SAND2015-5296J

# Efficient use of *ab initio* calculations to generate accurate Newtonian dynamics

M. C. Shaughnessy[†,¶] and R. E. Jones[*,‡]

*Materials Physics Department, Sandia National Laboratories, Livermore, CA 94550, USA, and Mechanics of Materials Department, Sandia National Laboratories, Livermore, CA 94550, USA*

E-mail: rjones@sandia.gov

**Abstract**

We develop and demonstrate a method to efficiently use density functional calculations to drive classical dynamics of complex atomic and molecular systems. The method has the potential to scale to systems and time-scales unreachable with current *ab initio* molecular dynamics schemes. It relies on a dataset of previously computed Hellmann-Feynman forces for atomic configurations endowed with a distance metric. The metric on configurations enables fast database lookup and robust interpolation of the stored forces. We discuss mechanisms for the database to adapt to the needs of the evolving dynamics while maintaining accuracy, and other extensions of the basic algorithm.

[*]To whom correspondence should be addressed
[†]Materials Physics Department, Sandia National Laboratories, Livermore, CA 94550, USA
[‡]Mechanics of Materials Department, Sandia National Laboratories, Livermore, CA 94550, USA
[¶]currenty at: Shroudbase, Berkeley, CA 94701 USA

# 1 Introduction

The need for simulation of complex materials and processes with atomic resolution is ever-increasing as materials and devices are constructed with nanoscale structure and principles. Examples include the ubiquitous vapor deposition techniques, nanoscale assembly of supramolecular materials, self-assembled surface coatings, and fabrication of topological insulator microelectronics. These systems are typically too large for current computers to model directly with *ab initio* methods such as density functional theory (DFT); and, molecular methods based on empirical potentials suffer from issues of accuracy as well as applicability to compositionally and structurally complex environments. There have been many efforts to develop empirical potentials for molecular dynamics (MD) and statics with the full accuracy of the DFT data typically used to tune the potentials, *e.g.* the bond-order potentials.[1–6] One of the more notable recent efforts in this arena is the potential representation developed by Bartók *et al.*,[7–9] which uses a general basis to represent the local atomic environment in order to construct a potential that is accurate across all probable configurations. The task of finding a suitable form for the empirical potential[10] and calibrating it is vastly complicated by the fact that most potentials of this type are pairwise for elements and hence all binary combinations need to be tuned.

Inspired by these challenges, we take a different tack based on the premise that the system can be decomposed into local atomic environments, as in classical MD, and, if the neighborhoods are sufficiently large, the Hellmann-Feynman (HF) forces that generate standard Born-Oppenhiemer (BO) dynamics are sufficiently accurate and representative of the forces of corresponding neighborhoods in a much larger system. In effect, we make the plausible assumption that atoms closer to any atom of interest have more influence and atoms sufficiently far away have negligible effect. In-line with this assumption of *locality*, we construct local representations of interatomic forces on-the-fly from the most relevant *ab initio* data instead of attempting to form a globally-accurate potential *a priori* . For efficiency, we also rely on the fact that the dynamics of the overall system will revisit the vicinity in phase

space of previous local configurations[11–13] so that eventually new configurations will be in regions densely sampled by previous configurations. [a].

We stress, given the need for on-the-fly DFT calculations and the focus on large-scale systems, that this is an algorithm most suitable for distributed supercomputers, not desktop computers. Nevertheless, for the present work we focus on the theoretical and practical aspects accessible on workstations. The fundamentals of the proposed algorithm are:

1. As dynamics ensues, new configurations are generated and compared to a neighborhood-force database.

2. If sufficient stored neighborhoods are close to a new/query neighborhood, the stored forces are interpolated at the query neighborhood; otherwise, HF forces are calculated for the new neighborhood and stored for subsequent searches.

Fig. 1 depicts how these steps generate dynamics. This algorithm relies on an abstract neighborhood space, with symmetries and physical constraints, endowed with a distance metric to enable fast queries and robust interpolation of the resulting data. We will hereafter refer to these neighborhoods as *configurational neighborhoods* or simply *clusters*. The fundamentals of this approach (a) draws from basic pattern-matching metrics, *e.g.* Ref. 15 and Ref. 16; (b) has some similarity to tabular methods developed in other arenas;[14,17–19] and (c) is related to the on-lattice cluster expansion method used in a Monte Carlo context.[20–23]

Four questions will be answered in this work:

Q1. Does the force on the central atom in a neighborhood converge to that on the central atom of a corresponding sub-domain embedded in the full system?

Q2. Do the neighborhood-neighborhood distances directly correlate with the forces on the atoms in the clusters and how a large a neighborhood needs to be considered for the neighborhood-neighborhood comparison?

[a]The idea of local interpolation functions of local configurations is not new, in fact Ref. 14 used interpolation to represent the potential energy surface.

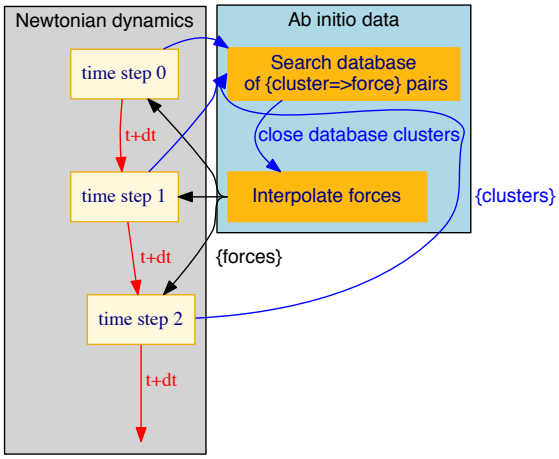



Figure 1: Schematic of the basic algorithm. The database search and force interpolation takes on the traditional role of an empirical potential.

Q3. Is a database search fast enough to enable a feasible and efficient algorithm?

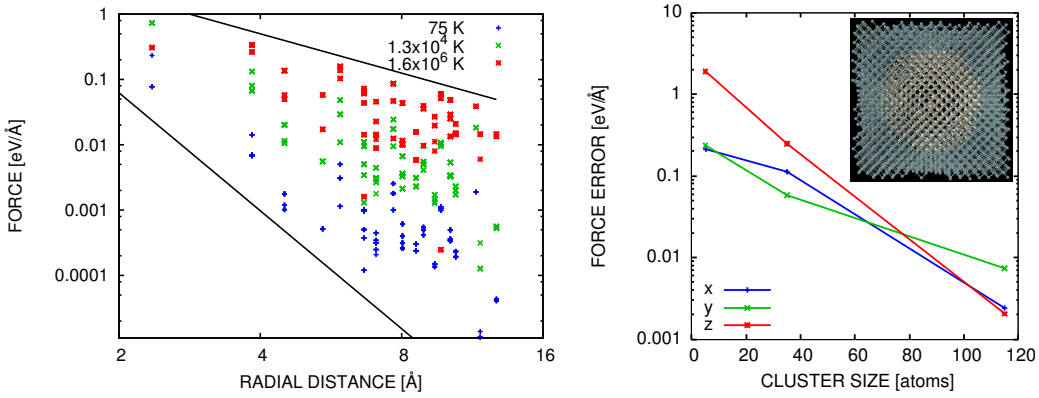
Q4. Is the error in the interpolation of forces controllable and convergent with increasing local density of samples in cluster space?

Since the first question is directly related to the basic premise and is generally accepted, we will put it to rest here in the introduction. (In fact the locality implied in question Q1 is the basic predicate of all empirical potentials.) For traditional BO dynamics,[24,25] the forces are derived from the Hellman-Feynman theorem, *i.e.* from derivatives with respect to the nuclear coordinates of the energy formed from the electronic wave-functions. Momentum is conserved because the forces are derived from this (implicit) potential dependent on all the nuclear positions; whereas, energy conservation is primarily due to the choice of time-integrator, as in classical MD. Hence, convergence of central force with neighborhood size implies conservation of momentum on par with BO dynamics. As an illustration, a plot of the magnitude of the force $\|\mathbf{f}(r)\|$ on atoms surrounding a perturbed atom as a function of distance from the perturbed atom $r$ is shown in Fig. 2a. The samples obey the decay of force with distance

$$\|\mathbf{f}(r)\| \leq C\frac{\|{}_{\Delta}\mathbf{f}\|}{r^2} \tag{1}$$

expected from Coulombic interactions, where ${}_{\Delta}\mathbf{f}$ is the force on displaced atom and the

constant $C$ depends on the size of the perturbation. Conversely, Fig. 2b shows that the force on the central atom of the system converges with the size of the isolated spherical sub-system relative to force calculated in the full periodic system. These results give us confidence that for materials with a sufficiently large dielectric constant, only a limited number of nearest neighbors are needed to determine the force on a central atom. When we use the proposed neighborhood decomposition in a large scale simulation, we lose the explicit dependence of the force on the coordinates of the atoms outside the neighborhood. The variation in the true, full-system force for neighborhoods in the database that are zero (cluster) distance apart is a measure of the error introduced by the cluster approximation. This phenomenon is explored further in the Results section.




(a) (b)

Figure 2: (a) Spatial decay of force due to a small perturbation in a perfect Si lattice as a function of the distance ($r$) from the perturbed atom. The upper trend line is $1/r^2$, the lower is $1/r^6$ and the equivalent temperatures are estimated from the energy of the perturbation $\triangle E$ as $T \equiv \triangle E/k_B$. (b) Convergence of the difference in forces on the central (perturbed) atom in an isolated cluster with increasing neighborhood size (in number of atoms) and in a fully periodic system. (Inset) periodic 1000 atom system (cube containing blue and tan atoms) and sub-configuration (sphere containing tan atoms) with a perturbed central atom.

With regard to questions Q 2-4, we need a metric to measure the distance between neighborhoods and thus their similarity. There are many similarity measures proposed and in use for structure comparison and reconstruction,[26,27] potential energy exploration[8,28] and the related field of sequence comparison,[29,30] but not all are metrics and hence do not enable

efficient geometric database queries. We concentrate on two metrics: (a) the well-known *Root Mean Squared Deviation* (RMS-D) and (b) our adaptation of the overlap of Gaussian type orbitals (OGTO)[31] and its spherical harmonic decomposition, the so-called *Smooth Overlap of Atomic Positions* (SOAP)[8] measure. which is a spherical harmonic decomposition of the overlap of Gaussian type orbitals (OGTO) in Ref. 31. The RMS-D metric is commonly used in protein folding and drug discovery to quantify geometric similarity between molecules.[26,27,31–34] The origins of the SOAP representation trace to the *bond-order* formalism.[3] [b]

In the next section, Sec. 2, we will develop metrics to compare the similarity of a given local configuration and those stored in a database and then, based on this measure, interpolate the associated forces to give an estimate for the force on any given atom. In Sec. 3 we will outline an algorithm to efficiently search through a large database of configuration-force pairs using the metric properties of the distance measure and a hierarchical graph representation of the database. Then, in Sec. 4 in addition to testing the conservation properties of the algorithm, we address questions Q2-4 with results generated with the proposed method. Finally, in Sec. 5 we conclude with a discussion of the results and ideas for future work.

# 2 Theory

In this section we introduce the two chosen metrics, RMS-D and OGTO/SOAP, including means of determining the optimal rotation and permutation for appropriately invariant comparison of clusters. At the heart of both metrics is a representation of local atomic configuration. Then we apply these metrics to the task of approximating the force on the central atom of given neighborhood by interpolation of forces of nearby neighborhoods with

[b] Other measures include: (a) so-called *fingerprint* differences between eigenvector representations of the symmetric matrix formed from interatomic distance of neighborhoods[28,31] (*i.e.* the spectral content of the edge distance-weighted adjacency matrix of graph theory[35,36]), (b) the Weyl matrix of inner products of position vectors,[8] (c) the Kullback-Leibler and Jensen-Shannon divergences of the probability density representations of the point densities,[34] as well as (d) straightforward central moments of the point masses comprising the neighborhood configuration.

pre-computed forces. These developments form the basis for database construction and search algorithms given in the following Methods section.

## 2.1 Cluster distance metrics

Given a configurational neighborhood $A$ of all atoms, $\alpha \in \mathcal{N}_A \equiv \{b_1 \dots b_{N_A}\}$, in a ball around a central atom $a$:

$$\mathcal{C}_A = \{\mathbf{x}_{\alpha_1 a}, \mathbf{x}_{\alpha_2 a}, \mathbf{x}_{\alpha_3 a}, \dots, \mathbf{x}_{\alpha_{N_A} a}\} \tag{2}$$

where $\mathbf{x}_{\alpha a} \equiv \mathbf{x}_\alpha - \mathbf{x}_a$ are distance vectors relative to the position of the central atom $\mathbf{x}_a$, the basic properties that a distance metric $d(A, B)$ between neighborhoods $A$ and $B$ must satisfy are:

M1. **coincidence**: $d(A, B) = 0$ if and only if $\mathcal{C}_A = \mathcal{C}_B$
M2. **positivity**: $d(A, B) > 0$
M3. **symmetry**: $d(A, B) = d(B, A)$
M4. **triangle inequality**: $d(A, C) \leq d(A, B) + d(B, C)$
M5. **reverse triangle inequality**: $d(A, C) \geq |d(A, B) - d(B, C)|$

in addition to the physical invariances $d(A, B) = d(A, B')$:

I1. **translation**: $\mathcal{C}_B \to \mathcal{C}_{B'} = \mathcal{C}_B + \mathbf{a}$
I2. **rotation**: $\mathcal{C}_B \to \mathcal{C}_{B'} = \mathsf{R}\mathcal{C}_B$
I3. **permutation**: $\mathcal{C}_B \to \mathcal{C}_{B'} = \mathsf{P}\mathcal{C}_B$

Note that the use of relative distances $\mathbf{x}_{\alpha a}$ is a simple means of enforcing translational invariance; and, the reverse triangle inequality (M5) is an elementary implication of the triangle inequality (M4).

With an inner product a distance metric can be generated directly. In general, an inner product has three properties:

P1. **symmetry**: $\langle A, B\rangle = \langle B, A\rangle$
P2. **linearity**: $\langle sA + B, C\rangle = s\langle A, C\rangle + \langle B, C\rangle$

P3. **induction of a norm**: $\|A\|^2 \equiv \langle A, A\rangle \geq 0$ and $\langle A, A\rangle = 0$ only if $A = 0$

the last of which induces a metric $d(A, B) = \|A - B\|$.

### 2.1.1 RMS-D

Assuming that the size of all clusters $N$ is the same $N_A = N_B \equiv N$, a weighted root mean square deviation (RMS-D) comparison metric is simply

$$\begin{aligned} d_{\text{RSMD}}(A, B) &= \min_{\mathbf{R},\mathsf{P}} \sqrt{\sum_{\alpha,\beta=1}^{N} \underbrace{\|\mathbf{x}_{\alpha a} - \mathbf{R}\mathbf{x}_{\beta b}\|^2}_{r^2_{\alpha\beta'}} \; P_{\alpha\beta}} \\ &= \min_{\mathbf{R},\mathsf{P}} \sqrt{(\mathsf{X}_A - \mathsf{X}_B\mathbf{R}^T) \cdot \mathsf{P}\,(\mathsf{X}_A - \mathsf{X}_B\mathbf{R}^T)} = \min_{\mathbf{R},\mathsf{P}} \|\mathsf{X}_A - \mathsf{X}_B\mathbf{R}^T\|_{\mathsf{P}} \\ &= \min_{\mathbf{R},\mathsf{P}} \sqrt{\|\mathsf{X}_A\|^2_{\mathsf{P}} + \|\mathsf{X}_B\|^2_{\mathsf{P}} - 2\mathsf{X}_A \cdot \mathsf{P}\mathsf{X}_B\mathbf{R}^T} \end{aligned} \tag{3}$$

where $\mathsf{X}_A = [\mathbf{x}_{\alpha_1 a}, \mathbf{x}_{\alpha_2 a}, \mathbf{x}_{\alpha_3 a}, \ldots, \mathbf{x}_{\alpha_{N_A} a}]^T$ and $\mathsf{X}_B$ are $(N \times 3)$ matrices of the neighborhood vectors, $\mathbf{R} \in \mathsf{SO}(3)$ is a (physical $3 \times 3$) rotation of the neighborhood and $\mathsf{P}$ is a $(N \times N)$ permutation of ordering of atoms in the neighborhood, *i.e.* a binary orthogonal matrix which is simply the rearrangement of the rows of the identity matrix [c]. The permutation matrix $\mathsf{P}$ selects which of the pair-wise distances between atoms of the two configurations will contribute to the metric $d_{\text{RSMD}}$.

To determine the optimal rotation $\mathbf{R}$, Kabsch[37] devised the following solution to this version of Wahba's problem.[38] Starting with Eq. (3) and noticing the last term is the only one dependent on $\mathbf{R}$ leads to:

$$2\max_{\mathbf{R}} \mathsf{X}_A \cdot \mathsf{P}\mathsf{X}_B\mathbf{R}^T = 2\max_{\mathbf{R}} \operatorname{tr} \underbrace{\mathsf{X}_A^T\mathsf{P}\mathsf{X}_B}_{\mathbf{U}\mathbf{S}\mathbf{V}^T} \mathbf{R}^T = 2\max_{\mathbf{R}} \operatorname{tr} \mathbf{S}\mathbf{V}^T\mathbf{R}^T\mathbf{U} = 2\operatorname{tr}\mathbf{S} \tag{4}$$

after application a singular value decomposition (SVD) of $\mathsf{X}_A^T\mathsf{P}\mathsf{X}_B = \mathbf{U}\mathbf{S}\mathbf{V}^T$ where $\mathbf{U}, \mathbf{V} \in \mathsf{SO}(3)$ and $\mathbf{S}$ is diagonal and also of dimension 3. This term in the metric is maximized when

[c] Regarding notation, a boldface font will be used for vectors and tensors in real space, *e.g.* $\mathbf{R}$ and $\mathbf{a}$; whereas, a sans-serif font will be employed for general matrices, *e.g.* $\mathsf{P}$.

$\mathbf{V}^T\mathbf{R}^T\mathbf{U}$ is the identity tensor and the resulting optimal rotation $\mathbf{R} = \mathbf{V}\mathbf{U}^T$ reduces the last term Eq. (3) to the trace of the matrix of eigen-values $\mathrm{tr}\,\mathbf{S}$. Consequently

$$\min_{\mathbf{R}} \|\mathsf{X}_A - \mathsf{X}_B\mathbf{R}^T\|_{\mathsf{P}} = \sqrt{\|\mathsf{X}_A\|^2_{\mathsf{P}} + \|\mathsf{X}_B\|^2_{\mathsf{P}} - 2\,\mathrm{tr}\,\mathbf{S}(\mathsf{P})} \tag{5}$$

for a fixed permutation $\mathsf{P}$ [d].

It remains to solve the permutation problem, *i.e.* an optimal match is possibly only one of the $N!$ permutations of the ordering of the labels $\{1, 2, \ldots, N\}$ for neighborhood $B$, which is clearly not independent of the optimization with respect rotation. This is a complex problem, as recognized in this and related fields, see *e.g.* Ref. 30 and, in general, most applicable integer optimization methods are known to have difficulties with symmetries. Possible solutions include using combinatorial programming algorithms like *branch-and-bound*.[31,40]

Given the geometry of our application, our approach is to segregate the permutation problem into smaller sub-problems involving only the sets that have (nearly) identical radial sorts, *i.e.* equivalent shells. As an alternative to this straight-forward $r$-sort and permute, we can use the spectral representation of the clusters to perform what we call a $\lambda$-sort:

1. Form the adjacency matrices $[\mathsf{A}]_{ij} = r^2_{\alpha_i\alpha_j}$ for both $A$ and $B$ and compute eigen-system for $\mathsf{A}_A$ and $\mathsf{A}_B$
2. Compare the eigenvalues $\lambda_i$ of $\mathsf{A}_A$ and $\mathsf{A}_B$, starting with the largest. If $\lambda^A_i \approx \lambda^B_i$ then sort and match the components of the corresponding eigenvectors $\mathbf{e}^A_i$ and $\mathbf{e}^B_i$ to form a order map $m : a \in Q \to b \in B$ and hence a trial permutation matrix $\mathbf{P}_{a,m(a)} = 1,\ \forall a \in Q$.

In preliminary tests the $\lambda$ sort, albeit more expensive, succeeded in finding the best match cases where the $r$-sort failed *e.g.* a tetrahedral arrangement of neighbors with only one atom perturbed. In any event the overall method does not require the optimal match in all queries

[d] The rotation $\mathbf{R} = \mathbf{I}$ maximizes since $\mathrm{tr}\,\mathbf{SR} = \mathrm{tr}\,\mathbf{R}\Sigma_i\lambda_i\mathbf{e}_i\otimes\mathbf{e}_i = \Sigma_i\lambda_i\mathbf{R}\mathbf{e}_i\cdot\mathbf{e}_i = \Sigma_i\lambda_i\mathbf{I}\mathbf{e}_i\cdot\mathbf{e}_i = \Sigma_i\lambda_i$, using an eigenvalue representation of $\mathbf{S}$. Alternatives to the SVD-based solution exist most notably quaternion-based solutions.[31,39]

if the database is dense and perhaps redundant with permutations of a given configurations.

Lastly, the assumption of locality motivates the use of a weighted norm within the radial cutoff implicit in the chosen cluster size. We can set the non-zero entries of of the permutation operator $P_{\alpha\beta}$ to the value of positive weight function dependent on the radial coordinates of $\alpha$ and $\beta$ so that the distance metric is less sensitive to neighbors far away from the central atom. Also the usual normalization by number of atoms, $1/(N+1)$, can be accommodated in $\mathsf{P}$ which creates an interpretation of $d$ as an average atom-to-atom distance. As proved in Ref. 41 via equivalence classes induced by the optimal rotation, the RMS-D measure is a valid distance metric.

### 2.1.2 OGTO

As we will see, the so-called overlap of Gaussian type orbitals (OGTO) is more complex than RMS-D; however, it has the advantage of being intrinsically permutationally invariant by comparing all atom locations in one cluster to those in the other in a weighted fashion (instead of specific pairs with RMS-D). OGTO starts with the representation of atomic density of a neighborhood surrounding a central atom (whose density, $\Delta(\mathbf{0})$, we omit in the representation, as in the RMS-D metric):

$$\rho_A(\mathbf{x}) = \sum_{\alpha \in \mathcal{N}_A} \Delta(\mathbf{x}_{\alpha a}) \tag{6}$$

in terms of Gaussian smeared point densities

$$\Delta(\mathbf{x}) = \frac{1}{(\sqrt{2\pi}\sigma)^3} \exp\left(-\frac{\|\mathbf{x}\|^2}{2\sigma^2}\right), \tag{7}$$

with width $\sigma$ [e]. With this representation, we can form the product between the atomic densities of neighborhoods $A$ and $B$:

$$\begin{aligned}\langle \rho_A, \rho_B \rangle &= \int \mathrm{d}\mathbf{x}\, \rho_A(\mathbf{x})\rho_B(\mathbf{x}) = \int \mathrm{d}\mathbf{x} \sum_{\alpha\in\mathcal{N}_A} \Delta(\mathbf{x}-\mathbf{x}_\alpha) \sum_{\beta\in\mathcal{N}_B} \Delta(\mathbf{x}-\mathbf{x}_\beta) \\ &= \sum_{\alpha\in\mathcal{N}_A,\beta\in\mathcal{N}_B} \int \mathrm{d}\mathbf{x}\, \Delta(\mathbf{x}-\mathbf{x}_\alpha)\Delta(\mathbf{x}-\mathbf{x}_\beta) \\ &= \frac{1}{(2\sqrt{\pi}\sigma)^3} \sum_{\alpha\in\mathcal{N}_A,\beta\in\mathcal{N}_B} \exp\left(-\frac{r_{\alpha\beta}^2}{4\sigma^2}\right)\end{aligned} \tag{8}$$

and thenceforth, an OGTO-based metric:

$$\begin{aligned}&d_{\mathrm{OGTO}}(A,B) = \min_{\mathbf{R}} \sqrt{\langle \rho_A, \rho_A \rangle + \langle \rho_B, \rho_B \rangle - 2\langle \rho_A, \mathbf{R}\rho_B \rangle} \\ &= \frac{1}{(2\sqrt{\pi}\sigma)^{\frac{3}{2}}} \sqrt{\sum_{\alpha_1,\alpha_2\in\mathcal{N}_A} \exp\left(-\frac{r_{\alpha_1\alpha_2}^2}{4\sigma^2}\right) + \sum_{\beta_1,\beta_2\in\mathcal{N}_B} \exp\left(-\frac{r_{\beta_1\beta_2}^2}{4\sigma^2}\right) - 2\min_{\mathbf{R}} \sum_{\alpha\in\mathcal{N}_A,\beta\in\mathcal{N}_B} \exp\left(-\frac{r_{\alpha\beta'}^2}{4\sigma^2}\right)}\end{aligned} \tag{9}$$

where $r_{\alpha\beta'}^2 = \|\mathbf{x}_\alpha - \mathbf{R}\mathbf{x}_\beta\|^2$ as in Eq. (3). (Clearly, this product (8) is non-negative given positive densities but given its form does not satisfy the linearity requirement, P2, of a true inner product.)

The rotationally dependent term can be expressed in terms of *Wigner D matrices* [42] $\mathsf{D}_l = \mathsf{D}_l(\mathbf{R})$

$$\langle \rho_A, \mathbf{R}\rho_B \rangle = \sum_{l=0}^{\infty} \operatorname{tr} \mathsf{D}_l(\mathbf{R})\mathsf{J}_l \approx \underbrace{\operatorname{tr} \mathsf{D}_0(\mathbf{R})\mathsf{J}_0}_{\mathsf{J}_0} + \operatorname{tr} \mathsf{D}_1(\mathbf{R})\mathsf{J}_1 \tag{10}$$

and corresponding rotationally independent terms $\mathsf{J}_l$ based on the representation of (smeared) atomic densities in SOAP which we use as a convenient decomposition of the Gaussian overlap Eq. (8), see App. 5 for more details. Since $\mathsf{D}_0 = 1$ is constant and the dimensions of $\mathsf{D}_1$ are the same as $\mathbf{R}$ albeit being complex valued, we can find the optimal rotation $\mathbf{R}$ that

[e] The smearing $\sigma$ could be assigned in a per-atom/element fashion for greater specificity.

maximizes the inner product $\langle \rho_A, \mathbf{R}\rho_B \rangle$

$$\text{argmax}_{\mathbf{R}} \langle \rho_A, \mathbf{R}\rho_B \rangle = \text{argmax}_{\mathbf{R}} \sum_{l=0}^{p} \text{tr}\, \mathsf{J}_p \mathsf{D}_p(\mathbf{R}) \approx \text{argmax}_{\mathbf{R}}\, \text{tr}\, \mathsf{J}_1 \mathsf{D}_1(\mathbf{R}) \tag{11}$$

using an SVD of the matrix $\mathsf{J}_1$ in a manner analogous the solution for the optimal rotation of the RMS-D metric. First, we apply the mapping $\mathsf{D}_1 = \mathsf{A}^\dagger \mathbf{R} \mathsf{A}$ from Ref. 42 (Eq. 3.63):

$$\text{tr}\, \mathsf{J}_1 \mathsf{D}_1 = \text{tr}\, \mathsf{J}_1 \mathsf{A}^\dagger \mathbf{R} \mathsf{A} = \text{tr} \underbrace{\mathsf{A} \mathsf{J}_1 \mathsf{A}^\dagger}_{\mathbf{U}\mathbf{S}\mathbf{V}^T} \mathbf{R} \tag{12}$$

where the transformation matrix $\mathsf{A} : \mathsf{D} \rightarrow \mathbf{R} = \mathsf{A}\mathsf{D}\mathsf{A}^*$ is constant and unitary. Hence, the optimal (real-valued) rotation is $\mathbf{R} = \mathbf{V}\mathbf{U}^T$. Unfortunately the truncated sum in Eq. (11) does not encode all the configurational information, and hence we use it as an initial guess to a Newton solver based on the derivative

$$\frac{\partial}{\partial \mathbf{R}} d_{\text{OGTO}}(A, B) = -\frac{2\sigma\sqrt{\pi}^3}{d_{\text{OGTO}}(A, B)} \sum_{\alpha \in \mathcal{N}_A, \beta \in \mathcal{N}_B} \exp\left(-\frac{r_{\alpha\beta'}^2}{4\sigma^2}\right) \mathbf{x}_\alpha \otimes \mathbf{x}_\beta \tag{13}$$

Given the fact that the interatomic distances $r_{\alpha\beta}$ appear in an exponential function in Eq. (8), this product does not satisfy the linear property (P2) of true inner products. Nevertheless, Eq. (9) appears to satisfy the properties of a metric (M1-5). As an illustration, Fig. 3 shows, for a range of values of $\sigma$ on order of atomic spacing, that both the OGTO and RMS-D metrics satisfy the triangle inequality (M4) for the distances between a two arbitrary configurations $\mathcal{C}_1$, $\mathcal{C}_2$, and a third $\mathcal{C}_3$ which is interpolation of the two: $\mathsf{X}_3 = \alpha \mathsf{X}_1 + (1-\alpha)\mathsf{X}_2$. The RMS-D metric shows the expected linearity with $\alpha$: $d_{13} = (1-\alpha)d_{12}$ and $d_{23} = \alpha\, d_{12}$ [f], whereas the OGTO distances only regresses to this behavior with larger $\sigma$. Unfortunately, this asymptotic satisfaction of condition P2 comes with decreasing discrimination of close atom pairs, since all $r_{\alpha\beta}^2$ become weighed equally as $\sigma$ is increased. Nevertheless, for any given atomic system a range of suitable $\sigma$ is easy to find.

[f] For RMS-D, the fact that $r_{13}^2 = ||\mathbf{x}_1 - (\alpha\mathbf{x}_1 + (1-\alpha)\mathbf{x}_2)||^2 = (1-\alpha)^2 r_{12}^2$ and $r_{23}^2 = ||\mathbf{x}_2 - (\alpha\mathbf{x}_1 + (1-$

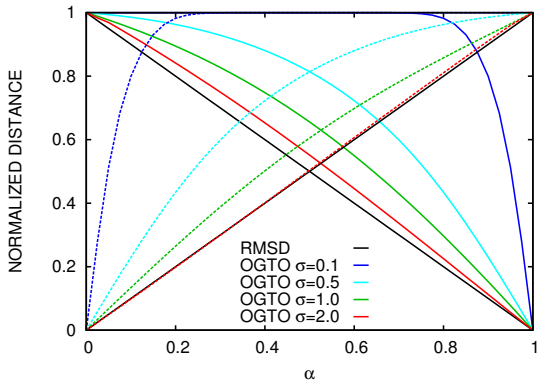



Figure 3: Demonstration of the triangle inequality (M4) and illustration of OGTO distance sensitivity to width of smearing function $\sigma$. Distances shown are from $\mathcal{C}_3 : \mathsf{X}_3 = \alpha\mathsf{X}_1 + (1-\alpha)\mathsf{X}_2$ to $\mathcal{C}_1$ (dashed) and $\mathcal{C}_2$ (solid).

## 2.2 Interpolation

Given a viable metric $d_{QA} = d(Q, A)$, all the neighborhoods $\mathcal{B}_Q = \{A \mid d_{QA} \leq R\}$ within a ball $\mathcal{B}_Q$ around a query neighborhood $\mathcal{C}_Q$ can be found (via the search algorithm described in Sec. 3.2). The forces on the central atoms of these nearby neighborhoods can be interpolated. Radial basis functions (RBF)[43] are a general class of functions amenable to this task. We choose the inverse quadratic form

$$\phi(d) = \frac{1}{1 + (\varsigma d)^2} \tag{14}$$

motivated by the expected force decay in real space and choose the parameter $\varsigma \ll 1/R$ based on the size of the trust region $\mathcal{B}_Q$. Then, to construct the interpolation solely based on the inter-cluster distances $d_{QB}$ and the neighboring forces $\mathbf{f}_B$, we solve the consistency equations

$$\mathbf{R}_{QA}\mathbf{f}_A = \sum_{B \in \mathcal{B}_Q} \phi(d_{AB})\, \mathbf{a}_B \text{ for every } A \in \mathcal{B}_Q \tag{15}$$

for the coefficients $\{\mathbf{a}_B\}$. Here, the matrix $\phi(d_{AB})$ is symmetric and positive definite (where $d_{AB}$ is the edge-weighted *adjacency* matrix of the sub-graph in the ball around $Q$ in cluster

$\alpha)\mathbf{x}_2)||^2 = \alpha^2 r_{12}^2$ for every atom pair leads to the linearity shown.

space), and $\mathbf{R}_{QA}$ is the optimal rotation from reference neighborhood $A$ to query neighborhood $Q$ which aligns the neighboring forces $\mathbf{f}_A$ to the orientation of the cluster $Q$. The local interpolation of the databased HF forces

$$\mathbf{f}_Q = \sum_{B \in \mathcal{B}_Q} \phi(d_{QB})\, \mathbf{a}_B \tag{16}$$

results.

# 3 Method

Beyond the basic theoretical developments there are a few more pragmatic components of the overall paradigm. Specifically, how to construct a viable dataset and an efficient search.

## 3.1 Database generation

The database is in effect a discrete representation/sampling of neighborhood configurational space as a graph with edges weighted by cluster-cluster distances. Specifically, given a set of $M$ neighborhoods $\{\mathcal{C}_A,\ A = 1, \dots M\}$, an atlas in the form of a complete edge-weighted graph: $\{d(A,B) \mid \forall A, B \in \{1, \dots M\}, A \neq B\}$, where the $^1/_2\, M(M-1)$ edges represent the inter-cluster distances, can be generated. For this work we generated samplings of cluster space by post-processing standard BO simulations of the dynamics of feasibly-sized periodic systems to extract HF forces and atomic neighborhoods [g].

Since a naïve graph that includes all possible edges would grow like the square of the number of stored clusters, we chose an $n$-level hierarchical representation where each level spans/covers a sub-graph, with the top-most level being the whole graph. This representation, which maintains a fixed bandwidth adjacency matrix, is constructed via an algorithm based on $k$-medoids clustering:[44]

[g] Alternately, we also generated configurations with (a) a random walk, via a Monte Carlo algorithm, and (b) a classical MD surrogate of the system of interest.

1. Group graph containing $\{\mathcal{C}_A\}$ using $k$-means based on distance

   (a) initialize by randomly picking $k$ candidate medoids $\{\mathcal{C}_{A_i},\ i = 1, k\}$

   (b) assign each $\mathcal{C}_B$ to group $\mathcal{G}_i^n$ based on closest medoid $\mathcal{C}_{A_i} = \underset{A_i \in \mathcal{G}_i^n}{\operatorname{argmin}}\, d(A_i, B)$

   (c) update medoids $\{\mathcal{C}_{A_i}, i = 1, k\}$ by $A_i = \underset{A_i \in \mathcal{G}_i}{\operatorname{argmin}} \sum_{B \in \mathcal{G}_i^n} d(A_i, B)$

   (d) repeat (b) and (c) until groups $\{\mathcal{G}_i^n,\ i = 1, k\}$ do not change

2. For each group $\mathcal{G}_i$, select medoid $A_i \in \mathcal{G}_i^n$ as the representative reference for this sub-graph at this level.

3. Repeat for each sub-graph containing $\mathcal{G}_i^n$ until each sub-graph of the current level $n$ is of most size $k$.

Note that each level is fully connected, *i.e.* all edges for the sub-graph are stored, so that metric search described in the next section will converge. Also, at the lowest level each neighborhood $\mathcal{C}_A$, in addition to being connected to all other neighborhoods on its level, is connected to all close neighborhoods within a radius twice that of the interpolation trust ball $\mathcal{B}_A$ to ensure the force interpolation described in Sec. 2.2 captures all local clusters. So, at the lowest level the groupings $\mathcal{G}_i$ form an overlapping cover of the whole set.

In general, the chemical composition of the neighborhoods segregate the database into separate/disjoint spaces and hence herein we treat a single species for simplicity. It should be noted that the configurational space visited by dynamics will typically be bounded by energy or temperature and hence the database samples will be dense in some region of configuration space. Also, the symmetries I1-I3 will tend to compact the database into a sub-manifold of the basic $\mathbb{R}^{3N}$ cluster space by creating equivalence classes,[41] *e.g.* setting the central atom at the origin removes the translational ambiguity, the optimal rotation reduces the dimensionality by another 3 dimensions, and optimal permutations condense segregated, equivalent high symmetry neighborhoods into dense groupings. Ultimately, the density of configuration-force samples determines the accuracy of the force interpolation that drives the

dynamics. Appropriate adaption alleviates concerns about having an dense sampling from the outset but here we treat the task of pre-computing a static database (one that could be used as an initial database for an adapting method). More ideas regarding this issue will be discussed in the concluding section.

## 3.2 Metric-based search

Recent work[26] has shown that an atlas of precomputed distances can accelerate molecular structure searches by orders of magnitude versus a simple exhaustive search of a configurational database. The associated algorithm relies on the reverse triangle inequality (M5) which implies that the distances between two pairs $(Q, A)$ and $(A, B)$ with an element $A$ in common puts bounds on the distance $d(Q, B)$. In fact, given $d(A, B) > 2d(Q, A)$ and M5, $d(Q, B) \geq |d(A, B) - d(Q, A)|$, then $d(Q, B) > |2d(Q, A) - d(Q, A)| = d(Q, A)$. If $Q$ is a query cluster this fact allows the search to omit all clusters $\{B \mid d(A, B) > 2d(Q, A)\}$, using the distances $d(A, B)$ stored in the database and the newly computed distance $d(Q, A)$.

A version of this metric-based search algorithm adapted from Ref. 26 to our hierarchical database is:

1. Given a query cluster $Q$, initialize $d_{\min} = d(Q, A_{\min})$ where $A_{\min}$ is the first element in $\mathcal{G}^0$, the set of all configurations in the highest level $n = 0$
2. For each level $n$ find $A^n_{\min} = \underset{A \in \mathcal{G}^n_i}{\operatorname{argmin}}\, d(Q, A)$ via iteration over $A \in \mathcal{G} = \mathcal{G}^n_i$:
   (a) compute $d = d(Q, A)$
   (b) if $d < d_{\min}$, set $A_{\min} = A$, $d_{\min} = d$ and, using the sorted distances stored in the database, update $\mathcal{G} = \{B \,|\, d(A_{\min}, B) < 2d_{\min}\}$
   (c) if $\mathcal{G}$ contains only $A_{\min}$, move to the next lowest level.
3. If at the lowest level, search over the region of radius $2R$ around $A_{\min}$ to obtain $\mathcal{B} = \{A_j \mid d(Q, A_j) < R\}$
4. Store set $\mathcal{B}$ for subsequent searches for the atom associated with $Q$.

and is illustrated in Fig. 4. A proof of convergence to the optimal $A_{\min}$ and the group around it can be constructed along the lines of convergence of sequence in a complete metric space: the graph is complete and the candidate $A_{\min}$ moves closer to the optimal $A_{\min}$ with iteration. With regard to the last step of the algorithm, unlike the task of matching molecules and chemical compounds by structure, with dynamics we have history and continuity in time and the assumption that in a small time-step $_{\triangle}t$ the particular cluster of interest has not moved very far through cluster space. So at time $t + _{\triangle}t$ rather than searching over the entire space of neighborhoods, we will begin the search in the ball about the best match for the previous time $t$.

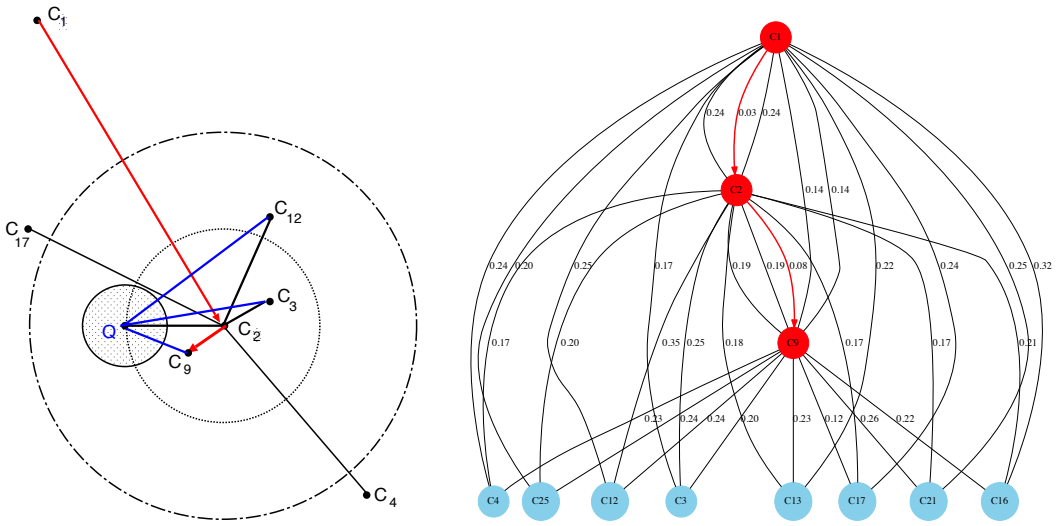



Figure 4: Metric search algorithm: (a) Illustration of metric search for query point Q. Given a current closest cluster, say, $\mathcal{C}_2$, the subset $\{\, \mathcal{C}_3, \mathcal{C}_9, \mathcal{C}_{12} \,\}$ near $\mathcal{C}_2$ are intermediate candidates for inclusion in the (shaded) interpolation ball around Q; whereas, the other neighbors, $\mathcal{C}_i$, of $\mathcal{C}_2$, $\mathcal{C}_3$ $\mathcal{C}_9$, can be ignored since $d(\mathcal{C}_i, \mathcal{C}_2) > 2d(Q, \mathcal{C}_2)$. After direct computation of distances from the candidate subset to Q in the next search iteration, $\mathcal{C}_9$ will be selected as the new closest configuration to Q. The distances that need to be computed are shown in blue, the distances that are retrieved from the database are in black and the search path connecting successive closest stored configurations is shown in red. (b) The corresponding search path through a portion of the database visualized as a graph.

# 4 Results

All the following results use two samples of the *ab initio* dynamics of a Si system, one crystalline at 300 K and the other amorphous at 2500K. The sampled system is periodic with 216 atoms ($4^3$ unit cells) and sampled for 200-1000 steps of 1 fs with the first 100 steps discarded since the starting state was a perfect crystal. Given that the first shell of the diamond cubic structure has 4 neighbors and the second 12 neighbors, we choose clusters with 4 and 16 neighbors.

As a preliminary to examining questions Q2-4, we examined sensitivities to algorithmic parameters. Given the results shown in Fig. 3, for all the following studies we used the OGTO smearing parameter $\sigma = 1$Å in $d_{\mathrm{OGTO}}$. For the scale parameter $\varsigma$ of the RBF used in the force interpolation Eq. (14) we used $\varsigma = 1$Å so that the RBF is sufficiently peaked in the interpolation ball (whose radius is on the order 2-4 Å). Also, since $\lambda$-sort modification of the RMS-D metric is not a perfect solution to the permutation optimization we only used the less computationally expensive $r$-sort in our demonstrations.

Here we also examine the behavior of the two chosen metrics over the database of clusters provided by the BO trajectories. Fig. 5 is an illustration that the BO dynamics path in cluster space is continuous. The bell shaped distributions of cluster RMS-D distances relative to a perfect lattice configuration broaden and shift to greater distances with increasing temperatures, as expected. The increase in number of neighbors from 4 to 16 increases the specificity of the metric as evidenced by the relatively sharper distributions for distances based on $N = 16$. Fig. 6a shows the characteristic radial distributions for a crystalline material at 300K and an amorphous material at 2500K. In Fig. 6b we compare the distribution of distances as a function of temperature and number of neighbors for both the RMS-D and OGTO measures for the whole database of configuration samples (as opposed to following the configuration associated with one atom through time, as in the inset of Fig. 5). Although quantitatively different, the distributions generated by both metrics show peaks that are closer to zero distance for colder systems and fewer neighboring atoms in the cluster

comparisons.

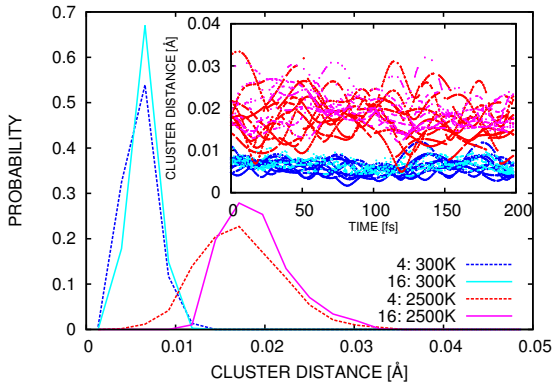
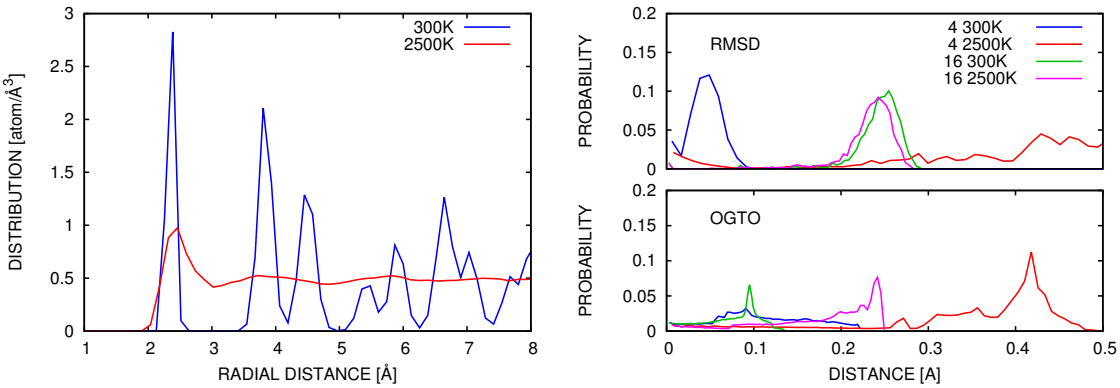



Figure 5: Dynamics path (inset) and associated distance distribution of a single neighborhood through cluster space, where the (RMS-D) distance is measured from a perfect crystalline configuration.




(a) radial distribution (b) histogram of inter-cluster distances

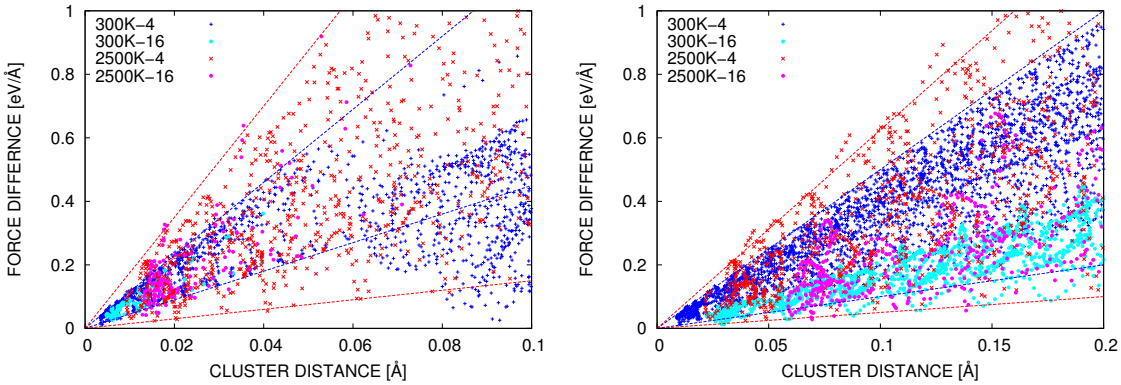
Figure 6: Spatial structure compared to cluster space structure. (a) the radial distribution of Si at 300 K (crystalline) and 2500 K (melt). (b) the distribution of pair-distances in the corresponding cluster space.

## 4.1 Force-distance correlation

In this section we address Q2 posed in the introduction: one of the fundamental assumptions of the method is that the cluster distance between a pair of configurations is correlated with the similarity in the forces

$$\|\mathbf{f}_A - \mathbf{R}_{AB}\mathbf{f}_B\| \propto d_{AB} \tag{17}$$

The force error *vs.* cluster distance correlation, Fig. 7, clearly shows that the HF forces and cluster distances are correlated and in fact that as $d \to 0$ the forces become effectively identical up to a rotation $\mathbf{R}_{AB}$. Also, as expected, using a larger number of atomic neighbors in a given cluster is more discriminating in terms of force errors and there is more scatter in the higher temperature results. Lastly, it appears that the OGTO samples are denser and more linearly correlated with distance than those of the RMS-D over the same range of force error. In part this is likely due to the permutation mismatches in our implementation of RMS-D which leads to a density of low force error samples appearing at large cluster-cluster distances (not shown in Fig. 7a).




(a) RMS-D (b) OGTO

Figure 7: Force error-cluster distance correlation for (a) RMS-D and (b) OGTO. ( Note the RMS-D and OGTO distances for a pair of clusters are not equal in general and so the data was plotted for a chosen range of force error.)

## 4.2 Search performance

In answer to question Q3 regarding search efficiency: Fig. 8 shows that the efficiency, as measured by the number of new distance evaluations of the query cluster to a member of the database relative to the size of the database, can be 1:100 or better depending on the size of the database. Also the efficiency appears to improve with increasing number of clusters in the hierarchical database, although clearly there is a limit to this improvement. Note that the results shown are an average over all the random parameters: starting points for the

search, seeds for the database grouping/clustering algorithm, *etc.* and the error bars shown are a result of these variances. Also note that we used the RMS-D metric, given that it is worse case and cheaper to compute; results for OGTO based searches are essentially the same. It also worth mentioning that with a complete graph the metric search is robust, *i.e.* it finds the closest cluster/s in the database without fail.

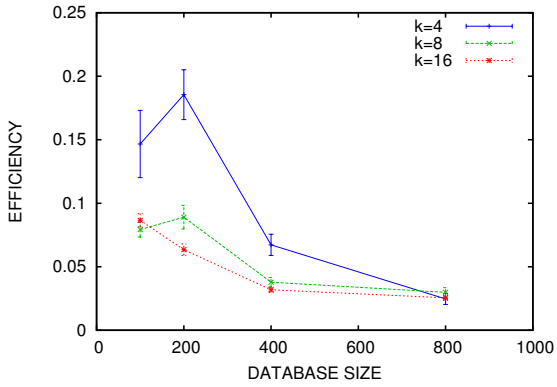



Figure 8: Search efficiency as function of the number of groups $k$ in any given level the hierarchical graph representation of the database, showing increased efficiency with appropriate selection of $k$. Efficiency is measured by number of new distance computations relative to the size of the database.

## 4.3 Interpolation convergence

Regarding question Q4, Fig. 9 shows force convergence with the size of the interpolation ball in Eq. (15). The volume of this ball roughly correlates with the number of neighboring clusters used in the RBF interpolation (15) given the fixed cluster database sampling density. Up to a certain radius, the interpolation converges in accuracy but the improvement eventually decreases with additional, more distant cluster neighbors. The convergence is strongly dependent on the sampling density of the database, *i.e.* the force error is strongly proportional to the distance to the closest sample (refer to Fig. 7 which indicates an upper bound of the convergence with the distance to the closest sample and hence to the density of samples in the database). In a practical sense, given a static/non-adapting database with a fixed sampling density, Fig. 9 show the limits to which the interpolation behavior can be

controlled by choosing the interpolation ball radius $R$ and the associated decay parameter $\varsigma$.

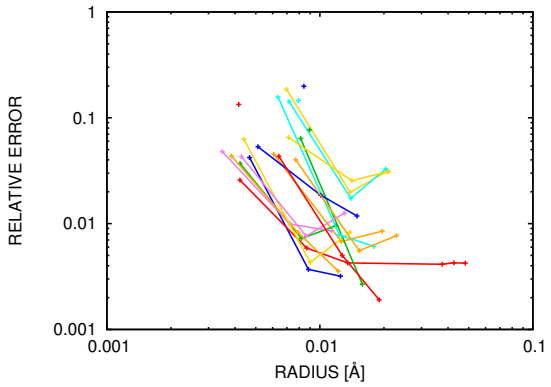



Figure 9: Force interpolation accuracy as a function of radius of the interpolation ball $\mathcal{B}_Q$ (which correlates with the number of neighboring clusters used in the interpolation). Each curve is generated by a different configuration and the error is highly correlated with the distance to the closest cluster in the database.

## 4.4 Consistency of dynamics

To test the assembled algorithm and determine the growth rate of the accumulated errors, we created a cluster database from a 100 step (100 fs) *ab initio* MD run on a 300 K, 2×2×2 unit cell Si system and used this to simulate the dynamics of a $3\times3\times3$ system for 200 steps. In this demonstration no new DFT samples were generated and hence a relatively large interpolation ball was needed (radius 0.02 Å, in the OGTO metric). Fig. 10 shows the temperature and momentum errors relative to a standard *ab initio* MD run of the $3\times3\times3$ system as a function of time. Clearly, the temperature error is increasing (linearly) but the momentum error seems to be oscillatory but bounded. This apparently unbounded temperature error could be controlled with a thermostat and both errors could be reduced with a denser sampling of cluster space (or an adaptive sampling).

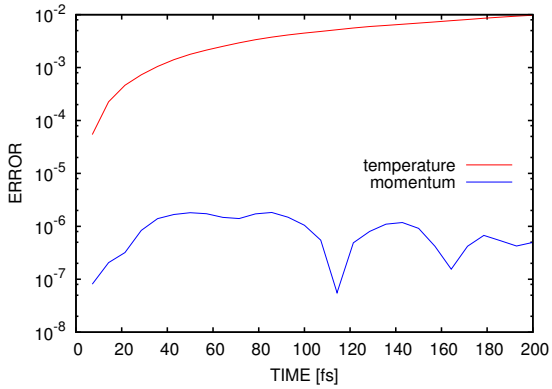



Figure 10: Dynamics consistency: error relative to a full *ab initio* dynamics. Temperature error normalized by $T = 300$ K, and momentum error by the average thermal velocity times the mass of the system.

# 5 Discussion

We have set the foundations of a method to simulate molecular dynamics with the accuracy of *ab initio* calculations and on the scale of classical molecular dynamics. The basis of this development is the connection of the local arrangement of atoms to the HF forces on individual atoms. Certainly there are materials that violate this condition, *e.g.* those with long range Coulomb interactions and/or extended electronic orbitals, but a large class of important materials are amenable to this treatment and, as mentioned, this connection is the basis for the widely-employed classical molecular dynamics method. In particular, the proposed method may be effective in simulating metals given the success of the local, empirical embedded atom potential.[45]

In this work we have focussed on the fundamental problem of mapping a local atomic arrangement/configuration to the HF force on central atom. The basic issue is developing a (I1) translationally, (I2) rotationally, (I3) permutationally invariant method that can be used to drive dynamics or search a database. We have developed two such methods, one based on the well-known RMS-D metric and the other developed from the OGTO and SOAP similarity measures. Each has relative advantages. RMS-D is cheap to compute including the optimal rotation but finding the optimal permutation remains an issue. The OGTO metric,

on the other hand, is intrinsically permutationally invariant and has the expected decreasing sensitivity to far away atoms in the local neighborhood at the expense of a more difficult determination of the optimal rotation. In addition to the two metrics and a version of metric search we provide an accurate means of interpolating forces in neighborhood space. We have shown: (Q1) the central atom force in a cluster converges to the full system HF force with increasing cluster size, (Q2) the cluster-cluster distances correlate well with the atom forces, (Q3) a metric search is orders of magnitude faster than a brute-force query, and (Q4) with interpolation the error in forces retrieved from the database and the optimal density of the database can be controlled. These developments allow us to treat the presently infeasible problem of simulating large compositionally and configurationally complex materials with *ab initio* accuracy - avoiding the tedious task of constructing a globally applicable empirical potential. It should be clear that the proposed algorithm reduces to *ab initio* molecular dynamics in limit of large neighborhood size and no interpolation.

We leave a number of topics and extensions for future work. First, although we have some confidence that momentum conservation will be enforced to the degree of error in the HF forces, we will need a large-scale implementation of the algorithm for a complete test of physical momentum and energy conservation with dynamics. In finite temperature applications we foresee that the conservation properties will be obscured and benefited by thermostat control schemes. Second, we would like to investigate how to construct a comparison metric appropriate for multi-element materials, which are very challenging to represent in classical formulations of interatomic potentials. Also, we see that there may be advantages to alternate means of representing cluster space, *e.g.* using the eigen-basis of the cluster graph Laplacian as a basis to embed the cluster manifold in a vector space and drawing upon ideas developed in the context of locality sensitive hashing[46] to improve the cluster search and force interpolation. In addition, in following work we will focus on database structure and adaptation. Adaptation will be driven by the goal of obtaining optimal density in region of configuration space occupied by the dynamics by pruning configuration samples irrele-

vant to current state and those in regions that are overly dense that reduce efficiency and adding samples necessary to maintain accuracy. The basic issue of generating new HF force samples may turn into a significant subsidiary problem in and of itself. Beyond adapting the database, there are also significant issues in scaling the algorithm to supercomputers, primarily distributing the database across a parallel file-system and managing it in limited memory.


## Acknowledgement

We appreciate helpful discussions with Aidan Thompson, Kevin Young, Ali Pinar, Jeremy Templeton and Peter Schultz (Sandia), as well as funding from Sandia Laboratories. Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.


# Appendix: Decomposition of the Gaussian overlap product

A spherical harmonic expansion of the Gaussian-smeared point densities Eq. (6), after dropping the normalization $1/(\sqrt{2\pi}\sigma)^3$ for convenience, can been written:

$$\rho_A(\mathbf{x}) = \sum_{\alpha \in \mathcal{N}_A} \exp\left(-\frac{\|\mathbf{x}-\mathbf{x}_\alpha\|^2}{4\sigma^2}\right) = \sum_{\alpha \in \mathcal{N}_A} \left[\sum_{l=0}^{\infty} \sum_{m=-l}^{l} c_\alpha^{lm}(r) Y_{lm}(\mathbf{r})\right], \tag{18}$$

in terms of the spherical harmonic basis $Y_{lm}$ and the coefficients $c_\alpha^{lm}$ containing the corresponding radial series in terms of modified spherical Bessel functions of the 1st kind, $\imath_l$ :

$$c_\alpha^{lm}(r) = 4\pi \exp\left(-\frac{r^2+r_\alpha^2}{2\sigma^2}\right) \imath_l\left(\frac{r r_\alpha}{\sigma^2}\right) Y_{lm}^*(\mathbf{r}_\alpha) \tag{19}$$

With this expansion we can reformulate the product Eq. (8) and associated norm Eq. (9)

$$\begin{aligned}
\langle \rho_A, \rho_B \rangle &= \int \mathrm{d}\mathbf{x} \rho_A(\mathbf{x}) \rho_B(\mathbf{x}) \\
&= 4\pi \sum_{\alpha \in \mathcal{N}_A} \sum_{\beta \in \mathcal{N}_B} \exp\left(-\frac{\sigma}{2}(r_\alpha^2 + r_\beta^2)\right) \sum_{l=0}^{\infty} \iota_l(\sigma r_\alpha r_\beta) \sum_{m=-l}^{l} Y_{lm}(\mathbf{r}_\alpha) Y_{lm}^*(\mathbf{r}_\beta) \\
&= 4\pi \sum_{l=0}^{\infty} \left[ \sum_{\alpha \in \mathcal{N}_A} \sum_{\beta \in \mathcal{N}_B} \sum_{m=-l}^{l} Y_{lm}(\mathbf{r}_\alpha) \exp\left(-\frac{\sigma}{2} r_\alpha^2\right) \iota_l(\sigma r_\alpha r_\beta) \exp\left(-\frac{\sigma}{2} r_\beta^2\right) Y_{lm}^*(\mathbf{r}_\beta) \right] \\
&= \sum_{l=0}^{\infty} \mathrm{tr} \underbrace{\left[ \mathsf{Y}_A \mathsf{W}_{AB} \mathsf{Y}_B^\dagger \right]_l}_{\mathsf{J}_l}
\end{aligned} \tag{20}$$

with matrices $[\mathsf{Y}_l]_{\alpha m} = Y_{lm}(\mathbf{r}_\alpha)$ and $[\mathsf{W}_l]_{\alpha\beta} = 4\pi \exp\left(-\frac{\sigma}{2}(r_\alpha^2)\right) \iota_l(\sigma r_\alpha r_\beta) \exp\left(-\frac{\sigma}{2}(r_\beta^2)\right)$.

The rotation of $\mathcal{C}_B$ relative to $\mathcal{C}_A$ in the inner product

$$\begin{aligned}
\langle \rho_A, \mathbf{R}\rho_B \rangle &= \sum_{l=0}^{p} \sum_{m_A, m_B = -l}^{l} D_{l, m_A m_B}(\mathbf{R}) \int \mathrm{d}r \sum_{\beta_A, \beta_B} c_{\beta_A}^{l, m_A}(r) c_{\beta_B}^{l, m_B}(r) \int \mathrm{d}\mathbf{r}\, Y_{l, m_A}^*(\mathbf{r}) Y_{l, m_B}(\mathbf{r}) \\
&= \sum_{l=0}^{p} \mathrm{tr}\, \mathsf{D}_l(\mathbf{R}) \mathsf{J}_l \ .
\end{aligned} \tag{21}$$

can be expressed in terms of *Wigner D* matrices $\mathsf{D}_l = \mathsf{D}_l(\mathbf{R})$ ($\mathsf{D}_l(\mathbf{I})$ is $2l+1$ square identity matrix) An operational definition of the (unitary) Wigner D matrices is $\mathsf{Y}_l(\mathbf{R}\mathbf{r}) = \mathsf{D}_l(\mathbf{R})\mathsf{Y}_l(\mathbf{r})$. Since $\mathsf{D}_0 = 1$ is constant and the dimensions of $\mathsf{D}_1$ are the same as $\mathbf{R}$ albeit being complex valued, we can find the optimal rotation $\mathbf{R}$ that maximizes the similarity $\langle \rho_A, \mathbf{R}\rho_B \rangle$ via

$$\mathrm{argmax}_{\mathbf{R}} \langle \rho_A, \mathbf{R}\rho_B \rangle = \mathrm{argmax}_{\mathbf{R}} \sum_{l=0}^{p} \mathrm{tr}\, \mathsf{J}_p \mathsf{D}_p(\mathbf{R}) \approx \mathrm{argmax}_{\mathbf{R}}\, \mathrm{tr}\, \mathsf{J}_1 \mathsf{D}_1(\mathbf{R}) \tag{22}$$

First, we apply the mapping $\mathsf{D}_1 = \mathsf{A}^\dagger \mathbf{R} \mathsf{A}$ from Eq. 3.63 in Ref. 42:

$$\mathrm{tr}\, \mathsf{J}_1 \mathsf{D}_1 = \mathrm{tr}\, \mathsf{J}_1 \mathsf{A}^\dagger \mathbf{R} \mathsf{A} = \mathrm{tr} \underbrace{\mathsf{A} \mathsf{J}_1 \mathsf{A}^\dagger}_{\mathbf{U}\mathbf{S}\mathbf{V}^T} \mathbf{R} \tag{23}$$

where the transformation matrix $\mathsf{A} : \mathsf{D} \rightarrow \mathbf{R}$ is constant and unitary. (We use $\mathsf{A}^*$ to denote the complex conjugate of $\mathsf{A}$ and $\mathsf{A}^\dagger$ to denote the transpose of $\mathsf{A}^*$.) Hence, an approximate optimal (real-valued) rotation is $\mathbf{R} = \mathbf{V}\mathbf{U}^T$. Note, given the angle conventions in Biedenharn and Louck's text,[42] $\mathsf{A}$ has the components:

$$[\mathsf{A}] = \begin{bmatrix} -1/\sqrt{2} & 0 & 1/\sqrt{2} \\ -\imath/\sqrt{2} & 0 & -\imath/\sqrt{2} \\ 0 & 1 & 0 \end{bmatrix} \tag{24}$$

Also note given that components of $\mathsf{Y}_l$ can be negative implies product formed from truncated SOAP expansion is not metric.

## References


(1) Tersoff, J. *Physical Review Letters* **1988**, *61*, 2879.

(2) Tersoff, J. *Physical Review B* **1989**, *39*, 5566–5568.

(3) Pettifor, D. *Physical review letters* **1989**, *63*, 2480.

(4) Horsfield, A.; Bratkovsky, A.; Fearn, M.; Pettifor, D.; Aoki, M. *Physical Review B* **1996**, *53*, 12694.

(5) Brenner, D. W.; Shenderova, O. A.; Harrison, J. A.; Stuart, S. J.; Ni, B.; Sinnott, S. B. *Journal of Physics: Condensed Matter* **2002**, *14*, 783.

(6) Pettifor, D.; Finnis, M.; Nguyen-Manh, D.; Murdick, D.; Zhou, X.; Wadley, H. *Materials Science and Engineering: A* **2004**, *365*, 2–13.

(7) Bartók, A. P.; Payne, M. C.; Kondor, R.; Csányi, G. *Physical review letters* **2010**, *104*, 136403.

(8) Bartók, A. P.; Kondor, R.; Csányi, G. *Physical Review B* **2013**, *87*, 184115.

(9) Thompson, A.; Swiler, L.; Trott, C.; Foiles, S.; Tucker, G. *Journal of Computational Physics* **2014**,

(10) Martin, J. *Journal of Physics C: Solid State Physics* **1975**, *8*, 2858.

(11) Mazur, P.; Montroll, E. *Journal of mathematical physics* **1960**, *1*, 70–84.

(12) Saitô, N.; Ooyama, N.; Aizawa, Y.; Hirooka, H. *Progress of Theoretical Physics Supplement* **1970**, *45*, 209–230.

(13) Hoover, W. G. *Physical Review A* **1988**, *37*, 252.

(14) Ischtwan, J.; Collins, M. A. *The Journal of chemical physics* **1994**, *100*, 8080–8088.

(15) Gunaratne, G. H.; Jones, R. E.; Ouyang, Q.; Swinney, H. L. *Physical review letters* **1995**, *75*, 3281.

(16) Weinberger, K. Q.; Saul, L. K. *The Journal of Machine Learning Research* **2009**, *10*, 207–244.

(17) Pope, S. *Combustion Theory and Modelling* **1997**, *1*, 41–63.

(18) Arsenlis, A.; Barton, N.; Becker, R.; Rudd, R. *Computer methods in applied mechanics and engineering* **2006**, *196*, 1–13.

(19) Barton, N. R.; Knap, J.; Arsenlis, A.; Becker, R.; Hornung, R. D.; Jefferson, D. R. *International Journal of Plasticity* **2008**, *24*, 242 – 266.

(20) Sanchez, J. *Physical review B* **1993**, *48*, 14013.

(21) de Fontaine, D. *Alloy Phase Stability*; Springer, 1989; pp 177–203.

(22) Asta, M.; Wolverton, C.; De Fontaine, D.; Dreyssé, H. *Physical Review B* **1991**, *44*, 4907.

(23) Wolverton, C.; Asta, M.; Dreyssé, H.; De Fontaine, D. *Physical Review B* **1991**, *44*, 4914.

(24) Marx, D.; Hutter, J. *Modern methods and algorithms of quantum chemistry* **2000**, *1*, 301–449.

(25) Kühne, T. D. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **2013**,

(26) Fogolari, F.; Corazza, A.; Viglino, P.; Esposito, G. *Algorithms for Molecular Biology: AMB* **2012**, *7*, 16.

(27) Hung, L.-H.; Samudrala, R. *Bioinformatics* **2012**, *28*, 2191–2192.

(28) De, S.; Schaefer, B.; Sadeghi, A.; Sicher, M.; Kanhere, D.; Goedecker, S. *Physical Review Letters* **2014**, *112*, 083401.

(29) Levitt, M.; Gerstein, M. *Proceedings of the National Academy of sciences* **1998**, *95*, 5913–5920.

(30) Baudet, C.; Dias, Z.; Sagot, M.-F. *Algorithms for Molecular Biology* **2012**, *7*.

(31) Sadeghi, A.; Ghasemi, S. A.; Schaefer, B.; Mohr, S.; Lill, M. A.; Goedecker, S. *The Journal of chemical physics* **2013**, *139*, 184118.

(32) Drineas, P.; Javed, A.; Magdon-Ismail, M.; Pandurangant, G.; Virrankoski, R.; Savvides, A. Distance matrix reconstruction from incomplete distance information for sensor network localization. Sensor and Ad Hoc Communications and Networks, 2006. SECON'06. 2006 3rd Annual IEEE Communications Society on. 2006; pp 536–544.

(33) Carugo, O. *Protein Engineering Design and Selection* **2007**, *20*, 33–37.

(34) Lindorff-Larsen, K.; Ferkinghoff-Borg, J. *PLoS One* **2009**, *4*, e4203.

(35) Vishveshwara, S.; Brinda, K.; Kannan, N. *Journal of Theoretical and Computational Chemistry* **2002**, *1*, 187–211.

(36) Gutman, I.; Trinajstić, N. *Graph theory and molecular orbitals*; Springer, 1973.

(37) Kabsch, W. *Acta Crystallographica Section A* **1976**, *32*, 922–923.

(38) Wahba, G. *SIAM review* **1965**, *7*, 409–409.

(39) Liu, P.; Agrafiotis, D. K.; Theobald, D. L. *Journal of computational chemistry* **2010**, *31*, 1561–1563.

(40) Hong, E.-J.; Lee, K.-H.; Wenzel, W. *International Journal of Biology and Biomedical Engineering* **2007**, *1*, 46–48.

(41) Steipe, B. *Acta Crystallographica Section A: Foundations of Crystallography* **2002**, *58*, 506–506.

(42) Biedenharn, L.; Louck, J. D. *Angular Momentum in Quantum Physics: Theory and Application, Encyclopedia of Mathematics and its Applications*; Addison-Wesley, Englewood Cliffs, 1981.

(43) Buhmann, M. D. *Radial basis functions: theory and implementations*; Cambridge university press, 2003; Vol. 12.

(44) Kaufman, L.; Rousseeuw, P. In *Statistical Data Analysis based on L1 Norm*; Dodge, Y., Ed.; Elsevier/North-Holland, 1987; pp 405–416.

(45) Daw, M. S.; Baskes, M. I. *Physical Review B* **1984**, *29*, 6443.

(46) Leskovec, J.; Rajaraman, A.; Ullman, J. D. *Mining of massive datasets*; Cambridge University Press, 2014.