Machine learning for a finite size correction in periodic coupled cluster theory calculations

Laura Weiler,[1] Tina N. Mihm,[1] and James J. Shepherd[1, a]
Department of Chemistry, University of Iowa

(Dated: March 30, 2022)

We introduce a straightforward Gaussian process regression (GPR) model for the transition structure factor of metal periodic coupled cluster singles and doubles (CCSD) calculations. This is inspired by the method introduced by Liao and Grüneis for interpolating over the transition structure factor to obtain a finite size correction for CCSD [J. Chem. Phys. 145, 141102 (2016)], and by our own prior work using the transition structure factor to efficiently converge CCSD for metals to the thermodynamic limit [Nat. Comput. Sci. 1, 801 (2021)]. In our CCSD-FS-GPR method to correct for finite size errors, we fit the structure factor to a 1D function in the momentum transfer, G. We then integrate over this function by projecting it onto a k-point mesh to obtain comparisons with extrapolated results. Results are shown for lithium, sodium, and the uniform electron gas.

I. INTRODUCTION

Calculations at the thermodynamic limit (TDL) are necessary for understanding properties of solids. Wavefunction-based methods such as coupled cluster theory are popular for treating electron correlation in many-body solid-state systems. However, TDL calculations for periodic solids suffer from slow convergence with increased system size, especially at the coupled cluster level of theory. Due to the scaling of the coupled cluster method, energies are often calculated at reduced system sizes and extrapolated to the TDL. Calculating energies at reduced system sizes introduces finite size effects: errors resulting from the modeling of a real solid with a finite number of periodic cells.

Despite this challenge, coupled cluster with single and double excitations (CCSD) has previously been applied to three-dimensional atomistic solids,[1,2] solids in metallic, insulating, and semiconducting phases,[3] and rare-gas solids,[4] and used as a benchmark for comparison against density functional theory (DFT),[5] full configuration interaction quantum Monte Carlo (FCIQMC),[6] and explicitly correlated second-order perturbation theory.[7] Work has also been done to develop equation-of-motion[8–11] and time-dependent[12] coupled cluster formalisms for applications to solids and warm dense matter.

Efforts to extend coupled cluster to solids have benefited from the various finite size corrections developed by the quantum Monte Carlo (QMC) community.[13–16] These developments include extrapolation methods,[17,18] the identification of a special twist angle,[19,20] the replacement of the Coulomb potential with Ewald interactions,[13,21] and corrections based on the random phase approximation (RPA)[22] and the KZK functional.[23] Twist averaging has also been developed for addressing finite size effects in variational,[24] diffusion,[15] and full configuration interaction Monte Carlo methods.[18]

———

a)Electronic mail: james-shepherd@uiowa.edu

In addition to developments in finite size corrections, embedding methods[25] aim to alleviate finite size error by injecting high-level theory into finite regions of systems where an overall high-level treatment is infeasible. CCSD may be used to attain accurate descriptions of correlation in finite regions of systems without incurring the full overhead of a general CCSD calculations. To this aim, CCSD has recently been embedded in RPA.[26] Despite advantages of embedding theories, finite size corrections are still necessary for applications where an overall high-level treatment is desired.

Liao and Grüneis recently introduced a finite size correction for periodic coupled cluster calculations (CCSD-FS) based on the transition structure factor.[2] Their CCSD-FS theory has demonstrated a reduction in finite size error for insulating solids such that energies are converged to within chemical accuracy with $3 \times 3 \times 3$ k–meshes. CCSD-FS has since been used in applications such as the adsorption of water molecules on boron nitride,[1] and for studying the electronic properties of solid hydrogen crystals.[27] However, only insulators were included in the original work by Liao and Grüneis due to the exacerbation of finite size effects in small and zero gap systems. In our recent work locating a special twist angle to converge CCSD correlation energies for metals,[3] we applied CCSD-FS to small gap systems but our application required an additional extrapolation to fully remove finite size error.

This paper is intended as a direct extension of the CCSD-FS theory. In part inspired by recent machine learning approaches across the chemical sciences, we seek to determine whether structure factor interpolation via Gaussian process regression (GPR) can be used to provide a CCSD-FS-level correction for metals. We benefit from the GPR approach not requiring an explicit parametrization of the transition structure factor. For this proof of principle work we took the ground truth for GPR to be the structure factor of the largest system size we could calculate. Then energies were calculated by, first, projecting the resulting continuous function onto auxiliary grids of varying sizes up to this maximum grid

Accepted to J. Chem. Phys. 10.1063/5.0086580

2

size, then, second, extrapolating these to an infinite grid size. These CCSD-FS-GPR energies were then compared with CCSD, extrapolated CCSD over the same grid sizes, CCSD-FS, and CCSD-FS with additional extrapolation.

II.   CCSD AND THE TRANSITION STRUCTURE FACTOR

We employ the coupled cluster method in the typical fashion as described in previous studies.[28–30] Here, we will only provide a brief summary overview of the method.

Coupled cluster expresses the ground-state wavefunction using an exponential operator, $e^{\hat{T}}$, acting on the reference Hartree–Fock wavefunction. The coupled cluster energy is typically expressed as a sum of the Hartree–Fock energy plus the coupled cluster correlation energy correction calculated using the following equation:

$$E_{\text{corr}} = \frac{1}{2} \sum_{ijab} T_{ijab} V_{ijab} \tag{1}$$

Here, T_{ijab} are the individual t-amplitudes for the excitation operator, $\hat{T}$, and V_{ijab} is the antisymmeterized electron repulsion integral defined as $V_{ijab} = 2v_{ijab} - v_{ijba}$. In this paper, we only work with the coupled cluster singles and doubles (CCSD) energy, so T_{ijab} is therefore defined as $T_{ijab} = t_{ijab} + t_{ia}t_{jb}$, where t_{ia} and t_{ijab} correspond to the singles and doubles amplitudes, respectively. Note that for the UEG, singles amplitudes are zero by momentum symmetry and CCSD therefore reduces to CCD. For all equations, i and j refer to occupied orbital indices and a and b refer to unoccupied orbital indices in the wavefunction.

Following the derivation laid out in the paper by Liao and Grüneis,[2] the correlation energy can be re-written in terms of the transition structure factor, $S_{\mathbf{G}}$:

$$E_{\text{corr}} = \frac{1}{2} \sum_{\mathbf{G}}' S_{\mathbf{G}} V_{\mathbf{G}} \tag{2}$$

where $\mathbf{G}$ is the plane wave vector that is a sum of the reciprocal space lattice vector, $\mathbf{g}$, and the difference between two crystal lattice vectors chosen within the first Brillouin zone, $\Delta\mathbf{k}$. Here $V_{\mathbf{G}}$ is the electron repulsion integral as a function of $\mathbf{G}$, taking the form $4\pi/G^2$ (where G is the magnitude of $\mathbf{G}$).

Crucially, in order to correct for finite size errors, we wish to recover the CCSD correlation energy *not captured* by the sum in Eq. (2). The missing energy comes from the discretization of $\mathbf{G}$, leading to quadrature error, and the missing component of the integral between some minimum $|\mathbf{G}|$ and $|\mathbf{G}| = 0$: $\int_{|\mathbf{G}|=0}^{|\mathbf{G}|=G_{\min}} S_G V_G \, dG$. Here, $G_{\min}$ is the minimum non-zero value of G present in the calculation grid.

III.   CONNECTIVITY TWIST AVERAGING

To help reduce finite size effects (FSE) in general, we used our connectivity twist averaging (cTA) method from previous work.[31] This method is based on twist averaging, which reduces finite size effects through reducing the fluctuations in the wavefunction. This is done through using offsets (also called twist angles) which are applied to the grid points such that $(\phi_j \propto \exp(i(\mathbf{k}_j - \mathbf{k}_s) \cdot \mathbf{r}))$, where k_s is the offset. Typically, N_s twist angles are run, and the resulting energies across twist angles are then averaged over to obtain the final twist averaged energy, such that:

$$\langle E_{\text{corr}} \rangle_{k_s} = \frac{1}{N_s} \sum_{l=1}^{N_s} E_{\text{corr}}(k_{s,l})$$

This results in a linear cost scaling factor of N_s for twist averaging. In contrast, cTA aims to reduce this linear cost by finding a single twist angle that reproduces the twist averaged energy. This is done through looking at the momentum transfer vectors between the occupied (i) and virtual (a) orbital manifolds, $x = |\mathbf{n_i} - \mathbf{n_a}|^2$, Here, $\mathbf{n_i}$ is the integer equivalent of the quantum number associated with momentum: $\mathbf{k_i} = \frac{2\pi}{L}\mathbf{n_i}$. For each twist angle, these transfers are then histogrammed over the same indices as the correlation energy in Eq. (1) to form a vector with elements h_x. This vector measures what we call the connectivity. This histogramming is equivalent in cost to an MP2-like calculation and is repeated for 100 randomly chosen twist angles The histogramming vector is then averaged over all twist angles to give $\langle \mathbf{h} \rangle_{k_s}$, which is then used to find the twist angle with the minimum residual difference to the average using the following relationship:

$$S_{\text{res}}(\mathbf{k_s}) = \sum_x \frac{1}{x^2} |h_x(\mathbf{k}_s) - \langle h_x \rangle_{k_s}|^2$$

Here, $S_{\text{res}}(\mathbf{k_s})$ is the residual difference to the average for a single twist angle and $\frac{1}{x^2}$ weights towards the shorter momentum transfers as they contribute more to the energy. The $\mathbf{k_s}$ found with the smallest residual difference to the averaged connectivity is the cTA special twist angle, $\mathbf{k_s}^*$, which we have found can reproduce the twist average energy and, therefore, has the smallest amount of FSE in the energy compared to the other twist angles. With this special twist angle, we can then run a single coupled cluster calculation using Eq. (2), such that:

$$E_{corr}(\mathbf{k_s}^*) = \frac{1}{2} \sum_G' S_G(k_s^*) V_G(k_s^*)$$

As this method depends on the momentum transfers that occur in a given system, a different special twist angle is found for each system.

It is important to note here that our cTA method has only been developed for use in the uniform electron gas so, while all UEG calculations shown were run directly

Accepted to J. Chem. Phys. 10.1063/5.0086580

3

with our cTA method to get the structure factors and energies used in the GPR fits, we had to use another strategy for lithium and sodium. For Li and Na, we used a UEG calculation to find the cTA special twist angle before running a CCSD calculation using the realistic Hamiltonian to obtain the transition structure factors and energies used in the GPR fits. The parameters for the UEG calculations were set to have the same number of electrons as the supercell equivalent to the k-point grid (e.g. for a $2 \times 2 \times 2$ grid we have 16 electrons). We note that the density does not affect the twist angle chosen, and so the same special twist angle (per k-point grid) was used for both the Li and Na systems.

IV. GPR PROTOCOL

Gaussian process regression makes use of the "kernel trick"[32] to flexibly interpolate data for a wealth of applications. There has been a large amount of recent work using GPR for molecular and materials chemistry[33] including the use of GPR to predict molecular crystal energies,[34,35] MP2 and CCSD energies,[36] potential energy surfaces,[37] exchange spin coupling,[38] and for autonomous materials discovery.[39] Recently, GPR was used in a finite size correction for the coupled cluster and auxiliary field quantum Monte Carlo energies of one-dimensional hydrogen chains.[40]

One reason for GPRs growth in popularity is that it is nonparametric, or requires no *a priori* assumption about the distribution of data. For our application this quality is beneficial because we do not need to assume a functional form for the fit. GPR models can also be successfully trained on small data sets, making them a relatively efficient surrogate for many-body methods as GPR scales as $\mathcal{O}(n^3)$ with input size n.

We use GPR because it is well-suited to our design constraints, such as the provision of a smooth, continuous, one-dimensional function which is integrable on the G domain. We tested both GPR and neural networks for fitting transition structure factors and found the GPR approach to be less prone to overfitting. We also tried multiple frameworks for GPR in python, and settled on the open-source `scikit-learn`[41] library for its simple user interface.

Our GPR models require the specification of a kernel and an alpha (or noise) value. The GPR approach is known to be sensitive to the choice of kernel.[32] For our kernel specification we sought a single kernel which would provide reasonable fits across systems for simplicity. Prior to this design choice, we tried a kernel selection via grid search but found this impractical due to a lack of validation metric which accounts for the physics of the transition structure factor. In practice, we found that a sum of two radial basis function (RBF) kernels provided enough flexibility to fit all of our systems. Our kernel is initially set with a length scale of 0.5 and a variance of 1.0, and then tuned to optimal values by maximizing
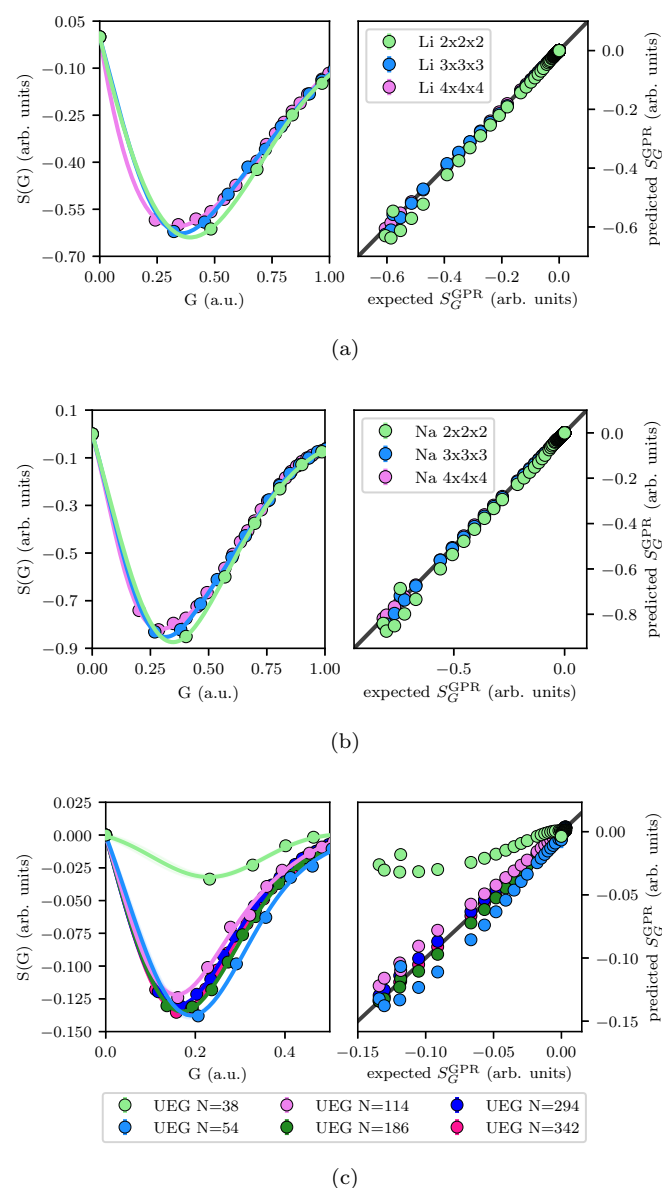

Figure 1. Data from the GPR approach to fitting the transition structure factor for: (a) bcc Li, (b) bcc Na, and (c) UEG $r_S = 5$ a.u. In each panel we show GPR transition structure factor fits for a range of grid sizes on the left, and on the right we compare the transition structure factor for the largest grid (expected) against the GPR fits for each grid (predicted) where an $x = y$ line represents perfect agreement between the expected and predicted transition structure factors. Note that 2σ confidence intervals are plotted on each of the GPR fits but are too small to see.

the log-marginal likelihood during training. We use alpha values of 10^{-5} to 10^{-6} to account for noise in the transition structure factor.

We train our GPR model on G, $\overline{S_G}$ pairs in order to learn a continuous, 1D representation of the transition structure factor, $S(G)$, which we can then project onto discrete grids for use in Eq. (2) (see the Supplementary Material for details concerning the transformation from $S_{\mathbf{G}}$ to $\overline{S_G}$). To generate structure factor predictions,

Accepted to J. Chem. Phys. 10.1063/5.0086580

4

we perform a weighted average over ten trained models' output using a weighting of $1/\sigma^2$, where σ is one standard error returned with each prediction from the `scikit-learn predict` method when the `return_std` flag is set to `True`. This averaging scheme rewards large confidence in predictions, and penalizes predictions with large uncertainty. By averaging over multiple models we additionally increase confidence in predictions and encourage reproducibility without having to seed the random number generator in the L-BFGS-B optimizer used by `scikit-learn` to maximize the log marginal likelihood.

We train a GPR model for each Li, Na, and UEG grid size and then use each trained model to predict transition structure factors on all grid sizes (see Sec. V for details on the systems used). We treat the largest grid size for each system as the ground truth for assessing the performance of our models after training (see the right panels of Fig. 1(a), Fig. 1(b), and Fig. 1(c) for a visual depiction and Section Sec. VI for discussion). We then use the GPR-predicted transition structure factors to calculate correlation energies at each auxiliary grid size by taking the sum in Eq. (2). We then extrapolate the prediction from each model across different grid sizes to the infinite grid limit to calculate the CCSD-FS-GPR energy. The technical details for this are in the Supplementary Material.

V. CALCULATIONS

We calculated coupled cluster transition structure factors and correlation energies for the lithium and sodium solids. The UEG systems were parameterized by a Wigner-Seitz radius of $r_s = 5$ a.u. and system sizes from $N = 38$ electrons to 342 electrons. The bcc Li and bcc Na systems were parameterized by k-point mesh grids from $2 \times 2 \times 2$ to $4 \times 4 \times 4$ (equivalently, $N = 16$ to 128), and lattice constants $a = 3.363$ Å and $a = 4.047$ Å, respectively. Lattice constants were taken from the NOMAD database.[42] The UEG calculations are from an in-group uegccd code,[29,30] and Li and Na calculations are from the Vienna Ab initio Simulation Package (VASP).[43,44] Arbitrary units (denoted arb. units) are used for the transition structure factor and Hartree atomic units are used for G and r_s values.

VI. RESULTS

A summary of our protocol is: 1) to fit the continuous function $S(G)$ using the GPR procedure outlined in Sec. IV, 2) to compare these to transition structure factors at the largest system size (which is our best known data), and 3) to use these fits to estimate a TDL energy and compare with known results. For Li, CCSD-FS is the state-of-the-art and our benchmark for comparison, these data were found to require additional extrapolation to reach the TDL by a previous study.[3] The benchmarks used for comparison for Na and the UEG are extrapolated CCSD and CCD, respectively. Of these benchmarks, the Li and UEG numbers are likely the closest to true TDL benchmarks – from CCSD-FS plus extrapolation and extrapolation over large system sizes, respectively. Na is benchmarked against extrapolated CCSD without the FS correction and is therefore less likely to be at the TDL.

In Fig. 1(a) we show a complete set of GPR fits for Li. The left panel of Fig. 1(a) shows transition structure factors from the $2 \times 2 \times 2$, $3 \times 3 \times 3$, and $4 \times 4 \times 4$ k-meshes. The number of data points these have, respectively, is: 41, 90, and 159. GPR fits these curves well and a typical deviation between the GPR line and the points is 0.0004 at each grid size.

The prediction of the point at the origin is particularly important for this method as we know that the transition structure factor must fall to zero. The predictions are: $-0.0008(10)$ for $2 \times 2 \times 2$ and $3 \times 3 \times 3$, and $-0.0015(10)$ for $4 \times 4 \times 4$. The numbers in parentheses are the error on each value. In general, these values seem reasonable. It is noteworthy that there is a slight, visible change in the position of the structure factor minimum as the grid is increased in size; this is likely due to relaxation of the coupled cluster t amplitudes in the presence of the increased number of basis functions. This kind of minimum drifting can make finite size effects more severe.

The right panel of figure Fig. 1(a) shows these same bcc Li transition structure factors with the GPR function evaluated at grid points corresponding to $4 \times 4 \times 4$. The GPR predictions are plotted against the expected (ground truth) S_G which is taken as the $4 \times 4 \times 4$ transition structure factor. The high G tail of the transition structure factor makes up the vast majority (97.4%, 96.7%, and 96.9% in order of increasing grid size) of the points, and the function is well-fit in this region of the curve (i.e. within 0.049, 0.012, 0.007 of the ground truth mean). While not the focus of this paper, this is important because the tail is responsible for basis set convergence.[45] Its lack of change as the system size changes (either in the GPR or in the presumed t amplitudes) lends credence to the approach of separating out the basis set and finite size errors because they commute.[46] At larger transition structure factor values, $|S_G| > 0.5$, we see that there are discrepancies between what GPR on $2 \times 2 \times 2$ data predicts and what is found at $4 \times 4 \times 4$. Owing to the low fit error in GPR (0.0004) relative to the scale of the transition structure factors, we attribute this to the effects of the structure factor minimum drifting described above rather than to the fitting procedure.

The GPR protocol is repeated for Na and the UEG in the remainder of Fig. 1. The transition structure factor fits for both systems (Fig. 1(b) and Fig. 1(c)) are comparable with Li, except that Na shows slightly less relaxation with increasing system size. The opposite tendency is seen in the UEG. In the right panel of Fig. 1(b) the average deviation between the GPR-predicted Na $4 \times 4 \times 4$

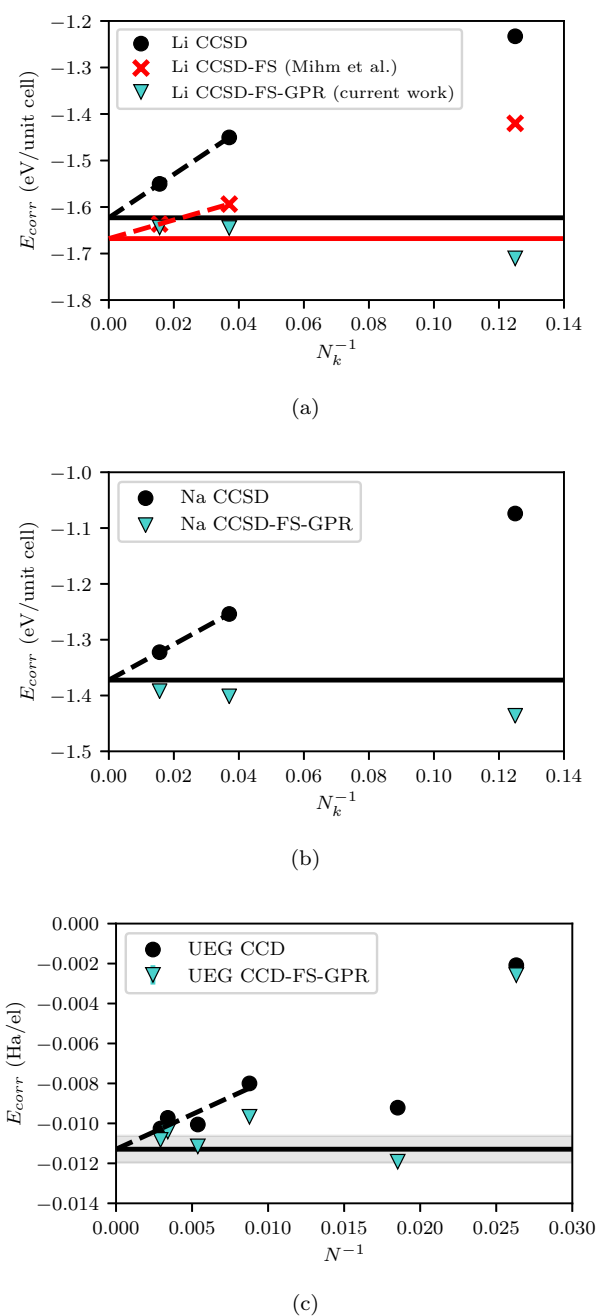*Accepted to J. Chem. Phys. 10.1063/5.0086580*

5

(a)

(b)

(c)

Figure 2. CCSD-FS-GPR energy results compared to CCSD for: (a) bcc Li, (b) bcc Na, and (c) UEG $r_S = 5$ a.u. Horizontal lines show results extrapolated to the thermodynamic limit, with dashed lines showing the points used and function fit for the extrapolation. In panel (a), Mihm et al. CCSD-FS benchmark data comes from Ref. 3 and these authors perform the extrapolation shown here in red (for more discussion, see main text). In panel (c), the shaded region corresponds to plus/minus one standard error on the 4-point CCD extrapolation; the remaining extrapolations in panels (a) and (b) are 2-point extrapolations and therefore do not have a standard error.

point and the line is comparable to the Li result: 0.022, 0.008, and 0.004 in order of increasing grid size. The deviation is slightly larger for the UEG in Fig. 1(c), ranging from 0.003 at large N to 0.109 at small N. Both Na and UEG predictions at the $G = 0$ point are reasonable at all grid sizes and comparable to those of Li.

To get a better sense of these GPR fit errors, we can convert them to energies by comparing the CCSD energy of a given k-point grid with an energy evaluation of the GPR-trained S_G projected onto the same grid. We find that these energies are generally small compared to the overall energy. For Li and Na they are < 3 meV/unit cell and < 0.4 meV/unit cell respectively; for the UEG they are < 3 mHa/el decreasing to < 1 mHa/el at $N > 186$. A full list of these values is presented in the supplementary material.

Figure 2 compares CCSD-FS-GPR results with extrapolated CCSD. In Fig. 2(a) bcc Li TDL energy estimates from GPR predictions at the three grid sizes are compared with CCSD-FS literature values from our recent paper.[3] In that paper, the CCSD-FS method was used on a different twist angle selection protocol analogous to our cTA method here (see Sec. II). The CCSD-FS data were then extrapolated to the TDL using $1/N$ owing to a remaining error in the functional form of the fit.[3] This is our best estimate for the TDL of this system. We also include another power-law fit from conventional $1/N$ extrapolation to our CCSD data.

All three of our bcc Li CCSD-FS-GPR energies are considerably improved over the raw CCSD correlation energies for each grid size. The CCSD-FS-GPR energies for $3 \times 3 \times 3$ and $4 \times 4 \times 4$ lie between extrapolated CCSD and CCSD-FS (with additional extrapolation). We see this as consistent with the need to extrapolate after CCSD-FS corrections in metals. We also see that the GPR correction for $2 \times 2 \times 2$ is almost twice as large as that of CCSD-FS. This trend in how CCSD-FS-GPR behaves for Li matches the behavior of CCSD-FS in insulators and improves over CCSD-FS without extrapolation in metals.

CCSD-FS-GPR results are presented for Na and the UEG in the remainder of Fig. 2. As with Li, all three of the bcc Na CCSD-FS-GPR predictions improve over both the raw CCSD results. In Fig. 2(b), the $3 \times 3 \times 3$ and $4 \times 4 \times 4$ Na grid sizes predict similar energies of -1.402 eV/unit cell and -1.392 eV/unit cell, respectively, which falls slightly lower than the extrapolated CCSD number (-1.372 eV/unit cell). A downward trend in the TDL predictions is visible across the three grids, with $2 \times 2 \times 2$ predicting a lower value of -1.436 eV/unit cell. We attribute this to a similar effect seen in Fig. 2(a) for Li and in general believe the $3 \times 3 \times 3$ and $4 \times 4 \times 4$ predictions to be reasonable and the $2 \times 2 \times 2$ prediction to be greatly improved over the CCSD number.

In Fig. 2(c), we show CCSD-FS-GPR results for the UEG. The UEG CCSD-FS-GPR result converges nicely to the extrapolated CCSD with increased system size and agrees with the extrapolated CCSD within error for grid sizes $N = 54$, 186, and 342. Unlike for Na and Li, the

smallest grid size ($N = 38$) barely improves over the raw CCSD energy. This is likely due to the points in $N = 38$ being severely different than the other system sizes (see Fig. 1(c)). However, the CCSD-FS-GPR energies generally improve over the raw CCSD correlation energies and the discrepancies between predictions at different grid sizes is likely due to the structure factor minimum drifting which is most severe for the UEG.

VII. CONCLUSIONS

In this paper we provide a straightforward application of Gaussian process regression (GPR) for correcting finite size effects in small metallic systems. This application is an extension of work previously done by Liao and Grüneis on transition structure factor interpolation for correcting finite size errors in coupled cluster calculations for insulators.[2] We have demonstrated in this application that transition structure factor interpolation via GPR can correct finite size error in simple metals to achieve CCSD-FS level accuracy; we call the resulting method CCSD-FS-GPR.

We note that we did not attempt to address the accuracy of CCSD in this manuscript. CCSD was our model of choice for finite size effects because of a trade off between accuracy and cost. However, it is also believed that CCSD will capture the physical effects of what is happening on approach to the thermodynamic limit (i.e. the long-range effects), in part because it contains the same physical information as the random phase approximation (RPA). This paper highlights the importance of the minimum of the structure factor, which appears to change with system size (and presumably method). Our initial investigations have found this to be a complicated phenomenon and we will save its discussion for a separate study.

We believe these results will generalize to different electronic structures, atomic compositions, and lattice symmetries; though there are formalism challenges related to including anisotropy that will need to be overcome. This is where our attention will focus in our next study.

As previously referenced, a similar GPR approach for finite size correction in solids was very recently published by Borda and Rubenstein as a preprint.[40] While the work of Borda and Rubenstein is similar to our approach in the motivation for using the GPR algorithm, a fundamental difference in the two approaches is the subject of the regression. Where we learn a representation of the CCSD transition structure factor which is then summed to attain CCSD correlation energies, Borda and Rubenstein learn correlation energies from GPR models trained on smooth overlap of atomic positions (SOAP) descriptors. We believe that our work is consistent with the evidence provided by Borda and Rubenstein that GPR is a fine tool for finite size correction in costly many-electron calculations.

VIII. SUPPLEMENTARY MATERIAL

The reader is directed to the Supplementary Material for details concerning extrapolation methods.

IX. ACKNOWLEDGEMENTS

X. DATA AVAILABILITY

The data that support the findings of this study are available within the article.

REFERENCES

[1] T. Gruber, K. Liao, T. Tsatsoulis, F. Hummel, and A. Grüneis, Physical Review X **8**, 021043 (2018).

[2] K. Liao and A. Grüneis, The Journal of Chemical Physics **145**, 141102 (2016).

[3] T. N. Mihm, T. Schäfer, S. K. Ramadugu, L. Weiler, A. Grüneis, and J. J. Shepherd, Nature Computational Science , 1 (2021).

[4] K. Rościszewski, B. Paulus, P. Fulde, and H. Stoll, Physical Review B **62**, 5482 (2000).

[5] V. O. Kiohara, E. F. Carvalho, C. W. Paschoal, F. B. Machado, and O. Roberto-Neto, Chemical Physics Letters **568-569**, 42 (2013).

[6] G. H. Booth, A. Grüneis, G. Kresse, and A. Alavi, Nature **493**, 365 (2013).

[7] A. Grüneis, Physical Review Letters **115**, 066402 (2015).

[8] A. M. Lewis and T. C. Berkelbach, Physical Review Letters **122**, 226402 (2019).

[9] X. Wang and T. C. Berkelbach, arXiv:2107.04740 [cond-mat, physics:physics] (2021), arXiv: 2107.04740.

[10] X. Wang and T. C. Berkelbach, Journal of Chemical Theory and Computation **16**, 3095 (2020).

[11] J. McClain, Q. Sun, G. K.-L. Chan, and T. C. Berkelbach, Journal of Chemical Theory and Computation **13**, 1209 (2017).

[12] A. F. White and G. K.-L. Chan, Journal of Chemical Theory and Computation **14**, 5690 (2018).

[13] L. M. Fraser, W. M. C. Foulkes, G. Rajagopal, R. J. Needs, S. D. Kenny, and A. J. Williamson, Physical Review B **53**, 1814 (1996).

[14] M. Holzmann, R. C. Clay III, M. A. Morales, N. M. Tubman, D. M. Ceperley, and C. Pierleoni, Physical Review B **94**, 035126 (2016).

[15] N. D. Drummond, R. J. Needs, A. Sorouri, and W. M. C. Foulkes, Physical Review B **78**, 125106 (2008).

Accepted to J. Chem. Phys. 10.1063/5.0086580

7

[16] S. Azadi and W. Foulkes, The Journal of chemical physics **143**, 102807 (2015).

[17] M. Holzmann, B. Bernu, C. Pierleoni, J. McMinis, D. M. Ceperley, V. Olevano, and L. Delle Site, Physical Review Letters **107**, 110402 (2011).

[18] M. Ruggeri, P. L. Ríos, and A. Alavi, Physical Review B **98**, 161105 (2018).

[19] G. Rajagopal, R. J. Needs, S. Kenny, W. M. C. Foulkes, and A. James, Physical Review Letters **73**, 1959 (1994), publisher: American Physical Society.

[20] M. Dagrada, S. Karakuzu, V. L. Vildosola, M. Casula, and S. Sorella, Physical Review B **94**, 245108 (2016).

[21] A. J. Williamson, G. Rajagopal, R. J. Needs, L. M. Fraser, W. M. C. Foulkes, Y. Wang, and M.-Y. Chou, Physical Review B **55**, R4851 (1997).

[22] S. Chiesa, D. M. Ceperley, R. M. Martin, and M. Holzmann, Physical Review Letters **97**, 076404 (2006).

[23] H. Kwee, S. Zhang, and H. Krakauer, Physical Review Letters **100**, 126404 (2008).

[24] C. Lin, F. H. Zong, and D. M. Ceperley, Physical Review E **64**, 016702 (2001).

[25] Q. Sun and G. K.-L. Chan, Accounts of Chemical Research **49**, 2705 (2016).

[26] T. Schäfer, F. Libisch, G. Kresse, and A. Grüneis, The Journal of Chemical Physics **154**, 011101 (2021).

[27] K. Liao, T. Shen, X.-Z. Li, A. Alavi, and A. Grüneis, Physical Review B **103**, 054111 (2021).

[28] J. J. Shepherd and A. Grüneis, Physical Review Letters **110**, 226401 (2013).

[29] J. J. Shepherd, T. M. Henderson, and G. E. Scuseria, Physical Review Letters **112**, 133002 (2014).

[30] J. J. Shepherd, T. M. Henderson, and G. E. Scuseria, The Journal of Chemical Physics **140**, 124102 (2014).

[31] T. N. Mihm, A. R. McIsaac, and J. J. Shepherd, The Journal of Chemical Physics **150**, 191101 (2019).

[32] C. K. Williams and C. E. Rasmussen, *Gaussian processes for machine learning*, Vol. 2 (MIT press Cambridge, MA, 2006).

[33] V. L. Deringer, A. P. Bartók, N. Bernstein, D. M. Wilkins, M. Ceriotti, and G. Csányi, Chemical Reviews **121**, 10073 (2021).

[34] M. Krynski and M. Rossi, arXiv preprint arXiv:2106.08612 (2021).

[35] N. Raimbault, A. Grisafi, M. Ceriotti, and M. Rossi, New Journal of Physics **21**, 105001 (2019).

[36] M. Welborn, L. Cheng, and T. F. Miller, Journal of Chemical Theory and Computation **14**, 4772 (2018).

[37] A. Kamath, R. A. Vargas-Hernández, R. V. Krems, T. Carrington Jr, and S. Manzhos, The Journal of chemical physics **148**, 241702 (2018).

[38] M. P. Bahlke, N. Mogos, J. Proppe, and C. Herrmann, The Journal of Physical Chemistry A **124**, 8708 (2020).

[39] M. M. Noack, G. S. Doerk, R. Li, J. K. Streit, R. A. Vaia, K. G. Yager, and M. Fukuto, Scientific reports **10**, 1 (2020).

[40] E. J. L. Borda and B. Rubenstein, arXiv preprint arXiv:2112.10334 (2021).

[41] 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, Journal of Machine Learning Research **12**, 2825 (2011).

[42] M. Scheffler, "The NoMaD (Novel Materials Discovery) Center of Excellence: Challenges and Solutions to Build a Code-Independent Data Base for Computational Materials Science and to Identify Causal Models in Machine Learning," (2015).

[43] G. Kresse and J. Hafner, Journal of Physics: Condensed Matter **6**, 8245 (1994).

[44] G. Kresse and J. Furthmüller, Physical Review B **54**, 11169 (1996).

[45] K. Liao and A. Grüneis, The Journal of Chemical Physics **145**, 141102 (2016).

[46] J. J. Shepherd, The Journal of Chemical Physics **145**, 031104 (2016).