regularization norm (also known as *norm-balls*). The solution geometry for regularized regression using ℓ_2 (Ridge $||\boldsymbol{J}|_2^2$) and ℓ_1 (Lasso $||\boldsymbol{J}||_1$) regularization in $\mathbb{R}^3$ is shown in Figure 4.2. Shrinkage can be understood as arising from the monotonically increasing nature of the regularizing norm. This means that the level sets for any of the norm balls depicted are physically larger for increasing values of the norm. Hence the norm penalization will tend to drive solutions closer to the origin compared to the OLS solution, which in turn increases model *bias* but decreases model *variance*.

[3] A level set or isosurface is the locus of points where the value of a function is equal to a constant k, $f(\boldsymbol{x}) = k$.



(a) Ridge (ℓ_2 norm) (b) Lasso (ℓ_1 norm)

Figure 4.1: Regularized regression solution geometry for $\boldsymbol{J} \in \mathbb{R}^3$. (a) Ridge ($\ell_2$) solution geometry. (b) Lasso (ℓ_1) solution geometry.

Sparsity inducing regularization norms

Another especially important reason for using regularization involves *feature selection*. The use of specific sparsity-inducing norms for regularization results in feature selection by shrinking coefficients for less important features exactly to zero. This allows fitting much sparser applied lattice models in one shot. Although there exist other methods for feature selection that do not rely on regularization [88, 91], feature selection by regularization is now overwhelmingly used over other methods for fitting applied lattice models [2, 148, 153, 227].

Although shrinkage and regularization can be obtained with any norm, even if its norm-ball is smooth everywhere, feature selection from regularization can only be obtained by regularizing with non-smooth norms. Feature selection occurs when the elliptical level sets (isosurfaces) of the least squares objective in Equation 4.6 impinge on singular points (sharp edges or vertices) of the norm-ball for the regularization term used. Solutions for problems using non-smooth norms will appear with very high probability at a singular point [7]. This behavior yields sparse solutions precisely because many elements of the solution vector are exactly zero at those singular points. This solution geometry is shown in $\mathbb{R}^3$ for the case of the Lasso in Figure 4.1b, where sections of the isosurfaces of the least squares objective are shown in different colors. Additional norms with different feature selection properties are also shown in Figure 4.2; where one can observe that sharp edges and vertices occur at axes and/or planes spanned by the axes.

To further understand feature selection in regularized regression problems, it is useful to

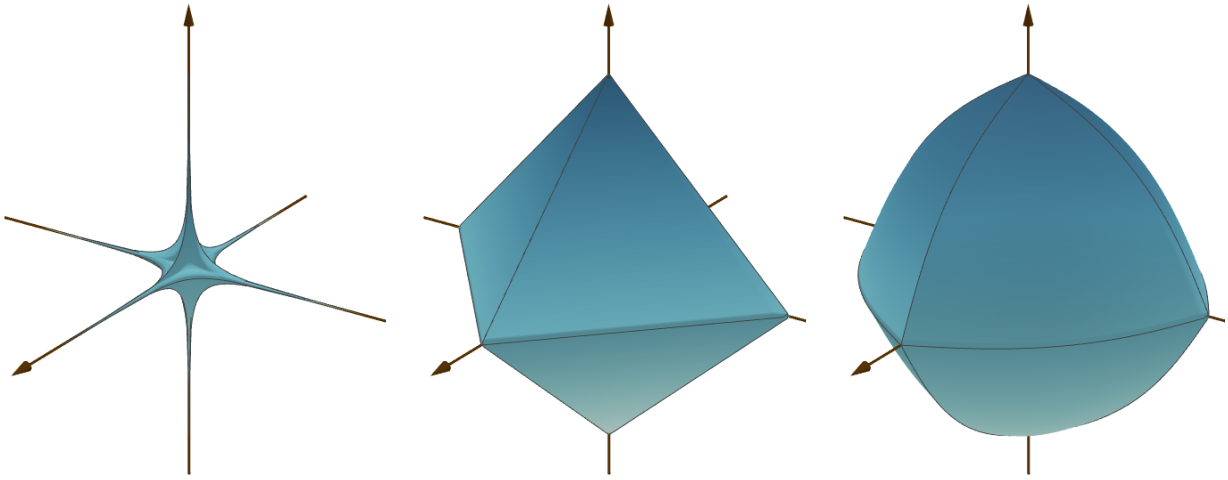


Figure 4.2: Unit norm-balls for the solution vector $\boldsymbol{J} \in \mathbb{R}^3$ for three different types of regularization: The ℓ_0 pseudo-norm, Lasso (ℓ_1 norm) and Elastic net (convex combination of ℓ_2 and ℓ_1 norms).

introduce the ℓ_0 pseudo-norm of a vector in $\mathbb{R}^n$, which is shown in Figure 4.2 for $\mathbb{R}^3$. The ℓ_0 pseudo-norm can be formally defined by the following limiting procedure [64],

$$||\boldsymbol{J}||_0 = \lim_{p \to 0} ||\boldsymbol{J}||_p^p = |\{i, : J_i \neq 0 \;\; \forall i \in [n]\}|$$

The ℓ_0 norm essentially counts the number of non-zero coefficients in a vector[4] and so is a direct measurement of *sparsity*. However, the regression problem with ℓ_0 regularization is non-convex, which is a direct result of the non-convexity of the ℓ_0 norm-ball shown in Figure 4.2b. As a result, obtaining solutions constitutes a complex combinatorial search and is in general an NP-hard problem [63].

The most common approach to obtain an approximate solution to ℓ_0 regularized regression is to solve the corresponding *convex relaxation* of the problem by replacing the ℓ_0 norm with an ℓ_1 norm. In this sense, feature selection via regularization can be thought of as a convex relaxation of ℓ_0 selection. Linear regression using an ℓ_1 norm for regularization is known as the Least Absolute Shrinkage and Selection Operator (Lasso) [220]. The Lasso has become a popular and efficient method for fitting sparse cluster expansions [9, 152,

[4]It is not formally a norm, since it does not satisfy all of the mathematical requirements for a norm. Most notable it is not sensitive to scale, $||k\boldsymbol{J}||_0 = ||\boldsymbol{J}||_0$ for any scalar k.

153]. However, the Lasso has some notable limitations, including its lack of strict convexity and selection irregularity. In addition, the Lasso can have reduced prediction performance (compared to the Ridge) in cases with highly correlated features [262].

The Elastic Net, which uses a convex combination of the ℓ_2 and ℓ_1 norms as shown in Figure 4.2, has been specifically developed to address some of these shortcomings. Although the Elastic Net has not seen much use for fitting applied lattice models, the use of an ℓ_2 norm along with a norm resulting in feature selection is usually beneficial. Apart from further improving the conditioning of the regression problem, it also makes the regression problem strictly convex [90, 93]. Further, an ℓ_2-norm can also be motivated from a *physical feasibility ansatz*. The ansatz requires that the total variance of a Hamiltonian be bounded by a *reasonable value*. In particular, as we have shown in Chapter 2.3, the square of the ℓ_2 norm (without including the offset term $\boldsymbol{J}_0$) of the Fourier correlation function coefficients corresponds exactly to the variance of the fitted Hamiltonian, $||\boldsymbol{J}||_2^2 = \text{Var}_{\boldsymbol{\rho}}[H(\boldsymbol{\sigma})]$. As a result, one can think of any regression that includes an ℓ_2 regularization term, as the equivalent regression problem with a prescribed bound on the total energy variance, written out as follows,

$$\begin{aligned} \boldsymbol{J}^* &= \underset{\boldsymbol{J}}{\text{argmin}}\ ||\boldsymbol{\Pi J} - \boldsymbol{E}||_2^2 + \gamma||\boldsymbol{J}||_2^2 \\ &= \underset{\boldsymbol{J}}{\text{argmin}}\ ||\boldsymbol{\Pi J} - \boldsymbol{E}||_2^2 \\ &\quad \text{subject to} \quad \text{Var}_{\boldsymbol{\rho}}[H(\boldsymbol{\sigma})] \leq s_{\text{max}}^2 \end{aligned} \tag{4.13}$$

where s_{max}^2 is the maximum variance that the resulting Hamiltonian can have; and there is a one-to-one correspondence between the values of the hyper-parameter λ and the values of s_{max}^2 [90]. In general terms, we insist that an ℓ_2 regularization should be used when fitting the majority of applied lattice models, with perhaps the exception of redundant representations and those dealing with very well-conditioned linear systems.

Hyper-parameter selection

Regularized regression models will have at least one hyper-parameter associated with the regularization term, and models that mix more than one norm, such as the Elastic-Net, have two hyper-parameters. Selecting appropriate hyper-parameters is critically important because the hyper-parameters control the importance given to a regularization term, and consequently the resulting amount of shrinkage and/or feature selection. Accordingly, the hyper-parameters strongly affect the resulting prediction accuracy. The standard way to determine these hyper-parameters is by using Cross-Validation (CV) optimization [90, 93].

Determining a hyper-parameter value using CV optimization involves minimizing a CV score, most commonly the root mean square error (RMSE), with respect to the relevant hyper-parameter. CV involves splitting the available training data randomly into k sets of equal size. Subsequently, k fits are computed using the data from all combinations involving $k-1$ sets. The CV score is the average RMSE for all k fits computed with respect to the

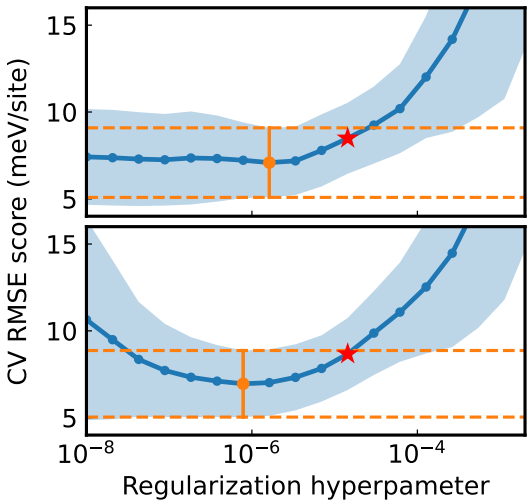


Figure 4.3: CV score regularization paths for Sparse Group Lasso fits of an LMTOF rocksalt system. The top plot shows the path for a fit using pairs up to 7Å, triplets up to 4.2Å, and quadruplets up to 4.2Å. The bottom plot shows the path for a fit using pairs up to 7Å, triplets up to 5.6Å, and quadruplets up to 5.6Å.

k-th set that was not included in each fit. When using k sets, the procedure is called k-fold CV, and the most commonly used values of k are 1, 5, and 10. For $k = 1$ the procedure is known as Leave-One-Out CV (LOOCV), and its use in learning applied lattice models has been extensively discussed [149, 227].

Choosing the number of folds for CV is a choice left to the practitioner, and there are no hard rules on which and when to choose a particular value of k. Nevertheless, we can say that in general smaller values of k tend to produce models with lower bias (and higher variance) and may have a tendency to exhibit over-fitting. Larger values of k will show lower variance but higher bias which may affect overall model performance, particularly considering the fact that training data for CEs is expensive and often scant. A good recommended compromise for the number of folds is 5 or 10 [91].

In addition, we emphasize what is known as the *one-standard error rule* [91], because it is particularly applicable to CE fitting. The one-standard error rule states that when choosing a hyper-parameter with feature selection (sparsity) as one of the goals, it is recommended to choose the largest value of the hyper-parameter for which the CV-RMSE is within one standard deviation of the minimum CV-RMSE [91] and results in better sparsity. The reason behind this is that the hyper-parameter value that minimizes CV error optimizes for prediction accuracy but not for feature selection, and a sufficient reduction in model complexity may be well worth the cost of a slightly larger CV-RMSE.

Figure 4.3 shows the regularization paths for two fits of a $LiMnO_2$-Li_2TiO_3-LiF system using Sparse Group Lasso regression with different sets of cutoffs. The mean CV score is

shown in blue, and the standard deviation is shaded. The minimum CV score is marked in yellow, and the corresponding standard deviation region is marked with dashed lines. According to the *one-standard error rule*, the models that should be chosen are marked with a red star.

Although the *one-standard error rule* by itself should not be taken as a definitive rule for applied lattice model selection, it serves as general guidance for practitioners to select models that are both accurate and parsimonious, rather than solely optimizing CV-score at the cost of sparsity. This is particularly important to keep in mind for the commonly occurring *CV-plateau* scenario, which is present in the top plot of Figure 4.3. In systems exhibiting a *CV-plateau* the CV minimum can often occur at hyper-parameter values far into the plateau region, and as a result, using the hyper-parameters for the CV minimum results in models with severely compromised sparsity and only marginal improvements in CV score compared to those obtained following the *one-standard error rule*.

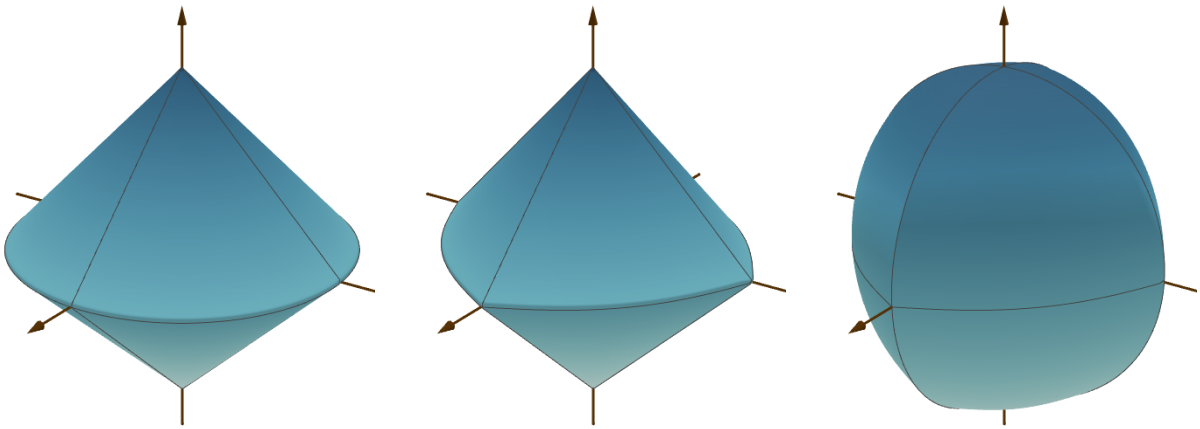
4.3 Structured sparsity

The feature selection discussed thus far leads to *unstructured sparsity*, meaning that coefficients are set to zero individually without regard to the degree of the corresponding correlation function or any relationships between correlation functions. However, having thoroughly explored details of representation and implementations of lattice Hamiltonians in Chapters 2 and 3, we can motivate the use of *structural priors* or patterns that expansion coefficients should follow. Feature selection that follows structural priors leads to regression solutions that exhibit *structured sparsity*. A wide variety of structural priors on coefficients can be obtained by generalizing norms over disjoint or overlapping groups of coefficients [8] or equivalently by introducing linear constraints between coefficients. We will first introduce regression methodology based on the Group Lasso and mixed integer quadratic programming (MIQP) formulations of ℓ_0 regularization that can be used to introduce such structural priors. Subsequently, we will introduce specific structural priors that we have found to yield sparser and more accurate applied lattice models for complex multi-component materials compared to models with unstructured sparsity.

The first structured sparsity regression model we consider is the Group Lasso. The Group Lasso is a generalization of the Lasso, where feature selection occurs in groups, such that all coefficients in a group are zero or all are nonzero. The Group Lasso problem uses a sum of ℓ_2 norms of groups of coefficients as follows,

$$\boldsymbol{J}^* = \underset{\boldsymbol{J}}{\operatorname{argmin}} \; ||\boldsymbol{\Pi J} - \boldsymbol{E}||_2^2 + \lambda \sum_{\boldsymbol{g} \in G} \sqrt{|\boldsymbol{g}|} ||\boldsymbol{J}_{\boldsymbol{g}}||_2, \tag{4.14}$$

where G is a set of groups of coefficient indices $\mathbf{g}$. $\boldsymbol{J}_{\boldsymbol{g}} \in \mathbb{R}^{|\boldsymbol{g}|}$ is a vector of only the coefficients in group $\boldsymbol{g}$. The scaling $\sqrt{|\boldsymbol{g}|}$ is commonly used to consider all groups equally regardless of size, however other weighting schemes can be used [93].



Group Lasso Sparse Group Lasso Overlap Group Lasso

Figure 4.4: Unit norm-balls for the solution vector $\boldsymbol{J}$ corresponding to each of the regularization models described. Feature selection occurs when the OLS problem level sets contact singular points of the corresponding norm-ball. The figure for the ℓ_0 pseudo-norm is actually for a small value of ℓ_p, $p \to 0$ and not exactly 0. When the value of $p = 0$ exactly, the surface becomes 6 singular points only at values of ± 1 along each of the axes.

When the groups G in Equation 4.14 are individual coefficients, the Group Lasso reduces to the Lasso. However, when groups include more than one coefficient, feature selection occurs in a group-wise manner, where either all of the coefficients in a group $\boldsymbol{g}$ are non-zero or all of them are exactly zero. This can be visualized in the corresponding Group-Lasso regularization norm-ball shown in Figure 4.4. In the Figure, two coefficients are grouped and thus the norm-ball has a continuous (circular) locus of singular points on the plane that they span.

The Group Lasso objective is convex [255], but not necessarily strictly so. In order for the Group Lasso to have unique solutions, each group must be full column rank (i.e. the feature vectors in a group must be linearly independent) [201].

A further extension of the Group Lasso that allows for in-group sparsity, called the Sparse Group Lasso [73, 202] can yield results with improved sparsity and provide within-group regularization. In Sparse Group Lasso, an ℓ_1 norm over all coefficients is added to the ℓ_2 norm over groups as a convex combination of regularization terms as follows,

$$\boldsymbol{J}^* = \underset{\boldsymbol{J}}{\operatorname{argmin}} \, ||\boldsymbol{\Pi J} - \boldsymbol{E}||_2^2 + (1-\alpha)\lambda \sum_{\boldsymbol{g} \in G} \sqrt{|\boldsymbol{g}|} ||\boldsymbol{J_g}||_2 + \alpha\lambda ||\boldsymbol{J}||_1 \tag{4.15}$$

Intuitively, the Sparse Group Lasso combines both the regular Lasso and the Group Lasso, as seen in both the curved edges on a group plane and the sharp vertices on all axes in the respective norm-ball in Figure 4.4. Within-group sparsity can be particularly useful for large complex lattice models where charge neutrality constraints, inaccessible configurations, or simply insufficient sampling give rise to rank deficiencies within groups. The ℓ_1 penalization at the individual correlation level yields an additional level of regularization that improves the conditioning of the overall regression problem [253]. ℓ_2 penalization within groups has also been proposed for this reason [201]. We make note, however, that when estimating expansion coefficients for a Potts frame expansion, any structural priors must result in within-group sparsity to effectively make use of the redundancy in the representation and permit sparse reconstruction of coefficients within a compressed sensing framework, as we detail in Section 4.6.

Another extension of the Group Lasso, known as the *Overlap Group Lasso* allows for structure with overlapping groups [102]. In order to achieve this, variables are replicated depending on the number of groups they are part of, and their final value is the sum of all replicated or latent variables. Whenever a coefficient J_i appears in n groups, each group contains a replicated variable $\tilde{J}_i^{(\boldsymbol{g}_j)}$; and the final value of the coefficient is the sum of all the replicated variables,[5]

$$J_i = \sum_{j=1}^{n} \tilde{J}_i^{(\boldsymbol{g}_j)} \tag{4.16}$$

Accordingly, norm-balls for the overlap lasso penalty will have intersecting loci of singular points, as shown for the case of three overlapping groups of two coefficients in $\mathbb{R}^3$ in Figure 4.4.

The overlap group lasso can be used to enforce hierarchical relationships between coefficients. This allows the inclusion of coefficients or groups of coefficients only if another group of coefficients is also included in the resulting model [93]. For example, if a group of coefficients in a group $\boldsymbol{g}_2$ are only allowed to be included if the set of coefficients in another group $\boldsymbol{g}_1$ are also included, then a group of replicated coefficients for group $\boldsymbol{g}_1$ and a *new* group of replicated variables that includes variables for both groups $\boldsymbol{g}_{1,2}$ is used to determine the final set of coefficients. Doing so means that coefficients associated with group $\boldsymbol{g}_1$ can be nonzero independently, but if coefficients in group $\boldsymbol{g}_2$ are nonzero then so will those in $\boldsymbol{g}_1$, by virtue of including replicated latent variables in group $\boldsymbol{g}_{1,2}$. These hierarchical patterns can be constructed by *staggering* the overlap of groups of replicated variables [133].

Grouped and hierarchical structural priors can also be obtained by a *mixed integer quadratic programming* (MIQP) formulation of the ℓ_0 regularized regression problem. However, since the ℓ_0 norm only provides feature selection but no shrinkage or regularization, including a convex combination of a norm that does so, is practically advantageous [98, 259], and is also motivated by the bounded variance ansatz given in Equation 4.13. Furthermore,

[5] This formulation is slightly different than the original [102] where a latent vector $\boldsymbol{v}_{\boldsymbol{g}}\mathbb{R}^p$ is used for all groups, and each $\boldsymbol{v}_{\boldsymbol{g}}$ is constrained to be zero in all elements $i \neq \boldsymbol{g}$.

the ℓ_0-norm can trivially be generalized to count the number of nonzero groups of coefficients $\boldsymbol{J}_{\boldsymbol{g}}$ for all groups $\boldsymbol{g} \in G$ in a set of groups G, by generalizing Equation 4.2 as follows,

$$||\boldsymbol{J}||_{0,G} = |\{\boldsymbol{g} \ : \boldsymbol{J}_{\boldsymbol{g}} \neq \boldsymbol{0} \ \ \forall \boldsymbol{g} \in G\}| \tag{4.17}$$

Using the ℓ_0-norm over groups, a resulting generalized $\ell_2\ell_0$-regularized regression problem can be expressed as follows,

$$\boldsymbol{J}^* = \underset{\boldsymbol{J}}{\operatorname{argmin}} \ ||\boldsymbol{\Pi}\boldsymbol{J} - \boldsymbol{E}||_2^2 + \alpha\lambda||\boldsymbol{J}||_{0,G} + (1-\alpha)\lambda||\boldsymbol{J}||_2 \tag{4.18}$$

where the regularization hyper-parameter $\alpha \in [0, 1]$ is constrained to lie between zero and one.

The ℓ_0 regularized regression problem in Equation 4.18 is an NP-hard problem, but suitable near-optimal solutions can be found for moderately sized applied lattice models (up to 500 groups of coefficients[6]) using a MIQP formulation [17, 21]. The problem in Equation 4.18 transformed to a MIQP takes the following form,

$$\begin{aligned} \min_{\boldsymbol{J}} \ \ & \boldsymbol{J}^\top \left(\boldsymbol{\Pi}^\top\boldsymbol{\Pi} + (1-\alpha)\lambda\boldsymbol{I}\right)\boldsymbol{J} - 2\boldsymbol{E}_{\mathrm{S}}^\top\boldsymbol{\Pi}_S\boldsymbol{J} + \alpha\lambda\sum_{\boldsymbol{g}\in G} z_{\boldsymbol{g}} \\ \text{subject to} \quad & M z_{\boldsymbol{g}} \geq J_i \ \ \forall i \in \boldsymbol{g} \\ & M z_{\boldsymbol{g}} \geq -J_i \ \ \forall i \in \boldsymbol{g} \\ & z_{\boldsymbol{g}} \in \{0, 1\} \end{aligned} \tag{4.19}$$

where $\boldsymbol{I}$ is the identity matrix; and $z_{\boldsymbol{g}}$ is slack variable that describes whether a group of coefficients $\boldsymbol{J}_{\boldsymbol{g}}$ is zero or non-zero, or equivalently if a group of correlation functions $\Theta_{\beta(i)} \ \ \forall i \in \boldsymbol{g}$ is active ($z_{\boldsymbol{g}} \neq 0$) or inactive ($z_{\boldsymbol{g}} = 0$).

This transformation of the regression problem into a MIQP is also what allows the introduction of *hierarchical* constraints as linear constraints on the auxiliary slack variables $z_{\boldsymbol{g}}$. Since we have that $z_{\boldsymbol{g}} \in \{0, 1\}$, hierarchical constraints can be expressed as inequality constraints between slack variables. For example, if a group of coefficients $\boldsymbol{J}_{\boldsymbol{g}_2}$ can only be nonzero if another group of coefficients $\boldsymbol{J}_{\boldsymbol{g}_1}$ is also nonzero, one would add the following constraint between their corresponding slack variables to the problem in Equation 4.19: $z_{\boldsymbol{g}_2} \leq z_{\boldsymbol{g}_1}$.

Having described two effective methods that allow the incorporation of a wide variety of different *structural priors* to sparse linear regression, we will now motivate and described two useful structural priors that result in robust, accurate, and sparse expansions of lattice Hamiltonians. In doing so will use the mathematical structure, properties, and statistical interpretations of the cluster decomposition detailed in Chapter 2.4 to derive, motivate and rationalize these structural priors. In Chapter 5 we show examples of improved fits of ternary alloys and lithium transition metal oxyfluorides using the structured sparsity-based methods described.

[6] At the time of writing this number is more than sufficient for any of the applied lattice models reported in the literature.

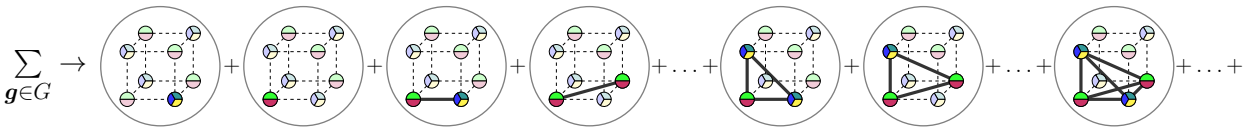


Figure 4.5: Illustration of group regularization by grouping correlation functions that act over the same orbits of site space clusters. $\boldsymbol{g}$ labels group of correlation functions. Each circled figure in the sum represents a different group of correlation functions, analogous to the one shown in (b). G is the set of all groups considered in the expansion (i.e. all cluster interactions). The same convention as Figures 2.7 and 2.9 are used: site coloring in the images represent non-constant site functions. In the illustration, there are two types of site spaces, one with 4 allowed species (3 non-constant site functions); and another with 3 allowed species (2 non-constant site functions).

Group sparsity

Selecting coefficients based on grouping correlation functions that make up each cluster interaction term H_B, or more generally in any representation, by the orbits of site space clusters B over which they operate on, is a judicious form of structured sparsity. Broadly, this approach to structure sparsity is mathematically and physically motivated to regularize over groups of correlation functions that represent a single multiple-body term in the expansion of a lattice Hamiltonian. The concept of grouped sparsity by correlation interactions is shown schematically in Figure 4.5 following the same conventions introduced in Figure 2.9, where the circled figures represent groups $\boldsymbol{g}$ of correlation functions that operate over the same orbit of site space clusters B—that is that make up the same cluster interaction H_B,

We have formally shown that the particular choice of standard basis used to represent a particular lattice Hamiltonian is irrelevant since the underlying unique cluster decomposition is invariant to any arbitrary transformation between standard basis sets. However, in practice, one has to choose a particular standard basis set in order to estimate expansion coefficients. Hence it is worthwile to seek out regression methods that result in the same (up to the available estimation accuracy) cluster decomposition irrespective of the standard basis sets used to construct the Fourier correlation functions.

Group regularization can be used precisely to construct an estimation algorithm that is independent of the particular choice of standard site basis. Based on the fact that the effective cluster weights $\mathrm{W}[H_B]$ are invariant to standard basis transformations as is shown in Chapter 2.4, regularization penalties written in terms of $\mathrm{W}[H_B]$ will likewise be invariant. This can be trivially introduced into ℓ_2 regularization by including a diagonal weighting matrix W_m with the appropriate cluster multiplicities such that each element of the vector

used for regularization is given by,[7]

$$(W_m \boldsymbol{J})_i = \sqrt{\hat{m}_{\beta_i}} J_{\beta_i} \tag{4.20}$$

When doing so, the resulting ℓ_2 norm squared $||W_m \boldsymbol{J}||_2^2$ will correspond to the sum of effective cluster weights,

$$\begin{aligned} ||W_m \boldsymbol{J}||_2^2 &= \sum_{\beta} \hat{m}_{\beta_i} J_{\beta_i}^2 \\ &= \sum_{B} \sum_{\beta \in B} \hat{m}_{\beta_i} J_{\beta_i}^2 \\ &= \sum_{B} \mathrm{W}[H_B] \end{aligned}$$

Furthermore, structured sparsity involving the selection of cluster interaction terms H_B that is agnostic the choice of standard site basis can be obtained by the Group Lasso in the following manner,

$$\boldsymbol{J}^* = \underset{\boldsymbol{J}}{\operatorname{argmin}} \; ||\boldsymbol{\Pi} \boldsymbol{J} - \boldsymbol{E}||_2^2 + \lambda \sum_{B \in G} w(B) ||W_{m_B} \boldsymbol{J}_B||_2 \tag{4.21}$$

where $w(B)$ is a scalar weight associated with the site space cluster B (i.e. such as its multiplicity m_B, the size of the clusters in B or a product of both). W_{m_B} is a diagonal block of W_m for the rows and columns corresponding to the correlation functions that make up the cluster interaction H_B. Similarly, $\boldsymbol{J}_B$ is a vector of the coefficients for said correlation functions.

An equivalent site basis agnostic estimation of coefficients can be obtained using the grouped $\ell_2\ell_0$ MIQP regression as follows,

$$\begin{aligned} \min_{\boldsymbol{J}} \;\; & \boldsymbol{J}^\top \left(\boldsymbol{\Pi}^\top \boldsymbol{\Pi} + (1-\alpha)\lambda W_m^2 \right) \boldsymbol{J} - 2 \boldsymbol{E}_{\mathrm{S}}^\top \boldsymbol{\Pi}_S \boldsymbol{J} + \alpha\lambda \sum_B z_B \\ \text{subject to} \quad & M z_B \geq J_\beta \;\; \forall \beta \in B \\ & M z_B \geq -J_\beta \;\; \forall \beta \in B \\ & z_B \in \{0, 1\} \end{aligned} \tag{4.22}$$

When using a cluster group type regularization penalty with representations other than a Fourier cluster expansion, such as the Potts frame or a non-orthonormal correlation basis, the resulting regression problem will no longer give basis agnostic solutions. However, the use of cluster group regularization can still be heuristically motivated by the selection of multiple-body terms in the expansion, i.e. those that operate over the same orbits of sites. Doing so in practice may also result in increased sparsity and accuracy for complex materials with high dimensional configuration spaces [253].

[7] In fact this would correspond to the more general Tikhonov regularization [222].

Hierarchically constrained sparsity

Another compelling form of structured sparsity involves establishing hierarchical relations between correlation functions. This can be used to enforce physically motivated heuristics, such as the inclusion of correlation functions over larger clusters only if correlation functions over all subclusters are included. These heuristic based structural priors have been known and discussed to great length in the cluster expansion literature [98, 131, 227, 256, 259].

Apart from intuition and practical heuristics, establishing hierarchical relationships that ensure large degree expansion functions (i.e. those that operate over clusters with many sites) are included only if certain smaller degree functions are also included can be motivated by appealing to a notion of *stability* or *smoothness* of the resulting lattice Hamiltonian. In broad terms, we say a function of configuration is *stable* when the values predicted for two configurations that are different only by relatively few occupation variables are similar relative to the total variance of the function. For a given value of the total variance, a function where most of the variance is concentrated on low degree terms is more stable than one with more variance carried by higher degree terms. Although we will not do so here, this line of argument can be formally constructed using the notions of *noise stability* and *noise sensitivity* of functions over probability product spaces [146, 157].

We can further rationalize and justify including hierarchical priors by invoking the accepted statistical principle that interaction terms should only be included in a model if their main effects are already included. These hierarchy structural priors have been developed for various statistical applications under names including, the *heredity principle* [42, 86], *marginality constraints* [140], and *well-formulated* models [172]. Under these principles, there is a notion of *weak hierarchy* and *strong hierarchy*. Weak hierarchy is satisfied when at least one of the main effects of an interaction is included when that interaction is included the resulting expansion. Strong hierarchy, on the other hand, is only satisfied if all main effects are included along with an interaction [18].[8].

The principle of *hierarchically well-formulated* models is almost self-evident in the connection of the cluster decomposition and functional ANOVA. With the aforementioned reasoning of establishing hierarchical priors to obtain *stable* and *well-formulated* models, we describe two different forms of hierarchically constrained sparsity that prove to be quite effective for fitting applied lattice models of complex multi-component materials.

The first involves imposing hierarchical constraints between higher degree correlation functions and their lower degree factors. Higher degree correlation functions are only allowed to have nonzero coefficients if all of their lower degree factors do so too. This form of structured sparsity is shown schematically in Figure 4.6a, where the constraints between correlation functions and their factors are represented by edges connecting them. This form of hierarchically constrained sparse regression has been recently shown to be quite promising for fitting applied lattice models of ternary alloys and disordered ionic materials [98, 131, 259]. However, this form of hierarchy prior only satisfies weak hierarchy. Additionally, hierarchical

[8]Many argue that models that violate strong hierarchy are not *statistically* sensible and violating strong hierarchy amounts to postulating a special position to the origin [18, 140].

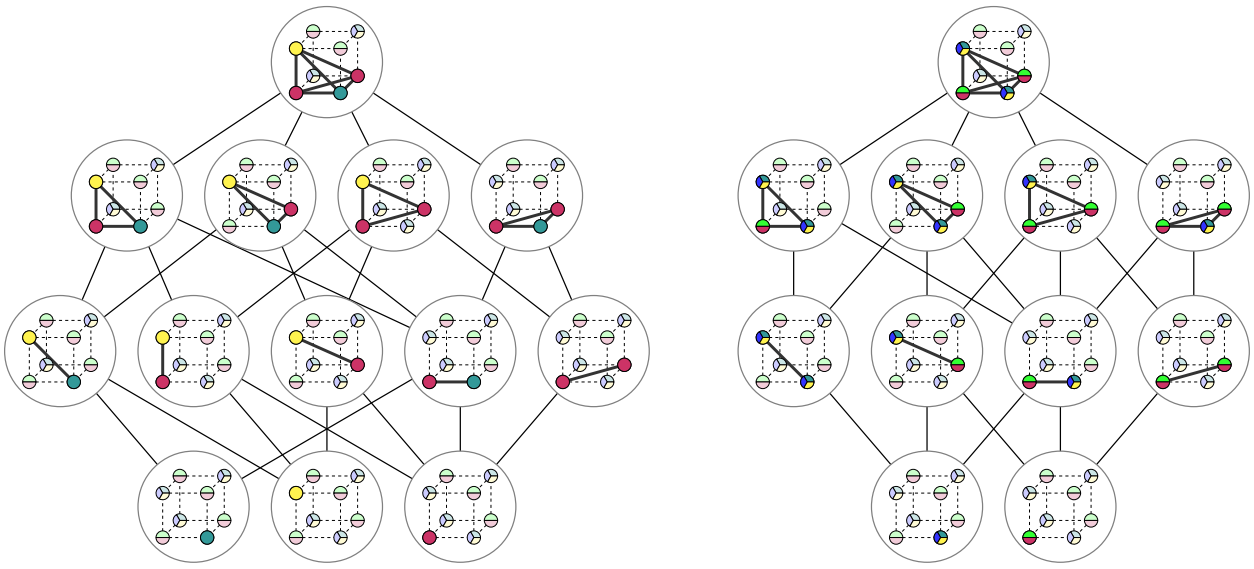


Figure 4.6: Schematic illustrations of hierarchically constrained sparsity. The site coloring in the images represents non-constant site functions. In the illustration, there are two types of site spaces, one with 4 allowed species (3 non-constant site functions); and another with 3 allowed species (2 non-constant site functions). (a) Hierarchical relations for a specific quadruplet correlation function and all its possible factors ensure the recovery of a model with weak hierarchy. (b) Hierarchical relations between quadruplet cluster interactions and lower order interactions; or equivalently between groups of correlation functions acting over the same orbits of quadruplet clusters and all correlation function groups acting over the orbits of sub-clusters of the quadruplet cluster. These illustrated constraints result in a model with strong hierarchy.

constraints between correlation functions and their factors result in the estimation of Fourier cluster expansion coefficients that depend on the choice of standard site basis.

An alternative form of hierarchy constraints that satisfies strong hierarchy and leads to coefficient estimates that are invariant to the choice of standard site basis, is obtained by establishing hierarchy constraints among cluster interactions H_B. In this case, the hierarchical constraints are between the groups of correlation functions that are grouped according to the group structure previously described and shown in Figure 4.5. A representation of this hierarchy structure is shown in Figure 4.6b as a graph representing the hierarchical relations between cluster interactions (orbit of correlation functions). To the best of our knowledge, this regression model has not been previously used for fitting applied lattice models[9].

[9]For binary cluster expansions there is no distinction between the two forms of hierarchical constraints described since there is only one correlation function associated with each orbit of site space clusters.

Weak hierarchy constraints can be obtained using $\ell_2\ell_0$ regression by including slack variables for every correlation function Θ_β considered in the fit (i.e. using singleton groups $\boldsymbol{g} = \{\beta\}$), and including inequality constraints between each correlation function Θ_β of degree d and all correlation functions Θ_γ involving clusters of degree $d-1$ which are its factors, as follows,

$$z_\beta \leq z_\gamma \; \forall \gamma \text{ s.t. } \text{ctr}(\boldsymbol{\eta}) \subset \text{ctr}(\boldsymbol{\alpha}) \; \forall \boldsymbol{\eta} \in \gamma \text{ and } \boldsymbol{\alpha} \in \beta$$

Similarly, strong hierarchy constraints can be obtained by using a slack variable z_B per group of correlation functions associated with the same orbit of sites B, and including inequality constraints between a slack variable z_B, and all slack variables z_D for site orbits D of clusters that are subclusters of the clusters in B, denoted $D \sqsubset B$,

$$z_B \leq z_D \; \forall D \text{ s.t. } D \sqsubset B$$

Weak and strong hierarchy constraints can also be obtained by using the Overlap Group Lasso and using *staggered* groups of replicated variables to introduce the same structural priors [133] as depicted in Figure 4.7.

The particular implementation of the regression model that ensures weak or strong hierarchy constraints by way of an ℓ_0 MIQP formulation or an Overlap Group Lasso formulation is a practical choice only. Both implementations can effectively introduce the structural priors we have described. However, using the Overlap Group Lasso, which is a convex optimization problem, for obtaining lattice Hamiltonian fits for materials with complex and very large configuration spaces may be more appropriate when NP-hard ℓ_0 MIQP near-optimal solutions are not practical. On the other hand, when near-optimal solutions to $\ell_2\ell_0$ can be obtained in a reasonable time, such a formulation can result in improved sparsity and in some cases also improved accuracy.

4.4 Adaptive regularization

Adaptive or iteratively re-weighted regularization in Lasso and Group Lasso regression has been shown to lead to enhanced sparsity, improved model selection consistency, with possibly improved prediction accuracy [28, 238, 240]. The adaptive form of a Group Lasso estimator consists of using a weighted norm in the regularization term. The adaptive norm can be written as follows,

$$||\boldsymbol{J}|| = \sum_{\boldsymbol{g} \in G} w_{\boldsymbol{g}} ||\boldsymbol{J}_{\boldsymbol{g}}||_2 \tag{4.23}$$

where $w_{\boldsymbol{g}}$ are the components of the corresponding weight vector. When the groups $\boldsymbol{g}$ are singletons, the index i runs over the individual elements of the coefficient vector $\boldsymbol{J}$ for the case of the standard adaptive Lasso [28, 261]. More generally, when $\boldsymbol{g}$ are groups with more than one coefficient regularization corresponds to the adaptive variant for Group Lasso [240].

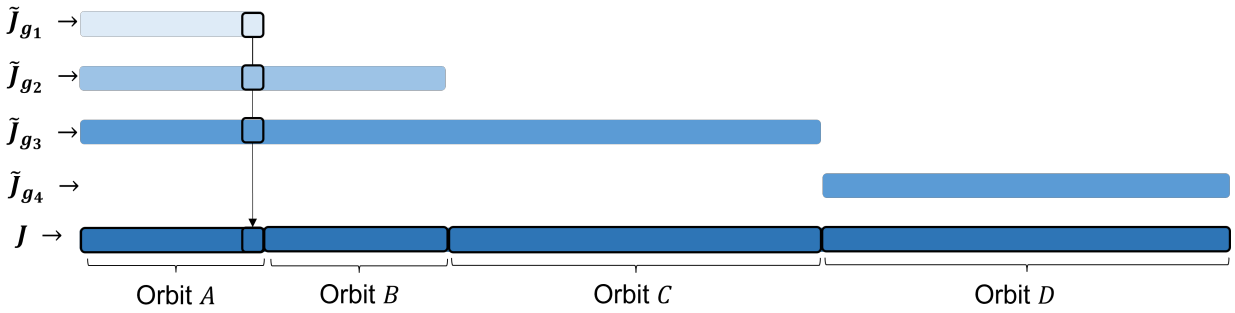


Figure 4.7: Illustration of how site space cluster orbit hierarchical constraints can be established by way of latent variables using the Overlap Group Lasso. In the example shown if coefficients corresponding to orbit C are nonzero, then coefficients for orbits B and A are nonzero as well by virtue of selecting the latent variable $\boldsymbol{J}_{g_3}$; and if coefficients for orbit B are nonzero, coefficients for orbit A are necessarily non-zero as well from latent variable $\boldsymbol{J}_{g_2}$. This would respect the hierarchy $A \sqsubset B \sqsubset C$. Coefficients for orbit D are independent of the rest.

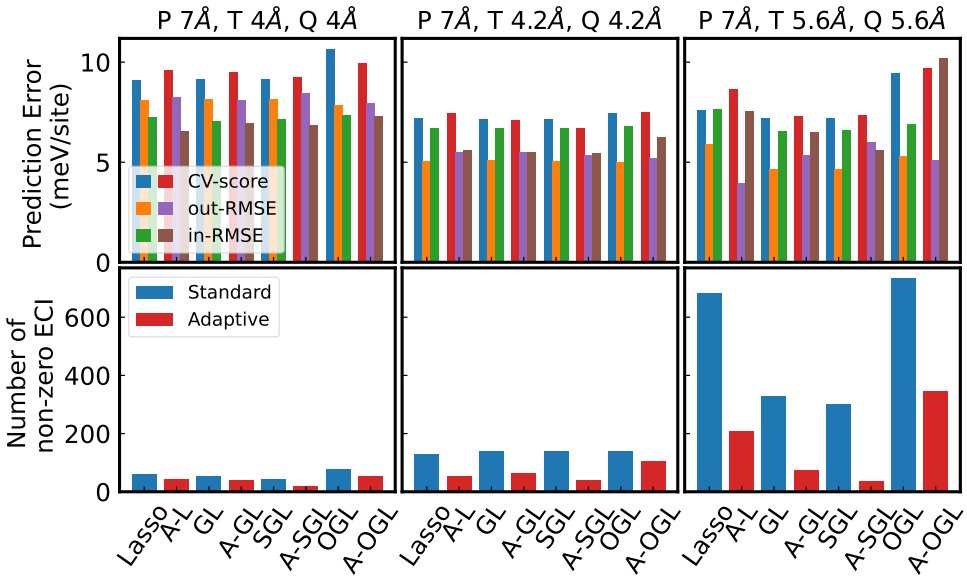


Figure 4.8: Fit metrics for Adaptive Lasso variants and standard Lasso variants using 3 different cutoff sets for generating cluster expansion terms of a $LiMnO_2$-Li_2TiO_3-LiF material. (L) Lasso, (GL) Group Lasso, (SGL) Sparse Group Lasso, (OGL) Overlap Group Lasso, (A-) are adaptive variants.

During the fitting process, the weights are updated iteratively using a decreasing function of the norm of the current estimate of the coefficients $\boldsymbol{J_g}$. The most common form used in practice is given by,

$$w_{\boldsymbol{g}}^{(l+1)} = \frac{1}{||\boldsymbol{J}_{\boldsymbol{g}}^{(l)}||_2 + \varepsilon}$$

where l is the current iteration in the iterative fitting and $\varepsilon > 0$ is a small offset to allow $||\boldsymbol{J}_i||_2$ to take on values of zero [28].

Results comparing root mean squared error (RMSE) accuracy and sparsity metrics for standard and adaptive versions of Lasso-based estimators for a set of fitted Hamiltonians of a $LiMnO_2$-Li_2TiO_3-LiF ceramic material are shown in Figure 4.8. In all cases, we see that the adaptive versions give substantial improvements in sparsity with minimal to no change in accuracy.

4.5 Fitting mixed models with explicit pair potentials

Although in theory any materials property of crystalline configuration can be fitted to any of the representations introduced in Chapter 2, certain properties, in particular those that involve long-range atomic interactions, require expansion terms up to very high spatial cutoffs and a large number of training configurations for large supercell structures that quickly becomes prohibitively expensive to compute by most first-principles electronic structure methods.

In many cases, long-range interactions can be mostly captured as pair-wise interactions only. In those cases using a *mixed model* that includes an empirical pair potential in addition to the lattice Hamiltonian, can prove to be an effective means to obtain a lattice model that includes fewer expansion terms and can be fitted with smaller supercell structures and often results in higher predictive accuracy. A mixed model of this kind is expressed as follows,

$$H(\boldsymbol{\sigma}) = \sum_{\beta} m_\beta J_\beta \Theta_\beta(\boldsymbol{\sigma}) + \lambda E_P(\boldsymbol{\sigma}) \tag{4.24}$$

Where the first term is any representation of a lattice Hamiltonian. $E_P(\boldsymbol{\sigma})$ is a pair potential that possibly includes structural parameters which are treated as constants determined by the underlying random crystal structure.[10] And λ is a mixing parameter, which depending on the particular pair potential used, can have an effective physical meaning.

Particularly, using a mixed model with a Coulomb electrostatic potential term has been shown to be a very effective way to construct applied lattice models of ionic materials with substantially improved prediction accuracy [181, 196]. In what follows we will focus our discussion specifically on the practical benefits of using a Coulomb pair electrostatic potential in addition to a lattice Hamiltonian expansion, and how to appropriately estimate model parameters using the previously presented regression methods.

[10] A pair potential may also include additional model parameters that must also be estimated, however, we do not discuss this more complicated scenario.

Explicit electrostatic pair potentials

The physical nature of long-range electrostatic interactions complicates a critical underlying premise for fitting lattice Hamiltonians under which an expansion truncation is justified by the rapid decay of correlations with respect to the physical distance between sites. For example, in the case of a system with only Coulomb interactions on a rigid lattice, the terms of the Fourier cluster expansion can be easily solved for analytically [31]. For the specific case of a binary system with sites with either positive charge q_+ or negative charge q_-, an expansion using a polynomial site basis will have pair correlation terms with coefficients given by,

$$J_{ij} = \frac{\kappa(q_+ - q_-)^2}{4r_{ij}} \tag{4.25}$$

The coefficients for pair terms decay slowly as the underlying Coulomb potential $\sim r^{-1}$. This slow decay in theory requires longer pair clusters to be included in a cluster expansion to correctly capture the long-range electrostatic interactions.

It was demonstrated that when only structures with electrostatic energy below a prescribed energy cutoff are considered for the simple binary system with only $+q$ and $-q$ charged species, an expansion with rapidly converging coefficients can be obtained [31]. Furthermore, such CE was shown to have low prediction error for out-of-sample structures below the prescribed energy cutoff [31]. This can be attributed to the locally neutral environments associated with low electrostatic energy structures [31].

For more complex ionic systems, such as those with hetero-valent species and/or cases including the effects of structural relaxations, considering only low electrostatic energy configurations and simply truncating the CE is usually not sufficient to ensure accurate and sufficiently sparse applied lattice models with only short-range terms [196]. Even applied lattice models with acceptable cross-validation (CV) scores may result in erroneous MC sampling—such as states with unphysical charge segregation—for large supercells when long-range interactions are not correctly accounted for.

A very effective way to handle systems with strong electrostatic interactions has been proposed and tested empirically [181, 196, 228]. By including an electrostatic term along with the CE Hamiltonian, a sparse and accurate applied lattice model can be constructed much more reliably, and MC sampling is improved by more accurately computing long-range electrostatics even in large supercell sizes that were absent in the training set. To do so, the CE and electrostatic interaction Hamiltonian is expressed as the following mixture model,

$$H(\boldsymbol{\sigma}) = \sum_{\beta} m_\beta J_\beta \Theta_\beta(\boldsymbol{\sigma}) + \frac{1}{\varepsilon_r} E_C(\boldsymbol{\sigma}) \tag{4.26}$$

where E_C represents the point electrostatic energy for a Coulomb potential, which can be computed efficiently and with high accuracy using the Ewald summation method [223] or

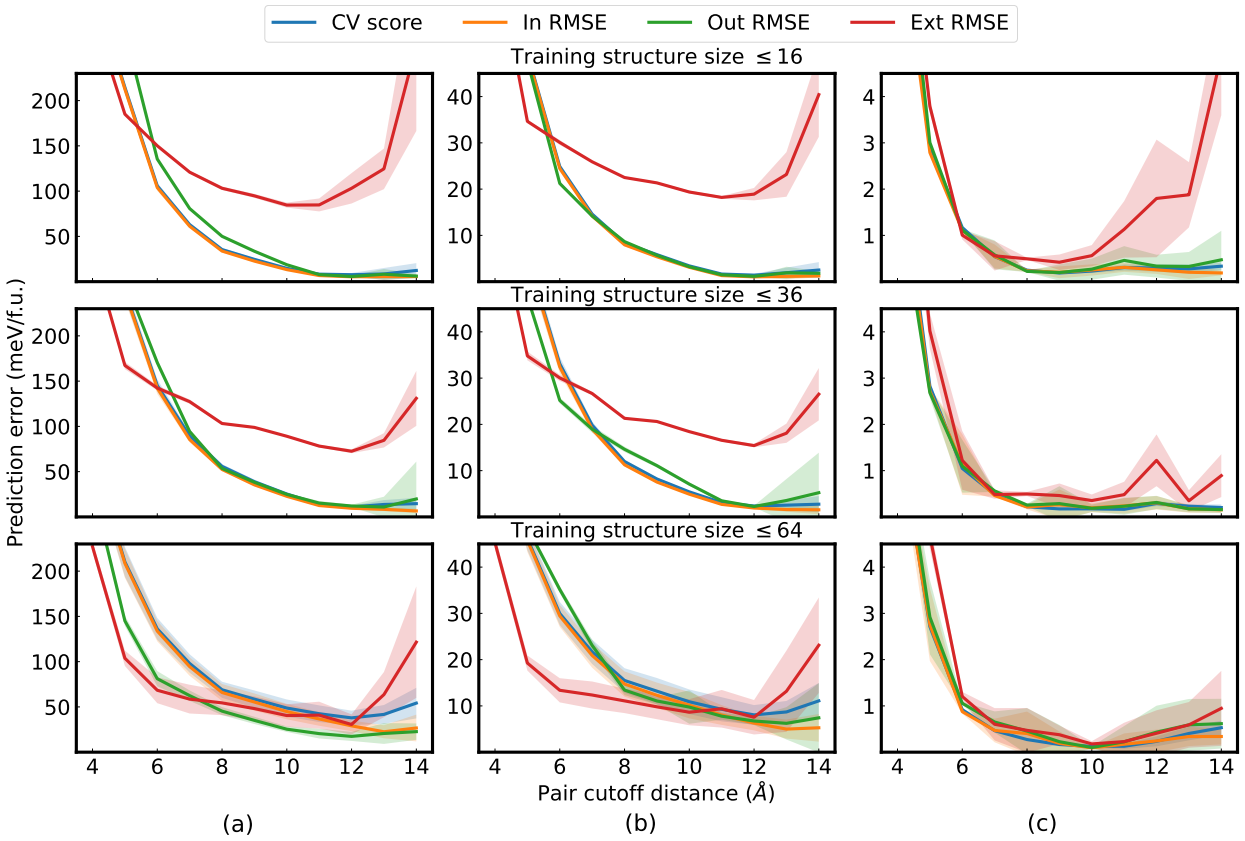


Figure 4.9: Prediction accuracy metrics for CE fits of empirical pair potentials of heterovalent (+1, +3 cation) and (-1, -2 anion) charges in a rocksalt structure. The metrics shown are RMSE cross-validation score (CV), in sample RMSE (in), and out of sample RMSE (out) with supercells with the same number of sites as the listed training structure size, and extrapolation RMSE to larger supercell sizes up to 144 sites (ext). Shaded areas denote ± one standard deviation for 50 different fits. (a) Accuracy metrics for CE fits of a Coulomb potential only. (b) Accuracy metrics for CE fits of a Buckingham-Coulomb potential. (c) Accuracy metrics for a Fourier cluster expansion and electrostatic model fits of a Buckingham-Coulomb potential.

the Fast Multipole Method [82]. The constant ε_r, which can be interpreted as an effective dielectric constant, should be fitted simultaneously by including the electrostatic term directly as a feature in the regression problem.

To illustrate the shortcomings of a cluster expansion in capturing electrostatics in heterovalent systems and the improvements obtained when including an explicit electrostatic term, we carried out several cluster expansion fits of a Coulomb electrostatic potential, as well as a sum of a Coulomb and a Buckingham pair potential. Both potentials were computed for a

system with hetero-valent ($+1, +3$ cation) and ($-1, -2$ anion) charges in a rocksalt structure. Further details, parameters, and results of the calculations are given in Appendix C.4.

Figure 4.9 shows curves for the resulting prediction accuracy metrics with their corresponding standard deviation for the previously described fits. Based on Figure 4.9 (a) and (b), we see that including longer range pairs in the applied lattice models without explicit electrostatics monotonically improves prediction accuracy for samples of similar-sized supercells. However, the extrapolation prediction of longer-period superstructures is severely compromised. The extrapolation prediction accuracy can only be reduced by adding pairs with distances up to those sampled in the training set structures. Including pairs with longer distances that are not present in the training, set ruins the extrapolation accuracy.

Figure 4.9 also shows that including larger period superstructures in training improves fits, as previously suggested in similar work [196]. However, doing so, such that the resulting CE converges to an acceptable level of accuracy, requires very large (and in many cases prohibitively large) data sets that must include large supercell structures. Furthermore, when fitting applied lattice models of ionic systems with more complex physical interactions in addition to long-range electrostatics, these issues can become worse such that fitting a reliable cluster expansion that captures both short and long-range interactions is not straightforward.

In contrast, Figure 4.9(c) shows fit metrics for the same Buckingham-Coulomb potential as 4.9(b), but for a fit using the mixed model cluster expansion with an explicit point electrostatic term computed with the Ewald summation method. The results show that the addition of the point electrostatic term in the cluster expansion substantially improves the resulting accuracy. Furthermore, by setting the cutoff for cluster expansion terms relatively shorter (≤ 8 Å), the resulting fit is substantially improved and has high prediction accuracy even for longer period superstructures. These results are also consistent with previous results computed for a similar point charge system with a spinel structure [196].

In addition, Figure 4.10 shows prediction accuracy metrics, the fitted value for the effective dielectric constant, and the in terms of the regularization hyper-parameter for a fit using Ridge (ℓ_2) regression. As the error in the fit converges, the value of the fitted dielectric constant approaches the true value used in the Buckingham-Coulomb potential, meaning that the electrostatic interactions are exactly captured by the electrostatic term, and the cluster expansion needs only to capture the short-range Buckingham interactions.

In the case of this simple additive potential, the convergence of the dielectric is only illustrative and the use of regularization is actually not necessary since the fit converges at the lowest values of the regularization hyper-parameter. In fact, ordinary least squares (OLS) can be used to correctly fit the Buckingham-Coulomb potential, since the short-ranged Buckingham interactions can be almost exactly captured by short-range correlation functions, and the electrostatic energy is exactly captured by the Ewald summation. Table 4.1 lists the fitted dielectric values and accuracy metrics for the same CE + point electrostatic term using OLS, Ridge regression, and the Lasso.

The results in Figures 4.9 and 4.10 show convincing evidence that including a point electrostatic term in a cluster expansion effectively captures the long-range point electrostatics

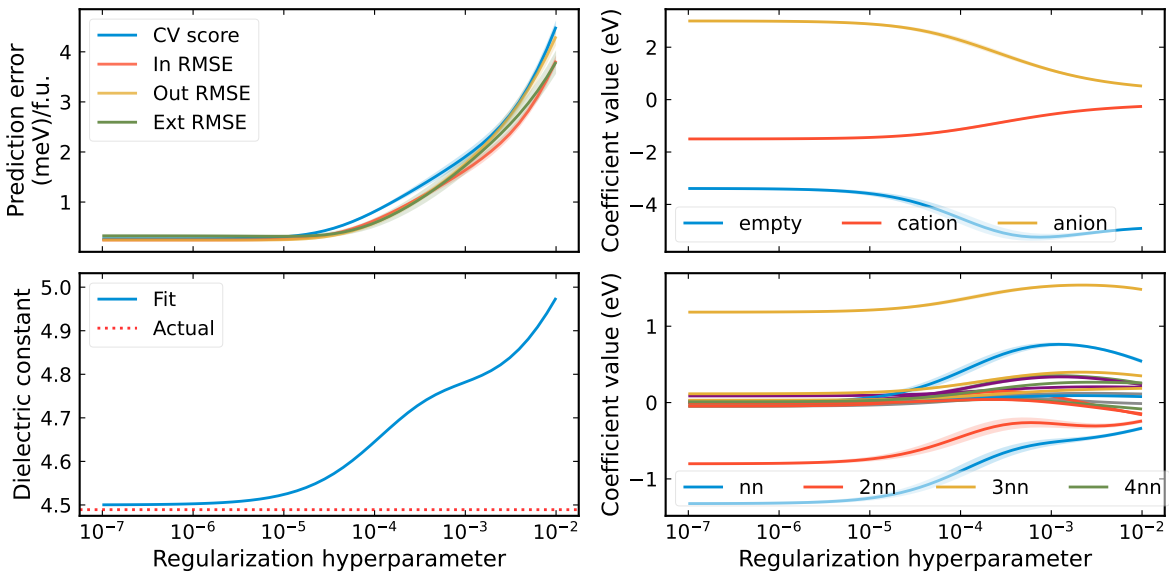


Figure 4.10: Convergence of error, correlation coefficients, and effective dielectric constant with respect to hyperparameter selection for a Ridge regression fit of a Buckingham-Coulomb potential.

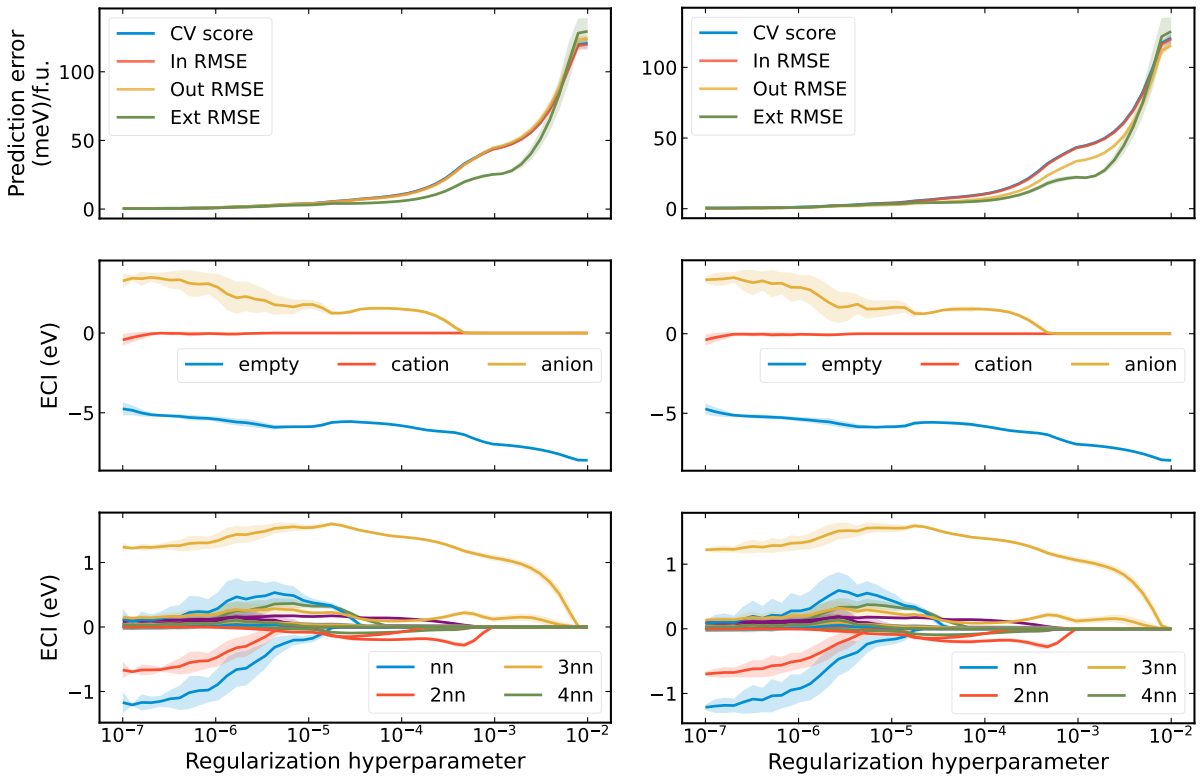
Regression	Fitted dielectric	CV score	In RMSE	Out RMSE	Extrapolation RMSE
OLS	4.494	NA	0.235	0.242	0.365
Lasso	4.543	0.372	0.297	0.347	0.343
Ridge	4.502	0.314	0.263	0.272	0.337

Table 4.1: Fitted dielectric constant and accuracy metrics in meV/f.u. using ordinary least squares (OLS), Ridge and Lasso regression. The exact dielectric value in the model is 4.5.

and allows the CE to represent short-range interactions. While we cannot expect to recover a true dielectric constant when using first-principles calculations of a real material, it can be considered an *effective* dielectric constant for the model. Further, the addition of the point electrostatic term has been shown to substantially improve the stability and performance of a cluster expansion fit using DFT energies, particularly for prediction values of longer period superstructures [181, 196].

Regularization and centering in mixed models

We have shown for the case of electrostatic interactions how using a mixed model that includes a Coulomb interaction pair potential can result in significantly better-converged Hamiltonians with substantially improved prediction accuracy, especially for longer period



(a) Buckingham-Coulomb potentials fit with electrostatic term centering data

(b) Buckingham-Coulomb fit subtracting out exact Coulomb term and centering data

Figure 4.11: Coefficient regularization paths for fits of a Buckingham-Coulomb potential with an electrostatic term, and a Buckingham-Coulomb potential by subtracting the exact coulomb potentials and centering the remaining training data.

superstructures. In the simple example of a Buckingham-Coulomb potential, the fitted mixed model converges to the exact underlying dielectric. This indicates that since the electrostatic interactions are captured exactly, the cluster expansion terms are left to capture the shorter-range Buckingham interactions.

Stronger evidence showing that the Buckingham interactions are captured entirely by cluster expansion terms is shown in Figure 4.11. Figure 4.11 shows the convergence of accuracy metrics and cluster expansion coefficients for (a) the fit of the Buckingham-Coulomb potential using a mixed model with a Coulomb electrostatic term and (b) for the fit of only the Buckingham potential (by subtracting the exact electrostatic interaction term) using only a cluster expansion. The converged values of all coefficients for both cases and their respective selection paths with respect to hyper-parameter match almost exactly for both cases; which implies that the cluster expansion portion of the mixed model ends up capturing the Buckingham interactions only.

The above results, and the effective use of an electrostatic pair term in a system fit to DFT energies, requires that the fitting data be centered appropriately. Centering target data and centering/normalizing features is often standard and recommended practice in regularized linear regression [66, 90]. Centering target data $\boldsymbol{E}$ amounts to subtracting an offset such that the centered target data has mean zero. Similarly, centering features amounts to subtracting their mean. In a like manner, *standardizing* features requires centering them and normalizing them to have unit norm. Standardization is recommended, particularly when using regularization and when features are measured in different units, in order to treat all features equally [90].

Although the cluster correlation functions and the pair potential in a mixed model clearly have different units—correlation functions are *unit-less*, and pair potentials have units of energy—we contend that standardization is not appropriate. Correlation functions are unitless and Fourier correlations are essentially already *standardized*, in fact, they are whitened with respect to the product distribution $\boldsymbol{\rho}$.[11] Furthermore, normalizing the target values of the external pair potential $E_P(\sigma)$ to lie in the same range as that of correlation functions amounts to claiming that the pair potential captures a similar amount of information as that captured by a single correlation function. This in a sense contradicts the very reason motivating the inclusion of a pair potential in the mixed model. Namely, one must keep in mind the *mixed model* concept, and treat the whole cluster expansion and the same grounds as the additional pair potential.

In contrast, centering target energies and features is appropriate, and essentially necessary to allow the pair potential to fully account for its respective interactions and leave the remaining portion of the energy to be captured by the expansion terms in the Hamiltonian representation. To make this explicit, we can re-write the OLS objective problem in terms of the centered energies $\widetilde{\boldsymbol{E}}$, the centered correlation matrix $\widetilde{\boldsymbol{\Pi}}$ and the centered values of the additional pair potential $\widetilde{\boldsymbol{E}}_P$, where each of its elements is given by,

$$\widetilde{\boldsymbol{E}}_i = \boldsymbol{E}_i - \frac{1}{n}\sum_{j=1}^{n} \boldsymbol{E}_j = \boldsymbol{E}_i - \bar{E}\mathbf{1} \tag{4.27}$$

$$\widetilde{\boldsymbol{E}}_{Pi} = \boldsymbol{E}_{Pi} - \frac{1}{n}\sum_{j=1}^{n} \boldsymbol{E}_{Pj} = \boldsymbol{E}_{Pi} - \bar{E}_P\mathbf{1} \tag{4.28}$$

$$\widetilde{\boldsymbol{\Pi}}_{ij} = \boldsymbol{\Pi}_{ij} - \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\Pi}_{ij} = \boldsymbol{\Pi}_{ij} - \bar{\boldsymbol{\Pi}}_j \tag{4.29}$$

where $\bar{E}, \bar{E}_P \in \mathbb{R}$ are the average of the individual energies in the vector $\boldsymbol{E} \in \mathbb{R}^n$ and the evaluated pair potentials in the vector $\boldsymbol{E}_P \in \mathbb{R}^n$ respectively. $\mathbf{1} \in \mathbb{R}^n$ is a vector of all ones. $\bar{\boldsymbol{\Pi}} \in \mathbb{R}^{p-1}$ is a vector where each element is the average of each of the $p-1$ nonconstant correlation function samples that make up the columns of the correlation matrix $\boldsymbol{\Pi}$.

[11] Fourier correlation functions are by construction centered in expectation, i.e. $\mathbb{E}_{\boldsymbol{\rho}}[\Theta_\beta] = 0$; and standardized $\mathbb{E}_{\boldsymbol{\rho}}\left[\Theta_\beta^2\right] = 1$.

We can re-write the OLS objective for centered inputs and data as follows [93],

$$||\boldsymbol{\Pi J} + \lambda \boldsymbol{E}_p + J_0 \mathbf{1} - \boldsymbol{E}||_2^2 = ||\widetilde{\boldsymbol{\Pi}} \boldsymbol{J} + \lambda \widetilde{\boldsymbol{E}}_p - \widetilde{\boldsymbol{E}} + (J_0 + \bar{\boldsymbol{\Pi}}^\top \boldsymbol{J} + \lambda \bar{E}_P - \bar{E})\mathbf{1})||_2^2 \tag{4.30}$$

where the coefficient vector $\boldsymbol{J} \in \mathbb{R}^{p-1}$ only contains coefficients for non-constant correlation functions (the constant J_0 is included separately).

Minimizing the objective on the left-hand side of Equation 4.30 over $\boldsymbol{J}$ can be done by minimizing the least squares of the centered inputs and data,

$$\boldsymbol{J}^*, \lambda^* = \underset{\boldsymbol{J}, \lambda}{\operatorname{argmin}} \, ||\widetilde{\boldsymbol{\Pi}} \boldsymbol{J} + \lambda \widetilde{\boldsymbol{E}}_p - \widetilde{\boldsymbol{E}}||_2^2 \tag{4.31}$$

And the resulting constant coefficient J_0 can be then set to,

$$J_0^* = \bar{E} - \bar{\boldsymbol{\Pi}}^\top \boldsymbol{J}^* - \lambda^* \bar{E}_P \tag{4.32}$$

Apart from the general statistical estimation benefits, solving for mixed model coefficients according to Equations 4.31 and 4.32 has two particular practical benefits. First, including a regularization penalty in terms of $\boldsymbol{J} \in \mathbb{R}^{p-1}$ avoids penalizing both the offset J_0 and the pair potential parameter λ. And setting the offset J_0 according to Equation 4.32, in contrast to setting it equal to the mean of the uncentered data $J_0 = \bar{E}$, allows for the pair potential term to fully capture the interactions it corresponds to, leaving the correlation functions to capture the rest. Furthermore, according to Equation 4.32, J_0 is the mean energy not accounted for by the external pair potential.[12] In other words, J_0 captures the expectation of the energy represented by the Fourier cluster expansion which is congruent with its statistical interpretation given in Chapter 2.3. Precisely, this form of centering is how the correctly converging results shown in Figure 4.11 were obtained.

4.6 Compressed sensing

The regression models we have introduced can be used for learning lattice models using overdetermined (more training structures than correlation functions) and underdetermined (more correlation functions than training structures) linear systems. Many linear regression algorithms have been proposed and benchmarked in literature [2, 148, 153, 227]. However, the formal analysis of solutions and mathematical guarantees on the accuracy of coefficients can be very different and are for the most part carried out for each type of linear system separately [25, 92]. The motivation for using an overdetermined system comes from several studies showing decreasing cross-validation errors with an increasing number of training structures [2, 131]. The case of underdetermined systems is usually motivated as a way to avoid a nearsighted pre-selection of correlation functions. In this section, we will focus solely on reconstructing lattice Hamiltonians from underdetermined linear systems, and often

[12] Shifted by a *sampling bias* $\bar{\boldsymbol{\Pi}}^\top \boldsymbol{J}^*$, which should be much smaller than $\lambda^* \bar{E}_P$.

severely so. Additional considerations for fitting lattice Hamiltonians using overdetermined linear systems are treated in Chapter 5.1.

For relatively small dimensional systems, such as binary or ternary alloys, that allow accurate fits using only a small number of short-range clusters of low degree (number of sites in a cluster) such that the number of total correlation functions is manageable, practitioners have the option to choose between using an overdetermined system or an underdetermined system approach. However, for more complex high-dimensional systems there will often be no such choice. Since the number of expansion terms that need to be considered for an acceptable fit quickly becomes too large, computing DFT energies for enough structures to obtain an overdetermined system becomes untenable. The reason for this is twofold. First of all, the number of terms in a truncated expansion grows polynomially with the number of allowed species at each site. Additionally, complex physical interactions may require longer range and/or higher degree clusters. For example, these situations frequently occur in the study of high entropy materials which has recently gained much attention in the design of metallic alloys [78] and ceramic materials [136]. Although theoretically the representations developed in Chapter 2 are well suited for the study of high entropy materials, the large high dimensional configuration spaces involved render an *overdetermined* system approach prohibitive. For such scenarios, there is no choice but to work with underdetermined systems only and hope that *betting on sparsity* [92] applies favorably to high dimensional cluster expansions for the increasing number of materials being studied that require fits using severely underdetermined systems. This approach can be formally cast and analyzed within the framework of Compressed Sensing (CS) [27, 29].

The focus of CS is to recover a signal or function from a very small set of measurements. The key idea behind CS is that the function sought has a sparse representation in an appropriate basis. Although any given underdetermined linear system has infinitely many solutions, if the function indeed is sparse, the true underlying sparse set of coefficients can be recovered exactly (up to measurement accuracy) under suitable constraints on the samples [27, 53]. In fact, classical CS guarantees that for a suitable set of n measurements the sought coefficients can be recovered exactly using an ℓ_1 minimization problem if n is of order $s \log p$ where p is the number of basis functions measured, and s is the number of nonzero basis function coefficients [27, 53]. The classical CS ℓ_1 minimization problem is expressed as follows,

$$\hat{\boldsymbol{J}} = \underset{\widetilde{\boldsymbol{J}}}{\operatorname{argmin}} ||\widetilde{\boldsymbol{J}}||_1 \quad \text{subject to } ||A\boldsymbol{\Pi}\widetilde{\boldsymbol{J}} - \boldsymbol{E}||_2 \leq \varepsilon \tag{4.33}$$

where the measurement matrix $M = A\boldsymbol{\Pi} \in \mathbb{R}^{n \times p}$, with $n \ll p$ is made up of the values of a relatively small set of p truncated correlation vectors for each of the n training structures.

The notion of classical CS has been previously shown to yield accurate cluster expansions of metallic alloys [152, 153]. Classical CS relies strictly on the concept of *incoherence* [27] in order to guarantee accurate recovery of the underlying coefficients. The need for incoherent measurements—which in the present case correspond to correlation function values evaluated for selected training structures—is clear when the goal is to accurately recover the exact

coefficients in the expansion of a function. This requirement can be made intuitive by thinking of *coherence* as a measurement of how similar correlation function samples (i.e., columns of Π_S) are to each other. High *incoherence* (low *coherence*), means that the set of correlation function samples used are more uniformly spread out in their span (they are closer to mutually orthogonal), and as a result, the portion of the energy represented by each, can be easily distinguished.

In the limit of zero coherence, the correlation matrix is orthogonal, and in principle, the portion of the energy for each correlation function can be identified exactly. For the cases with high coherence, correlation function samples will be much more closely correlated and it is no longer possible to distinguish which correlation function a specific portion of the energy comes from.

Formally, the *coherence* of a measurement matrix M—which for an applied lattice model corresponds to the truncated correlation matrix Π_S—is defined as [53],

$$\mu(M) = \max_{i<j} \frac{|\langle M_i, M_j \rangle|}{||M_i||_2 ||M_j||_2} \tag{4.34}$$

where M_i and M_j are columns of matrix M. The normalized definition given above bounds the coherence of a measurement matrix to values between zero and one, $0 \leq \mu(M) \leq 1$.

The guarantees of classical compressed sensing depend on the coherence of the measurement matrix M being as close to zero as possible in order to maximize the probability of accurate reconstruction of coefficients. As suggested above, incoherent measurements improve the chances of accurate recovery of coefficients, specifically by requiring a lower number of necessary training structures for a given level of accuracy [27]. Furthermore, it is necessary for M to satisfy the *restricted isometry property* (RIP) with a small isometry constant δ [27]. In broad terms, the RIP is a condition that ensures that most of the possible sparse solutions for the linear system lie outside the nullspace of the measurement matrix M [53].

It has been shown that matrices made up of random measurements, such as those composed of Gaussian vectors on the unit hypersphere, satisfy the RIP with overwhelming probability and lead to high levels of incoherence [27]. Random measurements have been previously reported to lead to effective training structure selection and resulting accurate underdetermined fits of cluster expansions for binary metallic alloys [152, 153]. Nevertheless, selecting training structures such that the resultant measurement matrix has high incoherence and ideally satisfies the RIP with a small constant, becomes difficult and in some cases almost impossible for materials systems with more complex physics. This can be generally understood as a result of physically imposed sampling constraints. Usually, the vast majority of all possible configurations will have high energies that are complicated if not impossible to compute with first principles methods. For example, systems with configurations that undergo large structural relaxations can no longer be mapped back to the fixed structure underlying the lattice model. As another example, certain cluster occupations in ionic systems are difficult to access when very high electrostatic repulsion exists. Finally, charge neutrality constraints in heterovalent ionic systems restrict the possible configurations that can be sampled. These

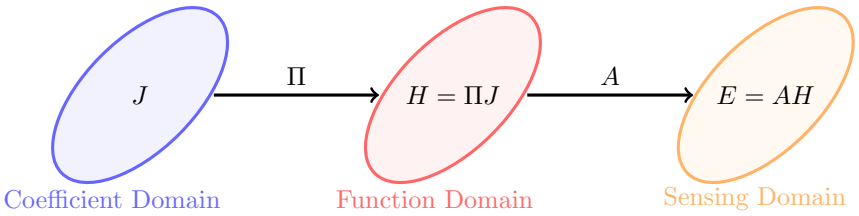


Figure 4.12: Schematic of different domains involved in compressed sensing. Classical CS seeks to approximate the set of *exact* coefficients $\boldsymbol{J}$. CS with redundancy seeks to recover function H in the function domain in the center. Adapted from Candes *et al* [29].

phenomena complicate and in many cases prevent the possibility of obtaining appropriate correlation matrices with minimal coherence required for classical CS recovery, even with previously proposed methods, such as the aforementioned use of uniformly random vectors over the hypersphere.

Compressed sensing with coherency and redundancy

A variant of compressed sensing has been shown to give accurate reconstructions of sparse or compressible signals with only a small set of *coherent* measurements by including redundancy in these measurements [29]. The essence of CS with coherent and redundant measurements is not to recover the model coefficients as accurately as possible, but rather to recover an approximation for the actual function as accurately as possible by way of a redundant representation and possibly highly coherent measurements [29]. For systems constructed using measurements with high coherence and a redundant representation, solutions will always be highly degenerate, but again the focus is on the accuracy of predictions obtained using the recovered expansion and not the exact values of the fitted coefficients. However, we will still seek only sparse solutions such that the large degeneracy of solutions becomes manageable and the resulting expansion can be used efficiently for subsequent predictions and simulations such as in Monte Carlo studies. Figure 4.12 shows a schematic of the pertinent mathematical objects, the corresponding domains, and their relationships as it pertains to CS. For the purposes of CS with coherent and redundant measurements, we seek an accurate representation in the function domain made up of a small number of coefficients but without much regard to whether or not the obtained coefficients correspond to those of any underlying *exact* expansion.

In contrast to classical CS, the goal of CS with redundant and coherent representations is to reconstruct a sparse representation or approximation of a function by solving the following optimization problem.

$$\hat{H} = \underset{\widetilde{H}}{\operatorname{argmin}} ||\boldsymbol{\Pi}^* \widetilde{H}||_1 \quad \text{subject to } ||A\widetilde{H} - \boldsymbol{E}||_2 \leq \varepsilon \tag{4.35}$$

where H is the function sought—in our case a function of atomic configuration. $\mathbf{\Pi}$ is a linear operator mapping coefficients J to the function H as depicted in Figure 4.12. In the case of an applied lattice model, $\mathbf{\Pi}$ is the matrix of all correlation basis functions. A is a sensing matrix used in selecting the training data, and E is the measured value of the function H for the training configurations. The measurements are of the form $E = AH + z$, where z represents an additive noise with some upper bound $||z||_2 \leq \varepsilon$.

The main theorem of CS with coherency and redundancy gives the following error bounds for an s-sparse reconstruction of H,

$$||\hat{H} - H||_2 \leq C_0\varepsilon + C_1 \frac{||\Pi^* H - (\Pi^* H)_s||_1}{\sqrt{s}} \tag{4.36}$$

where $\hat{H}$ is the s-sparse approximation to H. $(\Pi^* H)_s$ denotes a vector with the largest s nonzero coefficients of H and zeros elsewhere. C_0 and C_1 are constants that depend only on the sensing matrix A. The theorem holds if the measurement matrix satisfies a restricted isometry property (D-RIP) adapted to the union of the span of all sets of s columns of Π with isometry constant $\delta_{2d} < 0.08$ [29]. The D-RIP prevents possible solutions from being highly distorted by the measurement matrix, and similarly to the standard RIP, also prevents them from falling in its null space.

The theorem from Equation 4.36 essentially says that the solution to the optimization problem in Equation 4.35 gives accurate reconstructions of the function H, when the coefficients $\Pi^* H$ are sparse and/or decay rapidly—i.e., H is *compressible*. These results suggest that if functions of configuration for multicomponent materials are indeed *compressible*, meaning they can be represented with only a small set of functions, then accurate reconstructions can be obtained with coherent and redundant measurements. Although a rigorous proof along with bounds on the compressibility of functions of configuration—which to the best of our knowledge has not been reported in literature—would be of immense value, it is beyond the scope of the present work. We do however provide a numerical indication that such functions are indeed favorably compressible. This result is not unanticipated considering established knowledge regarding the physics of locality, the success of atomic potentials with small numbers of multiple body terms, and the general tenet of parsimony in physics. This translates to our intuition that such expansions of atomic configuration should have low degree and small associated cluster diameters, and as such, we should be able to represent them using a small number of terms. The crux is finding an optimal subspace spanned by a small set of functions that allows an accurate and sparse representation.

In the context of learning applied lattice models, the generalized Potts frame constitutes a redundant representation that permits successful recovery of lattice Hamiltonians. Since the generalized Potts frame is highly redundant, the motivation is that introducing more expansion functions than strictly necessary and using an appropriate algorithm—such as for solving Equation 4.35—can yield both accurate and sparse representations of functions of crystalline configurations without the need of maximally incoherent measurements. An intuition for this can be formed by picturing the union of all subspaces spanned by size s

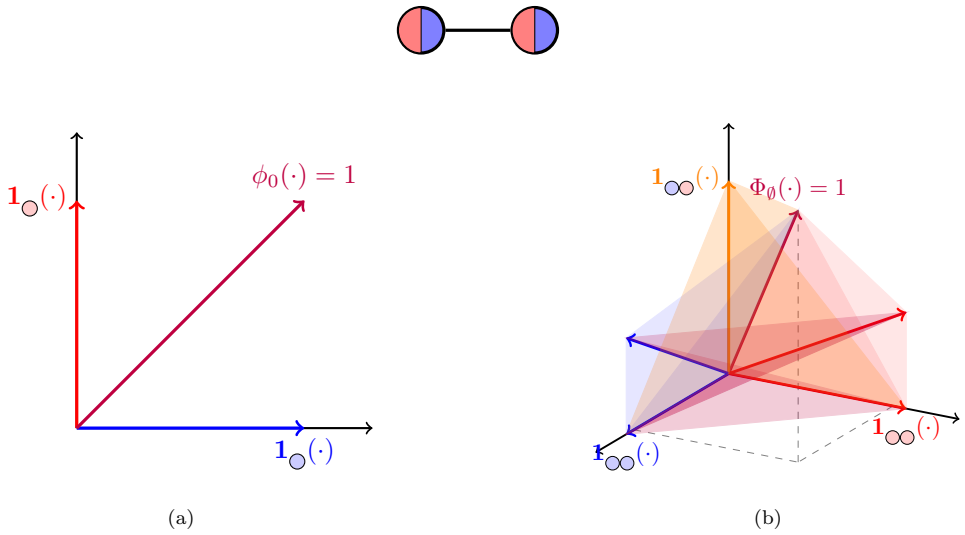


Figure 4.13: Function space representations over the configurations of a single binary site and a symmetric binary diatomic molecule. (a) Function space over a single binary site space. The two different choices for site bases to construct a standard CE are colored red and blue, and each set also includes the purple $\phi_0 \equiv 1$ function. (b) Function space over symmetrically distinct configurations of the molecule. The generalized Potts frame includes all colored functions (blue/red/yellow). All function sets also include the magenta-colored constant function. The D-RIP for a 2-sparse representation, in this case, is adapted to the union of all colored planes in (b).

subsets of functions in the generalized Potts frame. If the function sought lies on any of these subspaces or close enough, then the function can be accurately represented by only s terms.

As an illustration, let us reconsider the example of the binary diatomic molecule. The site function space and resulting symmetrically invariant product space $L^2(\mathbf{\Omega}, \boldsymbol{\rho})^{\mathcal{G}}$ are depicted again in Figure 4.13. Additionally, the subspaces spanned by all possible sets of two cluster indicator correlation functions (2D planes) are highlighted. These 2D planes, constitute $s = 2$-dimensional subspaces. In this case, the function space over configurations of the binary diatomic molecule has three dimensions, however, when a function lies close to or on any of the highlighted planes that function can be well approximated using only two terms. In Chapter 5.3, we show how using the Potts frame under the context of redundant and coherent CS, indeed permits successful recovery of lattice Hamiltonians with underdetermined linear systems—sometimes severely so—that often surpass accuracy and sparsity compared to fits using a basis representation.

Chapter 5

Training data preparation & applications

The regularized regression methodology developed in Chapter 4 requires a series of data preparation and prepossessing steps to ensure that the fitted lattice Hamiltonian is an accurate representation of the material under study. In light of the growing interest in multi-principal element alloy and ionic materials, training data preparation and processing methods have been undergoing continuous reevaluation and new development in order to allow accurate lattice Hamiltonian fits for increasingly large configuration spaces. In this chapter, we present a handful of data preparation and pre-processing methods that result in accurate and sparse fits using the linear regression models presented in Chapter 4. The methodology discussed here is not meant to be exhaustive or conclusive. More so, it represents methodology that can be further optimized, but that we have found particularly effective in dealing with challenges that occur when fitting applied lattice models over high dimensional configuration spaces. Subsequently, we provide various results illustrating relevant methodology using technologically relevant materials under active study. Several results and illustrative applications given in this chapter are based on published work [9, 10, 253, 259].

5.1 Training data generation & preprocessing

The overall process necessary to obtain a converged, sparse, and accurate lattice Hamiltonian for a complex multi-component material usually requires an iterative procedure. Figure 5.1 shows a general workflow diagram of the steps necessary to successfully fit an applied lattice model. Obtaining an adequate feature matrix $\mathbf{\Pi}$ to fit a cluster expansion of a real system, requires a sequence of nuanced preparation steps that are still the subject of active study. Additionally, the choice of the regularization is of critical importance such that recovered expansion coefficients should in principle follow predefined priors, sparsity patterns, and/or hierarchical relations.

In this section, we will first describe the training data preparation and pre-processing

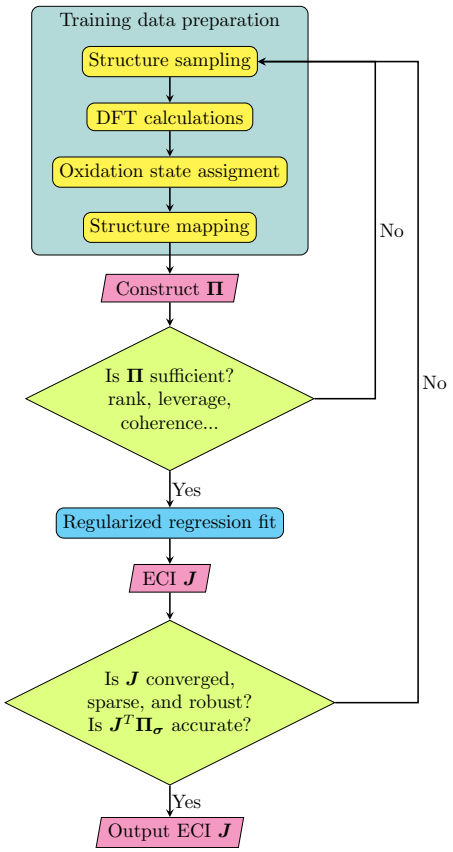


Figure 5.1: General workflow diagram depicting the necessary steps required to generate and prepare training data and successfully fit a converged, sparse, and accurate cluster expansion of a complex ionic material.

necessary to obtain an appropriate training set consisting of a correlation matrix and its corresponding target energy vector $(\boldsymbol{\Pi}, \boldsymbol{E})$. In particular, we will briefly introduce structure sampling methods geared to obtaining well-conditioned feature matrices. We also touch on methodology to effectively assign oxidation states in ionic systems using DFT magnetic moment results. We additionally describe structure matching methodology to account for large structure relaxations that commonly occur in materials that include vacancies and ionic materials systems that may also include oxidation states. We also discuss the effects and offer practical solutions for handling systems with physically *inaccessible* configurations, such as those that undergo substantial relaxation and can no longer be mapped to the underlying disordered structure.

Training structure sampling

The sampling of representative training structures is a critical step to obtain useful lattice Hamiltonians for real materials. Ideally, structure sampling can cover all of the relevant areas of configuration space—areas such that predictions are interpolating rather than extrapolating. However, completely covering all relevant areas of very large configuration spaces is usually not possible.

The vast majority of all possible configurations tend to be concentrated at particular correlation function values [197]. These have been previously named majority structures. And structures far from those correlation values have been named minority structures [197]. The values of orthogonal correlation functions concentrate at the origin for uniformly sampled configurations in systems without any composition constraints. This has been well known in studies of metal alloys [197]. However, this is not the case in ionic systems because charge neutrality constraints reduce the number of allowed points in configuration space as detailed in Chapter 2.1. Instead, correlation functions in ionic systems, or more generally systems with composition constraints, will concentrate on different values based on the particular set of constraints.

Figure 5.2 shows the number of structures in a disordered rocksalt material system for several different correlation function values for a set of charge neutral and a set of unconstrained (including charged structures) uniformly randomly sampled structures. The vast majority of structures are concentrated around particular correlation values. These distributions of correlation function values can differ substantially between the case of unconstrained and constrained configuration spaces. The highly biased distribution of correlation functions in ionic systems, which results in higher coherence/similarity between correlation functions, should be considered when using structure sampling mechanisms that have developed considering unconstrained configuration spaces only [149, 152, 153, 195, 197, 227].

Structure sampling approaches generally depend on the relationship between the number of structures m and the number of correlation functions d that will be used in fitting a lattice Hamiltonian. Based on relationship between m and d (i.e. the shape of the correlation matrix $\mathbf{\Pi}$) the linear system in Equation 4.6 can be categorized as an overdetermined problem ($m > d$) or an underdetermined one ($m < d$). For an ionic material, the full linear system is always underdetermined; but based on the cluster cutoffs used to truncate the expansion, the resulting linear system can be made overdetermined. Structure sampling methods and their mathematical rationalization differ accordingly based on the relationship between m and d.

Most theoretical properties as well as the practical stability of regression depend on the correlation matrix $\mathbf{\Pi}$ being full rank, $\text{rank}(\mathbf{\Pi}) = \min(m, d)$. In other words, the rank is equal to the number of columns for the overdetermined case (when $m > d$), and it is equal to the number of rows in the underdetermined case (when $m < d$). We briefly discuss the two situations and how they pertain to structure sampling to effectively fit applied lattice models. Figure 5.3 shows overview flow diagrams for sampling procedures in underdetermined and overdetermined cases.

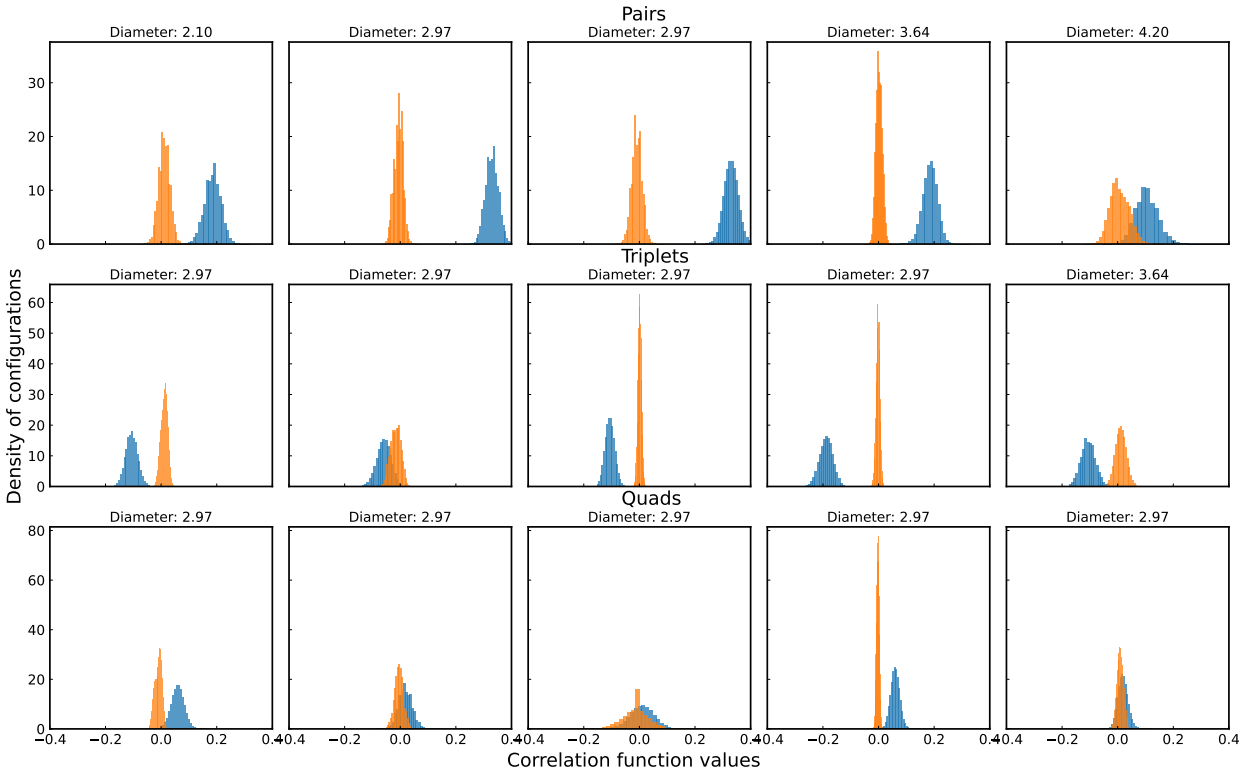


Figure 5.2: Histograms of pair, triplet, and quadruplet correlation function values for uniformly random sampled structures basis correlation values for charge-neutral configurations only and unconstrained (any possible) configurations.

For the overdetermined case, a full rank matrix is one in which the sampled values for each correlation function (i.e feature vectors) are linearly independent. For any finite set of samples there is likely to be a combination of insufficient sampling and in the case of ionic materials intrinsic linear dependencies (those introduced in Chapter 2.1) that contribute to rank deficiencies in $\mathbf{\Pi}$. Rank deficiency can be further aggravated by configurations with energies that are inaccessible to first principle calculations, which we address in more detail in Section 5.1. Further, based on the aforementioned effects of charge neutrality constraints, obtaining a full rank overdetermined feature matrix in ionic systems is technically never possible (unless as previously mentioned correlation functions that give rise to intrinsic linear dependencies are removed). Appropriate sampling should seek to minimize the former effects and improve the overall rank of the correlation matrix.

Sampling in overdetermined linear systems

In overdetermined cases, even though $m > d$, the rank($\mathbf{\Pi}$) can be smaller than d. Under such circumstances, the rank($\mathbf{\Pi}$) can be increased by adding more structures to cover a wider

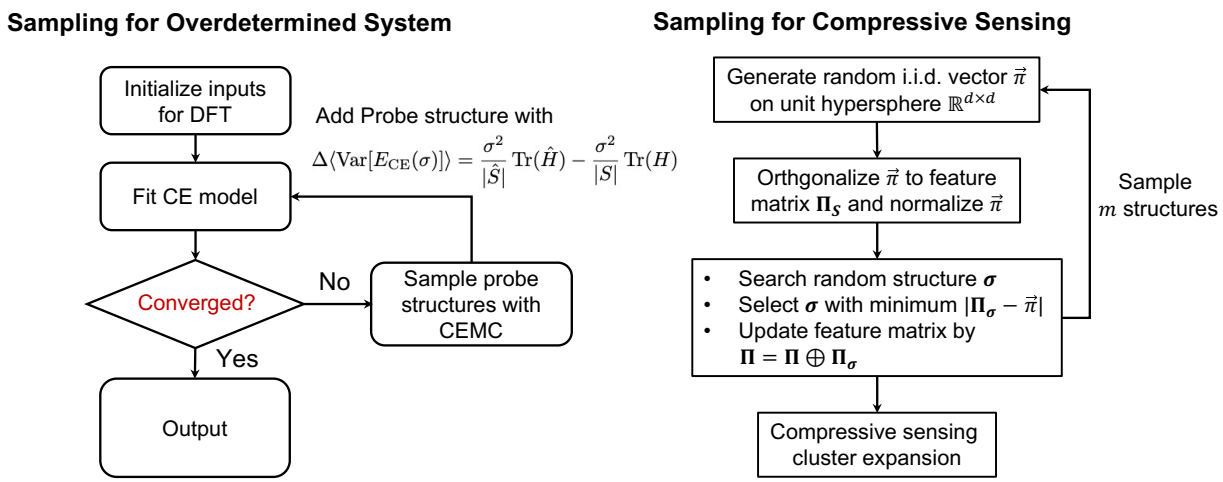


Figure 5.3: (a) Sampling procedure for overdetermined problems, including initialization of inputs for DFT calculations, fit of the lattice Hamiltonian, convergence checks, and addition of probe (additional) structures [195]. The probe structures are selected by maximizing the reduction of leverage score (uncertainty) between the previous set S and the new set $\hat{S}$. (b) Sampling procedure for the compressive sensing cluster expansion. In such a procedure, structures are selected by selecting correlation vectors $\Pi_{\boldsymbol{\sigma}}$ that most closely align with uniformly random vectors over the hyper-sphere $\vec{\pi}$ [153].

range of correlation values, and/or including additional correlation functions that introduce new linear independent features. In simple systems, this can minimize the rank deficiency up to only the trivial linear dependency between point functions from the constraint of charge neutrality, but for more complex high dimensional systems this procedure may not be tenable based on the large number of structures that would be required to appropriately sample the fast-growing charge constrained configuration space.

Nevertheless, overdetermined penalized linear regression, in particular variants of the Lasso (ℓ_1-norm regularization), with rank deficient matrices still yield valid solutions with which useful cluster expansions can be constructed. As explained previously, the solutions will be degenerate (i.e., certain linear transformations of the estimated coefficients will represent the exact same cluster expansion) [93, 221], but this degeneracy is not by itself a practical point of concern. Instead, the focus of structure sampling should be on improving the predictions and variances for a fitted lattice Hamiltonian for any acceptable estimates of expansion coefficients.

To simplify our analysis of prediction variance, we assume that a fitted lattice Hamiltonian is fitted with an overdetermined, full rank correlation matrix and captures the real target energy as follows,

$$\boldsymbol{E}(\boldsymbol{\sigma}) = \mathbf{\Pi}_{\boldsymbol{\sigma}} J + \boldsymbol{\varepsilon} \tag{5.1}$$

where $\boldsymbol{E}(\boldsymbol{\sigma})$ is the real energy, and $\boldsymbol{\varepsilon}$ is a random error with *heteroskedastic* uncorrelated

variances, $\text{cov}(\boldsymbol{\varepsilon}) = s^2 I$.

Under the assumptions above, the variance of the predicted energy by a cluster expansion fitted with least squares regression can be expressed as [148, 195, 227],

$$\text{Var}[E_{\text{CE}}(\boldsymbol{\sigma})] = s^2 \boldsymbol{\Pi}_{\boldsymbol{\sigma}}^{\top} (\boldsymbol{\Pi}^{\top} \boldsymbol{\Pi})^{-1} \boldsymbol{\Pi}_{\boldsymbol{\sigma}} \tag{5.2}$$

where s^2 represents the variance from intrinsic noise in the DFT calculations for a given population of structures, and $\boldsymbol{\Pi}_{\boldsymbol{\sigma}}$ is the truncated correlation vector for the particular occupancy $\boldsymbol{\sigma}$ used in prediction. The expression above can be adjusted for penalized regression models under a Bayesian interpretation [148]. However, for the purpose of our current explanation, Equation 5.2 is sufficient.

According to Equation 5.2, the average variance for predicted energies is given as,

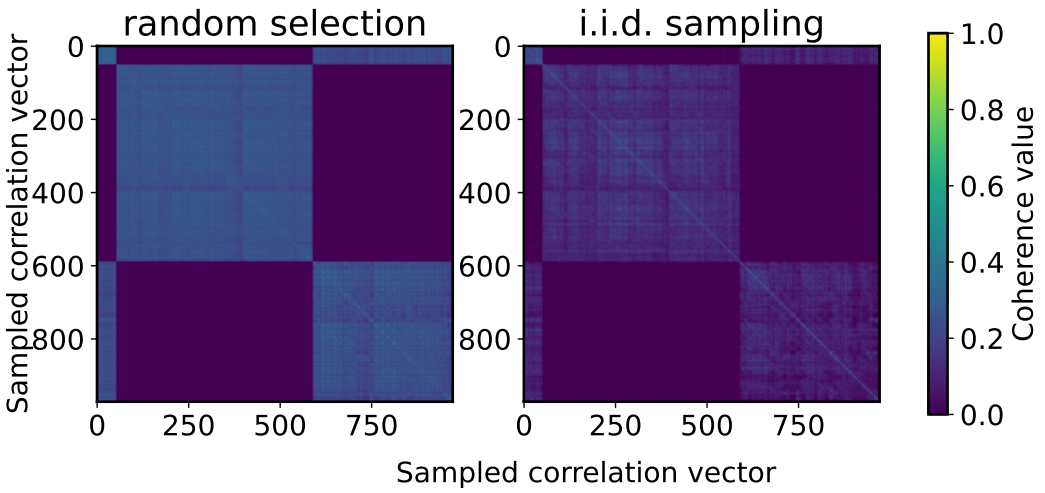
$$\begin{aligned} \langle \text{Var}[E_{\text{CE}}(\boldsymbol{\sigma})] \rangle &= \frac{\sigma^2}{|S|} \sum_{\boldsymbol{\sigma} \in S} \boldsymbol{\Pi}_{\boldsymbol{\sigma}}^{\top} (\boldsymbol{\Pi}^{\top} \boldsymbol{\Pi})^{-1} \boldsymbol{\Pi}_{\boldsymbol{\sigma}} \\ &= \frac{\sigma^2}{|S|} \text{trace}(H) \end{aligned} \tag{5.3}$$

where S is the number of training structures. $H = \boldsymbol{\Pi}^{\top} (\boldsymbol{\Pi}^{\top} \boldsymbol{\Pi})^{-1} \boldsymbol{\Pi}$ is the so-called hat matrix [66], and its diagonal elements H_{ii} are the predicted variances for a particular structure; which are also known in the statistics literature as *leverage scores*. The leverage score ranks the uncertainty of the corresponding probe occupancy $\boldsymbol{\sigma}$ into high-leverage or low-leverage points according to regression diagnostics [59]. A handful of methods for structure sampling have been proposed that seek to minimize the average leverage score, or equivalently maximize the reduction in average predicted variance, for each additional structure included [148, 195, 227]. These methods can lead to improved robustness and accuracy in cluster expansion fits of complex multicomponent materials.

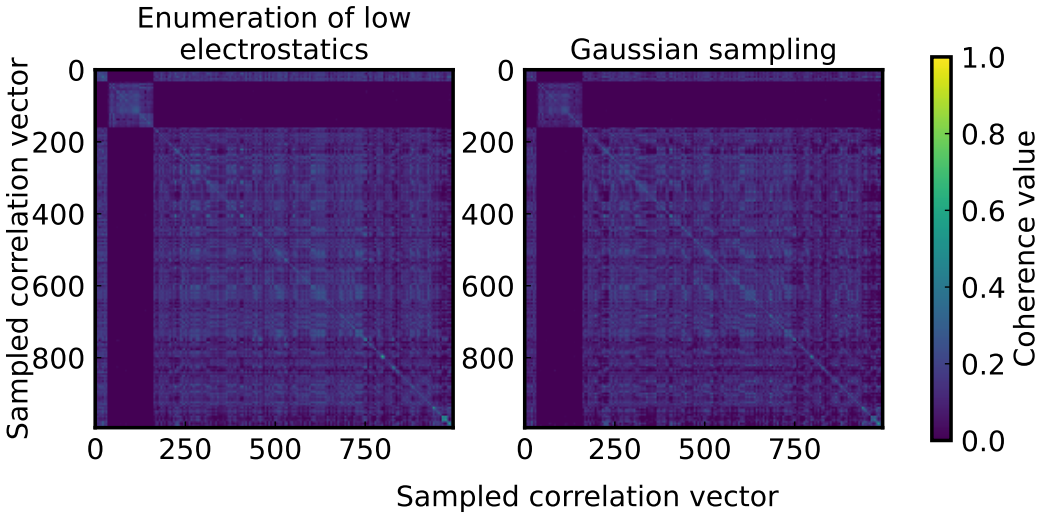
Sampling in underdetermined linear systems

For the underdetermined linear regression case ($m < d$), obtaining a full rank correlation matrix is much more straightforward. An underdetermined system has full rank when all correlation vectors (rows of $\boldsymbol{\Pi}$ are linearly independent), as opposed to linearly independent correlation functions. In such a case, maximizing the $\text{rank}(\boldsymbol{\Pi}) \leq m$ instead requires obtaining m structures with linearly independent correlation vectors.

Since there are more unknowns than samples, sampling and regression for an underdetermined linear system are suitably addressed within the framework of Compressive Sensing (CS). As described in Chapter 4.6, a CS approach to cluster expansions can result in accurate and sparse solutions of coefficients using a relatively small amount of DFT measurements compared to the number of correlation functions ($m \ll d$) [9, 153]. However, the necessary structure sampling for classical CS that maximizes the probability of accurate coefficient recovery has strict requirements based on the coherence—a measure of the degree of similarity—among the sampled correlation functions [27, 152].



(a) Ni-Co-Cr



(b) $LiMnO_2$-Li_2TiO_3-LiF

Figure 5.4: Gram matrices (coherence) for randomly sampled structures and Gaussian sampled orthogonal correlation vectors for a ternary alloy system and an ionic rocksalt system.

Sampling methods resulting in correlation matrices appropriate for CS have been proposed in the context of the cluster expansion method for metallic alloys. Specifically, correlation matrices appropriate for CS can be obtained by sampling correlation vectors that are random, independent, and identically distributed (i.i.d.) over the unit hyper-sphere [152, 153]. Figure 5.4a shows two normalized Gram matrices $G = \mathbf{\Pi}^\top \mathbf{\Pi}$ for the correlation functions of a Ni-Co-Cr ternary alloy structures for a case of random sampling from a large set of enumerated structures and for a case of i.i.d. random sampling over the hypersphere. The elements G_{ij} of Gram matrices are the dot product of sampled correlation function values i and j, which measure the level of coherence between correlation functions i and j. High coherence or similarity between sampled correlation functions is visualized as the off-diagonal yellow pixels. From left to right, a clear reduction in the overall coherence—pixel colors closer to the blue end of the spectrum—is visible when using the i.i.d. sampling proposed for classical CS cluster expansion fits [152, 153].

However, charge neutrality constraints and strong electrostatic interactions complicate such random sampling in ionic systems. As an illustration of this, Figure 5.4b shows the two Gram matrices for a $LiMnO_2$-Li_2TiO_3-LiF rocksalt system. The left-hand matrix corresponds to low electrostatic energy enumeration for cells up to 64 atoms, and the right-hand matrix corresponds to structures with correlations as close as possible to i.i.d. random vectors unit hyper-sphere. From left to right, although a slight decrease of the coherence values between sampled correlation vectors is successfully obtained, the maximal coherence, which is often taken as the coherence value for the full matrix, remains unchanged; and the coherence is likely too high to reliably use CS recovery of expansion coefficients. The comparison indicates that generating structures to obtain correlation matrices that approximate i.i.d. random matrices, may not be an effective way to minimize the coherence for classical CS. Nonetheless, as described in Chapter 4.6 the over-complete nature of the correlation basis can be leveraged under a newer variant of CS that relies on redundant expansion terms[29]. This form of CS with redundancy can be used to fit sparse and accurate CEs even with highly coherent sampling [9].

Sampling for group-wise structured sparsity

Though it is hard to obtain a full-rank feature matrix for overdetermined systems or a low-coherency matrix for compressive sensing in an overdetermined system, it is still feasible to obtain accurate and well-converged lattice Hamiltonians by also relying on the appropriate use of the previously described group-wise and hierarchically constrained structured sparsity regularization. By classifying correlation functions into groups based on orbits B as described in Chapter 4.3, the correlation matrix can be analyzed in terms of *orbit sub-matrices*. An *orbit sub-matrix* of a given correlation matrix is made up of all the column vectors that correspond to the same orbit B of site space clusters. Figure 5.5 shows a schematic illustration of a correlation matrix and its orbit sub-matrices. For such regularization, structure sampling should strive to keep the orbit sub-matrices of the training correlation matrix $\mathbf{\Pi}$ full rank or as close to full rank as possible. Without full rank (or near full rank) orbit sub-matrices

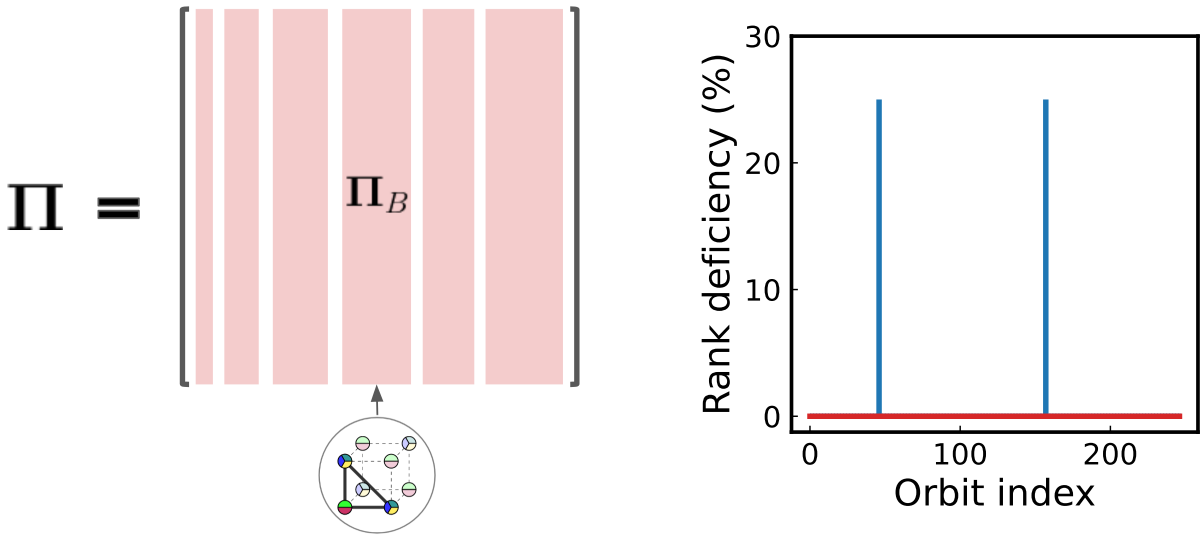


Figure 5.5: Illustration of orbit sub-matrices making up a correlation matrix. Orbit sub-matrices correspond to all correlation functions that act over the same set of symmetrically equivalent clusters, as depicted by the schematic triplet cluster below. Orbit submatrix rank deficiency for a set of sampled correlation vectors for a template rocksalt system

grouped regularized regression may result in poorly conditioned problems and non-unique solutions. In cases where this is unavoidable, group-level and within-group regularized regression, such as using the Sparse Group Lasso or Ridged Group Lasso, can be used to help avoid degenerate solutions [201, 202, 253]. Figure 5.5 also shows the orbit rank degeneracy (defined as one minus the ratio between the sub-matrix orbit rank and the total number of correlation functions in the orbit) for a set of structures of a $LiMnO_2$-Li_2TiO_3-LiF disordered rocksalt material system. In this example, only 3/248 orbits show a small amount of rank deficiency ($\leq 25\%$), which is sufficient to obtain accurate fits with grouped regularization as detailed in Chapter 4.3.

Oxidation State Assignment

In ionic materials containing hetero-valent transition metals, it is necessary to assign formal valence to ions, since the same ion can behave differently when it has a different valence. For instance, according to crystal field theory, valence electron *d*-filling of the transition metal-oxygen states is one factor controlling whether a transition metal ion prefers tetrahedral or octahedral coordination. Furthermore, size and charge effects can cause metal ions to have different kinds of short-range order [108]. This thermodynamic preference arising from different formal valence necessitates treating ions with hetero-valent oxidation states as different species.

However, in determining the formal valence of an ion, the DFT charge density on a metal

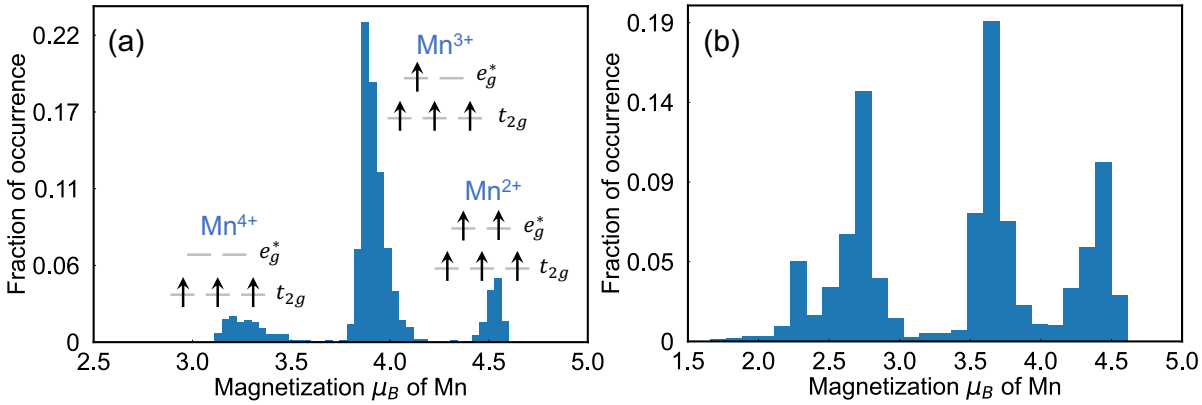


Figure 5.6: (a) The magnetization distribution of Mn calculated with GGA+U in the system of $Li_{1.2}Mn_{0.6}Nb_{0.2}O_{2.0}$. The valence of each Mn atom is determined by the on-site Bohr magnetization μ_B. From the histogram, we can manually estimate the boundary for $Mn^{4+/3+}$ and $Mn^{3+/2+}$ classification to be $3.6\mu_B$ and $4.2\mu_B$. (b) The magnetization distribution of Mn is calculated with the SCAN density functional in the system of Li-Mn-O-F, and a more continuous distribution is observed. The boundary for $Mn^{4+/3+}$ and $Mn^{3+/2+}$ classification is $3.22\mu_B$ and $4.08\mu_B$, determined by Bayesian optimization via Gaussian Processes.

cannot be directly used as it is invariant to the valence state due to hybridization with the anion [32]. Instead, we can rely on the magnetic moment for a given metal site to assign a formal charge and can either use the sum of s, p, and d local orbital contributions or the individual d orbital contribution to assign this charge state. This local contribution can be obtained by integrating the spin up minus spin down magnetic moment around each atom.

Figure 5.6(a) presents a histogram of the magnetic moments on the ions in structures with composition $Li_{1.2}Mn_{0.6}Nb_{0.2}O_{2.0}$, by taking the sum of s, p, and d orbital contributions. In this example, the values $\approx 3.6\ \mu_B$ (differentiates Mn^{4+} from Mn^{3+}) and $\approx 4.2\ \mu_B$ (Mn^{3+} from Mn^{2+}), is enough separation in the magnetic moments to clearly delineate oxidation states.

In other cases, the separation of oxidation states is not as obvious. For example, the histogram of Mn d-orbital magnetic moment in the Li-Mn-O-F system [253] is shown in Figure 5.6(b). It is not straightforward to define cut-off values to classify the different Mn oxidation states. In this case, one can use black-box optimization approaches (such as Bayesian optimization via Gaussian Processes [205]) to assign oxidation states that are optimally consistent with a maximal number of charge-neutral structures.

More specifically, the loss function for Bayesian oxidation state assignment can be formulated as the sum of the absolute value of each structure's charge, taken over all structures in a DFT computed dataset. The loss function depends on a black box function f, which is the mapping function between any local magnetic moment for a metal to its formal valence.

Table 5.1: Magnetic moments for Mn in three configurations of $Li_7Mn_7O_{12}F_2$ calculated with DFT-SCAN [211], and sorted into their oxidation states as determined by Bayesian optimization. The d orbital magnetic moments and energy above hull (eV/atom) are listed.

configuration	Mn^{2+}	Mn^{3+}	Mn^{4+}	Energy above hull (eV/atom)
A	4.207, 4.26, 4.31	3.602, 3.629, 4.017	2.916	0.133
B	4.169, 4.208, 4.264,	3.615, 3.65, 3.982	3.217	0.137
C	4.169, 4.278, 4.33, 4.366	4.07	2.703, 2.974	0.157

The exact form of f is neither known nor differentiable, but it depends solely on the magnetic moment upper cutoff for each different metal species of interest. For the dataset used in Figure 5.6(b) the function is $f(c_1, c_2, c_3)$ where c_1, c_2, and c_3 are three upper magnetic moment cutoffs for Mn^{2+}, Mn^{3+} and Mn^{4+}. After black-box optimization, the upper cutoffs, corresponding to a minimal loss of structures with non-zero total charge for a given DFT dataset, can be used to assign the formal valence for any structure.

Table 5.1 additionally shows three configurations of $Li_7Mn_7O_{12}F_2$, with oxidation states assigned using the recently published Bayesian optimized solution [253]. The cutoffs are 3.228 μ_B (differentiating Mn^{4+} from Mn^{3+}) and 4.0815 μ_B (differentiating Mn^{3+} from Mn^{2+}).

Configurations A and B both have three Mn^{2+}, three Mn^{3+}, and one Mn^{4+}. It is less straightforward to determine where the Mn^{3+} and Mn^{2+} cutoff lies for configuration A because 4.017 μ_B is closer to the magnetic moments assigned to Mn^{2+} atoms (4.207 μ_B, 4.26 μ_B, 4.31 μ_B) than to the moments assigned to Mn^{3+} atoms (3.602 μ_B, 3.629 μ_B). Using Bayesian optimization circumvents this complication.

Interestingly, within configuration B the magnetic moments are more clearly separated, as the ranges of magnetic moments for Mn^{2+} and Mn^{3+} are notably less than that for configuration A, but this is not associated with a lower energy since configuration B is 4 meV/atom higher in energy. Configuration C has an entirely different set of charge orderings (four Mn^{2+}, one Mn^{3+}, and two Mn^{4+}) which can be recognized and assigned by the algorithm.

This optimization approach to assign charge states was successfully used in other chemical systems, including Li-$Mn^{2+/3+/4+}$-Ti-O [40], and Li-$V^{4+/5+}$-O [104], further supporting how Bayesian optimization can find non-trivial solutions for charge state assignments onto magnetic moments and increase the efficiency of using DFT-calculated configurations to train ionic CE.

Structure Mapping

In practice, DFT calculations performed to obtain a set of training structures for fitting an applied lattice model involve calculations for structures that have different supercell sizes and shapes. In many available packages [231], initial structures of the *ab initio* calculations are generated from the cluster expansion, the occupancy strings are obtained from the cluster expansion-generated initial structures, and the energies (or other properties) are obtained

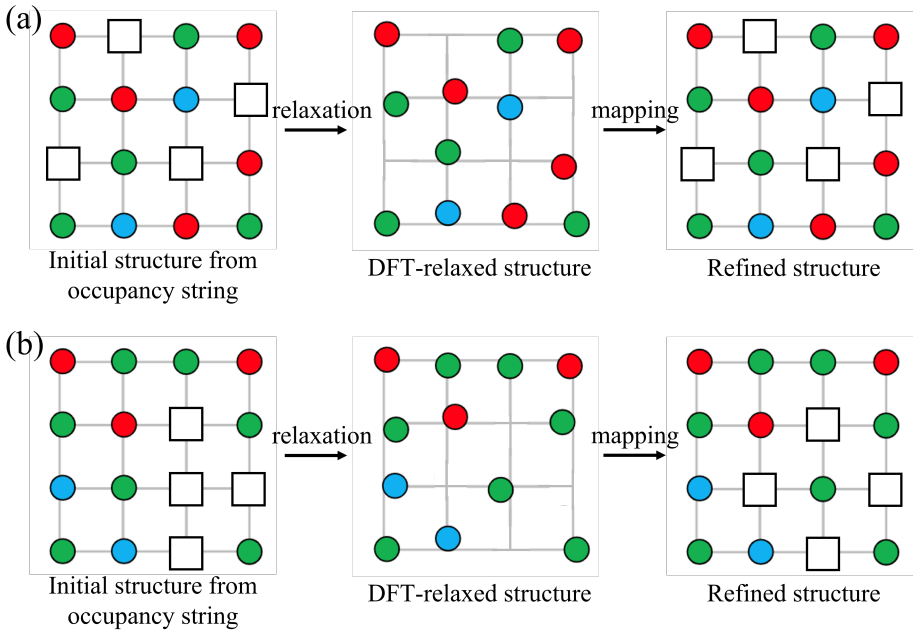


Figure 5.7: Schematics of an input structure corresponding to an occupancy string $\boldsymbol{\sigma}$, the resulting relaxed (DFT-calculated) structure and a *refined* structure. The refined structure is represented by the sites of the relaxed structure mapped to the locations of the sites of the rigid disordered structure underlying the CE. The different colors represent multiple species on the lattice. The empty boxes are explicit representations of vacancies (which in the lattice model are treated as a species). (a) An example case where the refined structure effectively maps back to the initial structure and occupancy string. (b) An example case where the refined structure does correspond to the initial structure or occupancy string due to substantial relaxation.

from the relaxation of the ionic and electronic structure. However, doing so requires that the relaxed structure still corresponds to the occupancy string from which the initial unrelaxed structure was obtained. In many cases encountered in ionic systems, ions relax too far away from their initial site, such that re-assigning them to sites corresponding to the unrelaxed initial structure is infeasible. This is especially noticeable in systems containing vacancies, which allow atoms to relax towards the vacant lattice site. This is also common in structures with large electrostatic or repulsive interactions because the strong interactions often force ions to maximize the distance between the interacting ions.

In cases where the structure relaxation is significant, converting the relaxed structure itself (after *ab initio* relaxation) to an occupancy string is a more appropriate way to capture the configurational energy landscape. A practical implementation of this requires a mapping between sites of the underlying disordered crystal structure and the training structure that

has been relaxed by first-principles calculations. We call the ordered structure that has been appropriately mapped to the rigid lattice the *refined* structure. This mapping can then be used to construct the corresponding configuration strings $\boldsymbol{\sigma}$ for the relaxed structures. A schematic illustration of the relationship between the initial, relaxed, and refined structures is shown in Figure 5.7.

Formally, the procedure of structure mapping for purposes of fitting an applied lattice model can be stated as follows. We represent the *disordered structure* (that represents the domain of the Hamiltonian) using a set of lattice vectors $L_U = \left[\vec{l}_1\ \vec{l}_2\ \vec{l}_3\right]$ and a set of fractional coordinates $P_U = \{\vec{p}_i, \ldots, \vec{p}_{N_U} \mid \vec{p}_i \in [0,1]^3\}$ for N_U sites. To each site, we assign a site space Ω_i. Similarly, for a given ordered structure, we label the set of fractional coordinates P_Q for N_Q sites and refer to the corresponding set of lattice vectors as L_Q. Each site in the ordered structure is occupied by a specific species σ_i. A supercell of the disordered primitive cell must be obtained to enable a one-to-one mapping between the sites of an ordered structure and the sites of the corresponding supercell of the disordered structure. We write the lattice vectors of this supercell L_{UQ}. The number of atomic sites, or equivalently the set of fractional coordinates P_{UQ} of the disordered supercell must be the same as in the ordered structure $|P_{UQ}| = |P_Q|$. Having obtained the disordered supercell L_{PQ}, a map between the sites P_Q of an ordered structure and the sites P_{UQ} of the appropriate disordered supercell is represented by the following bijection,

$$A : P_Q \to P_{UQ} \text{ s.t. } \sigma_i \in \Omega_{A(i)}\ \forall i \in \{1, \ldots, N_Q\} \tag{5.4}$$

The map A can be practically established within reasonable tolerances for structural deformations of the lattice L_Q. In practice, performing these two steps (finding the disordered structure supercell, and finding the map between sites of the ordered structure and the disordered supercell) requires a crystallographic structure matching algorithm, such as the `StructureMatcher` in the `pymatgen` library [161]. A handful of other effective algorithms for crystallographic matching are freely available [94, 209, 218].

However, most approaches treat the inputs of allowed tolerances for all sites on equivalent grounds. For many ionic systems, and in particular, those including vacancies, cations tend to undergo larger displacement than anions during DFT relaxation. Usually, the anion sublattice undergoes less distortion, and as a result, can be more easily mapped with the predefined primitive cell. This practical observation can be revealed by comparing the drift force in DFT outputs for cations and anions, respectively. As a result structure mapping methods may fail for many ordered structures that may still have well-defined structure mappings A. One method for correcting this during mapping involves first performing a search over varying lattices to map the relaxed anions to the fixed anion sublattice sites within a fractional tolerance. Subsequently, cation centers within anion polyhedra (based on the relaxed anion to anion lattice site mapping) can be used to map the cation sublattice sites [253].

Effective structure mapping methods allow practical calculations of the minimum or relaxed energy landscape in terms of atomic configuration. However, it is well known—and

has been numerically quantified—that the extent of structural relaxations affects the number of correlation functions required to obtain a robust and well-converged cluster expansion [155]. Rigorous quantification of strain and a corresponding metric for structure mapping may prove very useful to further establish a formal understanding of the effects of structural relaxations when fitting lattice Hamiltonians of configuration. The majority of available crystallographic matching algorithms lack a rigorous quantification of the strains and symmetry breaking involved. This has only been recently addressed in a newly proposed matching algorithm [218], where cost functions for lattice strain and atomic displacement are constructed for scale-invariant geometric distortions and symmetric breaking distortions.

Physically Inaccessible Configurations

When fitting a lattice Hamiltonian of a complex material, there will usually exist configurations that cannot be reached due to convergence issues in DFT calculations. Inaccessibility can arise in metallic alloys and ionic materials. For the most part, we will focus our attention on issues arising in ionic materials. There are two main categories of configurations that can be inaccessible to DFT: geometrical inaccessibility and charge-valence inaccessibility.

Geometrical inaccessibility occurs when the DFT-relaxed structures drift far from their original lattice sites and cannot be correctly mapped. Although Section 5.1 addresses some ways to find mappings when the cations relax substantially, large anion drift can make the mapping impossible. Consider, for example, anion drift that destroys the FCC anion framework of a rock-salt. Although the initial configuration may have been in the rock-salt configuration space, the resulting relaxed structure no longer is. This becomes a very notable problem when considering configurations with a large number of vacancies.

Charge-valence inaccessibility happens when the DFT-relaxed configuration can be appropriately mapped back to a lattice model with oxidation-assigned ionic species; however, charge transfer prevents specific oxidation states for particular configurations of the predefined lattice model. This happens mostly in transition metal oxides when the valence of the transition metal cannot be well assigned and results in non-charge-balanced configurations. This can also be the result of internal charge transfer in configurations with very high electrostatic energy.

The efficiency of structure sampling is thus reduced depending on how many physically inaccessible states occur in the sampled training configurations. For example, as shown in Figure 5.8, the blue sites in the cluster figures are occupied by high-valent transition metal (such as Nb^{5+}, Mo^{6+}), which have strong repulsion in a single tetrahedron. Such features cannot be appropriately computed by DFT calculations. The effect on sampling is most clear when using an indicator basis since this will result in a void correlation function in the feature matrix $\mathbf{\Pi}$. The void correlation function manifests itself as a column with all elements equal to zero. This happens since no information has been obtained for those particular configurations, such that this correlation function is rendered uninformative and should be removed prior to fitting. For lattice Hamiltonians with orthogonal correlation functions, the effect manifests itself more subtly. In the orthogonal case, inaccessible states

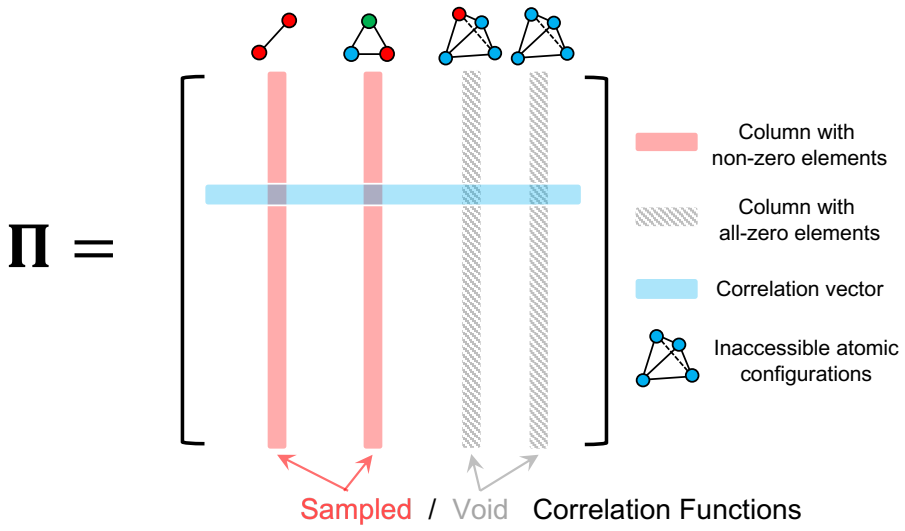


Figure 5.8: Illustration of feature matrix $\mathbf{\Pi}$ with inaccessible (non-sampled) configurations using an indicator basis. The red columns represent the correlation functions that are covered by DFT calculations, while the gray (shaded) columns represent the inaccessible atomic configurations. (e.g., the blue sites are occupied by high-valent transition metals such as Nb^{5+}, Mo^{6+}, which have strong repulsion in one tetrahedron and cannot be well evaluated via DFT. And the blue row represents the correlation vector of one specific structure.

will manifest as linear dependencies or equivalently rank deficiency of the corresponding orbit sub-matrix.

Such inaccessibility can further induce configuration sampling problems in Monte Carlo simulations. This occurs because the lattice Hamiltonian, fitted as described above, has no information regarding the coefficients associated with the inaccessible high-energy configurations. Consider the case in which a configuration with one or more inaccessible features lies close in configuration space to a low-energy configuration (i.e., a few MC steps away). The configuration with inaccessible features may be accepted since its energy will be incorrectly predicted. The end result is that unfavorable configurations can be incorrectly sampled in MC and will distort ensemble statistics and computed thermodynamic properties.

To resolve this issue, one should include as many configurations to reduce the number of under-sampled correlation functions. However, since inaccessible states are in principle caused by DFT instability, under-sampled correlation functions may remain. We suggest two approaches that are useful to deal with the remaining inaccessible sampling issues. First, the coefficients can be regularized with more importance given to those corresponding to lower degree clusters (such as pair-wise interactions). This can be achieved by using hierarchy constraints or group-wise regularization as detailed in Chapter 4.3. These fitting strategies are effective when the configuration energy can be well depicted by correlations of clusters

with small support, therefore void or under-sampled correlation functions for clusters with larger support will contribute minimally to the total energy.

If the resulting lattice Hamiltonian still under-predicts the energy of configurations that are likely to be high energy, rejection of these configurations can be easily achieved in MC. The rejection can be done by including a cluster indicator function of the orbit β associated with such inaccessible atomic configurations. The probability evaluated in Monte Carlo that guarantees the rejection of inaccessible configurations is,

$$p \propto \exp\left(-\frac{1}{k_B T}(E_{\text{CE}} + \sum_{\beta \in \text{void}} M \cdot \mathbf{1}_\beta)\right), \tag{5.5}$$

where E_{CE} is the predicted energy evaluated with actual coefficients, M is a large positive number, and $\mathbf{1}_\beta$ is the indicating function of orbit β. Since the cluster indicator function will only be nonzero when the specific inaccessible cluster configuration is present, all other configurations that do not include such configuration will not be affected. However, this approach requires practitioners to explicitly detect the inaccessible configurations in the first place.

5.2 Structured-sparsity fits of $LiMnO_2$-Li_2TiO_3-LiF ceramics

Cluster expansion fits of the $LiMnO_2$-Li_2TiO_3-LiF (LMTOF) disordered rocksalt system were computed using standard Lasso and the structured sparsity-based regression models previously introduced. The LMTOF rocksalt system comprises a binary face-centered cubic (FCC) anion lattice with O^{2-} and F^{-} disorder, and a FCC cation lattice with Li^{+}, Mn^{3+} and Ti^{4+} disorder. All fits include an explicit electrostatic term as expressed in Equation 4.25 which is computed using the Ewald summation method. We compare the resulting model prediction accuracy, sparsity, and ECI structure of the various fits using a training set of DFT calculated energies for 983 structures with supercells up to 72 atoms. An additional test set of 247 structures of supercell sizes 128 and 132 atoms is used for validation. Additional details of the DFT training structure calculations and fitting are reported in Appendix C.2.

Figure 5.9 shows prediction accuracy metrics for fits using each regression model with three different sets of cluster size cutoffs for pair, triplet, and quadruplet clusters respectively. Hyper-parameter tuning curves for the various regression models are given in Appendix C.2.

From the results in Figure 5.9 we see that all regression models yield similar levels of predictive accuracy. However, although all regression models achieve some degree of feature selection, Sparse Group Lasso and $\ell_2\ell_0$ regression are the most effective in reducing the total number of features required to achieve similar levels of accuracy. We make note specifically that Overlap Group Lasso has the worst performance in feature selection due to the restrictive hierarchical constraints imposed, as described in Chapter 4.3. However, this structure-

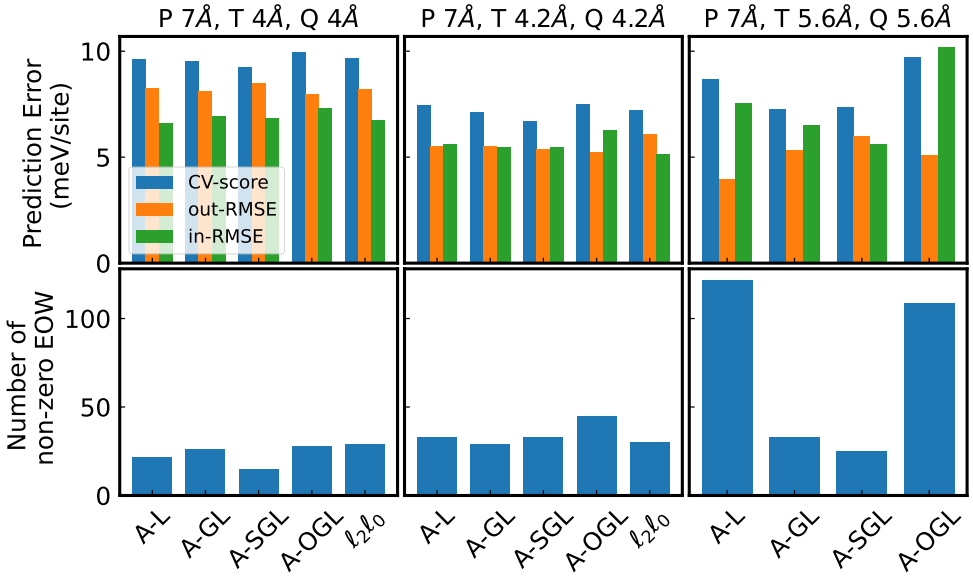


Figure 5.9: Fitted LMTOF CE accuracy metrics and resulting model sparsity using Lasso and structured sparsity-based regression algorithms. (A-) adaptive variants, (L) Lasso, (GL) Group Lasso, (SGL) Sparse Group Lasso, (OGL) Overlap Group Lasso. All fits shown were done using correlation functions for cluster size cutoffs for pair (P), triplet (T), and quadruplet (Q) clusters listed above the figures using a primitive cell of the rocksalt structure with lattice parameter $a = 3$ Å.

sparsity form yields solutions that are competitively accurate and better aligned with physical heuristics.

Figure 5.10 shows the resulting sets of fitted ECI $\boldsymbol{J}$ and effective cluster weights $\mathrm{W}[H_B]$ for each regression model. There are some notable observations and trends regarding Figure 5.10. Structured sparsity models on average result in lower magnitude coefficients compared to the Lasso. Furthermore, although the solutions obtained with these regression models are not unique since the ionic configuration space has charge neutrality constraints, the different models tend to identify a few apparently important correlation functions, in particular short-range pair correlations and some larger diameter triplet correlations. Lastly, hierarchy-based regularization, and in particular the orbit level hierarchy implemented with the Overlap Group Lasso, results in coefficients that much better align with physical intuition and heuristics (i.e., decay with physical distance and cluster size) and the principles of a statistically *well-formulated model* [172].

All in all, the results from the fitted expansions for the LMTOF system shown in Figures 5.9 and 5.10 and the accompanying results in Appendix C.2 demonstrate how expansions with structured sparsity have similar or improved levels of accuracy as those from the Lasso, and additionally tend to have higher sparsity and trends in the resulting coefficients that

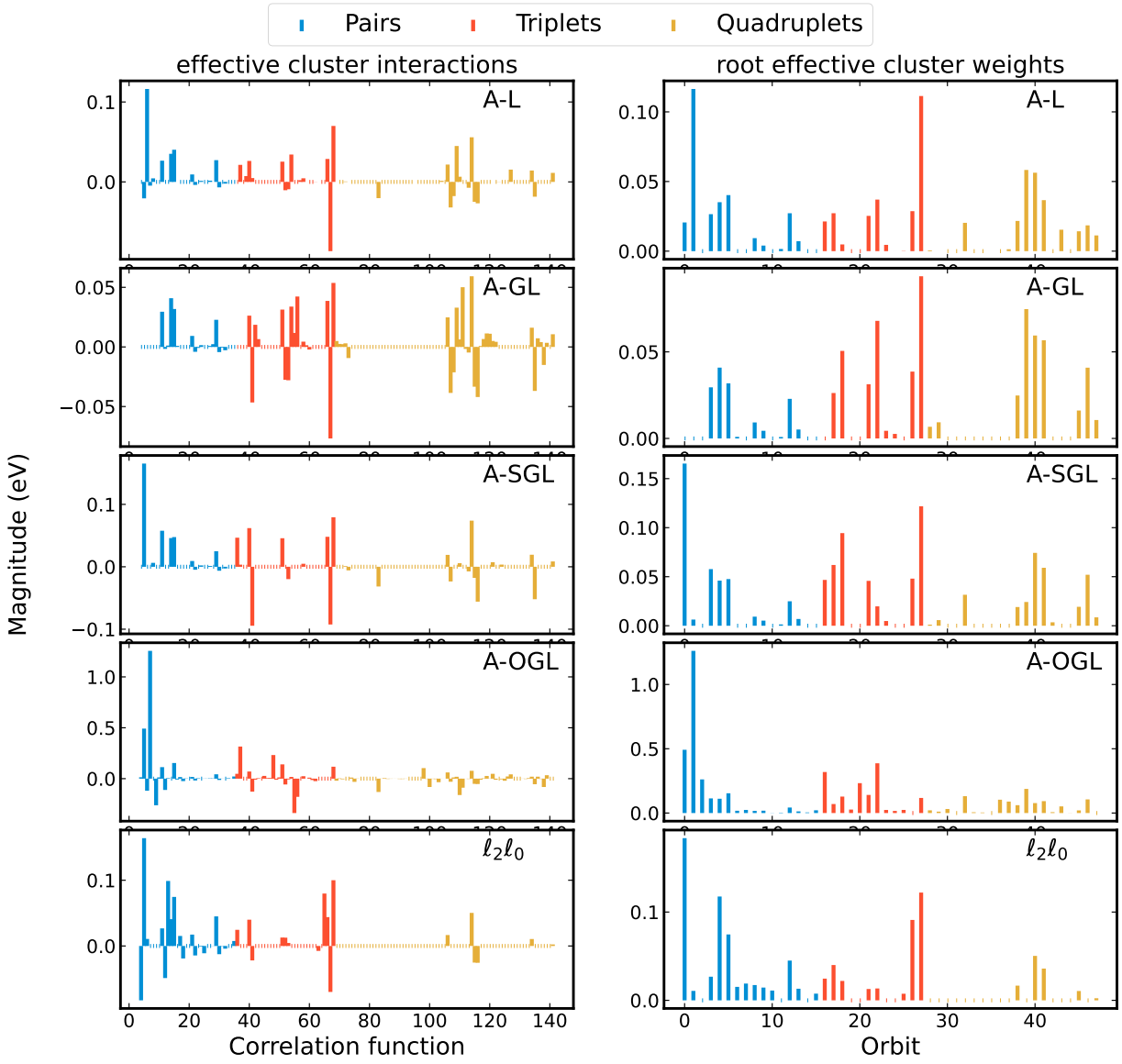
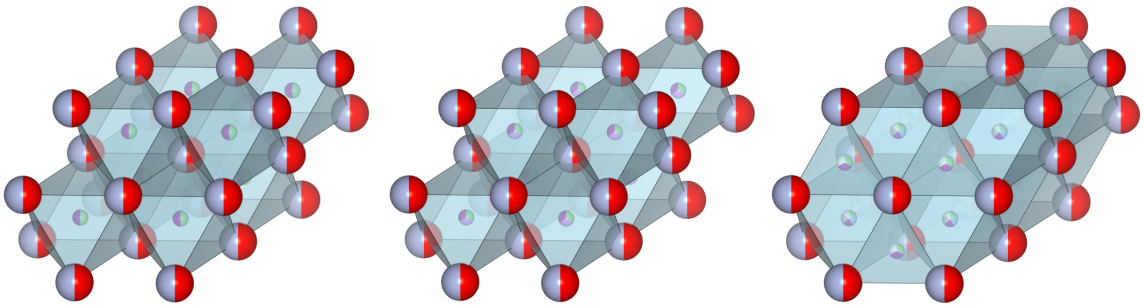


Figure 5.10: Fitted LMTOF effective cluster interactions and square root effective cluster weights using adaptive Lasso and structured sparsity-based regression algorithms. (A-) adaptive variants, (L) Lasso, (GL) Group Lasso, (SGL) Sparse Group Lasso, (OGL) Overlap Group Lasso. All fits shown were done using correlation functions for cluster size cutoffs of 7 Å, 4.2 Å, and 4.2 Å for pair, triplet, and quadruplet clusters respectively using a primitive cell of the rocksalt structure with lattice parameter $a = 3$ Å.



(a) 2-2 LiMnOF binary-binary rocksalt

(b) 3-2 LiMnTiOF ternary-binary rocksalt

(c) 5-3-2 LiMnOF quinary-ternary-binary spinel-like

Figure 5.11: (a) System 2-2: Li-Mn-O-F rocksalt system with binary (Li+/Mn3+) cation sites and binary (O2-/F-) anion sites. (b) System 3-2: Li-Ti-Mn-O-F rocksalt system with ternary (Li+/Mn3+/Ti4+) cation sites and binary (O2-/F-) anion sites. (c) System 5-3-2: Li-Mn-O-F spinel-like system with quinary (Li/Mn2+/Mn3+/Mn4+/vacancy) octahedral cation sites, ternary (Li/Mn2+/vacancy) tetrahedral cation sites, binary (O2-/F-) anion sites.

much better align with physical priors and heuristics.

5.3 Compressed sensing fits of Li-(M_1M_2)-OF ceramics

We demonstrate the performance of the generalized Potts frame by comparing fits for three fluorinated lithium-transition metal oxide systems with fits obtained using a site indicator-based cluster expansion (with site basis functions given in Equation 2.62), and a cluster expansion using orthogonal sinusoid site basis functions [230]. The configuration spaces for the materials considered increase in both size and complexity (larger number of allowed species and number of symmetrically distinct sites) as shown in Figure 5.11.

The total number of expansion terms considered for each fit is listed as the model size in Table 5.2. The total number of terms is obtained by using the same cutoff radius for clusters with up to four sites. Functions that evaluate to the same value (remain constant) for the training structures used in the corresponding fit are subsequently removed from the final measurement matrix used for each fit. Particularly, removal of constant functions was only required for the 5-3-2 system, for which the total number of columns in the measurement matrix ended up being 4194 and 17350 for the indicator basis correlation basis based and Potts frame models respectively. Additionally, we include an electrostatic energy term [181, 196] as an additional feature in every fit. Finally, for the fits using the Potts frame, we

System	Expansion Type	Training Set Size	Test Set Size	Model Size
2-2	Correlation basis	112	337	121
	Potts frame	112	337	520
3-2	Correlation basis	195	456	312
	Potts frame	195	456	1040
5-3-2	Correlation basis	312	56	7030*
	Potts frame	312	56	23070*

Table 5.2: Regression model and training/test data size specifications for the three fluorinated lithium-transition metal oxide systems. *Removal of correlation functions that remained constant for structures in the training set reduced the number of columns in the measurement matrices in 5-3-2 system to 4194 and 17350 for the indicator basis and Potts frame models respectively.

remove one cluster indicator function from each set associated with the same orbit; this is simply to do away with the trivial linear relation in Equation 2.67 applied to cluster concentrations. Doing so has a minimal effect in reducing redundancy, however, we found this slightly improved efficiency for obtaining full rank measurement matrices. It is clear by construction that for the same spatial cutoffs, the number of terms in the generalized Potts frame far exceeds the number of terms in a standard cluster expansion, and this difference grows exponentially with the configuration space complexity.

A total of 50 different fits for each system are computed by selecting a random set of training structures that gives a full rank underdetermined system. The remaining structures are used as a test set. The Lasso solutions were obtained for each fit in a two-step process. First, a 10-fold hyperparameter cross-validation optimization search is done with the training set. Subsequently, a finer hyperparameter search is done centered at the previously obtained value now optimizing for out-of-sample error with respect to the test set but training only with the training set data. From the resulting fits, those with sparsity (number of nonzero coefficients) above the third quartile of the set are considered outliers and removed from the results.

Values for accuracy metrics and sparsity results for the fitted expansion obtained from the set of fits for each of the three expansion types and for each of the three materials systems are shown in boxplots in Figure 5.12. The average prediction accuracy metrics given include cross-validation root mean squared error (CV RMSE) for the initial cross-validation hyperparameter search, the out-of-sample RMSE for the final fit (test structures only), and the full data RMSE for both the training and test structures combined. The average sparsity value for the resulting fits is also listed.

Figure 5.12 shows boxplots depicting the resulting fit statistics in terms of cross-validation, out-of-sample and full sample (both training and testing structures) root mean squared errors, along with the corresponding model sparsity values. The results depicted in Figure 5.12

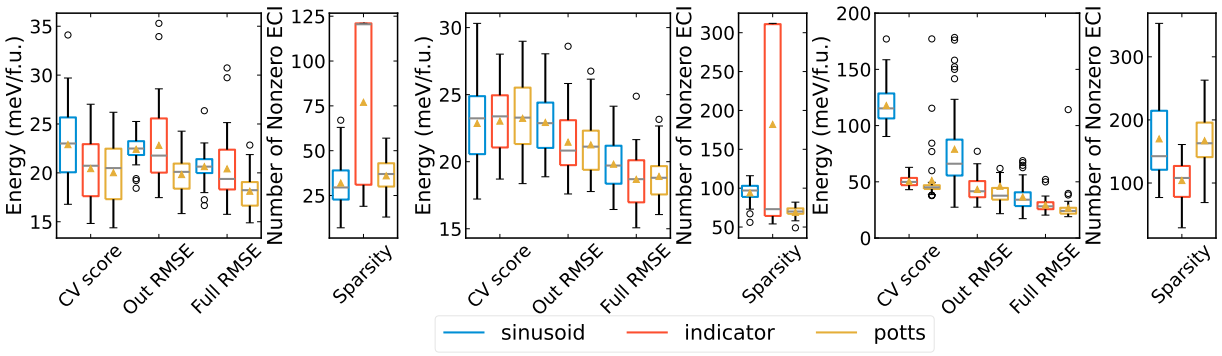


Figure 5.12: Fit metric statistics for the systems tested using standard correlation basis with sinusoid site basis, indicator site basis, and generalized Potts frame. The plotted metrics include cross validation RMSE (CV score), out-of-sample RMSE (Out RMSE), full data RMSE (Full RMSE) for both the training and test structures combined, and the number of nonzero ECI in the fits (sparsity). LiMnOF binary-binary with two sites per formula unit (top), LiMnTiOF ternary-binary with two sites per formula unit (middle), LiMnOF quinary-ternary-binary with four sites per formula unit (bottom).

show that for all systems, the expansions resulting from the Potts frame tend to have either a lower minimum, median, and mean accuracy metrics, a lower spread in these accuracy metrics, or both. In all cases, the accuracy metrics are either very competitive (very close to standard cluster expansions), or better. As for the resulting model sparsity, although considerably more terms are included in the Potts frame fit, the resulting expansions have similar sparsity along with lower spread in sparsity values. These results demonstrate the applicability of using coherency and redundancy to obtain expansions that match and even exceed the accuracy and sparsity of those obtained with standard cluster expansion basis sets.

The boxplots in Figure 5.12 provide a summary of the average values obtained for each type of fit, however, they do not allow us to see which models have both high accuracy and high sparsity (low number of terms). Tables 5.3 and 5.4 list the same fit metrics as Figure 5.12 but for the single sparsest (smallest number of nonzero coefficients) fit and the fit resulting in the highest accuracy in terms of the full data RMSE for each materials system.

From Table 5.3, we see that for the sparsest expansions obtained, the ones fitted using the Potts frame result in the lowest error metrics with only a few exceptions where the Potts frame fits are still of comparable accuracy. Furthermore, the expansions based on the Potts frame result in the lowest or second lowest sparsity for all three systems. The results for the most accurate models in terms of full dataset RMSE in Table 5.4 show that the Potts frame results in both sparse and accurate models. For the cases where one of the correlation basis-based fits results in a lower full RMSE, the sparsity of that model is compromised and substantially worse than the corresponding Potts frame fit. This behavior can also be

System	Expansion	CV RMSE	Out RMSE	Full RMSE	Sparsity
	Sinusoid	34.10	25.26	26.36	7
2-2	Indicator	23.16	19.37	18.79	19
	Potts Frame	23.82	22.31	21.5	13
	Sinusoid	24.13	27.62	24.14	56
3-2	Indicator	25.14	21.69	19.55	54
	Potts Frame	22.26	23.43	20.93	49
	Sinusoid	131.76	158.16	68.80	77
5-3-2	Indicator	45.52	49.99	52.12	29
	Potts Frame	44.58	40.51	39.65	69

Table 5.3: Fitted model accuracy metrics and sparsity of sparsest models. Cross-validation RMSE (CV RMSE), out of sample RMSE (out RMSE), and full dataset RMSE (full RMSE) in meV per formula unit (random structure primitive cell).

System	Expansion	CV RMSE	Out RMSE	Full RMSE	Sparsity
	Sinusoid	25.68	18.42	16.66	50
2-2	Indicator	18.63	17.87	15.76	121
	Potts Frame	21.42	16.11	14.89	43
	Sinusoid	23.10	18.88	16.44	105
3-2	Indicator	28.02	17.89	15.07	310
	Potts Frame	23.09	17.77	16.06	72
	Sinusoid	104.39	36.72	17.24	291
5-3-2	Indicator	55.69	27.45	20.38	147
	Potts Frame	115.42	24.52	18.98	192

Table 5.4: Fitted model accuracy metrics and sparsity of most accurate models in terms of the root mean squared error on the whole dataset. Cross-validation RMSE (CV RMSE), out of sample RMSE (out RMSE), and full dataset RMSE (full RMSE) in formula unit (random structure primitive cell).

observed in the results in Figure 5.12, where the box plot for sparsity obtained with the Potts frame has a smaller interquartile range than that of the correlation basis-based fits (with the exception of the 5-3-2 system where it is only slightly larger than that of the indicator based cluster expansion).

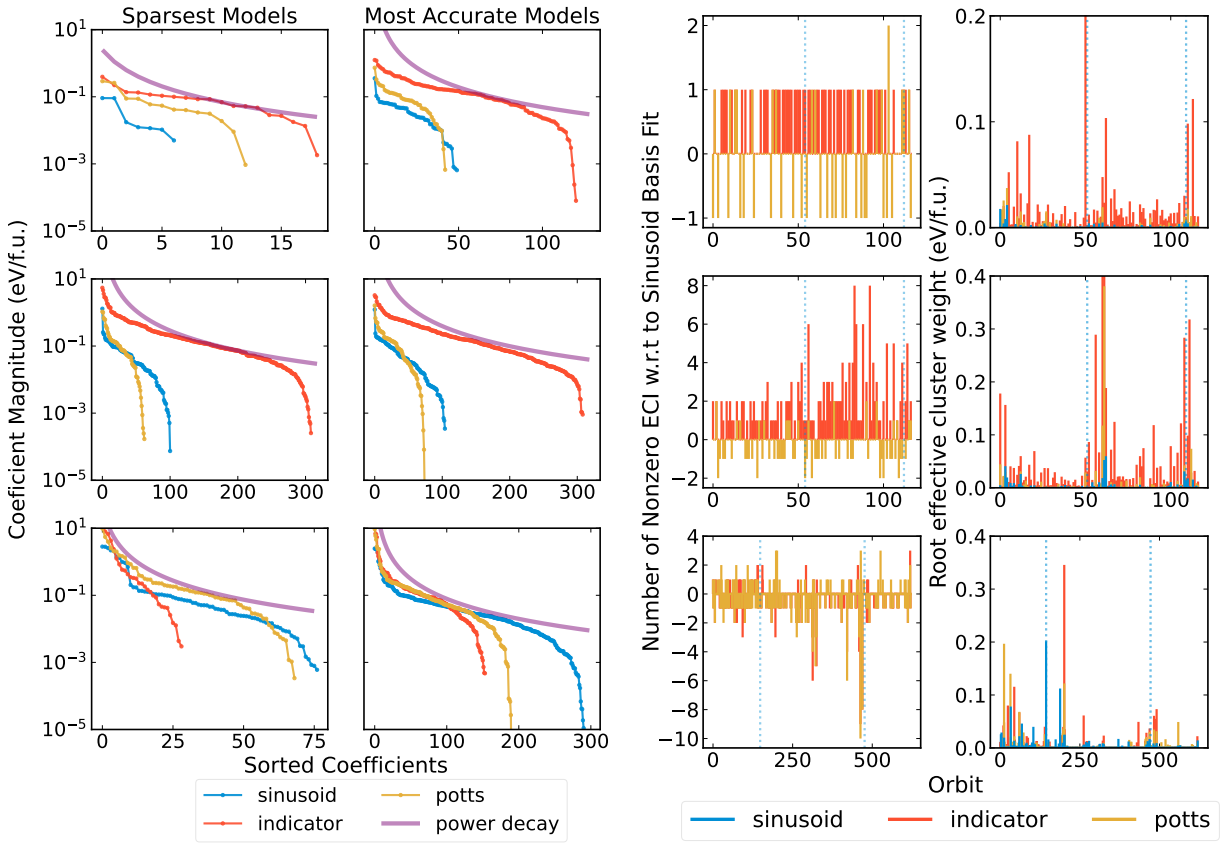
Figure 5.13a shows the sorted magnitudes of the fitted coefficients (magnitude of the ECI times the multiplicity of the correlation function) for each expansion type and materials system for both the sparsest models and most accurate models obtained. Based on the adequacy of the resulting fit error metrics and the fast decay of coefficients shown in Figure 5.13a, we can conclude that the configuration energy if not exactly sparse is highly compressible—since signals or functions with power-law decaying (or faster) coefficients can be well approximated by a small subset of terms [53]. Specifically, a series of coefficients obey a power law decay if the sorted sequence satisfies the following,

$$|c_i| \leq Ci^{-q} \tag{5.6}$$

where c_i are the coefficients and $C, q > 0$ are constants. For all expansion types, we see that the coefficient magnitude decay is faster than the power law decay shown.

We also see that in the results in Figure 5.13a, the coefficient decay of fits using the Potts frame are, at worst, the second fastest decaying series for both the sparsest and most accurate fits, such that expansions using the Potts frame are arguably more reliable in yielding sparser models than standard cluster expansions. This seems surprising considering that the total number of terms in the underdetermined system is exceedingly larger than that of the standard cluster expansions. However, considering the geometry of the union of s-dimensional subspaces as illustrated in Figure 4.13 in Chapter 4.6, we posit that indeed functions of configurational energy lie close to one of these subspaces with high probability, and we can therefore accurately represent the function with a much smaller number of terms than the total considered in the underdetermined system. Additionally, in light of the Theorem in Equation 4.36, the rapid decay of coefficients suggests that an s-sparse set of coefficients is close to the real coefficients and as a result, the second term in the function approximation error is likely to be very small.

We can take the previous results as ex post facto evidence that configuration energy is a compressible function and that CS with redundancy and coherent measurements works well for fitting expansions of configuration energy. Additionally, we observed that expansions fitted using the Potts frame also tend to follow expectations driven by physical considerations. Figure 5.13b shows the number of nonzero coefficients for each crystallographic orbit considered for the most accurate indicator basis and Potts frame-based models with respect to the number of nonzero coefficients obtained in the most accurate sinusoid basis-based model. We notice from the plots, that using the Potts frame, despite having a much larger number of total coefficients associated with each orbit, results in fits that set a similar number of nonzero coefficients within each orbit and never exceed three additional coefficients per orbit when compared with the sinusoid correlation basis fit. The significance is that, not only do we recover accurate and sparse models, but the models themselves also have a similar



(a) (b)

Figure 5.13: (a) Sorted fitted coefficient magnitudes (multiplicity times ECI) for the sparsest and most accurate model (Full RMSE). (b) Number of nonzero coefficients relative to the sinusoid basis fit for each orbit, and norm of coefficients for each orbit for the most accurate models (Full RMSE). The vertical dotted lines separate the degree of the orbit (pairs/triplets/quadruplets). In both (a) and (b) LiMnOF binary-binary (top), LiMnTiOF ternary-binary (middle), LiMnOF quinary-ternary-binary (bottom).

sparsity structure to the correlation basis-based models. Figure 5.13b also shows the root sum of squares or norm of the fitted ECI for each orbit. We observed that for the most accurate models the fitted coefficient weight associated with each orbit is less erratic than that of the indicator-based cluster expansion. Additionally, the fitted coefficients tend to follow the heuristic of coefficient decay with orbit size more fittingly than the indicator-based cluster expansion. The aforementioned observations again demonstrate using the Potts frame can result in fitted expansions that are not only accurate and sparse but also produce *well-behaved* coefficients that align with practices and heuristics based on physical insights used in the field.

5.4 Structured-sparsity fits of NiCoCr alloys

As an application demonstrating how the uniqueness of a cluster decomposition can be leveraged with invariant regression algorithms developed in Chapter 4 to obtain expansions with improved accuracy and sparsity, we have fitted Fourier cluster expansion Hamiltonians for a NiCoCr alloy system. Expansion fits were carried out using the Lasso [220], and $\ell_2\ell_0$ estimation with a correlation function hierarchy prior [259] and a cluster interaction hierarchy as detailed in Chapter 4.3. Fits were carried out using two sets of cutoffs: a set including only pairs up to 8 Å(501 training structures), and a set with pairs up to 10 Åand triplets up to 6 Å(502 training structures). Training structure sampling was done by random sampling over a unit-hypersphere as detailed in Chapter 5.1 [152]. A holdout test set of 1000 structures was used for model validation. Additional details on DFT calculations and fits are reported in Appendix C.3.

Estimator	CV RMSE (meV/site)	Out RMSE (meV/site)	Cluster Interaction Sparsity
Lasso	17.72	32.35	11
$\ell_2\ell_0$	17.94	32.11	11
Grouped-$\ell_2\ell_0$	17.78	32.22	8

Table 5.5: Cross validation root mean squared error (CV RMSE), out of sample root mean squared error (Out RMSE) and number of nonzero cluster interactions (Cluster Interaction Sparsity) for Fourier cluster expansion fits of a NiCoCr with pairs up to 8 Åwhich amounted to a total of 11 pair cluster interactions consisting of 33 pair correlation functions.

All models resulted in basically the same cross-validation root mean squared error (18 meV/site) and out-of-sample root mean squared error (30 meV/site) regardless of the regression algorithm used. The resulting model cross-validation root mean squared error, out-of-sample error, and orbit sparsity (number of orbits with nonzero coefficients) are listed in Tables 5.5 and 5.6. The accuracy metrics obtained are consistent with those reported previously for a NiCoCr cluster expansion [171]. Beyond the comparable prediction accuracy, a substantial improvement in sparsity is obtained by using hierarchically structured sparsity

Estimator	CV RMSE (meV/site)	Out RMSE (meV/site)	Cluster Interaction Sparsity
Lasso	18.85	31.38	36
$\ell_2\ell_0$	19.36	31.34	26
Grouped-$\ell_2\ell_0$	18.43	31.38	5

Table 5.6: Cross validation root mean squared error (CV RMSE), out of sample root mean squared error (Out RMSE) and number of nonzero cluster interactions (Cluster Interaction Sparsity) for Fourier cluster expansion fits of a NiCoCr with pairs up to 10 Åand triplets up to 6 Åwhich amounted to a total of 37 pair and triplet cluster interactions consisting of 180 pair and triplet correlation functions.

priors. Overall the best fit is obtained using the cluster interaction hierarchical prior (i.e. a well-formulated mode with strong hierarchy) which obtains the best sparsity for both the expansion with pairs only and the expansion with pairs and triplets.

Additionally, the resulting parameter structure shows some level of selection similarity for each of the regression models used. But in particular, for the expansion using pair and triplet correlation functions, the resulting structure satisfying strong, weak, and no hierarchical constraints has some noticeable feature selection differences. The models with weak and strong hierarchy lead to visibly improved selection of the interactions with the largest cluster sensitivity (i.e. the most expressive features). Figure 5.14 shows the resulting expansion parameters (effective cluster interactions) and corresponding effective cluster weights. Considering the comparable accuracy obtained with all regression models, the model with cluster interaction hierarchy (grouped $\ell_2\ell_0$), that satisfies the strong hierarchy prior results in the sparsest model and only selects cluster interactions with the largest cluster sensitivities. These highly informative cluster interactions are captured by all regression models, but both Lasso and $\ell_2\ell_0$ with correlation hierarchy (weak hierarchy) end up selecting substantially more features that make little difference in improving predictive accuracy.

The two fits (pairs only and pairs + triplets) using cluster interaction hierarchical constraints result in a quite competitive prediction accuracy and sparsity. However, the resulting thermodynamic behavior can still be noticeably different. In the present example, it indeed happens to result in slightly different thermodynamic behavior. Figure 5.15 shows the nearest neighbor and second nearest neighbor cluster probabilities, as well as the cluster energies for temperatures between 100 to 1000 K using Wang-Landau density of states sampling [237]. In addition, the fitted nearest neighbor and second nearest neighbor interaction values for both expansions are shown.

The results show that the atomic ordering behavior with temperature exhibits broadly similar trends but noticeable differences in the finer details. Both expansions exhibit a phase transition near 600 K, however, the expansion including triplets also exhibits a transition at a lower temperature of around 200 K. The nearest neighbor probabilities show somewhat similar trends as well, but the second nearest neighbor probabilities are noticeably distinct.

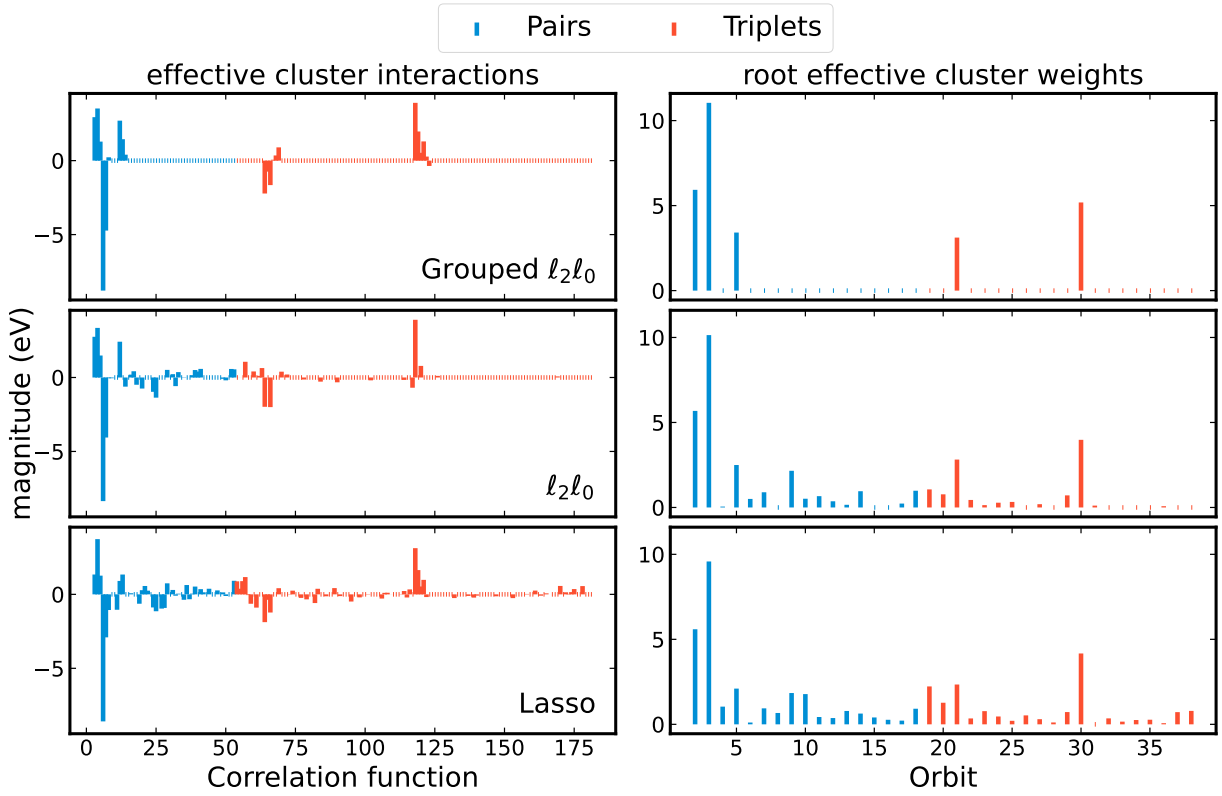


Figure 5.14: Fourier cluster expansion parameters (effective cluster interactions) and root effective cluster weights for three fits of an expansion involving pair and triplet terms up to 10 Åand 6 Årespectively. using the Lasso, and $\ell_2\ell_0$ regression with correlation function hierarchical constraints ($\ell_2\ell_0$) and cluster interaction constraints (grouped $\ell_2\ell_0$).

This can be rationalized to some extent simply by considering the relative differences in cluster interactions. Both expansions result in very similar nearest neighbor interactions, but relatively different second nearest neighbor pair interactions. The pair expansion results in a Co−Co and Ni−Ni second nearest neighbor interactions that are more favorable than their mixed counterparts. In contrast, second nearest neighbor interactions in the pair + triplet expansion give more favorable Co−Ni interactions.

More careful inspection of Figure 5.15 shows that at least another cluster, in addition to the first and second nearest neighbors, plays an important role in the thermodynamic behavior. The cluster energies for the remaining clusters are shown as dashed curves in Figure 5.15. In both expansions, there is a dashed curve that appears to contribute an important part of the total energy.

We can readily identify and gain further understanding of the relative importance of the contributions of the different interactions by looking at the corresponding cluster sensitivity

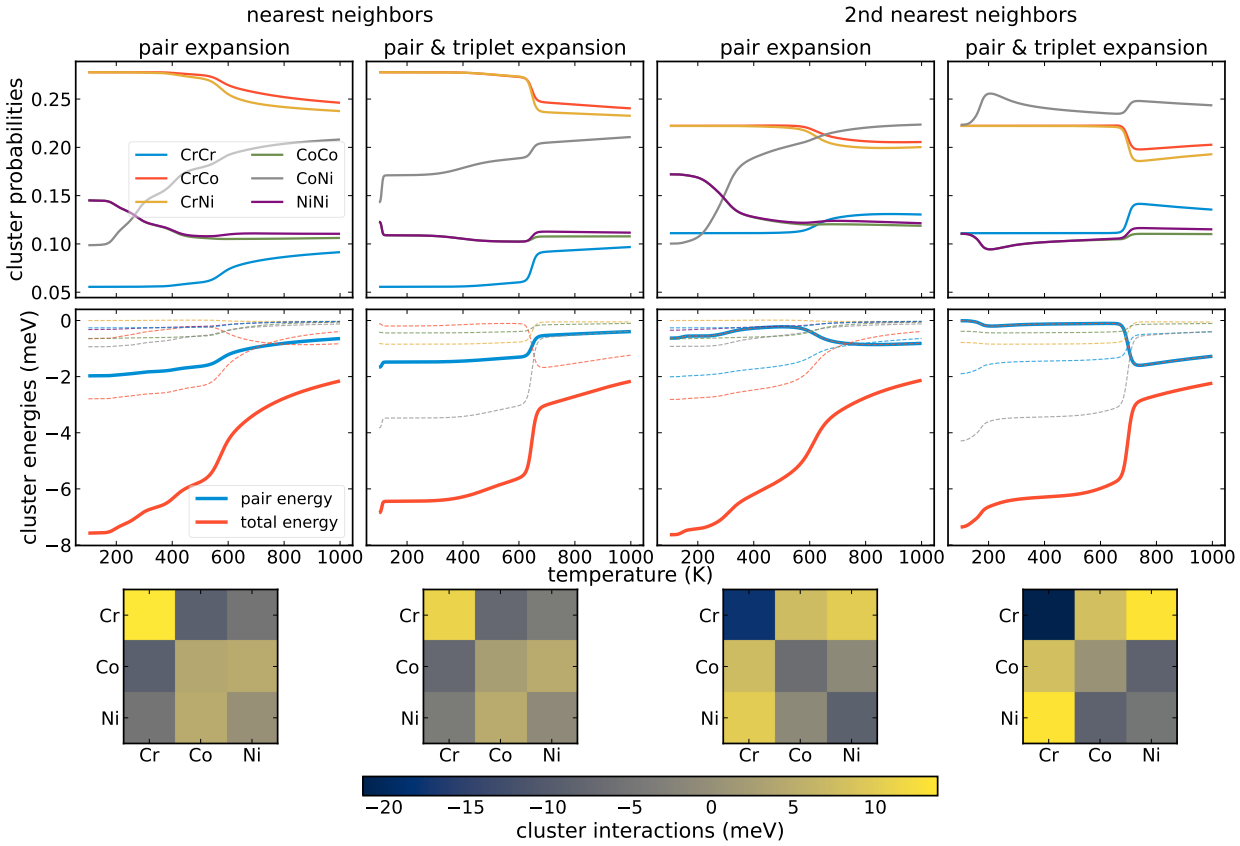
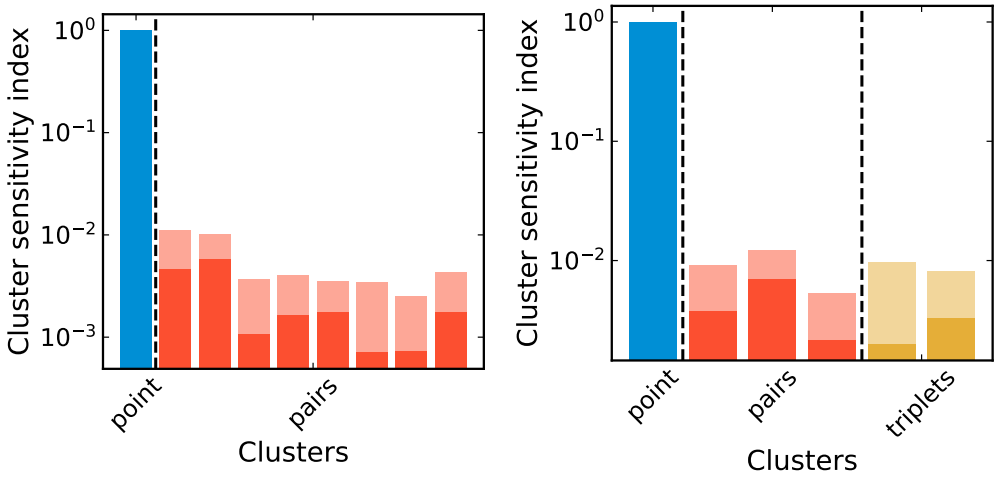


Figure 5.15: Nearest neighbor and second nearest neighbor cluster probabilities (top); cluster energies and total energy (middle); and nearest neighbor and second nearest neighbor mean cluster interactions. The cluster energies of the first and second nearest neighbors are plotted with a solid blue curve, the total energy with a solid red curve, and the remaining cluster energies are plotted with dashed curves

indices. Figure 5.16 shows the effective (solid color) and total (translucent) cluster sensitivity indices. We can observe that the most important interactions in both expansions indeed come from the first and second nearest neighbors. However, the longest range pair in the pair-only expansion and the triplets in the other expansion have also comparable sensitivity values. The longest range pair and the largest triplet are the clusters contributing to those unidentified interactions in Figure 5.15. The resulting *modulation* of short-range ordering by longer range or higher degree clusters has been previously observed and used to rationalize the phase transition of NiCoCr medium entropy alloys [171].

This example serves as a basic illustration of the additional insight into the energetic contributions and resulting atomic ordering of different clusters that can be obtained from the cluster decomposition and cluster sensitivity indices presented in Chapter 2.4. We have



(a) Pairs to 8 Å

(b) Pairs to 10 Å& triplets to 6 Å

Figure 5.16: Effective (solid color) and total (translucent) cluster sensitivity indices for two expansions of a NiCoCr medium entropy alloy.

no doubt these concepts can be used in a variety of different and novel ways beyond what we have illustrated here to analyze additional details and provide much richer insight into the effects of partial and full atomic disorder in the thermodynamics of complex multi-component crystals.

Chapter 6

Conclusion & outlook

In this work, we investigated the formal representation of generalized Lattice models and developed novel regression methods as a means to effectively capture the energy of multi-component crystalline materials in terms of their atomic configuration. The cluster expansion method [188, 192] is an essential part of the core formalism developed in this work. However, we have further formalized and extended the methodology. In doing so, we have added further insight into the mathematical structure and established its connection to well-developed mathematical fields including discrete harmonic analysis, analysis of variance, sensitivity analysis, and frame theory.

Two main representations of generalized lattice models were established. Fourier cluster expansions—which correspond to the original formalism of the cluster expansion method—were constructed from a symmetrized tensor product basis constructed from *standard site basis* sets. We further recast Fourier cluster expansions into a unique *cluster decomposition*, which we have shown constitutes a symmetrized version of the Sobol decomposition otherwise known as the functional ANOVA [96, 206]. This connection allows a formal interpretation of expansion terms as the conditional expectation of energy contributions from interactions between sites in a cluster. The second representation we have developed is a departure from the original cluster expansion method and involves a redundant representation in terms of a mathematical frame. We formulated the Potts frame expansion as a direct generalization of the Potts model to arbitrary interactions. The motivation behind the Potts frame is that its redundant nature allows robust and highly sparse estimation of applied lattice Hamiltonians from scant data.

Fourier cluster expansions and Potts frame expansions were then used to develop novel structured-sparsity-based regression methods that enable accurate, robust, and interpretable expansions of lattice Hamiltonians. The regression methods have been developed with the goal of allowing accurate estimation in high dimensional configuration spaces using relatively small samples of training structures and their respective energies computed by way of density functional theory. In addition, a series of practical implementations and auxiliary methods suitable for the effective implementation and learning of applied lattice models were presented. Finally, we have illustrated the successful application of the methodology to learn

lattice models of several technologically and scientifically relevant Li transition metal oxides and medium entropy alloys.

The ability to accurately capture the high dimensional configurational energy landscapes as lattice models permits the use of well-developed and efficient Monte Carlo sampling and related methodology to calculate thermodynamic properties of multi-principal element materials. Indeed, the cluster expansion method coupled with Monte Carlo (CE-MC) sampling has become a standard tool in the study of disordered and partially disordered metal alloy and ceramic materials [185, 212, 231, 232]. The continued exploration of materials with a larger number of components has compelled the continued innovation and development of statistical estimation and data sampling methods [2, 131, 148, 152, 153, 195, 226] (some that we presented in this work) to keep up with the increasing gap between the combinatorial growth of configuration spaces, and the relatively fixed number of training structures for which electronic calculations can be realistically computed to a sufficient level of accuracy. In accordance with such efforts, we have also focused in the presented work on refining and extending the underlying mathematical formalism that can be used to represent lattice Hamiltonians in ways that further enable novel high-dimensional parameter estimation methods and subsequent statistical thermodynamic calculations.

Although the work presented in this dissertation, along with a large body of previous work cited herein, has addressed to a great extent the construction and parametrization of lattice Hamiltonians for complex multi-component materials, we must acknowledge that (apart from a few exceptions) not much has been done that departs from the (now standard) CE-MC *recipe* to study disordered materials; namely: fit a lattice model to first-principles calculated data using generalized linear regression and use Metropolis Monte Carlo sampling to calculate statistical thermodynamic properties. Some notable exceptions that have established methodology beyond the standard recipe include aforementioned special quasi-random structures (SQS) [263] (methodology which itself has been establishes for a long time), the related special quasi-ordered structures (SQoS) [135, 184] and (SSOS) *small sets of ordered structures* [109], (provable) ground state search algorithms [100, 103, 128], and the use of Wang-Landau sampling [170, 171, 213].

6.1 Enhancements, extensions and new directions

We end this dissertation with a few suggestions on possible improvements to the methods listed above by leveraging the work presented in this dissertation. Finally, we also suggest new learning and thermodynamic inference paradigms that extend the use of applied lattice models beyond the standard CE-MC recipe.

Enhanced structure generation

Special structures

As already mentioned, using representative structures that approximate disordered or partially disordered materials is now a well-established computational method in studying multi-component materials. For the most part, pertinent methodology, including SQS [263], SQoS [135, 184] and SSOS [109], is based on searching for structures that match the values of correlation functions of fully disordered or partially disordered states. As we have discussed at length, any such property, such as a material's energy, at any temperature can be expanded in a suitable representation as follows,

$$\langle H(\boldsymbol{\sigma})\rangle_T = \sum_{\beta} m_\beta J_\beta \langle \Theta_\beta(\boldsymbol{\sigma})\rangle_T \tag{6.1}$$

The essence of special structure-based methods is to approximate the properties of a thermally disordered state using a single or a small set of representative structures. This objective can be achieved by searching for structures that minimize the difference between the thermodynamic internal energy sought and the corresponding energy of one or a few specific configurations. This objective can be written as an optimization problem as follows,

$$\begin{aligned}\min_{\{(w_i,\boldsymbol{\sigma}^{(i)})\}} \left| \langle H(\boldsymbol{\sigma})\rangle_T - \sum_{i=1}^{m} w_i H(\boldsymbol{\sigma}^{(i)}) \right| &= \min_{\{(w_i,\boldsymbol{\sigma}^{(i)})\}} \sum_{\beta\neq\emptyset} |m_\beta J_\beta| \left| \langle \Theta_\beta(\boldsymbol{\sigma})\rangle_T - \sum_{i=1}^{m} w_i \Theta_\beta(\boldsymbol{\sigma}^{(i)}) \right| \\ &= \min_{\{(w_i,\boldsymbol{\sigma}^{(i)})\}} \sum_{\beta\neq\emptyset} \left| \langle \Theta_\beta(\boldsymbol{\sigma})\rangle_T - \sum_{i=1}^{m} w_i \Theta_\beta(\boldsymbol{\sigma}^{(i)}) \right| \end{aligned} \tag{6.2}$$

where the minimization can be any suitable positive valued objective function, such as the square of the difference instead of the absolute value used above.

When the number of representative structures in Equation 6.2 is $m = 1$ and the fully disordered energy ($T \to \infty$) is sought, the above problem is the SQS objective [263]. When $m = 1$, but the correlations are matched to those at a finite temperature the SQoS objective is obtained [135, 184]. Finally, when $m > 1$ and $T \to \infty$ the problem represents the SSOS objective [109].

As we have established in Chapter 2, when using a Fourier correlation basis[1], the fully disordered correlation functions are equal to zero $\langle \Theta_\beta(\boldsymbol{\sigma})\rangle_{T\to\infty} = \langle \Theta_\beta(\boldsymbol{\sigma})\rangle_{\boldsymbol{\rho}} = 0$. Such that the minimization in Equation 6.2 simply involves making correlation functions $\Theta_\beta(\boldsymbol{\sigma})$ as close to zero as possible. Indeed, this is how special structures are most commonly generated for equiatomic compositions [229, 263]. However, for all other compositions, the practice is to use the values of correlation functions $\langle \Theta_\beta\rangle_T$ in the random (i.e. non-interacting state) but still under a uniform (equiatomic) apriori measure only, which results in nonzero target correlation function values. Although there is nothing formally unsound in using this

[1] One can actually get away with using an orthogonal cluster correlation basis as well.

approach, the correlation function values in the non-interacting limit need to be computed for the particular basis chosen, and their distributions may be generally centered at non-zero values. By instead using Fourier correlation basis under an apriori measure corresponding to the composition sought as detailed in Chapter 2.3, the correlation function expectation values will always be zero and the optimization problem is the same as that for equiatomic compositions (uniform apriori distributions).

Perhaps even more importantly, the size of the problem (the number of correlation functions optimized over) as stated in Equation 6.2, scales polynomially with the number of components or species. Furthermore, there is no reason to purport special importance to minimizing any subset of correlation functions acting over the same site space clusters. These two factors can be directly addressed by simply re-casting the optimization problem in terms of a cluster decomposition as follows,

$$\min_{\{(w_i,\boldsymbol{\sigma}^{(i)})\}} \sum_{B\neq\emptyset} \left| \langle H_B(\boldsymbol{\sigma})\rangle_T - \sum_{i=1}^{m} w_i H_B(\boldsymbol{\sigma}^{(i)}) \right| \tag{6.3}$$

where the problem in Equation 6.3 is now constant in the number of components. As we have shown in Chapter 2.4, the expected cluster interactions $\langle H_B(\boldsymbol{\sigma})\rangle_T$ are equal to zero in the random limit. However, the expansion coefficients $\hat{m}_\beta J_\beta$ are now included implicitly in the cluster interactions H_B, and cannot simply be factored out of the new objective as done in Equation 6.2. Although this is a nonissue, since the expected cluster interactions are equal to zero in the random limit, irrespective of specific values of the coefficients J_β. This fact can instead be used to control the importance of cluster interactions in the optimization problem. For example, one can simply set all $J_\beta = 1/\hat{m}_\beta^p$ with $p > 1$ to set the importance of interactions inversely proportional to their degree.

Finally, to obtain representative partially disordered structures, without the need of fitting a lattice model and running MC sampling to obtain values of correlation functions for partially disordered states, one can directly use the methods described in Chapter B.5 to seek out structures with specific short-range ordering, by instead minimizing the sum of differences of cluster probabilities and the cluster averages for representative structures,

$$\min_{\{(w_i,\boldsymbol{\sigma}^{(i)})\}} \sum_{B} \left| \mathbb{P}_B(\boldsymbol{\sigma}_S \mid T) - \sum_{i=1}^{m} w_i M_{DB} \mathbb{V}_S^{+} \boldsymbol{\Pi}_B(\boldsymbol{\sigma}^{(i)}) \right| \tag{6.4}$$

The above optimization allows using target short-range order values obtained from other means rather than from fitting and sampling a cluster expansion. The same objective function in Equation 6.4 can also be suitably written using a Potts frame in terms of cluster indicator correlation functions.

Ground states

The *ground-state problem* in lattice models has been an ongoing research effort and a notably challenging problem for generalized lattice models [100]. The problem is concerned with

finding the lowest energy configurations for a given generalized lattice model in some suitable representation. Several approaches have been proposed to generate ground states or near ground states for lattice models of varying complexity [60, 69, 100, 103, 111, 217]. State-of-the-art methods are now able to find exact ground state configurations over finite domains and in some cases provably exact ground state configurations for bulk structures [100]. In the majority of methods, specifically, those most recently proposed [100, 103], the ground state problem is expressed in a generalized lattice gas representation[2] which allows the problem to be converted into a pseudo-Boolean or integer optimization problem. Such a transformation allows leveraging significant recent advancements in Boolean and integer optimization solvers [100, 103].

However, the constraint to using lattice gas representations has limited the application of such methods to binary lattice models [103], or to generalized models fit solely in a lattice gas representation [99]. The main impediment in using such methods with other representations is not fundamental, it has been simply a lack of practical transformation procedures between general representations of lattice models. Although in this work, we have only detailed a practical way to transform any representation to a Fourier cluster expansion, we have stated how any lattice model represented as a (pseudo) cluster decomposition, can be trivially rewritten as a symmetrized Potts model. To be more precise, one can express Equation 3.21 from Chapter 3.2 in terms of reduced cluster indicator correlation tensors as follows,

$$m_B N J_\beta I_\beta(\boldsymbol{\sigma}) = \sum_{S \in B} \left[\hat{H}_{B(\beta)}\right]_{\boldsymbol{\sigma}_\beta} \times \hat{m}_\beta \hat{I}_\beta(\boldsymbol{\sigma}_S) \tag{6.5}$$

where $\boldsymbol{\sigma}_\beta$ represents the cluster configuration indicated by $I_\beta(\boldsymbol{\sigma})$. The terms on the left-hand-side of Equation 6.5 are effectively Boolean variables, $\hat{m}_\beta \hat{I}_\beta(\boldsymbol{\sigma}_S) \in \{0, 1\}$.

By using Equation 6.5 the ground state problem can be written for any generalized representation as follows,

$$\min_{X_\beta} \sum_{S \in [N]} \sum_{S \in B} \left[\hat{H}_{B(\beta)}\right]_{\boldsymbol{\sigma}_\beta} \times X_\beta \tag{6.6}$$

where the Boolean variables X_β are shorthand for the reduced cluster indicator functions $X_\beta = \hat{m}_\beta \hat{I}_\beta(\boldsymbol{\sigma}_S)$.

The expression in Equation 6.6 can be directly used in the recently proposed linear mixed integer programming method [103]. Further, by simply expanding X_β as products of species indicator functions corresponding to the sites given by the support of the multi-indices in β, the problem can be re-written as a satisfiability problem and provable ground states searches can be carried out according to recently developed MAX-SAT methodology [100]. The above prescription allows leveraging state-of-the-art ground state search methods using any lattice Hamiltonian representation.

[2] Equivalently in a cluster expansion using site indicator functions.

New learning paradigms

Revisiting direct configurational averaging

In Chapter 4 the method of direct configurational averaging (DCA) was briefly described in the historical context of the development of learning methodology for applied lattice models. As mentioned, the DCA has been mostly abandoned due to the comparably larger number of training structures required to obtain similar levels of accuracy compared to the linear regression-based structure inversion method. However, the DCA provides a mathematically rigorous avenue for estimating expansion coefficients independently. Revisiting the DCA may be a worth-wile endeavor to rigorously explore the extent and limitations of learning applied lattice models as well as some of the unsettled subtleties in doing so, such as the effects of structural relaxations [155], species concentrations [147, 187, 191], re-normalized interactions [190], and long-range interactions.

In light of the substantial advancements in computing energies of materials, specifically by way of machine learning methods, it is now possible to obtain substantially larger training sets with levels of accuracy that are quickly approaching those of density functional theory (DFT). By using recently developed and highly accurate machine learning potentials (MLPs) [12, 39, 57] energies for configurations with a fixed structure can be obtained to similar levels of accuracy but orders of magnitude faster than DFT [265]. In addition, energies for structurally relaxed configurations can be obtained by coupling MLPs with molecular dynamics or one of many available structural relaxation algorithms [20, 77, 168, 264]. This approach has very recently been undertaken to explore the convergence and predictions of cluster expansion models trained using linear regression with respect to varying training set sizes [251].

Using substantially larger MLP-generated training sets may well be an opportune approach to leverage the DCA as a revitalized learning algorithm that can be used to obtain theoretical guarantees on the resulting accuracy and structure of the estimated coefficients. As a matter of fact, the DCA can be analyzed within the *probably approximately correct* (PAC) learning framework [158]. Specifically, probabilistic bounds on the accuracy of DCA estimation of expansion coefficients following Equation 4.2, can be obtained by way of a Chernoff bound under the *a-priori* distribution $\boldsymbol{\rho}$ as follows [138],

$$\mathbb{P}\left[|J_\beta - \widetilde{J}_\beta| \leq \varepsilon\right] \leq 2e^{-2m\varepsilon^2/r_\beta^2} \tag{6.7}$$

where $J_\beta, \widetilde{J}_\beta$ are the ground-truth and DCA estimated expansion coefficients respectively. ε is an arbitrary accuracy bound; $r_\beta = b - a$ is the range the expansion coefficient lies in $a \geq J_\beta \leq b$, and m is the number of training structures used to estimate $\widetilde{J}_\beta$

By requiring that the probability in Equation 6.7 be at most some value $\delta < 1$, the following PAC bound on the number of training samples is obtained,

$$m \geq \log(2/\delta)\frac{r_\beta^2}{2\varepsilon^2} \tag{6.8}$$

Meaning that with m satisfying the above, the estimated coefficient $\widetilde{J}_\beta$ will be within ϵ of the ground truth with probability at least $1-\delta$.

In this manner, using DCA allows the estimation of expansion coefficients with guaranteed theoretical bounds. Finally, if the goal is to increase the convergence or accuracy concerning the number of training samples, the DCA approach can be suitably extended by way of established non-parametric estimation, such as using *modulation* [242]. Introducing modulation allows further optimization of prediction accuracy by estimating expansion coefficients as $J_\beta \approx b_\beta \widetilde{J}_\beta$ and optimizing the modulated estimator over the modulation parameters b_β using suitable risk functions [242].

Thermodynamic inference

Optimization based inference

Sampling-based inference of applied lattice models by way of Monte Carlo has become the de facto method for computing free energies and associated thermodynamic properties. Understandingly so, Monte Carlo sampling is a straightforward, general and effective computation technique. However, Monte Carlo can also be computationally expensive (millions of samples are usually necessary for sufficiently accurate calculations), inefficient at exploring rough energy landscapes, suffers from *critical slowing down* near phase transitions [127], and often does not allow direct calculations of free energy functions. In many situations, more efficient exploration—computationally faster and/or with improved coverage of phase space—or direct approximations of free energy may be necessary to obtain a better understanding of the thermodynamic behavior of materials. In such cases, inference based on optimization can be a compelling computation technique complementary to sampling-based inference.

The original motivation behind constructing a rigorous basis for functions of atomic configuration in crystals was to develop a way to solve the *cluster variation method* (CVM) [192, 193]. The CVM, which is a generalization of the Bethe approximation beyond pair-wise interactions [174], entails a *variational* approximation to the free energy [114]. Therefore, the CVM is essentially an optimization-based (mean-field) inference method. Although the CVM has been largely abandoned in computational materials research, contemporaneous development of optimization-based inference has continued and expanded in the statistics community. Furthermore, a formal connection between prominent inference algorithms, such as belief propagation and generalized belief propagation and the Bethe approximation and CVM has been established [254], and now a rich class of optimization inference methods, generally known as *region-based free energy* approximations, has been developed in statistics [121, 234].

Optimization-based inference in statistics is used in the context of *probabilistic graphical models*, which are multivariate probabilistic models represented by graphs where variables are expressed as nodes and statistical dependence is represented by edges [121]. It should be no surprise, that applied lattice models are distinctly amenable for use with modern variational inference algorithms since they fall squarely within the definition of probabilis-

tic graphical models. The applied lattice models presented in this work can be trivially expressed as *undirected* graphical models, otherwise known as *Markov random fields* [121, 234]. Undirected graphical models are usually expressed as follows,

$$\mathbb{P}(\boldsymbol{\sigma}) = \frac{1}{Z} \prod_{S \in \mathcal{M}} \psi_S(\boldsymbol{\sigma}_S) \tag{6.9}$$

where the functions ψ_S are called potentials and $\mathcal{M}$ a set of cliques in the undirected graph representation of the model.

Equation 6.9 can be written as a Boltzmann distribution simply by setting $\psi_{\boldsymbol{\sigma}_S} = e^{-\beta H_S(\boldsymbol{\sigma}_S)}$, in which case the graphical model is referred to as a member of an *exponential family* [121, 234]. More specifically, thermodynamic ensembles defined by a Fourier cluster expansion or a Potts frame expansion are *linear exponential families* [121]. Fourier cluster expansions are *minimal* exponential families since there is exactly one set of expansion coefficients to express a particular distribution given a set of correlation functions. On the other hand, Potts frame expansions are *overcomplete* exponential families since there always exists an affine transformation between sets of coefficients that represent the exact same distribution [234].

Directly using many of the now well-developed tools for optimization-based inference is a straightforward avenue to extend the applicability of applied lattice models in materials research. Specific benefits from doing so include substantial improvement in computation time by using iterative message passing algorithms such as loopy belief propagation (BP) and generalized belief propagation (GBP), which have been shown to provide over an order of magnitude speed-up compared to the natural iteration method originally proposed to solve the CVM [114, 174]. Furthermore, despite the approximate nature of such methods, in some cases, such as with belief propagation, guaranteed bounds on the free energy can be obtained [234]. On the other hand, notable weaknesses are that BP and GBP have no convergence guarantees for graphs with loops (basically any applied lattice model); and although GBP can provide much better approximations there are no guarantees that the approximation is an upper or lower bound on the free energy. Nevertheless, progress in addressing these issues has been achieved as of late. The recently proposed *neural-enhanced* BP and BP neural networks rely on a neural network to correct messages and thus obtain much tighter lower bounds on the free energy with faster convergence time [124, 194]. Other alternatives have been developed that do not rely on iterative message passing altogether, and instead perform direct minimization of BP using gradient descent [244, 252], as well as direct minimization of GBP using *region-based neural networks* [134]. Finally, a most recent proposal which aims to directly account for invariance/equivariance of the underlying graphical model [210], may be notably applicable for use with applied lattice models.

Generative models

As a fundamentally related modeling paradigm to variational inference methods, generative modeling represents a further route for advancing how applied lattice models can be used in

statistical thermodynamic calculations, and some cases even obtaining improved learning of model coefficients.

Generative modeling, in particular *deep generative modeling*, has generated much attention in machine learning research and has recently received an enormous level of hype from the general public with the recent release of several language, audio, and image generation algorithms [183]. As the name suggests, a generative model is a model that can be used to generate samples from a target probability distribution that is usually unknown or intractable. In deep generative modeling, neural networks with several layers are used to approximate these complex high-dimensional probability distributions [183]. A variety of recently developed algorithms and methods including variational autoencoders (VAE) [115, 179], generative adversarial networks [51], autoregressive networks [83], normalizing flows [116, 178], and probabilistic diffusion models [95, 117, 207], have shown exceptional, flexible, and tractable sample generation capabilities from arbitrarily complex distributions.

A majority of these methods are almost directly applicable to use with generalized lattice models. The use of generative models with lattice models has already led to the development of intriguing computation techniques for sampling and variational inference. For example, deep generative models have been recently explored as a way to improve MC sampling efficiency by generating transition proposals from actively learned (latent) distributions. Improvements in MC sampling using VAEs [141, 145], autoregressive neural networks [248], and normalizing flows [75] have been recently studied in classical and spin-glass lattice models. We can expect similar improvements to translate for MC sampling of generalized lattice models, which may be a fruitful way to address slow Markov chain mixing and inefficient sampling that can occur in rugged energy landscapes associated with more complex lattice hamiltonians.

In addition, generative models can be used as an alternative to MC sampling altogether, either by generating samples directly or by providing variational free energy approximations. Likewise, initial work is already underway paving new directions for methodological extensions in lattice model inference. For instance, autoregressive neural networks have been shown to give accurate free energy approximations and provide approximate but unbiased sampling from Boltzman distributions for a handful of (binary) lattice models and moderately sized system sizes [52, 249]. Furthermore, hierarchical variational models have been proposed as a means to scale generative sampling and free energy approximations to much larger system sizes [1]. Although the usage of deep generative modeling in materials science and in particular with lattice models is still nascent and limited in scale, when coupled with the framework of generalized lattice models we have presented in this dissertation, it may well constitute another powerful paradigm to extend and/or complement CE-MC calculations.

6.2 Closing remarks

With the continued growth of available data and the advent of highly accurate, transferable, and general machine learning (ML) models, we must ask ourselves: Are methods based on

applied lattice models still relevant? Indeed, state-of-the-art machine learning potentials (MLPs) have been shown to predict a variety of properties of multi-component systems and their derivatives with respect to structural parameters to extraordinary levels of accuracy [12, 39, 57]. Furthermore, these MLPs can be used with molecular dynamics to perform statistical thermodynamic calculations. These MLPs have been shown to yield substantial improvements in accuracy compared to classical molecular dynamics, and substantial performance improvements compared to ab-initio molecular dynamics [12, 40, 137].

Nevertheless, we maintain that applied lattice models are still quite relevant and will probably remain so for the foreseeable future. First off, MLPs and molecular dynamics are not effective alternatives to computing equilibrium properties of atomic configuration. Despite the exceptional increase in performance and accuracy, the time scales that are attainable by molecular dynamics are in many cases on completely different scales to that of excitations of atomic configuration. In contrast, Monte Carlo and variational inference methods based on applied lattice models deal (almost) directly with the equilibrium distributions of atomic configuration. Though one could consider using Monte Carlo directly with an MLP, structural coordinates must still be accounted for, either directly with Monte Carlo or by performing a structural relaxation for each new atomic configuration. All in all, the simple, extensible, and robust nature of applied lattice models, coupled with sampling or optimization-based inference, renders them a valuable method in the study of multi-component materials that is unlikely to be superseded anytime soon.

Perhaps even more importantly, lattice models provide an avenue to improve machine learning models. Applied lattice models have already been used as direct input features to neural network models [151]. And as a matter of fact, the mathematical formalism of Fourier expansions over product spaces underlying cluster expansions of configuration has been directly used in the development of the atomic cluster expansion [56], which at its core is an extension of a configuration cluster expansion that includes the positions of sites by introducing $\mathbb{R}^3$ vector spaces into the product space in Equation 2.2.

On account of their simple, effective, and established use in statistical thermodynamics calculations, the many possible methodological extensions and enhancements, and many promising ways to compliment and inspire novel state-of-the-art ML models, applied Lattice models are indispensable to our ability to compute thermodynamic properties of atomic configuration and advance our understanding of disordered and partially disordered multi-component materials. We expect that the mathematical formalism and framework we have established, as well as the practical data preparation and structured sparsity linear regression methods we have presented, serve as a further foundation from which continued progress can be made.

Bibliography

[1] Jaan Altosaar. "Probabilistic Modeling of Structure in Science: Statistical Physics to Recommender Systems". PhD thesis. Princeton University, 2020.

[2] Mattias Ångqvist et al. "ICET – A Python Library for Constructing and Sampling Alloy Cluster Expansions". In: *Advanced Theory and Simulations* 2.7 (2019), p. 1900015. ISSN: 2513-0390. DOI: 10.1002/adts.201900015.

[3] M. Asta, V. Ozolins, and C. Woodward. "A First-Principles Approach to Modeling Alloy Phase Equilibria". In: *JOM* 53.9 (Sept. 2001), pp. 16–19. ISSN: 1543-1851. DOI: 10.1007/s11837-001-0062-3.

[4] M. Asta et al. "Effective Cluster Interactions from Cluster-Variation Formalism. I". In: *Physical Review B* 44.10 (Sept. 1991), pp. 4907–4913. DOI: 10.1103/PhysRevB.44.4907.

[5] "Lp-Spaces". In: *Measure Theory and Probability Theory*. Ed. by Krishna B. Athreya and Soumendra N. Lahiri. Springer Texts in Statistics. New York, NY: Springer, 2006, pp. 83–111. ISBN: 978-0-387-35434-7. DOI: 10.1007/978-0-387-35434-7_4.

[6] "Product Measures, Convolutions, and Transforms". In: *Measure Theory and Probability Theory*. Ed. by Krishna B. Athreya and Soumendra N. Lahiri. Springer Texts in Statistics. New York, NY: Springer, 2006, pp. 147–188. ISBN: 978-0-387-35434-7. DOI: 10.1007/978-0-387-35434-7_6.

[7] Francis Bach et al. "Optimization with Sparsity-Inducing Penalties". In: *Foundations and Trends® in Machine Learning* 4.1 (Jan. 2012), pp. 1–106. ISSN: 1935-8237, 1935-8245. DOI: 10.1561/2200000015.

[8] Francis Bach et al. "Structured Sparsity through Convex Optimization". In: *Statistical Science* 27.4 (Nov. 2012), pp. 450–468. ISSN: 0883-4237, 2168-8745. DOI: 10.1214/12-STS394.

[9] Luis Barroso-Luque, Julia H. Yang, and Gerbrand Ceder. "Sparse Expansions of Multicomponent Oxide Configuration Energy Using Coherency and Redundancy". In: *Physical Review B* 104.22 (Dec. 2021), p. 224203. DOI: 10.1103/PhysRevB.104.224203.

[10] Luis Barroso-Luque et al. "Cluster expansions of multicomponent ionic materials: Formalism & Methodology". In: *Physical Review B* (Oct. 2022).

[11] Luis Barroso-Luque et al. "smol: A Python package for cluster expansions and beyond". In: *Journal of Open Source Software* 7.77 (2022), p. 4504. DOI: 10.21105/joss.04504. URL: https://doi.org/10.21105/joss.04504.

[12] Simon Batzner et al. "E(3)-Equivariant Graph Neural Networks for Data-Efficient and Accurate Interatomic Potentials". In: *Nature Communications* 13.1 (May 2022), p. 2453. ISSN: 2041-1723. DOI: 10.1038/s41467-022-29939-5.

[13] Alexei Bazavov, Bernd A. Berg, and Santosh Dubey. "Phase Transition Properties of 3D Potts Models". In: *Nuclear Physics B* 802.3 (Oct. 2008), pp. 421–434. ISSN: 0550-3213. DOI: 10.1016/j.nuclphysb.2008.04.020.

[14] Paul D. Beale. *Statistical Mechanics*. Academic Press, Apr. 2011. ISBN: 978-0-12-382189-8.

[15] A. D. Beath and D. H. Ryan. "Fcc Antiferromagnetic Ising Model in a Uniform External Field Solved by Mean-Field Theory". In: *Physical Review B* 72.1 (July 2005), p. 014455. DOI: 10.1103/PhysRevB.72.014455.

[16] A. D. Beath and D. H. Ryan. "Thermodynamic Properties of the Fcc Ising Antiferromagnet Obtained from Precision Density of States Calculations". In: *Physical Review B* 73.17 (May 2006), p. 174416. DOI: 10.1103/PhysRevB.73.174416.

[17] Dimitris Bertsimas and Angela King. "OR Forum—An Algorithmic Approach to Linear Regression". In: *Operations Research* 64.1 (Feb. 2016), pp. 2–16. ISSN: 0030-364X. DOI: 10.1287/opre.2015.1436.

[18] Jacob Bien, Jonathan Taylor, and Robert Tibshirani. "A LASSO FOR HIERARCHICAL INTERACTIONS". In: *The Annals of Statistics* 41.3 (2013), pp. 1111–1141. ISSN: 0090-5364.

[19] K. Binder and A. P. Young. "Spin Glasses: Experimental Facts, Theoretical Concepts, and Open Questions". In: *Reviews of Modern Physics* 58.4 (Oct. 1986), pp. 801–976. DOI: 10.1103/RevModPhys.58.801.

[20] Erik Bitzek et al. "Structural Relaxation Made Simple". In: *Physical Review Letters* 97.17 (Oct. 2006), p. 170201. DOI: 10.1103/PhysRevLett.97.170201.

[21] Sébastien Bourguignon et al. "Exact Sparse Approximation Problems via Mixed-Integer Programming: Formulations and Computational Performance". In: *IEEE Transactions on Signal Processing* 64.6 (Mar. 2016), pp. 1405–1419. ISSN: 1941-0476. DOI: 10.1109/TSP.2015.2496367.

[22] STEPHEN G. BRUSH. "History of the Lenz-Ising Model". In: *Reviews of Modern Physics* 39.4 (Oct. 1967), pp. 883–893. DOI: 10.1103/RevModPhys.39.883.

[23] Keith T. Butler et al. "Machine Learning for Molecular and Materials Science". In: *Nature* 559.7715 (July 2018), pp. 547–555. ISSN: 1476-4687. DOI: 10.1038/s41586-018-0337-2.

[24] Richard H. Byrd et al. "A Limited Memory Algorithm for Bound Constrained Optimization". In: *SIAM Journal on Scientific Computing* 16.5 (Sept. 1995), pp. 1190–1208. ISSN: 1064-8275. DOI: 10.1137/0916069.

[25] Giuseppe C. Calafiore and Laurent El Ghaoui. *Optimization Models*. Cambridge University Press, Oct. 2014. ISBN: 978-1-139-99293-0.

[26] Herbert B. Callen. *Thermodynamics and an Introduction to Thermostatistics*. John Wiley & Sons, Jan. 1991. ISBN: 978-0-471-86256-7.

[27] Emmanuel J. Candes and Michael B. Wakin. "An Introduction To Compressive Sampling". In: *IEEE Signal Processing Magazine* 25.2 (Mar. 2008), pp. 21–30. ISSN: 1558-0792. DOI: 10.1109/MSP.2007.914731.

[28] Emmanuel J. Candès, Michael B. Wakin, and Stephen P. Boyd. "Enhancing Sparsity by Reweighted 1 Minimization". In: *Journal of Fourier Analysis and Applications* 14.5 (Dec. 2008), pp. 877–905. ISSN: 1531-5851. DOI: 10.1007/s00041-008-9045-x.

[29] Emmanuel J. Candès et al. "Compressed Sensing with Coherent and Redundant Dictionaries". In: *Applied and Computational Harmonic Analysis* 31.1 (July 2011), pp. 59–73. ISSN: 1063-5203. DOI: 10.1016/j.acha.2010.10.002.

[30] Tullio Ceccherini-Silberstein, Fabio Scarabotti, and Filippo Tolli. *Discrete Harmonic Analysis: Representations, Number Theory, Expanders, and the Fourier Transform*. Cambridge Studies in Advanced Mathematics. Cambridge: Cambridge University Press, 2018. ISBN: 978-1-107-18233-2. DOI: 10.1017/9781316856383.

[31] G. Ceder, G. D. Garbulsky, and P. D. Tepesch. "Convergent Real-Space Cluster Expansion for Configurational Disorder in Ionic Systems". In: *Physical Review B* 51.17 (May 1995), pp. 11257–11261. DOI: 10.1103/PhysRevB.51.11257.

[32] G. Ceder et al. "Identification of Cathode Materials for Lithium Batteries Guided by First-Principles Calculations". In: *Nature* 392.6677 (Apr. 1998), pp. 694–696. ISSN: 1476-4687. DOI: 10.1038/33647.

[33] Gerbrand Ceder. "Alloy Theory and Its Applications to Long Period Superstructure Ordering in Metallic Alloys and High-Temperature Superconductors". PhD thesis. United States – California: University of California, Berkeley, 1991. ISBN: 9798208009635.

[34] Gerbrand Ceder et al. "Thermodynamics of Oxides with Substitutional Disorder: A Microscopic Model and Evaluation of Important Energy Contributions". In: *Journal of the American Ceramic Society* 81.3 (1998), pp. 517–525. ISSN: 1551-2916. DOI: 10.1111/j.1151-2916.1998.tb02369.x.

[35] A. V. Ceguerra et al. "Short-Range Order in Multicomponent Materials". In: *Acta Crystallographica Section A: Foundations of Crystallography* 68.5 (Sept. 2012), pp. 547–560. ISSN: 0108-7673. DOI: 10.1107/S0108767312025706.

[36] Anna V. Ceguerra et al. "Quantitative Description of Atomic Architecture in Solid Solutions: A Generalized Theory for Multicomponent Short-Range Order". In: *Physical Review B* 82.13 (Oct. 2010), p. 132201. DOI: 10.1103/PhysRevB.82.132201.

[37] David Chandler. *Introduction to Modern Statistical Mechanics*. Oxford University Press, 1987. ISBN: 978-0-19-504277-1.

[38] Jin Hyun Chang et al. "CLEASE: A Versatile and User-Friendly Implementation of Cluster Expansion Method". In: *Journal of Physics: Condensed Matter* 31.32 (May 2019), p. 325901. ISSN: 0953-8984. DOI: 10.1088/1361-648X/ab1bbc.

[39] Chi Chen and Shyue Ping Ong. *A Universal Graph Deep Learning Interatomic Potential for the Periodic Table*. Aug. 2022. DOI: 10.48550/arXiv.2202.02450. arXiv: 2202.02450 [cond-mat, physics:physics].

[40] Tina Chen et al. "Removing the two-phase transition in spinel $LiMn_2O_4$ through cation disorder". In: *Energy Environtal Science* ("submitted 2022").

[41] Xuefei Chen et al. "Direct Observation of Chemical Short-Range Order in a Medium-Entropy Alloy". In: *Nature* 592.7856 (Apr. 2021), pp. 712–716. ISSN: 1476-4687. DOI: 10.1038/s41586-021-03428-z.

[42] Hugh Chipman. "Bayesian Variable Selection with Related Predictors". In: *The Canadian Journal of Statistics / La Revue Canadienne de Statistique* 24.1 (1996), pp. 17–36. ISSN: 0319-5724. DOI: 10.2307/3315687.

[43] Ole Christensen. *Frames and Bases: An Introductory Course*. Applied and Numerical Harmonic Analysis. Birkhäuser Basel, 2008. ISBN: 978-0-8176-4677-6. DOI: 10.1007/978-0-8176-4678-3.

[44] F. Claro and G. D. Mahan. "Ising Model with Magnetic Field for 3D Cubic Structures". In: *Journal of Physics C: Solid State Physics* 10.4 (Feb. 1977), pp. L73–L77. ISSN: 0022-3719. DOI: 10.1088/0022-3719/10/4/004.

[45] R. J. Clément, Z. Lun, and G. Ceder. "Cation-Disordered Rocksalt Transition Metal Oxides and Oxyfluorides for High Energy Lithium-Ion Cathodes". In: *Energy & Environmental Science* 13.2 (Feb. 2020), pp. 345–373. ISSN: 1754-5706. DOI: 10.1039/C9EE02803J.

[46] Raphaële J. Clément et al. "Short-Range Order and Unusual Modes of Nickel Redox in a Fluorine-Substituted Disordered Rocksalt Oxide Lithium-Ion Cathode". In: *Chemistry of Materials* 30.19 (Oct. 2018), pp. 6945–6956. ISSN: 0897-4756. DOI: 10.1021/acs.chemmater.8b03794.

[47] Marvin L. Cohen and Steven G. Louie. *Fundamentals of Condensed Matter Physics*. Cambridge University Press, May 2016. ISBN: 978-0-521-51331-9.

[48] J. W. D. Connolly and A. R. Williams. "Alloy Phase Diagrams From First Principles". In: *The Electronic Structure of Complex Systems*. Ed. by P. Phariseau and W. M. Temmerman. NATO ASI Series. Boston, MA: Springer US, 1984, pp. 581–592. ISBN: 978-1-4613-2405-8. DOI: 10.1007/978-1-4613-2405-8_9.

[49] J. W. D. Connolly and A. R. Williams. "Density-Functional Theory Applied to Phase Transformations in Transition-Metal Alloys". In: *Physical Review B* 27.8 (Apr. 1983), pp. 5169–5172. DOI: 10.1103/PhysRevB.27.5169.

[50] J. M. Cowley. "An Approximate Theory of Order in Alloys". In: *Physical Review* 77.5 (Mar. 1950), pp. 669–675. DOI: 10.1103/PhysRev.77.669.

[51] Antonia Creswell et al. "Generative Adversarial Networks: An Overview". In: *IEEE Signal Processing Magazine* 35.1 (Jan. 2018), pp. 53–65. ISSN: 1558-0792. DOI: 10.1109/MSP.2017.2765202.

[52] James Damewood, Daniel Schwalbe-Koda, and Rafael Gómez-Bombarelli. "Sampling Lattices in Semi-Grand Canonical Ensemble with Autoregressive Machine Learning". In: *npj Computational Materials* 8.1 (Apr. 2022), pp. 1–10. ISSN: 2057-3960. DOI: 10.1038/s41524-022-00736-4.

[53] Mark A. Davenport et al. "Introduction to Compressed Sensing". In: *Compressed Sensing*. Ed. by Yonina C. Eldar and Gitta Kutyniok. Cambridge: Cambridge University Press, 2012, pp. 1–64. ISBN: 978-0-511-79430-8. DOI: 10.1017/CBO9780511794308.002.

[54] D. de Fontaine. "The Number of Independent Pair-Correlation Functions in Multi-component Systems". In: *Journal of Applied Crystallography* 4.1 (Feb. 1971), pp. 15–19. ISSN: 0021-8898. DOI: 10.1107/S0021889871006174.

[55] Didier de Fontaine. "Cluster Variation and Cluster Statics". In: *Theory and Applications of the Cluster Variation and Path Probability Methods*. Ed. by J. L. Morán-López and J. M. Sanchez. Boston, MA: Springer US, 1996, pp. 125–144. ISBN: 978-1-4613-0419-7. DOI: 10.1007/978-1-4613-0419-7_8.

[56] Ralf Drautz. "Atomic Cluster Expansion for Accurate and Transferable Interatomic Potentials". In: *Physical Review B* 99.1 (Jan. 2019), p. 014104. DOI: 10.1103/PhysRevB.99.014104.

[57] Ralf Drautz. "Atomic Cluster Expansion of Scalar, Vectorial, and Tensorial Properties Including Magnetism and Charge Transfer". In: *Physical Review B* 102.2 (July 2020), p. 024104. DOI: 10.1103/PhysRevB.102.024104. arXiv: 2003.00221.

[58] H. Dreyssé et al. "Determination of Effective-Pair Interactions in Random Alloys by Configurational Averaging". In: *Physical Review B* 39.4 (Feb. 1989), pp. 2442–2452. DOI: 10.1103/PhysRevB.39.2442.

[59] Petros Drineas et al. "Fast Approximation of Matrix Coherence and Statistical Leverage". In: *Journal of Machine Learning Research* 13.111 (2012), pp. 3475–3506. ISSN: 1533-7928.

[60] Yu. I. Dublenych. "Ground States of the Ising Model on the Shastry-Sutherland Lattice and the Origin of the Fractional Magnetization Plateaus in Rare-Earth-Metal Tetraborides". In: *Physical Review Letters* 109.16 (Oct. 2012), p. 167202. DOI: 10.1103/PhysRevLett.109.167202.

[61] F. Ducastelle and F. Gautier. "Generalized Perturbation Theory in Disordered Transitional Alloys: Applications to the Calculation of Ordering Energies". In: *Journal of Physics F: Metal Physics* 6.11 (Nov. 1976), pp. 2039–2062. ISSN: 0305-4608. DOI: 10.1088/0305-4608/6/11/005.

[62] Hugo Duminil-Copin, Vladas Sidoravicius, and Vincent Tassion. "Continuity of the Phase Transition for Planar Random-Cluster and Potts Models with $1 \leq q \leq 4$". In: *Communications in Mathematical Physics* 349.1 (Jan. 2017), pp. 47–107. ISSN: 1432-0916. DOI: 10.1007/s00220-016-2759-8.

[63] Michael Elad. "Pursuit Algorithms – Practice". In: *Sparse and Redundant Representations: From Theory to Applications in Signal and Image Processing*. Ed. by Michael Elad. New York, NY: Springer, 2010, pp. 35–54. ISBN: 978-1-4419-7011-4. DOI: 10.1007/978-1-4419-7011-4_3.

[64] Michael Elad. "Uniqueness and Uncertainty". In: *Sparse and Redundant Representations: From Theory to Applications in Signal and Image Processing*. Ed. by Michael Elad. New York, NY: Springer, 2010, pp. 17–33. ISBN: 978-1-4419-7011-4. DOI: 10.1007/978-1-4419-7011-4_2.

[65] Kai-Tai Fang, Runze Li, and Agus Sudjianto. *Design and Modeling for Computer Experiments*. New York: Chapman and Hall/CRC, Oct. 2005. ISBN: 978-0-429-14376-2. DOI: 10.1201/9781420034899.

[66] Julian J. Faraway. *Linear Models with Python*. CRC Press, Jan. 2021. ISBN: 978-1-351-05339-6.

[67] A. Fernández-Caballero et al. "Short-Range Order in High Entropy Alloys: Theoretical Formulation and Application to Mo-Nb-Ta-V-W System". In: *Journal of Phase Equilibria and Diffusion* 38.4 (Aug. 2017), pp. 391–403. ISSN: 1863-7345. DOI: 10.1007/s11669-017-0582-3.

[68] Yuval Filmus. "An Orthogonal Basis for Functions over a Slice of the Boolean Hypercube". In: *The Electronic Journal of Combinatorics* (Feb. 2016), P1.23–P1.23. ISSN: 1077-8926. DOI: 10.37236/4567.

[69] A. Finel and F. Ducastelle. "On the Phase Diagram of the FCC Ising Model with Antiferromagnetic First-Neighbour Interactions". In: *Europhysics Letters (EPL)* 1.3 (Feb. 1986), pp. 135–140. ISSN: 0295-5075. DOI: 10.1209/0295-5075/1/3/007.

[70] D. De Fontaine. "Cluster Approach to Order-Disorder Transformations in Alloys". In: *Solid State Physics*. Ed. by HENRY Ehrenreich and DAVID Turnbull. Vol. 47. Academic Press, Jan. 1994, pp. 33–176. DOI: 10.1016/S0081-1947(08)60639-6.

[71] Didier De Fontaine. *Principles of Classical Thermodynamics: Applied to Materials Science*. World Scientific Publishing Company Pte. Limited, 2019. ISBN: 978-981-322-268-7.

[72] Daan Frenkel and Berend Smit. *Understanding Molecular Simulation: From Algorithms to Applications*. Elsevier, Oct. 2001. ISBN: 978-0-08-051998-2.

[73] J. Friedman, T. Hastie, and R. Tibshirani. "A Note on the Group Lasso and a Sparse Group Lasso". In: *arXiv:1001.0736 [math, stat]* (Jan. 2010). arXiv: 1001.0736 [math, stat].

[74] "Representation Theory". In: *Young Tableaux: With Applications to Representation Theory and Geometry*. Ed. by William Fulton. London Mathematical Society Student Texts. Cambridge: Cambridge University Press, 1996, pp. 79–82. ISBN: 978-0-521-56724-4. DOI: 10.1017/CBO9780511626241.010.

[75] Marylou Gabrié, Grant M. Rotskoff, and Eric Vanden-Eijnden. "Adaptive Monte Carlo Augmented with Normalizing Flows". In: *Proceedings of the National Academy of Sciences* 119.10 (Mar. 2022), e2109420119. DOI: 10.1073/pnas.2109420119.

[76] Michael C. Gao et al. "High-Entropy Functional Materials". In: *Journal of Materials Research* 33.19 (Oct. 2018), pp. 3138–3155. ISSN: 0884-2914, 2044-5326. DOI: 10.1557/jmr.2018.323.

[77] Estefanía Garijo del Río, Jens Jørgen Mortensen, and Karsten Wedel Jacobsen. "Local Bayesian Optimizer for Atomic Structures". In: *Physical Review B* 100.10 (Sept. 2019), p. 104103. DOI: 10.1103/PhysRevB.100.104103.

[78] Easo P. George, Dierk Raabe, and Robert O. Ritchie. "High-Entropy Alloys". In: *Nature Reviews Materials* 4.8 (Aug. 2019), pp. 515–534. ISSN: 2058-8437. DOI: 10.1038/s41578-019-0121-4.

[79] A. Gonis and J. W. Garland. "Multishell Method: Exact Treatment of a Cluster in an Effective Medium". In: *Physical Review B* 16.6 (Sept. 1977), pp. 2424–2436. DOI: 10.1103/PhysRevB.16.2424.

[80] A. Gonis et al. "Electronic Structure, Alloy Phase Stability and Phase Diagrams". In: *Journal of the Less Common Metals* 168.1 (Feb. 1991), pp. 127–144. ISSN: 0022-5088. DOI: 10.1016/0022-5088(91)90040-B.

[81] Paul Gordon. *Principles of Phase Diagrams in Materials Systems*. McGraw-Hill, 1968. ISBN: 978-0-07-023793-3.

[82] L. Greengard and V. Rokhlin. "A Fast Algorithm for Particle Simulations". In: *Journal of Computational Physics* 135.2 (Aug. 1997), pp. 280–292. ISSN: 0021-9991. DOI: 10.1006/jcph.1997.5706.

[83] Karol Gregor et al. "Deep AutoRegressive Networks". In: *Proceedings of the 31st International Conference on Machine Learning*. PMLR, June 2014, pp. 1242–1250.

[84] Chong Gu. *Smoothing Spline ANOVA Models*. Second. Springer Series in Statistics. New York, NY: Springer, 2013. ISBN: 978-1-4614-5368-0.

[85] B. L. Gyorffy and G. M. Stocks. "Concentration Waves and Fermi Surfaces in Random Metallic Alloys". In: *Physical Review Letters* 50.5 (Jan. 1983), pp. 374–377. DOI: 10.1103/PhysRevLett.50.374.

[86] M. Hamada and C. F. J. Wu. "Analysis of Designed Experiments with Complex Aliasing". In: *Journal of Quality Technology* 24.3 (July 1992), pp. 130–137. ISSN: 0022-4065. DOI: 10.1080/00224065.1992.11979383.

[87] Gus L. W. Hart and Rodney W. Forcade. "Generating Derivative Structures from Multilattices: Algorithm and Application to Hcp Alloys". In: *Physical Review B* 80.1 (July 2009), p. 014120. DOI: 10.1103/PhysRevB.80.014120.

[88] Gus L. W. Hart et al. "Evolutionary Approach for Determining First-Principles Hamiltonians". In: *Nature Materials* 4.5 (May 2005), pp. 391–394. ISSN: 1476-4660. DOI: 10.1038/nmat1374.

[89] Gus L. W. Hart et al. "Machine Learning for Alloys". In: *Nature Reviews Materials* 6.8 (Aug. 2021), pp. 730–755. ISSN: 2058-8437. DOI: 10.1038/s41578-021-00340-w.

[90] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. "Linear Methods for Classification". In: *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Ed. by Trevor Hastie, Robert Tibshirani, and Jerome Friedman. Springer Series in Statistics. New York, NY: Springer, 2009, pp. 101–137. ISBN: 978-0-387-84858-7. DOI: 10.1007/978-0-387-84858-7_4.

[91] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. "Model Assessment and Selection". In: *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Ed. by Trevor Hastie, Robert Tibshirani, and Jerome Friedman. Springer Series in Statistics. New York, NY: Springer, 2009, pp. 219–259. ISBN: 978-0-387-84858-7. DOI: 10.1007/978-0-387-84858-7_7.

[92] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Series in Statistics. New York: Springer-Verlag, 2001. ISBN: 978-0-387-21606-5. DOI: 10.1007/978-0-387-21606-5.

[93] Trevor Hastie, Robert Tibshirani, and Martin Wainwright. *Statistical Learning with Sparsity: The Lasso and Generalizations*. New York: Chapman and Hall/CRC, May 2015. ISBN: 978-0-429-17158-1. DOI: 10.1201/b18401.

[94] David Hicks et al. "AFLOW-XtalFinder: A Reliable Choice to Identify Crystalline Prototypes". In: *npj Computational Materials* 7.1 (Feb. 2021), pp. 1–20. ISSN: 2057-3960. DOI: 10.1038/s41524-020-00483-4.

[95] Jonathan Ho, Ajay Jain, and Pieter Abbeel. "Denoising Diffusion Probabilistic Models". In: *Advances in Neural Information Processing Systems*. Vol. 33. Curran Associates, Inc., 2020, pp. 6840–6851.

[96] Giles Hooker. "Generalized Functional ANOVA Diagnostics for High-Dimensional Functions of Dependent Variables". In: *Journal of Computational and Graphical Statistics* 16.3 (2007), pp. 709–732. ISSN: 1061-8600.

[97] Jianhua Z. Huang. "Projection Estimation in Multiple Regression with Application to Functional Anova Models". In: *The Annals of Statistics* 26.1 (1998), pp. 242–272. ISSN: 0090-5364.

[98] Wenxuan Huang et al. "An L_0L_1-Norm Compressive Sensing Paradigm for the Construction of Sparse Predictive Lattice Models Using Mixed Integer Quadratic Programming". In: *arXiv:1807.10753 [cond-mat, physics:physics]* (July 2018). arXiv: 1807.10753 [cond-mat, physics:physics].

[99] Wenxuan Huang et al. "Construction of Ground-State Preserving Sparse Lattice Models for Predictive Materials Simulations". In: *npj Computational Materials* 3.1 (Aug. 2017), pp. 1–9. ISSN: 2057-3960. DOI: 10.1038/s41524-017-0032-0.

[100] Wenxuan Huang et al. "Finding and Proving the Exact Ground State of a Generalized Ising Model by Convex Optimization and MAX-SAT". In: *Physical Review B* 94.13 (Oct. 2016), p. 134424. DOI: 10.1103/PhysRevB.94.134424.

[101] Bertrand Iooss and Paul Lemaître. "A Review on Global Sensitivity Analysis Methods". In: *Uncertainty Management in Simulation-Optimization of Complex Systems: Algorithms and Applications*. Ed. by Gabriella Dellino and Carlo Meloni. Operations Research/Computer Science Interfaces Series. Boston, MA: Springer US, 2015, pp. 101–122. ISBN: 978-1-4899-7547-8. DOI: 10.1007/978-1-4899-7547-8_5.

[102] Laurent Jacob, Guillaume Obozinski, and Jean-Philippe Vert. "Group Lasso with Overlap and Graph Lasso". In: *Proceedings of the 26th Annual International Conference on Machine Learning - ICML '09*. Montreal, Quebec, Canada: ACM Press, 2009, pp. 1–8. ISBN: 978-1-60558-516-1. DOI: 10.1145/1553374.1553431.

[103] Karsten Wedel Jacobsen, Jakob Schiøtz, and Peter Mahler Larsen. "Rich Ground-State Chemical Ordering in Nanoparticles: Exact Solution of a Model for Ag-Au Clusters". In: *Physical Review Letters* 120.25 (June 2018), p. 256101. DOI: 10.1103/PhysRevLett.120.256101.

[104] Zinab Jadidi et al. "Ab-initio study of short-range-ordering in vanadium-based disordered rocksalt structures". In: ("in prep. 2022").

[105] Anubhav Jain et al. "Commentary: The Materials Project: A Materials Genome Approach to Accelerating Materials Innovation". In: *APL Materials* 1.1 (July 2013), p. 011002. DOI: 10.1063/1.4812323.

[106] Anubhav Jain et al. "Formation Enthalpies by Mixing GGA and GGA + U Calculations". In: *Physical Review B* 84.4 (July 2011), p. 045115. DOI: 10.1103/PhysRevB.84.045115.

[107] Svante Janson and Professor of Mathematics Svante Janson. *Gaussian Hilbert Spaces*. Cambridge University Press, June 1997. ISBN: 978-0-521-56128-0.

[108] Huiwen Ji et al. "Hidden Structural and Chemical Order Controls Lithium Transport in Cation-Disordered Oxides for Rechargeable Batteries". In: *Nature Communications* 10.1 (Feb. 2019), p. 592. ISSN: 2041-1723. DOI: 10.1038/s41467-019-08490-w.

[109] Chao Jiang and Blas P. Uberuaga. "Efficient Ab Initio Modeling of Random Multicomponent Alloys". In: *Physical Review Letters* 116.10 (Mar. 2016), p. 105501. DOI: 10.1103/PhysRevLett.116.105501.

[110] Tian Jin et al. "Mechanochemical-Assisted Synthesis of High-Entropy Metal Nitride via a Soft Urea Strategy". In: *Advanced Materials* 30.23 (2018), p. 1707512. ISSN: 1521-4095. DOI: 10.1002/adma.201707512.

[111] Makoto Kaburagi and Junjiro Kanamori. "Ground State Structure of Triangular Lattice Gas Model with up to 3rd Neighbor Interactions". In: *Journal of the Physical Society of Japan* 44.3 (Mar. 1978), pp. 718–727. ISSN: 0031-9015. DOI: 10.1143/JPSJ.44.718.

[112] Mehran Kardar. *Statistical Physics of Particles*. Cambridge University Press, June 2007. ISBN: 978-1-139-46487-1.

[113] A. Ya Khinchin. *Mathematical Foundations of Statistical Mechanics*. Courier Corporation, Jan. 2013. ISBN: 978-0-486-13873-2.

[114] Ryoichi Kikuchi. "A Theory of Cooperative Phenomena". In: *Physical Review* 81.6 (Mar. 1951), pp. 988–1003. DOI: 10.1103/PhysRev.81.988.

[115] Diederik P. Kingma and Max Welling. "An Introduction to Variational Autoencoders". In: *Foundations and Trends® in Machine Learning* 12.4 (Nov. 2019), pp. 307–392. ISSN: 1935-8237, 1935-8245. DOI: 10.1561/2200000056.

[116] Diederik P. Kingma et al. *Improving Variational Inference with Inverse Autoregressive Flow*. Jan. 2017. DOI: 10.48550/arXiv.1606.04934. arXiv: 1606.04934 [cs, stat].

[117] Diederik P. Kingma et al. *Variational Diffusion Models*. June 2022. DOI: 10.48550/arXiv.2107.00630. arXiv: 2107.00630 [cs, stat].

[118] Daniil A. Kitchaev et al. "Design Principles for High Transition Metal Capacity in Disordered Rocksalt Li-ion Cathodes". In: *Energy & Environmental Science* 11.8 (Aug. 2018), pp. 2159–2171. ISSN: 1754-5706. DOI: 10.1039/C8EE00816G.

[119] David A. Kofke and Eduardo D. Glandt. "Monte Carlo Simulation of Multicomponent Equilibria in a Semigrand Canonical Ensemble". In: *Molecular Physics* 64.6 (Aug. 1988), pp. 1105–1131. ISSN: 0026-8976. DOI: 10.1080/00268978800100743.

[120] W. Kohn and L. J. Sham. "Self-Consistent Equations Including Exchange and Correlation Effects". In: *Physical Review* 140.4A (Nov. 1965), A1133–A1138. DOI: 10.1103/PhysRev.140.A1133.

[121] Daphne Koller and Nir Friedman. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, July 2009. ISBN: 978-0-262-25835-7.

[122] G. Kresse and J. Furthmüller. "Efficiency of Ab-Initio Total Energy Calculations for Metals and Semiconductors Using a Plane-Wave Basis Set". In: *Computational Materials Science* 6.1 (July 1996), pp. 15–50. ISSN: 0927-0256. DOI: 10.1016/0927-0256(96)00008-0.

[123] G. Kresse and D. Joubert. "From Ultrasoft Pseudopotentials to the Projector Augmented-Wave Method". In: *Physical Review B* 59.3 (Jan. 1999), pp. 1758–1775. DOI: 10.1103/PhysRevB.59.1758.

[124] Jonathan Kuck et al. "Belief Propagation Neural Networks". In: *Advances in Neural Information Processing Systems*. Vol. 33. Curran Associates, Inc., 2020, pp. 667–678.

[125] Akai Kurbanovich Murtazaev and Albert Babaevich Babaev. "Phase Transitions and Critical Phenomena in a Three-Dimensional Site-Diluted Potts Model". In: *Journal of Magnetism and Magnetic Materials* 324.22 (Nov. 2012), pp. 3870–3875. ISSN: 0304-8853. DOI: 10.1016/j.jmmm.2012.06.038.

[126] D. P. Landau. "Critical Behavior of a Bcc Ising Antiferromagnet in a Magnetic Field". In: *Physical Review B* 16.9 (Nov. 1977), pp. 4164–4170. DOI: 10.1103/PhysRevB.16.4164.

[127] David Landau and Kurt Binder. *A Guide to Monte Carlo Simulations in Statistical Physics*. Cambridge University Press, July 2021. ISBN: 978-1-108-49014-6.

[128] Peter Mahler Larsen et al. "Alloy Design as an Inverse Problem of Cluster Expansion Models". In: *Acta Materialia* 139 (Oct. 2017), pp. 254–260. ISSN: 1359-6454. DOI: 10.1016/j.actamat.2017.08.008.

[129] Eunseok Lee and Kristin A. Persson. "Revealing the Coupled Cation Interactions behind the Electrochemical Profile of LixNi0.5Mn1.5O4". In: *Energy & Environmental Science* 5.3 (Mar. 2012), pp. 6047–6051. ISSN: 1754-5706. DOI: 10.1039/C2EE03068C.

[130] Eunseok Lee, Friedrich B. Prinz, and Wei Cai. "Enhancing Ionic Conductivity of Bulk Single-Crystal Yttria-Stabilized Zirconia by Tailoring Dopant Distribution". In: *Physical Review B* 83.5 (Feb. 2011), p. 052301. DOI: 10.1103/PhysRevB.83.052301.

[131] Zhidong Leong and Teck Leong Tan. "Robust Cluster Expansion of Multicomponent Systems Using Structured Sparsity". In: *Physical Review B* 100.13 (Oct. 2019), p. 134108. DOI: 10.1103/PhysRevB.100.134108.

[132] D. Lerch et al. "UNCLE: A Code for Constructing Cluster Expansions for Arbitrary Lattices with Minimal User-Input". In: *Modelling and Simulation in Materials Science and Engineering* 17.5 (June 2009), p. 055003. ISSN: 0965-0393. DOI: 10.1088/0965-0393/17/5/055003.

[133] Michael Lim and Trevor Hastie. "Learning Interactions via Hierarchical Group-Lasso Regularization". In: *Journal of Computational and Graphical Statistics* 24.3 (July 2015), pp. 627–654. ISSN: 1061-8600. DOI: 10.1080/10618600.2014.938812.

[134] Dong Liu, Ragnar Thobaben, and Lars K. Rasmussen. "Region-Based Energy Neural Network for Approximate Inference". In: *arXiv:2006.09927 [cs, stat]* (June 2020). arXiv: 2006.09927 [cs, stat].

[135] Jian Liu, Maria V. Fernández-Serra, and Philip B. Allen. "Special Quasiordered Structures: Role of Short-Range Order in the Semiconductor Alloy $(GaN)_{1-x}(ZnO)_x$". In: *Physical Review B* 93.5 (Feb. 2016), p. 054207. DOI: 10.1103/PhysRevB.93.054207.

[136] Zhengyan Lun et al. "Cation-Disordered Rocksalt-Type High-Entropy Cathodes for Li-ion Batteries". In: *Nature Materials* 2 (Oct. 2020), pp. 1–8. ISSN: 1476-4660. DOI: 10.1038/s41563-020-00816-0.

[137] Yury Lysogorskiy et al. "Performant Implementation of the Atomic Cluster Expansion (PACE): Application to Copper and Silicon". In: *arXiv:2103.00814 [cond-mat, physics:physics]* (Mar. 2021). arXiv: 2103.00814 [cond-mat, physics:physics].

[138] Yishay Mansour. "Learning Boolean Functions via the Fourier Transform". In: *Theoretical Advances in Neural Computation and Learning*. Ed. by Vwani Roychowdhury, Kai-Yeung Siu, and Alon Orlitsky. Boston, MA: Springer US, 1994, pp. 391–424. ISBN: 978-1-4615-2696-4. DOI: 10.1007/978-1-4615-2696-4_11.

[139] Samuel W. McAlpine, Julie V. Logan, and Michael P. Short. "Predicting Single Phase Stability and Segregation in the NbMoTaTi–(W,V) High Entropy Alloy System with the Vacancy Exchange Potential". In: *Scripta Materialia* 191 (Jan. 2021), pp. 29–33. ISSN: 1359-6462. DOI: 10.1016/j.scriptamat.2020.08.043.

[140] P. McCullagh and J. A. Nelder. *Generalized Linear Models*. Second. New York: Routledge, Jan. 2019. ISBN: 978-0-203-75373-6. DOI: 10.1201/9780203753736.

[141] B. McNaughton et al. "Boosting Monte Carlo Simulations of Spin Glasses Using Autoregressive Neural Networks". In: *Physical Review E* 101.5 (May 2020), p. 053312. DOI: 10.1103/PhysRevE.101.053312.

[142] Marc Mézard and Andrea Montanari. *Information, Physics, and Computation*. OUP Oxford, Jan. 2009. ISBN: 978-0-19-857083-7.

[143] D. B. Miracle and O. N. Senkov. "A Critical Review of High Entropy Alloys and Related Concepts". In: *Acta Materialia* 122 (Jan. 2017), pp. 448–511. ISSN: 1359-6454. DOI: 10.1016/j.actamat.2016.08.081.

[144] Tetsuo Mohri. "Cluster Variation Method". In: *JOM* 65.11 (Nov. 2013), pp. 1510–1522. ISSN: 1543-1851. DOI: 10.1007/s11837-013-0738-5.

[145] Jacob I. Monroe and Vincent K. Shen. "Learning Efficient, Collective Monte Carlo Moves with Variational Autoencoders". In: *Journal of Chemical Theory and Computation* 18.6 (June 2022), pp. 3622–3636. ISSN: 1549-9618. DOI: 10.1021/acs.jctc.2c00110.

[146] E. Mossel, R. O'Donnell, and K. Oleszkiewicz. "Noise Stability of Functions with Low Influences: Invariance and Optimality". In: *46th Annual IEEE Symposium on Foundations of Computer Science (FOCS'05)*. Oct. 2005, pp. 21–30. DOI: 10.1109/SFCS.2005.53.

[147] Tim Mueller. "Comment on "Cluster Expansion and the Configurational Theory of Alloys"". In: *Physical Review B* 95.21 (June 2017), p. 216201. DOI: 10.1103/PhysRevB.95.216201.

[148] Tim Mueller and Gerbrand Ceder. "Bayesian Approach to Cluster Expansions". In: *Physical Review B* 80.2 (July 2009), p. 024103. DOI: 10.1103/PhysRevB.80.024103.

[149] Tim Mueller and Gerbrand Ceder. "Exact Expressions for Structure Selection in Cluster Expansions". In: *Physical Review B* 82.18 (Nov. 2010), p. 184107. DOI: 10.1103/PhysRevB.82.184107.

[150] Chiraag M. Nataraj, Axel van de Walle, and Amit Samanta. "Temperature-Dependent Configurational Entropy Calculations for Refractory High-Entropy Alloys". In: *Journal of Phase Equilibria and Diffusion* 42.5 (Oct. 2021), pp. 571–577. ISSN: 1863-7345. DOI: 10.1007/s11669-021-00879-9.

[151] Anirudh Raju Natarajan and Anton Van der Ven. "Machine-Learning the Configurational Energy of Multicomponent Crystalline Solids". In: *npj Computational Materials* 4.1 (Nov. 2018), p. 56. ISSN: 2057-3960. DOI: 10.1038/s41524-018-0110-y.

[152] Lance J. Nelson et al. "Cluster Expansion Made Easy with Bayesian Compressive Sensing". In: *Physical Review B* 88.15 (Oct. 2013), p. 155105. DOI: 10.1103/PhysRevB.88.155105.

[153] Lance J. Nelson et al. "Compressive Sensing as a Paradigm for Building Physics Models". In: *Physical Review B* 87.3 (Jan. 2013), p. 035125. DOI: 10.1103/PhysRevB.87.035125.

[154] M. Nespolo. "Lattice versus Structure, Dimensionality versus Periodicity: A Crystallographic Babel?" In: *Journal of Applied Crystallography* 52.2 (Apr. 2019), pp. 451–456. ISSN: 1600-5767. DOI: 10.1107/S1600576719000463.

[155] Andrew H. Nguyen et al. "Robustness of the Cluster Expansion: Assessing the Roles of Relaxation and Numerical Error". In: *Physical Review B* 96.1 (July 2017), p. 014107. DOI: 10.1103/PhysRevB.96.014107.

[156] "Boolean Functions and the Fourier Expansion". In: *Analysis of Boolean Functions*. Ed. by Ryan O'Donnell. Cambridge: Cambridge University Press, 2014, pp. 1–25. ISBN: 978-1-107-03832-5. DOI: 10.1017/CBO9781139814782.002.

[157] "Generalized Domains". In: *Analysis of Boolean Functions*. Ed. by Ryan O'Donnell. Cambridge: Cambridge University Press, 2014, pp. 197–239. ISBN: 978-1-107-03832-5. DOI: 10.1017/CBO9781139814782.009.

[158] "Spectral Structure and Learning". In: *Analysis of Boolean Functions*. Ed. by Ryan O'Donnell. Cambridge: Cambridge University Press, 2014, pp. 54–78. ISBN: 978-1-107-03832-5. DOI: 10.1017/CBO9781139814782.004.

[159] Eric C. O'Quinn et al. "Predicting Short-Range Order and Correlated Phenomena in Disordered Crystalline Materials". In: *Science Advances* 6.35 (Aug. 2020), eabc2758. DOI: 10.1126/sciadv.abc2758.

[160] Shyue Ping Ong et al. "Li-Fe-P-O2 Phase Diagram from First Principles Calculations". In: *Chemistry of Materials* 20.5 (Mar. 2008), pp. 1798–1807. ISSN: 0897-4756. DOI: 10.1021/cm702327g.

[161] Shyue Ping Ong et al. "Python Materials Genomics (Pymatgen): A Robust, Open-Source Python Library for Materials Analysis". In: *Computational Materials Science* 68 (Feb. 2013), pp. 314–319. ISSN: 0927-0256. DOI: 10.1016/j.commatsci.2012.10.028.

[162] Yoshitsugu Oono. *Perspectives on Statistical Thermodynamics*. Cambridge University Press, Dec. 2017. ISBN: 978-1-108-36524-6.

[163] Corey Oses, Cormac Toher, and Stefano Curtarolo. "High-Entropy Ceramics". In: *Nature Reviews Materials* 5.4 (Apr. 2020), pp. 295–309. ISSN: 2058-8437. DOI: 10.1038/s41578-019-0170-8.

[164] F. Otto et al. "Relative Effects of Enthalpy and Entropy on the Phase Stability of Equiatomic High-Entropy Alloys". In: *Acta Materialia* 61.7 (Apr. 2013), pp. 2628–2638. ISSN: 1359-6454. DOI: 10.1016/j.actamat.2013.01.042.

[165] Bin Ouyang et al. "Effect of Fluorination on Lithium Transport and Short-Range Order in Disordered-Rocksalt-Type Lithium-Ion Battery Cathodes". In: *Advanced Energy Materials* 10.10 (2020), p. 1903240. ISSN: 1614-6840. DOI: 10.1002/aenm.201903240.

[166] V. Ozoliņš, B. Sadigh, and M. Asta. "Effects of Vibrational Entropy on the Al–Si Phase Diagram". In: *Journal of Physics: Condensed Matter* 17.13 (Mar. 2005), pp. 2197–2210. ISSN: 0953-8984. DOI: 10.1088/0953-8984/17/13/017.

[167] V. Ozoliņš, C. Wolverton, and Alex Zunger. "First-Principles Theory of Vibrational Effects on the Phase Stability of Cu-Au Compounds and Alloys". In: *Physical Review B* 58.10 (Sept. 1998), R5897–R5900. DOI: 10.1103/PhysRevB.58.R5897.

[168] David Packwood et al. "A Universal Preconditioner for Simulating Condensed Phase Materials". In: *The Journal of Chemical Physics* 144.16 (Apr. 2016), p. 164109. ISSN: 0021-9606. DOI: 10.1063/1.4947024.

[169] Fabian Pedregosa et al. "Scikit-Learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12.85 (2011), pp. 2825–2830. ISSN: 1533-7928.

[170] Zongrui Pei et al. "Error Controlling of the Combined Cluster-Expansion and Wang–Landau Monte-Carlo Method and Its Application to FeCo". In: *Computer Physics Communications* 235 (Feb. 2019), pp. 95–101. ISSN: 0010-4655. DOI: 10.1016/j.cpc.2018.09.017.

[171] Zongrui Pei et al. "Statistics of the NiCoCr Medium-Entropy Alloy: Novel Aspects of an Old Puzzle". In: *npj Computational Materials* 6.1 (Aug. 2020), pp. 1–6. ISSN: 2057-3960. DOI: 10.1038/s41524-020-00389-1.

[172] Julio L. Peixoto. "A Property of Well-Formulated Polynomial Regression Models". In: *The American Statistician* 44.1 (1990), pp. 26–30. ISSN: 0003-1305. DOI: 10.2307/2684952.

[173] Luca Peliti. *Statistical Mechanics in a Nutshell*. Princeton University Press, Aug. 2011. ISBN: 978-1-4008-3936-0.

[174] Alessandro Pelizzola. "Cluster Variation Method in Statistical Physics and Probabilistic Graphical Models". In: *Journal of Physics A: Mathematical and General* 38.33 (Aug. 2005), R309–R339. ISSN: 0305-4470. DOI: 10.1088/0305-4470/38/33/R01.

[175] John P. Perdew, Kieron Burke, and Matthias Ernzerhof. "Generalized Gradient Approximation Made Simple". In: *Physical Review Letters* 77.18 (Oct. 1996), pp. 3865–3868. DOI: 10.1103/PhysRevLett.77.3865.

[176] R. B. Potts. "Some Generalized Order-Disorder Transformations". In: *Mathematical Proceedings of the Cambridge Philosophical Society* 48.1 (Jan. 1952), pp. 106–109. ISSN: 1469-8064, 0305-0041. DOI: 10.1017/S0305004100027419.

[177] *Prisms-Center/CASMcode*. PRISMS Center. Mar. 2022.

[178] Danilo Rezende and Shakir Mohamed. "Variational Inference with Normalizing Flows". In: *Proceedings of the 32nd International Conference on Machine Learning*. PMLR, June 2015, pp. 1530–1538.

[179] Danilo Jimenez Rezende and Fabio Viola. *Taming VAEs*. Oct. 2018. DOI: 10.48550/arXiv.1810.00597. arXiv: 1810.00597 [cs, stat].

[180] John A. Rice. *Mathematical Statistics and Data Analysis*. Cengage Learning, 2007. ISBN: 978-0-534-39942-9.

[181] William D. Richards et al. "Design of Li1+2xZn1-xPS4, a New Lithium Ion Conductor". In: *Energy & Environmental Science* 9.10 (Oct. 2016), pp. 3272–3278. ISSN: 1754-5706. DOI: 10.1039/C6EE02094A.

[182] William D. Richards et al. "Fluorination of Lithium-Excess Transition Metal Oxide Cathode Materials". In: *Advanced Energy Materials* 8.5 (2018), p. 1701533. ISSN: 1614-6840. DOI: 10.1002/aenm.201701533.

[183] Lars Ruthotto and Eldad Haber. "An Introduction to Deep Generative Modeling". In: *GAMM-Mitteilungen* 44.2 (2021), e202100008. ISSN: 1522-2608. DOI: 10.1002/gamm.202100008.

[184] A. M. Saitta, S. de Gironcoli, and S. Baroni. "Structural and Electronic Properties of a Wide-Gap Quaternary Solid Solution: \(Zn, Mg\) \(S, Se\)". In: *Physical Review Letters* 80.22 (June 1998), pp. 4939–4942. DOI: 10.1103/PhysRevLett.80.4939.

[185] V. Saltas et al. "Modelling Solid Solutions with Cluster Expansion, Special Quasirandom Structures, and Thermodynamic Approaches". In: *Applied Physics Reviews* 4.4 (Oct. 2017), p. 041301. DOI: 10.1063/1.4999129.

[186] Andrea Saltelli et al. *Global Sensitivity Analysis: The Primer*. John Wiley & Sons, Feb. 2008. ISBN: 978-0-470-72517-7.

[187] J. M. Sanchez. "Cluster Expansion and the Configurational Theory of Alloys". In: *Physical Review B* 81.22 (June 2010), p. 224202. DOI: 10.1103/PhysRevB.81.224202.

[188] J. M. Sanchez. "Cluster Expansions and the Configurational Energy of Alloys". In: *Physical Review B* 48.18 (Nov. 1993), pp. 14013–14015. ISSN: 0163-1829, 1095-3795. DOI: 10.1103/PhysRevB.48.14013.

[189] J. M. Sanchez. "Foundations and Practical Implementations of the Cluster Expansion". In: *Journal of Phase Equilibria and Diffusion* 38.3 (June 2017), pp. 238–251. ISSN: 1863-7345. DOI: 10.1007/s11669-017-0521-3.

[190] J. M. Sanchez. "Renormalized Interactions in Truncated Cluster Expansions". In: *Physical Review B* 99.13 (Apr. 2019), p. 134206. DOI: 10.1103/PhysRevB.99.134206.

[191] J. M. Sanchez. "Reply to "Comment on 'Cluster Expansion and the Configurational Theory of Alloys'"". In: *Physical Review B* 95.21 (June 2017), p. 216202. DOI: 10.1103/PhysRevB.95.216202.

[192] J. M. Sanchez, F. Ducastelle, and D. Gratias. "Generalized Cluster Description of Multicomponent Systems". In: *Physica A: Statistical Mechanics and its Applications* 128.1 (Nov. 1984), pp. 334–350. ISSN: 0378-4371. DOI: 10.1016/0378-4371(84)90096-7.

[193] J. M. Sanchez and T. Mohri. "Approximate Solutions to the Cluster Variation Free Energies by the Variable Basis Cluster Expansion". In: *Computational Materials Science* 122 (Sept. 2016), pp. 301–306. ISSN: 0927-0256. DOI: 10.1016/j.commatsci.2016.05.035.

[194] Víctor Garcia Satorras and Max Welling. "Neural Enhanced Belief Propagation on Factor Graphs". In: *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*. PMLR, Mar. 2021, pp. 685–693.

[195] Atsuto Seko, Yukinori Koyama, and Isao Tanaka. "Cluster Expansion Method for Multicomponent Systems Based on Optimal Selection of Structures for Density-Functional Theory Calculations". In: *Physical Review B* 80.16 (Oct. 2009), p. 165122. DOI: 10.1103/PhysRevB.80.165122.

[196] Atsuto Seko and Isao Tanaka. "Cluster Expansion of Multicomponent Ionic Systems with Controlled Accuracy: Importance of Long-Range Interactions in Heterovalent Ionic Systems". In: *Journal of Physics: Condensed Matter* 26.11 (Mar. 2014), p. 115403. ISSN: 0953-8984. DOI: 10.1088/0953-8984/26/11/115403. arXiv: 1309.2516.

[197] Atsuto Seko and Isao Tanaka. "Grouping of Structures for Cluster Expansion of Multicomponent Systems with Controlled Accuracy". In: *Physical Review B* 83.22 (June 2011), p. 224111. DOI: 10.1103/PhysRevB.83.224111.

[198] Atsuto Seko et al. "First-Principles Study of Cation Disordering in $MgAl_2O_4$ Spinel with Cluster Expansion and Monte Carlo Simulation". In: *Physical Review B* 73.9 (Mar. 2006), p. 094116. DOI: 10.1103/PhysRevB.73.094116.

[199] Atsuto Seko et al. "Structure and Stability of a Homologous Series of Tin Oxides". In: *Physical Review Letters* 100.4 (Jan. 2008), p. 045702. DOI: 10.1103/PhysRevLett.100.045702.

[200] James Sethna. *Statistical Mechanics: Entropy, Order Parameters and Complexity*. OUP Oxford, Apr. 2006. ISBN: 978-0-19-156621-9.

[201] Noah Simon and Robert Tibshirani. "STANDARDIZATION AND THE GROUP LASSO PENALTY". In: *Statistica Sinica* 22.3 (2012), pp. 983–1001. ISSN: 1017-0405.

[202] Noah Simon et al. "A Sparse-Group Lasso". In: *Journal of Computational and Graphical Statistics* 22.2 (Apr. 2013), pp. 231–245. ISSN: 1061-8600. DOI: 10.1080/10618600.2012.681250.

[203] Steven H. Simon. *The Oxford Solid State Basics*. OUP Oxford, June 2013. ISBN: 978-0-19-150210-1.

[204] Prashant Singh, A. V. Smirnov, and D. D. Johnson. "Atomic Short-Range Order and Incipient Long-Range Order in High-Entropy Alloys". In: *Physical Review B* 91.22 (June 2015), p. 224204. DOI: 10.1103/PhysRevB.91.224204.

[205] Jasper Snoek, Hugo Larochelle, and Ryan P Adams. "Practical Bayesian Optimization of Machine Learning Algorithms". In: *Advances in Neural Information Processing Systems*. Vol. 25. Curran Associates, Inc., 2012.

[206] I. M Sobol′. "Global Sensitivity Indices for Nonlinear Mathematical Models and Their Monte Carlo Estimates". In: *Mathematics and Computers in Simulation*. The Second IMACS Seminar on Monte Carlo Methods 55.1 (Feb. 2001), pp. 271–280. ISSN: 0378-4754. DOI: 10.1016/S0378-4754(00)00270-6.

[207] Jascha Sohl-Dickstein et al. "Deep Unsupervised Learning Using Nonequilibrium Thermodynamics". In: *Proceedings of the 32nd International Conference on Machine Learning*. PMLR, June 2015, pp. 2256–2265.

[208] Charles J. Stone. "The Use of Polynomial Splines and Their Tensor Products in Multivariate Function Estimation". In: *The Annals of Statistics* 22.1 (1994), pp. 118–171. ISSN: 0090-5364.

[209] Chuanxun Su et al. "Construction of Crystal Structure Prototype Database: Methods and Applications". In: *Journal of Physics: Condensed Matter* 29.16 (Mar. 2017), p. 165901. ISSN: 0953-8984. DOI: 10.1088/1361-648X/aa63cd.

[210] Fan-Yun Sun et al. "Equivariant Neural Network for Factor Graphs". In: *arXiv:2109.14218 [cs]* (Sept. 2021). arXiv: 2109.14218 [cs].

[211] Jianwei Sun, Adrienn Ruzsinszky, and John P. Perdew. "Strongly Constrained and Appropriately Normed Semilocal Density Functional". In: *Physical Review Letters* 115.3 (July 2015), p. 036402. DOI: 10.1103/PhysRevLett.115.036402.

[212] Christopher Sutton and Sergey V. Levchenko. "First-Principles Atomistic Thermodynamics and Configurational Entropy". In: *Frontiers in Chemistry* 8 (2020). ISSN: 2296-2646.

[213] Kazuhito Takeuchi, Ryohei Tanaka, and Koretaka Yuge. "New Wang-Landau Approach to Obtain Phase Diagrams for Multicomponent Alloys". In: *Physical Review B* 96.14 (Oct. 2017), p. 144202. DOI: 10.1103/PhysRevB.96.144202.

[214] J. C. Taylor. "Independence and Product Measures". In: *An Introduction to Measure and Probability*. Ed. by J. C. Taylor. New York, NY: Springer, 1997, pp. 86–136. ISBN: 978-1-4612-0659-0. DOI: 10.1007/978-1-4612-0659-0_3.

[215] P. D. Tepesch, G. D. Garbulsky, and G. Ceder. "Model for Configurational Thermodynamics in Ionic Systems". In: *Physical Review Letters* 74.12 (Mar. 1995), pp. 2272–2275. DOI: 10.1103/PhysRevLett.74.2272.

[216] Patrick D. Tepesch, Mark Asta, and Gerbrand Ceder. "Computation of Configurational Entropy Using Monte Carlo Probabilities in Cluster-Variation Method Entropy Expressions". In: *Modelling and Simulation in Materials Science and Engineering* 6.6 (Nov. 1998), pp. 787–797. ISSN: 0965-0393. DOI: 10.1088/0965-0393/6/6/009.

[217] Max Teubner. "Ground States of Classical One-Dimensional Lattice Models". In: *Physica A: Statistical Mechanics and its Applications* 169.3 (Dec. 1990), pp. 407–420. ISSN: 0378-4371. DOI: 10.1016/0378-4371(90)90111-5.

[218] John C. Thomas, Anirudh Raju Natarajan, and Anton Van der Ven. "Comparing Crystal Structures with Symmetry and Geometry". In: *npj Computational Materials* 7.1 (Oct. 2021), pp. 1–11. ISSN: 2057-3960. DOI: 10.1038/s41524-021-00627-0.

[219] Aidan P. Thompson et al. "LAMMPS - a Flexible Simulation Tool for Particle-Based Materials Modeling at the Atomic, Meso, and Continuum Scales". In: *Computer Physics Communications* 271 (Feb. 2022), p. 108171. ISSN: 0010-4655. DOI: 10.1016/j.cpc.2021.108171.

[220] Robert Tibshirani. "Regression Shrinkage and Selection via the Lasso". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 58.1 (1996), pp. 267–288. ISSN: 0035-9246.

[221] Ryan J. Tibshirani. "The Lasso Problem and Uniqueness". In: *Electronic Journal of Statistics* 7.0 (2013), pp. 1456–1490. ISSN: 1935-7524. DOI: 10.1214/13-EJS815.

[222] A. N. Tikhonov et al. *Numerical Methods for the Solution of Ill-Posed Problems*. Springer Science & Business Media, June 1995. ISBN: 978-0-7923-3583-2.

[223] Abdulnour Y. Toukmaji and John A. Board. "Ewald Summation Techniques in Perspective: A Survey". In: *Computer Physics Communications* 95.2 (June 1996), pp. 73–92. ISSN: 0010-4655. DOI: 10.1016/0010-4655(96)00016-1.

[224] Alexander Urban, Jinhyuk Lee, and Gerbrand Ceder. "The Configurational Space of Rocksalt-Type Oxides for High-Capacity Lithium Battery Electrodes". In: *Advanced Energy Materials* 4.13 (2014), p. 1400478. ISSN: 1614-6840. DOI: 10.1002/aenm.201400478.

[225] A. van de Walle. "A Complete Representation of Structure–Property Relationships in Crystals". In: *Nature Materials* 7.6 (June 2008), pp. 455–458. ISSN: 1476-4660. DOI: 10.1038/nmat2200.

[226] A. van de Walle, M. Asta, and G. Ceder. "The Alloy Theoretic Automated Toolkit: A User Guide". In: *Calphad* 26.4 (Dec. 2002), pp. 539–553. ISSN: 0364-5916. DOI: 10.1016/S0364-5916(02)80006-2.

[227] A. van de Walle and G. Ceder. "Automating First-Principles Phase Diagram Calculations". In: *Journal of Phase Equilibria* 23.4 (Aug. 2002), p. 348. ISSN: 1054-9714. DOI: 10.1361/105497102770331596.

[228] A. van de Walle and D. E. Ellis. "First-Principles Thermodynamics of Coherent Interfaces in Samarium-Doped Ceria Nanoscale Superlattices". In: *Physical Review Letters* 98.26 (June 2007), p. 266101. DOI: 10.1103/PhysRevLett.98.266101.

[229] A. van de Walle et al. "Efficient Stochastic Generation of Special Quasirandom Structures". In: *Calphad* 42 (Sept. 2013), pp. 13–18. ISSN: 0364-5916. DOI: 10.1016/j.calphad.2013.06.006.

[230] Axel van de Walle. "Multicomponent Multisublattice Alloys, Nonconfigurational Entropy and Other Additions to the Alloy Theoretic Automated Toolkit". In: *Calphad*. Tools for Computational Thermodynamics 33.2 (June 2009), pp. 266–278. ISSN: 0364-5916. DOI: 10.1016/j.calphad.2008.12.005.

[231] A. Van der Ven et al. "First-Principles Statistical Mechanics of Multicomponent Crystals". In: *Annual Review of Materials Research* 48.1 (July 2018), pp. 27–55. ISSN: 1531-7331. DOI: 10.1146/annurev-matsci-070317-124443.

[232] Anton Van der Ven et al. "Rechargeable Alkali-Ion Battery Materials: Theory and Computation". In: *Chemical Reviews* (Feb. 2020). ISSN: 0009-2665. DOI: 10.1021/acs.chemrev.9b00601.

[233] Pauli Virtanen et al. "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python". In: *Nature Methods* 17.3 (Mar. 2020), pp. 261–272. ISSN: 1548-7105. DOI: 10.1038/s41592-019-0686-2.

[234] Martin J. Wainwright and Michael I. Jordan. "Graphical Models, Exponential Families, and Variational Inference". In: *Foundations and Trends® in Machine Learning* 1.1–2 (Nov. 2008), pp. 1–305. ISSN: 1935-8237, 1935-8245. DOI: 10.1561/2200000001.

[235] Shayne Waldron. *An Introduction to Finite Tight Frames*. Applied and Numerical Harmonic Analysis. Birkhäuser Basel, 2018. ISBN: 978-0-8176-4814-5. DOI: 10.1007/978-0-8176-4815-2.

[236] Da Wang et al. "β-MnO2 as a Cathode Material for Lithium Ion Batteries from First Principles Calculations". In: *Physical Chemistry Chemical Physics* 15.23 (May 2013), pp. 9075–9083. ISSN: 1463-9084. DOI: 10.1039/C3CP50392E.

[237] Fugao Wang and D. P. Landau. "Efficient, Multiple-Range Random Walk Algorithm to Calculate the Density of States". In: *Physical Review Letters* 86.10 (Mar. 2001), pp. 2050–2053. DOI: 10.1103/PhysRevLett.86.2050.

[238] Hansheng Wang and Chenlei Leng. "A Note on Adaptive Group Lasso". In: *Computational Statistics & Data Analysis* 52.12 (Aug. 2008), pp. 5277–5286. ISSN: 0167-9473. DOI: 10.1016/j.csda.2008.05.006.

[239] Lei Wang, Thomas Maxisch, and Gerbrand Ceder. "Oxidation Energies of Transition Metal Oxides within the GGA + U Framework". In: *Physical Review B* 73.19 (May 2006), p. 195107. DOI: 10.1103/PhysRevB.73.195107.

[240] Mingqiu Wang and Guo-Liang Tian. "Adaptive Group Lasso for High-Dimensional Generalized Linear Models". In: *Statistical Papers* 60.5 (Oct. 2019), pp. 1469–1486. ISSN: 1613-9798. DOI: 10.1007/s00362-017-0882-z.

[241] Shun Wang et al. "Phase Transitions of Ferromagnetic Potts Models on the Simple Cubic Lattice". In: *Chinese Physics Letters* 31.7 (July 2014), p. 070503. ISSN: 0256-307X. DOI: 10.1088/0256-307X/31/7/070503.

[242] Larry Wasserman. *All of Nonparametric Statistics*. Springer Science & Business Media, Sept. 2006. ISBN: 978-0-387-30623-0.

[243] Joachim Weidmann. *Linear Operators in Hilbert Spaces*. First. Graduate Texts in Mathematics. New York, NY: Springer, 1980. ISBN: 978-1-4612-6029-5.

[244] Sam Wiseman and Yoon Kim. "Amortized Bethe Free Energy Minimization for Learning MRFs". In: *Advances in Neural Information Processing Systems*. Vol. 32. Curran Associates, Inc., 2019. URL: https://proceedings.neurips.cc/paper/2019/hash/dc554706afe4c72a60a25314cbaece80-Abstract.html (visited on 09/12/2022).

[245] C. Wolverton and Alex Zunger. "Cation and Vacancy Ordering in Li_xCoO_2". In: *Physical Review B* 57.4 (Jan. 1998), pp. 2242–2252. DOI: 10.1103/PhysRevB.57.2242.

[246] C. Wolverton and Alex Zunger. "Comparison of Two Cluster-Expansion Methods for the Energetics of Pd-V Alloys". In: *Physical Review B* 50.15 (Oct. 1994), pp. 10548–10560. DOI: 10.1103/PhysRevB.50.10548.

[247] C. Wolverton and Alex Zunger. "First-Principles Prediction of Vacancy Order-Disorder and Intercalation Battery Voltages in Li_xCoO_2". In: *Physical Review Letters* 81.3 (July 1998), pp. 606–609. DOI: 10.1103/PhysRevLett.81.606.

[248] Dian Wu, Riccardo Rossi, and Giuseppe Carleo. "Unbiased Monte Carlo Cluster Updates with Autoregressive Neural Networks". In: *Physical Review Research* 3.4 (Nov. 2021), p. L042024. DOI: 10.1103/PhysRevResearch.3.L042024.

[249] Dian Wu, Lei Wang, and Pan Zhang. "Solving Statistical Mechanics Using Variational Autoregressive Networks". In: *Physical Review Letters* 122.8 (Feb. 2019), p. 080602. DOI: 10.1103/PhysRevLett.122.080602.

[250] F. Y. Wu. "The Potts Model". In: *Reviews of Modern Physics* 54.1 (Jan. 1982), pp. 235–268. DOI: 10.1103/RevModPhys.54.235.

[251] Jun-Zhong Xie et al. "Machine Learning Force Field Aided Cluster Expansion Approach to Configurationally Disordered Materials: Critical Assessment of Training Set Selection and Size Convergence". In: *Journal of Chemical Theory and Computation* 18.6 (June 2022), pp. 3795–3804. ISSN: 1549-9618. DOI: 10.1021/acs.jctc.2c00017.

[252] Hao Xiong et al. "One-Shot Marginal MAP Inference in Markov Random Fields". en. In: *Proceedings of The 35th Uncertainty in Artificial Intelligence Conference*. PMLR, Aug. 2020, pp. 102–112. URL: https://proceedings.mlr.press/v115/xiong20a.html (visited on 09/12/2022).

[253] Julia H. Yang et al. "Approaches for Handling High-Dimensional Cluster Expansions of Ionic Systems". In: *npj Computational Materials* 8.1 (June 2022), pp. 1–11. ISSN: 2057-3960. DOI: 10.1038/s41524-022-00818-3.

[254] J.S. Yedidia, W.T. Freeman, and Y. Weiss. "Constructing Free-Energy Approximations and Generalized Belief Propagation Algorithms". In: *IEEE Transactions on Information Theory* 51.7 (July 2005), pp. 2282–2312. ISSN: 1557-9654. DOI: 10.1109/TIT.2005.850085.

[255] Ming Yuan and Yi Lin. "Model Selection and Estimation in Regression with Grouped Variables". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 68.1 (2006), pp. 49–67. ISSN: 1467-9868. DOI: 10.1111/j.1467-9868.2005.00532.x.

[256] Nikolai A. Zarkevich and D. D. Johnson. "Reliable First-Principles Alloy Thermodynamics via Truncated Cluster Expansions". In: *Physical Review Letters* 92.25 (June 2004), p. 255702. DOI: 10.1103/PhysRevLett.92.255702.

[257] Rui-Zhi Zhang et al. "Data-Driven Design of Ecofriendly Thermoelectric High-Entropy Sulfides". In: *Inorganic Chemistry* 57.20 (Oct. 2018), pp. 13027–13033. ISSN: 0020-1669. DOI: 10.1021/acs.inorgchem.8b02379.

[258] Xi Zhang and Marcel H. F. Sluiter. "Cluster Expansions for Thermodynamics and Kinetics of Multicomponent Alloys". In: *Journal of Phase Equilibria and Diffusion* 37.1 (Feb. 2016), pp. 44–52. ISSN: 1863-7345. DOI: 10.1007/s11669-015-0427-x.

[259] Peichen Zhong et al. "An $\ell_0\ell_2$-Norm Regularized Regression Model for Construction of Robust Cluster Expansions in Multicomponent Systems". In: *Physical Review B* 106.2 (July 2022), p. 024203. DOI: 10.1103/PhysRevB.106.024203.

[260] Ciyou Zhu et al. "Algorithm 778: L-BFGS-B: Fortran Subroutines for Large-Scale Bound-Constrained Optimization". In: *ACM Transactions on Mathematical Software* 23.4 (Dec. 1997), pp. 550–560. ISSN: 0098-3500. DOI: 10.1145/279232.279236.

[261] Hui Zou. "The Adaptive Lasso and Its Oracle Properties". In: *Journal of the American Statistical Association* 101.476 (Dec. 2006), pp. 1418–1429. ISSN: 0162-1459. DOI: 10.1198/016214506000000735.

[262] Hui Zou and Trevor Hastie. "Regularization and Variable Selection via the Elastic Net". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67.2 (2005), pp. 301–320. ISSN: 1467-9868. DOI: 10.1111/j.1467-9868.2005.00503.x.

[263] Alex Zunger et al. "Special Quasirandom Structures". In: *Physical Review Letters* 65.3 (July 1990), pp. 353–356. DOI: 10.1103/PhysRevLett.65.353.

[264] Yunxing Zuo et al. "Accelerating Materials Discovery with Bayesian Optimization and Graph Deep Learning". In: *Materials Today* 51 (Dec. 2021), pp. 126–135. ISSN: 1369-7021. DOI: 10.1016/j.mattod.2021.08.012.

[265] Yunxing Zuo et al. "Performance and Cost Assessment of Machine Learning Interatomic Potentials". In: *The Journal of Physical Chemistry A* 124.4 (Jan. 2020), pp. 731–745. ISSN: 1089-5639. DOI: 10.1021/acs.jpca.9b08723.

Appendix A

Notation & auxiliary definitions

A.1 Notation conventions

Notion conventions that are not formally introduced in the main text are listed here.

- We denote tensor products with $\bigotimes$, and Cartesian products with $\bigtimes$.
- We write $[N] = \{1, \ldots, N\}$ for the set of all positive integers up to N.
- We use multi-indices to index a specific sequence from a set of sequences. When the choices for each element in the sequences are finite we can write a multi-index as
  - $\boldsymbol{\alpha} \in \mathbb{N}^N_{<n}$ where each element $\alpha_i \in \{0, \ldots, n-1\}$.
  - $\boldsymbol{\alpha} \in \mathbb{N}^N_{<\boldsymbol{n}}$ where each element $\alpha_i \in \{0, \ldots, n_i - 1\}$, with $\boldsymbol{n} = (n_1, n_2, \ldots, n_N)$.
- The support of a multi-index is the set of indices of multi-index elements that are nonzero: $\mathrm{supp}(\boldsymbol{\alpha}) = \{i \ : \ \alpha_i \neq 0\}$.
- We usually use $\boldsymbol{\alpha}$ and $\boldsymbol{\eta}$ to denote multi-indices unless otherwise noted.
- We write the powerset of set X, as $\mathrm{P}(X) = \{Y \ : \forall Y \subseteq X\}$
- We will use the concept of an *orbit* generated by the operations of a symmetry group $\mathcal{G}$ on elements of a set X. The orbit generated from an element $A \in X$ by $\mathcal{G}$ is the set of all *symmetrically equivalent* elements to A,

$$B = \mathrm{Orb}_{\mathcal{G}}(A) = \{g \cdot A \ : \ \forall g \in \mathcal{G}\}$$

- We will write the multiplicity of an orbit B normalized by a stated unit N as $m_B = \frac{|B|}{N}$
- When dealing with a finite domain—i.e. a crystal supercell with N sites specified by a primitive cell and 3×3 integer supercell matrix—we will use $|B|$ to denote the total number of elements $A \in B$, where A is a subset of sites $A \subseteq [N]$.

- We use the notation $\mathcal{G}(X) = \{\text{Orb}_{\mathcal{G}}(x) : \forall x \in X\}$, for the set of all orbits generated by each of the elements x in the set X by symmetry operations of the group $\mathcal{G}$.
  - We write the set of all orbits generated from the elements in the powerset of a set X as $\mathcal{G}\,\text{P}(X)$.
- We use the binary relation symbol $D \sqsubset B$, to denote the relationship between an orbit D whose elements are subsets of an element in B. Meaning,

$$D \sqsubset B \iff \forall\, T \in D,\ \exists S \in B \text{ s.t. } T \subset S$$

- We will often work with orbits of multi-indices $\beta = \text{Orb}_{\mathcal{G}}(\boldsymbol{\alpha})$ and orbits of subsets of indices $S \subset [N]$, $B = \text{Orb}_{\mathcal{G}}(S)$. We will use the notation $B(\beta)$ to refer to the orbit of indices generated from the support of a multi-index $\boldsymbol{\alpha} \in \beta$, i.e. $B(\beta) = \text{Orb}_{\mathcal{G}}(\text{supp}(\boldsymbol{\alpha}))$.
- We usually use β and γ to denote orbits of multi-indices unless otherwise noted.
- We usually use B and D to denote orbits of subsets of indices $S \subset [N]$ for a given N, which we use to specify orbits of site clusters.

A.2 Site basis generation recipes

Any *recipe* can be used to obtain an initial basis set and subsequently converted to a *standard basis set* by a Gram-Schmidt orthonormalization process. Further, as discussed in the main text, all standard site basis sets are equivalent, such that whatever *recipe* was used to build the basis set has no theoretical importance. Nevertheless, we list the most commonly used *recipes* to generate site basis sets for fitting cluster expansions in practice, since these *recipes* are widely used by practitioners without orthonormalizing.

- **Polynomial** [192] (is a standard basis)

$$\phi_j(\sigma_i) = \begin{cases} \sum_{k=0}^{j/2} c_k \sigma_i^{2k} & \text{if } j \text{ is even} \\ \sum_{k=0}^{(j-1)/2} c_k \sigma_i^{2k+1} & \text{if } j \text{ is odd} \end{cases}$$

  Where the coefficients c_k are chosen to satisfy conditions (1) and (2) for a standard site basis.

- **Trigonometric** [225] (is an orthogonal basis, but is not normalized for functions over site spaces $|\Omega_i| > 2$)

$$\phi_j(\sigma_i) = \begin{cases} 1 & \text{if } j = 0 \\ -\cos\left(\frac{\pi(j+1)\sigma_i}{n_i}\right) & \text{if } j \text{ is odd} \\ -\sin\left(\frac{\pi j \sigma_i}{n_i}\right) & \text{if } j \text{ is even} \end{cases}$$

- **Indicator** [258] (also referred to as *occupancy* basis; is not a standard basis, since basis functions are not orthogonal to $\phi_0 = 1$)

$$\phi_j(\sigma_i) = \begin{cases} 1 & \text{if } j = 0 \\ \mathbf{1}_{\sigma_j}(\sigma_i) & \text{if } j > 0 \end{cases}$$

  Here we use $\sigma_j \in \text{enc}(\Omega_i)$ for $j = 1, \ldots, n_i - 1$. That means we use indicators for all but one species in the encoded site space.[1]

[1] In this case, since we use indicator functions we could just use the site space Ω_i without any encoding.

Appendix B

Additional proofs & derivations

B.1 Degree of fixed lattice expansions

Under the *ansatz* that any multi-body Hamiltonian $\mathcal{H}(\{\boldsymbol{p}_i, \boldsymbol{r}_i\}, \{\boldsymbol{R}_i\}; \boldsymbol{\sigma})$ can be expressed as a sum of multi-body terms V_i, for example in the case of a structure with all symmetrically equivalent sites,

$$\begin{aligned}\mathcal{H}(\{\boldsymbol{R}_i\}; \boldsymbol{\sigma}) &= V_0 + \sum_i \mathcal{V}_1(\boldsymbol{R}_i; \sigma_i) + \sum_{i<j} \mathcal{V}_2(\boldsymbol{R}_i, \boldsymbol{R}_j; \sigma_i, \sigma_j) + \sum_{i<j<k} \mathcal{V}_3(\boldsymbol{R}_i, \boldsymbol{R}_j, \boldsymbol{R}_k; \sigma_i, \sigma_j, \sigma_k) + \dots \\ &= V_0 + \sum_S \mathcal{V}_{|S|}(\{\boldsymbol{R}_i, i \in S\}; \boldsymbol{\sigma}_S)\end{aligned}$$

We can compute the Fourier expansion by computing the expansion for each multi-body term V_i individually, and subsequently, sum the computed expansions for each term to obtain the expansion for the multi-body Hamiltonian. Moreover, for a fixed structure $\mathcal{S}$, $\{\boldsymbol{R}_i\} = \boldsymbol{R}_\mathcal{S}$, each multi-body term is a discrete function of the possible configurations $\boldsymbol{\sigma}_S$,

$$\mathcal{V}_{|S|}(\{\boldsymbol{R}_i, i \in S\}; \boldsymbol{\sigma}_S) \to V_{|S|}(\boldsymbol{\sigma}_S)$$

And so each $V_{|S|}(\boldsymbol{\sigma}_S)$ can always be represented by a Fourier expansion where the largest possible degree is $d = |S|$, i.e. the order of the multi-body term.

It follows that the maximum degree term in Fourier cluster expansion of a multi-body Hamiltonian will be at most equal to the highest order multi-body term in the Hamiltonian.

This does hold however when structural relaxations are allowed because the values of relaxed structural parameters depend on the full configuration $\boldsymbol{\sigma}$ and so one can construct the total Hamiltonian from a sum of individual Fourier expansions for each multi-body term.

B.2 Fourier basis sets

Standard site basis

Let the expression of $f \in L^2(\Omega, \rho)$ in a standard site basis $\{\phi_i, i \in [n]\}$ (where $n = |\Omega|$) be,

$$f(\sigma) = \sum_{i=0}^{n-1} a_i \phi_i(\sigma)$$

The proofs for the properties of the expansion of f listed in Chapter 2.3 are as follows,

1.

$$\begin{aligned}
\mathbb{E}_\rho [f] &= \mathbb{E}_\rho \left[\sum_{i=0}^{n-1} a_i \phi_i(\sigma) \right] \\
&= \sum_{i=0}^{n-1} a_i \mathbb{E}_\rho [\phi_i(\sigma)] \\
&= a_0
\end{aligned}$$

Where we have used the fact that $\mathbb{E}_\rho [\phi_i] = \langle \phi, 1 \rangle_\rho = 0$ for all $i > 0$ by orthogonality of a standard basis.

2. This is Parseval's theorem,

$$\begin{aligned}
||f||_2^2 &= \langle f^2 \rangle_\rho \\
&= \left\langle \left(\sum_{i=0}^{n-1} a_i \phi_i(\sigma) \right)^2 \right\rangle_\rho \\
&= \left\langle \sum_{i=0}^{n-1} \sum_{j=0}^{n-1} a_i a_j \phi_i(\sigma) \phi_j(\sigma) \right\rangle_\rho \\
&= \sum_{i=0}^{n-1} \sum_{j=0}^{n-1} a_i a_j \langle \phi_i(\sigma), \phi_j(\sigma) \rangle_\rho \\
&= \sum_{i=0}^{n-1} \sum_{j=0}^{n-1} a_i a_j \delta_{ij} \\
&= \sum_{i=0}^{n-1} a_i^2
\end{aligned}$$

$\langle \phi_i, \phi_j \rangle_\rho = \delta_{ij}$ by orthonormality of a standard basis.

3. Using formulas (1) and (2),

$$\begin{aligned}\mathrm{Var}_\rho[f] &= \langle f^2\rangle_\rho - \langle f\rangle_\rho^2 \\ &= \sum_{i=0}^{n-1} a_i^2 - a_0 \\ &= \sum_{i=1}^{n-1} a_i^2\end{aligned}$$

4. First we show Plancherel's theorem following (2), using $g(\sigma) = \sum_{i=0}^{n-1} b_i\phi_i(\sigma)$,

$$\begin{aligned}\langle f, g\rangle_\rho &= \left\langle \sum_{i=0}^{n-1} a_i\phi_i(\sigma), \sum_{i=0}^{n-1} b_i\phi_i(\sigma)\right\rangle_\rho \\ &= \left\langle \sum_{i=0}^{n-1}\sum_{j=0}^{n-1} a_i b_j\phi_i(\sigma)\phi_j(\sigma)\right\rangle_\rho \\ &= \sum_{i=0}^{n-1}\sum_{j=0}^{n-1} a_i b_j\langle\phi_i(\sigma),\phi_j(\sigma)\rangle_\rho \\ &= \sum_{i=0}^{n-1}\sum_{j=0}^{n-1} a_i b_j\delta_{ij} \\ &= \sum_{i=0}^{n-1} a_i b_i\end{aligned}$$

Now using Parseval's and Plancherel's theorem,

$$\begin{aligned}\mathrm{Cov}_\rho[f, g] &= \langle f, g\rangle_\rho - \langle f\rangle_\rho\langle g\rangle_\rho \\ &= \sum_{i=0}^{n-1} a_i b_i - a_0 b_0 \\ &= \sum_{i=1}^{n-1} a_i b_i\end{aligned}$$

Graham-Schmidt process to obtain binary and ternary standard site basis sets

We can obtain a standard site basis for a binary site space $|\Omega| = 1$, starting from any other basis $\{\psi_0, \psi_1\}$ as follows:

1. Set $\phi_0(\sigma) = 1$

2. Choose any of ψ_0, ψ_1 that is not constant for the remainder of the Graham-Schmidt process,

   a) Compute the projection of ψ_1 onto ϕ_0,

$$\begin{aligned}\frac{\mathbb{E}_\rho[\psi_1\phi_0]}{\mathbb{E}_\rho[\phi_0\phi_0]}\phi_0 &= \frac{\mathbb{E}_\rho[\psi_1 \times 1]}{\mathbb{E}_\rho[1 \times 1]} \times 1 \\ &= \mathbb{E}_\rho[\psi_1]\end{aligned}$$

   b) Compute the norm squared of $\psi_1 - \mathbb{E}_\rho[\psi_1]$,

$$\begin{aligned}\mathbb{E}_\rho\left[(\psi_1 - \mathbb{E}_\rho[\psi_1])^2\right] &= \mathbb{E}_\rho\left[\psi_1^2\right] - 2\mathbb{E}_\rho[\psi_1]^2 + \mathbb{E}_\rho[\psi_1]^2 \\ &= \mathbb{E}_\rho\left[\psi_1^2\right] - \mathbb{E}_\rho[\psi_1]^2 \\ &= \mathrm{Var}_\rho[\psi_1]\end{aligned}$$

   c) Set,

$$\phi_1(\sigma) = \frac{\psi_1(\sigma) - \mathbb{E}_\rho[\psi_1(\sigma)]}{\sqrt{\mathrm{Var}_\rho[\psi_1(\sigma)]}}$$

In order to obtain a standard basis for a ternary site starting from another basis $\{\psi_0, \psi_1, \psi_2\}$, we simply follow the steps above for the binary case and carry out an additional Graham-Schmidt step to obtain ϕ_2.

1. The projection of ψ_2 onto ϕ_0 is $\mathbb{E}_\rho[\psi_2]$.

2. Compute the projection of ψ_2 onto ϕ_1,

$$\begin{aligned}\frac{\mathbb{E}_\rho[\psi_2\phi_1]}{\mathbb{E}_\rho[\phi_1\phi_1]}\phi_1 &= \mathbb{E}_\rho[\psi_2\phi_1]\,\phi_1 \\ &= \frac{\mathbb{E}_\rho[\psi_2(\psi_1 - \mathbb{E}_\rho[\psi_1])]}{\sqrt{\mathrm{Var}_\rho[\psi_1]}}\phi_1 \\ &= \frac{\mathrm{Cov}_\rho[\psi_2\psi_1]}{\sqrt{\mathrm{Var}_\rho[\psi_1]}}\phi_1\end{aligned}$$

   Where we have used the fact that ϕ_1 is normalized, and we will keep ϕ_1 as is.

3. Compute the norm squared of $u = \psi_2 - \mathbb{E}_\rho[\psi_2] - \mathbb{E}_\rho[\psi_2\phi_1]\,\phi_1$,

$$\begin{aligned}
&\mathbb{E}_\rho\left[(\psi_2 - \mathbb{E}_\rho[\psi_2] - \frac{\mathrm{Cov}_\rho[\psi_2\psi_1]}{\sqrt{\mathrm{Var}_\rho[\psi_1]}}\phi_1)^2\right] \\
&= \mathbb{E}_\rho\left[\psi_2^2 - 2\mathbb{E}_\rho[\psi_2]\,\psi_2 + \mathbb{E}_\rho[\psi_2]^2 + \frac{\mathrm{Cov}_\rho[\psi_2\psi_1]^2}{\mathrm{Var}_\rho[\phi_1]}\phi_1^2 - \frac{2\,\mathrm{Cov}_\rho[\psi_2\psi_1]}{\sqrt{\mathrm{Var}_\rho[\psi_1]}}\psi_2\phi_1 + \frac{2\,\mathrm{Cov}_\rho[\psi_2\psi_1]}{\sqrt{\mathrm{Var}_\rho[\psi_1]}}\mathbb{E}_\rho[\psi_2]\,\phi_1\right] \\
&= \mathbb{E}_\rho\left[(\psi_2 - \mathbb{E}_\rho[\psi_2])^2\right] + \frac{\mathrm{Cov}_\rho[\psi_2\psi_1]^2}{\mathrm{Var}_\rho[\psi_1]}\mathbb{E}_\rho\left[\phi_1^2\right] - \frac{2\,\mathrm{Cov}_\rho[\psi_2\psi_1]}{\sqrt{\mathrm{Var}_\rho[\psi_1]}}\mathbb{E}_\rho[\psi_2\phi_1] \\
&= \mathrm{Var}_\rho[\psi_2] + \frac{\mathrm{Cov}_\rho[\psi_2\psi_1]^2}{\mathrm{Var}_\rho[\psi_1]} - \frac{2\,\mathrm{Cov}_\rho[\psi_2\psi_1]}{\mathrm{Var}_\rho[\psi_1]}\mathbb{E}_\rho[\psi_2\psi_1 - \mathbb{E}_\rho[\psi_1]\,\psi_2] \\
&= \mathrm{Var}_\rho[\psi_2] - \frac{\mathrm{Cov}_\rho[\psi_2\psi_1]^2}{\mathrm{Var}_\rho[\psi_1]}
\end{aligned}$$

4. Set $\phi_2 = \frac{u}{||u||_2}$,

$$\phi_2 = \frac{\psi_2 - \mathbb{E}_\rho[\psi_2] - \frac{\mathrm{Cov}_\rho[\psi_2\psi_1]}{\mathrm{Var}_\rho[\psi_1]}(\psi_1 - \mathbb{E}_\rho[\psi_1])}{\sqrt{\mathrm{Var}_\rho[\psi_2] - \frac{\mathrm{Cov}_\rho[\psi_2\psi_1]^2}{\mathrm{Var}_\rho[\psi_1]}}}$$

Fourier product basis

Normalized: First we show that Fourier product basis functions $\Phi_{\boldsymbol{\alpha}}$ for $L^2(\boldsymbol{\Omega}, \boldsymbol{\rho})$ (with $|\boldsymbol{\Omega}| = \prod_{i=1}^N |\Omega_i|$) are normalized,

$$\begin{aligned}
\langle \Phi_{\boldsymbol{\alpha}}, \Phi_{\boldsymbol{\alpha}} \rangle_{\boldsymbol{\rho}} &= \left\langle \prod_{i=0}^{N-1} \phi_{\alpha_i}^{(i)}(\sigma_i), \prod_{i=0}^{N-1} \phi_{\alpha_i}^{(i)}(\sigma_i) \right\rangle_{\boldsymbol{\rho}} \\
&= \left\langle \prod_{i=0}^{N-1} \phi_{\alpha_i}^{(i)}(\sigma_i)\phi_{\alpha_i}^{(i)}(\sigma_i) \right\rangle_{\boldsymbol{\rho}} \\
&= \prod_{i=0}^{N-1} \langle \phi_{\alpha_i}^{(i)}, \phi_{\alpha_i}^{(i)} \rangle_\rho \\
&= 1
\end{aligned}$$

Where we have used the fact that sums over configurations commute with products of site basis functions[1], and site basis functions are orthonormal.

[1] Which means that site basis functions are *uncorrelated* under a probabilistic interpretation

Orthogonal: Now we show that Fourier basis functions are orthogonal following the same procedure,

$$\begin{aligned}
\langle \Phi_{\boldsymbol{\alpha}}, \Phi_{\boldsymbol{\eta}} \rangle_{\boldsymbol{\rho}} &= \left\langle \prod_{i=0}^{N-1} \phi^{(i)}_{\alpha_i}(\sigma_i), \prod_{i=0}^{N-1} \phi^{(i)}_{\eta_i}(\sigma_i) \right\rangle_{\boldsymbol{\rho}} \\
&= \left\langle \prod_{i=0}^{N-1} \phi^{(i)}_{\alpha_i}(\sigma_i) \phi^{(i)}_{\eta_i}(\sigma_i) \right\rangle_{\boldsymbol{\rho}} \\
&= \prod_{i=0}^{N-1} \langle \phi^{(i)}_{\alpha_i}, \phi^{(i)}_{\eta_i} \rangle_{\rho} \\
&= \prod_{i=0}^{N-1} \delta_{\alpha_i, \eta_i} \\
&= \delta_{\boldsymbol{\alpha}, \boldsymbol{\eta}}
\end{aligned}$$

Complete: Finally, it follows that the set of all possible $\{\Phi_{\boldsymbol{\alpha}} \ : \ \forall \boldsymbol{\alpha} \in \mathbb{N}^N_{<\boldsymbol{n}}\}$ is a basis[2] for $L^2(\boldsymbol{\Omega}, \boldsymbol{\rho})$, since all $\Phi_{\boldsymbol{\alpha}}$ are linearly independent (orthogonal), and there are a total $|\mathbb{N}^N_{<\boldsymbol{n}}| = \prod_{i=1}^{N} |\Omega_i| = |\boldsymbol{\Omega}|$ such functions, which is precisely the dimension of $L^2(\boldsymbol{\Omega}, \boldsymbol{\rho})$ [157].

Fourier formulas for an expansion of $F \in L^2(\boldsymbol{\Omega})$: The Fourier formulas for a function expanded using a Fourier product basis listed in Chapter 2.3 can be derived following the same procedure as done for the analogous formulas for a site function given in Appendix 2.19.

Orthogonality of Fourier correlation functions

Fourier correlation functions can be shown to be orthogonal simply by expanding them in terms of Fourier product basis functions and using their orthonormality.

$$\begin{aligned}
\langle \Theta_\beta, \Theta_\gamma \rangle_{\boldsymbol{\rho}} &= \frac{1}{N^2} \left\langle \frac{1}{m_\beta} \sum_{\boldsymbol{\alpha} \in \beta} \Phi_{\boldsymbol{\alpha}}(\boldsymbol{\sigma}), \frac{1}{m_\gamma} \sum_{\boldsymbol{\eta} \in \gamma} \Phi_{\boldsymbol{\eta}}(\boldsymbol{\sigma}) \right\rangle_{\boldsymbol{\rho}} \\
&= \frac{1}{N^2 m_\beta m_\gamma} \sum_{\boldsymbol{\alpha} \in \beta} \sum_{\boldsymbol{\eta} \in \gamma} \langle \Phi_{\boldsymbol{\alpha}}(\boldsymbol{\sigma}), \Phi_{\boldsymbol{\eta}}(\boldsymbol{\sigma}) \rangle_{\boldsymbol{\rho}} \\
&= \frac{1}{N^2 m_\beta m_\gamma} \sum_{\boldsymbol{\alpha} \in \beta} \sum_{\boldsymbol{\eta} \in \gamma} \delta_{\boldsymbol{\alpha}, \boldsymbol{\eta}} \\
&= \frac{\delta_{\beta,\gamma}}{N m_\beta}
\end{aligned}$$

For which we used the fact that a multi-index $\boldsymbol{\alpha}$ never appears in two different orbits $\beta \neq \gamma$.

[2] In other words is complete or spans $L^2(\boldsymbol{\Omega}, \boldsymbol{\rho})$

Orthogonality of cluster interactions

The proof of orthogonality of cluster interactions follows almost directly from the orthogonality of Fourier correlation basis functions,

$$\begin{aligned}\langle H_B, H_D\rangle &= \left\langle \sum_{\beta\in L(B)} \hat{m}_\beta J_\beta \Theta_\beta(\boldsymbol{\sigma}), \sum_{\gamma\in L(D)} \hat{m}_\gamma J_\gamma \Theta_\gamma(\boldsymbol{\sigma}) \right\rangle \\ &= \sum_{\beta\in L(B)} \sum_{\gamma\in L(D)} \hat{m}_\beta \hat{m}_\gamma J_\beta J_\gamma \langle \Theta_\beta(\boldsymbol{\sigma}), \Theta_\gamma(\boldsymbol{\sigma})\rangle \\ &= \sum_{\beta\in L(B)} \sum_{\gamma\in L(D)} \hat{m}_\beta \hat{m}_\gamma J_\beta J_\gamma \frac{\delta_{\beta\gamma}}{m_B \hat{m}_\beta N} \\ &= \begin{cases} ||H_B||_2^2 \text{ if } B = D \\ 0 \text{ if } B \neq D \end{cases}\end{aligned}$$

Additionally, a cluster interaction H_B is orthogonal to any function $F \in L^2(\boldsymbol{\Omega}, \boldsymbol{\rho})^{\mathcal{G}}$ that can be expressed with correlation functions for $\gamma \in \bigcup_D L(D) \;\; \forall D \neq B$. To show this we need to simply expand such a function in a Fourier correlation basis and use the fact that all basis functions will be orthogonal to those in the expansion of H_B.

Uniqueness of cluster decomposition

The proof of the uniqueness of a cluster decomposition is simple and follows established proofs for the uniqueness of the Sobol and the functional ANOVA decomposition [96, 206]. The proof is by contradiction, so we start by considering two different cluster decompositions for the same Hamiltonian $H \in L^2(\boldsymbol{\Omega}, \boldsymbol{\rho})^{\mathcal{G}}$,

$$\begin{aligned} H(\boldsymbol{\sigma}) &= N \sum_{\mathcal{G}\,\mathrm{P}([N])} m_B H_B(\boldsymbol{\sigma}) \\ H(\boldsymbol{\sigma}) &= N \sum_{\mathcal{G}\,\mathrm{P}([N])} m_B \widetilde{H}_B(\boldsymbol{\sigma}) \end{aligned}$$

We can use the two expression above as an expansion of a function everywhere zero, $F(\sigma) = 0 \;\forall\sigma$,

$$\begin{aligned} F(\sigma) &= N \left(\sum_{\mathcal{G}\,\mathrm{P}([N])} m_B H_B(\boldsymbol{\sigma}) - \sum_{\mathcal{G}\,\mathrm{P}([N])} m_B \widetilde{H}_B(\boldsymbol{\sigma}) \right) \\ &= N \sum_{\mathcal{G}\,\mathrm{P}([N])} m_B \left(H_B(\boldsymbol{\sigma}) - \widetilde{H}_B(\boldsymbol{\sigma}) \right) \end{aligned}$$

Now, if we consider the norm squared of each term in the expansion F of the zero function,

$$\begin{aligned}\left\langle \left(H_B(\boldsymbol{\sigma}) - \widetilde{H}_B(\boldsymbol{\sigma})\right), \left(H_D(\boldsymbol{\sigma}) - \widetilde{H}_D(\boldsymbol{\sigma})\right)\right\rangle_{\boldsymbol{\rho}} &= \langle H_B^2(\boldsymbol{\sigma})\rangle_{\boldsymbol{\rho}} - \langle \widetilde{H}_B^2(\boldsymbol{\sigma})\rangle_{\boldsymbol{\rho}} \\ &= 0\end{aligned}$$

Where we used the orthogonality properties of cluster interactions, and that the norms of cluster interactions are invariant: $\langle H_B^2(\boldsymbol{\sigma})\rangle_{\boldsymbol{\rho}} = \langle \widetilde{H}_B^2(\boldsymbol{\sigma})\rangle_{\boldsymbol{\rho}}$ from Chapter 2.4.

Finally, since the norm of each term is equal to zero, then each term in the expansion is itself a zero function,

$$\begin{aligned}H_B(\boldsymbol{\sigma}) - \widetilde{H}_B(\boldsymbol{\sigma}) &= 0 \\ H_B(\boldsymbol{\sigma}) &= \widetilde{H}_B(\boldsymbol{\sigma})\end{aligned}$$

This shows that the cluster decomposition of $H(\boldsymbol{\sigma})$ is unique.

Change of basis matrix

Proving that the change of basis matrix between two Fourier basis sets is orthogonal follows the fact that it is constructed from rotation matrices, which are themselves orthogonal,

$$\begin{aligned}(U^T U)_{\alpha\beta} &= \sum_{\gamma} U_{\alpha\gamma} U_{\gamma\beta} \\ &= \sum_{\gamma} \prod_{i}^{N} \langle \phi_{\alpha_i}, R\phi_{\gamma_i}\rangle \langle R\phi_{\gamma_i}, \phi_{\beta_i}\rangle \\ &= \prod_{i}^{N} \sum_{\gamma_i} \langle \phi_{\alpha_i}, R\phi_{\gamma_i}\rangle \langle R\phi_{\gamma_i}, \phi_{\beta_i}\rangle \\ &= \prod_{i}^{N} \langle \phi_{\alpha_i}, \phi_{\beta_i}\rangle \\ &= \delta_{\alpha\beta} \\ \implies U^T &= U^{-1}\end{aligned}$$

Where we have used the resolution of the identity.

To prove that the blocks are diagonal we can follow the same prescription above. However, it actually follows that a block diagonal matrix is orthogonal if and only if the blocks are orthogonal. The proof is straightforward.

Write U in terms of the diagonal blocks $U_S, \ldots, U_T$,

$$U = \begin{bmatrix} U_S & & 0 \\ & \ddots & \\ 0 & & U_T \end{bmatrix}$$

Then U orthogonal means $U^T U = I$, and thus by expanding,

$$\begin{aligned} U^T U &= \begin{bmatrix} U_S^T & & 0 \\ & \ddots & \\ 0 & & U_T^T \end{bmatrix} \begin{bmatrix} U_S & & 0 \\ & \ddots & \\ 0 & & U_T \end{bmatrix} \\ &= \begin{bmatrix} U_S^T U_S & & 0 \\ & \ddots & \\ 0 & & U_T^T U_T \end{bmatrix} \\ \implies & U_V^T U_V = I \ \forall V \in \{S, \ldots, T\} \end{aligned}$$

To prove the reverse, we simply go backward starting from $U_V^T U_V = I \ \forall V \in \{S, \ldots, T\}$.

B.3 Frame bounds for the generalized Potts frame

To derive frame bounds for the generalized Potts frame, we work with the product-basis of cluster indicator functions without taking symmetry-adapted averages. The bounds obtained apply equally to the functions obtained from symmetry-adapted averages. The lower frame bound A is obtained by splitting up the sum of projections into the frame elements as follows,

$$\begin{aligned}\sum_{\gamma\in I}|\langle H,\mathbf{1}_\gamma\rangle|^2 &= \sum_{\gamma_{\max}}|\langle H,\mathbf{1}_{\gamma_{\max}}\rangle|^2+\sum_{\gamma\in I\setminus\{\gamma_{\max}\}}|\langle H,\mathbf{1}_\gamma\rangle|^2\\ &= ||H||^2+\sum_{\gamma\in I\setminus\{\gamma_{\max}\}}|\langle H,\mathbf{1}_\gamma\rangle|^2\\ &\geq ||H||^2\end{aligned}$$

Where $\{\gamma_{\max}\}$ denotes the set of all maximal clusters and thus represents the canonical orthonormal basis for the Hilbert space $\mathcal{H}$. The second sum over all smaller clusters is always greater than or equal to zero, and so the obtained lower frame bound is $A = 1$.

To obtain an upper frame bound we start by writing the Fourier expansion for cluster indicator functions,

$$\begin{aligned}\mathbf{1}_\gamma(\boldsymbol{\sigma}) &= \sum_\alpha\langle\mathbf{1}_\gamma,\Phi_\alpha\rangle\Phi_\alpha(\boldsymbol{\sigma})\\ &= \sum_\alpha\prod_i\langle\mathbf{1}_{\gamma_i},\phi_{\alpha_i}\rangle\Phi_\alpha(\boldsymbol{\sigma})\\ &= \sum_{\substack{\alpha;\\ \text{supp}(\alpha)\subseteq\text{supp}(\gamma)}}\frac{1}{n^{\#\gamma}}\prod_i\phi_{\alpha_i}(\sigma_{\gamma_i})\Phi_\alpha(\boldsymbol{\sigma})\\ &= \sum_{\substack{\alpha;\\ \text{supp}(\alpha)\subseteq\text{supp}(\gamma)}}\frac{1}{n^{\#\gamma}}\Phi_\alpha(\boldsymbol{\sigma}_{\boldsymbol{\gamma}})\Phi_\alpha(\boldsymbol{\sigma})\end{aligned}$$

where supp$(\cdot)$ represents the support or indices of the nonzero entries of α or γ, and $\#\gamma$ represents the total number of nonzero entries—i.e. the number of sites in a cluster. n is the number of species allowed at a site and $\boldsymbol{\sigma}_\gamma$ is any occupancy string that includes the cluster represented by γ. Note the above expression is for a system with a single type of site (i.e. with the same set of allowed species in all sites). The general expression simply involves one over the product of the different number of species for each site instead of the factor $1/n^{\#\gamma}$.

Using the expansion given above for cluster indicator functions, we obtain an upper frame

bound as follows,

$$\begin{aligned}
\sum_{\gamma\in I}|\langle H,\mathbf{1}_\gamma\rangle|^2 &= \sum_{\gamma\in I}\left|\left\langle H,\sum_{\substack{\alpha;\\ \mathrm{supp}(\alpha)\subseteq\mathrm{supp}(\gamma)}}\frac{1}{n^{\#\gamma}}\Phi_\alpha(\boldsymbol{\sigma_\gamma})\Phi_\alpha(\boldsymbol{\sigma})\right\rangle\right|^2\\
&= \sum_{\gamma\in I}\left(\frac{1}{n^{\#\gamma}}\right)^2\left|\sum_{\substack{\alpha;\\ \mathrm{supp}(\alpha)\subseteq\mathrm{supp}(\gamma)}}\hat{H}_\alpha\Phi_\alpha(\boldsymbol{\sigma_\gamma})\right|^2\\
&= \sum_{S}\frac{1}{n^{2|S|}}\sum_{\substack{\boldsymbol{\sigma}_\gamma;\\ \mathrm{supp}(\gamma)=S}}\left|\sum_{\substack{\alpha;\\ \mathrm{supp}(\alpha)\subseteq S}}\hat{H}_\alpha\Phi_\alpha(\boldsymbol{\sigma_\gamma})\right|^2\\
&= \sum_{S}\frac{1}{n^{|S|}}\sum_{\substack{\alpha;\\ \mathrm{supp}(\alpha)\subseteq S}}\hat{H}_\alpha^2\\
&= \sum_{\alpha}\left(\sum_{S\supseteq\mathrm{supp}(\alpha)}\frac{1}{n^{|S|}}\right)\hat{H}_\alpha^2\\
&\leq \left(\sum_{S}\frac{1}{n^{|S|}}\right)||H||^2\\
&= \left(1+n^{-1}\right)^N||H||^2
\end{aligned}$$

where $\hat{H}_\alpha$ are the Fourier coefficients of H. The sets S contain site indices and the sum over these contains all possible subsets of site indices—i.e. all clusters of un-labeled sites. The upper frame bound obtained is $B=(1+n^{-1})^N$, where N is the total number of sites in the structure. The bound obtained is an improvement over the bound 2^N given by the Cauchy-Schwarz inequality.

B.4 Reduced correlations and cluster interactions

As discussed in Chapter 3 correlation functions and mean cluster interaction can be practically computed using a different expression than the expressions used to define them in Chapter 2. For convenience, we reproduce Equation 2.27 used to define a correlation function,

$$\Theta_\beta(\boldsymbol{\sigma}) = \frac{1}{|\beta|}\sum_{\boldsymbol{\alpha}\in\beta}\Phi_{\boldsymbol{\alpha}}(\boldsymbol{\sigma})$$

However, for practical purposes, it is more effective to use reduced correlation functions following Equation 3.3, which we can obtain as follows,

$$\begin{aligned}\Theta_\beta(\boldsymbol{\sigma}) &= \frac{1}{|\beta|}\sum_{\boldsymbol{\alpha}\in\beta}\Phi_{\boldsymbol{\alpha}}(\boldsymbol{\sigma}) \\ &= \frac{1}{|B|}\sum_{S\in B}\frac{1}{\hat{m}_\beta}\sum_{\hat{\boldsymbol{\alpha}}\in\hat{\beta}}\Phi_{\hat{\boldsymbol{\alpha}}}(\boldsymbol{\sigma}_S) \\ &= \frac{1}{|B|}\sum_{S\in B}\hat{\Theta}_\beta(\boldsymbol{\sigma}_S)\end{aligned}$$

Where the site cluster $B = B(\beta) = \mathrm{orb}(\mathcal{G}(\mathrm{supp}(\boldsymbol{\alpha})))$ for any $\boldsymbol{\alpha} \in \beta$; $\hat{\boldsymbol{\alpha}} = \mathrm{ctr}(\boldsymbol{\alpha})$ are contracted multi-indices; and $\hat{m}_\beta = |\hat{\beta}|$ is the number of symmetrically equivalent contracted multi-indices $\hat{\boldsymbol{\alpha}}$.

Many of the expressions given in Chapter 3 for practical calculations and their derivations are based on computing correlation functions or cluster interactions that act over a given orbit D of site space clusters T as averages over clusters $S \in B$ that contain the clusters $T \in D$, i.e. any $T \in D$ is a sub-cluster $T \subseteq S$ of one of the clusters $S \in B$. This practical transformation is derived based on the expression given in Equation 3.3 (reproduced above) to compute a correlation, cluster interaction, or any function that is *symmetrized* over an equivalent set of site space clusters.

$$\begin{aligned}F_D(\boldsymbol{\sigma}) &= \frac{1}{|D|}\sum_{S\in D}\hat{F}_D(\boldsymbol{\sigma}_S) \\ &= \frac{1}{|B|}\frac{|D|}{|B|}\times\frac{1}{N_{DB}}\sum_{S\in B}\sum_{T\leftarrow S}\hat{F}_D(\boldsymbol{\sigma}_T) \\ &= \frac{c_{DB}}{|B|}\sum_{S\in B}\sum_{T\leftarrow S}\hat{F}_D(\boldsymbol{\sigma}_T)\end{aligned}$$

Where $T \leftarrow S$, refers to all clusters T in orbit D that are sub-clusters of cluster $S \in B$, i.e. $T \in \{T \subset S \text{ for } T \in D\}$. $N_{DB} = |\{T \ : \ \forall T \leftarrow S\}|$ are the number of subclusters $T \in D$ contained in cluster $S \in B$.

The *counting factor* c_{DB} is the inverse of the number of times a cluster $T \in D$ was included in the sum over clusters $S \in B$, and can be expressed as,

$$\begin{aligned} c_{DB} &= \frac{|D|}{N_{DB}|B|} \\ &= \frac{m_D}{N_{DB} m_B} \end{aligned}$$

Where we used the definition of a site cluster multiplicity: $m_B = |B|/N$

For practical purposes, it is useful to express the inner sum over subclusters $T \leftarrow S$ as a *broadcasted* tensor of dimensions $\times_{i \in S} |\Omega_i|$. Broadcasting simply amounts to aligning the indices of the clusters T to their counterpart of the subset of indices of cluster S, i.e. following the mapping between sites in T to those in S. This must be done for all $T \leftarrow S$ and then each broad-casted array is summed element-wise, such that evaluating the inner sum over $T \leftarrow S$ amounts to accessing the corresponding configuration $\boldsymbol{\sigma}_S$,

$$\sum_{T \leftarrow S} \hat{F}_D(\boldsymbol{\sigma}_T) = \left[\hat{F}_{DB}\right]_{\boldsymbol{\sigma}_S} \tag{B.1}$$

Where $\left[\hat{F}_{DB}\right]$ is the final broadcasted tensor of the reduced function $\hat{F}_D$ that operates over clusters $T \in D$ broadcasted to a reduced function $\hat{F}_{DB}$ acting over a cluster $S \in B$.

Using the above expressions, we can now derive the inner products used in Equation 3.20, to compute the projections of a (pseudo) mean cluster interaction $\widetilde{H}_B$ onto a correlation function Θ_γ, where every $\text{supp}(\boldsymbol{\alpha}) \in \gamma$ is a subcluster $\text{supp}(\boldsymbol{\alpha}) \subseteq S$ of some cluster $S \in B$.

$$\begin{aligned} m_B N \langle \widetilde{H}_B, \Theta_\gamma \rangle_{\boldsymbol{\rho}} &= m_B N \left\langle \frac{1}{|B|} \sum_{S \in B} \hat{H}_B(\boldsymbol{\sigma}_S) \frac{1}{|D(\gamma)|} \sum_{T \in D(\gamma)} \hat{\Theta}_\gamma(\boldsymbol{\sigma}_T) \right\rangle_{\boldsymbol{\rho}} \\ &= \frac{m_B N}{|B||D(\gamma)|} \sum_{S \in B} \sum_{T \in D(\gamma)} \left\langle \hat{H}_B(\boldsymbol{\sigma}_S) \hat{\Theta}_\gamma(\boldsymbol{\sigma}_T) \right\rangle_{\boldsymbol{\rho}} \\ &\overset{(1)}{=} \frac{m_B}{m_D |B|} \sum_{S \in B} \left\langle \hat{H}_B(\boldsymbol{\sigma}_S) \sum_{T \leftarrow S} \hat{\Theta}_\gamma(\boldsymbol{\sigma}_T) \right\rangle_{\boldsymbol{\rho}} \\ &= \frac{m_B}{m_D |B|} \sum_{S \in B} \left\langle \hat{H}_B(\boldsymbol{\sigma}_S) \hat{\Theta}_{\gamma B}(\boldsymbol{\sigma}_S) \right\rangle_{\boldsymbol{\rho}} \\ &= \frac{m_B}{m_D} \left\langle \hat{H}_B(\boldsymbol{\sigma}_S) \hat{\Theta}_{\gamma B}(\boldsymbol{\sigma}_S) \right\rangle_{\boldsymbol{\rho}} \end{aligned}$$

Where to obtain (1) we use the fact that the inner product of any function $\hat{H}_B(\boldsymbol{\sigma}_S)$ with a reduced correlation over a site cluster $T \not\subseteq S$ is zero. The last line can be calculated in practice using reduced tensors following Equation 3.20, which is what we sought to derive.

B.5 Cluster probabilities from correlation function expectations

In order to derive Equation 3.31, which allows computing cluster concentrations and probabilities using any correlation function basis, let us first derive the cluster expansion of a cluster indicator correlation function.

$$\begin{aligned}
\hat{m}_\beta I_\beta(\boldsymbol{\sigma}) &= \frac{1}{|B|}\sum_{S\in B}\hat{m}_\beta \hat{I}_\beta(\boldsymbol{\sigma}_S) \\
&\overset{(1)}{=} \frac{1}{|B|}\sum_{S\in B}\sum_{\gamma\in\mathcal{G}(\mathbb{N}^{|S|}_{<\boldsymbol{n}})} I_{\gamma\beta}\hat{\Theta}_{\gamma B}(\boldsymbol{\sigma}_S) \\
&= \frac{1}{|B|}\sum_{S\in B}\sum_{\gamma\in\mathcal{G}(\mathbb{N}^{|S|}_{<\boldsymbol{n}})} I_{\gamma\beta}\sum_{T\leftarrow S}\hat{\Theta}_\gamma(\boldsymbol{\sigma}_T) \\
&\overset{(2)}{=} \sum_{\gamma\in\mathcal{G}(\mathbb{N}^{|S|}_{<\boldsymbol{n}})} I_{\gamma\beta}\frac{1}{|B|}\sum_{S\in B}\sum_{T\leftarrow S}\hat{\Theta}_\gamma(\boldsymbol{\sigma}_T) \\
&= \sum_{\gamma\in\mathcal{G}(\mathbb{N}^{|S|}_{<\boldsymbol{n}})} I_{\gamma\beta}\frac{1}{c_{DB}|B|}\sum_{T\in D(\gamma)}\hat{\Theta}_\gamma(\boldsymbol{\sigma}_T) \\
&= \sum_{\gamma\in\mathcal{G}(\mathbb{N}^{|S|}_{<\boldsymbol{n}})} I_{\gamma\beta}\frac{N_{DB}}{|D|}\sum_{T\in D(\gamma)}\hat{\Theta}_\gamma(\boldsymbol{\sigma}_T) \\
&= \sum_{\gamma\in\mathcal{G}(\mathbb{N}^{|S|}_{<\boldsymbol{n}})} N_{DB} I_{\gamma\beta}\Theta_\gamma(\boldsymbol{\sigma})
\end{aligned}$$

Where $B(\beta)$ and $D(\gamma)$ are the orbits of site clusters given by the support of the multi-indices in β and γ respectively. (1) Where we used the expansion of a reduced indicator correlation given in Equation 3.28. (2) The sums are rearranged since $|S|$ is constant for any $S \in B$. By taking thermodynamic expectations, we obtain the probability of the clusters $\boldsymbol{\sigma}_S$ that are included in the orbit of reduced multi-indices $\hat{\beta}$ (see Equation 3.27),

$$\mathbb{P}_B(\boldsymbol{\sigma}_S \in \hat{\beta}|T) = \langle \hat{m}_\beta I_\beta(\boldsymbol{\sigma})\rangle_T = \sum_{\gamma\in\mathcal{G}(\mathbb{N}^{|S|}_{<\boldsymbol{n}})} N_{DB} I_{\gamma\beta}\langle\Theta_\gamma(\boldsymbol{\sigma})\rangle_T$$

Finally, by using the last expression for all symmetrically distinct cluster occupancies indicated by each I_β expressed in vector form, we obtain,

$$\mathbb{P}_B = \mathrm{diag}(N_{DB})\mathbb{V}_S^+\langle\boldsymbol{\Pi}_B\rangle_T$$

Which is precisely Equation 3.31, where we identify $M_{DB} = \mathrm{diag}(N_{DB})$ as a diagonal matrix with the number of clusters of orbit $D(\gamma)$ that are sub-clusters of clusters in the orbit $B(\beta)$.

Appendix C

Numerical calculations & expansion fits

C.1 Li transition metal oxifluorides

Density functional theory calculations

The 2-2 LiMnOF binary-binary rocksalt and 3-2 LiMnTiOF ternary-binary rocksalt structures were generated using Monte Carlo with an electrostatic potential to sample configurations with low electrostatic energy. Formation energy was computed using a plane-wave basis set with an energy cutoff of 520 eV, and reciprocal space discretization of 25 k-points per Å. All calculations were converged to 10^{-6} eV in total energy for electronic loops and 0.02 eV/Åin interatomic forces for ionic loops using the Perdew–Burke–Ernzerhof (PBE) generalized gradient approximation exchange-correlation functional[175] with rotationally-averaged Hubbard U correction (GGA+U) to compensate for the self-interaction errors as done for the LMTOF rocksalt.

The Li-Mn-O-F spinel-like structures were generated by using an initial set of structures from the Materials Project[105], and subsequently using this set and Monte Carlo to generate additional structures of up to 216 atoms. The formation energy for each training/test structure using the projector augmented wave (PAW) method,[123] with reciprocal space discretization of 25 k-points per Å and a plane wave energy cutoff of 520 eV. Spin-polarized calculations were done using the SCAN meta-GGA exchange-correlation [211] and pseudopotentials which include semicore states: Li_sv, Mn_pv, O, and F. Structures are converged to 10^{-6} eV in total energy and 0.01 eV/Å on atomic forces.

Cluster and Potts frame expansion fits

A total of 50 different fits for each system listed in Table 5.2 were computed by selecting a random set of training structures that gives a full rank underdetermined system with the given number of training structures. The remaining structures listed are used as a test set.

The fits were all carried out using the Lasso regression model implemented in the Python package `scikit-learn` [169].

C.2 $LiMnO_2$-Li_2TiO_3-LiF ceramics

Density functional theory calculations

DFT calculations were performed with the *Vienna ab initio simulation package* (VASP) using the projector-augmented wave method[122, 123], a plane-wave basis set with an energy cutoff of 520 eV, and a reciprocal space discretization of 25 k-points per Å. All calculations were converged to 10^{-6} eV in total energy for electronic loops and 0.02 eV/Åin interatomic forces for ionic loops. We used the Perdew–Burke–Ernzerhof (PBE) generalized gradient approximation exchange-correlation functional[175] with rotationally-averaged Hubbard U correction (GGA+U) to compensate for the self-interaction error on all transition metal atoms except titanium[239]. The U parameters were obtained from the literature, where they were calibrated to transition metal oxide formation energies (3.9 eV for Mn). The GGA+U computational framework is believed to be reliable in determining the formation enthalpies of similar compounds[106]. DFT calculations were done for a total of 1230 structures with supercells ranging from 4 atoms to 132 atoms.

Cluster expansion fits

In the construction of CE, models with different numbers of ECIs were considered. The number of ECIs is set by changing the radii cutoffs for different interactions. The CE models are labeled with (pair, triplet, quadruplet) in Å to represent the cutoff radius of different types of interactions. Consequently, (7, 4, 4) results in 83 ECIs, (7, 4.2, 4.2) results in 143 ECIs, (7, 5.6, 4.2) results in 235 ECIs, and (7, 5.6, 5.6) results in 995 ECIs. All CE models are constructed based on a primitive cell of rocksalt structure with lattice parameter $a = 3$ Å. All models include an explicit electrostatic term computed using the Ewald summation method as implemented in `pymatgen` [161]. Fits with $\ell_2\ell_0$ regression were only carried out with the first two sets of cutoffs based on the substantial compute time necessary compared to Lasso-based regression. All fits were carried out with a training set of 983 structures with supercells up to 72 atoms. A test set of 247 structures of supercell sizes 128 and 132 atoms was used for validation.

The hyper-parameter tuning paths for all linear regression models used are shown in Figure C.1. We observe that a plateau region exists for most models for the two shorter sets of cluster cutoffs. For the fits with the largest set of cutoffs, all regression models show a clear cross-validation score minimum. Further, the location of the minimum is relatively constant for all regression models.

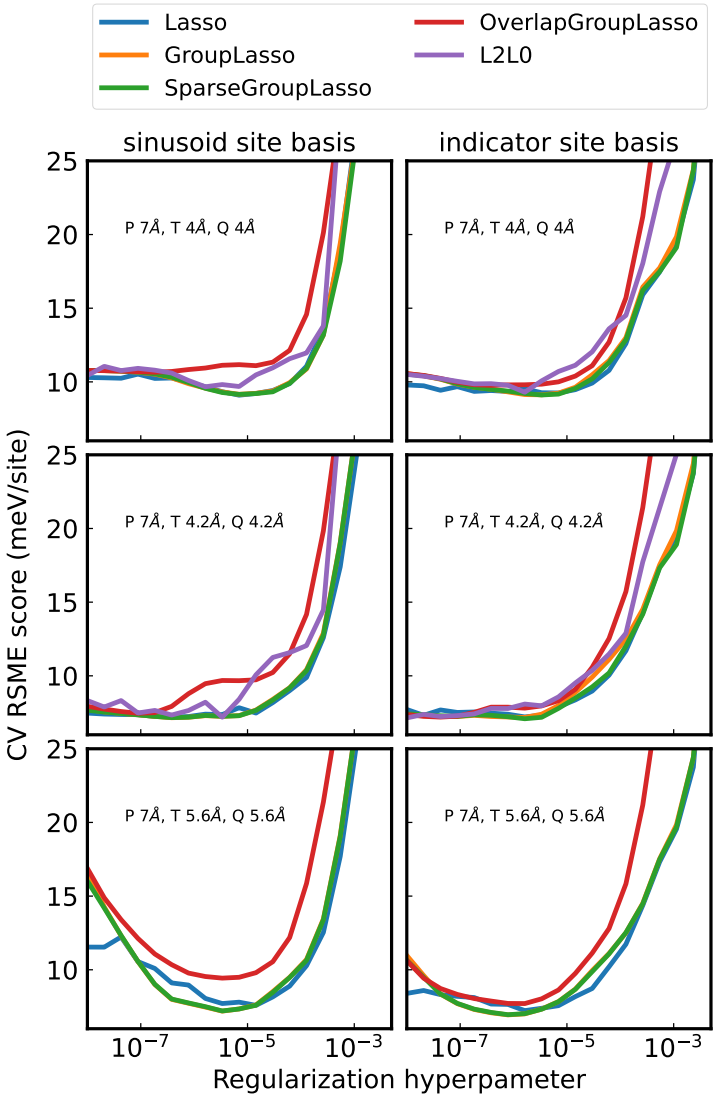


Figure C.1: Regularization paths for LMTOF CE fits.

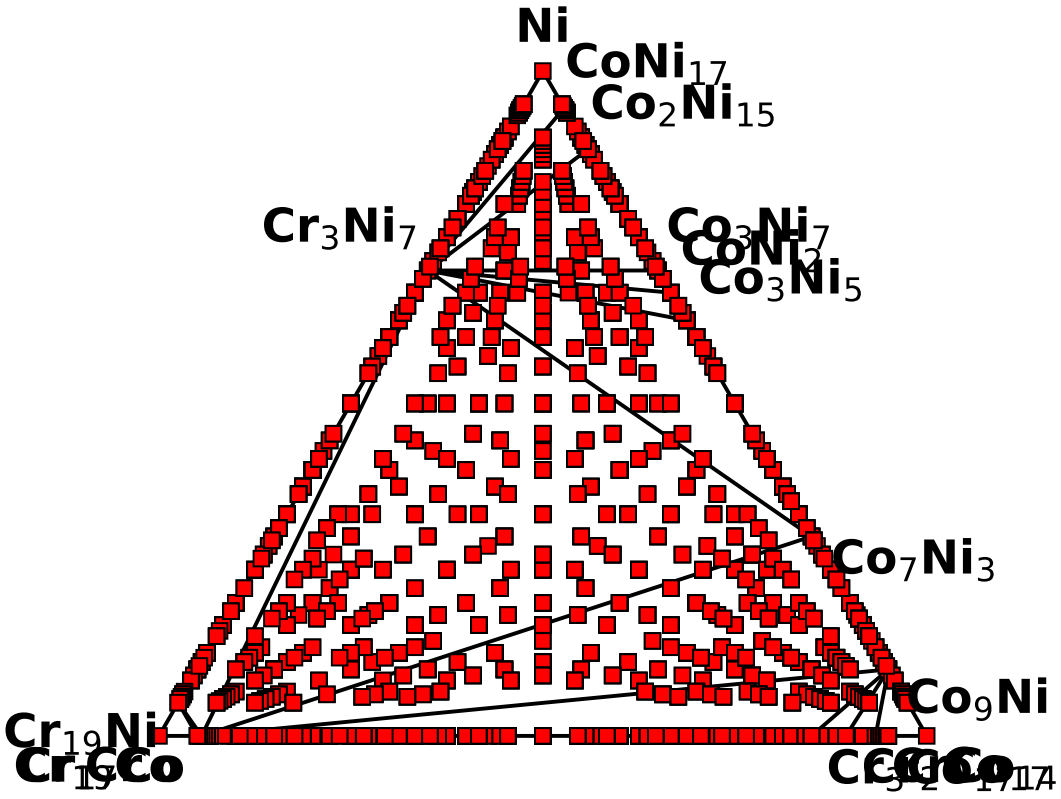


Figure C.2: Ternary phase diagram of compositions sampled for NiCoCr training structures

C.3 NiCoCr alloys

Density functional theory calculations

DFT calculations for NiCoCr structures were performed following the Materials Project [105] `MetalRelaxSet` defined in the `pymatgen` package [161]. Calculations were done using *Vienna ab initio simulation package* (VASP) using the projector-augmented wave method[122, 123], a plane-wave basis set with an energy cutoff of 520 eV, and a reciprocal space discretization of 200 k-points per Å. Electronic exchange-correlation effects are described using used the Perdew–Burke–Ernzerhof (PBE) generalized gradient approximation exchange-correlation functional [175]. All calculations were converged to 10^{-5} eV in total energy for electronic loops and 0.01 eV/Å.

DFT calculations were done for training structures covering a wide range of compositions. Figure C.2 shows a ternary composition diagram of compositions for which electronic structure calculations were performed. At least 4 distinct structures were computed at each composition.

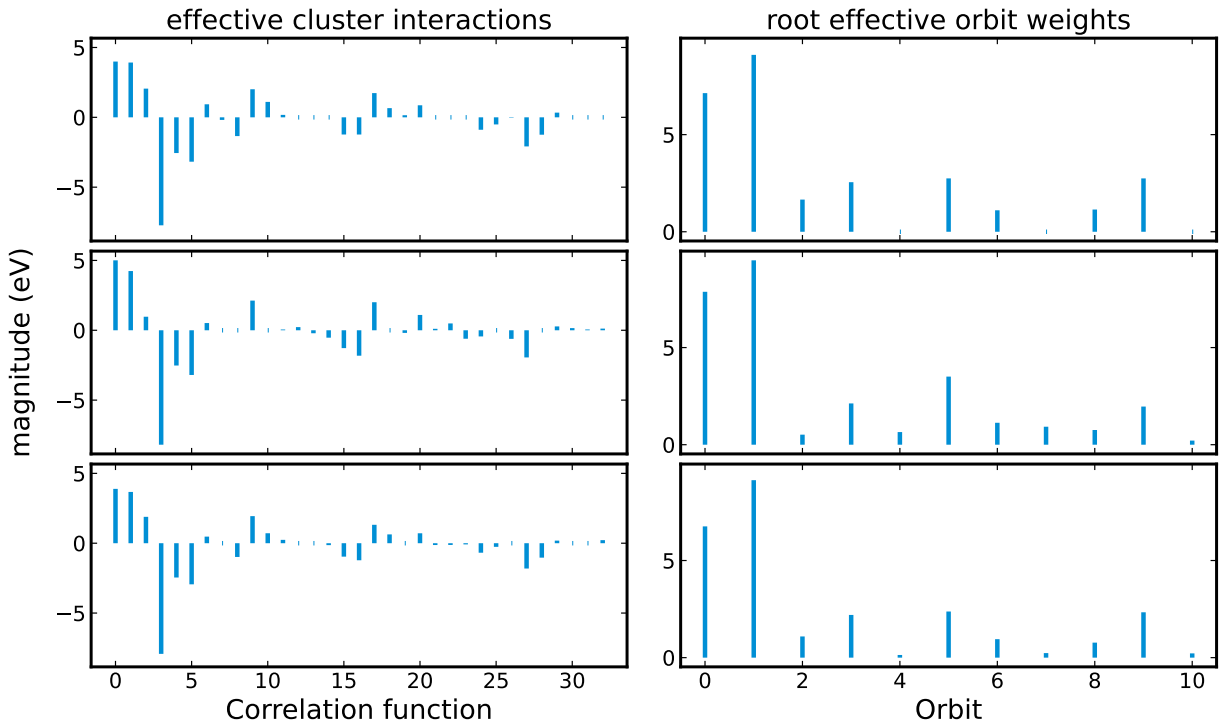


Figure C.3: Fourier cluster expansion parameters (effective cluster interactions) and root effective cluster weights for three fits of an expansion involving only pair terms using the Lasso, and $\ell_2\ell_0$ regression with correlation function hierarchical constraints ($\ell_2\ell_0$) and cluster interaction constraints (Grouped $\ell_2\ell_0$).

Cluster Expansion Fits

Expansion parameters were fitted using the Lasso, $\ell_2\ell_0$ regression with correlation function hierarchy constraints (i.e. singleton groups in Equation 4.19), and Grouped $\ell_2\ell_0$ regression with cluster interaction (per orbit) hierarchy constraints. Fits were done with terms including only pairs up to 8 Å(501 training structures), and terms including pairs up to 10 Åand triplets up to 6 Å(502 training structures). Training structure sampling was done by random sampling over a unit-hyper-sphere as detailed in Chapter 5.1 [152]. A holdout test set of 1000 structures was used for model validation. Hyperparameter optimization was done using 10-fold cross-validation with a two-dimensional grid search to determine the two hyper-parameters in Equation 4.19, a single parameter grid search for the Lasso fits.

Figure C.3 shows the resulting expansion parameters (effective cluster interactions) and corresponding effective cluster weights for the expansion using only pair terms. Feature selection remains constant for all regression methods. However Grouped $\ell_2\ell_0$ is the only method that fully sets to zero some of the interactions, resulting in a sparser model.

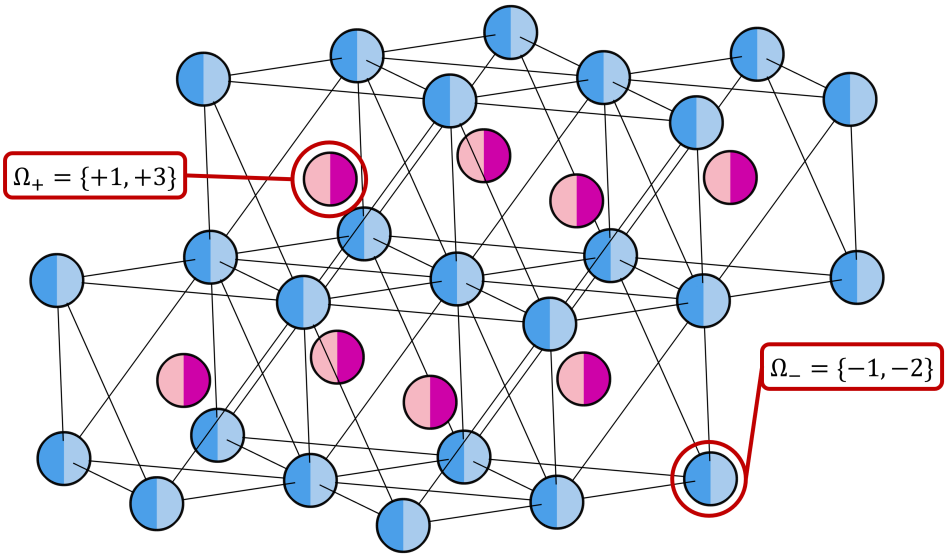


Figure C.4: Disordered rocksalt structure used in computing Coulomb and Buckingham/-Coulomb interaction potentials. The structure includes an FCC anion lattice with allowed species having oxidation -1 or -2 and an FCC cation lattice with allowed species having oxidation +1 or +3.

C.4 Empirical ionic potentials

Coulomb and Buckingham-Coulomb empirical potentials

The Coulomb point electrostatic and Buckingham/Coulomb pair potentials were calculated using `LAMMPS` [219] for a point charge structure with a FCC binary positive charge substructure of $+1$ and $+3$ charges, and a FCC binary negative charge substructure of -1 and -2, as shown in Figure C.4. A total of 3470 ordered structures for supercells ranging from 4 to 144 atoms were enumerated, and the energy based on the two pair potentials was computed for each ordered structure.

The electrostatic calculations were calculated using an Ewald summation method for a Coulomb pair interaction potential.

The Buckingham-Coulomb potential is a commonly used atomic pair potential in the study of ceramic materials. The pair interaction potential between two particles is,

$$\Phi(r_{ij}) = A\exp(-r_{ij}/\rho) - \frac{C}{r_{ij}^6} + \frac{\kappa q_i q_j}{\epsilon r_{ij}} \tag{C.1}$$

where A (eV), ρ (Å), C $(eV \times \text{Å}^6)$ are interaction constants, $\kappa = 1/4\pi\epsilon_0$, and ϵ is a dielectric

Interaction	A (eV)	ρ (Å)	C $(eV \times \text{Å}^6)$
$(+1) \leftrightarrow (+1)$	968.4720	0.2277	94.7183
$(+3) \leftrightarrow (+3)$	4999.2056	0.1685	14.2347
$(-1) \leftrightarrow (-1)$	1206.1667	0.1488	34.3060
$(-2) \leftrightarrow (-2)$	22956.6158	0.2510	11.2060
$(+1) \leftrightarrow (+3)$	4966.7186	0.1085	6.1160
$(+1) \leftrightarrow (-1)$	506.6165	0.2284	62.2407
$(+1) \leftrightarrow (-2)$	476.8532	0.2408	9.9168
$(+3) \leftrightarrow (-1)$	499.0849	0.1990	100.0
$(+3) \leftrightarrow (-2)$	708.8818	0.2103	27.6586
$(-1) \leftrightarrow (-2)$	4987.3754	0.2738	99.2494

Table C.1: Buckingham potential interaction parameters.

constant. The dielectric constant value used in the calculations is $\epsilon = 4.4892$. The parameters corresponding to Buckingham pair interaction are listed in Table C.1.

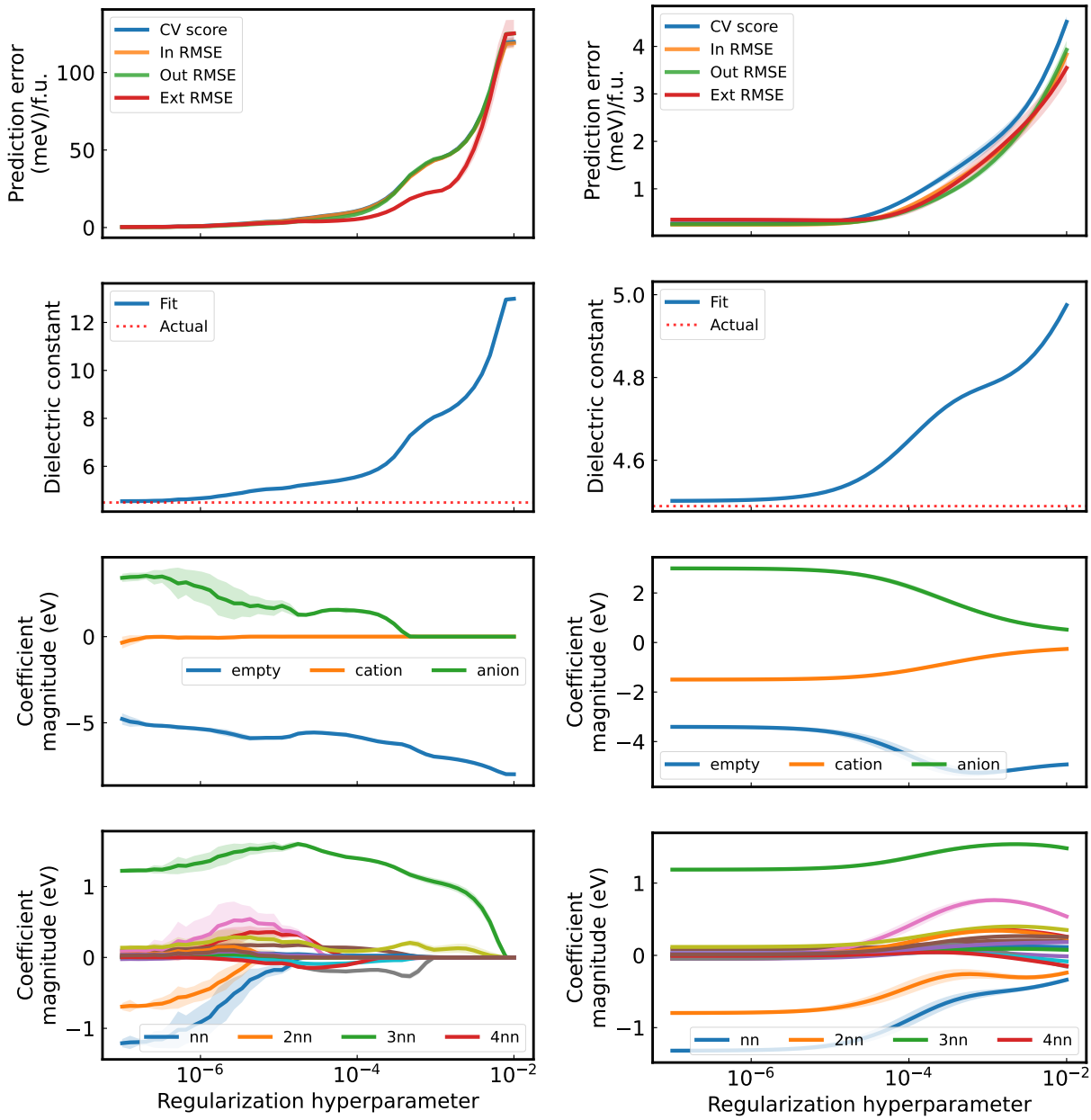
The interaction parameters in Table C.1 were obtained from a multivariate fit to a set of DFT calculated configurations of a Li+/Fe3+/O2−/F− rocksalt structure. The mean squared error between the Buckingham/Coulomb computed energy via `LAMMPS` and the DFT energy was used as the optimization objective function. The fit was done using the limited memory Broyden–Fletcher–Goldfarb–Shanno algorithm with box constraints (L-BFGS-B)[24, 260] algorithm via the `scipy.optimize` Python module [233]. The obtained fit has root mean squared error 40.259 meV/atom. The actual accuracy of the fit is not particularly meaningful. The fit was only carried out to obtain a set of parameters that would result in energy values and variances within an applicable range. The obtained parametrization for the Buckingham/Coulomb potential is taken as a ground truth model for the purpose of precisely quantifying the accuracy and fitted coefficient values of cluster expansions.

Cluster expansion fits

The expansions used to fit the Coulomb potential included correlation functions only (i.e no explicit electrostatic term). An expansion with correlation functions only and one with correlation functions and an electrostatic term were used to fit the Buckingham-Coulomb potential. All fits were carried out including the constant term, all point correlations, and various sets of pair correlations with increasing pair distance. Fits were also done with 3 different training sets: one with structures only up to 16 sites, another with structures up to 36 sites, and the last one with structures up to 64 sites. In all cases, an out-of-sample set of structures with up to the same number of sites used in training was kept for validation. An additional set with structures up to 144 sites was used to test the accuracy of extrapolated predictions to larger super-structures. For all the cases a total of 50 fits randomly shuffling

training and validation structures were carried out.

The Buckingham-Coulomb model is simple enough that feature selection and regularization is not necessary. This can be seen from the results comparing ordinary least squares (OLS), ridge regression, and lasso listed in the main text. Furthermore, the convergence of the model and resulting dielectric at very low hyperparameter values for the regularized regression shown in Figure C.5, similarly suggests that for this simple case regularization is not necessary since both Lasso and Ridge regression converge to effectively the same solution at very low values of the regularization hyperparameter.



(a) Lasso

(b) Ridge

Figure C.5: Convergence of error, correlation function coefficients, and effective dielectric constant with respect to hyperparameter selection for Lasso and Ridge regression models.