Exploring the polymorphism of dicalcium silicates using transfer learning enhanced machine learning atomic potentials

Jon López-Zorrilla,* Xabier M. Aretxabaleta, and Hegoi Manzano*

Physics department, University of the Basque Country UPV/EHU, 48940 Leioa, Basque Country, Spain

E-mail: jon.lopezz@ehu.eus; hegoi.manzano@ehu.eus

Abstract

Belitic cements are a greener alternative to Ordinary Portland Cements due to the lower CO_2 associated to their production. However, their low reactivity with water is currently a drawback, resulting in longer setting times. In this study, we utilize a combination of evolutionary algorithms and machine learning atomic potentials (MLPs) to identify previously unreported belite polymorphs that may exhibit higher hydraulic reactivity than the known phases. To address the high computational demand of this methodology, we propose a novel transfer learning approach for generating MLPs. First, the models are pre-trained on a large set of classical data (ReaxFF) and then re-trained with Density Functional Theory (DFT) data. We demonstrate that the transfer learning enhanced potentials exhibit higher accuracy, require less training data, and are more transferable than those trained exclusively on DFT data. The generated machine learning potential enables a fast, exhaustive, and reliable exploration of the dicalcium silicate polymorphs. This includes studying their stability through phonon analysis and calculating their structural and elastic properties. Overall, we identify ten new belite

polymorphs within the energy range of the existing ones, including a layered phase with potentially high reactivity.

1 Introduction

Cement is the most extensively manufactured product globally in terms of mass. In 2022, its global production reached an astounding 4.2 billion tons, well over 500 kg per capita.[1] Cement and concrete, characterized by their versatility, cost-effectiveness, abundance, and local availability, are indispensable components in modern societies, playing a fundamental role in various construction applications. However, despite their ubiquity, the prevalent technologies associated with their production contribute significantly to carbon dioxide (CO_2) emissions, accounting for 8% of total per capita emissions. Most of the emissions, up to 60%, are intrinsic to the material.[2] The raw minerals, mainly clays and limestone, are melted at high temperatures to form the clinker phases: calcium silicate and aluminate phases, which mixed with other components such as gypsum and additives, form the cement. During melting, the calcination of limestone ($CaCO_3$) releases a considerable amount of CO_2, which is unavoidable.[3] Therefore, the strategies for cement's environmental impact reduction are based on Carbon Capture and Utilization or a reduction of the clinker in cement through the use of Supplementary Cementitious Materials. But there is a third alternative that researchers have long pursued: the so-called *belite cements*.[4]

Belite cements (BCs) are, as the name indicates, cements in which the main component is dicalcium silicate (also known as belite or C_2S in cement chemistry notation), in contrast to Ordinary Portland Cement (OPC) in which tricalcium silicate or alite predominates. Due to its lower Ca content, BCs require less limestone, reducing by up to 1/3 the OPC CO_2 emissions upon calcination. Furthermore, the sintering temperature is lower, also reducing fuel consumption. The resulting cement paste after BC hydration is equal to or even outperforms OPC pastes in terms of durability and mechanical properties.[5,6] However, BCs are

unpractical for most industrial applications due to the low dissolution rate of belite. Therefore, a practical transition from OPC to BC requires an acceleration of belite dissolution, often called activation.

Dicalcium silicate, Ca_2SiO_4, can be found in nature as an orthosilicate called larnite, the Ca end member of the olivine mineral group. In cement chemistry, this stable polymorph is denoted as γ, and it is not desirable due to its low hydraulic activity. Four additional polymorphs are found during clinker production, named β, α, α'_L, and α'_H. The β form is predominant in cement, stabilized by the presence of guest ions in the structure, mainly Mg.[7] A partial activation of belite has been achieved by chemical and mechanical means, as well as by the use of additives.[6,8] Besides these strategies, a new polymorph denoted as X has been recently obtained by thermal decomposition of a hydrated calcium silicate α-C_2SH.[9] The X polymorph is obtained together with a considerable amount of amorphous phase, and the mixture is reported to hydrate faster than the conventional β and α forms.[10] Large-scale production of amorphous and X-belite is currently impractical, although laboratory-scale synthesis is feasible.

The discovery of the X polymorph motivates the current work: could other metastable and highly reactive polymorphs of belite exist? To answer this question, we have used Evolutionary Algorithms (EA) to explore the configurational space of dicalcium silicate and search for unreported belite polymorphs. Performing an exhaustive search using EA requires thousands of Density Functional Theory (DFT) simulations, which can be prohibitive due to their high computational cost. The recent outburst of machine learning atomic potentials[11,12] (MLPs) provides a new alternative, enabling simulations with DFT precision at a significantly reduced computational cost. However, training a MLP demands, in turn, a substantial number of DFT calculations to build the database. To break the deadlock, we have used the transfer learning (TL) methodology,[13,14] which involves pre-training the models on low-quality data before training on the DFT data to enhance the performance of machine learning potentials. In practice, the re-training on the smaller set of high-quality data can be

accomplished by either fine-tuning all the parameters or keeping some of the layers frozen.

Transfer learning is a valuable tool in computational materials science for predicting various properties, where models pre-trained on different levels of computational data are used to improve performance when only a few data points are available.[13–15] In particular, in the field of machine learning potentials, a common approach involves fine-tuning Density Functional Theory models to achieve post-Hartree-Fock accuracy,[16,17] mainly by training on the difference between the two methods (commonly known as Δ-learning).[18] Although transfer learning models are typically used to achieve coupled cluster accuracy from DFT,[19–21] some research has been conducted to reduce the amount of DFT data required by pre-training on more primitive DFT approximations.[22] Nevertheless, to the best of our knowledge, none of the published works demonstrate the feasibility of transferring the physical knowledge from classical potentials to *ab initio* quantum methods. Even low-quality data based on empirical potentials contains significant, even if not very accurate, physical information about the systems at the atomic scale that can be used to enhance the MLP while minimizing the required amount of data. Thus, our approach involves exploiting the speed of empirical potentials to thoroughly sample the phase space and pre-train the machine learning models. We then select a small subset of those configurations to include in the DFT training database. In this work, we choose to pre-train the models using the ReaxFF reactive force field,[23,24] which is itself fitted to reproduce *ab initio* calculations. As for the high-level method, we consider DFT under the PBE exchange-correlation functional[25] sufficient for our purpose.

2 Methods

Reference data generation

For each phase in the data set, the same sampling technique was followed, consisting of different cell deformations and molecular dynamics simulations. First, several MD runs were performed using the ReaxFF forcefield[24] with the Ca/Si/O/H set of parameters from

Refs.[26,27] in LAMMPS[28] under the NVT ensemble. Various simulations were performed at different combinations of temperatures ($T = 300K, 600K, 900K$) and cell volumes ($\Delta V/V = 0.9, 1, 1.1$). A time step of 0.2 fs was used, and snapshots of the trajectory were saved every 500 steps. Second, all the non-symmetric axes were deformed from 10% compression to 10% expansion, including hydrostatic deformations with a maximum of 10% variation in volume, and angles were varied from -10º to 10º. The DFT data set was generated by randomly selecting structures from the ReaxFF data set, and evaluating their energy and forces.

Density Functional Theory

DFT calculations were performed using the quantum ESPRESSO software[29,30] using ON-CVPSP pseudopotentials[31] from pseudodojo,[32] under the PBE exchange-correlation functional,[25] and with the empirical dispersion by Grimme.[33] The plane wave energy cutoff was set to 100 Ry, and calculations were converged to 10^{-6} eV. Geometry optimizations were converged to 10^{-5} eV and 10^{-4}eVÅ^{-1} for energy and forces, respectively. Taking into account that systems of very different sizes have been studied, the number of k points was systematically selected such that the distance between points in the reciprocal space was about 0.25Å^{-1}.

MLP architecture and training

The machine learning atomic potentials used in this work are based on artificial neural networks.[11,12] All of them were trained using the ænet-PyTorch software,[34,35] using all the energies and 50% of the atomic forces. Chebyshev polynomials were used as descriptors for the atomic environments,[36] with a $N_{\mathrm{rad}} = 18$ and $N_{\mathrm{ang}} = 6$ order expansion for the radial and angular basis respectively. The radial cutoff distance was 6.5Å, while the angular distance was 4Å. This leads to a fingerprint with 52 components for each atomic environment. The MLP architecture for all models was $52 - 10 - 10 - 1$, with hyperbolic tangent as activation function. The only exception is the initial toy model of the calcium ion and the silicon

dioxide molecule, where the architecture was reduced to $40-3-3-1$. Transfer learning is performed by fine-tuning all the parameters of the pre-trained models.

The Supplementary Information contains a detailed analysis of the transferring methodology by freezing all combinations of the layers, and a comparison between different network architecture and descriptor sizes, showing that fine-tuning all layers is the optimal choice in our case.

Note that for every training data set of each experiment presented throughout this work, several MLPs have been trained, and the results shown correspond to the average of all MLPs.

Evolutionary algorithms

The exploration of the dicalcium silicate polymorphs was done using evolutionary algorithms as implemented in the USPEX code[37–39] (version 10.5). For each system size, EA runs were performed for enough generations until all experimentally known phases were found. Each structure was relaxed using the ænet-LAMMPS interface,[40,41] first minimizing the energy using ReaxFF to avoid random structures far from the included in the training data, and then using MLPs.

Phonon and elastic properties

Phonons were computed under the finite difference approximation to build the dynamical matrix. The phonopy[42,43] software was employed to generate the appropriate atomic displacements for each crystal structure, to build the dynamical matrix, and to compute the force constants and phonon dispersion along the high-symmetry path of the corresponding space group. The atomic displacements were set to 0.1Å, and supercells of at least 13Å (twice the cutoff distance of the descriptors) along each crystallographic axis were used in order to guarantee convergence.

The elastic tensor of all the structures was computed fitting the stress-strain relationship

$\sigma_i = C_{ij}\epsilon_j$. Each crystal parameter (cell-vector lengths and angles) was deformed independently 10 times, with a deformation in a range $\epsilon_0 \in (-0.01, 0.01)$. The elastic properties are computed under the Hill scheme. The forces and stress tensors for each of the structures were evaluated using the LAMMPS interface of ænet.

Annealing and amorphous dicalcium silicate

All MLP-based molecular dynamics simulations were performed with the ænet-LAMMPS interface.[40,41] The annealing to refine the polymorphs was performed under the NPT ensemble with a time step of 0.5 fs, both heating from 0 K to 400 K and cooling back at a rate of $2 \cdot 10^{12}$ Ks^{-1}.

The amorphous dicalcium silicate models were generated by heating a $4 \times 2 \times 4$ supercell of γ-C_2S up to 2000 K. The three different amorphous models were obtained by cooling the heated structure at three rates: $2 \cdot 10^{12}$, $2 \cdot 10^{13}$, and $2 \cdot 10^{14}$ Ks^{-1}. In all three cases, the time step was lowered to 0.1 fs to ensure the stability of the high-temperature molecular dynamics.

3 Results and discussion

Our results are organized as follows: first, the advantages of transfer learning are qualitatively introduced with a simple toy model. Second, we consider a more complex and realistic dataset to train a MLP for dicalcium silicates, while quantifying the benefits of the transfer learning approach. Finally, we use the trained MLP to explore the polymorphism of dicalcium silicates.

A simple transfer learning model

Let us first consider a simplified scenario to illustrate the capabilities of the transfer learning methodology: a system formed by a calcium ion and a silicon dioxide molecule, with the Si

and Ca atoms fixed at a distance of 5Å. We will explore the potential energy surface (PES) of the system as a function of the distance $d_{\text{Si-O}}$ from the silicon atom to a mobile oxygen, which breaks its bond to move toward the calcium atom. Sampling that PES using ReaxFF and DFT yields two similar landscapes, with two possible bound states for the oxygen and an energy barrier for the oxygen transfer. According to DFT, being bonded to the calcium ion is the lowest-energy configuration, while ReaxFF predicts the Si-O bond to be more stable, as shown in Figure 1 (a).

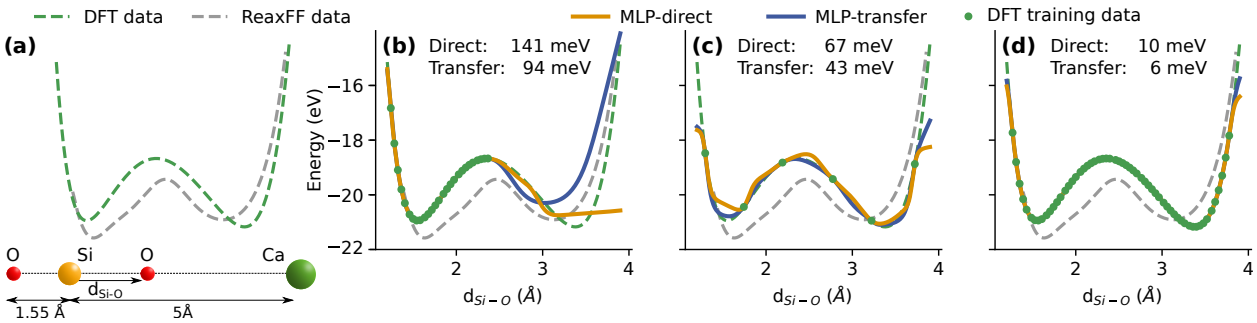

Figure 1: **Simple transfer learning model. (a)** Potential energy surface of the calcium ion and the silicon dioxide molecule computed with ReaxFF (grey dashed lines) and DFT (green dashed lines) as a function of the Si-O distance. The rest of the figure shows machine-learned potential energy surface for both direct learning (orange lines) and transfer learning (blue lines), for different distributions of DFT training data (green dots). The mean absolute error of the energy of both transfer and direct learning is displayed in each case.

We evaluate the advantage of the TL strategy in three scenarios with different distributions of the DFT training data. Two independent models are trained for each of them: one trained on all the ReaxFF data and subsequently re-trained on the selected DFT data points (MLP-transfer), and the second trained exclusively on that DFT data (MLP-direct).

First, we focus on a scenario where one of the bound states is correctly sampled by DFT (i.e. the Si-O bound state) while no DFT training data about the second state is included, see Figure 1(b). In this case, the direct training leads to an incorrect representation of the Ca-O region, even predicting an unphysical PES. On the contrary, the transferred model does predict a bound state resembling that of the ReaxFF data. Second, we explore the case where the DFT training points cover both regions of the PES but they are sparse (with only

3 points·Å^{-1}). In this case, both the direct and transfer learning protocols give a reasonable answer, but the transfer learning results are clearly smoother and the error with respect to the DFT data is considerably lower, see Figure 1(c). By increasing the DFT training configurations to 30 points·Å^{-1} [Figure 1(d)], both direct and transfer models yield similar results, but the error on the validation set is still lower for TL.

Overall, this simple model illustrates the ability of TL to reach accurate predictive capability with a reduced DFT dataset by employing empirical potentials to pre-train the MLP, and even predict energies for unsampled areas of the phase-space.

Transfer Learning MLP for dicalcium silicates

We now focus on the construction of a large dataset to train the MLP for dicalcium silicates, containing the 12 polymorphs available in Materials Project[44] as of December 2022, including experimental and theoretical phases. To pre-train the model, we sample a total of 10000 configurations using ReaxFF, by performing molecular dynamics simulations for each polymorph under different conditions and deforming the equilibrium cell along all independent crystallographic axes, as detailed in the Methods section. The DFT dataset is built from this data, by randomly selecting structures to be evaluated by DFT.

First, we study the impact of the transfer learning protocol on a realistic system like this. We consider several subsets of the database with an increasing amount of data and train models within both MLP-direct and MLP-transfer approaches. Figures 2(a) and (c) show the mean absolute error (MAE) of the energy and force as a function of the amount of DFT training data. Very interestingly, the transferred model outperforms its direct counterpart for any given amount of training data. The improvement (shown in the lower panels) is most significant with only a few hundred DFT data available for training, reaching up to a 40% reduction in both energy and force errors. With approximately 2000 DFT training data (half the total available set), the transfer learning model already reaches the same level of accuracy as the model trained directly on the full data set. Additionally, the model pre-

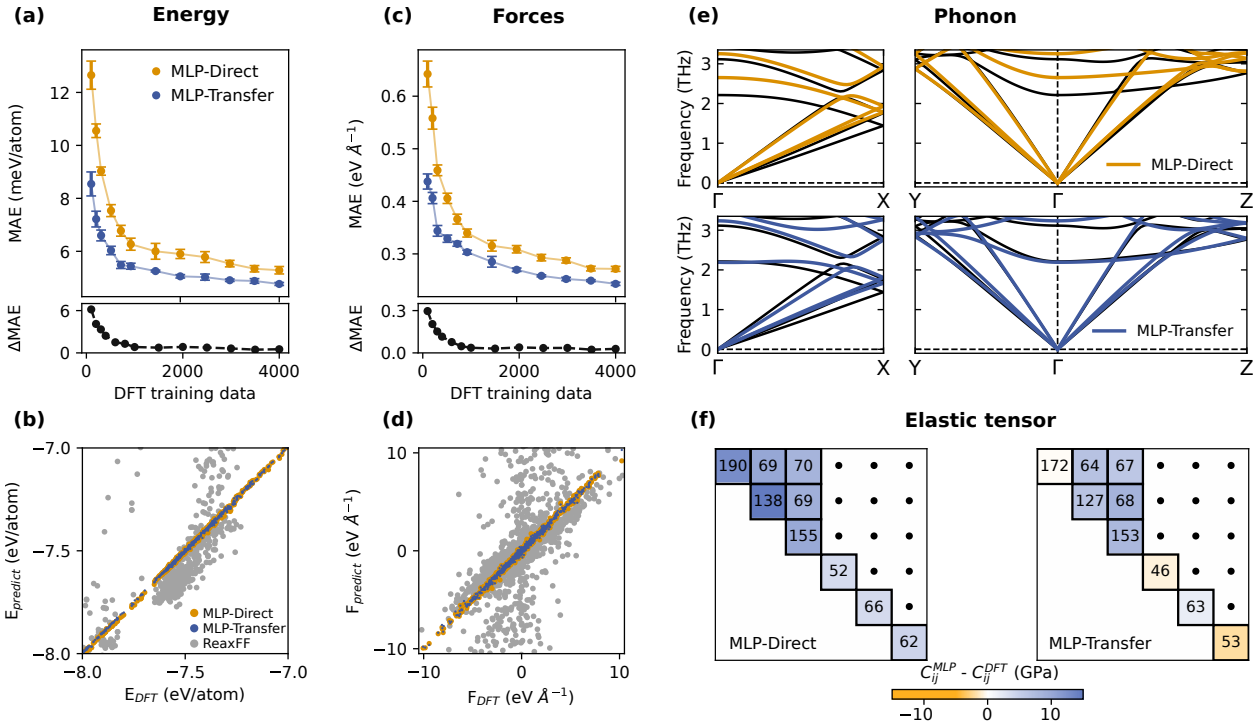

Figure 2: **Transfer learning results for dicalcium silicates.** **(a)** Energy error of transfer (blue) and direct learning (orange), as a function of the amount of DFT data used for training. The bottom panel shows the decrease of the error due to transfer learning. **(b)** Energy of a test set of 700 structures evaluated using transfer and direct learning and ReaxFF (grey), compared to the DFT value. **(c)-(d)** Same as the previous figure, in the case for the error of the forces. **(e)** Phonon dispersion of γ-belite computed using DFT (black), MLP-direct (orange) and MLP-transfer (blue). **(f)** Elastic tensor for the same polymorph. The color indicates the deviation with respect to the DFT value.

trained on ReaxFF data and trained on the full DFT data set is about 10% more accurate than the best-performing MLP-direct model.

Let us further characterize the performance of the transferred and direct MLPs trained on all the DFT data. Figures 2 (b) and (d) display the correlation of energy and forces with the actual DFT calculations, for both machine learning approaches and for the ReaxFF potential used for pre-training. Both MLP models outperform the ReaxFF potential and demonstrate high accuracy, resulting in a nearly perfect correlation with the DFT data. However, the MLP-transfer is still more accurate than the MLP-direct, by approximately 1 meV/atom in energies and 0.05 eVÅ^{-1} in forces, as quantified in Figure 2 (a) and (c). This is a clear indication of the capabilities of the transfer learning enhanced MLPs and

their suitability for the exploration of the C_2S phase space at a similar computational cost to empirical potentials. A simple efficiency check with a supercell of $4 \times 2 \times 4$ of γ-belite containing 896 atoms shows that the MLP is as fast as ReaxFF for 4 cores, but it scales better and is twice as fast for 32 cores, as shown in the Supplementary Information.

Finally, we explore the ability of our models to describe magnitudes related to higher-order derivatives of the PES, in particular, phonons [Figure 2(e)] and elastic properties [Figure 2(f)], which are of paramount importance for the discussion of our main results in the subsequent sections. As an illustrative example, we compute those magnitudes for the most stable belite polymorph, γ, within both training approaches, and compare the results to DFT calculations. The phonon dispersion curves are closer to DFT within the TL approach (blue lower panel) than directly training (orange upper panel). This is particularly noticeable for the optical modes, but it is also significant for acoustic modes. Moreover, the TL model excels at describing the phonon dispersion near the selected high symmetry points. Regarding the elastic tensor, direct MLPs are relatively accurate, within a 15% deviation from the DFT reference values. Additionally, the TL model further reduces the error in all the elastic constants, with a mean absolute error of 8.9 GPa on the elastic tensor, compared to the original 13.4 GPa of the direct model.

Thus, the TL approach improves the prediction of the PES and its first and second-order derivatives over directly training on all the available data. The Supplementary Information includes results from a similar study where only some phases are undersampled in the training data, demonstrating that our transfer learning approach also improves the performance in such scenarios. This is also interesting for the exploration of the phase space using evolutionary algorithms, where many atomic arrangements not included in the database are likely to be encountered. Hence, pre-training the MLPs in a ReaxFF dataset as diverse as possible will enhance the predictive power of the resulting potentials.

Therefore, to further improve the flexibility of the MLP on those high-energy regions of the energy landscape, we incorporated several new structures into the previous training set:

several polymorphs of silicon dioxide (SiO_2), calcium oxide (CaO), calcium silicates (CS), and tricalcium silicates (C_3S), as well as 1000 data points taken from a preliminary DFT-EA run to account for pseudorandom and high-energy conformations. Finally, to prevent the system from collapsing if the interatomic distances are too small, we included dimers for each pair of chemical elements in the system. The distance was reduced until repulsive forces exceeded 20 eVÅ^{-1} and expanded up to 4Å. The final ReaxFF training database comprises 20000 structures, while the DFT database is a subset of 8000 data points, as detailed in the S.I. Given the previous performance analysis, all the subsequent results are computed only with the MLP generated by the TL approach.

Belite polymorph search and computational screening

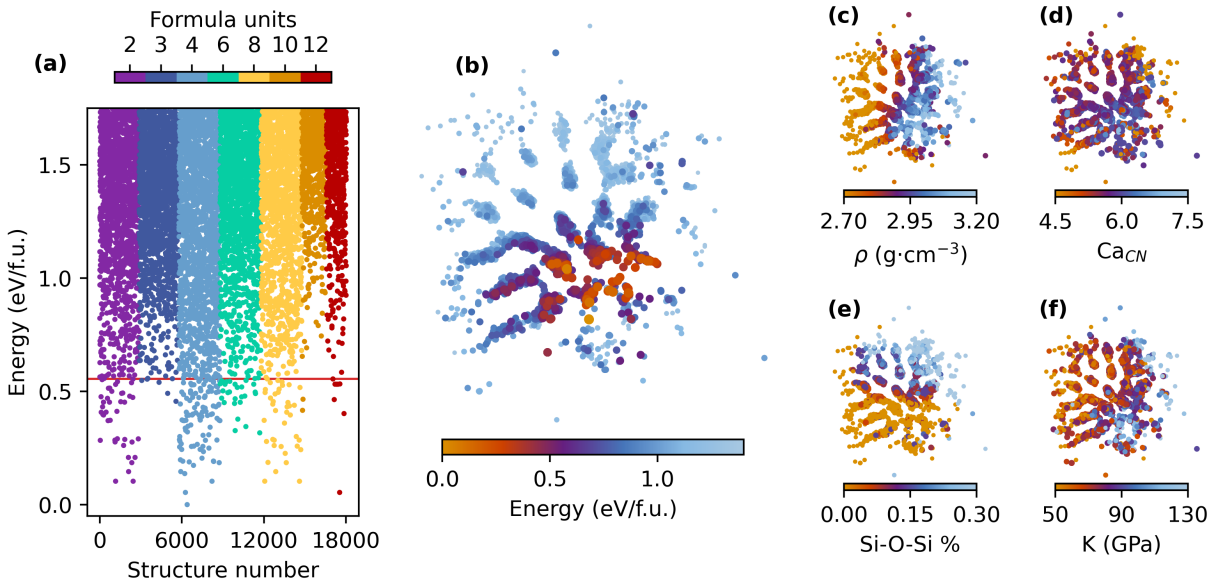

Figure 3: **Sketch-map of the generated structures. (a)** Energy of the found structures for all the considered numbers of formula units. **(b)** Sketch-map of the lowest-energy 5000 structures. Each point represents one of the structures, and the distance between them represents their structural similarity, i.e., the closer the points, the more similar the structures. **(c)** - **(f)** Several structural and mechanical properties represented using the same sketch-map: density (ρ), average coordination number of calcium (Ca_{CN}), percentage of oxygen atoms bonded to more than one silicon atom (Si-O-Si), and bulk modulus (K).

To discover new metastable polymorphs of belite, we conducted several independent EA

searches for different system sizes, ranging from 2 to 12 formula units per cell, at least until the experimental phases are found. This led to about 18000 potential structures, as displayed in Figure 3(a). Additional information regarding the number of structures for each size can be found in the Supplementary Information.

As shown in Figure 3(a), most of the generated structures have an energy considerably above the range of the experimental structures, delimited by the energy of the α polymorph, indicated by a red line. Considering the large amount of structures generated, we focus on the 5000 with the lowest energy, which are more likely to contain metastable phases. For these phases, we calculated their structural dissimilarities as detailed in the Supplementary Material. We then reduced their dimensionality using a sketch-map,[45,46] leading to the representation displayed in Figures 3(b)-(f). Each point of the figure represents one structure, and the distance between points indicates the similarity between structures: the closer two points are, the more similar the structures are. Over the maps in Figures 3(b)-(f), we projected several structural properties and elastic properties obtained from the elastic tensor. The low-energy structures generally exhibit high density and high bulk modulus values. In particular, all the structures with a density above 2.86 $\mathrm{gcm^{-3}}$, the lowest experimental density corresponding to X-C_2S, have energies below the higher energy experimental polymorph α-belite. The calcium coordination number (Ca_{CN}) distribution is centered at 6. Structures with Ca_{CN} below 5.5 are generally high-energy phases, while higher Ca_{CN} are more favorable. Finally, most of the low-energy polymorphs are orthosilicates, i.e. they have isolated silicate monomers, like the already known experimental phases.

The number of structures found in our initial search is too large, so we devise a computational screening procedure to systematically filter the unique and most stable polymorphs.

- In the first step, we identified duplicate structures and superstructures, by examining the structures with matching energies, densities, and space groups. Furthermore, the dissimilarity analysis described in the previous section was used to discard phases with a structural distance lower than 0.05. At this step, we identified 3000 unique structures.

- A large number of structures were still considerably above the highest energy of the known experimental polymorphs. To narrow down the searching space, we discarded all the phases with cohesive energy 7.5% over the energy of the highest-energy experimental phase, i.e. α, reducing the number of polymorphs to 215. The phonon dispersion was computed for each of them along the high-symmetry path corresponding to their crystal symmetry, rejecting any phase displaying imaginary modes. Although these phases may give rise to lower energy structures after undergoing phase transitions, we did not explore such possibilities due to the complexity of the problem. Instead, we assume that the initial set of 18000 structures already includes any of those lower-energy structures. This leaves 70 dynamically stable phases.

- Finally, an annealing process was performed, which involved increasing the temperature up to 400K and then cooling it down to 0K, followed by a geometry optimization. Since MD simulations might break the crystal symmetry, a lousy symmetry check was performed to identify the symmetry group using the ASE interface of `Spglib`.[47,48] The process concludes with one last structure optimization with fixed symmetry. After this refinement, the structural dissimilarity analysis was performed again, removing similar and identical structures.

After the computational screening, only 12 possible candidates remain from the initial 18000 structures. All the candidates are orthosilicates with IV-coordinated silicon, consistent with the experimental phases. It is worth noting that two structures (S_{12} and S_5) present only translational symmetry (P1 space group) and have large unit cells; therefore, we argue that they could be classified as glasses. These structures were generated in the EA stage and survived the annealing stage and the stability checks. Furthermore, none of the candidate structures were included in the Materials Projects database and therefore were not part of the training set. As a matter of fact, all non-experimental phases in the training set have energies above the α phase and all of the polymorphs found by EA, except for the S_{12} phase. However, this phase has already been discarded. For instance, the energy of the next

polymorph with the highest energy (S_{11}) predicted by the MLP is 0.64 eV/f.u. above γ, while the remaining non-experimental phases in the database are approximately 0.7 eV/f.u. above that reference.

Reactivity analysis of the C_2S candidates

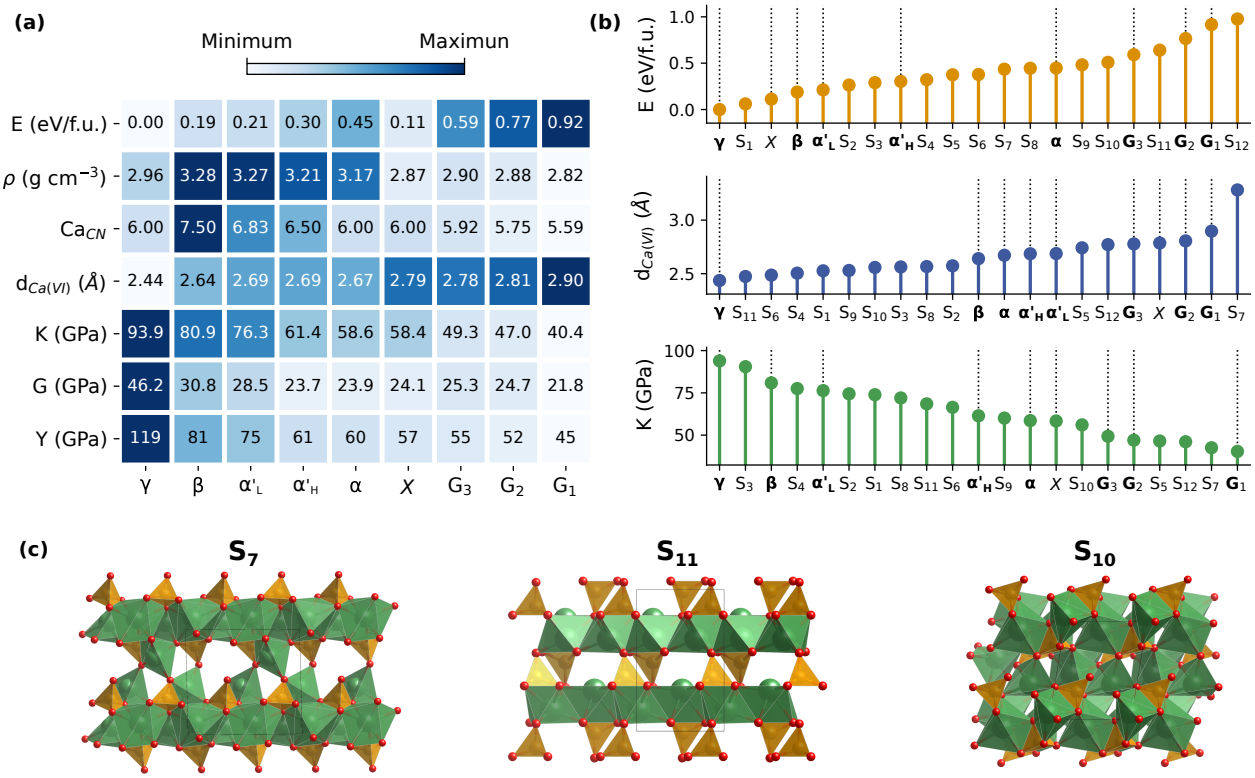

Figure 4: **Reactivity analysis of the candidates.** **(a)** Several properties computed for the known polymorphs and the three generated glassy structures: energy per formula unit (E), density (ρ), averaged coordination number for calcium (Ca_{CN}), average Ca-O distance to fulfill the Ca coordination shell ($d_{\text{Ca-O (VI)}}$) , bulk modulus (K), shear modulus (G) and Young's modulus (Y). The color ranges from the minimum to the maximum value of each magnitude, and the polymorphs are ordered from least to most reactive. **(b)** Energy, average Ca-O distance of the Ca coordination shell and bulk modulus of all the candidates. **(c)** Selected candidates with the potential to display high reactivity.

The final aim of this work is to find potential C_2S polymorphs with high dissolution rates, and for that, we need an atomic scale reactivity indicator. Unfortunately, a quantitative prediction of mineral dissolution rates based on atomistic simulations is a complicated task. Electronic structure calculations are suitable for surface chemisorption reactions, for

example, to predict the catalytic properties of materials.[49,50] However, they are less appropriate to predict dissolution, as the individual water chemisorption at the mineral surface do not correlate with the dissolution rates. For instance, it has been shown that the water dissociation reaction in γ-C_2S can be barrier-less despite being the polymorph with the lowest dissolution rate. In contrast, water dissociation at the fast-dissolving β-C_2S surface presents energy barriers between 2.5 and 47.2 kJ mol^{-1}. Furthermore, electronic properties are not accessible to the MLP, so we need a structure or energy-based descriptor.[51]

A key factor that determines mineral dissolution rates is the ligand-cation exchange ability.[52–54] There is a clear correlation between the dissolution rates of oxides, orthosilicates, and carbonates, with the water exchange rate of the forming cation in solution.[52,54] We assume that the ligand-cation exchange mechanism applies to phases with the same cation, and the exchange tendency is related to how strongly bonded is the cation in the structure, reflected in factors such as coordination and flexibility. To find correlations we plotted in Figure 4 (a) the binding energy and several structural and elastic properties of the experimental polymorphs in increasing order of hydraulic reactivity. Unfortunately, there are no quantitative sample-independent values for the dissolution rate of these polymorphs, but it is well-known that they follow the order (from least to most reactive) $\gamma < \beta < \alpha$.[6] Moreover, the high dissolution rate of X-C_2S samples has been suggested to correspond actually to an amorphous coexisting phase. Therefore, three amorphous dicalcium silicate models have been constructed from MD simulations by heating a γ-C_2S supercell to 700K and rapid cooling at different cooling rates (see Methods for details). These models are denoted as G1, G2 and G3, from lowest to highest cooling rate. The final expected order for the dissolution rate is $\gamma < \beta < \alpha <$ X $<$ G3 $<$ G2 $<$ G1, with uncertainty about the actual ranking of the X phase.

Figure 4 (a) shows that there is a correlation between the cohesive energy and the reactivity: the more energetically stable the polymorph is, the less reactive they are. The X phase is an exception to the trend, but it has already been discussed that its reactivity may

not have been properly quantified. In fact, all three amorphous models have higher energy than the rest of the polymorphs. Nevertheless, the X phase correlates with the remaining properties, while the γ phase, undoubtedly the least reactive phase, breaks the trends in density. For all the other phases a lower density, higher Ca-O distance to fulfill the Ca coordination shell $d_{Ca(VI)}$, lower Ca_{CN}, and lower elastic properties correlate with a higher reactivity. It is interesting to note that the amorphous structures ranked as a function of the cooling rate follow the correlations, which suggests that amorphous C_2S could be indeed responsible for the high dissolution rate in samples with X-C_2S.

Considering the previous trends, we propose three magnitudes as potential reactivity descriptors of the dicalcium silicates: the cohesive energy, the Ca-O distance ($d_{Ca(VI)}$), and the bulk modulus (K). Figure 4 (b) ranks the 12 new polymorphs according to these three reactivity indicators. In theory, the phases located at the right end (high energy, high $d_{Ca(VI)}$, and low K) should be the most reactive polymorphs, and the objective of our search. Despite being in the range of interest, we will not consider the S_5 and S_{12} because they are amorphous systems as mentioned before, and therefore it is natural that they lie close to the amorphous models. Out of the remaining candidates, the S_7 structure shows the most promise. It has the lowest bulk modulus among the crystalline structures, half the value of γ-C_2S, and the largest $d_{Ca(VI)}$. The S_7 is a layered structure, with CaO forming a central sheet and silicate groups at both sides with three of their four oxygen atoms coordinated to the CaO central sheet. The interlayer space contains a Ca atom that links consecutive layers. The S_{11} is a similar layered structure, without Ca in the interlayer space. Its energy is in the upper range, only surpassed by the amorphous structures. However, its overall potential is limited due to low values of other indicators. Other possible candidates that may display higher reactivity than the known phases include S_{10}, a bulk phase with a monoclinic axis, which also exhibits high energy and low bulk modulus.

4 Conclusions

In this work, we have introduced a hitherto unexplored approach to efficiently generating accurate MLPs based on transfer learning (TL) from the ReaxFF reactive force field to DFT. Previous attempts to reuse a lower-quality training set to reduce the amount of high-quality data were limited to using quantum methods as both low- and high-quality data. The present study demonstrates that TL from a classical force fields to DFT is both feasible and effective. In particular, we find that building the MLPs from models pre-trained on ReaxFF data can boost their accuracy from 10% to 40% in both energy and forces. Very importantly, the generation of the data for pre-training is virtually free of computational cost, and the methodology has no drawback: the TL-enhanced MLPs outperform those trained exclusively on DFT data in every tested scenario. In addition, MLPs pre-trained on large datasets made by empirical potentials can cover larger regions of the configurational space, providing flexibility and generality to the potential.

The TL methodology has been applied to build a MLP for calcium silicates. First, the MLPs were pre-trained on a dataset of 20000 ReaxFF configurations, followed by a refinement on 8000 DFT data points. The resulting MLP can successfully reproduce the DFT energies and forces with a mean absolute error of 4.8 meV/atom and 0.25 eVÅ^{-1} respectively, as well as phonon spectra and elastic properties of calcium silicate crystals. This potential has been used to search for new dicalcium silicate polymorphs, aiming to find new (and hopefully highly reactive) belite phases. The combination of the DFT-like accuracy with the efficiency of classical potentials permits to examine and sieve thousands of polymorphs. In particular, we generated 18000 structures using EA, which were filtered using a computational screening protocol to discard duplicates, supercells, and dynamically unstable structures according to their phonon spectra and annealing at 400K. From the initial 18000, we identified 10 new crystalline C_2S polymorphs that are potentially stable. Based on our mechanical and structural descriptors of reactivity, a layered structure, denoted as S_7 in this work, is particularly promising for displaying higher hydraulic activity than the

currently known belite phases.

The next step will be to investigate the hydration of these structures by performing molecular dynamics simulations at the crystal/water interfaces. If the proposed polymorphs are indeed highly reactive phases, it will be essential to test their thermodynamic stabilization by guest ions, in order to guide the synthesis and eventual production of highly reactive belitic cements. To conduct these studies, the computational work should focus on exploiting the presented TL methodology to include large and complex systems beyond the DFT capabilities. This could include belite/water interfaces,[55] complexes and clusters in solution,[56] an extension of the MLP to new chemical species etc., allowing quantitative studies of the C_2S reactivity and stability.

Acknowledgement

The authors thank Professor Iñigo Etxebarria for his valuable comments on the manuscript. This work was supported by the "Departamento de Educación, Política Lingüística y Cultura del Gobierno Vasco" (Grant No. IT1458-22), the "Ministerio de Ciencia e Innovación" (TED2021-130860B-I00), the University of the Basque Country UPV/EHU (Colab22/06) and the Transnational Common Laboratory "Aquitaine-Euskadi Network in Green Concrete and Cement-based Materials" (LTC-Green Concrete). The authors thank for technical and human support provided by SGIker (UPV/EHU/ ERDF, EU). J.L.-Z. acknowledges the financial support from the Basque Country Government (PRE_2019_1_0025).

Supporting Information Available

The database created for this work is available in Zenodo `https://doi.org/10.5281/zenodo.10796241`, together with the generated MLPs and the structures of the final candidates. All the software used in this work is open source. The modifications to the ænet-PyTorch code to perform transfer learning can be found at `https://github.com/`

`jlopez141/TL_aenet-PyTorch`. The Supplementary Information contains:

- Chebyshev descriptors and structural similarity.
- Performance of machine learning potentials.
- Training data set details.
- Hyperparameters for training and transfer learning.
- Transfer Learning with different compositions.
- Details about the structures generated by EA.
- Validation of the MLPs.

References

(1) Activity Report 2022. `https://cembureau.eu/media/m3jcyfre/cembureau-activity-report-2022-light.pdf`, 2023; Cembureau: The European Cement Association.

(2) Environment, U.; Scrivener, K. L.; John, V. M.; Gartner, E. M. Eco-efficient cements: Potential economically viable solutions for a low-CO2 cement-based materials industry. *Cement and concrete Research* **2018**, *114*, 2–26.

(3) Barcelo, L.; Kline, J.; Walenta, G.; Gartner, E. Cement and carbon emissions. *Materials and structures* **2014**, *47*, 1055–1065.

(4) Ghosh, S. N.; Rao, P. B.; Paul, A.; Raina, K. The chemistry of dicalcium silicate mineral. *Journal of Materials Science* **1979**, *14*, 1554–1566.

(5) Wang, L.; Yang, H.; Zhou, S.; Chen, E.; Tang, S. Hydration, mechanical property and CSH structure of early-strength low-heat cement-based materials. *Materials Letters* **2018**, *217*, 151–154.

(6) Cuesta, A.; Ayuela, A.; Aranda, M. A. Belite cements and their activation. *Cement and Concrete Research* **2021**, *140*, 106319.

(7) Zhao, Y.; Lu, L.; Wang, S.; Gong, C.; Huang, Y. Modification of dicalcium silicates phase composition by BaO, SO 3 and MgO. *Journal of Inorganic and Organometallic Polymers and Materials* **2013**, *23*, 930–936.

(8) Kim, Y.-M.; Hong, S.-H. Influence of minor ions on the stability and hydration rates of β-dicalcium silicate. *Journal of the American Ceramic Society* **2004**, *87*, 900–905.

(9) Link, T.; Bellmann, F.; Ludwig, H.; Haha, M. B. Reactivity and phase composition of Ca2SiO4 binders made by annealing of alpha-dicalcium silicate hydrate. *Cement and Concrete Research* **2015**, *67*, 131–137.

(10) Pirvan, A. A.; Haha, M. B.; Boehm-Courjault, E.; Scrivener, K. L. Calcium-silicate-hydrates from reactive dicalcium silicate binder. 39 th Cement and Concrete Science Conference 2019. 2019; p 215.

(11) Behler, J.; Parrinello, M. Generalized neural-network representation of high-dimensional potential-energy surfaces. *Physical review letters* **2007**, *98*, 146401.

(12) Behler, J. Neural network potential-energy surfaces in chemistry: a tool for large-scale simulations. *Physical Chemistry Chemical Physics* **2011**, *13*, 17930–17955.

(13) Yamada, H.; Liu, C.; Wu, S.; Koyama, Y.; Ju, S.; Shiomi, J.; Morikawa, J.; Yoshida, R. Predicting materials properties with little data using shotgun transfer learning. *ACS central science* **2019**, *5*, 1717–1730.

(14) Jha, D.; Choudhary, K.; Tavazza, F.; Liao, W.-k.; Choudhary, A.; Campbell, C.; Agrawal, A. Enhancing materials property prediction by leveraging computational and experimental data using deep transfer learning. *Nature communications* **2019**, *10*, 5316.

(15) Cai, C.; Wang, S.; Xu, Y.; Zhang, W.; Tang, K.; Ouyang, Q.; Lai, L.; Pei, J. Transfer learning for drug discovery. *Journal of Medicinal Chemistry* **2020**, *63*, 8683–8694.

(16) Smith, J. S.; Nebgen, B. T.; Zubatyuk, R.; Lubbers, N.; Devereux, C.; Barros, K.; Tretiak, S.; Isayev, O.; Roitberg, A. E. Approaching coupled cluster accuracy with a general-purpose neural network potential through transfer learning. *Nature communications* **2019**, *10*, 2903.

(17) Kaser, S.; Boittier, E. D.; Upadhyay, M.; Meuwly, M. Transfer learning to CCSD (T): Accurate anharmonic frequencies from machine learning models. *Journal of Chemical Theory and Computation* **2021**, *17*, 3687–3699.

(18) Ramakrishnan, R.; Dral, P. O.; Rupp, M.; Von Lilienfeld, O. A. Big data meets quantum chemistry approximations: the Δ-machine learning approach. *Journal of chemical theory and computation* **2015**, *11*, 2087–2096.

(19) Bogojeski, M.; Vogt-Maranto, L.; Tuckerman, M. E.; Müller, K.-R.; Burke, K. Quantum chemical accuracy from density functional approximations via machine learning. *Nature communications* **2020**, *11*, 5223.

(20) Ruth, M.; Gerbig, D.; Schreiner, P. R. Machine learning of coupled cluster (T)-energy corrections via delta (Δ)-Learning. *Journal of Chemical Theory and Computation* **2022**, *18*, 4846–4855.

(21) Maier, S.; Collins, E. M.; Raghavachari, K. Quantitative Prediction of Vertical Ionization Potentials from DFT via a Graph-Network-Based Delta Machine Learning Model Incorporating Electronic Descriptors. *The Journal of Physical Chemistry A* **2023**, *127*, 3472–3483.

(22) Sun, G.; Sautet, P. Toward fast and reliable potential energy surfaces for metallic Pt clusters by hierarchical delta neural networks. *Journal of chemical theory and computation* **2019**, *15*, 5614–5627.

(23) Van Duin, A. C.; Dasgupta, S.; Lorant, F.; Goddard, W. A. ReaxFF: a reactive force field for hydrocarbons. *The Journal of Physical Chemistry A* **2001**, *105*, 9396–9409.

(24) Senftle, T. P.; Hong, S.; Islam, M. M.; Kylasa, S. B.; Zheng, Y.; Shin, Y. K.; Junkermeier, C.; Engel-Herbert, R.; Janik, M. J.; Aktulga, H. M.; others The ReaxFF reactive force-field: development, applications and future directions. *npj Computational Materials* **2016**, *2*, 1–14.

(25) Perdew, J. P.; Burke, K.; Ernzerhof, M. Generalized gradient approximation made simple. *Physical review letters* **1996**, *77*, 3865.

(26) Van Duin, A. C.; Strachan, A.; Stewman, S.; Zhang, Q.; Xu, X.; Goddard, W. A. ReaxFFSiO reactive force field for silicon and silicon oxide systems. *The Journal of Physical Chemistry A* **2003**, *107*, 3803–3811.

(27) Manzano, H.; Pellenq, R. J.; Ulm, F.-J.; Buehler, M. J.; Van Duin, A. C. Hydration of calcium oxide surface predicted by reactive force field molecular dynamics. *Langmuir* **2012**, *28*, 4187–4197.

(28) Thompson, A. P.; Aktulga, H. M.; Berger, R.; Bolintineanu, D. S.; Brown, W. M.; Crozier, P. S.; in 't Veld, P. J.; Kohlmeyer, A.; Moore, S. G.; Nguyen, T. D.; Shan, R.; Stevens, M. J.; Tranchida, J.; Trott, C.; Plimpton, S. J. LAMMPS - a flexible simulation tool for particle-based materials modeling at the atomic, meso, and continuum scales. *Comp. Phys. Comm.* **2022**, *271*, 108171.

(29) Giannozzi, P.; Baroni, S.; Bonini, N.; Calandra, M.; Car, R.; Cavazzoni, C.; Ceresoli, D.; Chiarotti, G. L.; Cococcioni, M.; Dabo, I.; others QUANTUM ESPRESSO: a modular and open-source software project for quantum simulations of materials. *Journal of physics: Condensed matter* **2009**, *21*, 395502.

(30) Giannozzi, P.; Andreussi, O.; Brumme, T.; Bunau, O.; Nardelli, M. B.; Calandra, M.; Car, R.; Cavazzoni, C.; Ceresoli, D.; Cococcioni, M.; others Advanced capabilities for

materials modelling with Quantum ESPRESSO. *Journal of physics: Condensed matter* **2017**, *29*, 465901.

(31) Hamann, D. Optimized norm-conserving Vanderbilt pseudopotentials. *Physical Review B* **2013**, *88*, 085117.

(32) van Setten, M. J.; Giantomassi, M.; Bousquet, E.; Verstraete, M. J.; Hamann, D. R.; Gonze, X.; Rignanese, G.-M. The PseudoDojo: Training and grading a 85 element optimized norm-conserving pseudopotential table. *Computer Physics Communications* **2018**, *226*, 39–54.

(33) Grimme, S. Semiempirical GGA-type density functional constructed with a long-range dispersion correction. *Journal of computational chemistry* **2006**, *27*, 1787–1799.

(34) López-Zorrilla, J.; Aretxabaleta, X. M.; Yeu, I. W.; Etxebarria, I.; Manzano, H.; Artrith, N. ænet-PyTorch: A GPU-supported implementation for machine learning atomic potentials training. *The Journal of Chemical Physics* **2023**, *158*.

(35) Artrith, N.; Urban, A. An implementation of artificial neural-network potentials for atomistic materials simulations: Performance for TiO2. *Computational Materials Science* **2016**, *114*, 135–150.

(36) Artrith, N.; Urban, A.; Ceder, G. Efficient and accurate machine-learning interpolation of atomic energies in compositions with many species. *Physical Review B* **2017**, *96*, 014112.

(37) Oganov, A. R.; Glass, C. W. Crystal structure prediction using ab initio evolutionary techniques: Principles and applications. *The Journal of chemical physics* **2006**, *124*.

(38) Oganov, A. R.; Lyakhov, A. O.; Valle, M. How Evolutionary Crystal Structure Prediction Works and Why. *Accounts of chemical research* **2011**, *44*, 227–237.

(39) Lyakhov, A. O.; Oganov, A. R.; Stokes, H. T.; Zhu, Q. New developments in evolutionary structure prediction algorithm USPEX. *Computer Physics Communications* **2013**, *184*, 1172–1182.

(40) Chen, M. S.; Morawietz, T.; Mori, H.; Markland, T. E.; Artrith, N. AENET–LAMMPS and AENET–TINKER: Interfaces for accurate and efficient molecular dynamics simulations with machine learning potentials. *The Journal of Chemical Physics* **2021**, *155*.

(41) Mori, H.; Tsuru, T.; Okumura, M.; Matsunaka, D.; Shiihara, Y.; Itakura, M. Dynamic interaction between dislocations and obstacles in bcc iron based on atomic potentials derived using neural networks. *Physical Review Materials* **2023**, *7*, 063605.

(42) Togo, A.; Chaput, L.; Tadano, T.; Tanaka, I. Implementation strategies in phonopy and phono3py. *J. Phys. Condens. Matter* **2023**, *35*, 353001.

(43) Togo, A. First-principles Phonon Calculations with Phonopy and Phono3py. *J. Phys. Soc. Jpn.* **2023**, *92*, 012001.

(44) Jain, A.; Ong, S. P.; Hautier, G.; Chen, W.; Richards, W. D.; Dacek, S.; Cholia, S.; Gunter, D.; Skinner, D.; Ceder, G.; others Commentary: The Materials Project: A materials genome approach to accelerating materials innovation. *APL materials* **2013**, *1*.

(45) Oganov, A. R.; Valle, M. How to quantify energy landscapes of solids. *The Journal of chemical physics* **2009**, *130*.

(46) Ceriotti, M.; Tribello, G. A.; Parrinello, M. Simplifying the representation of complex free-energy landscapes using sketch-map. *Proceedings of the National Academy of Sciences* **2011**, *108*, 13023–13028.

(47) Larsen, A. H.; Mortensen, J. J.; Blomqvist, J.; Castelli, I. E.; Christensen, R.; Dułak, M.; Friis, J.; Groves, M. N.; Hammer, B.; Hargus, C.; others The atomic sim-

ulation environment—a Python library for working with atoms. *Journal of Physics: Condensed Matter* **2017**, *29*, 273002.

(48) Togo, A.; Tanaka, I. Spglib: a software library for crystal symmetry search. *arXiv preprint arXiv:1808.01590* **2018**,

(49) Nørskov, J. K.; Abild-Pedersen, F.; Studt, F.; Bligaard, T. Density functional theory in surface chemistry and catalysis. *Proceedings of the National Academy of Sciences* **2011**, *108*, 937–943.

(50) Nørskov, J. K.; Bligaard, T.; Rossmeisl, J.; Christensen, C. H. Towards the computational design of solid catalysts. *Nature chemistry* **2009**, *1*, 37–46.

(51) Rejmak, P.; Dolado, J. S.; Aranda, M. A.; Ayuela, A. First-principles calculations on polymorphs of dicalcium silicate—Belite, a main component of Portland cement. *The Journal of Physical Chemistry C* **2019**, *123*, 6768–6777.

(52) Ohlin, C. A.; Villa, E. M.; Rustad, J. R.; Casey, W. H. Dissolution of insulating oxide materials at the molecular scale. *Nature materials* **2010**, *9*, 11–19.

(53) Rustad, J. R.; Casey, W. H. Metastable structures and isotope exchange reactions in polyoxometalate ions provide a molecular view of oxide dissolution. *Nature materials* **2012**, *11*, 223–226.

(54) Casey, W. H.; Swaddle, T. W. Why small? The use of small inorganic clusters to understand mineral surface and dissolution reactions in geochemistry. *Reviews of Geophysics* **2003**, *41*.

(55) Qi, C.; Manzano, H.; Spagnoli, D.; Chen, Q.; Fourie, A. Initial hydration process of calcium silicates in Portland cement: A comprehensive comparison from molecular dynamics simulations. *Cement and Concrete Research* **2021**, *149*, 106576.

(56) Aretxabaleta, X. M.; López-Zorrilla, J.; Etxebarria, I.; Manzano, H. Multi-step nucleation pathway of CSH during cement hydration from atomistic simulations. *Nature Communications* **2023**, *14*, 7979.