Masterarbeit
Zur Erlangung des akademischen Grades Master of Science

Combining Cluster Expansion with Machine Learning Towards Nonlinear Modeling of Materials Properties

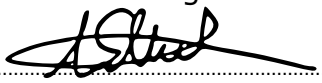
eingereicht von:	Adrian Stroth
Gutachter/innen:	Prof. Dr. Claudia Draxl
	Prof. Dr. Igor Sokolov

Eingereicht am Institut für Physik der Humboldt-Universität zu Berlin am:

8. September 2023

Selbstständigkeitserklärung

Ich erkläre hiermit, dass ich die vorliegende Arbeit selbstständig verfasst und noch nicht für andere Prüfungen eingereicht habe. Sämtliche Quellen einschließlich Internetquellen, die unverändert oder abgewandelt wiedergegeben werden, insbesondere Quellen für Texte, Grafiken, Tabellen und Bilder, sind als solche kenntlich gemacht. Mir ist bekannt, dass bei Verstößen gegen diese Grundsätze ein Verfahren wegen Täuschungsversuchs bzw. Täuschung eingeleitet wird.

Berlin, 

Combining Cluster Expansion with Machine Learning Towards Nonlinear Modeling of Materials Properties

Mathematisch-Naturwissenschaftliche Fakultät
Institut für Physik

Author	Adrian Stroth
Supervisors	Dr. Santiago Rigamonti
	Prof. Dr. Claudia Draxl

September 8, 2023

Contents

Abstract

A focus of material science is the discovery of novel materials for the use in high-tech devices, recently driven by sustainability concerns. One such material is the thermoelectric clathrate $Ba_8Al_xSi_{46-x}$, a potential alternative to GaGe-based compounds. GaGe-based compounds have low thermal and high electrical conductivity, beneficial for thermoelectricity, but both elements are scarce resources. Al and Si, instead, are abundant and cheap. The theoretical description of materials properties is commonly conducted from first-principles using density functional theory. However, this approach becomes unfeasible due to the clathrate's complex configurational space. Instead, the cluster expansion (CE) method is employed. In CE, properties of materials are modeled in terms of substructure interactions within the crystal. While this method has been successfully applied to multicomponent alloys, we show that the CE does not converge and introduces spurious interactions when the property demonstrates a nonlinear configuration dependence. Such nonlinearities occur in the energies of mixing and bandgaps of $Ba_8Al_xSi_{46-x}$. Two solutions with limited usability have been reported in the literature. We propose a novel approach where we treat the standard CE method as solving a linear problem, allowing us to incorporate common machine learning techniques to treat nonlinearities. We call this new solution the nonlinear CE. The idea is to utilize polynomial feature expansion to create nonlinear features (out of the cluster correlations) and employ linear models in combination with cross-validation to identify the optimal feature subset. We use linear models such as LASSO and orthogonal matching pursuit. This method exactly reproduces toy data models with varying degrees of nonlinear configuration dependence while significantly reducing the required number of features compared to standard CE models. In a systematic study of the energy of mixing of $Ba_8Al_xSi_{46-x}$, we find that a 3rd-order nonlinear CE with a small number of initial clusters returns more accurate predictions than standard CE models. Furthermore, our study reveals concentration-dependent 2-point interactions, which exhibit temperature dependence. When fitting bandgaps, our approach demonstrates comparable precision to a standard CE with many clusters despite being initialized with significantly fewer clusters. This points to higher computational efficiency. Finally, we explore the adaptation of CE to classification tasks using support vector machines, extending its potential utility.

1

Introduction

In recent years, in our society, a discussion has arisen about the sustainability of raw materials used in high-tech products. In particular, the commercial distribution of semiconductor-based devices has increased the need for technology-critical elements, which typically occur in host minerals scarcely present in Earth's crust. Their extraction as by-products is energy-intensive and costly, causes large amounts of greenhouse emissions, and negatively impacts biodiversity. One focus of current materials research is therefore on investigating the physical and chemical properties of materials containing technology-critical elements in order to replace them in the long term with not-yet-synthesized materials for use in respective high-tech devices. Such materials ideally show comparable properties without causing adverse effects on the ecology.

One example of such devices is thermoelectrics, which enable the conversion of electricity into heat and vice versa. Thermoelectric generators are highly scalable and free of moving parts and are viewed as potential solutions to recover electric energy from the vast amounts of waste heat generated, for example, by computers, server farms, engines, or lasers. The underlying mechanism, thermoelectricity, is based on carrier diffusion in a material owing to a temperature gradient. For good conversion abilities, thermoelectric materials need a high figure of merit zT, which is approximately given as the temperature-dependent ratio of electric conductivity σ and thermal conductivity κ. Consequently, increasing zT can be achieved by materials with high electrical and low thermal conductivity. However, finding such materials is very challenging since high zT values are hindered by the counterplay of electrical and thermal conductivity: the thermal conductivity κ is the sum of two contributions, namely κ_l, originating from phonons traveling through the crystal, and κ_e, resulting from charge carriers transporting heat. κ_e and σ are roughly proportional through the Wiedemann-Franz law, which means that, in general, minimizing the total thermal conductivity also reduces the electrical conductivity. Thus, the ratio of σ and κ to optimize in the figure of merit would not be affected. Possible solutions are doped small-band-gap semiconductors, where the phonon contribution is larger than the electronic contribution, enabling the decoupling of the electronic and vibrational degrees of freedom. This allows for individual optimization of electronic and thermal conductivity. A class of materials that help to achieve this goal are phonon-glass electron-crystals [1].

An example of phonon-glass electron-crystals are the clathrates, which exhibit a substitutional crystal structure that can be used as a playground for tuning its properties. Clathrates based on gallium and germanium were found to have glass-like thermal conductivity, i.e., very low κ_l, making them promising for thermoelectric applications [2], [3]. Unfortunately, Ga and Ge are only

sparsely present in other minerals and are extracted as their by-products, making their production energy-intensive. Additionally, they depended on the mining of the primary material and, thus, are potentially unsustainable. Therefore, it is of considerable interest to substitute these elements with more abundant, cheaper, and lighter elements while retaining the beneficial properties. In the case of GaGe-based clathrates, this could be achieved by replacing Ga and Ge with the isoelectronic species Al and Si, respectively. AlSi-based clathrates have been the subject of experimental [4]–[7] and theoretical studies [8]–[10], and are taken as an application example for the methods introduced in this thesis.

The standard approach for the *ab initio* theoretical description of materials properties is density functional theory (DFT). DFT utilizes a fictitious system of non-interacting electrons to compute the electronic density and ground state energy of the material. Hence, in principle, it can be employed to find the stability of different clathrate compositions in the search for new clathrate materials. For clathrates, not only the amount per species in the compound, the composition, but also their location in the crystal lattice, the configuration, are relevant for the theoretical description of their properties. This was demonstrated in Ref. [8], where it was found that properties such as the total energy and the electronic structure are very sensitive to the configuration. However, the compositional and configurational space of clathrates is astronomically large. Just for the unit cell of the clathrate, finding the structure of lowest energy would require the computation of around 10^{10} different structures. Due to the computational effort, such a task is infeasible for DFT.

One way to approach this problem is to use the cluster expansion (CE) method [11]. It enables the modeling of materials properties that depend on the configuration, which in turn determines the structure. This method has been employed in a wide range of problems and systems, including the calculation of phase diagrams [12]–[15], surface alloying [16], and catalysis [17]. The models obtained with CE enable the evaluation of materials properties at a minimal computational cost, making it an ideal tool to address the above-mentioned problem.

Nevertheless, CE encounters difficulties when modeling properties with nonlinear dependencies on configuration. For instance, it has been demonstrated that the CE of properties that depend quadratically on the concentration of a binary alloy cannot be converged [18]. This problem has also been the topic of recent discussions in the literature [19], [20]. In fact, for clathrate materials, it was found that the CE model of the energy of formation was hindered by the presence of a nonlinear kink in the energy-versus-concentration curve [8], [10].

In this work, in order to propose a novel solution to the problem, we show that the CE method can be conveniently formulated such that standard machine learning (ML) techniques can be applied. This novel solution is inspired by methods common in the ML community. After describing the thermoelectric clathrate $Ba_8Al_xSi_{46-x}$ in more detail in the subsequent section, Sec. 2.2 and Sec. 2.3 introduce the theoretical grounds of the CE method and the nonlinear problem, respectively. Sec. 3 presents the novel ML-based solution. Its application to the energy of mixing and band gap energies of an AlSi-based clathrate are described in Sec. 4.3 and Sec. 4.4, respectively. In Sec. 4.5, we utilized the nonlinear CE to obtain a materials descriptor that can be used for metal-semiconductor classification. Finally, the implementation of the novel method into the Python CE package CELL is shown in Sec. 5.

2

Theoretical Background

In the following, we introduce density functional theory (DFT) in Sec. 2.1, as it is a precursor to most cluster expansion (CE) applications and also has been employed to obtain the dataset that will be used for the studies in Sec. 4. Subsequently, Sec. 2.2 and Sec. 2.3 introduce the CE method and discuss the difficulties that arise when expanding nonlinear properties, respectively.

2.1 Density Functional Theory

Density functional theory is a first-principles method to compute the electronic structure of materials and their properties. Since its introduction in Ref. [21] it has been successfully employed for describing the ground state properties of various crystal and molecular systems.

In DFT, the system of interacting electrons is replaced by an auxiliary system of non-interacting electrons, yielding the same electronic density as the interacting one. Kohn and Sham introduced the auxiliary system [22] and, together with the Hohenberg Kohn (HK) theorem [21], it constitutes the core of DFT. The HK theorem postulates the one-to-one relationship between external potentials and electronic densities. This allows an enormous simplification since instead of having to deal with the many-body quantum wave function, which depends on the coordinates of all electrons in the system, the problem can be recast in terms of the electronic density, which only depends on three spatial coordinates. The Kohn-Sham equation in atomic units is given by

$$\left[-\frac{1}{2}\Delta^2 + v_{KS}(\boldsymbol{r})\right]\Psi_i(\boldsymbol{r}) = \epsilon_i \Psi_i(\boldsymbol{r}) \ .$$

Here, the $\Psi_i(\boldsymbol{r})$ represent the single-particle wavefunctions that yield the electron density by

$$n(\boldsymbol{r}) = \sum_{i=1}^{N} |\Psi_i(\boldsymbol{r})|^2 \ ,$$

where ϵ_i are the eigenenergies, and v_{KS} is the Kohn-Sham potential. The Kohn-Sham potential is the sum of the external potential v_{ext} (e.g., the electrostatic potential of the fixed nuclei, as seen

by the electron), the Hartree potential v_H, which represents the classical Coulomb electrostatic interaction of the electrons with the electronic cloud, and the exchange-correlation potential v_{xc}:

$$v_{KS} = v_{ext} + v_H + v_{xc} \; .$$

The exchange-correlation potential is a functional derivative defined as

$$v_{xc}(\boldsymbol{r}) = \frac{\delta E_{xc}[n(\boldsymbol{r})]}{\delta n(\boldsymbol{r})} \; .$$

The exchange-correlation energy E_{xc} contains all remaining classical and quantum correlation effects and is defined in terms of the difference between the ground state energy computed with the full many-body wave function and the one computed with the single Kohn-Sham Slater determinant.

In principle, the Kohn-Sham scheme yields the exact ground state energies, provided that the exact E_{xc} is known. Unfortunately, E_{xc} and $v_{xc}(\boldsymbol{r})$ are unknown, requiring their approximation. Over the years, multiple approximations have been developed, differing in levels of sophistication. The simplest is the local density approximation (LDA), which is based on replacing the exchange-correlation energy density at point $\boldsymbol{r}$ in space by the exchange-correlation energy density of the homogeneous electron gas with density $n(\boldsymbol{r})$, $\varepsilon_{xc}^{hom}(n(\boldsymbol{r}))$. Thus, the exchange-correlation energy in the LDA approximation is given by

$$E_{xc}^{LDA} = \int \varepsilon_{xc}^{hom}(n(\boldsymbol{r}))\, dr \; .$$

A more sophisticated approximation is the generalized gradient approximation (GGA) in which the exchange-correlation energy density depends on the electron density $n(\boldsymbol{r})$ and its gradient $\nabla_{\boldsymbol{r}} n$. The GGA functional used in Ref. [8], relevant for Sec. 4, is a version of the functional introduced by Perdew, Burke, and Ernzerhof [23] that was specifically adapted to solids, called PBEsol [24].

2.2 Cluster Expansion

The theoretical description of materials properties often begins with a ground state search. However, for materials with a large unit cell, the problem of finding the ground state is complicated because the configurational space becomes too vast for DFT to computationally handle. However, it is possible to cheaply predict property values of arbitrary configurations with the CE method, which only requires the *a priori* calculated values for a subset of structures. Thus, CE can be used for the theoretical description of properties presenting a combinatorial explosion of the configurational space.

Cluster expansion builds on the relationship between the property of a material and the configuration of its structure. The idea was first introduced by Kikuchi in 1955 [25], and a general formalism for the cluster expansion for multicomponent systems was described by Sanchez, Ducastelle, and Gratias in 1984 [11]. A mulitcomponent system refers to a system composed of more than one constituents. For materials, these could be different elements, chemical compounds, phases of matter, or vacancies.

A multicomponent material is described by the number of crystal lattice sites N and their occupation with different species. The occupation of a lattice site is indicated by the occupation variable $\sigma_i \in \{0, ..., M-1\}$ which in the example of a binary alloy of atoms A and B results in $\sigma = 0$ and $\sigma = 1$ describing the occupation with A and B, respectively. An arbitrary configuration is then

represented by the vector of occupation variables at all lattice sites $\boldsymbol{\sigma} = (\sigma_1, \sigma_2, ..., \sigma_N)$, and thus, any configuration-dependent property of a material can be written as a function of $\boldsymbol{\sigma}$. In the cluster expansion, any such configuration-dependent property is expanded in terms of cluster functions $\Gamma_{\boldsymbol{\alpha}}(\boldsymbol{\sigma})$, which form a basis set in configurational space. The expansion is given by

$$P(\boldsymbol{\sigma}) = \sum_{\boldsymbol{\alpha}} J_{\boldsymbol{\alpha}} \Gamma_{\boldsymbol{\alpha}}(\boldsymbol{\sigma}) \,, \tag{1}$$

where the sum runs over all clusters $\boldsymbol{\alpha}$. Clusters are sets of crystal sites representing n-body interactions. The expansion coefficients $J_{\boldsymbol{\alpha}}$ are called the *effective cluster interactions* (ECI), and the cluster functions form a complete and orthonormal basis with respect to the scalar product

$$\langle\, \alpha \mid \beta \,\rangle = \sum_{\sigma} \Gamma_\alpha(\sigma)\, \Gamma_\beta(\sigma) = \delta_{\alpha\beta}. \tag{2}$$

A cluster function corresponding to $\boldsymbol{\alpha}$, $\Gamma_{\boldsymbol{\alpha}}(\boldsymbol{\sigma})$, is given by the product of site basis functions $\gamma_i(\sigma_i)$ over all lattice sites $i \in \{1, ..., N\}$,

$$\Gamma_{\boldsymbol{\alpha}}(\boldsymbol{\sigma}) = \prod_{i=1}^{N} \gamma_{\alpha_i}(\sigma_i) \,. \tag{3}$$

The site basis functions fulfill the orthonormality condition:

$$\sum_{\sigma=0}^{M-1} \gamma_\alpha(\sigma)\gamma_\beta(\sigma) = \delta_{\alpha\beta} \,. \tag{4}$$

Note that the domain of the functions γ comprises integer scalars (i.e., the occupation variable of a single site), while the domain of functions Γ comprises integer vectors (i.e., the crystal configuration). The orthonormality property of the γ functions, Eq. 4, guarantees the orthonormality of the cluster functions Γ [11], [26]. In the binary case, the set of site basis functions per lattice site σ_i is $\{\gamma_0(\sigma_i), \gamma_1(\sigma_i)\}$. Generally, site basis functions are chosen such that they include a constant function. For $\sigma = 0, 1$, a possible set is given by

$$\begin{aligned} \gamma_0(\sigma) &= 1 \\ \gamma_1(\sigma) &= -cos(\pi\sigma) = -1, 1 \,. \end{aligned} \tag{5}$$

The choice of the constant basis function $\gamma_0(\sigma_i) = 1$ ensures that in Eq. 3 only the evaluation of the site basis functions of sites with $\alpha_i \neq 0$ are relevant. Thus, the product in Eq. 3 can be restricted to the *cluster* of sites for which $\alpha_i \neq 0$. This allows the definition of clusters as the set

$$\boldsymbol{\alpha} = \{\alpha_i | \alpha_i \neq 0\} \,. \tag{6}$$

More generally, clusters may be defined by the vector $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, ..., \alpha_N)$, where the variables $\alpha_i \in \{0, ..., M-1\}$ indicate the choice of basis function per lattice site. This more general definition does not require the introduction of the constant function in the set of site basis functions.

Using the basis defined in Eq. 5, an example of a 2-point cluster is a vector of N elements with $\alpha_2 = 1$, $\alpha_3 = 1$, and $\alpha_{i\neq 2,3} = 0$. Consequently, because of Eq. 6, Γ_{2p} reduces to the product of the site basis functions at sites σ_2 and σ_3,

$$\begin{aligned} \Gamma_{2p}(\boldsymbol{\sigma}) &= \gamma_0(\sigma_1) \times \gamma_1(\sigma_2) \times \gamma_1(\sigma_3) \times \cdots \times \gamma_0(\sigma_N) \\ &= \gamma_1(\sigma_2) \times \gamma_1(\sigma_3) \,. \end{aligned}$$

Note that so far, all clusters in a crystal that are equivalent under its symmetry operations count as distinct clusters. However, two such symmetrically equivalent clusters add an equal contribution to the expansion, i.e., they have equal ECIs. To understand this, one must keep in mind that the vector of indices $\boldsymbol{\alpha}$ of a cluster function transforms in the same way as the configuration vector $\boldsymbol{\sigma}$ under the symmetry operations of the crystal. That is, for any crystal symmetry S, $\Gamma_{\boldsymbol{\alpha}}(\boldsymbol{\sigma}) = \Gamma_{S\boldsymbol{\alpha}}(S\boldsymbol{\sigma})$. Toghether with $P(\boldsymbol{\sigma}) = P(S\boldsymbol{\sigma})$, this implies $J_{\boldsymbol{\alpha}} = J_{S\boldsymbol{\alpha}}$. Consequently, the summation over all clusters in Eq. 1 can be divided into a sum over symmetrically inequivalent (s.i.) and a sum over symmetrically equivalent clusters,

$$\begin{aligned} P(\boldsymbol{\sigma}) &= \sum_{\boldsymbol{\alpha}} J_{\boldsymbol{\alpha}} \Gamma_{\boldsymbol{\alpha}}(\boldsymbol{\sigma}) \\ &= \sum_{\boldsymbol{\alpha}}^{s.i.} J_{\boldsymbol{\alpha}} \sum_{\beta \equiv \boldsymbol{\alpha}} \Gamma_{\beta}(\boldsymbol{\sigma}). \end{aligned}$$

Here, $\boldsymbol{\beta} \equiv \boldsymbol{\alpha}$ denotes the sum over all clusters $\boldsymbol{\beta}$ symmetrically equivalent to $\boldsymbol{\alpha}$. Furthermore, the number of symmetrically equivalent clusters is introduced as the *multiplicity* of cluster $\boldsymbol{\alpha}$, $M_{\boldsymbol{\alpha}}$. As the sum over all clusters $\boldsymbol{\beta}$ symmetrically equivalent to $\boldsymbol{\alpha}$ has $M_{\boldsymbol{\alpha}}$ terms, by inserting $M_{\boldsymbol{\alpha}}/M_{\boldsymbol{\alpha}}$, one obtains the average of all symmetrically equivalent cluster functions:

$$\begin{aligned} P(\boldsymbol{\sigma}) &= \sum_{\boldsymbol{\alpha}}^{s.i.} J_{\boldsymbol{\alpha}} M_{\boldsymbol{\alpha}} \frac{1}{M_{\boldsymbol{\alpha}}} \sum_{\beta \equiv \boldsymbol{\alpha}} \Gamma_{\beta}(\boldsymbol{\sigma}) \\ &= \sum_{\boldsymbol{\alpha}}^{s.i.} J_{\boldsymbol{\alpha}} M_{\boldsymbol{\alpha}} \langle \Gamma_{\beta}(\boldsymbol{\sigma}) \rangle_{\boldsymbol{\alpha}}. \end{aligned}$$

The average defines the correlation of cluster $\boldsymbol{\alpha}$ with the configuration $\boldsymbol{\sigma}$,

$$X_{\boldsymbol{\alpha}}(\boldsymbol{\sigma}) = \frac{1}{M_{\boldsymbol{\alpha}}} \sum_{\beta \equiv \boldsymbol{\alpha}} \Gamma_{\beta}(\boldsymbol{\sigma}) \quad . \tag{7}$$

It can be read as counting how often cluster $\boldsymbol{\alpha}$ is present in $\boldsymbol{\sigma}$ divided by how many possibilities exist in the crystal for $\boldsymbol{\alpha}$ to exist. Finally, with the cluster correlations, the cluster expansion of a property of configuration is written as

$$P(\boldsymbol{\sigma}) = \sum_{\boldsymbol{\alpha}}^{s.i.} J_{\boldsymbol{\alpha}} M_{\boldsymbol{\alpha}} X_{\boldsymbol{\alpha}}(\boldsymbol{\sigma}) \, . \tag{8}$$

Inspecting the vectors of configuration $\boldsymbol{\sigma}$, one can see that the configurational space is of dimension M^N. The same is true for the number of possible clusters in a crystal. Thus, the number of configurations is equal to the number of clusters, which becomes infinite in the thermodynamic limit. Consequently, Eq. 8 is an expansion over an infinite number of clusters. As this cannot realistically be computed, it is approached by limiting the expansion to finite sets of compact clusters [12], [27]–[29]. These sets usually contain clusters with only a few sites and close-by neighbors, such as 1-point, 2-point first neighbor, or 3-point first neighbor clusters. This consideration can be regarded as equivalent to the truncation of an infinite Fourier transform after the lowest and most relevant frequencies. In the case of the energy of formation of an alloy, a converging CE that can be reduced to a finite sum of expansion coefficients casts the energy of formation in the form of an Ising-type model.

Lastly, it should be mentioned that in the CE literature, different choices can be found for the domain of configuration variables σ_i. For instance, in Ref. [11], $\sigma = -1, 1$ for a binary and $\sigma = -1, 0, 1$

for a ternary are used, instead of $\sigma = 0, 1$ and $\sigma = 0, 1, 2$ as above. Also, occasionally, alternative site basis functions to the trigonometric basis of Eq. 5 are employed. These include Chebyschev polynomials [11], [30], indicator functions [31], and pair probabilities [30]. A very convenient basis for the case of binary systems is the indicator function basis (IFB) defined by

$$\{\gamma_0(\sigma_i), \gamma_1(\sigma_i)\} = \{1, \sigma_i\} \quad \text{with} \quad \sigma = 0, 1 \; . \tag{9}$$

Although lacking orthogonality, the IFB is complete and enables the direct representation of a substituent's concentration with the 1-point cluster correlation $X_{1p}(\boldsymbol{\sigma})$. That is, the cluster correlation of the 1-point cluster is identical to the substituent concentration of the alloy defined by $x = N_B/(N_A + N_B)$. It also allows for the identification of clusters as distinct substructures of substituents at the pristine crystal sites.

2.3 Cluster Expansion of Nonlinear Properties

The cluster expansion method has been used successfully to study structure stabilities, phase diagrams, order-disorder transitions, electrochemical properties, or thermodynamics of substitutional materials, such as alloys [12]–[15], [32]–[34]. However, if the property has a nonlinear dependence on the concentration of substituents in the material, its CE does not always converge, and a truncated CE might not accurately represent the material's property. This can be shown, for example, in a binary alloy. In this case, while the CE of the energy of alloy formation normally converges, one can show that, in general, it cannot necessarily be reduced to a finite sum of expansion terms [35]. In fact, a cluster expansion of a property with nonlinearities typically results in an infinite expansion [18], [19].

To show this, we assume the crystal of a binary alloy with N atoms, consisting of N_A atoms A and N_B atoms B, and a property of configuration given by

$$P(\boldsymbol{\sigma}) = x(\boldsymbol{\sigma})^2 \; , \tag{10}$$

where x is the concentration of substituent B in the crystal. Next, we will demonstrate how x depends on the configuration. For this example, we use the domain of site occupation variable $\sigma = 0, 1$ with the IFB as in Eq. 9. Then, with Eq. 7 and Eq. 3, the 1-point cluster correlations X_{1p} are given by

$$\begin{aligned} X_{1p}(\boldsymbol{\sigma}) &= \frac{1}{M_{1p}} \sum_{\beta \in 1p} \prod_{i=1}^{N} \gamma_{\beta_i}(\sigma_i) \\ &= \frac{1}{M_{1p}} \sum_{\beta \in 1p} \prod_{i \in \{\beta_i \neq 0\}} \gamma_{\beta_i}(\sigma_i) \\ &= \frac{1}{N} \sum_{i=1}^{N} \sigma_i \\ &= \frac{N_B}{N} = x \; . \end{aligned} \tag{11}$$

Here in the second line, we have used the fact that $\gamma_0(\boldsymbol{\sigma}) = 1$. Next, we consider that there exist N symmetrically equivalent clusters $\boldsymbol{\beta}$ to the 1-point cluster, $M_{1p} = N$. Thus, summing σ_i at all lattice sites is equivalent to counting the number of substituents N_B.

Now, we will derive the *exact* CE of the property given by Eq. 10. From Eq. 11, we see that the concentration is given by the 1-point cluster correlations, and thus,

$$\begin{aligned} P(\boldsymbol{\sigma}) &= X_{1p}(\boldsymbol{\sigma})X_{1p}(\boldsymbol{\sigma}) \\ &= (\frac{1}{N}\sum_{i=1}^{N}\sigma_i)(\frac{1}{N}\sum_{j=1}^{N}\sigma_j) \\ &= \frac{1}{N^2}\sum_{i=1}^{N}\sigma_i\sigma_i + \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j\neq i}\sigma_i\sigma_j \\ &= \frac{1}{N}X_{1p}(\boldsymbol{\sigma}) + \frac{1}{N^2}\sum_{\boldsymbol{\alpha}\in 2p}\Gamma_{\boldsymbol{\alpha}}(\boldsymbol{\sigma})\ . \end{aligned} \tag{12}$$

Here, we obtain a double sum that can be split into a part over equal lattice sites $i = j$ and a part over unequal lattice sites $i \neq j$. It is evident that the first part is equal to the derived 1-point cluster correlations. The double sum of unequal lattice sites evaluates the cluster function of all possible 2-point clusters in the structure, with the sum containing $N(N-1)$ terms. The ECIs are given by $J_{1p} = 1/N^2$ for the single 1-point cluster and by $J_{2p} = 1/N^2$ for *all* 2-point clusters. Now, in the thermodynamic limit, the number of 2-point clusters becomes infinite, whereas all their coefficients equally approach zero. Consequently, the CE of x^2 is exactly represented by the 1-point cluster and an infinite expansion in 2-point clusters, each with the same infinitesimally small coefficient.

As mentioned in Sec. 2.2, a CE of the property of a crystal is, in principle, always an infinite expansion in the number of clusters and is usually truncated. However, as demonstrated in Eq. 12 the CE of the property given in Eq. 10 does not converge, and a truncation of the expansion results in spurious interactions. This can be illustrated by assuming the exact CE of an arbitrary property $P(\boldsymbol{\sigma})$, containing infinite terms, and a summation over clusters in a finite set $\mathcal{C}$, representing the cutoff. Then the truncated CE can be expressed by

$$\begin{aligned} P_{truncated}(\boldsymbol{\sigma}) &= \sum_{\boldsymbol{\alpha}\in\mathcal{C}} J_{\boldsymbol{\alpha}} M_{\boldsymbol{\alpha}} X_{\boldsymbol{\alpha}}(\boldsymbol{\sigma}) \\ &= \sum_{\boldsymbol{\alpha}} J_{\boldsymbol{\alpha}} M_{\boldsymbol{\alpha}} X_{\boldsymbol{\alpha}}(\boldsymbol{\sigma}) - \sum_{\boldsymbol{\alpha}\notin\mathcal{C}} J_{\boldsymbol{\alpha}} M_{\boldsymbol{\alpha}} X_{\boldsymbol{\alpha}}(\boldsymbol{\sigma}) \\ &= P_{exact}(\boldsymbol{\sigma}) - \sum_{\boldsymbol{\alpha}\notin\mathcal{C}} J_{\boldsymbol{\alpha}} M_{\boldsymbol{\alpha}} X_{\boldsymbol{\alpha}}(\boldsymbol{\sigma})\ . \end{aligned} \tag{13}$$

It can be seen that the truncated CE is an approximation of the infinite expansion only if the ECIs of clusters not included in $\mathcal{C}$ are negligible. In the example of Eq. 12, this is not the case; hence, the truncated ECIs introduce a spurious configuration dependence given by the second term on the right-hand side of Eq. 13.

Next, we want to consider the energy of formation of the random alloy, which is given for the totally disordered configuration. The random alloy is characterized by the fact that a lattice site occupation is independent of the occupation of all other lattice sites, i.e., two sites i and j are uncorrelated. This occurs, for instance, in the high-temperature limit. This fact can be used to show that, in a random alloy, the cluster correlation of any 2-point cluster is identical to the squared concentration. For simplicity, we prove this for the 2-point cluster corresponding to k-nearest neighbor sites in a linear chain. The proof for any other p-point cluster in a general crystal is analogous.

$$X_{2p}(\sigma) = \frac{1}{N}\sum_{i}\sigma_i\sigma_{i+k} = Pr(\sigma_\bullet = 1 \wedge \sigma_{\bullet+k} = 1)\ , \tag{14}$$

where we use the fact that for an infinite system, the sum on the r.h.s. is equal to the joint probability of having simultaneous occupation of an arbitrary site (denoted by $\bullet$ and $\bullet + k$) with atoms of type B, which is denoted by $Pr(\sigma_\bullet = 1 \wedge \sigma_{\bullet+k} = 1)$. A random alloy is defined by the property of having uncorrelated site occupations, i.e.,

$$Pr(\sigma_\bullet = 1 \wedge \sigma_{\bullet+k} = 1) = Pr(\sigma_\bullet = 1)Pr(\sigma_{\bullet+k} = 1) \; . \tag{15}$$

Again, for an infinite alloy, we can identify

$$Pr(\sigma_\bullet = 1) = \frac{1}{N}\sum_i \sigma_i = X_{1p}(\sigma) \; . \tag{16}$$

Using equations Eq. 15 and Eq. 16 in Eq. 14, we obtain

$$X_{2p}(\sigma) = X_{1p}(\sigma)^2 \; . \tag{17}$$

Analogously, one derives for the 3-point and any p-point cluster $X_{3p}(\sigma) \to x^3$ and $X_{pp}(\sigma) \to x^p$, respectively. This means Eq. 8 for the energy of formation of the random alloy is a power expansion in concentration,

$$P(\boldsymbol{\sigma}) = (\sum_{\alpha \in 1p} J_\alpha m_\alpha)x + (\sum_{\alpha \in 2p} J_\alpha m_\alpha)x^2 + (\sum_{\alpha \in 3p} J_\alpha m_\alpha)x^3 + \dots \; , \tag{18}$$

where $1p$, $2p$, and $3p$, denote sets of all 1-point, 2-point, and 3-point clusters, respectively.

Equation 18 has been derived in Ref. [19] to demonstrate that the standard CE is able to describe nonlinear behavior. This equation also shows that for a property as given in Eq. 10, and in the case of a random alloy, a CE containing any number of 2-point clusters would suffice to capture the x^2 dependence. This expression of the CE is used in Ref. [19] to argue that one can use the standard CE to represent nonlinearities of concentration in a property. It is pointed out that this is the case for the CE of the regular solution model, which can be used to describe the enthalpy of formation in alloys and is given by

$$H(\boldsymbol{\sigma}) = \omega\frac{(1-x^2)}{4} \; ,$$

where x is the concentration and ω a constant. However, it is essential to understand that the expression in Eq. 18 only holds for the special case of random alloys, and hence it alone does not solve the problem of nonlinearities in CE.

A possible solution is proposed in Ref. [18] and is called the variable basis cluster expansion. Its core idea is the introduction of a set of concentration-dependent basis functions, where the concentration dependence is set to match the concentration of substituents in the structure whose property is expanded. With regard to the formation energy of alloys, the variable basis cluster expansion separates the expansion into a concentration-dependent random alloy energy term and a concentration-independent ordering energy term. The former then incorporates the nonlinear concentration dependence that has been shown to converge slowly, while the latter converges rapidly and can be treated with the standard CE methods but has concentration-dependent expansion coefficients. For practical purposes, this approach implies applying a different basis set per alloy composition, questioning its usefulness.

From this discussion, it is evident that the main reason nonlinearities in materials properties can hinder the effectiveness of a CE lies in obtaining an infinite expansion that cannot be effectively truncated. Accordingly, to efficiently model the properties of materials, a modification to the CE method must be found that enables converging expansions of nonlinear properties. This is the main goal of this work.

3

THE NONLINEAR CLUSTER EXPANSION

In this section, we discuss the utilization of CE in practice and expose how this usage can be regarded as a machine learning problem in Sec. 3.1 and Sec. 3.2, respectively. This serves as a foundation for introducing the novel method developed in this work — the nonlinear CE — in Sec. 3.3. Lastly, we apply the nonlinear CE to an illustrative toy problem in Sec. 3.4.

3.1 Cluster Expansion in Practice

The previous two sections have shown that the CE of a property (Eq. 8) is generally an expansion over infinite clusters. For a well-converging CE, the cluster basis can be truncated while retaining a good accuracy of the expansion. However, in the case of a property with nonlinearities, the CE can necessitate an infinite expansion with infinitesimally small coefficients, where a truncated cluster basis might not result in a converging CE. That being said, in both cases, the ECIs are still unknown.

Mathematically, the ECI of an arbitrary cluster $\boldsymbol{\alpha}_k$, $J_{\boldsymbol{\alpha}_k}$, can be derived from Eq. 1 by use of the scalar product Eq. 2, performing a basis transformation from α-space to σ-space:

$$\begin{aligned}
\langle\, \boldsymbol{\alpha}_k \mid P \,\rangle &= \sum_{\boldsymbol{\sigma}} \Gamma_{\boldsymbol{\alpha}_k}(\boldsymbol{\sigma}) P(\boldsymbol{\sigma}) \\
&= \sum_i J_{\boldsymbol{\alpha}_i} \sum_{\sigma} \Gamma_{\boldsymbol{\alpha}_k}(\boldsymbol{\sigma}) \Gamma_{\boldsymbol{\alpha}_i}(\boldsymbol{\sigma}) \\
&= \sum_i J_{\boldsymbol{\alpha}_i} \delta_{k,i} \\
&= J_{\boldsymbol{\alpha}_k} \ .
\end{aligned}$$

Thus, evaluating the ECIs requires knowing the property value of every configuration, which for a crystal is an infinite number of values. Furthermore, even if all these values were known, this would also include the property value of the configuration to be estimated with CE in the first place, which contradicts the whole idea.

In practice, the ECIs are approximated by obtaining the materials properties for a subset of structures $\mathcal{S} = \{\boldsymbol{\sigma}_1, \boldsymbol{\sigma}_2, \ldots, \boldsymbol{\sigma}_{N_s}\}$ and fitting the property values $P(\mathcal{S}) = \{P(\boldsymbol{\sigma}_1), P(\boldsymbol{\sigma}_2), \ldots, P(\boldsymbol{\sigma}_{N_s})\}$ to find

optimal ECIs. Here, N_s is the number of structures in the subset. Employing *ab-intio* methods such as DFT is the standard procedure to calculate $P(\mathcal{S})$. Note that the accuracy of these DFT calculations defines the maximum possible precision that can be reached with CE. Additionally, as mentioned in Sec. 2.2, the summation over infinite clusters in Eq. 8 is cut off by considering finite sets of clusters $\mathcal{C} = \{\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \ldots, \boldsymbol{\alpha}_{N_c}\}$. Then, with a subset of structures $\mathcal{S}$ and a finite set of clusters $\mathcal{C}$, one can write Eq. 8 in matrix from as

$$\hat{\boldsymbol{P}} = \boldsymbol{X}\boldsymbol{\mathcal{J}} \,. \tag{19}$$

Here, $\boldsymbol{\mathcal{J}}$ denotes the N_c-dimensional vector of effective cluster interactions and $\hat{\boldsymbol{P}}$ is the N_s-dimensional vector of predicted property values, i.e., the values are approximations of the true values $P(\mathcal{S})$ due to using subsets of all clusters and structures. Note that $\boldsymbol{\mathcal{J}}$ includes the multiplicities: $\boldsymbol{\mathcal{J}}_{\boldsymbol{\alpha}} = m_{\boldsymbol{\alpha}} J_{\boldsymbol{\alpha}}$. $\boldsymbol{X}$ is the $N_s \times N_c$-dimensional matrix of correlations,

$$\boldsymbol{X} = \begin{array}{c} \\ \text{structures} \downarrow \end{array} \overset{\text{clusters} \longrightarrow}{\begin{pmatrix} X_{1,1} & \ldots & X_{1,N_c} \\ \vdots & \ddots & \vdots \\ X_{N_s,1} & \ldots & X_{N_s,N_c} \end{pmatrix}} ,$$

where the matrix element $X_{i,j}$ is given by the cluster interaction between structure $\boldsymbol{\sigma}_i$ and cluster $\boldsymbol{\alpha}_j$, $X_{i,j} = X_{\boldsymbol{\alpha}_j}(\boldsymbol{\sigma}_i)$.

3.2 CE Viewed as Machine Learning Problem

The representation of the CE method as the linear problem in Eq. 19 is convenient because it suggests the utilization of machine learning (ML) methods, such as linear regression or compressed sensing [36] for finding the model coefficients J yielding optimal property predictions. For the connection between CE and ML, we introduce ML terminology in Table 3.1 that will be used interchangeably with their CE counterparts. Additionally, we will call the CE of a property with a distinct set of ECIs a CE model or, simply, a model.

Table 3.1. Synonyms between terms used in cluster expansion and machine learning as used in this work.

CE term	ML term
Correlation matrix	Input matrix
Structures	Samples
Clusters	Features
Property values	Targets
Predicted properties	Predictions
ECIs	Coefficients
CE model	Model

In ML, the simplest approach to solve the linear problem posed by Eq. 19 is linear regression. In linear regression, one minimizes the residual sum of squares (RSS) with respect to the coefficients $\boldsymbol{\mathcal{J}}$. The RSS serves as cost function ($C(\boldsymbol{\mathcal{J}})$) and is given by

$$C_{LR}(\boldsymbol{\mathcal{J}}) = \sum_{i=1}^{N_s} (P_i - \hat{P}_i)^2 = \|\boldsymbol{P} - \hat{\boldsymbol{P}}\|_2^2 \,. \tag{20}$$

Here $\|\cdot\|_2^2$ is the squared ℓ^2-norm, which is given by using $p = 2$ in the ℓ^p-norm defined by

$$||u||_p = (\sum_i |u_i|^p)^{\frac{1}{p}} .$$

After inserting Eq. 19, Eq. 20 can be cast as a quadratic form,

$$C_{LR}(\boldsymbol{\mathcal{J}}) = \boldsymbol{P}^\top \boldsymbol{P} - \boldsymbol{P}^\top \boldsymbol{X}\boldsymbol{\mathcal{J}} - \boldsymbol{\mathcal{J}}^\top \boldsymbol{X}^\top \boldsymbol{P} + \boldsymbol{\mathcal{J}}^\top \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\mathcal{J}} . \tag{21}$$

Then, the optimal coefficient $\boldsymbol{\mathcal{J}}$ can be obtained by minimizing the cost function,

$$\frac{\partial}{\partial \boldsymbol{\mathcal{J}}^*} C_{LR}(\boldsymbol{\mathcal{J}}^*) = 0 , \tag{22}$$

which yields

$$\boldsymbol{\mathcal{J}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{P} . \tag{23}$$

Graphically, the contour surfaces of Eq. 21 can be represented as hyperellipsoid in feature space. The solution yielding the smallest RSS lies in the center of the ellipsoids and is denoted by $\mathcal{J}_{LR}$ in Fig. 3.1. Furthermore, Eq. 23 tells us that finding $\boldsymbol{\mathcal{J}}$ requires inverting $\boldsymbol{X}^\top \boldsymbol{X}$, and thus it must be of full rank, i.e., $\text{rank}(\boldsymbol{X}^\top \boldsymbol{X}) = N_c$. This can only be fulfilled if the number of samples is equal to or larger than the number of features.

However, in computational material science, generating many samples is computationally expensive, and often the feature set size exceeds the sample set size. This results in an underdetermined problem, where the matrix cannot be inverted and infinite solutions exist. To solve this problem, the minimization problem posed by Eq. 22 must be regularized. One form of regularization consists of penalizing large values of the coefficients $\boldsymbol{\mathcal{J}}$.

A widely used regularization is called Ridge regression and penalizes possible solutions with the ℓ^2-norm of the coefficients [37],

$$\begin{aligned} C_{Ridge}(\boldsymbol{\mathcal{J}}) &= \|\boldsymbol{P} - \boldsymbol{X}\boldsymbol{\mathcal{J}}\|_2^2 + \lambda \|\boldsymbol{\mathcal{J}}\|_2^2 \\ \Rightarrow \boldsymbol{\mathcal{J}} &= (\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I})^{-1} \boldsymbol{X}^\top \boldsymbol{P} , \end{aligned} \tag{24}$$

where λ is the regularisation parameter and $\boldsymbol{I}$ the identity. Ridge regression shrinks the regression coefficients toward zero, where λ controls the amount of shrinkage applied to the regression coefficients.

Penalization of linear regression is also possible with the ℓ^1-norm, called the least absolute shrinkage and selection operator (LASSO) [38], and the ℓ^0-norm. In LASSO, the absolute values of the coefficients are penalized, $\|\mathcal{J}\|_1 = \sum_j |\mathcal{J}_j|$. When using the ℓ^0-norm, the penalized quantity is the number of nonzero coefficients, $\|\mathcal{J}\|_0 = \#\{\mathcal{J}_i \mid \mathcal{J}_i \neq 0\}$. Their cost functions are, respectively,

$$\begin{aligned} C_{LASSO}(\mathcal{J}) &= \sum_{i=1}^{N_s} (p_i - \hat{p}_i)^2 + \lambda \sum_{j=1}^{N_f} |\mathcal{J}_j| \\ C_{\ell_0}(\mathcal{J}) &= \sum_{i=1}^{N_s} (p_i - \hat{p}_i)^2 + \lambda \sum_{j=1}^{N_f} \mathbf{1}_{x\neq 0}(\mathcal{J}_j) . \end{aligned} \tag{25}$$

Here $\mathbf{1}_{x\neq 0}$ is an indicator function such that $\mathbf{1}_{x\neq 0}(\mathcal{J}) = 0(1)$ for $\mathcal{J} = 0(\mathcal{J} \neq 0)$.

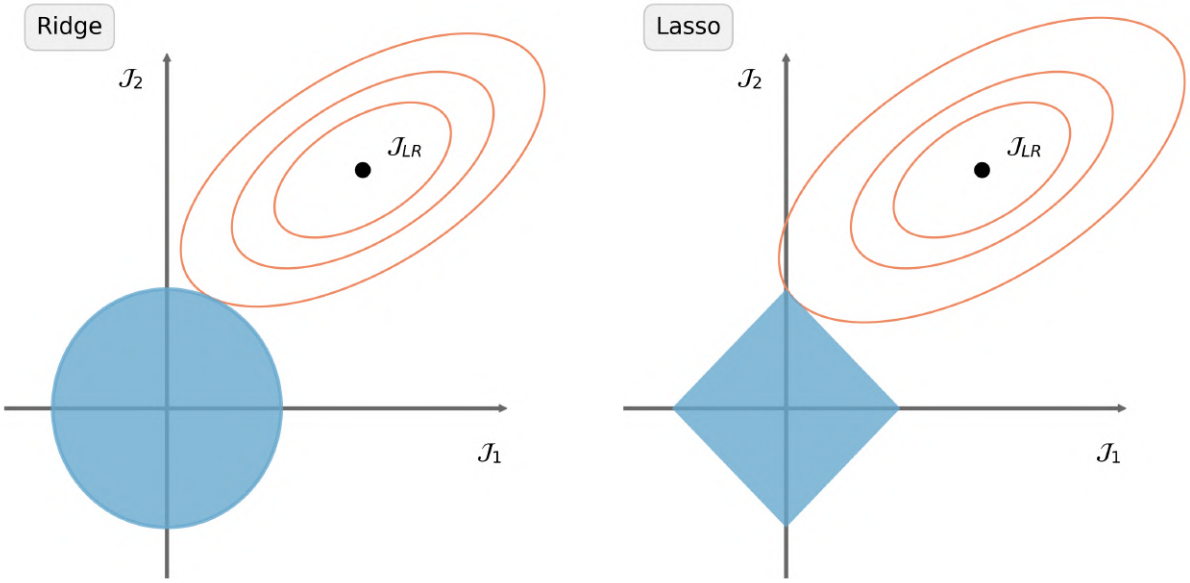


Figure 3.1. Estimation process of RSS with penalization. The blue circle and square represent the constraint regions $\beta_1^2 + \beta_2^2 \leq t^2$ and $|\beta_1| + |\beta_2| \leq t$ of ridge regression (left) and LASSO (right), respectively. The orange ellipses indicate the contours of constant RSS solutions.

Figure 3.1 illustrates how the penalization works when finding an optimal solution for $\mathcal{J}$ with least squares fitting in the two-dimensional feature space of the coefficients $\mathcal{J}_1$ and $\mathcal{J}_2$. The colored patches describe the constraint regions introduced by the ℓ^p-norms. Possible solutions have to lie within the patches, and thus, the optimum solution will be found at the intersection of the patch edge and contour of constant RSS. In contrast to the circle, the square has corners positioned directly on the axes; for a small λ, a solution will most likely be found on the square corner, with the result of setting a coefficient equal to zero. This effectively discards certain features.

In summary, while Ridge regression only increases or decreases the influence of the features of the model, LASSO and the ℓ^0-norm set some feature coefficients to zero, thereby actively reducing the feature space. Hence, the latter two belong to the compressed sensing methods [39], [40]; they reduce the dimensionality of the solution and therefore create sparse models. In terms of CE, LASSO and ℓ_0 regularization perform a *cluster selection*.

Another type of regression model that performs cluster selection is known as orthogonal matching pursuit (OMP) [41]. It is a type of greedy algorithm that is used for sparse signal recovery. The OMP algorithm works by iteratively selecting the most relevant feature column vectors in Eq. 19 to represent the property until a maximum level of sparsity has been reached or the residual falls below a certain threshold. Such a set of most relevant features is called the support $\mathcal{S}$. The relevance of a feature is chosen such that it minimizes the correlation between its vector and the residual of the targets calculated with the model predictions of the previous iteration, i.e., the inner product between the feature and residual vectors is maximized. Next, the feature is added to $\mathcal{S}$, and a new residual of the targets and the predictions made with the updated support model is evaluated. The updating of the support model coefficients is performed with least squares fitting and ensures that the new residual is orthogonal to all feature vectors in $\mathcal{S}$, which consequently cannot be chosen in the next iteration.

In order to obtain an estimate of the model performance on unknown data, cross-validation (CV) is used. The idea is to split the available data into a training and a validation set. Subsequently, the model is fitted to the training data and then tested on the validation set to get an estimate of its performance on the unseen validation data. A commonly used form of CV is known as k-fold CV and is characterized by dividing the data into k subsets (or folds), with each subset consisting of roughly the same number of samples. CV is also used to choose the hyperparameters of the model, e.g., λ in Eq. 24 and Eq. 25. An optimal model finds the sweet spot between underfitting and overfitting (i.e., high and low λ), which is the minimum of the CV error as a function of the hyperparameter.

Now in order to find a CE model that can make good predictions on materials properties, Eq. 19 can be solved by employing ridge regression, LASSO, or regression with the ℓ_0-norm. An optimal value for the strength of regularization can be found with CV, which, depending on the regularization method used, also selects a set of optimal clusters out of the initial clusters set $\mathcal{C}$. The prediction of the property of a new structure is then made with the optimal model.

3.3 Nonlinear Features in Cluster Expansion

The method developed for this work builds on the idea to augment the initial feature space of the linear problem in Eq. 19, and then use regularization to find an optimal set of features. Extending the feature space is a common procedure in machine learning and aims at retaining the simplicity of linear models while introducing nonlinearities to the feature space [42].

To introduce the nonlinear behavior of data into linear models, the feature space is extended with new features that are nonlinear transformations of the existing ones. The intention is to turn the feature space into a new space where a linear model can easily be fitted. The expansion of a function $f(X)$, where X is, for now, the vector of input features, is then denoted by

$$f(X) = \sum_i \beta_i h_i(X) \, , \tag{26}$$

where $h_i(X)$ are the augmented features. In principle, $h_i(X)$ can take the form of any nonlinear transformation, but for the purpose of demonstration, we restrict ourselves to polynomial features, such as $h_m(X) = X_k^2$ or $h_m(X) = X_k X_l$, which is called *polynomial feature expansion*.

X in Eq. 26 is one row of cluster correlations from the correlation matrix $\boldsymbol{X}$. Then, nonlinearities can be introduced into the model by expanding the columns with new columns that represent the nonlinear cluster correlations created by multiplying the columns of initial correlations with each other. For example, a polynomial expansion up to the polynomial degree 2 of a $N_s \times 2$ correlation matrix with two clusters and N_s structures yields the $N_s \times 5$ correlation matrix

$$\begin{pmatrix} X_{1,1} & X_{1,2} \\ \vdots & \vdots \\ X_{N_s,1} & X_{N_s,2} \end{pmatrix} \Rightarrow \begin{pmatrix} X_{1,1} & X_{1,2} & X_{1,1}^2 & X_{1,1}X_{1,2} & X_{1,2}^2 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ X_{N_s,1} & X_{N_s,2} & X_{N_s,1}^2 & X_{N_s,1}X_{N_s,2} & X_{N_s}^2 \end{pmatrix} \, . \tag{27}$$

As the new correlations are created out of the initial cluster correlations, one can speak of polynomial features that are created out of the initial clusters pool. The number of features N_f that are created out of an initial clusters pool of size N_c by a polynomial expansion of degree d is given by

$$N_f(N_c, d) = \frac{(N_c + d)!}{N_c! d!} \, . \tag{28}$$

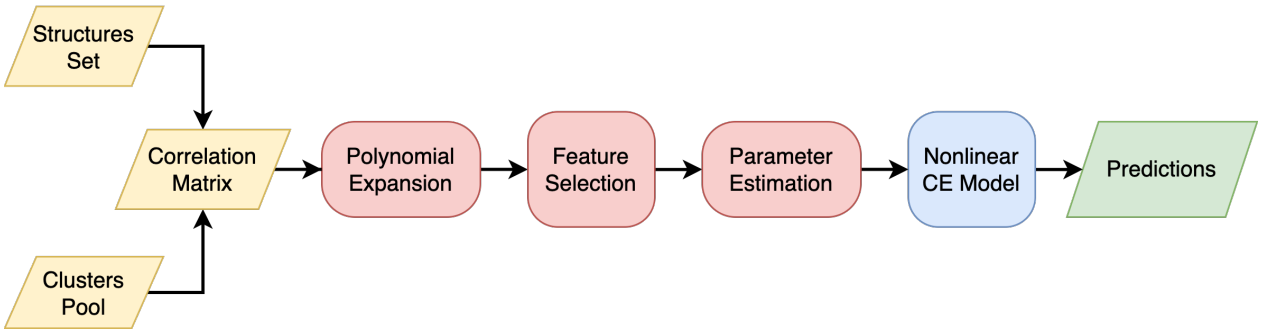


Figure 3.2. Workflow for predicting with the nonlinear CE.

We call this novel method the "nonlinear CE".

A workflow of how the nonlinear CE can be used for making predictions is shown in Fig. 3.2. It describes the procedure used in the following sections. A more detailed introduction to the nonlinear CE in the Python package CELL will be given in Sec. 5. The creation of a CE model starts with obtaining a set of sample structures with known property values, typically calculated with DFT. Next, one defines a pool of clusters. Both the structures set and the clusters pool are derivative structures from a common parent lattice, which defines the primitive lattice, the crystal symmetry, and the possible substituents. Both sets are used to calculate the correlation matrix. As described in Eq. 27, the matrix is then polynomially expanded to form the input for the feature selection methods. Feature selection is performed using compressed sensing methods, such as LASSO. With the optimal feature set, the parameters of the nonlinear CE model can be estimated. Finally, to perform property predictions of a previously unseen structure, the nonlinear features for that structure must be computed and used as input for the nonlinear CE model. The nonlinear features are computed by first determining the cluster correlations for that structure and then performing the polynomial expansion, retaining only the terms given by the feature selection before.

If we return to the problem of the binary alloy discussed in Sec. 2.3, the advantages of this method become evident. When expanding the squared concentration with a nonlinear CE of degree two, $x^2 = X_{1p}(\boldsymbol{\sigma})X_{1p}(\boldsymbol{\sigma})$ is exactly given by the squared 1-point cluster correlations shown in the third column of the expanded correlation matrix in Eq. 27. Consequently, the exact representation of the squared concentration in the binary alloy with the nonlinear CE is given only by this single feature. This is an important benefit compared to the non-converging infinite expansion of 2-point clusters with the standard CE. This idea also holds for more sophisticated nonlinearities. Thus, the novel approach of introducing nonlinear features drastically reduces the number of expansion terms needed for an exact representation of properties with nonlinearities. As will be shown in the following sections, our method, in combination with feature selection methods, such as LASSO, is able to find compact and accurate CE models for the efficient prediction of materials properties.

Classification

The idea of viewing the CE as a common ML task opens up its application to another major ML field: classification. In materials science, it is of interest to categorize new structures with little computational effort. For instance, in solids, classifying structures as metals or semiconductors with a minimum number of materials descriptors would be beneficial.

A common method to classify data is to find linear boundaries that separate the data into various classes in feature space. A well-established supervised ML method for linear classification is called the support vector machine (SVM) [43]. SVMs can also be employed to find nonlinear decision boundaries by the use of the kernel trick [42], [44]. The idea behind the SVM method is to find a separating hyperplane where the margin, i.e., the distance from the plane to the closest data points, is maximal. In the linearly separable case, all training points lie on their respective side of the boundary and outside their margin. New observations are then classified according to the sign of their distance from the decision boundary. The intuition is that the margin maximization will also lead to a good separation of new data that has not been used for training.

However, in cases when the data is not easily separable, or outliers would enforce a bad decision boundary, the method can be improved by allowing some training data points to lie on the wrong side of their margin, i.e., they can be misclassified in training. The misclassified points are defined by their offset into the wrong side of the margin. Thus, by limiting the maximum size of this offset (through the number of points or their distance from the margin), one effectively introduces a regularisation to the maximization of the margin. The offset of misclassifications can be formalized by the *hinge loss*,

$$t = \max(0, 1 - yf(x)) \ .$$

Here,

$$f(x) = x^T\beta + \beta_0 \tag{29}$$

is the function used to evaluate the distance of a point x from the plane, and y yields the class label $1(-1)$. For $f(x) = 0$, Eq. 29 defines the hyperplane. The margin is given by $1/|\beta|$, and thus maximizing M is equivalent to minimizing β.

With these elements, we obtain an optimization problem under two constraints, formalized by

$$\begin{aligned} &\min_{\beta,\beta_0} \frac{1}{2}\beta^\top\beta + C\sum_{i=1}^{n} t_i \ , \quad \text{subject to} \\ &t_i \geq 1 - y(x^\top\beta + \beta_0)) \ , \\ &t_i \geq 0 \ . \end{aligned} \tag{30}$$

Here, C is the regularization coefficient, and n is the number of features. This is a convex optimization problem and can be solved with Lagrange multipliers. After solving, for predicting the class of an observation, one obtains

$$\text{class of } x = sign\left(\sum_i \alpha_i y_i x_i^\top x + \beta_0\right) \ . \tag{31}$$

Here, α_i are the Lagrange multipliers; they are zero for all points on the correct side of their margin and larger than zero for all other points. The latter are called the *support vectors* because these are used to construct the margin.

The support vector machine expressed by Eq. 30 and Eq. 30 is restricted to identifying linear decision boundaries in the input feature space. As shown for linear regression, one can augment this method by introducing nonlinear features, for example, with polynomial expansion. In higher-dimensional space, it is often possible to find a hyperplane that achieves better training-class separation or separates previously inseparable data. This linear boundary in higher-dimensional space translates to a nonlinear boundary in the original space.

3.4 Nonlinear CE Demonstrated on Toy Model

To illustrate the shortcomings of the CE method, Ref. [18]–[20] introduced a toy data model with squared concentration dependence. In this section, we show that the nonlinear CE can capture such a dependence exactly. For this, we consider both a simple toy model similar to that of Refs. [18]–[20], and a generalization obtained by introducing configuration dependence through concentration-dependent two-body interactions. In the generalized toy data model, the physical situations found in, e.g., binary alloys, are better represented.

The first property is simply the squared concentration as used in the discussion, adjusted to yield zero for the pristine crystal (x=0) as well as the fully substituted one (x=1). It is given by

$$P(\sigma) = ax(1-x) \ , \tag{32}$$

where x is the concentration of configuration σ, and a is a real number, the so-calleed bowing factor.

Next, we introduce a configuration dependence into the property by adding a 2-point interaction with concentration dependence. The second property is given by

$$P(\sigma) = a_1 x(x-1) + J(x)X(\sigma)_{2p} + \mu x \ , \tag{33}$$

where $X(\sigma)_{2p}$ describes the 2-point first neighbor for configuration σ. The term μx is introduced to make $P(1) = 0$, with a properly selected value of the constant μ. $J(x)$ is a concentration-dependent 2-point interaction defined by

$$J(x) = b_1(1-x) + b_2 x + b_3 x(1-x) \ , \tag{34}$$

where the coefficients b_1, b_2, and b_3 can be used for adjusting the complexity of the added concentration dependence. Possible choices are

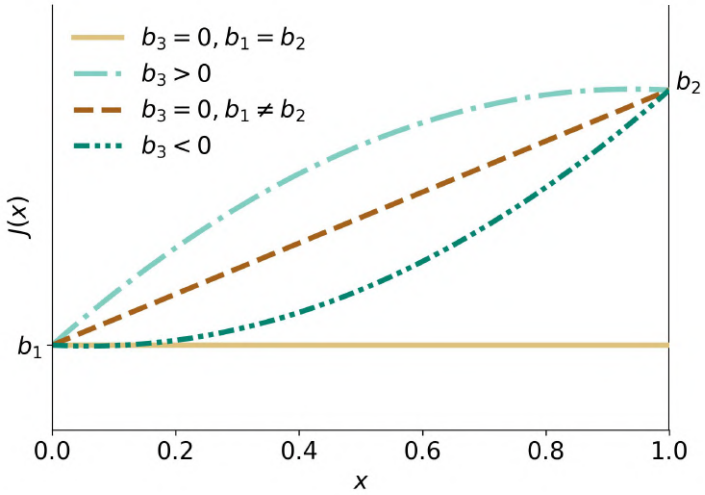


Figure 3.3. Complexity of the 2-point interaction concentration dependence. It is given by the $J(x) = b_1(1-x) + b_2 x + b_3 x(1-x)$. For $b_1 = b_2$ and $b_3 = 0$, the introduced 2-point interaction is constant for all concentrations (beige). Unequal b_2 and b_3 make it linearly increasing for higher concentrations (brown). A positive or negative nonlinear bowing is introduced by the choice of $b_3 > 0$ (light blue) or $b_3 < 0$ (teal).

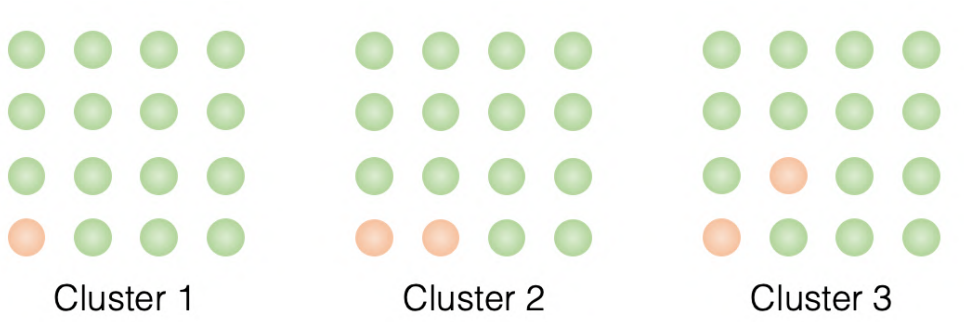


Figure 3.4. 1-point, 2-point first neighbor, and 2-point second neighbor cluster in the binary alloy. These three clusters comprise the smallest used clusters pool P1, and polynomial features for the nonlinear CE are created from their corresponding cluster correlations.

1. constant for $b_1 = b_2, b_3 = 0$,
2. linear for $b_1 \neq b_2, b_3 = 0$, and
3. nonlinear for $b_1 \neq b_2 \neq b_3$.

Fig. 3.3 illustrates four possibilities for $J(x)$.

We consider a binary alloy consisting of a pristine copper crystal and gold substituents that we describe using the IFB basis defined in Eq. 9. We recall that in this basis, $X(\sigma)_{1p} = x$, the concentration of Au, and $X(\sigma)_{2p}$ is just the concentration of the first neighbor Au-Au pairs. The supercell is an 8×8 two-dimensional lattice with 64 atoms. A training set consisting of 132 structures was created. Besides the pristine Cu and Au structures and those with a single Au or a single Cu atom, various concentrations corresponding to a number of Au substituents between 2 and 62 were considered. For each Au concentration, 8 random structures were created. LASSO with LOO CV is used for feature selection. Subsequently, the model parameters are fitted with linear regression.

The clusters can be visualized as the atoms occupying the lattice sites. The first three smallest clusters are illustrated in Fig. 3.4. These are the 1-point, the 2-point first neighbor, and the 2-point next neighbor cluster. The smallest clusters pool, which will be referred to as P1, is made up of these three clusters, which correspond to the cluster correlations from which nonlinear features are created for the nonlinear CE. For a polynomial of degree three, this results in features such as x^2, xX_{2p}, or x^2X_{2p}. Larger clusters pools contain larger 2-point clusters, i.e., 2-point interactions between further away atoms and 3-point clusters, starting with the first neighbors and then increasing in size.

Fig. 3.5 and Fig. 3.6 compare the standard CE and the nonlinear CE when modeling the properties defined by Eq. 32 and Eq. 33, respectively. In Fig. 3.5, the property values and the model predictions for the two considered properties are plotted as a function of the concentration of substituents x. Figure 3.6 shows the model coefficients for the two properties using standard CE and nonlinear CE.

The upper left panel shows the predictions of the property of Eq. 32 made with the standard CE. As the property does not have configuration dependence, all 8 configurations per composition yield the same value, as is evident from the fact that there is a single target value (black circle) per concentration. It can be seen that the predictions are inaccurate, especially in the region around $x = 0.5$. The RMSE is 0.0088. Different configurations per composition show differing predictions, suggesting the introduction of a spurious configuration dependence of the underlying property by

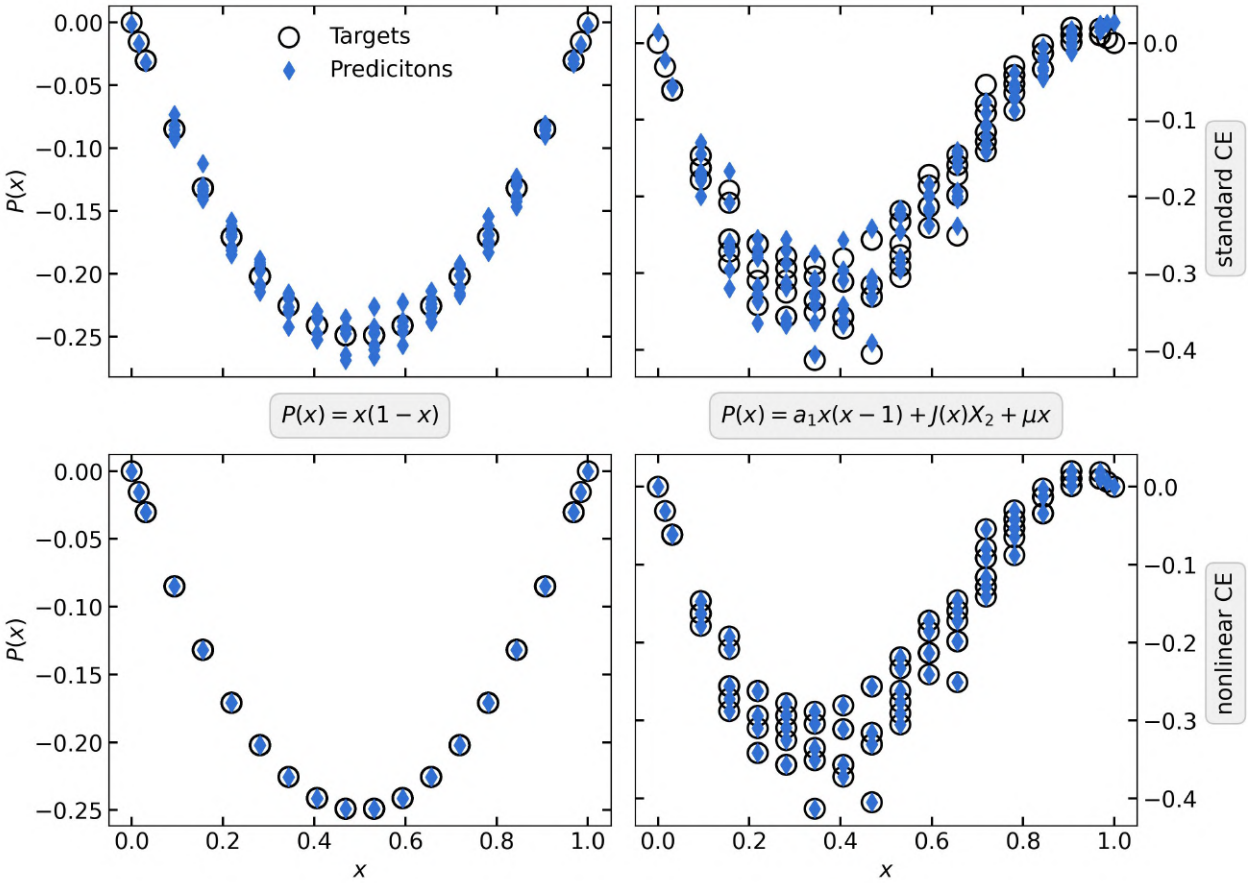


Figure 3.5. Predictions (blue diamonds) of standard (upper panels) and nonlinear (lower panels) CE models for configuration-independent (left panels) and configuration-dependent (right panels) property targets versus the substituent concentration x.

the standard CE, which does not converge. Additionally, it can be seen in the upper-left panel of Fig. 3.6 that all clusters contribute to the final model. The linear concentration term in Eq. 32 is exactly captured by the negative 1-point cluster ECI, while the squared concentration cannot be exactly expanded by the remaining 2-point cluster correlations. From Sec. 2.2, we know that an accurate representation would require an expansion over all possible clusters. In fact, fitting Eq. 32 with the standard CE and a clusters pool consisting of the 1-point cluster and all 14 symmetrically inequivalent 2-point clusters yields a perfect fit. Nevertheless, as argued in Sec. 3.3, the squared concentration dependence is exactly captured by the nonlinear CE, as can be seen in the prediction matches in the bottom left panel of Fig. 3.5. The fact that the model exactly reproduces the concentration dependence is expressed by the strength of the 1-point and 1-point squared cluster correlation features with -1 and 1, respectively, as shown in the lower left panel of Fig. 3.6.

Predictions of the configuration-dependent property with nonlinear concentration dependence are presented by the right panels of Fig. 3.5. In contrast to the previous case, now the target values show a dependence on configuration, as can be seen by the fact that different property target values are present at fixed concentrations. The target property is given by Eq. 33 and Eq. 34 with the coefficients $a_1 = -1,\ b_1 = 2,\ b_2 = 1,\ b_3 = 1.5,$ and $\mu = -b2$.

The upper panel displays the predictions obtained by a standard CE model using clusters pool P3. Despite the model having a relatively large number of clusters, the predictions were found to be

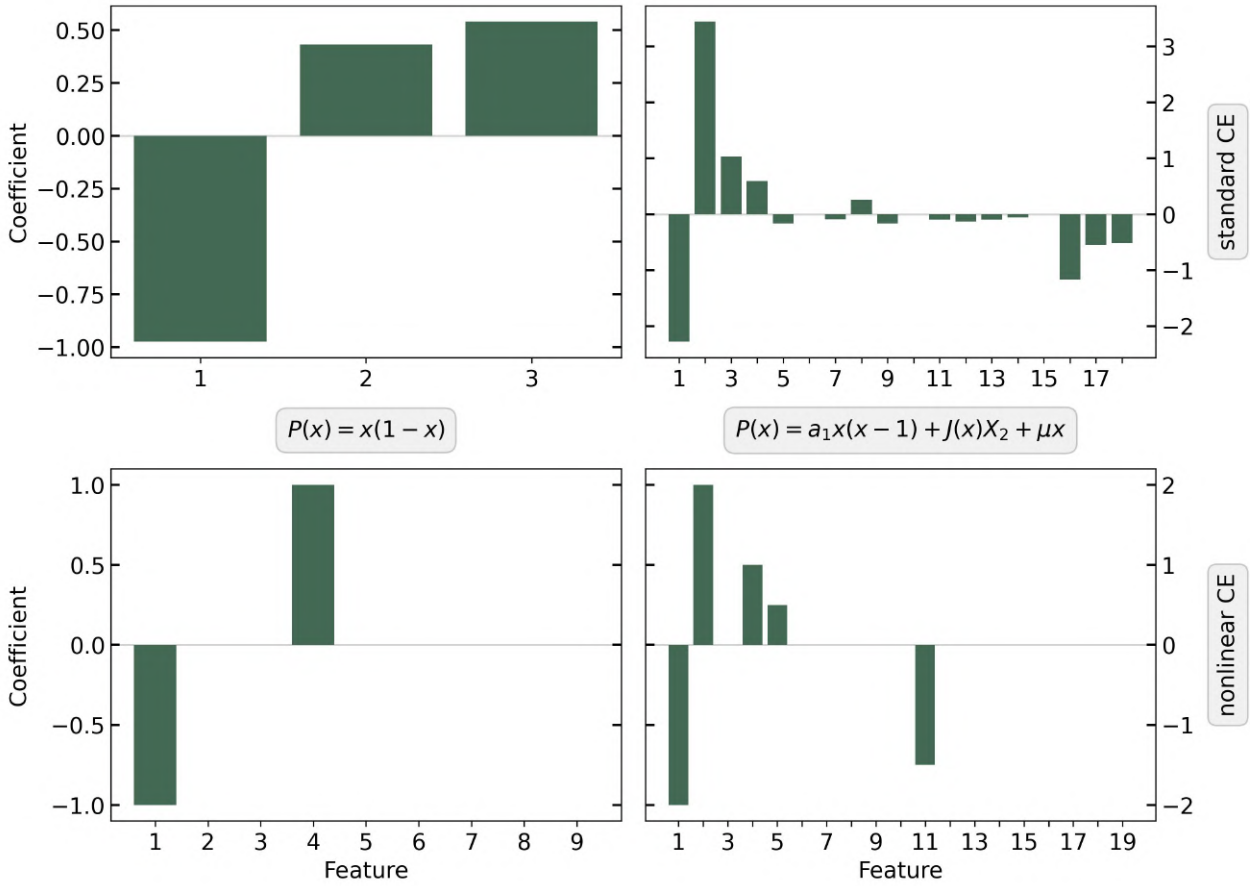


Figure 3.6. Effective cluster interaction of two standard (upper panel) and two nonlinear CE models (lower panel). The ECIs correspond to the models that were used to fit the configuration-independent and the configuration-dependent properties, with predictions shown in Fig. 3.5.

inaccurate and unable to capture the precise configuration dependence introduced, with a RMSE of 0.014. Taking a look at the coefficients in the upper right panel of Fig. 3.6, almost all clusters contribute to some degree. ECI 1 again corresponds to the 1-point cluster and captures the linear concentration dependence as in the case of the simpler property (left panels).

Again, the nonlinear CE has no trouble fitting this more sophisticated nonlinear configuration dependence in the property. All predictions are within machine precision, and fewer coefficients contribute to the model. The coefficients in the lower right panel of Fig. 3.6 are directly related to the above-chosen coefficients of the toy data model. The relation is found by recasting the toy data model in terms of the linear and nonlinear features x, X_{2p}, x^2, xX_{2p}, x^2X_{2p},

$$\begin{aligned} P(x) &= a_1x(1-x) + [b_1(1-x) + b_2x + b3x(1-x)]X_{2p} + \mu x \\ &= (a_1+\mu)x + b_1X_{2p} - a_1x^2 + (-b_1+b_2+b_3)xX_{2p} - b_3x^2X_{2p} \\ &= C_1x + C_2X_{2p} - C_4x^2 + C_5xX_{2p} - C_{11}x^2X_{2p}\ . \end{aligned}$$

Here, the coefficients C_i are numbered according to the position of the corresponding feature in the expansion. The relationship between the coefficients is given in Tab. 3.2. Inserting the returned coefficients $C_1 = -2$, $C_2 = 2$, $C_4 = 1$, $C_5 = 0.5$, and $C_{11} = 1.5$, the correct toy data model can be derived exactly,

$$P(x) = -1x(1-x) + [2(1-x) + 1x + 1.5x(1-x)]X_{2p} - 1x\ .$$

Table 3.2. Equalities between expansion coefficients of nonlinear CE model of $P(\sigma)$ of Eq. 33 (bottom row) and the coefficients of the underlying model (top row).

a_1	b_1	b_2	b_3	μ
c_4	C_2	$C_2 + C_5 + C_{11}$	$-C_{11}$	$C_4 - C_1$

These examples show that the nonlinear CE can accurately fit materials properties with varying complexities of nonlinear concentration and configuration dependence. Even if an exact fit can be obtained with the standard CE by using more clusters in the clusters pool, this requires a much larger number of features when using the standard CE instead of the nonlinear CE. This means the nonlinear CE can create simpler models with at least equal accuracy than the standard CE. Considering that the evaluation of the cluster correlation is the main computational effort in CE, the benefits of these simple models are evident.

In conclusion, the introduction of nonlinear features created out of the cluster correlations is an effective method to solve the problem of nonlinear dependence in alloy properties and solves the problem discussed in Ref. [18] and Ref. [19].

4

Clathrate Study

As a test of the nonlinear CE on real-world problems, we model the energy of mixing and the bandgaps of the thermoelectric type-I clathrate $Ba_8Al_xSi_{46-x}$. In Sec. 4.1, we introduce the crystal structure and electronic properties of type-I clathrates. Section 4.2 describes the dataset and justifies the selection of $Ba_8Al_xSi_{46-x}$ as an exemplary case for applying the nonlinear CE. Our results of modeling the energy of mixing and of fitting the bandgaps are presented in Sec. 4.3 and Sec. 4.4. Lastly, we demonstrate the use of the nonlinear CE for the classification of the $Ba_8Al_xSi_{46-x}$ bandgaps in Sec. 4.5.

4.1 Type-I Clathrates

As stated in Sec. 1, the phonon-glass electron-crystal concept can be realized in intermetallic type-I clathrates. In 1865, two cubic phases of NaSi were the first discovered intermetallic clathrates [45], and in 1998 Ge-based intermetallic clathrates were described using the phono-glass electron-crystal concept and predicted to be promising thermoelectrics with a figure of merit greater than one [3].

The structure of type-I clathrates is characterized by a cage-like framework of host atoms with loosely bound guest atoms inside the cages. It is shown in Fig. 4.1 taken from Ref. [8]. Its unit cell (Fig. 4.1 (a)) belongs to the cubic space group $Pm\bar{3}n$ and consists of 46 covalently bound atoms located at the three symmetrically distinct Wyckoff sites 24*k*, 16*i*, and 6*c*. They form eight cavities, six tetrakaidecahedra and two dodecahedra (Fig. 4.1 (a)), centered at the Wyckoff sites 6*d* and 2*c*, at which guest atoms can be located, resulting in a total of 54 atoms in the unit cell. Figure 4.1 (b) depicts the framework of bonds between each Wyckoff site and its four nearest neighbors. Four distinct types of bonds result: *k*-*k*, *k*-*i*, *k*-*c*, and *i*-*i*. The guest atoms, typically alkali or alkaline-earth metals, vibrate at low frequencies and scatter acoustic phonons, which are responsible for thermal transport. Consequently, the guest atoms reduce the lattice part of the thermal conductivity and, hence, are also called "rattlers" [46].

The pristine host crystal X_{46}, populated with group IV elements (e.g., Ge or Si), creates a covalently bound semiconductor [47]. However, the introduction of eight guest atoms, A_8X_{46}, each donating two valence electrons, turns the clathrate into an n-doped metal. This doping can be compensated by substituting group IV host atoms with group III atoms that have one valence electron less (p-doping).

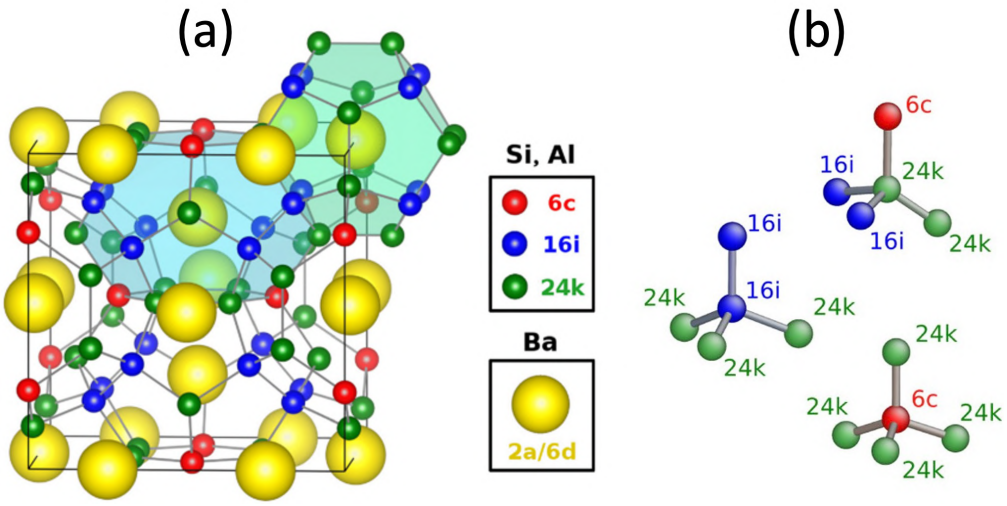


Figure 4.1. Unit cell with guest containing cages (a), and host site nearest neighbors (b) of $A_8M_xX_{46-x}$ type-I clathrates. The host framework consists of group IV atoms at the Wykoff sites 6*c*, 16*i*, and 24*k* in green, blue, and red, respectively, building tetrakaidecahedra (light blue) and dodecahedra (light green) cages with guest atoms at sites 2*a* and 6*d* at their center (yellow). Figures are taken from Ref. [9] (left panel) and Ref. [8] (right panel).

The electronic properties of $A_8M_xX_{46-x}$ can be well approximated by the Zintl-Klemm model [48]. Assuming pure ionic host-guest interactions, the charge-neutral composition at $x = 16$ is expected to be semiconducting [49]–[51], which is favorable for thermoelectric behavior.

Promising candidates for thermoelectric clathrates are the GaGe-based compounds $Sr_8Ga_xGe_{46-x}$ and $Ba_8Ga_xGe_{46-x}$ [52]–[54]. For thermoelectric materials, zT values greater than one are an important milestone, which has been surpassed for GaGe-based clathrates, with a zT up to 1.63 [55]. However, Ga and Ge are considered critical raw materials with limited production around the world. As such, it is of considerable interest to find clathrates with substitute elements for Ga and Ge, which retain their low thermal and high electric conductivities and the tunability of these properties in the material. Replacing Ga and Ge with Si and Al, respectively, are obvious choices and have already shown promising results [4]. Moreover, experimentalists have been able to synthesize compositions up to $x < 15$ [5], which is close to the desired charge neutral composition $x = 16$. It has been demonstrated that the charge-neutral ground state configuration indeed exhibits semiconducting behavior, while higher energy configurations undergo a semiconductor-to-metal transition [8].

The theoretical description of type-I clathrates is complicated, as the number of configurations increases exponentially when the number of substituents rises, which leads to a combinatorial explosion. For example, for 16 substituents, the number of possible structures in a unit cell exceeds 10^{10}. Consequently, the configurational space becomes too vast for *ab initio* methods or DFT to computationally handle.

4.2 $Ba_8Al_xSi_{46-x}$ Dataset

The ground state energies and electronic structure of $Ba_8Al_xSi_{46-x}$ were studied in Ref. [8]–[10]. These studies included the calculation of total energies for multiple compositions and configurations with the code `exciting`, which is a full-potential all-electron DFT package [56]. To represent the wavefunctions efficiently and accurately in both the atom and the inter-atom region, `exciting` employs the linearized augmented plane-wave method. Augmented plane waves are wavefunctions with a dual basis of, first, spherical harmonics and radial functions in the muffin-tin region close to the cores and, second, plane waves in the interstitial region between cores. This basis set successfully treats all electrons. We reuse the datasets used for these studies for various analyses of the energy of mixing and the bandgap with the nonlinear CE.

The energy of mixing for a given concentration x describes if an alloy mixes or forms separate phases. It represents the offset from the linear interpolation between the energies for both the pristine and the fully-substituted cases. The energy of mixing per atom is given by

$$E_{mix}(\boldsymbol{\sigma}) = \frac{1}{54}\left\{E_{total}(\boldsymbol{\sigma}) - \left[E^0_{total} + (E^{46}_{total} - E^0_{total})\frac{N_{Al}}{46}\right]\right\} . \tag{35}$$

Here, $E_{total}(\boldsymbol{\sigma})$ is the total energy of a configuration calculated with DFT and E^0_{total} and E^{46}_{total} denote reference energies of the pristine and the fully substituted crystal, respectively ($E^0_{total} = -78361.24\ Ha$ and $E^{46}_{total} = -76192.94\ Ha$). For this example, for convenience, the reference energies were not computed by DFT but predicted using a CE model. The conclusions of the analysis below do not depend on their precise value.

The thermoelectric clathrate $Ba_8Al_xSi_{46-x}$ introduced in Sec. 4.1 poses an interesting problem for multiple reasons. First, considering that doping occurs only in the host framework, it can be treated as a binary alloy with the two species Si and Al. Second, its mixing energy and band gap energy depend not only on the composition but also on the configuration of Al substituents. Third, as shown in Ref. [8], the energy of mixing has a nonlinear concentration dependence that can, to some extent, still be fitted with the standard CE. In that work, the nonlinear behavior given by a kink in the energy of mixing against Al concentration was attributed to an abrupt change of the occupied electronic states around $x = 13$. This change is reflected by the electronic density of states at the Fermi level, which presents a local minimum around $x = 13$ (see Fig. 4.2, left panel, reproduced from Ref. [8]). The energy of mixing is presented in Fig. 4.2 (right panel), where the two linear fits of the low-lying ground state configurations intersect at $x \sim 13$, putting in evidence the presence of the kink. Lastly, it was found that fitting the data with a single standard CE for the full concentration range considered ($x = 6 - 16$) leads to poor predictive performance. This suggests that improvements in predictive performance by using the nonlinear CE method developed in this thesis could be obtained.

In Ref. [8], to improve the accuracy of energy predictions, the data set was split into two regions: one for concentrations between 6 and 13 and the other for concentrations between 12 and 16. Then, for every split, an independent standard CE model was obtained. This approach led to a significant improvement in global accuracy. Such a procedure presents a great disadvantage, however, since prior knowledge of the behavior of the property is needed for selecting an adequate split. Another obvious disadvantage is that two different models are needed for predictions of the full compositional range. By using the nonlinear CE, we demonstrate that one can make energy predictions that are at least comparable in accuracy to split CE predictions while having a single model to describe the full concentration range.

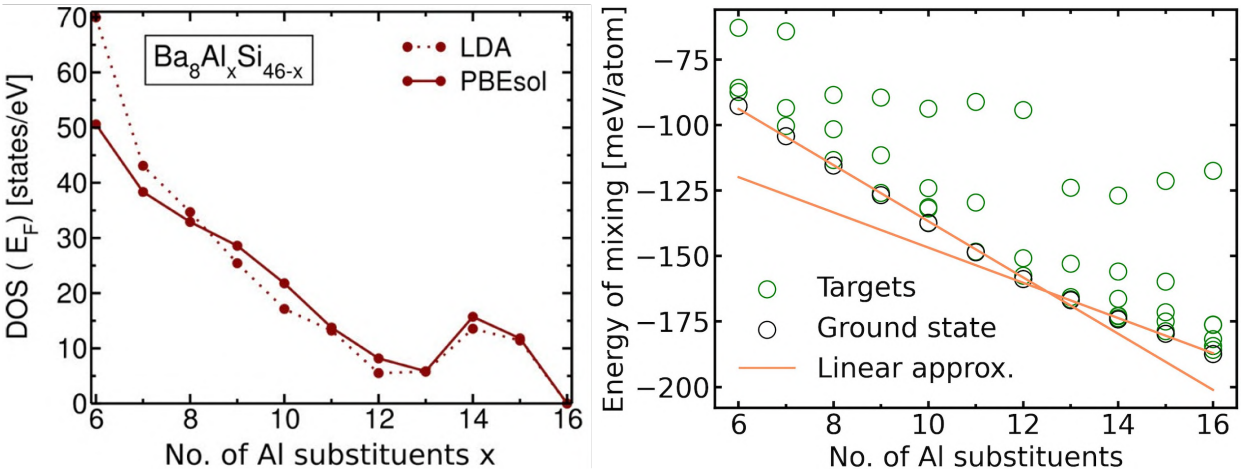


Figure 4.2. Left: Density of states of $Ba_8Al_xSi_{46-x}$ calculated with DFT exchange-correlation functionals LDA and PBEsol. The increase at around 13 Al substituents is attributed to the nonlinear kink in the energy of mixing. The figure was taken from Ref. [8]. Right: DFT calculated energy of mixing values (green) with ground state configurations (black). The two linear fits (orange) in the ranges 6 to 13 and 13 to 16 indicate the nonlinear kink at 13 substituents

Attempts to model the bandgap of clathrates with standard CE in Ref. [8] led to models with poor generalizability. That is to be expected since fitting the bandgap is a notoriously difficult task because of the discontinuity introduced by fitting metals and non-metals. With the help of the non-linear CE, we are able to improve on bandgap fitting with a single model for the full configurational range.

For the energy of mixing study, we employ a dataset of 56 structures from Ref. [8], calculated with the LDA functional. In Ref. [8], it was found that very similar results are obtained by using the PBEsol functional. Here, we use the LDA data set, as it contains more data points than PBEsol. This is convenient for the learning task in this study. The set consists of different compositions with a range of 6 to 16 Al substituents, with 16 representing the charge-neutral composition. Per composition, multiple configurations exist. The values of the energy of mixing of this dataset are shown in the right panel of Fig. 4.2.

The data set for constructing a CE model of the band gap is also in the range of 6 to 16 Al substituents. It combines the PBEsol data from Ref. [8] and the data from Ref. [9]. Both were used in Ref. [10], where it was found that the standard CE yields poor models for the band gap. At $x = 16$, there is a larger concentration of data points as for the other compositions, which is due to the fact that the charge-balanced composition $x = 16$ is more relevant for thermoelectric applications. The band gap energies were derived from the Kohn-Sham eigenvalues. Due to the Brillouin Zone discretization, it was necessary to set up a threshold such that all bandgaps below 5.44 meV are considered metals. After removing duplicate structures, the data set contained 78 structures in total. The values of the DFT bandgaps for this dataset are shown in Fig. 4.11.

4.3 Energy of Mixing

In this section, the energy of mixing of $Ba_8Al_xSi_{46-x}$ is modeled with a CE model trained by employing LASSO with CV for feature selection and ridge regression with CV for estimating the model coefficients. As in Sec. 3.4, the correlations are evaluated using the IFB (Eq. 9). Before obtaining an optimal CE model, the three main hyperparameters that influence model convergence are systematically studied. These are the range of λ values for LASSO, the degree of polynomial feature expansion, and the choice of the initial clusters pool.

Four clusters pools are created, for an increasing number of points and radii. The first pool consists of the three 1-point clusters at the three Wykoff sites 24k, 16i, and 6c, visualized in Fig. 4.1, and is called P0. By Eq. 11, these correspond to the fractional occupation factor of Al in the various Wyckoff sites of the host framework. To the next pool, P1, the four nearest neighbor 2-point clusters are added for a total of 7 clusters. These correspond to the first neighbor bonds depicted in Fig. 4.1 b). In the next pool, P2, larger 2-point clusters are added, which are clusters of two Al atoms with a Si atom in between. $Ba_8Al_xSi_{46-x}$ contains 8 such pairs, increasing the size of P2 to 16 clusters. The last clusters pool, P3, includes 3-point interactions and includes 36 clusters. Additional fixed hyperparameters are the range of regularisation parameters for ridge regression of 1 to 1×10^{-5}. Ridge regression was employed to allow for additional regularisation when estimating the accuracy of a model. For validating the model on unseen data, we used leave-one-out (LOO) CV.

The search of the optimal λ value for LASSO is performed by minimizing the MSE in a grid-search LOO-CV procedure. For this task, we employ the `LassoCV()` method of the Python library scikit-learn. Initially, a wide range of values (1×10^{-4} – 1×10^{-10}) is used for λ to ensure entering the regimes of under- and overfitting and not to miss any local minima in between. The polynomial degree 3 and clusters pool P1 were kept fixed, resulting in 120 features. Fig. 4.3 shows the convergence

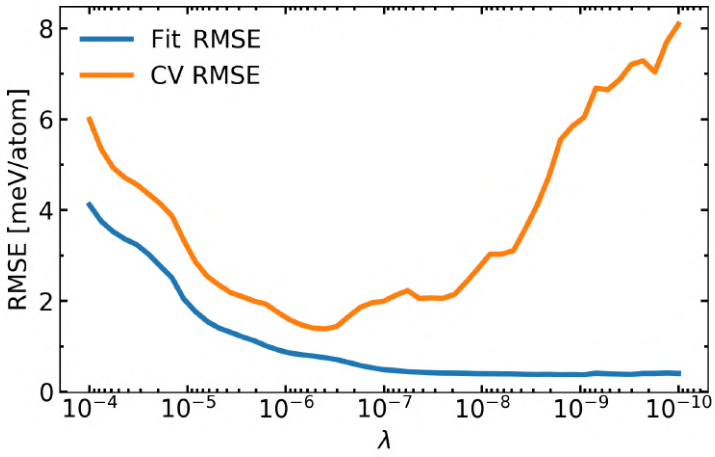


Figure 4.3. Root mean squared error versus LASSO regularization strength.

curve of the RMSE versus λ. The high RMSE-CVs for large and small values of λ describe under- and overfitting, respectively. Conversely, the RMSE-fit decreases monotonously for decreasing λ; this is expected since LASSO yields denser models as lambda is decreased. In the RMSE-CV curve, besides a few local minima (e.g., around 2×10^{-8}), a well-defined global minimum at the optimal $\lambda = 3.7 \times 10^{-7}$ is found. Thus, the lambda range for the grid-search CV in the results presented below is narrowed down to 5×10^{-5} to 1×10^{-7} for the subsequent studies.

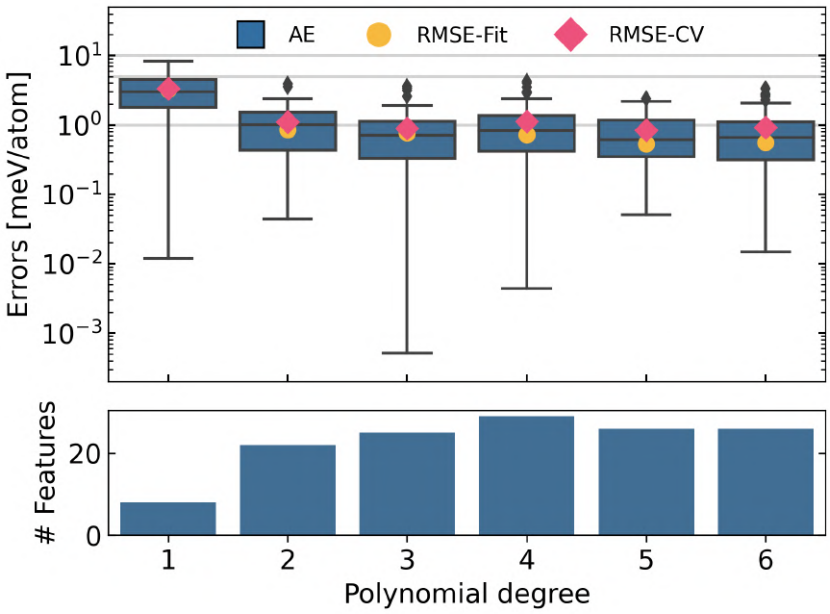


Figure 4.4. Boxplot of errors versus polynomial degree in nonlinear cluster expansion. The boxes show the absolute error (AE) per leave-one-out cross-validation (CV) split, and the yellow circle and red diamond display the fit and RMSE-CV in meV/atom. The number of features per model after feature selection is shown below. The three gray lines display 1 meV/atom, 5 meV/atom, and 10 meV/atom as visual guidelines. The best generalization with the lowest number of features is given for degree 3 (RMSE-Fit = 0.77meV/atom, RMSE-CV = 1.2meV/atoms, number of optimal features is 25).

Next, the influence of different degrees of the maximal order term in the polynomial feature expansion on the model performance is examined. As given by Eq. 28, higher-order polynomials vastly increase the model complexity. For instance, for the pool P1 used in this analysis, the polynomial degrees 2, 4, and 6 yield feature spaces of size 36, 330, and 1716, respectively. However, higher-order terms potentially capture behavior that otherwise requires many more clusters to describe. Fig. 4.4 shows a boxplot of the errors for models with varying polynomial degrees ranging from 1 to 6. The thin horizontal lines indicate the values 1 meV/atom, which corresponds to the estimated intrinsic error of the DFT data for this dataset (see Ref. [8]), 5 meV/atom, which is a moderate accuracy, and 10 meV/atom, which indicates a large error for this learning task. The boxplots show the distribution of absolute errors (AE) per LOO CV split, and RMSE-CV and -Fit are added to visualize the model convergence for increasing degrees. It can be seen that the RMSE-CV is usually higher than the median absolute error, which is expected since the RMSE is more sensitive to outliers than the MAE. As expected, the RMSE-Fit reduces for more complexity of the model induced by higher-order polynomials. By the reduction in the number of features after polynomial degree 4, it can be assumed that there exist higher-order features that are more suitable for fitting the energy. However, the RMSE-CV is comparable for degrees 3, 5, and 6, indicating convergence. Overall, the model with degree 3 appears to have the lowest CV error with a reasonable number of features. The errors are: RMSE-Fit = 0.77meV/atom, RMSE-CV = 1.2meV/atom, and the number of optimal features is 25. In conclusion, in the remaining calculations, we employ a polynomial degree of 3.

Lastly, the influence of the different clusters pools P0–P3 on the obtained CE models is analyzed. This analysis is done both for the nonlinear CE and the standard CE. In the second case, we distinguish two cases, namely, global standard CE models obtained using the whole data set and standard CE models using the split dataset as in Ref. [8] (see explanation at the start of Sec. 4).

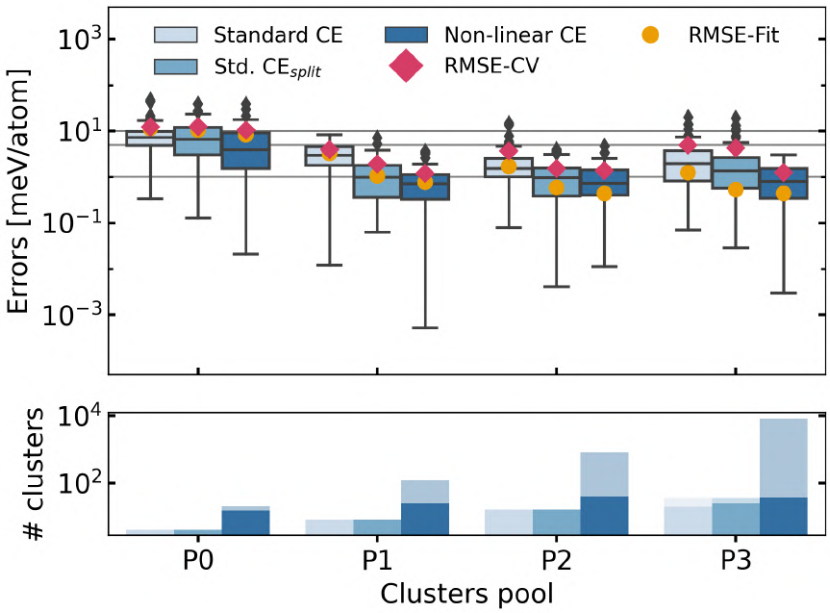


Figure 4.5. Boxplot of errors versus clusters pool (upper panel) and number of features in an optimal cluster expansion model (lower panel). The boxes show the distribution of absolute errors in LOO-CV, and the yellow circle and red diamond display the RMSE-CV and -Fit in meV/atom. The barplot below shows the number of features of the optimal CE model per CE type and clusters pool. The light-coloured bars represent the number of clusters before the cluster selection and the dark-coloured bars after the cluster selection, as used in the optimal model. With respect to RMSE-CV and model complexity, the nonlinear CE with P1 performs best.

These are denoted with "standard CE" and "standard CE$_{split}$ " in what follows. In Fig. 4.5, the prediction errors of the three CE models, namely, standard CE, standard CE$_{split}$, and nonlinear CE, for different clusters pools are presented as boxplots. It can be seen that the nonlinear CE and the standard CE$_{split}$ are more accurate than the standard CE for all four clusters pools. Furthermore, in all cases, the fit error decreases for larger clusters pools, as expected. The RMSE-CV for the nonlinear CE converges at pool P1, whereas in the case of the standard and the standard CE$_{split}$, the CV error increases after P2. For P2, the errors of the split CE and nonlinear CE are comparable. The best-performing models, regarding generalization error and robustness, are the ones obtained with the nonlinear CE for P1, P2, and P3. Hence, by favoring simplicity at comparable accuracy, clusters pool P1 appears to result in the best CE model for fitting the energy of mixing. Additionally, the RMSE is close to the 1 meV/atom mark, while the median absolute error is lower than 1 meV/atom. The errors for the optimal model, namely nonlinear CE using P1, are RMSE-Fit = 0.77meV/atom and RMSE-CV = 1.2meV/atom, with a number of optimal features of 25. The features per model are presented by the barplot in the lower panel of Fig. 4.5. For the nonlinear CE, an exponential increase in the amount of input features can clearly be seen. However, the amount of clusters of the optimal models only increase slightly from 25 to 39 for P1 to P2, respectively, and reduce again to 36 for P3. This shows that the vast number of features created by the nonlinear CE is not problematic as it is efficiently reduced by LASSO. Also, a larger, more diverse feature set does not necessarily improve performance. From the barplots of the standard and standard CE$_{split}$, we observe that until P3, all features contribute to the optimal model. Only for the largest clusters pool can a reduction be seen. It is different for both models, as the number of features of the standard CE$_{split}$ is the maximum of the two trained models as opposed to the standard CE.

The result presented in Fig. 4.5 is one of the central results of this thesis. It clearly illustrates how the nonlinear CE solves the problem of modeling a property with nonlinearities in a case where the standard CE fails unless physical intuition, namely the selection of the split point in the concentration axis, is employed. Our method does not require having such physical intuition beforehand.

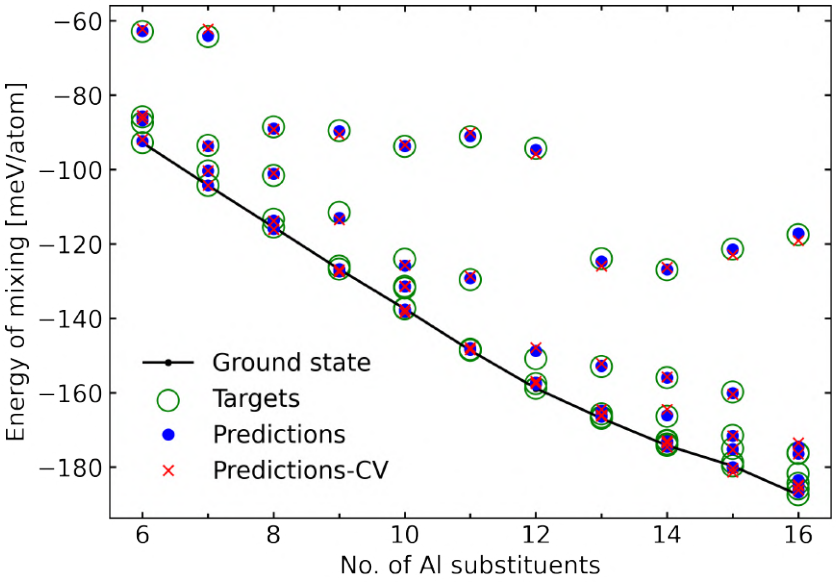


Figure 4.6. Energy of mixing versus the number of aluminum substituents fitted with the optimal nonlinear CE. The fit (blue dots), as well as the CV predictions (red crosses), lie accurately within the DFT energies (green circles), with a maximum absolute error of 3 meV/atom.

The energy predictions of the optimal nonlinear CE model are presented in Fig. 4.6. The predictions on the fitted data, as well as the predictions on CV, are very accurate, with a maximum absolute error of only 3 meV/atom. The high accuracy is especially prevalent for the ground state configurations, as indicated by the good overlap of circles and crosses inside circles, which indicate the target values computed with DFT. With 0.68 and 1.17 meV/atom, the RMSE-Fit and RMSE-CV are lower than the respective errors obtained with the standard CE_{split} of 1.06 meV/atom RMSE-Fit and 1.92 meV/atom RMSE-CV. Additionally, they are lower than the errors reported in Ref. [8] with 1.43 meV/atom and 1.45 meV/atom for RMSE-CV of the structures set from 6 to 13 substituents and from 13 to 16 substituents, respectively.

The optimal model contains 23 features that contribute to the expansion [57]. Their coefficients are displayed in Fig. 4.7. The three 1-point clusters are denoted with their Wykoff site letters k, i, and c, and the nearest neighbor pairs with the contributing 1-point clusters letters kk, ii, ki, and kc. The names of the nonlinear features are constructed as the combination of names of the contributing clusters, e.g., $X_{kc}X_{kk} = X_{kckk}$. One can see that the three 1-point and four 2-point nearest neighbors all contribute to the model, as is the case in the split CE model of Ref. [8]. Note that in Fig. 4.7, the coefficients are not normalized by their multiplicities. For comparison with Ref. [8], Tab. 4.1 shows the ECIs corresponding to the initial clusters pool of the nonlinear CE; those clusters make up the split CE of Ref. [8]. It can be seen that the 6c 1-point cluster has the lowest energy, followed by 24k and 16i. This ordering is also found for the split CE of the reference. Here, it is important to note that for those important interactions, the nonlinear CE gives similar results as the standard CE.

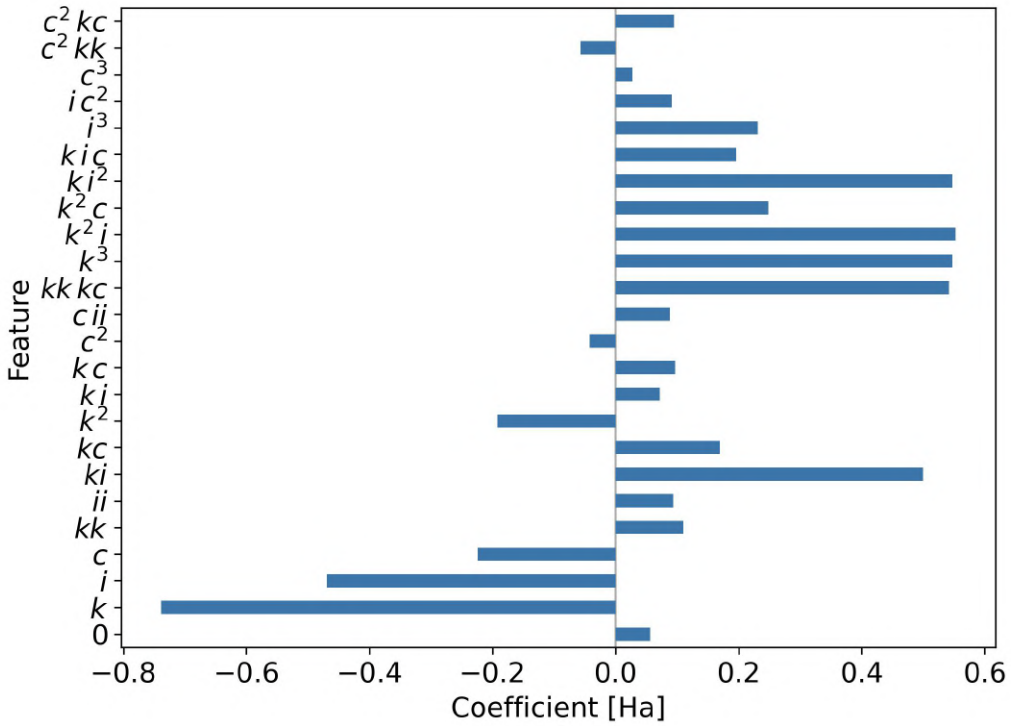


Figure 4.7. Coefficients of features as present in the optimal nonlinear CE model. The barplot gives an indication of the strength of feature importance in the model.

Furthermore, a great deal of insight can be obtained by re-arranging the interaction terms in the nonlinear CE in order to compare the behavior of the 2-point interaction ECIs as a function of Al concentration of the nonlinear CE and the split CE. This can be achieved by recasting the nonlinear CE in terms of an expansion in the initial clusters while factoring together nonlinear interaction terms containing the initial clusters. This is possible since the nonlinear features in the polynomial expansion are obtained by multiplying cluster correlations. Thus, for example, the following factoring is possible: $J_{kk\,kc}X_{kk\,kc} = (J_{kk\,kc}X_{kk})X_{kc}$. For the four 2-point interactions, this yields the coefficients

$$\begin{aligned} J_{kk}(\sigma) &= (J_{kk} + tJ_{kk\,kc}X_{kc}(\sigma) + J_{c^2\,kk}X_c(\sigma)^2)/m_{kk} \\ J_{kc}(\sigma) &= (J_{kc} + (1-t)J_{kk\,kc}X_{kk}(\sigma) + J_{c^2\,kc}X_c(\sigma)^2)/m_{kc} \\ J_{ii}(\sigma) &= (J_{ii} + J_{c\,ii}X_{ii}(\sigma))/m_{ii} \\ J_{ki}(\sigma) &= J_{ki}/m_{ki}\ . \end{aligned} \tag{36}$$

Table 4.1. Coefficients of nonlinear CE features and of the split CE clusters in Ref. [8] for splits 1 and 2. The values are given in meV. For the 1-point clusters at sites 16i and 24k, the difference between the lowest energy 1-point cluster 6c is given.

CE Model	$J_k - J_c$	$J_i - J_c$	J_c	J_{kk}	J_{ii}	J_{ki}	J_{kc}
Nonlinear	181.7	220.2	1019.5	250.1	318.5	283.4	192.4
standard CE_{split} 1	283.5	301.1	1281.9	265.7	296.4	213.6	258.8
standard CE_{split} 2	323.0	372.1	1282.1	150.5	327.0	163.2	259.8

Here, the coefficients $J_*(\sigma)$ are divided by the multiplicities of the respective 2-point interactions, which are 12, 24, 48, and 8 for m_{kk}, m_{kc}, m_{ii}, and m_{ki}, respectively. t is an arbitrary real number — any choice of t yields the same model. Here, we choose $t = 0.5$, such that the contribution of the correlation $X_{kk\,kc}$ is split in half for X_{kk} and X_{kc}. By performing this reordering of terms, the nonlinear CE can be interpreted as a CE with configuration-dependent interactions, which is an interesting finding, as it aligns with the derivation of concentration-dependent coefficients in Ref. [18].

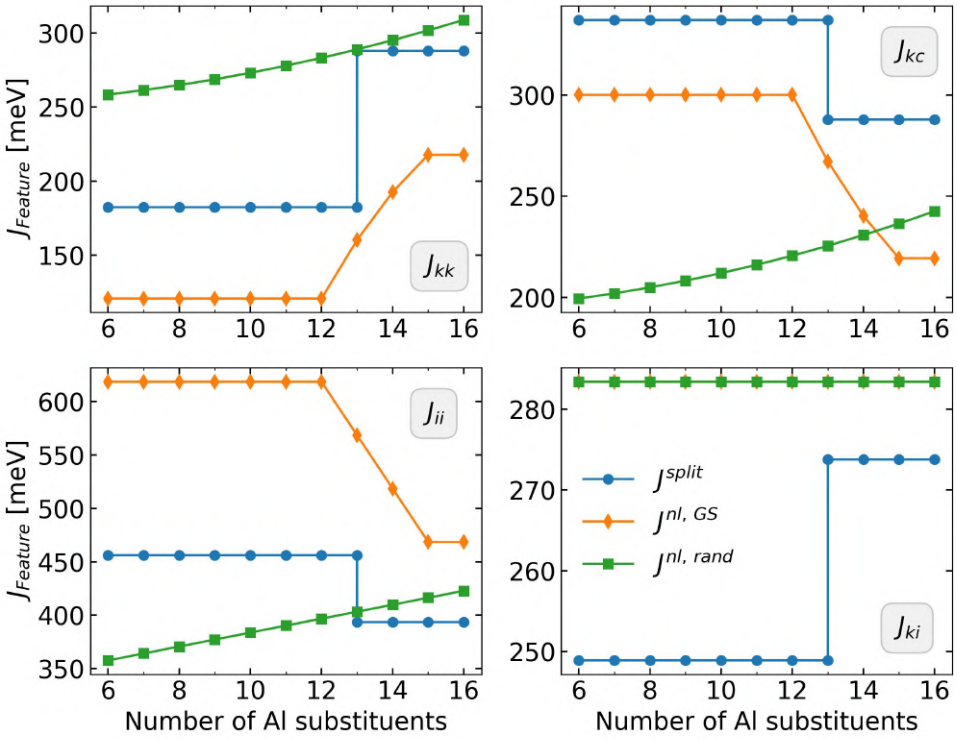


Figure 4.8. Comparison of the 2-point interaction ECIs over the range of Al substituents. The ECIs are given in meV. J_{kk} for the random alloy case, and the two constant ECIs of the split CE model taken from Ref. [8]. For the nonlinear CE, J_{ki} is a constant independent of concentration and of configuration. Thus the $J^{nl,GS}$ and $J^{nl,rand}$ overlap.

Employing the configuration dependence of cluster interactions of Eq. 36, it is interesting to consider two limiting cases, namely, the GS configurations and the random alloy. These limiting cases correspond, respectively, to the alloy in thermodynamical equilibrium at $T = 0$ and at $T \to \infty$, and are depicted in Fig. 4.8. They are compared with the piece-wise constant ECIs of the split CE models of $Ba_8Al_xSi_{46-x}$ calculated with LDA taken from Ref. [8]. The calculation of the ECIs for the random alloy is performed using the fact that the 1- and 2-point cluster correlations in that case are identical to x and x^2, respectively (see Sec. 2.3). It is observed that, in most cases, the $J^{nl,GS}$ show a similar trend to that of the split CE model, J^{split}, the main qualitative difference being a smooth transition of the interaction value around the split point, instead of an abrupt change. Note that this transition of the interaction value is obtained directly by the nonlinear CE method, while in the split CE, physical insight outside the CE formalism was needed. In general, the $J^{nl,rand}$ shows a smooth monotonous increase with concentration. Interestingly, in some cases (kk, ii), the interactions of the split CE lay in between the two limiting cases found with the nonlinear CE.

From these results, we can assume that the nonlinear CE is better able to capture the nonlinear behavior in the energy of mixing than a split standard CE. Most notably, no prior knowledge of the other material properties is required. Additionally, only one model is needed for training and for making predictions, which simplifies the application of the CE model, such as for calculating thermodynamic averages. To obtain thermodynamic properties, methods such as Markov Chain Monte Carlo [58] are used for sampling the configurational space of the material. These require many energy predictions of new configurations, where the cluster correlations have to be recalculated. Therefore, the nonlinear approach can be more efficient as it only requires the correlation evaluation for a small set of 7 clusters, P1, and can then expand the feature space at a low cost through polynomial expansion.

4.4 Band Gap

For $Ba_8Al_xSi_{46-x}$ to have good thermoelectric performance, semiconducting behavior is needed. However, in $Ba_8Al_xSi_{46-x}$, the band gap changes strongly depending on the Al concentration and the configuration. This also includes metal-to-semiconductor transitions within the same composition. Constructing CE models for the bandgap is hampered by such a nonlinear discontinuity of bandgap energies, which can take a continuous range of positive values or zero, but not negative values. This has been demonstrated in Ref. [10], where the bandgaps are predicted with the split CE technique explained in Sec. 4.3, and which is only trained on semiconducting samples, such to elude the problem of strong non-linearities in the regression problem. The quality of the found models in Ref. [10] was rather poor. As for the energy of mixing, we expect an improvement in bandgap fitting with the nonlinear CE.

For comparison, the bandgap energies were fitted with the standard and the non-linear CE. Cluster selection was performed with the OMP algorithm. As explained in Sec. 3.2, this algorithm efficiently finds a subset of relevant features for an optimal representation. The number of features is a hyperparameter and can be used to easily adjust the model size. A 10-fold CV with ten random runs was chosen for hyperparameter selection and model evaluation. More precisely, the training set (78 structures) is randomly split into ten folds for CV; this is repeated ten times to obtain a better estimate of the RMSE-CV in terms of its average and standard deviation. One small and one large clusters pool were used. The small clusters pool, P_s, is the same as P1 in the previous section, i.e., seven clusters: the three 1-point clusters and the four first neighbor 2-point clusters. In contrast, the large pool, P_l, consists of all 1-, 2-, and 3-point clusters that fit in the 54-atom unit cell of the clathrate. This sums to 368 clusters. Polynomial expansion of P_s to the order 2 and 3 results in 36 and 120 features, respectively. The CE models trained in the study were:

- a standard CE with P_s (7 features),
- a nonlinear CE with initial clusters pool P_s with polynomial degree 2 (36 features),
- a nonlinear CE with initial clusters pool P_s with polynomial degree 3 (120 features), and
- a standard CE with large clusters pool P_l (368 features).

We first compare the fitting accuracy of the standard CE with P_s and P_l and the nonlinear CE with degree 3. Fig. 4.9 shows prediction versus target and convergence plots for all three models.

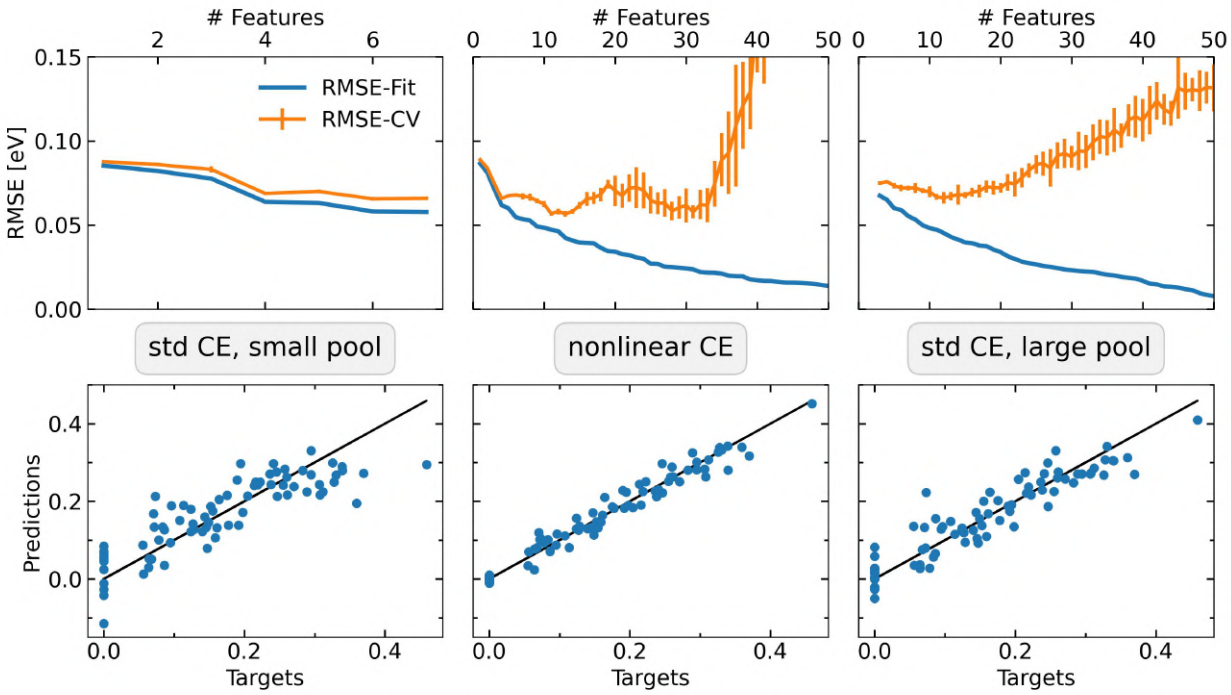


Figure 4.9. Top: Convergence of RMSE-Fit and RMSE-CV with one-sigma standard deviation intervals as vertical bars for RMSE-CV. Optimal models yielding the largest generalizability are obtained by selecting the number of features with the lowest RMSE-CV within one-sigma: 6, 33, and 14 features for the standard CE with P_s, the nonlinear CE, and the standard CE with P_l, respectively. Bottom: Band gap predictions versus DFT targets in eV.

Looking at the convergence of RMSE-CV for the nonlinear CE and the standard CE with P_l, we can see that both models show overfitting quickly. However, the nonlinear model displays two minima: a first minimum between 11 and 13 features, and a second between 28 and 33 features. Within the one-sigma standard deviation interval, models with comparable precision exist in both ranges. Therefore, we can conclude that the nonlinear CE model with 33 nonzero coefficients is the best model in terms of fit and CV error, with an RMSE-Fit of 0.023 eV and RMSE-CV of 0.057 eV. The standard CE with P_l also exhibits a global minimum at its first minimum. This indicates that the model does not generalize well with more clusters. Its RMSE-Fit and RMSE-CV at 14 clusters are 0.041 eV and 0.065 eV, respectively. In contrast, the standard CE with P_s yields comparable RMSE-CV, but the worst RMSE-fit. At six clusters, the RMSE-fit and RMSE-CV for standard CE with P_s are 0.058 eV and 0.064 eV, respectively. These results are summarized in Tab. 4.2.

Model	No. Features	RMSE-Fit	RMSE-CV
Std. CE, P_s	6	0.058	0.064
Nonlinear CE	33	0.023	0.057
Std. CE, P_l	14	0.041	0.065

Table 4.2. Number of features, RMSE-Fit, and RMSE-CV for the optimal models Fig. 4.9. An optimal model is chosen for the minimal RMSE-CV within a one-sigma standard deviation and the lowest RMSE-Fit. The errors are given in eV.

The predictions of these three optimal models are displayed in the three parity plots in the lower panel of Fig. 4.9. Note that the predictions are obtained by training and predicting on the whole

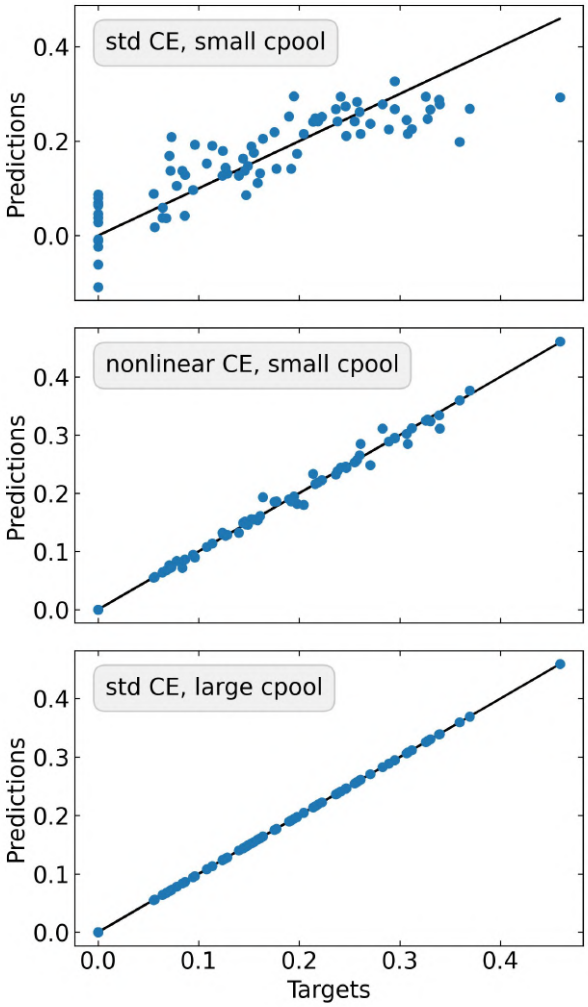


Figure 4.10. Predictions versus targets of the optimal fit predictions with three CE models: standard CE with small clusters pool (top), nonlinear CE with degree 3 and small initial clusters pool (middle), and standard CE with large clusters pool (bottom).

dataset; they reflect the RMSE-Fit at the optimal number of features in the upper panel. As expected, the accuracy rises with an increasing number of features. In the nonlinear case, this is particularly true for the metals, which appear as a collapsed cloud of points at the origin and are fitted much better than in the other two cases.

In the previous analysis, the models with the best generalisability were found. Nonetheless, it is also interesting to analyze the ability of the available features to fit the data, with disregard for the model's generalisability. Since ultimately, this determines the quality of the feature space. To this end, parity plots of the prediction of models converged on the RMSE-Fits are shown in Fig. 4.10. The data distribution demonstrates a significant improvement in the predictive performance of the nonlinear CE model compared to the standard CE with the small pool P_s, particularly for the metals. The RMSE-Fit values are recorded as 0.058 eV and 0.0089 eV, respectively. This indicates that by simply adding nonlinear features to a small clusters pool, the quality of the feature space can be highly improved. However, we observe that this 3rd-order feature space does not suffice to find an exact solution, as can be obtained with the standard CE with P_l. The reason is that, for this case, the number of clusters of the model matches the number of samples in the data set (78),

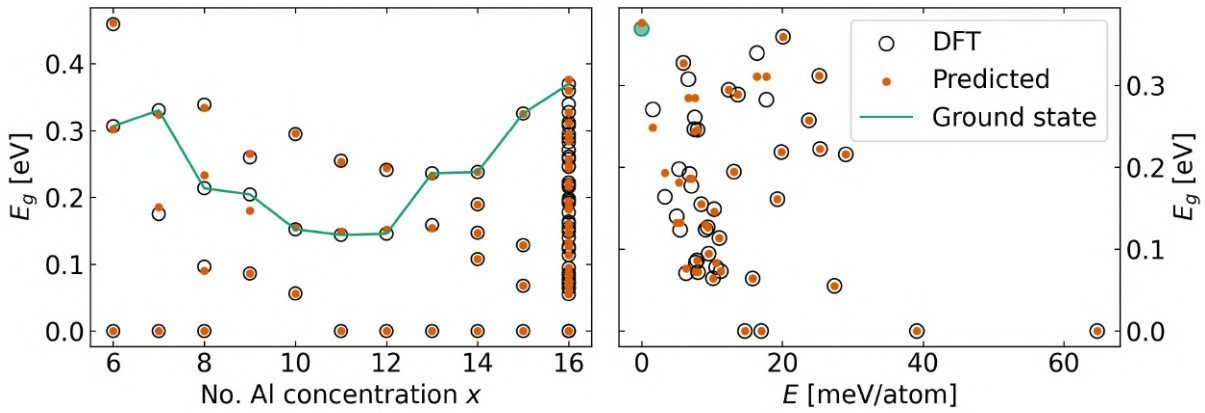


Figure 4.11. Predicted (orange dots) and *ab-initio* (black circles) band gap energies of the nonlinear CE with degree 3 and P_s. The model uses 63 features.

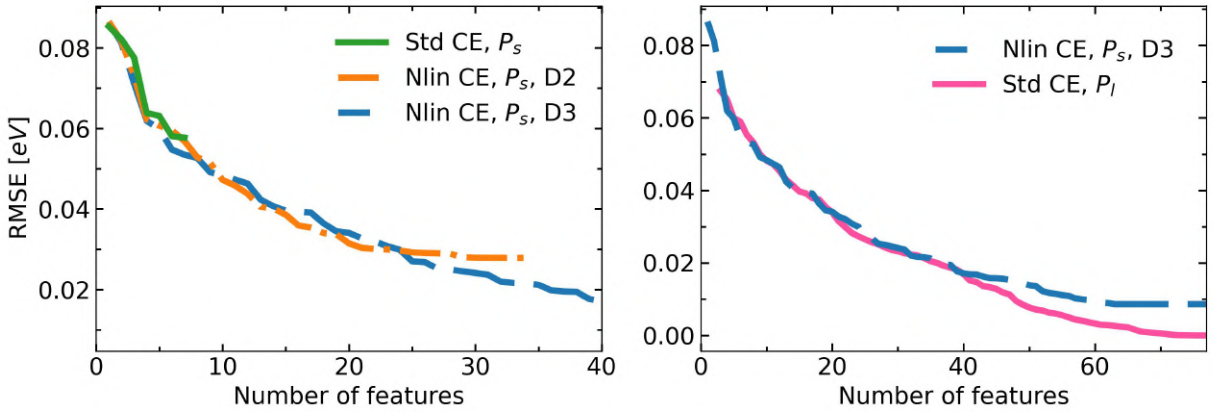


Figure 4.12. Convergence plots of four different CE models. Left: RMSE-Fit versus the number of clusters for standard CE, nonlinear CE with degree 2 (D2), and nonlinear CE with degree 3 (D3), all with clusters pool P_s, containing seven clusters. Right: RMSE-Fit versus the number of clusters for the nonlinear CE with D3 and P_s and the standard CE with P_l.

and the linear problem (assuming colinearities are absent) can be solved exactly. Thus, the bottom panel shows a perfect fit. Here, it must be stressed that to obtain the excellent fit of the middle panel of Fig. 4.10, one needs to compute only seven cluster correlations, while the bottom panel requires the computation of 368 cluster correlations.

Next, we want to analyze how the band gap predictions of the nonlinear CE distribute over the configurations. Fig. 4.11 depicts DFT (black circles) and predicted band gaps (orange dots) made with the nonlinear CE from Fig. 4.10. Over the full compositional range (left panel), the predictions are mostly accurate, especially for metals. This is also mostly true at the charge-neutral composition (right panel), although some predicted band gaps deviate visibly from the DFT targets.

Finally, it is instructive to compare the convergence of the RMSE-Fit against the number of features for the standard CE with P_s, the nonlinear CE to the polynomial order 2 and 3 and the standard CE with P_l. Fig. 4.12 describes how much the band gap data fit improves for an increasing number of features in the respective CE model. On the left, we see how much the nonlinear CE with P_s, containing only 7 clusters, can increase the number of features and how strongly this improves the

fit. While the standard CE can only reach a maximum accuracy of 0.058 eV, the nonlinear CE at the lowest degree, namely D2, is able to halve the error. A further reduction to nearly exact predictions (RMSE-Fit = 0.0089 eV) can be reached by expansion to the 3rd order. The right plot compares the RMSE-Fit of the nonlinear CE model with degree 3 and a standard CE with P_l. Up until around 40 features, both models are comparable in precision (RMSE-Fit: nonlinear CE = 0.019 eV, standard CE = 0.017 eV). Eventually, the standard CE with P_l converges to machine precision, which is not the case for the nonlinear CE. Nonetheless, one can then deduce that a CE model with features created out of a very small set of cluster correlations is comparable to a model created with a very large number of clusters. This is a key finding, as calculating the cluster correlations accounts for the highest computational effort in training, and, more importantly, when predicting, indicating a strong advantage in terms of computational efficiency for the nonlinear CE.

In conclusion, the comparative analysis of fitting accuracy and convergence behavior among different CE models highlights the advantages of using nonlinear features to improve the accuracy of a CE model when fitting $Ba_8Al_xSi_{46-x}$ band gaps. By polynomially expanding the small initial clusters pool P_s, it is possible to create models with a comparable level of accuracy to those generated using a much larger pool of clusters. This represents a significant advantage from the perspective of computational efficiency, given that calculating cluster correlations is the most computationally demanding aspect of generating predictions with CE. Overall, the study highlights the potential of this approach to improve the accuracy of CE models in a computationally efficient way.

4.5 Nonlinear CE for Classification

Determining if a structure of $Ba_8Al_xSi_{46-x}$ is a metal or semiconductor is a task that can be solved using classification techniques from ML. Here, we devise a method to combine CE with ML for classification. The idea is to create descriptors for the structures with CE and then find a representation in which data separation with the linear support vector machine described in Sec. 3.3 is possible. Next to using the standard CE, we employ the nonlinear CE to enable nonlinear decision boundaries in linear feature space.

For this study, we evaluated the cluster correlations of the data set with the three clusters of pool P0 (only the three 1-point clusters), and with the seven clusters of P1 (three 1-point and four 2-point clusters). By then polynomially expanding both correlation matrices to the degree of 2 and 3, we end up with a total of six classifiers. Their dimensionalities are shown in Tab. 4.3. The 12 metals and 66 semiconductors in the data set were labeled with 1 and -1, respectively.

	D1	D2	D3
P0	3	9	19
P1	7	35	119

Table 4.3. Dimension of feature space for clusters pool P0 and P1 and polynomial degrees D1, D2, and D3.

For the seven-dimensional classifier, we observed perfect classification when training the classifier and predicting the same data. The confusion matrix of its predictions is shown in Fig. 4.13 (right panel). The same results could also be obtained when using the 2- and 3-order classifiers. However, for P0, none of the three classifiers yields a perfect fit, as is presented by the four false-classified metals

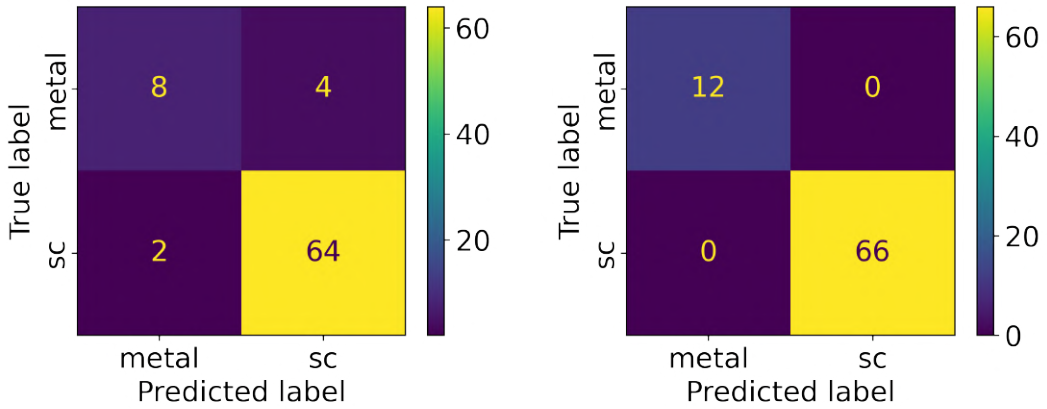


Figure 4.13. Confusion matrix of support vector classifier trained with three 1-point and four first neighbor 2-point clusters.

and two false-classified semiconductors of the 3-order classifier shown in Fig. 4.13 (left panel). This indicates that perfect separation of the data requires the inclusion of at least two-body interactions.

It is indeed possible to identify a descriptor that enables a representation where a linear decision boundary can effectively separate the data. To visualize the data, we map the data points into a hyperplane perpendicular to the decision boundary. We achieve this by projecting the data points along a vector parallel to the decision boundary and along the vector β of Eq. 29, which is perpendicular to the decision boundary.

This projection is used in Figure 4.14, which shows the semiconductor and metal regions, as predicted by the classifier, in colors green and red, respectively. The boundary between the two colors corresponds to the decision boundary found by the support vector machine. The classified target and prediction labels of the six classifiers are also depicted. For P1 (lower panels), we clearly see that the decision boundary can perfectly separate between the two classes, in accordance with the confusion matrix of Fig. 4.13, right panel. As expected, we observe that nonlinear features improve separability as the margin of the decision boundary (indicated with thin gray lines) increases while the clouds of data points contract. Conversely, for P0, we see that the representation is unable to transform the data to separate the true labels. Rather, perfect class separation is impossible, even at degree 3. This points to the fundamental importance of the presence of 2-point cluster correlations to capture the metal or semiconductor character of the samples. For increasing polynomial order, the misclassified points move closer to the boundary, as indicated by the thin gray lines, which could mean that the separation is still improved. It should also be noted that, using the small clusters pool P0, we were not able to perfectly classify the metals and semiconductors with more sophisticated fine-tuned models. These included support vector machines that utilize the kernel trick for efficiently exploiting higher dimensional feature spaces. We employed the polynomial kernel with high orders and the radial basis function kernel (Gaussian kernel), which theoretically moves to infinite dimensions.

From this, we conclude that it is not possible to perfectly classify metals and semiconductors by just using the site occupation factors (i.e., the one-point cluster correlations and features derived from them) as a descriptor and that many-body interactions are required. Indeed, when including many-body interactions, it is possible to use the CE in combination with the support vector machine to find descriptors that enable a representation in which separation by a liner boundary is possible.

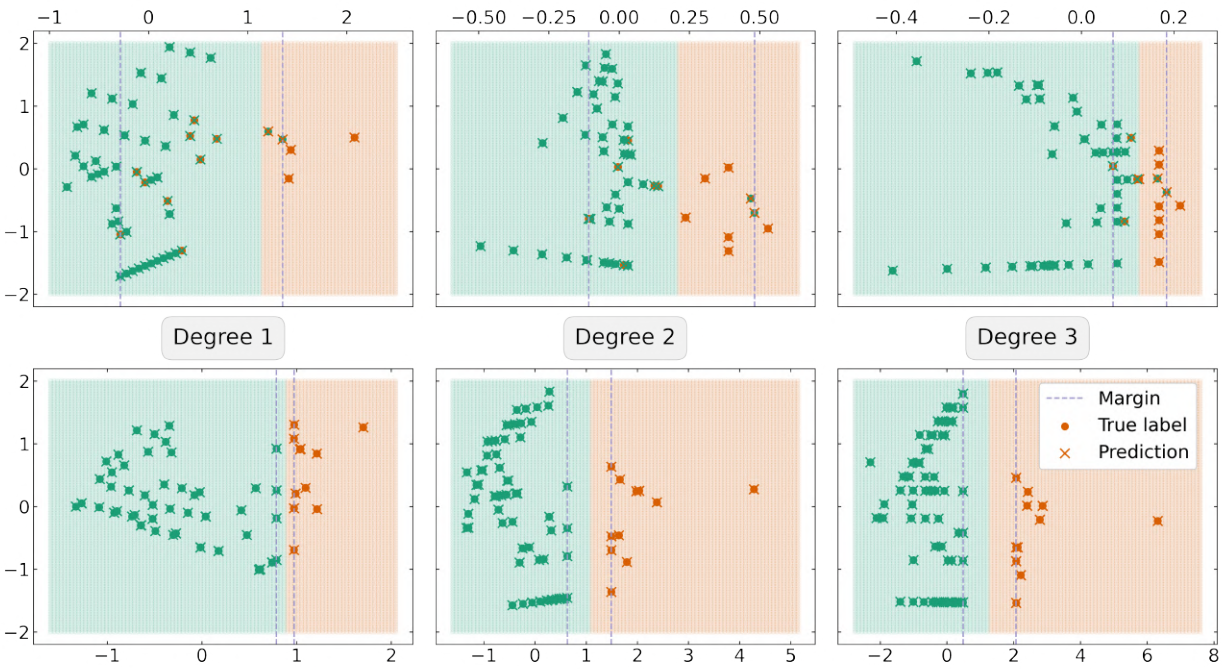


Figure 4.14. Classifier decision boundary in 2D representation of input data.

The new application of the CE method presented here has the potential to be applied to a wide range of problems in materials science.

5

Implementation in CELL

The studies presented in this work have been conducted with the Python package `CELL` (Cluster Expansion for large parent ceLLs) [59], [60]. `CELL` implements the CE method for alloy materials with a focus on complex unit cells. It is structured modularly such that the construction of workflows for CE models can be tailored to specific needs in a flexible way. The `Atoms` class of Atomic Simulation Environment is used as a base class for the creation of crystal structures, which allows for the creation of databases of structures and to interface to various *ab-initio* calculators. Futhermore, it provides an interface to estimators from the machine learning library `scikit-learn` [61].

In `CELL`, the easiest way to create a CE model is by using the helper class `ModelBuilder`. The `ModelBuilder` combines the main steps of the CE method and returns a fitted CE model with minimal user input. Prerequisites for a CE are a set of structures with known property values and a pool of clusters. These are realized in the `StructuresSet` and `ClustersPool` classes. An instance of the `StructuresSet` class also contains the known property values of each structure. All three can be created and calculated directly in `CELL`, or the structure parameters and property values can be imported from external files. The latter allows for the universal usage of precalculated data sets, such as DFT calculations. A more detailed description of the construction of sets of structures and pools of clusters with `CELL` is given in the tutorials on the CELL website [60] and in Ref. [59].

Listing 5.1 presents the `ModelBuilder` function call to create a nonlinear CE with polynomial features to the third order. Note that `CELL` is imported with the Python package name `clusterx` in Python code. It is assumed that a tructures set with total energies (`sset`) and a clusters pool (`\cpool`) have already been created. Two flags are relevant for the nonlinear CE: `nonlinear=True` initializes a nonlinear model, and `nonlinear_degree=3` sets the maximum polynomial degree to three. Note that using only one of the flags suffices to invoke a nonlinear CE. Only setting `nonlinear=True`, a default second-order expansion is performed, and passing a value to `nonlinear_degree`, the `nonlinear` flag is automatically activated. Passing `"LassoCV"` to the `estimator_type`, `scikit-learn`'s Lasso estimator with built-in CV is invoked. The desired parameters can be passed as a dictionary with `estimator_opts`. The method `build()` performs all steps to create a correlation matrix, perform feature expansion and selection, fit the estimator model, and return a CE model. The CE model can be used for predictions or for analysis, such as to report the fitting and CV errors.

To illustrate how the `ModelBuilder.build()` method works internally, List. 5.2 shows the individual steps if one were to build a CE model without using the helper function. Listing 5.2

```python
# Assuming a structures set (sset)
# with property values ("total_energy_emt")
# and a clusters pool (cpool) are created.

from clusterx.model import ModelBuilder

mb = ModelBuilder(
    estimator_type="LassoCV",
    estimator_opts={
        "fit_intercept": False, "cv": 10, "alphas": np.logspace(-2, -7, 50)
    },
    nonlinear=True,
    nonlinear_degree=3
)
ce_model = mb.build(sset, cpool, "total_energy_emt")
ce_model.report_errors(sset)
```

Listing 5.1. Creation of a nonlinear CE model with the helper class `ModelBuilder`.

exposes the main (internal) steps of constructing a nonlinear CE, omitting other features of the `ModelBuilder` implementation that would be distracting for this description. The first step in cluster expansion is to evaluate the cluster correlations. This starts with initializing the `CorrelationsCalculator` and calculating the correlation matrix between the structures set and the clusters. The `CorrelationsCalculator` is also needed to calculate single structure correlations when making predictions. In lines 21 and 22, the input data are then set up, namely the correlation matrix `X` and the array of material properties `y`.

When considering the usage of a fitted CE model for predictions on new data, the most efficient implementation of a polynomial expansion is a pipeline that fuses the expansion of the correlation matrix with an estimator. This pipeline can then be used interchangeably as an estimator. It can be fitted, and the fitted estimator can be passed to functions for making predictions. With the `EstimatorFactory` in List. 5.2 line 25 `scikit-learn`'s linear model `LassoCV` is initialized with the passed estimator parameters. It is combined in a scikit-learn `Pipeline` with the preprocessing step `PolynomialFeatures` in the following lines. The pipeline is then fitted to the input data, which invokes two steps. First, nonlinear features are created by expanding the matrix polynomially. Second, fitting `LassoCV`, the optimal nonlinear features are selected by finding the optimal regularization parameter (here `alpha`). Consequently, the fitted pipeline includes a fitted estimator with a number of nonzero coefficients that represent the optimal feature set. The last step of the `ModelBuilder` is to create a CE model. This is seen in lines 37–39, where the `Model` object is initialized with the initial correlations calculator, the fitted estimator and the corresponding property, and the `nonlinear` flag set to `True`. The flag is required for analysis steps, such as returning the actual estimator coefficients. When predictions are made with the CE model, first, correlations of a new structure with the initial clusters pool are calculated with the initial correlations calculator, and second, they are passed to the pipeline, expanded, and multiplied with the coefficients to evaluate the prediction.

In comparison with the standard case, when creating a nonlinear CE model with the `ModelBuilder`, two points must be noted: one of `nonlinear` or `nonlinear_degree` have to be set, and the estimator used should implement a CV scheme and perform feature selection. More advanced applications of

```python
# Assuming a structures set (sset) with property values ("total_energy_emt"),
# a parent lattice (platt) and a clusters pool (cpool) exist.

from clusterx.correlations import CorrelationsCalculator
from clusterx.clusters_selector import ClustersSelector
from clusterx.model import Model
from clusterx.estimators.estimator_factory import EstimatorFactory

from sklearn.pipeline import Pipeline
from sklearn.preprocessing import PolynomialFeatures

# ModelBuilder parameters:
estimator_type = "LassoCV"
estimator_opts = {"fit_intercept": False, "cv": 10, "alphas": np.logspace(-2, -7, 50)}
nonlinear_degree = 3 # max order of polynomial expansion

# Correlation matrix
corrcal = CorrelationsCalculator(
    basis="trigonometric", parent_lattice=platt, clusters_pool=cpool
)
X = corrcal.get_correlation_matrix(sset, outfile="corrmat.dat")
y = sset.get_property_values("total_energy_emt")

# Create estimator with pipeline
lasso = EstimatorFactory.create(estimator_type, **estimator_opts)
estimator = make_pipeline(
    PolynomialFeatures(include_bias=False, degree=nonlinear_degree),
    lasso
)
estimator.fit(X, y)

# Create CE model
ce_model = Model(
    corrc=corrcal, property_name="total_energy_emt", estimator=estimator, nonlinear=True
)

ce_model.report_errors(sset)
```

Listing 5.2. Manual creation of a nonlinear CE model with `CELL`.

the nonlinear CE can be realized by following the individual steps indicated here instead of using the helper class `ModelBuilder`. Most notably, for other machine learning tasks, such as classification, the relevant functionality of `CELL` is the creation of the correlation matrix and, thus, the implementation of the cluster function, the structures, the clusters, and their lower-level objects.

6

Conclusions

In this work, the cluster expansion method has been augmented with machine learning techniques to improve the modeling of materials properties that show nonlinear concentration dependencies. This is relevant as a contribution to the discovery of novel materials for high-tech devices. An example of such materials is the thermoelectric clathrate $Ba_8Al_xSi_{46-x}$, which is hoped to provide similar qualities for the usage in heat-energy conversion components as GaGe-based clathrates. To assess good thermoelectric behavior, the theoretical description of material properties over a range of compositions is required. Given the vast configurational space of the crystal, this is computationally unfeasible with common methods like DFT but can be accomplished with the cluster expansion method. However, this material exhibits nonlinear behavior of its properties as a function of the atomic configuration, specifically the energy of mixing and the bandgap. This poses insurmountable difficulties to the standard CE, which does not converge when nonlinearities are present. In this thesis, we show how this lack of convergence leads to the emergence of spurious interactions, which hamper an accurate description of materials properties with CE.

To solve the issue, we proposed an innovative solution that we call the "nonlinear CE". First, we demonstrated that one can conveniently reformulate the CE method as a linear problem, which makes cluster expansion amenable to the application of machine learning techniques. Most notably, this enables us to augment the feature space to include nonlinearities by polynomially expanding the cluster correlations. Moreover, we argue that this insight opens up the CE method for use in classification tasks with the support vector machine. To demonstrate the effectiveness of the nonlinear CE, we were able to exactly model the nonlinear properties of toy models as the squared concentration of the alloy. In this case, we observe the lack of convergence and the introduction of spurious interactions with standard CE. Conversely, our novel method shows exact convergence with a few expansion terms. From this, we can conclude that the nonlinear CE can effectively model nonlinear properties of materials, such as the ones solving the long-standing problems introduced in the discussion of Ref. [18] and Ref. [19].

As a real-world test, we studied the energy of mixing of $Ba_8Al_xSi_{46-x}$ with the nonlinear CE in comparison to the standard CE. The data set was taken from Ref. [8]. From a systematic study of the hyperparameters relevant to the nonlinear CE, we derived an optimal model, which only required the three 1-point and four first neighbor 2-point clusters expanded to the 3rd order. Excitingly, this model outperformed the standard CE and reached a RMSE-CV close to the estimated intrinsic error of the DFT data set. Upon examination of the ECIs of the nonlinear CE, we observed a

good agreement with the findings reported in Ref. [8]. Additionally, by analyzing the two-point interactions as a function of Al substituents, we found configuration-dependent interactions, which are consistent with the concentration-dependent coefficients introduced in Ref. [18] and introduce a temperature dependency. Based on these results, we conclude that the nonlinear CE is more proficient in capturing the nonlinear behavior of the energy of mixing compared to the standard CE. Most notably, no prior physical intuition is required. This is an important advantage of our method in comparison with standard CE, which required the specification of physically motivated domains of applicability to circumvent the issue of nonlinearities in Ref. [8]. Moreover, we have also shown that the nonlinear CE boosts the application of the CE method for computationally expensive tasks, as it suffices with the calculation of a small number of cluster correlations and their cheap nonlinear expansion. This is considerably advantageous when, for example, calculating thermodynamic averages, which require millions of property predictions where the cluster correlations have to be recalculated each time. This is underlined by fitting the $Ba_8Al_xSi_{46-x}$ bandgaps. Here, we show with a comparative analysis of fitting accuracy and convergence behavior that using nonlinear features strongly improves the ability of a CE model to fit $Ba_8Al_xSi_{46-x}$ band gaps. We demonstrate that comparable fit accuracies to those obtained with a standard CE with hundreds of clusters can be achieved with a nonlinear CE, which only requires the evaluation of seven cluster correlations.

Lastly, we showcased how the CE can be employed in a novel context: classification. For this, we used CE to construct a descriptor of the material, which allowed for the separation of the $Ba_8Al_xSi_{46-x}$ structures into metals or semiconductors. We observed that for a descriptor including the seven smallest clusters a linear decision boundary could perfectly separate the data and moving to higher dimensions simplified this separation visibly, whereas no separation was possible with only the three 1-point clusters. Based on these findings, we can deduce that using only the site occupation factors as descriptors is insufficient for accurately classifying metals and semiconductors. To achieve better results, the inclusion of many-body interactions was found to be imperative. In sum, we have demonstrated that the application of the novel CE method presented here has the potential to be successfully applied to various domains within materials science.

Even though the obtained results are promising, multiple angles for refinement exist. Assessing the generalizability of the computed models can be problematic if the datasets are small. Some of these difficulties were exposed in this work when assessing the CE models for the bandgap in Sec. 4.4. Thus, it would be interesting to employ this methodology in larger datasets, where the differences with the standard CE in terms of generalizability could be better tested. It is also worth exploring different ML algorithms. In order to introduce the nonlinear CE from the basics, we focussed on polynomial expansion; however, transformations in terms of other functions, such as the logarithm or the square root, have also been proposed in the ML literature [42]. Also, in the context of symbolic regression [62], the construction of descriptors by combining such functions and selecting the optimal mathematical expression could represent a generalization of our method. One can also access infinite dimensional feature spaces by employing kernel ridge regression with, e.g., Gaussian kernels, which take as input cluster correlations. Moreover, we showed how CE can be utilized as a materials descriptor, and with it, we introduced the CE method to classification tasks. As multicomponent materials often incorporate classifiable properties such as metal-to-semiconductor or order-to-unorder transitions, this realm of application should be further investigated.

In conclusion, the nonlinear CE represents a significant advancement that not only enhances the applicability of CE to existing use cases but also unveils opportunities for novel applications in the realm of materials discovery. As datasets grow and methodologies evolve, the future holds promise for even more robust and versatile CE-based approaches.

Bibliography

[1] G. A. Slack, "Thermoelectricity and thermal conductivity in semiconductors," in *CRC Handbook of Thermoelectrics*, D. M. Rowe, Ed., Boca Raton, FL: CRC Press, 1995, pp. 407–440.

[2] G. J. Snyder and E. S. Toberer, "Complex thermoelectric materials," *Nature Materials*, vol. 7, no. 2, pp. 105–114, 2008. DOI: 10.1038/nmat2090.

[3] G. S. Nolas, J. L. Cohn, G. A. Slack, and S. B. Schujman, "Semiconducting Ge clathrates: Promising candidates for thermoelectric applications," *Applied Physics Letters*, vol. 73, no. 2, pp. 178–180, Jul. 1998, ISSN: 0003-6951. DOI: 10.1063/1.121747. [Online]. Available: https://doi.org/10.1063/1.121747.

[4] J. Roudebush, C. Cruz, B. Chakoumakos, and S. Kauzlarich, "Neutron diffraction study of the type i clathrate ba8alxsi46-x: Site occupancies, cage volumes, and the interaction between the guest and the host framework," *Inorganic chemistry*, vol. 51, pp. 1805–12, Dec. 2011. DOI: 10.1021/ic202095e.

[5] C. L. Condron, S. M. Kauzlarich, T. Ikeda, G. J. Snyder, F. Haarmann, and P. Jeglič, "Synthesis, structure, and high-temperature thermoelectric properties of boron-doped ba8al14si31 clathrate i phases," *Inorganic Chemistry*, vol. 47, no. 18, pp. 8204–8212, 2008, PMID: 18710218. DOI: 10.1021/ic800772m. [Online]. Available: https://doi.org/10.1021/ic800772m.

[6] H. Anno, M. Hokazono, R. Shirataki, and Y. Nagami, "Crystallographic, thermoelectric, and mechanical properties of polycrystalline ba8alxsi46-x clathrates," *Journal of Electronic Materials*, vol. 42, no. 8, pp. 2326–2336, 2013.

[7] N. Tsujii, J. H. Roudebush, A. Zevalkink, C. A. Cox-Uvarov, G. Jeffery Snyder, and S. M. Kauzlarich, "Phase stability and chemical composition dependence of the thermoelectric properties of the type-i clathrate ba8alxsi46-x ($8 \leq x \leq 15$)," *Journal of Solid State Chemistry*, vol. 184, no. 5, pp. 1293–1303, 2011, ISSN: 0022-4596. DOI: https://doi.org/10.1016/j.jssc.2011.03.038. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0022459611001496.

[8] C. D. Maria Troppenz Santiago Rigamonti, "Predicting ground-state configurations and electronic properties of the thermoelectric clathrates ba8alxsi46-x and sr8alxsi46-x," *Chem. Mater.*, vol. 29, no. 6, pp. 2414–2424, 2017. DOI: https://doi.org/10.1021/acs.chemmater.6b05027.

[9] M. Troppenz, S. Rigamonti, J. O. Sofo, and C. Draxl, "Partial order-disorder transition driving closure of band gap: Example of thermoelectric clathrates," *Phys. Rev. Lett.*, vol. 130, p. 166 402, 16 Apr. 2023. DOI: 10.1103/PhysRevLett.130.166402. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevLett.130.166402.

[10] M. Troppenz, "Electronic transport properties of thermoelectric materials with a focus on clathrate compounds," Ph.D. dissertation, Humboldt-Universität zu Berlin, Mathematisch-Naturwissenschaftliche Fakultät, 2021. DOI: http://dx.doi.org/10.18452/23442.

[11] J. M. S. F. Ducastelle and D. Gratias, "Generalized cluster description of multicomponent systems," *Physica A*, vol. 128, no. 1–2, pp. 334–350, 1984. DOI: https://doi.org/10.1016/0378-4371(84)90096-7.

[12] A. van de Walle and M. Asta, "Self-driven lattice-model monte carlo simulations of alloy thermodynamic properties and phase diagrams," *Modelling and Simulation in Materials Science and Engineering*, vol. 10, no. 5, p. 521, Jul. 2002. DOI: 10.1088/0965-0393/10/5/304. [Online]. Available: https://dx.doi.org/10.1088/0965-0393/10/5/304.

[13] M. Asta, D. de Fontaine, M. van Schilfgaarde, M. Sluiter, and M. Methfessel, "First-principles phase-stability study of fcc alloys in the ti-al system," *Phys. Rev. B*, vol. 46, pp. 5055–5072, 9 Nov. 1992. DOI: 10.1103/PhysRevB.46.5055. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevB.46.5055.

[14] F. Zhou, T. Maxisch, and G. Ceder, "Configurational electronic entropy and the phase diagram of mixed-valence oxides: The case of li$_x$fepo$_4$," *Phys. Rev. Lett.*, vol. 97, p. 155 704, 15 Oct. 2006. DOI: 10.1103/PhysRevLett.97.155704. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevLett.97.155704.

[15] R. Malik, F. Zhou, and G. Ceder, "Phase diagram and electrochemical properties of mixed olivines from first-principles calculations," *Phys. Rev. B*, vol. 79, p. 214 201, 21 Jun. 2009. DOI: 10.1103/PhysRevB.79.214201. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevB.79.214201.

[16] M. Borg, C. Stampfl, A. Mikkelsen, J. Gustafson, E. Lundgren, M. Scheffler, and J. N. Andersen, "Density of configurational states from first-principles calculations: The phase diagram of al-na surface alloys," *ChemPhysChem*, vol. 6, no. 9, pp. 1923–1928, 2005. DOI: https://doi.org/10.1002/cphc.200400612. [Online]. Available: https://chemistry-europe.onlinelibrary.wiley.com/doi/abs/10.1002/cphc.200400612.

[17] Z.-K. Han, D. Sarker, M. Troppenz, S. Rigamonti, C. Draxl, W. A. Saidi, and S. V. Levchenko, "First-principles study of Pd-alloyed Cu(111) surface in hydrogen atmosphere at realistic temperatures," *Journal of Applied Physics*, vol. 128, no. 14, p. 145 302, Oct. 2020, ISSN: 0021-8979. DOI: 10.1063/5.0020625. [Online]. Available: https://doi.org/10.1063/5.0020625.

[18] J. M. Sanchez, "Cluster expansion and the configurational theory of alloys," *Phys. Rev. B*, vol. 81, p. 224 202, 22 Jun. 2010. DOI: 10.1103/PhysRevB.81.224202. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevB.81.224202.

[19] T. Mueller, "Comment on "cluster expansion and the configurational theory of alloys"," *Phys. Rev. B*, vol. 95, p. 216 201, 21 Jun. 2017. DOI: 10.1103/PhysRevB.95.216201. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevB.95.216201.

[20] J. M. Sanchez, "Reply to "comment on 'cluster expansion and the configurational theory of alloys"'," *Phys. Rev. B*, vol. 95, p. 216 202, 21 Jun. 2017. DOI: 10.1103/PhysRevB.95.216202. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevB.95.216202.

[21] P. Hohenberg and W. Kohn, "Inhomogeneous electron gas," *Phys. Rev.*, vol. 136, B864–B871, 3B Nov. 1964. DOI: 10.1103/PhysRev.136.B864. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRev.136.B864.

[22] W. Kohn and L. J. Sham, "Self-consistent equations including exchange and correlation effects," *Phys. Rev.*, vol. 140, A1133–A1138, 4A Nov. 1965. DOI: 10.1103/PhysRev.140.A1133. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRev.140.A1133.

[23] J. P. Perdew, K. Burke, and M. Ernzerhof, "Generalized gradient approximation made simple," *Phys. Rev. Lett.*, vol. 77, pp. 3865–3868, 18 Oct. 1996. DOI: 10.1103/PhysRevLett.77.3865. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevLett.77.3865.

[24] J. P. Perdew, A. Ruzsinszky, G. I. Csonka, O. A. Vydrov, G. E. Scuseria, L. A. Constantin, X. Zhou, and K. Burke, "Restoring the density-gradient expansion for exchange in solids and surfaces," vol. 100, no. 13, Apr. 2008. DOI: 10.1103/physrevlett.100.136406. [Online]. Available: https://doi.org/10.1103%2Fphysrevlett.100.136406.

[25] R. Kikuchi, "A theory of cooperative phenomena," *Phys. Rev.*, vol. 81, pp. 988–1003, 6 Mar. 1951. DOI: 10.1103/PhysRev.81.988. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRev.81.988.

[26] A. van de Walle, "Multicomponent multisublattice alloys, nonconfigurational entropy and other additions to the alloy theoretic automated toolkit," *Calphad*, vol. 33, no. 2, pp. 266–278, 2009, Tools for Computational Thermodynamics, ISSN: 0364-5916. DOI: https://doi.org/10.1016/j.calphad.2008.12.005. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0364591608001314.

[27] N. A. Zarkevich and D. D. Johnson, "Reliable first-principles alloy thermodynamics via truncated cluster expansions," *Phys. Rev. Lett.*, vol. 92, p. 255 702, 25 Jun. 2004. DOI: 10.1103/PhysRevLett.92.255702. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevLett.92.255702.

[28] T. Mueller and G. Ceder, "Exact expressions for structure selection in cluster expansions," *Phys. Rev. B*, vol. 82, p. 184 107, 18 Nov. 2010. DOI: 10.1103/PhysRevB.82.184107. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevB.82.184107.

[29] J. M. Sanchez, "Cluster expansions and the configurational energy of alloys," *Phys. Rev. B*, vol. 48, pp. 14 013–14 015, 18 Nov. 1993. DOI: 10.1103/PhysRevB.48.14013. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevB.48.14013.

[30] C. Wolverton and D. de Fontaine, "Cluster expansions of alloy energetics in ternary intermetallics," *Phys. Rev. B*, vol. 49, pp. 8627–8642, 13 Apr. 1994. DOI: 10.1103/PhysRevB.49.8627. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevB.49.8627.

[31] L. Barroso-Luque, J. H. Yang, and G. Ceder, "Sparse expansions of multicomponent oxide configuration energy using coherency and redundancy," *Phys. Rev. B*, vol. 104, p. 224 203, 22 Dec. 2021. DOI: 10.1103/PhysRevB.104.224203. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevB.104.224203.

[32] A. Berera and D. de Fontaine, "Oxygen-vacancy phase equilibria in $yba_2cu_3o_z$ calculated by the cluster variation method," *Phys. Rev. B*, vol. 39, pp. 6727–6736, 10 Apr. 1989. DOI: 10.1103/PhysRevB.39.6727. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevB.39.6727.

[33] T. Mueller and G. Ceder, "Ab initio study of the low-temperature phases of lithium imide," *Phys. Rev. B*, vol. 82, p. 174 307, 17 Nov. 2010. DOI: 10.1103/PhysRevB.82.174307. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevB.82.174307.

[34] Q. Wu, B. He, T. Song, J. Gao, and S. Shi, "Cluster expansion method and its application in computational materials science," *Computational Materials Science*, vol. 125, pp. 243–254, 2016, ISSN: 0927-0256. DOI: https://doi.org/10.1016/j.commatsci.2016.08.034. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0927025616304049.

[35] A. Gonis, P. P. Singh, P. E. A. Turchi, and X.-G. Zhang, "Use of the ising model in the study of substitutional alloys," *Phys. Rev. B*, vol. 51, pp. 2122–2131, 4 Jan. 1995. DOI: 10.1103/PhysRevB.51.2122. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevB.51.2122.

[36] L. J. Nelson, G. L. W. Hart, F. Zhou, and V. Ozoli ņ š, "Compressive sensing as a paradigm for building physics models," *Phys. Rev. B*, vol. 87, p. 035 125, 3 Jan. 2013. DOI: 10.1103/PhysRevB.87.035125. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevB.87.035125.

[37] A. E. Hoerl and R. W. Kennard, "Ridge regression: Biased estimation for nonorthogonal problems," *Technometrics*, vol. 12, no. 1, pp. 55–67, 1970. DOI: 10.1080/00401706.1970.10488634. [Online]. Available: https://www.tandfonline.com/doi/abs/10.1080/00401706.1970.10488634.

[38] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996. DOI: https://doi.org/10.1111/j.2517-6161.1996.tb02080.x. [Online]. Available: https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.2517-6161.1996.tb02080.x.

[39] D. Donoho, "Compressed sensing," *IEEE Transactions on Information Theory*, vol. 52, no. 4, pp. 1289–1306, 2006. DOI: 10.1109/TIT.2006.871582.

[40] E. J. Candès, J. K. Romberg, and T. Tao, "Stable signal recovery from incomplete and inaccurate measurements," *Communications on Pure and Applied Mathematics*, vol. 59, no. 8, pp. 1207–1223, 2006. DOI: https://doi.org/10.1002/cpa.20124. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/cpa.20124.

[41] Y. Pati, R. Rezaiifar, and P. Krishnaprasad, "Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition," 40–44 vol.1, 1993. DOI: 10.1109/ACSSC.1993.342465.

[42] T. H. R. T. J.Friedman, *Elements of Statistical Leanrning*, 2nd ed. Springer, New York, p. 140, ISBN: 978-0-387-84857-0.

[43] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995. DOI: 10.1007/BF00994018. [Online]. Available: https://doi.org/10.1007/BF00994018.

[44] M. Aizerman, E. Braverman, and L. Rozonoer, "Theoretical foundations of the potential function method in pattern recognition learning," *Automation and remote control*, vol. 25, no. 6, pp. 821–837, 1964.

[45] J. S. Kasper, P. Hagenmuller, M. Pouchard, and C. Cros, "Clathrate structure of silicon na_8si46 and na_xsi136 ($x < 11$)," *Science*, vol. 150, no. 3704, pp. 1713–1714, 1965. DOI: 10.1126/science.150.3704.1713. [Online]. Available: https://www.science.org/doi/abs/10.1126/science.150.3704.1713.

[46] T. Tadano, Y. Gohda, and S. Tsuneyuki, "Impact of rattlers on thermal conductivity of a thermoelectric clathrate: A first-principles study," *Phys. Rev. Lett.*, vol. 114, p. 095 501, 9 Mar. 2015. DOI: 10.1103/PhysRevLett.114.095501. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevLett.114.095501.

[47] D. Connétable, "First-principles calculations of carbon clathrates: Comparison to silicon and germanium clathrates," *Phys. Rev. B*, vol. 82, p. 075 209, 7 Aug. 2010. DOI: 10.1103/PhysRevB.82.075209. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevB.82.075209.

[48] B. Eisenmann, H. Schäfer, and R. Zagler, "Die verbindungen aii8biii16biv30 (aii ≡ sr, ba; biii ≡ al, ga; biv ≡ si, ge, sn) und ihre käfigstrukturen," *Journal of the Less Common Metals*, vol. 118, no. 1, pp. 43–55, 1986, ISSN: 0022-5088. DOI: https://doi.org/10.1016/0022-5088(86)90609-0. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0022508886906090.

[49] M. Christensen, S. Johnsen, and B. B. Iversen, "Thermoelectric clathrates of type i," *Dalton Trans.*, vol. 39, pp. 978–992, 4 2010. DOI: 10.1039/B916400F. [Online]. Available: http://dx.doi.org/10.1039/B916400F.

[50] S.-J. Kim, S. Hu, C. Uher, T. Hogan, B. Huang, J. D. Corbett, and M. G. Kanatzidis, "Structure and thermoelectric properties of ba6ge25-x, ba6ge23sn2, and ba6ge22in3: Zintl phases with a chiral clathrate structure," *Journal of Solid State Chemistry*, vol. 153, no. 2, pp. 321–329, 2000, ISSN: 0022-4596. DOI: https://doi.org/10.1006/jssc.2000.8777. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0022459600987772.

[51] A. Shevelkov and K. Kovnir, "Zintl clathrates," in *Zintl phases: Principles and recent developments*, T. Fässler, Ed., vol. 139, Springer, 2011, pp. 97–142. DOI: 10.1007/978-3-642-21150-8.

[52] N. P. Blake, S. Latturner, J. D. Bryan, G. D. Stucky, and H. Metiu, "Band structures and thermoelectric properties of the clathrates ba8ga16ge30, sr8ga16ge30, ba8ga16si30, and ba8in16sn30," *J. Chem. Phys.*, vol. 115, no. 18, pp. 8060–8073, 2001. DOI: https://doi.org/10.1063/1.1397324.

[53] N. P. Blake, D. Bryan, S. Latturner, L. Møllnitz, G. D. Stucky, and H. Metiu, "Structure and stability of the clathrates ba8ga16ge30, sr8ga16ge30, ba8ga16si30, and ba8in16sn30," *J. Chem. Phys.*, vol. 114, no. 22, pp. 10 063–10 074, 2001. DOI: https://doi.org/10.1063/1.1374123.

[54] A. Saramat, G. Svensson, A. E. C. Palmqvist, C. Stiewe, E. Mueller, D. Platzek, S. G. K. Williams, D. M. Rowe, J. D. Bryan, and G. D. Stucky, "Large thermoelectric figure of merit at high temperature in czochralski-grown clathrate," *Journal of Applied Physics*, vol. 99, no. 2, p. 023 708, 2006. DOI: https://doi.org/10.1063/1.2163979.

[55] E. S. Toberer, M. Christensen, B. B. Iversen, and G. J. Snyder, "High temperature thermoelectric efficiency in $ba_8ga16ge30$," *Phys. Rev. B*, vol. 77, p. 075 203, 7 Feb. 2008. DOI: 10.1103/PhysRevB.77.075203. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevB.77.075203.

[56] A. Gulans, S. Kontur, C. Meisenbichler, D. Nabok, P. Pavone, S. Rigamonti, S. Sagmeister, U. Werner, and C. Draxl, "Exciting: A full-potential all-electron package implementing density-functional theory and many-body perturbation theory," *Journal of Physics: Condensed Matter*, vol. 26, no. 36, p. 363 202, Aug. 2014. DOI: 10.1088/0953-8984/26/36/363202. [Online]. Available: https://dx.doi.org/10.1088/0953-8984/26/36/363202.

[57] The slight difference with the optimal model of the clusters pool study (Fig. 4.4.) for P1 in the nonlinear CE, which has one more cluster, is due to the fact that the regularization parameter was re-optimized for this final model.

[58] D. P. Landau and K. Binder, *A Guide to Monte Carlo Simulations in Statistical Physics*, 4th ed. Cambridge University Press, 2014.

[59] S. Rigamonti, M. Troppenz, M. Kuban, A. Hübner, and C. Draxl, "Cell: A python package for cluster expansion with a focus on complex alloys," In preparation, 2023.

[60] `CELL` documentation: https://sol.physik.hu-berlin.de/cell.

[61] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[62] L. Foppa, T. A. R. Purcell, S. V. Levchenko, M. Scheffler, and L. M. Ghiringhelli, "Hierarchical symbolic regression for identifying key physical parameters correlated with bulk properties of perovskites," *Phys. Rev. Lett.*, vol. 129, p. 055 301, 5 Jul. 2022. DOI: 10.1103/PhysRevLett.129.055301. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevLett.129.055301.