Efficient construction and applications of higher-order force constant models

Erik Fransson, Fredrik Eriksson, and Paul Erhart
Chalmers University of Technology, Department of Physics, Gothenburg, Sweden

The vibrational properties of solids are fundamental to a large number of physical phenomena, including phase stability and thermal conduction. The canonical approach to modeling these properties requires knowledge of the interatomic forces constants (FCs). The problem of extracting the parameters in the FC expansion from a set of reference forces can be cast in linear form making it amenable to linear regression techniques. Here, we consider the efficiency of various common regression methods for FC extraction and the efficacy of the resulting models for predicting various thermodynamic properties. The regression approach drastically reduces the required number of reference calculations, which constitute the computationally most demanding task in FC extraction, compared to explicit enumeration techniques. It thereby becomes possible to extract both harmonic and high-order anharmonic FCs for large systems with low symmetry, including defects and surfaces. It is shown that ordinary least-squares, especially in connection with feature elimination, often yields the best performance in terms of convergence with respect to training set size and sparsity of the solution. Regression based on the least absolute shrinkage and selection operator (LASSO) on the other hand, while useful in some cases, tends to yield a larger number of features, with a noise level that has a detrimental effect on the prediction of e.g., the thermal conductivity. Finally, we also consider methods for the prediction of the temperature dependence of vibrational spectra from high-order FC expansions via molecular dynamics simulations as well as self-consistent phonons.

I. INTRODUCTION

The vibrational degrees of freedom (DOFs) of materials are crucial for numerous thermodynamic properties, including phase stability and thermal conduction. In order to model these properties one requires an efficient representation of the potential energy surface (PES). Electronic structure calculations, often based on density functional theory (DFT), can in principle provide the necessary input, yet they are computationally rather demanding, which limits their application in particular in cases with strong anharmonicity and/or low symmetry. Here, effective models parametrized using, say DFT calculations, play a crucial role.

In crystals the vibrational atomic motion can be conveniently described in terms of phonons, quasi-particles that represent periodic and quantized excitations. Phonon theory is commonly formulated by starting from a Taylor expansion of the total energy, in which the expansion coefficients are referred to as force constants (FCs). Depending on material and property of interest the FC expansion must be carried out to different orders. Generally, it is preferable to keep the order as low as possible since the number of independent coefficients quickly increases with expansion order, lowering of the symmetry, and the size of the system of interest. For ideal materials with comparably small unit cells the FC up to third order can still be obtained by enumerating displacements and evaluating the derivatives numerically. This direct enumeration scheme becomes, however, tedious or impractical for larger systems (e.g., point defects, interfaces, or nanoparticles) and/or materials that require expansions beyond third order (e.g., metastable phases of transition metals or oxides). As an alternative, linear regression techniques have been proposed including singular value decomposition (SVD) and compressive sensing using either least absolute shrinkage and selection operator (LASSO) or the split-Bregman algorithm. There are various other linear regression techniques and feature selection algorithms that could be suitable for force constant regression. Further analysis of these techniques with regard to their efficiency, accuracy and reliability for constructing FC expansions is therefore in order.

We have recently introduced the HPHIVE package, which is ideally suited for this purpose. Thanks to being implemented in Python it can be readily interfaced with machine learning libraries such as scikit-learn that in turn provide efficient implementations of various optimization techniques. In this paper we present a comparison of linear regression methods and the direct enumeration approach for the extraction of FC of different order, including second-order FCs for large systems of low symmetry such as defects, third-order FCs for the prediction of the thermal conductivity, as well as higher-order FCs for bulk and surface (see Supplementary Information) systems. This approach enables us to determine the applicability of regression methods in different regimes. We also demonstrate the application of these FC models for studying anharmonic effects both in the framework of Boltzmann transport theory and molecular dynamics (MD) simulations. The following section provides a concise summary of the underlying theory, while sections thereafter present the different application examples named above.

II. FORCE CONSTANT EXTRACTION

The PES can be expanded in a Taylor series in the atomic displacements \( \mathbf{u} \) relative to a set of reference po-
where $\Phi$ are the force constants, Latin indices enumerate the atoms, Greek indices enumerate the Cartesian coordinates, and the Einstein summation convention is implied.

The number of components in the force constants scales as $O(N^n)$, where $N$ is the number of atoms and $n$ is the expansion order. There are, however, multiple constraints that reduce the number of free parameters, such as lattice symmetries and sum rules. Yet in the case of large systems, low symmetry, and/or higher expansion orders the number of unknowns is still very large.

### A. Direct approach

The conventional way of extracting FCs relies on the systematic evaluation of numerical derivatives. For example, for the second-order terms

$$V_i = V_0 + \Phi_{ij} u_i^\alpha u_j^\beta + \frac{1}{2} \Phi_{ijk} u_i^\alpha u_j^\beta u_k^\gamma + \ldots,$$

where $\Phi$ are the force constants, Latin indices enumerate the atoms, Greek indices enumerate the Cartesian coordinates, and the Einstein summation convention is implied.

The number of free parameters, i.e. the dimension of $\mathbf{x}$, can still be very large even for systems with high symmetry. The FC expansion is therefore often truncated. Firstly, as with most Taylor expansions, only few orders are usually needed to obtain an accurate representation of the PES in the range of relevant displacements. Secondly, the atomic interactions often decay rather quickly with interatomic distance, meaning a cutoff can be imposed. Thirdly, pair interactions are often stronger than three-body interactions, which in turn are often stronger than four-body interactions etc., i.e.

$$\Phi_{ijj} > \Phi_{ijk} > \Phi_{ijkl}.$$
the LASSO method often adopted in compressive sampling.

To evaluate the performance of a model obtained by solving eq. (1) one can employ cross-validation (CV). To this end, the available reference data set is split into training and validation (or test) sets. After the former has been used for fitting the parameter vector, one can evaluate the root-mean-square errors (RMSEs)

\[
\text{RMSE} = \sqrt{\frac{1}{N} \sum_i (F_{\text{model}} - F_{\text{target}})^2},
\]

where the summation extends over the \(N\) force components in the validation set. To reduce the statistical error, the RMSE is then averaged over several different splits of the reference data, yielding the CV score.

Efficient and generally applicable implementations of these methods are available in HIPHIVE thanks to the Python machine learning library SCIKIT-LEARN\(^{27}\).

4. Feature selection

In machine learning feature selection refers to the task of isolating the most important parameters (or features) during model construction. Reducing the number of parameters yields a less complex model, which in turn often leads to less overfitting and improved transferability. It can also reduce the computational cost of both training and sampling the model. Feature selection is especially interesting for FC models, for which only some interaction terms may be strong (Sect. II B 2).

Several feature selection methods are available for linear problems. One can employ for example a simple pruning condition based on the magnitude of the parameters. This is particularly effective in combination with regression techniques that include regularization, typically via the \(\ell_1\) or \(\ell_2\)-norm of the parameter vector (see eq. 2). One can also employ matching pursuit algorithms such as orthogonal matching pursuit (OMP), which allows one to impose a constraint on the number of non-zero coefficients in the solution vector, or techniques such as recursive feature elimination (RFE). In the latter case, a solution is determined using a fit method of choice, the weakest or least important parameters are removed, and the procedure is iteratively repeated until the target number of features is reached. In some cases, the optimal sparsity of a model can be determined by combining the above techniques with Bayesian regression.\(^{28}\)

It must be noted that the different methods can differ considerably with respect computational effort as well as memory requirements. SVD is the least demanding procedure in both regards. It is difficult to provide general guidelines with regard to the demands of different methods, since the effort can differ dramatically with the choice of hyperparameters and the conditioning of the sensing matrix. As a rough guideline, RFE with ordinary least squares (OLS) typically requires about 100 to 10,000 OLS fits depending on how accurately one wishes to perform the feature elimination. LASSO is comparable to RFE based on OLS with respect to computational effort.

III. SECOND-ORDER FORCE CONSTANTS: LARGE LOW-SYMMETRY SYSTEMS

A. Background and reference calculations

The second-order FCs of systems with many atoms and/or low symmetry can be tedious to obtain by the direct approach (Sect. II A). This applies in particular to the FCs of defect configurations, which are needed for example for computing the vibrational contribution to the free energy of defect formation\(^{12}\), analyzing the impact of defects on the thermal conductivity\(^{17,29}\) or predicting the vibrational broadening of optical spectra.\(^{30}\)

In this section, we therefore analyze the extraction of second-order FCs for the vacancy in body-centered cubic (BCC) Ta as a prototypical case, using both the direct approach (Sect. II A) as implemented in PHONOPY\(^{17}\) and the regression approach (Sect. II B) as implemented in HIPHIVE.\(^{24}\)

Calculations were carried out for supercells comprising \(N \times N \times N\) conventional BCC cells with \(N \in [4, \ldots, 10]\) that contain a single vacancy. Reference forces were computed using the embedded atom method (EAM) potential model TA1 from Ref. 31. Using an empirical potential rather than DFT calculations in this example allows us to compute reference second-order FCs for very large configurations. For the PHONOPY calculations we used a displacement amplitude of 0.01 Å while for the HIPHIVE calculations, we generated 30 structures for each \(N\) by drawing random displacements from a normal distribution with an standard deviation of 0.01 Å.

B. Scaling of regression methods

Several different methods were considered for constructing FC models for the Ta vacancy models, including SVD, RFE with OLS, LASSO, and automatic relevance detection regression (ARDR). SVD is by far the computationally least expensive method and exhibits a favorable scaling with training set size.\(^{32}\) RFE with OLS and LASSO exhibit a very similar scaling as SVD but are about 100 to 500 times more expensive. This is unsurprising as these methods carry out multiple OLS optimizations as part of their algorithms. Finally, ARDR, which we employed in earlier work to construct high accuracy FCs,\(^{10}\) exhibits a very unfavorable scaling with training set size (including both computational effort and memory requirements), which prevents its effective application for large sensing matrices including the present case.
Taking into account CV the computational effort required for the largest supercells considered here becomes notable for LASSO and RFE with OLS. This becomes apparent if one considers that both the size of the parameter space (see Supplementary Information) and the number of available force components increase approximately linearly with system size, giving a quadratic increase of the sensing matrix with the number of atoms. In the remainder of this section, we therefore primarily consider SVD, which will be demonstrated to work very well if combined with cutoff selection to avoid overfitting.

### C. Cutoff selection via cross-validation

The number of DOFs in a FC model grows rapidly as the cutoff increases. At the same time, as discussed in Sect. II B 2, interaction strength and hence the FCs decay with increasing interatomic distance. At some point an increase in cutoff will therefore lead to negligible improvement in accuracy but merely an increase in model complexity. Specifically in the absence of regularization terms in the objective function (α = β = 0 in eq. (2)), one can therefore observe a deterioration of model quality with the inclusion of more terms in the expansion due to overfitting. In this case one should therefore evaluate the performance of models with different cutoffs.

We employed CV using the shuffle-and-split method with 5 splits and 15 training structures for a system size of $N = 6$ and constructed a series of second-order FC models for a series of cutoffs [Fig. 2]. While the RMSE over the training set continues to decrease for larger parameter space, the CV-RMSE has a minimum around 6 to 7 Å. For standard SVD the validation score increases for larger cutoffs due to overfitting. This behavior can be counteracted by using RFE, which yields a slight improvement of the CV-RMSE. As discussed above RFE is, however, computationally substantially more expensive whence SVD with a judicious choice of cutoffs is a preferable choice. All subsequent analysis was therefore carried out using SVD and a second-order cutoff of 6 Å.

### D. Convergence of thermodynamic properties

In order to evaluate how accurately the regression approach reproduces the correct second-order force constants, we considered three different measures. First, we evaluated the absolute error of the zone-center (Γ) frequencies obtained by regression relative to the direct approach (Fig. 3a).

$$\Delta \omega = \sqrt{\frac{1}{3N} \sum_{i} (\omega_{\text{regression}} - \omega_{\text{direct}})^2},$$

(3)

where $\omega_i$ is frequency of mode $i$. Secondly, we considered the absolute difference in the harmonic free energy at 2250 K corresponding to 75% of the calculated melting temperature (Fig. 3b),

$$\Delta F = |F_{\text{regression}}^{\text{vib}} - F_{\text{direct}}^{\text{vib}}|.$$

(4)

Here, the free energies were computed within the harmonic approximation using PHONOPY.

Lastly, in order to obtain a computationally cheaper measure, we computed the relative error of the second-order FC matrices (Fig. 3c), defined as follows

$$\Delta \Phi = \|\Phi_{\text{regression}} - \Phi_{\text{direct}}\|/\|\Phi_{\text{regression}}\|.$$

(5)
FIG. 3. Evaluation of second-order FC models for vacancy in BCC Ta ($N = 6$). Convergence of (a) the zone-center frequencies according to eq. (3), (b) the free energy at 75% of the calculated melting temperature (2250 K) according to eq. (4), and (c) the elements of the second-order FC matrix according to eq. (5). All calculations were carried out using a second-order cutoff radius of 6 Å (compare Fig. 2).

The frequencies and free energies exhibit very similar convergence behavior. Both quantities reach convergence at about 3,000 force components, equivalent to two to three configurations, corresponding to the number of parameters in the model. Comparison with the measure based on the FC matrix itself, eq. (5), suggests that $\Delta \Phi \lesssim 5\%$ is sufficient to achieve convergence of the frequency spectrum and the free energy. Considering the convergence of $\Delta \Phi$ itself suggests a more conservative threshold of 2%.

The comparison includes both models with only second-order FCs terms and models with additional very short-ranged third-order FC terms using a cutoff of 3.0 Å. The latter perform consistently better than the “pure” second-order models. The inclusion of only a few third-order terms thus stabilizes the extraction of the second-order FCs, an observation that we have also made in various other applications. These terms enable to account for anharmonicity in the vicinity of the reference positions that would otherwise be effectively included in the second-order FCs. This principle can also be applied to higher-order terms, where we have found that adding a few terms of the respective next-highest order yields more accurate FCs.

E. Scaling with system size

Following the analysis in the previous section, computing $\Delta \Phi$ as a function of the training set size allows one to determine the number of training structures needed for recovering the second-order FCs at the accuracy level of the direct approach (Fig. 4). Using the more conservative threshold of $\Delta \Phi < 2\%$ (Fig. 3), we thereby determined the necessary number of supercell calculations as a function of system size (Fig. 4).

While in the direct approach the number of necessary calculations increases steeply with system size, in the regression approach it decreases with system size. This is possible partly due to the introduction of a cutoff but more importantly since the regression takes full advantage of the increase in information content with supercell size. In this context, it is important that all atoms are displaced at least by some small amount. By contrast, the information content in configurations employed in the direct approach decreases substantially with system size as only one atom is displaced at a time. One can anticipate this scaling effect to be even more pronounced for third or higher-order FCs due to the exponential increase of the number of parameters with order.
IV. THIRD-ORDER FORCE CONSTANTS: THERMAL CONDUCTIVITY

A. Background and reference calculations

Calculation of the thermal conductivity within the framework of Boltzmann transport theory requires knowledge of the second and third-order FCs, providing a sensitive test for the extraction of higher-order FCs. In this example we therefore analyze different methods for obtaining these FCs and the resulting thermal conductivity in silicon.

The first method considered here is SVD, which has already been used for the same purpose in Ref. 9. Since it requires an overdetermined problem in order to yield sensible results one must balance the number of DOFs in the FC model by tuning the cutoff radii according to the training set size (see Supplementary Information).

Secondly, we constructed a very large parameter space using maximum cutoffs for both the second and third-order terms and applied feature selection techniques to recover the correct solution. To this end, we considered the LASSO method, for which the sparsity of the solution, and hence the number of features, is controlled via the $\alpha$-parameter in eq. (2) with $\beta = 0$, and RFE in combination with OLS, for which the number of features $n_f$ is a direct input parameter of the algorithm. In both cases, optimal values for $\alpha$ and $n_f$ were determined by CV for each training set size. We note that LASSO was also employed recently to extract high-order FCs in clathrates.

Reference second and third-order FCs were calculated for 250-atom supercells ($5 \times 5 \times 5$ primitive unit cells) via the direct approach using PHONOPY and PHONOPY, respectively. No cutoff was imposed during the calculation of the third-order FCs, which therefore required 801 individual force calculations.

For the regression approach, we generated a total of 20 reference structures based on 250-atom supercells ($5 \times 5 \times 5$ primitive unit cells) with displacements drawn from a normal distribution yielding displacement amplitudes of about 0.03 Å.

Reference forces were obtained from DFT calculations using the projector augmented wave method as implemented in VASP and an exchange-correlation functional, based on the generalized gradient approximation. The Brillouin zone was sampled using only the $\Gamma$-point. The plane-wave energy cutoff was set to 245 eV, an additional support grid for fast-Fourier transformations was used during the force calculation, and the reciprocal projection scheme was employed throughout.

B. Evaluation of force constants

The optimal cutoffs when using SVD were found to be 9.65 Å (12 neighbor shells), 5.0 Å (3 shells), and 2.5 Å (1 shell) for second, third, and fourth-order terms, respectively, which are very similar to the cutoffs used in Ref. 22. (As described above, inclusion of a few terms of the next highest order, here the fourth, improves the accuracy of the extracted FCs.) This choice yields a total of 135 parameters. Further details concerning the cutoff convergence can be found in the Supplementary Information. For the feature selection study (LASSO and RFE with OLS), we used a cutoff of 9.65 Å for second and

FIG. 5. (a,b) Learning curves for second and third-order FCs in silicon showing (a) the RMSE calculated by CV and (b) the number of features (non-zero parameters) as a function of the training set size. The shaded regions indicate the standard deviation calculated by CV. (c) RMSE computed using CV vs the number of features obtained using seven training structures.
third-order terms and a cutoff of 4.0 Å for fourth-order terms, which yields 2611 parameters.

For each approach the RMSE was calculated by CV using ten splits (Fig. 5a). SVD with short cutoffs yields a well converged RMSE using only five structures, which is well inside the overdetermined region (about 3,750 force components vs 135 parameters). This is far smaller than the 95 calculations required by PHONO3PY when imposing the same third-order cutoff.

The RFE approach based on OLS fails in the underdetermined region, i.e. for fewer than about 4 structures (or 3,000 force components), since OLS includes no regularization terms. With five or more structures it, however, comfortably reaches the level of the SVD approach and does so using fewer features (Fig. 5b). Finally, LASSO consistently performs worse than either SVD or RFE with OLS while yielding the largest number of features (non-zero parameters) in the model.

To illustrate the difference in performance between LASSO and RFE further we explicitly computed the RMSE as a function of their respective pruning parameter using seven training structures (Fig. 5c). This shows that while both methods yield qualitatively similar curves and achieve comparable RMSE scores, LASSO yields a larger number of features at its respective RMSE minima, a behavior that we have also observed for other materials as well as similar linear models such as alloy cluster expansions.

The larger number of false positives obtained by LASSO can also be visualized by direct inspection of the FCs (Fig. 6). All three techniques yield practically identical second-order FCs. SVD with short cutoffs and RFE with OLS consistently find sparser solutions for the third-order FCs than LASSO, which produces a more noisy result with a substantial number of numerically small but non-zero FC terms. As will be shown next, this level of noise, albeit small, has a decisive impact on the thermal conductivity.

C. Thermal conductivity

While the FCs themselves cannot be directly accessed experimentally, measurable properties such as the phonon lifetimes and the thermal conductivity are directly derived from the third-order FCs. We therefore computed the (lattice) thermal conductivity within the framework of Boltzmann transport theory within the relaxation time approximation as implemented in PHONO3PY considering only phonon-phonon scattering as a possible scattering channel (i.e. ignoring e.g., isotope or boundary scattering). A Γ-centered 32 × 32 × 32 q-point mesh was employed for Brillouin zone integration.

The trends of the FC analysis largely translate to the thermal conductivity. SVD with short cutoffs and RFE with OLS both perform very well and reproduce the thermal conductivity obtained with the FCs generated by
the direct approach with as few as five to ten structures. LASSO on the other hand, fails to reproduce the reference data even when using as many as 15 structures for training. Here, we have also included SVD with long cutoffs to illustrate the danger of overfitting intrinsic to this approach (also compare Fig. 2).

In general, the performance of the different regression methods changes with the hyper-parameters ($\alpha$ approach (also compare Fig. 2).

offs to illustrate the danger of overfitting intrinsic to this training. Here, we have also included SVD with long cutoff selection and RFE based on OLS emerge as the most reliable approaches among the techniques considered here.

For the efficient extraction of FCs, in particular in high-throughput schemes, one must, however, not only consider the computational effort but also the amount of human intervention required. It is therefore desirable to set up protocols in the future that carry out the above analysis automatically and the implementation provided by HIPHIVE is very well suited for this purpose. To this end, one can also combine of course cutoff selection with RFE and employ other regression methods.

V. FOURTH-ORDER FORCE CONSTANTS: TEMPERATURE-DEPENDENT PHONONS

A. Background and reference calculations

While perturbation theory formally provides an expression for the temperature induced phonon frequency shifts caused by the third-order FC terms, one commonly carries the expansion at least to the next higher even order, when analyzing frequency shifts in fact, this is inevitable when dealing with metastable materials. In this context, we consider the inorganic clathrate Ba$_8$Ga$_{16}$Ge$_{30}$, for which the motion of some of the Ba atoms is strongly anharmonic.

Clathrates are inclusion compounds with a defined lattice structure that can trap atomic or small molecular species. Ba$_8$Ga$_{16}$Ge$_{30}$ belongs to the class of type-I clathrates with spacegroup Pm3n. In this case, the host lattice is made up of Ga and Ge atoms, which occupy Wyckoff sites 6c, 16i, and 24k, whereas Ba atoms reside on sites 2a and 6d inside the cages. Due to the size mismatch between the guest species and the cages, which is particular large for the 6d sites, the Ba atoms experience a very wide and flat PES with pronounced anharmonicity. We have previously analyzed the ordering of the host species and extracted the ordered ground state structure of Ba$_8$Ga$_{16}$Ge$_{30}$ which provides us with a prototype structure for the analysis of the FCs.

B. Model construction

First, we generated 50 structures based on the primitive 54-atom unit cell with an average displacement amplitude of 0.28 Å using the Monte Carlo (MC) rattling approach described in Ref. 11. Reference forces were obtained from DFT calculations using the computational approach described in Ref. 2. The exchange-correlation potential was represented using the vdW-DF-cx method, which combines semi-local exchange with non-local correlation, as it has been previously shown to provide a sound description of the vibrational modes of this system. The Brillouin zone was sampled using a Γ-centered $3 \times 3 \times 3$ $k$-point mesh and the plane-wave energy cutoff was set to 243 eV.

These data were used to train an initial fourth-order model, which was subsequently sampled by MD simulations at 300 and 650 K for 10 ps. Given the scaling analysis presented above (Fig. 1), we resorted to using SVD for regression. We extracted 50 structures for each temperature and obtained the “true” atomic forces using DFT calculations.

This approach provided us with a total of 150 reference structures and thus a sensing matrix with more than 20,000 rows. Using these data, we constructed a number of models by SVD and evaluated their performance by CV in order to identify a suitable combination of orders and cutoffs (Table). The smallest CV-RMSE score is obtained for model 4 (Fig. 8k), which yields a value of 68 meV/Å to be compared with maximum force components of about 4000 meV/Å (Fig. 8b). The fourth-order cutoff for this model is 4.35 Å, which is enough to include all Barium-cage interaction indicating that the anharmonicity of all of these interactions is important. The PES for Ba atoms along different direction calculated using the model is in fact in excellent agreement with DFT calculations (Fig. 9, see Supplementary Information for the PES along (100) and (110)) It is apparent that anharmonicity is more pronounced for 6d sites than for 2as sites, as the cages surrounding the former are larger.

C. Temperature dependent density of states

The low thermal conductivity of inorganic clathrates is often attributed to the rattling motion of the guest species. These modes exhibit a notable temperature

| Model | Two-body cutoffs | Three-body cutoffs |
|-------|------------------|--------------------|
|       | 2rd | 3rd | 4th | 5th | 6th | 3rd | 4th |
| 1     | 5.4 | 3.0 | 3.0 |
| 2     | 5.4 | 3.5 | 3.5 |
| 3     | 5.4 | 4.0 | 4.0 |
| 4     | 5.4 | 4.35| 4.35|
| 5     | 5.4 | 4.5 | 4.5 |
| 6     | 5.4 | 4.35| 4.35| 3.0| 3.0 |
| 7     | 5.4 | 4.0 | 4.0 | 4.0 | 4.0 |
MD data was also employed to generate EHMs. To this end, a harmonic model with a cutoff of 5.4 Å was fitted by SVD using 30 snapshots from the trajectory.

SCPHs can be constructed in several different ways. Here, we begin with a second-order FC model that is generated using only a few rattled configurations. This harmonic model is used to generate displaced configurations by superposing normal modes with amplitudes according to the Maxwell distribution at the temperature of interest (see Ref. 4 for details). Forces are then computed using the fourth-order model described above and a new harmonic model is constructed by regression. This procedure is carried out iteratively in a self-consistency loop until convergence is reached.

Regardless of approach, the DOS shows a significant temperature dependence of the low-frequency Ba modes around 4 meV, in line with experimental work. On closer inspection, the SCPHs exhibit a slightly more pronounced temperature dependence than the DOS generated from EHMs. This very slight difference (about 0.2 meV for the onset of the lowermost feature at 100 K) can be understood by considering the weighting of configurations in either method. The selection of configurations in the EHM approach is based on the full anharmonic PES as they are drawn directly from the MD trajectory. By contrast, in the case of the SCPH approach the selection of structures effectively occurs by the amplitudes generated according to a harmonic landscape. This difference is also apparent for example in the mean-square displacements (MSDs) generated by full MD simulations in comparison with the (zero Kelvin) harmonic approximation. The SCPH can be in principle improved by reweighing configurations during the self-consistency procedure. We have, however, found it difficult to achieve convergence with this approach.

The DOS obtained via the VACF differs more notably from the ones from EHMs or SCPHs. This behavior can be attributed to size effects, as the 4×4×4 supercell used
FIG. 10. Temperature dependent phonon density of states for Ba$_8$Ga$_{16}$Ge$_{30}$ from (a) the velocity auto-correlation function, (b) effective harmonic models, and (c) self-consistent phonons.

in the MD simulations only allows direct sampling of phonon modes with momentum vectors that are a multiple of $\pi/2a$.

One can also extract the phonon dispersion itself from MD simulations based on the fourth-order model described above. To this end, we compute the spectral energy density (SED) for a cell elongated in the ⟨100⟩ direction (Fig. 11). The low-frequency optical modes shift upward in frequency, as expected from the temperature dependence of the DOS while the two-fold degenerate transverse acoustic mode stiffens slightly. Overall, the lifetimes of the modes decrease with temperature as evident from the broadening of the dispersion. Phonon frequencies and lifetimes could now be in principle extracted via peak fitting.

VI. DISCUSSION

Regression techniques enable the efficient extraction of FCs, in particular in the case of higher-order expansions and/or large systems with low symmetry. In this study, we have assessed the efficiency of a selection of regression methods in different application scenarios using the HIPHIVE package, which serves as a simple yet flexible and computationally efficient tool for this purpose. From this analysis it is apparent that regression techniques can reduce the computational effort for FC extraction, which primarily originates from reference calculations, compared to explicit enumeration techniques by one or more orders of magnitude.

Specifically, we considered second-order FCs and derived properties for large systems (Sect. III), third-order FCs and thermal conductivity (Sect. IV) as well as fourth and higher-order FCs and their sampling for strongly anharmonic systems (Sect. V and Supplementary Information).

A. Second-order FCs in large systems

In the case of second-order FCs in large systems the regression approach can reduce the computational effort by more than one order of magnitude compared to the direct approach (Fig. 4). In this regard, we have shown SVD to perform very well if combined with cutoff selection. SVD yields prediction errors that are on par with more advanced regression techniques such as RFE with OLS, LASSO, or ARDR. The latter are, however, at least one order of magnitude more demanding in terms of computer time, which becomes a concern for very large systems.

For SVD to work properly the linear system to be solved must be overdetermined. The configurations used for regression are obtained by rattling the atomic positions. As a result, the information density, i.e. the number of force components that are sizable, is high, which is usually not the case for systematically enumerated struc-
tures such as the ones used in the direct approach. One therefore does not require a large number of configurations in order to obtain a well-conditioned sensing matrix.

We have furthermore found that inclusion of a few higher-order FC terms (here third-order FCs with a short cutoff) accelerates convergence of the lower-order FCs of interest with respect to training set size. Here, the third-order terms allow extraction of the “true” second-order expansion terms, which otherwise would have to effectively account for small anharmonicities in the PES.

B. Third-order FCs and thermal conductivity

As in the case of the second-order FCs the regression approach drastically reduces the number of reference calculations needed to recover the parameters of the FC expansion. It is possible to use SVD with cutoff selection as shown previously. Alternatively automatic feature selection is an attractive option. Here, RFE with OLS has been found to work very reliably and efficient, yielding both fast convergence and a small number of features (sparse solutions). LASSO generally produces less sparse solutions and more noisy solutions, an observation that was also made for other linear models such as cluster expansions. This behavior has a detrimental impact when predicting the thermal conductivity using LASSO-trained models.

C. Higher-order FC models and anharmonic PESs

We also considered the construction and sampling of FC expansions beyond third-order, which are impractical with the direct approach.

We constructed fourth-order FC models with up to 11,000 parameters for the inorganic clathrate $Ba_4Ga_{16}Ge_{30}$ using SVD with cutoff selection. Using the final model, we then applied several different methods to obtain the temperature dependence of the vibrational spectrum, including sampling the VACF from MD simulations, EHMs, SCPHs, and SED. While the methods yield consistent trends that agree with simulations, EHMs, SCPHs, and SED (Fig. 10). As shown by the eighth-order model for a Ni surface (see Supplementary Information), FC expansions can be practically carried to even higher-order if needed. Thereby, it becomes possible to account for even more anharmonic behavior, including for example the temperature dependence of the surface layer spacing.

VII. CONCLUSIONS AND OUTLOOK

For the applications considered here, SVD and RFE based on OLS provides close to optimal performance with regard to convergence relative to training set size and sparsity of the solution. LASSO converges more slowly and yields a larger number of features than RFE with OLS, a behavior that we have also observed in the case of cluster expansions, the construction of which can be cast in a linear form in analogous fashion to FCs, ARDR, which has only been considered for the extraction of second-order FCs, can perform very well but scales poorly with the size of the sensing matrix. This often precludes its application to FC extraction as both the number of data points per structure and the number of parameters are large.

Regression techniques are in principle very attractive for high-throughput schemes, since they require much less computational effort than direct approaches. For the regression approach to be viable one must, however, not only consider the computational effort but also the amount of human intervention required. In the future, it is therefore desirable to set up protocols that automatically construct and validate FC models. The HIPHIVE package is very well suited for this purpose as it can be easily integrated via Python and readily interfaced with DFT codes via ASE.

ACKNOWLEDGMENTS

This work was funded by the Knut and Alice Wallenberg Foundation (KAW), the Swedish Research Council (VR), and the Swedish Foundation for Strategic Research (SSF). Computer time allocations by the SNIC at C3SE (Gothenburg), NSC (Linköping), and PDC (Stockholm) are gratefully acknowledged.
