PHYSICAL REVIEW LETTERS

Molecular Dynamics with On-the-Fly Machine Learning of Quantum-Mechanical Forces

Zhenwei Li,[1,†] James R. Kermode,[1,2,*] and Alessandro De Vita[1,3]

[1]*King's College London, Physics Department, Strand, London WC2R 2LS, United Kingdom*
[2]*Warwick Centre for Predictive Modelling, School of Engineering, University of Warwick, Coventry CV4 7AL, United Kingdom*
[3]*CENMAT-UTS, Via Alfonso Valerio 2, 34127 Trieste, Italy*

We present a molecular dynamics scheme which combines first-principles and machine-learning (ML) techniques in a single information-efficient approach. Forces on atoms are either predicted by Bayesian inference or, if necessary, computed by on-the-fly quantum-mechanical (QM) calculations and added to a growing ML database, whose completeness is, thus, never required. As a result, the scheme is accurate and general, while progressively fewer QM calls are needed when a new chemical process is encountered for the second and subsequent times, as demonstrated by tests on crystalline and molten silicon.

The need to produce accurate dynamical representations of chemical processes has been ever increasing in recent years. Density-functional theory (DFT) provides a well established [1,2] framework to do this, notably through first-principles molecular dynamics (FPMD) simulations [3]. Standard DFT implementations are typically limited to a few hundred atoms (a few thousand in linear-scaling approaches [4–8]), and to the ~10 ps time scale, dramatically restricting the applicability of the approach. Much larger model systems and longer simulations are achievable using force fields (FFs), based on judiciously chosen parametrized functional forms fitted on experimental or FP data [9–14]. However, for many complex chemical situations accurate "reactive" FFs do not yet exist, would require long development times, and would be hard to validate systematically, particularly for use in truly predictive, extrapolative situations [15]. More recent approaches based on machine learning [16] use neural networks [17] or Gaussian processes (GPs)[18] to fit "once and for all" the DFT potential energy surface (PES), after which atomic forces are obtained by analytic differentiation. Similar to classical FFs, a fixed high-quality parametrized PES simultaneously ensures fast force evaluation and reliable interpolation. However, accuracy is still not guaranteed to transfer to chemical situations not represented in the fitting database.

Here, we propose an alternative machine-learning (ML) based scheme where we allow a stream of fresh quantum-mechanical (QM) calculations to augment the ML database during each MD simulation, enabling safe interpolation. The scheme could equally be viewed as an efficient FPMD approach where we seek to compute only the QM information necessary to progress the simulation, while retaining the very broad applicability of FPMD. To minimize the QM workload of the MD simulation, one can start by noticing that ML-predicted atomic forces will suffice as long as the dynamics visits configuration is "well represented" within the existing database. Thus, an ideal ML MD scheme should not attempt to increase its database through additional QM calculations until "something new" happens that necessitates this. This is a central guideline for the present work, significantly improving on an earlier scheme [19,20] where all QM information was used once and afterwards discarded. Our new scheme has the potential to reduce the cost of FPMD for the vast range of problems where it is already applied [2], and to extend its use to problems currently beyond reach because of prohibitive time and/or length scales. Here, we test it on standard benchmark physical systems [3,17,18,20], comprising crystalline and molten silicon over a wide range of temperatures and bonding geometries, in both insulating and metallic conditions (Figs. 1–3).
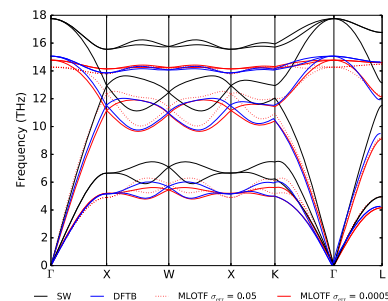

FIG. 1 (color online). Comparison of the bulk Si phonon spectrum calculated with DFTB (blue), and SW (black) and with the ML on-the-fly (MLOTF) approach (red), computed with the finite displacement Parlinksi-Li-Kawazoe method [21,22] using a standard $\sigma_{err} = 0.05$ eV/Å (dotted lines) and a high-accuracy $\sigma_{err} = 5 \times 10^{-4}$ eV/Å value (solid lines) for the ML data noise parameter. The ML database was constructed from a 300 K MD trajectory.
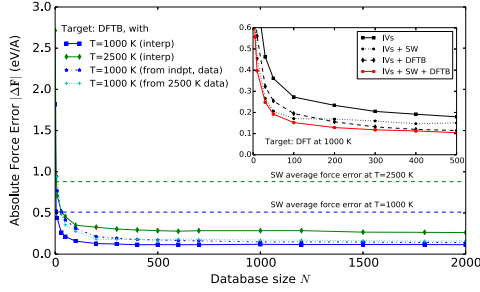
FIG. 2 (color online). Accuracy of forces predicted by the ML scheme as a function of database size. Teaching points are sampled from DFTB MD of silicon at 1000 K (blue squares) and at 2500 K (green diamonds) at 20 fs intervals. Inset: a similar test, using DFT forces sampled from MD at 1000 K as the target. Accuracy improves significantly when the set of IVs is augmented by classical or TB force vectors.

The scheme is based on direct ML prediction of atomic forces, rather than atomic energies or PESs. This ensures high force accuracy, e.g., allowing systematic convergence of trajectories to their FPMD limit by enhancing the QM fitting frequency, and avoiding any "blurring" effect connected with deriving forces from an intermediate PES representation. The forces on atoms are predicted by Bayesian inference using standard GP regression [16,23]. The QM database is progressively built during the MD run and, at the same time, queried to predict forces for those time steps where no QM calculation is made, using an n-time-step predictor-corrector algorithm [19,20]. As in the previous scheme, (free) energy barriers can be estimated by accurate thermodynamic integration [24]. Since the ongoing MD simulation continuously improves the database, as long as the system remains within the same phase space region, the accuracy of the predicted forces improves (Fig. 2). Alternatively, we may fix the target force accuracy so that the frequency of necessary QM calculations progressively decreases and n can be increased (Fig. 3).

To construct the covariance matrix required by the standard GP regression procedure [25], we need a symmetry-efficient representation to describe atomic configurations, and a function measuring the distance $d_{mn} = d(x_m, x_n)$ between any two such configurations x_m and x_n suitable for quantifying their "similarity for force prediction" [28]. As in PES-learning ML schemes, an efficient representation of an atomic environment x should be invariant under transformations to physically equivalent systems such as rotations and permutations of atoms of the same chemical species [29]. A special difficulty associated with a force-learning ML scheme is that the Cartesian force components depend on the choice of reference frame, unlike the (physically scalar, however defined) atomic energies, so that the best force components to be learned from a database configuration are only known after a rotation to its optimal alignment with the target configuration. As will be shown below, an efficient way to deal with this issue is to define a rotationally invariant "internal" representation for atomic
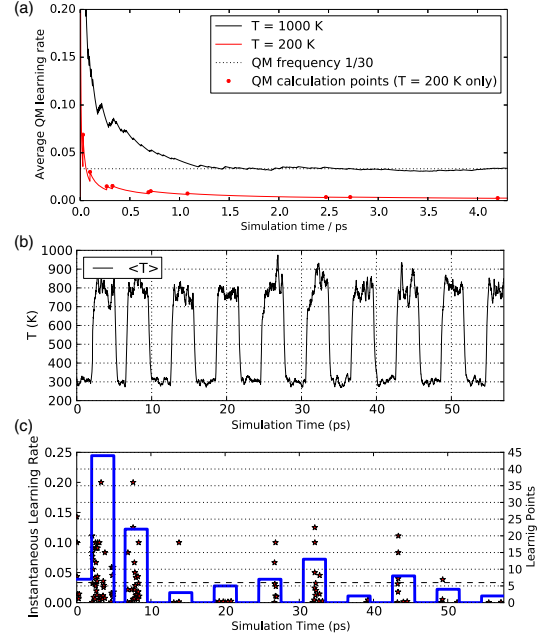

FIG. 3 (color online). (a) Average QM calling rate of low- and high-temperature MD "learning" simulations in bulk crystalline silicon. Red circles pinpoint QM calls, getting remarkably sparse after the initial learning phase. (b) Temperature profile of a MD simulation alternating between 300 and 800 K. (c) Instantaneous QM call frequency (left vertical axis, red stars) and total call count within each 800 K cycle (right axis, blue histograms) of the simulation of panel (b).

configurations and force vectors [30]. After carrying out ML in this representation, we transform the predicted force back into the Cartesian representation, so that they have the correct orientation for MD trajectory integration.

For each atom, k independent internal vectors (IVs) $\mathbf{V}_i$ for $i = 1, \ldots, k$ can be uniquely defined by the relative positions $\mathbf{r}_q$ of its neighbors, which makes them invariant under translations and any permutation of neighbors of the same chemical species. A possible choice is

$$\mathbf{V}_i = \sum_{q=1}^{N_{\text{neighb}}} \hat{\mathbf{r}}_q \exp\left[-\left(\frac{r_q}{r_{\text{cut}}(i)}\right)^{p(i)}\right], \qquad (1)$$

where each of these basis vectors is defined by different values of the parameters p and r_{cut}, chosen within a suitable range reflecting the decay rate or interaction range of forces in the system. This vector representation ensures that force components are null where this follows from symmetry, and that closer neighbors contribute more than far away ones. Crucially, to improve the prediction accuracy, this set can be expanded to include any additional vector presumed to carry useful information on target QM forces. These are typically force vectors obtained from well-established classical force fields or from QM models less computationally expensive than the current reference Hamiltonian (e.g., an empirical tight binding model if the main QM model is DFT based, see inset of Fig. 2). This offers a way to include precious

physical insight from established interatomic models in our descriptor. The directions $\hat{\mathbf{V}}_i = \mathbf{V}_i / \|\mathbf{V}_i\|$ form an internal coordinate system which, for $k > 3$, gives an overdetermined description of vector quantities. A representation for the atomic configuration x is given by the $k \times k$ matrix X with elements $X_{ij} = \mathbf{V}_i \cdot \hat{\mathbf{V}}_j = (VA^T)_{ij}$, where we have defined the vector and direction matrices

$$V^T = \begin{pmatrix} | & | & | \\ \mathbf{V}_1 & \dots & \mathbf{V}_k \\ | & | & | \end{pmatrix}, \quad A^T = \begin{pmatrix} | & | & | \\ \hat{\mathbf{V}}_1 & \dots & \hat{\mathbf{V}}_k \\ | & | & | \end{pmatrix}. \quad (2)$$

The feature matrix $X = VA^T$ is invariant under translations, permutations and rotations of the corresponding atomic configuration. The similarity between two atomic environments x_m and x_n is expressed by the distance

$$d_{mn}^2 = \frac{1}{k} \sum_{i,j=1}^k \left[\frac{X_{ij}^m}{\chi_i} - \frac{X_{ij}^n}{\chi_i} \right]^2, \quad (3)$$

where the scale factors χ_i are the standard deviation over the database of the Euclidean distance in the internal representation space between the $\mathbf{V}_i$ vectors from different configurations

$$\chi_i^2 = \sum_{n,m=1}^N \sum_{j=1}^k \frac{(X_{ij}^m - X_{ij}^n)^2}{N^2}. \quad (4)$$

This distance is used to build the covariance matrix [Supplemental Material [25] Eq. (S3)], which requires further tuning the correlation length and data noise hyperparameters σ_{cov} and σ_{err}. The former is always of the order of unity due to the χ_i distance normalization in Eq. (3), and $\sigma_{\mathrm{cov}} = 1.0$ is the value used in this work. For the latter, we typically impose $\sigma_{\mathrm{err}} = 0.05$ eV/Å, which has the effect of regularizing the linear algebra [31]. After predicting the components of the k-dimensional internal force vector $\mathcal{F}$ [Supplemental Material [25] Eq. (S1)], its Cartesian space coordinates can be reconstructed as $\mathbf{F} = A^+ \mathcal{F}$ using a least squares approach where the absolute orientation of the testing configuration is provided by the pseudoinverse matrix $A^+ = (A^T A)^{-1} A^T$.

As a first, stringent test of the robustness and accuracy of our matrix representation [Eq. (2)], we probed its performance in handling highly symmetric target configurations involving very small or null components of internal vectors and target forces, by computing the phonon dispersion of crystalline Si using the supercell approach [21]. The predicted phonon spectra closely track the target QM benchmark (Fig. 1), here obtained from a density-functional tight-binding (DFTB) Hamiltonian [32]. The spectrum obtained from the Stillinger-Weber (SW) classical potential [33] is also shown for comparison. In particular, using a standard $\sigma_{\mathrm{err}} = 0.05$ eV/Å data noise hyperparameter value already provides a reasonable approximation to the QM phonon spectrum, while a stricter $\sigma_{\mathrm{err}} = 5 \times 10^{-4}$ eV/Å value significantly

improves the precision, trading transferability for accuracy in a controlled fashion, appropriate to this particular application.

Next, we tested the accuracy of our approach for dynamics using a fixed database of 2000 configurations sampled at 20 fs intervals along an NVT 1000 K MD trajectory of a bulk silicon system, generated with the DFTB Hamiltonian. Testing sets containing up to 2000 configurations were sampled from the same trajectory. The testing points were chosen at the time midpoints between consecutive sampling points, to mimic the "locally worst" positions of a trajectory generated by a predictor-corrector scheme, where the force error typically reaches its maximum [34]. For each test point, we sort the database entries by the distance of Eq. (3). This makes the decreasing force error converge rapidly with the number N of entries used for force prediction, significantly improving the efficiency of the GP learning process, which involves an $O(N^3)$ inversion of the covariance matrix.

The results shown by the blue and green solid lines in Fig. 2 reveal an average force error of 0.1 eV/Å: around 6% of the average QM force magnitude of 1.8 eV/A. This "midpoint force error" is about twice the "time-averaged" error we would obtain when averaging over all the MD trajectory time steps, consistent with what is generally found in predictor-corrector schemes [34]. For comparison, the average force error of an SW classical potential fitted to reproduce the DFTB lattice parameter and bulk modulus is ~ 0.5 eV/Å, or almost 30% relative error. For the 2500 K Si liquid, metallic system (green lines in Fig. 2), our database configurations are more dispersed in phase space and the predicted forces are less accurate. However, a midpoint error below 0.25 eV/Å ($\sim 10\%$ relative error), can be obtained with the full 2000 configuration database, while the SW error remains large (~ 0.9 eV/Å, or 35% relative error). We carried out detailed tests of the effect of varying the sampling density and of using independent teaching and testing trajectories, finding that acceptable force accuracies can be achieved provided that ~ 1000 configurations or more are included (blue and cyan dashed lines in Fig. 2) [35].

Moving from an empirical QM Hamiltonian to target forces generated by standard DFT engines [36,37] does not qualitatively change the results obtained. The inset in Fig. 2 shows that the DFT forces for 1000 K bulk Si can be reproduced within a ~ 0.1 eV/Å midpoint error (relative error $\sim 6\%$) by using just the $N = 500$ most relevant configurations for each prediction. Here, the importance of including additional "relevant" vectors in the internal vector set can be clearly seen. Namely, the midpoint error decreases from 0.18 eV/Å (black squares in Fig. 2 inset) to 0.10 eV/Å when SW and DFTB force vectors are included in the IVs (red circles), which simply increases by two units the dimension k of all X matrices in the database. This suggests that force vectors computed from less accurate Hamiltonians, while typically differing from the DFT forces by much more than our target tolerance, still encode very valuable information for their prediction, which can be

successfully incorporated as an improved database metric structure produced by Eq. (3). Moreover, the initial "learning rate" (convergence rate of the curves in Fig. 2 inset) is systematically higher when the additional vectors are included, which helps to reduce the database sizes needed for any given target force error threshold. Preliminary work to investigate how robustly our approach can be expected to extend to more complex, multi-component systems such as crystalline SiC and amorphous SiO_2 is reported in the Supplemental Material [38].

Next, we test the scheme in MD simulations where the trajectory is generated by the ML forces. Rather than assuming that a sufficiently complete database is ever once and for all available, we allow the on-the-fly addition of newly computed CPU-intensive QM data to the database, within a predictor-corrector approach of variable stride (QM call frequency). This ensures that the corrector-loop forces are always effectively an interpolation of QM information relevant to the trajectory they generate, preventing any extrapolation accuracy issues, and enabling the use of rather strict target force error tolerances, typically 0.1 eV/A or lower. Figure 3(a) shows how the frequency of necessary QM calculations falls during dynamical simulations of a 64 atom bulk Si system, for two different temperatures. For the purpose of this test, QM forces are computed at all time steps but only incorporated in the database when a 0.09 eV/A threshold average deviation between predicted and QM-calculated forces at the end of the predictor stretch is exceeded. This is a safe criterion, as the "predictor error" is systematically larger than both the midpoint and time-averaged errors actually incurred along the final corrected trajectory in predictor-corrector schemes. Since the calculation starts with an empty database, database incorporation of QM results is initially needed with high frequency. The different "learning curves" in the figure reflect—and actually allow us to measure—the different complexity (needed database size) of the two systems for MD force learning. In general, how such complexity depends on the target accuracy will reflect the smoothness properties of the QM-force vector field in the system's phase space region of interest—that is, the underlying physical properties that enable ML prediction.

At 200 K (red solid line and points), after an initial 1 ps learning, long time intervals in excess of 1000 MD steps occur where no QM calculations are required. This suggests that the relatively small database accumulated in the first ps of dynamics encodes an essentially complete knowledge of the force repertoire of this system, for the given target error, making it a "low-complexity system" in the sense above. The observation that the necessary QM incorporation frequency approaches zero after this time simply reflects a lower rate of "chemically new" configurations occurring further along the trajectory [at 2.5 ps, 2.7 ps, and 4.2 ps in this test, cf. Fig. 3(a)]. The configuration space for the same system at 1000 K (black solid line) is more complex and force learning is

correspondingly slower. However, we find that after an initial 2 ps simulation time, the target predictor error can be met by carrying out QM calculations every 30 fs (black dotted line). Remarkably, this time interval is at least three times longer than that required by our previous approach without ML [39] set to the same target predictor error.

Figures 3(b) and 3(c) illustrate the overall learning process for a new MD simulation of the same system where a Langevin thermostat [40] is used to take the temperature from 300 to 800 K and back in ten identical 5 ps cycles. While the necessary QM database additions are quite frequent during the first cycle, their number (blue histogram) drops significantly in the next two cycles, as the 800 K system conditions are no more "new" to the learning algorithm when they are encountered for the second and third time. New database entries are still occasionally generated from here on, all within 800 K stretches, either in a sparse or in a time-correlated fashion (see, e.g., star symbols in the 4th and 5th cycles, respectively, measuring on the left axis the "instantaneous learning rate" defined as the inverse number of time steps between current and last entry).

We expect the present method to be particularly useful for the simulation of materials processes where complex but recurring chemical steps are encountered, which can be learned, while time-localized occurrences of previously unseen chemical bonding geometries cannot be ruled out but could be handled by an adaptive approach. Examples may include bond breaking and reforming during crack propagation, dislocation motion, or point-defect diffusion, tribochemical processes, repeated catalytic reaction steps, or atomic deposition or diffusion processes on surfaces. We note that the method cannot match purely classical potentials for accessible system sizes and speed to the extent that it keeps requiring high level "on-the-fly" calculations, albeit sparsely in space and time, nor can it improve on the underlying high level approach from which it is learning. On the other hand, the method is not limited to learning DFT forces: higher levels of reference theory such as quantum Monte Carlo or post-Hartree-Fock methods could equally well be used. Moreover, standard DFT forces could be substituted by higher level theory forces only in those systems and configuration space regions where improvement is clearly mandatory, the hybrid database so obtained optimally combining the two levels of theory [41]. In practical MD applications, monitoring the predictor error can be used to progressively lengthen, or if needed sometimes abruptly shorten, the predictor-corrector interval, as appropriate for the system investigated. Combining this with cautious use of an "internally estimated" error [Supplemental Material [25] Eq. (S2)] could be useful during system thermalization, or in more exploratory runs where a reduced accuracy will suffice or the issue of validation can be temporarily postponed. In all cases, every QM calculation performed on-the-fly during a dynamical simulation is automatically and permanently stored for later use in similar MD runs or across different projects. Scaling

---

[*]j.r.kermode@warwick.ac.uk

[†]Present address: Department of Chemistry, University of Basel, Klingelbergstrasse 80, CH-4056 Basel, Switzerland.

[1] M. C. Payne, M. P. Teter, D. C. Allan, T. Arias, and J. Joannopoulos, Rev. Mod. Phys. **64**, 1045 (1992).

[2] K. Burke, J. Chem. Phys. **136**, 150901 (2012).

[3] R. Car and M. Parrinello, Phys. Rev. Lett. **55**, 2471 (1985).

[4] E. Hernández, M. J. Gillan, and C. M. Goringe, Phys. Rev. B **55**, 13485 (1997).

[5] J. M. Soler, E. Artacho, J. D. Gale, A. García, J. Junquera, P. Ordejón, and D. Sánchez-Portal, J. Phys. Condens. Matter **14**, 2745 (2002).

[6] C.-K. Skylaris, P. D. Haynes, A. A. Mostofi, and M. C. Payne, J. Chem. Phys. **122**, 084119 (2005).

[7] D. R. Bowler and T. Miyazaki, Rep. Prog. Phys. **75**, 036503 (2012).

[8] N. Hine, P. Haynes, A. A. Mostofi, C.-K. Skylaris, and M. Payne, Comput. Phys. Commun. **180**, 1041 (2009).

[9] F. Ercolessi and J. B. Adams, Europhys. Lett. **26**, 583 (1994).

[10] M. Z. Bazant, E. Kaxiras, and J. F. Justo, Phys. Rev. B **56**, 8542 (1997).

[11] Y. Mishin, D. Farkas, M. J. Mehl, and D. A. Papaconstantopoulos, Phys. Rev. B **59**, 3393 (1999).

[12] A. C. T. van Duin, S. Dasgupta, F. Lorant, and W. A. Goddard, J. Phys. Chem. A **105**, 9396 (2001).

[13] P. Tangney and S. Scandolo, J. Chem. Phys. **117**, 8898 (2002).

[14] M. Finnis, Interatomic forces in condensed matter, in *Oxford Series on Materials Modelling* (Oxford University Press, Oxford, 2003).

[15] M. Payne, I. Stich, A. De Vita, M. Gillan, and L. Clarke, Faraday Discuss. **96**, 151 (1993).

[16] D. MacKay, *Information Theory, Inference and Learning Algorithms* (Cambridge University Press, Cambridge, England, 2003).

[17] J. Behler and M. Parrinello, Phys. Rev. Lett. **98**, 146401 (2007).

[18] A. P. Bartók, M. C. Payne, R. Kondor, and G. Csányi, Phys. Rev. Lett. **104**, 136403 (2010).

[19] A. De Vita and R. Car, Mater. Res. Soc. Symp. Proc. **491**, 473 (1997).

[20] G. Csányi, T. Albaret, M. C. Payne, and A. De Vita, Phys. Rev. Lett. **93**, 175503 (2004).

[21] K. Parlinski, Z.-Q. Li, and Y. Kawazoe, Phys. Rev. Lett. **78**, 4063 (1997).

[22] A. Togo, F. Oba, and I. Tanaka, Phys. Rev. B **78**, 134106 (2008).

[23] C. E. Rasmussen and C. K. I. Williams, in *International Journal of Neural Systems* (MIT Press, Cambridge, MA, 2006), Vol. 14.

[24] N. Bernstein, J. R. Kermode, and G. Csányi, Rep. Prog. Phys. **72**, 026501 (2009).

[25] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevLett.114.096405, which includes Refs. [26,27], specifically Section S1, for full details of the GP regression scheme.

[26] R. Taylor, Proc. Am. Math. Soc. **5**, 753 (1954).

[27] J. R. Kermode, S. Cereda, P. Tangney, and A. De Vita, J. Chem. Phys. **133**, 094102 (2010).

[28] Using a symmetry-faithful description of the physical environment around atoms is as crucial here as it is in classical theoretical approaches which use order parameters to represent deviations from reference states and symmetry breaking phenomena.

[29] A. P. Bartók, R. Kondor, and G. Csányi, Phys. Rev. B **87**, 184115 (2013).

[30] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevLett.114.096405 specifically Section S1 for a more complete discussion.

[31] In the $\sigma_{\mathrm{err}} \to 0$ limit the GP becomes equivalent to straightforward data point regression, where preventing overfitting is a common problem [16].

[32] M. Elstner, D. Porezag, G. Jungnickel, J. Elsner, M. Haugk, T. Frauenheim, S. Suhai, and G. Seifert, Phys. Rev. B **58**, 7260 (1998).

[33] F. H. Stillinger and T. A. Weber, Phys. Rev. B **31**, 5262 (1985).

[34] G. Csányi, T. Albaret, G. Moras, A. De Vita, and M. C. Payne, J. Phys. Condens. Matter **17**, R691 (2005).

[35] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevLett.114.096405, specifically Section S2.

[36] S. J. Clark, M. D. Segall, C. J. Pickard, P. J. Hasnip, M. I. J. Probert, K. Refson, and M. C. Payne, Zeitschrift für Kristallographie **220**, 567 (2009).

[37] G. Kresse and J. Hafner, Phys. Rev. B **47**, 558 (1993).

[38] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevLett.114.096405, specifically Section S3 and Figs. S1 and S2.

[39] G. Moras, R. Choudhury, J. R. Kermode, G. Csányi, M. C. Payne, and A. De Vita, in *Trends in Computational Nanomechanics* (Springer, New York, 2010) pp. 1–23.

[40] D. Quigley and M. I. J. Probert, J. Chem. Phys. **120**, 11432 (2004).

[41] J. C. Grossman and L. Mitas, Phys. Rev. Lett. **94**, 056403 (2005).