Multikernel similarity-based clustering of amorphous systems and machine-learned interatomic potentials by active learning

Firas Shuaib, Guido Ori, Philippe Thomas, Olivier Masson, Assil Bouzid

Received: 15 March 2024 | Revised: 22 August 2024 | Accepted: 29 August 2024

DOI: 10.1111/jace.20128

Journal of the American Ceramic Society

RESEARCH ARTICLE

Multikernel similarity-based clustering of amorphous systems and machine-learned interatomic potentials by active learning

Firas Shuaib[1] | **Guido Ori**[2,3] | **Philippe Thomas**[1] | **Olivier Masson**[1] | **Assil Bouzid**[1]

[1]Institut de Recherche sur les Céramiques (IRCER), UMR CNRS 7315, Université de Limoges, Limoges, France

[2]Institut de Physique et Chimie des Matériaux de Strasbourg, UMR CNRS 7504, Université de Strasbourg, CNRS, Strasbourg, France

[3]ADYNMAT CNRS consortium, Strasbourg, France

Correspondence
Firas Shuaib and Assil Bouzid, Institut de Recherche sur les Céramiques (IRCER), UMR CNRS 7315, F-87068 Université de Limoges, Limoges, France.
Email: firas.shuaib_mohammed@unilim.fr and assil.bouzid@cnrs.fr

Funding information
Région Nouvelle Aquitaine, Grant/Award Number: AAPR2021-2020-11779110; Agence Nationale de la Recherche, Grant/Award Number: ANR-20-CE08-0021

Abstract

We present a hybrid similarity kernel that exemplifies the integration of short- and long-range descriptors via the use of an average kernel approach. This technique allows for a direct measure of the similarity between amorphous configurations, and when combined with an active learning (AL) spectral clustering approach, it leads to a classification of the amorphous configurations into uncorrelated clusters. Subsequently, a minimum size database is built by considering a small fraction of configurations belonging to each cluster and a machine learning interatomic potential (MLIP), within the Gaussian approximation scheme, is fitted by relying on a Bayesian optimization of the potential hyperparameters. This step is embedded within an AL loop that allows to sequentially increase the size of the learning database whenever the MLIP fails to meet a predefined energy convergence threshold. As such, MLIP are fitted in an almost fully automatized fashion. This approach is tested on two diverse amorphous systems that were previously generated using first-principles molecular dynamics. Accurate potentials with less than 2 meV/atom root mean square energy error compared to the reference data are obtained. This accuracy is achieved with only 175 configurations sampling the studied systems at various temperatures. The robustness of these potentials is then confirmed by producing models with several thousands of atoms featuring a good agreement with reference ab initio and experimental data.

KEYWORDS
active learning clustering, average kernel, Bayesian optimization, Coulomb matrix, Gaussian approximation potential (GAP), hybrid similarity kernel, SOAP descriptor, spectral clustering

1 | INTRODUCTION

Machine learning interatomic potentials (MLIP)[1–7] have become a game changer in materials modeling as they allow to expand the size and time scales of computer simulations well beyond those achieved by quantum-mechanical-based techniques, yet keeping a similar accuracy.[8,9] In the particular case of amorphous systems, producing quantitative models usually requires the use of first-principles molecular dynamics (FPMD)[8,10,11] with a typical model size of the order of hundred atoms. FPMD simulations generate trajectories of a given system at various thermodynamical conditions (pressure, temperature, etc.) that represent a time-ordered concatenation of atomistic configurations of the studied system. While powerful, such technique remains heavy and costly as it requires the use of large computing facilities over a long periods of time that can go up to a year.[12]

At this level, machine learning (ML), whether in the framework of high-dimensional artificial neural networks (NNs)[1,13,14] or in the framework of kernel-based methods, such as the Gaussian approximation potential (GAP) framework,[2,15–21] offers a promising approach to construct fast and accurate MLIPs that enable to go beyond the limitations of first-principles simulations, yet keeping similar accuracy. MLIP can leverage these issues by exploiting the FPMD data (trajectories, energies, forces, stress, etc.) to fit the multidimensional potential energy surface of the explored phase space. Successful training enables atomistic simulations to achieve precision comparable to quantum mechanics, while significantly reducing the computational cost by many orders of magnitude. Nevertheless, the accuracy and transferability of these potentials, relies on, first, the construction of the database on which they are fitted, second, on the model hyperparameters and the way they are chosen.[1,22,23] Addressing these issues require the use of ML techniques at two levels: (i) a classification of the atomistic snapshots based on a measure of their similarity that enables to construct minimal, yet representative, fitting database,[24,25] and (ii) finding the optimal model hyperparameters required to achieve an accurate MLIP fit.[26] Focusing on the first point (i), generally, finding the optimal number of learning configurations is a key toward the development of general-purpose MLIPs, particularly when dealing with a large chemical space as this minimizes the required computational cost and ensures an efficient representativity of the data. In addition, efficient on-the-fly training and reinforcement learning schemes require a precise selection of uncorrelated training configurations. Therefore, developing ML schemes that optimize the data usage (i.e., data-efficient training and data-distillation schemes) is a field that is gaining considerable attention and is considered as the next challenge in computational materials science. In this context, very recently, Finkbeiner et al.[27] provided a scheme to identify uncorrelated atomic configurations from extensive data sets by relying on a nonstandard NN workflow, known as the random network distillation (RND), for training machine-learned interatomic potentials. Ben Mahmoud et al.[28] tackle the question of how exactly one chooses the structures that inform the model. In particular, the authors question the possibilities of making general-purpose cost-effective potentials trained on very large data sets. To reach this goal, the authors comment on "data set distillation" procedure to reduce large data sets and build multipurpose models. In another work, Speckhard et al.[29] tackled the question of "How big is Big Data?" with a particular focus on the generalization of models to similar data sets and on the possible ways to gather high-quality data sets from heterogenous sources. Finally, Kaur et al.[30] provide an approach that requires only a few tens of training structures to achieve sub-kJ/mol accuracy in the sublimation enthalpies and sub 1% error in densities for ice polymorphs at finite temperature and pressure. In this work, the training model size is optimized by studying the convergence of the average potential energy and density from simulations in the constant temperature, constant pressure sensemble (NPT) ensemble as a function of the size of the training set.

In the case of amorphous configurations, the root mean square displacement (RMSD) measured between the Cartesian coordinates of the atoms, which is readily made invariant to relative translations and rotations, is the most apparent option for a metric to compare atomic structures. However, it is very difficult to expand the RMSD to cope with scenarios where atoms in the two structures cannot be clearly mapped onto one another (due to combinatorial shift and scaling as a function of the size of the molecules being compared).[31] Starting with descriptors intended to represent atomic environments in a way that is unaffected by rotations, translations, and permutations of equivalent atoms and then combining them to produce a global measure of structural similarity is a particularly promising route for comparing structures.[32] This concept often depends on identifying the optimal correspondence between pairs of environments in the two configurations.[31,33] A particularly elegant framework for obtaining invariant local descriptors of atomistic environments is the smooth overlap of atomic positions (SOAP).[24,34] The SOAP descriptor represents atomic geometries by using a localized expansion of a Gaussian-smeared atomic density. This expansion is constructed using a weighted sum of orthonormal functions derived from spherical harmonics and radial basis functions. Therefore, the estimation of a specific physical property is broken down into individual contributions centered

on atoms. These contributions effectively capture correlations between atoms inside each localized environment. The assumption of locality, justified by the nearsightedness principle[35] of electronic matter, is very advantageous as it effectively reduces the complexity of the regression task, often encountered in ML schemes. The SOAP descriptor facilitates a thorough characterization of short-range order, allowing the identification of features that assess varying degrees of (dis)order in a given material.[36,37] However, it falls short in providing detailed quantification of medium-range order and network connectivity, crucial aspects for understanding amorphous systems. Long-range electrostatic interactions are well-recognized for their significant contribution to the characterization of several systems, including ionic systems,[38] interfaces with macroscopic polarization,[39] electrode surfaces,[40] and the field of nanoscience as a whole.[41] Capturing long-range effects without making any prior assumptions about the nature of the learning target is a challenging task that can be tackled utilizing a global representation of the studied system, such as Coulomb matrix,[42] many-body tensor representations,[43] and multiscale-invariant dictionaries.[44] A recent study conducted by Grisafi et al. has introduced a new framework for atomistic representations that can accurately capture long-range interactions by considering the local value of an atom-density potential. Thus, by combining a representation that captures long-range correlations with the transferability of an atom-centered additive model, this approach outperforms current ML methods and provides a conceptual framework for incorporating nonlocal physics into atomistic ML.[45] A particularly promising approach to compare amorphous structures is to combine descriptor that is designed to represent local atomic environments (such as SOAP) with another descriptor capturing the long-range effects (such as Coulomb matrix), to yield a global measure of similarity between structures.

Coming to the second point (ii), training MLIP continues to pose significant challenges. One of the primary difficulties lies in defining the optimal hyperparameters for the selected fitting technique and effectively selecting the appropriate training data set. When dealing with extensive data sets, the process of selecting training setups in a knowledgeable manner may become arduous. Active learning (AL), as defined in Refs. 18, 26, 46, 47 is an ML method in which a learning algorithm systematically finds the best hyperparameters and identifies the least number of training database necessary to develop a supervised ML model that achieves higher accuracy compared to models trained using a manual construction of the database.[26] AL has been employed in the generation of databases and the acceleration of the fitting process by relying on an iterative scheme that aims at enhancing the MLIP accuracy.[18,48–50] Recently, AL methods have been integrated with Gaussian process (GP)-based force fields, such as GAP,[51] and included in an FPMD framework.[52] This enables an on-the-fly fitting of force fields specifically tailored for a given studied system.[53,54] One prominent example within the field of glass materials modeling is the simulation of a Hafnium dioxide (HfO_2) system.[18] In this work,[18] the authors presented a technique that aims to achieve a clustering of an unlabeled data set of HfO_2 disordered snapshots by relying on a distance metric derived from the calculation of pairwise root mean square deviations between atomic positions. Subsequently, Bayesian optimization (BO) was used to select uncorrelated learning configurations from these obtained clusters. The results of this work show that the AL scheme was able to reach an energy fit tolerance of 5.0 meV/atom with a data set only containing ≈300 configurations. While useful, it is worth noting that the selection of the number of clusters is user-dependent. In addition, extending the applicability of the RMSD method to situations where the mapping of atoms between two structures is ambiguous poses significant challenges, as mentioned above.[31]

In our work, we provide an alternative modeling strategy to achieve the aforementioned goals, while minimizing the human bias. In particular, we introduce a hybrid similarity kernel that integrates local (SOAP)[34] and long-range (Coulomb matrix)[42] descriptors through an averaging kernel approach.[24] This method enhances the classification of amorphous configurations by leveraging both short- and long-range structural information, facilitating the construction of a representative training database. The similarity matrix is then converted into a distance matrix and fed to a classification algorithm that outputs a set of clusters, each containing structurally similar amorphous snapshots. Subsequently, we propose a strategy that automatically builds a database with a minimal number of uncorrelated configurations that will be used to fit GAP MLIP. Our AL workflow strategy involves the following steps: (1) Construction of a hybrid similarity matrix using SOAP and Coulomb matrix descriptors. (2) Conversion of the similarity matrix into a distance matrix for spectral clustering, iteratively adjusted to achieve optimal clustering. (3) Selection of diverse, uncorrelated training configurations from the identified clusters. (4) Fitting of the GAP MLIP using BO to refine the hyperparameters, ensuring the model meets predefined energy convergence criteria. This iterative process continues until the MLIP attains the desired accuracy This method is applied to two amorphous systems previously generated by FPMD.[55,56] We find that the AL approach is able to build databases with less than 175 configurations that efficiently describe the model at high temperature (liquid state), at room temperature (glassy state), and during the quenching process.

The optimized MLIPs achieve an energy accuracy of less than 2.0 meV/atom compared to FPMD reference data and are then used to generate large amorphous systems with several thousands of atoms, which were found to yield a good agreement with the reference FPMD data as well as the experimental measurements.

The paper is organized as follows: Methods are presented in Section 2, where we provide a description of the FPMD data sets, the clustering procedure, and the AL approach. In Section 3, we report our results on clustering and AL MLIP fit and discuss the real and reciprocal space properties of the obtained amorphous models. The conclusions of our work are presented in Section 4.

2 | METHODS

2.1 | FPMD data set

We focus on three data sets built on different glassy systems previously generated by FPMD.[55,56] The first data set (data set A) focuses on an amorphous $AsTe_3$ system that was generated by resorting to FPMD (more details available in Ref. 55 and Subsection 1.1 in supplementary material). In the original work, the modeling of $AsTe_3$ is conducted by resorting to the Becke, Lee, Yang, and Parr (BLYP)[57] exchange and correlation functional within the framework of density functional theory (DFT) and using a periodic cubic cell that has 240 atoms, consisting of 60 As and 180 Te. The glass was produced after a thermal annealing cycle followed by a residual stress calibration.[58]

The second data set (data set B) deals with amorphous TeO_2 generated in Ref. 56. We consider the system generated with BLYP exchange and correlation functional on 480 atoms (160 Te and 320 O). The TeO_2 glassy model was generated by quenching from the melt, followed by a subsequent residual stress calibration (more details available in Ref. 56 and Subsection 1.1 in supplementary material).

Finally, the third data set (data set C) is made of the second data set, to which configurations of the $\gamma - TeO_2$ crystalline system were added. $\gamma - TeO_2$ is the first stable polymorph obtained by crystallization of the glass. The $\gamma - TeO_2$ data set was generated by resorting to BLYP functional and a model of 48 atoms (16 Te and 32 O). In order to produce a high-temperature configuration of the crystal, the model was annealed in the constant temperature, constant volume ensemble (NVT) ensemble (Γ point) for 15 ps at T = 1000 K, 14 ps at T = 650 K, and 14 ps at T = 300 K.[56] For more details, refer to Subsection 1.1 in supplementary material.

To ensure the consistency of the database, we recalculated the DFT energies, forces, and virial stresses for all configurations in Refs. 55, 56 using a unique DFT setup and an energy cutoff of 1000 Ry.

2.2 | Descriptors and similarity measurement

2.2.1 | Smooth overlap of atomic orbitals

In the context of the SOAP descriptor, the representation of a given local atomic environment $\mathcal{Q}$ around an atom i within a cutoff distance r_{cut} involves the summation of Gaussian functions that represent the local density of atoms inside the environment $\mathcal{Q}$. These functions have a variance of σ^2 and are centered on each of the atoms belonging to $\mathcal{Q}$. The total atomic density is then given by:

$$\rho_{\mathcal{Q}}(r) = \sum_{i \in \mathcal{Q}} \exp\left(\frac{(x_i - r)^2}{2\sigma^2}\right), \tag{1}$$

where x_i is the position of the atom i. The SOAP kernel is thereafter characterized as the overlap of the local atomic neighbor densities, integrated over all three-dimensional (3D) rotations $\hat{R}$.

The measure of local atomic environment similarity requires the use of a kernel function[24] that is generally normalized in order to achieve a self-similarity value of unity for the given environment when compared to itself. The construction of the SOAP kernel, including the integration across all rotations, can be conducted analytically. First, the atomic neighbor density can be expanded as a function of spherical harmonics $Y_{lm}(\hat{r})$ and a collection of orthogonal radial basis functions $g_b(r)$ as follows:

$$\rho_{\mathcal{Q}}(r) = \sum_{blm} c_{blm} g_b(|r|) Y_{lm}(\hat{r}). \tag{2}$$

The expansion coefficients c_{blm} are collected and organized into a vector $\hat{P}(\mathcal{Q})$ that corresponds to the power spectrum. Subsequently, a polynomial kernel can be constructed to measure the similarity of two local atomic environments, Q and Q' as follows[34]:

$$\mathrm{k}(\mathcal{Q}, \mathcal{Q}') = \left[\ell [P(\mathcal{Q})]^{\top} \cdot \left[P(\mathcal{Q}')\right] + c_0\right]^d, \tag{3}$$

where the parameter c_0 is chosen to be equal to 1 to avoid homogeneous output results from the kernel, and the ℓ parameter is known as the kernel slope. To achieve high accuracy, we set the kernel degree to $d = 4$. This concept of similarity can also be seen as a distance measure between the two environments:

$$d(\mathcal{Q}, \mathcal{Q}') = \sqrt{2 - 2\mathrm{k}(\mathcal{Q}, \mathcal{Q}')}. \tag{4}$$

Within this definition, similar environments are close to each other (short distance) and vice versa. In this work, we resort to the SOAP descriptor as implemented in the DScribe software.[59] The radial basis function used in our

study is a polynomial basis set of order 4, as described in the original SOAP work. The polynomial basis function is guaranteed to decay to zero at r_{cut}.[24,34] SOAP calculations for measuring local atomic environment similarity are performed with a very effective numerical integration method, where the values of l_{max} and n_{max} are set to 8.

2.2.2 | Coulomb matrix

The Coulomb matrix[42] is a straightforward description that emulates the electrostatic interaction occurring among atomic nuclei within a given system. The pairwise, two-body matrix used in this approach is influenced by the Coulomb potential and serves to represent the atomic species and interatomic distances within atoms belonging to a finite system. The components of this matrix are defined by:

$$\mathcal{M}_{ij} = \begin{Bmatrix} 0.5Z_i^{2.4}, & \text{for } i = j \\ \frac{Z_i Z_j}{R_{ij}}, & \text{for } i \neq j \end{Bmatrix}, \quad (5)$$

where Z represents the atomic number, and R_{ij} denotes the Euclidean distance between atoms i and j. The determination of the shape of the diagonal terms was achieved by fitting the potential energy of neutral atoms.[60] The Coulomb descriptor is not designed in a manner that guarantees its dot product accurately depicts the overlap of atomic distributions. Therefore, a Laplacian kernel[61] is used to assess the long-range similarity between two environments as:

$$w(\mathcal{Q}, \mathcal{Q}') = \exp\left[-\zeta \|\mathcal{M}_{ij}(\mathcal{Q}) - \mathcal{M}_{ij}(\mathcal{Q}')\|\right], \quad (6)$$

where ζ is a hyperparameter that control the decay of the Laplacian kernel function. The similarity concept can be also converted to a distance measure using Equation (4). Similar to SOAP, the Coulomb descriptor is computed as implemented in DScribe package. In order to provide consistent dimensions for matrices representing systems with varying numbers of atoms, we consider zero-padding to have a unique matrix size corresponding to the biggest system in the data set.

For most applications, it is convenient to normalize the kernel. This ensures that the self-similarity of any environment is equal to one, resulting in the final kernel:[24]

$$\tilde{G}(\mathcal{Q}, \mathcal{Q}') = \frac{G(\mathcal{Q}, \mathcal{Q}')}{\sqrt{G(\mathcal{Q}, \mathcal{Q}) G(\mathcal{Q}', \mathcal{Q}')}}, \quad (7)$$

where $\tilde{G}(\mathcal{Q}, \mathcal{Q}')$ represents the normalized kernel and $G(\mathcal{Q}, \mathcal{Q}')$ refers to one of the kernels in Equation (3) or (6). For two identical environments, the kernel returns a value of 1, while it returns 0 for completely different ones. All values between 0 and 1 reflect the level of similarity.

2.3 | Hybrid similarity matrix

In order to build a similarity matrix between various configurations in a given data set, we construct a global similarity kernel. Figure S1 shows the general methodology used to construct the global similarity kernel. This global similarity description is either based on an averaged similarity measure between local atomic environments or between environments accounting for long-range interactions.

To this end, we consider a general case with two configurations containing N_A and N_B atoms. We compute for all pairs of atoms i in structure A and atoms j in structure B their SOAP similarity using Equation (3). As such pairwise similarity matrix of all the local atomic environments between the two configurations can be constructed[24]:

$$C_{ij}(A, B) = \mathrm{k}\left(\mathcal{Q}_i^A, \mathcal{Q}_j^B\right). \quad (8)$$

Subsequently, the average local similarity measure of the two configurations A and B can be achieved by computing the average of the C_{ij} matrix, yielding a single value. In practice, we resort to the average kernel as implemented in DScribe to measure the average similarity between the two structures (configurations) as follows[24]:

$$K_{SOAP}(A, B) = \frac{1}{N_A \times N_B} \sum_{ij} C_{ij}(A, B)$$

$$= \left[\ell\left[\frac{1}{N_A}\sum_i P\left(\mathcal{Q}_i^A\right)\right]^\top \cdot \left[\frac{1}{N_B}\sum_j P\left(\mathcal{Q}_j^B\right)\right] + c_0\right]^d \quad (9)$$

In this implementation, the computation of K_{SOAP} is achieved at a low cost by retaining the average SOAP fingerprint across all atom environments in both structures.

By repeating this procedure for all the configurations belonging to the considered data set, we construct the average SOAP similarly matrix of dimension $N \times N$, where N is the number of configurations in the data set.

The same methodology can also be applied to calculate an average measure of similarity between configurations

based on the Coulomb matrix:

$$K_{Coulomb}(A,B) = \frac{1}{N_A \times N_B} \sum_{ij} C'_{ij}(A,B) = \frac{1}{N_A \times N_B} \exp\left[-\zeta \sum_{ij} \|\mathcal{M}_{ij}(\mathcal{Q}^A) - \mathcal{M}_{ij}(\mathcal{Q}^B)\|\right], \tag{10}$$

where $C'_{ij}(A,B)$ represents the pairwise long-range similarity of all local atomic environments between two configurations calculated through the Laplacian kernel (Equation 6).

Following this procedure, we now have a global measure of the similarity between all the configurations within a given data set by relying on an atom-centered similarity measure (i) of local atomic environments (SOAP kernel) (ii) and of the long-range order (Coulomb matrix). Each of these quantities carries valuable information about the compared systems. We note that the sole use of either global similarity kernels as an input for ML classification of the atomistic configurations within the data set is not efficient as one loses a part of the information related to the structure during the averaging procedure. Instead, in this work, we propose a hybrid kernel approach that combines both the SOAP and Coulomb average kernels in a way to take advantage of the local and long-range descriptions of the studied systems. We show that this procedure leads to an efficient classification of amorphous configurations obtained during molecular dynamics (MD) simulations. The hybrid similarity kernel is defined as follows:

$$K_{Hybrid}(A,B) = (1-\delta) \times K_{SOAP}(A,B) + \delta \times K_{Coulomb}(A,B). \tag{11}$$

The hyperparameter δ controls the relative weight of the two kernels K_{SOAP} and $K_{Coulomb}$ in the description of a given system. This kernel can also be converted into a distance metric as follows[62]:

$$D(A,B) = \sqrt{2 - 2K_{Hybrid}(A,B)}. \tag{12}$$

$K_{Hybrid}(A,B)$ can be used to compute a similarity matrix between all the configurations and systems within a given data set, leading to a distance measure that can subsequently be used to classify the data set into clusters of structurally similar elements. The ultimate goal of this procedure is to extract from each cluster a given number of configurations that can then be gathered to build a minimalist data set for training MLIP. In this way, regardless of the size of the reference data set, one can always insure an efficient selection of the least number of required configurations that are needed to achieve a complete description of the studied systems and their properties. This can be achieved through a proper AL approach.

2.4 | Active learning

The objective of AL is to autonomously choose "n" diverse uncorrelated learning configurations from a reference data set[63,64] that will be later used for fitting MLIP using the GAP model.[2] The overall strategy implemented in this work is outlined in Figure 1. Once the hybrid similarity matrix is constructed on the FPMD data set (Figure 1, part A), it is first converted into a distance matrix using Equation (12) then fed to the spectral clustering algorithm[65–67] as implemented in Scikit-learn.[68]

Spectral clustering technique is based on a robust and powerful theoretical framework[69] that does not rely on any assumptions about the overall structure of the data. This algorithm has the ability to converge toward optimum solutions and exhibits strong performance when applied to sample spaces of variable shapes or data sets with a nonconvex nature.[70] In practice, the data clustering issue is considered as a problem of graph partitioning, where each data point in the data set is represented as a vertex. The weight of the edge linking two vertices corresponds to the similarity value between the respective data points.[71] The obtained network can be further decomposed into related components (or domains) using specific graph-cut techniques.[70] The obtained components are then referred to as clusters. This technique requires a prior definition of the number of clusters that the algorithm needs to achieve. Instead, here we iteratively vary the total number of clusters (N_c), and compute for each clustering iteration (I') the average distance between all the clusters as:

$$< d_{I'} > = \frac{2}{N_c(N_c-1)} \sum_{\substack{s,s'=1 \\ s \neq s'}}^{N_c} d_{s,s'} \tag{13}$$

where, $d_{s,s'}$ represent the distance between clusters s and s', which can be obtained as follows:

$$d_{s,s'} = \frac{1}{N_s \times N_{s'}} \sum_{\substack{A \in \{s\} \\ B \in \{s'\}}} D(A,B), \tag{14}$$

N_s and $N_{s'}$ refer to the total number of configurations in clusters s and s', respectively, and $D(A,B)$ denotes the distance between configurations A and B as computed in Equation (12). The total number of clusters (N_c) is increased with step of one until a distance convergence

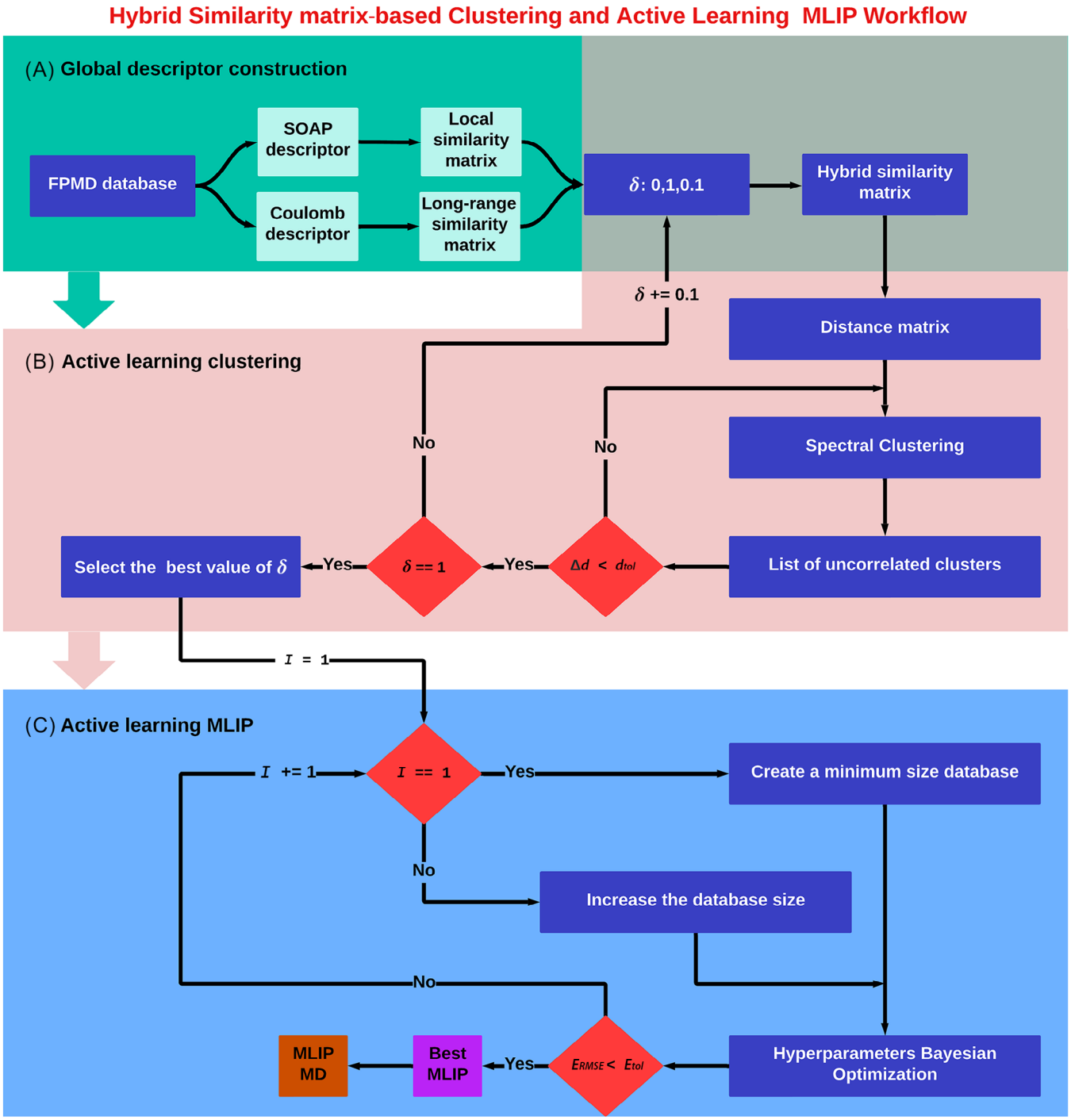

FIGURE 1 Hybrid similarity-based clustering and active learning machine learning interatomic potential (MLIP) workflow: (A) The computation involves the assessment of local (SOAP, smooth overlap of atomic positions) and long-range (Coulomb) similarities based on the average kernel method. The global similarity between two configurations is computed for a given value of δ and afterward transformed into a distance matrix. (B) The distance matrix is provided as input to the spectral clustering algorithm iteratively until the change in the average distance between obtained clusters reach the desired threshold d_{tol}. This clustering procedure is repeated for different values of δ. The value of δ in Equation (11) is varied from 0 to 1 with a step of 0.1 and at the end the best value of δ is selected. (C) The training and test data samples are built successively from the selected δ clusters in order to fit the Gaussian approximation potential (GAP) model until the desired level of accuracy is attained. During each cycle I of data sampling, Bayesian optimization does real-time hyperparameter adjustment of the GAP model. After convergence the best MLIP is used to run MD.

threshold d_{tol} is reached (Figure 1, part B). The change in the average distance between all the clusters between two consecutive iterations $\Delta d = < d_{I'} > - < d_{I'+1} >$ should be less than d_{tol}. Hence, we stop the clustering procedure at the smallest number of clusters that satisfy this convergence criteria. The obtained list of uncorrelated clusters can further be visualized in a two-dimensional (2D) map by resorting to the multidimensional scaling (MDS) method[72–74] as implemented in Scikit-learn library.[68] This map illustrates the spatial connections between different data points where items that share similarities are positioned in close proximity to each other, while those that are distinct are positioned farther away from one another.

After extracting the list of uncorrelated clusters, we now focus on fitting the GAP MLIP by relying on an AL scheme (Figure 1, part C). This scheme aims at constructing a minimalist data set and at the same time optimizing the GAP hyperparameters. First, we extract from each uncorrelated cluster a fraction I_{perc} of configurations in a way to build a small database representative of the structural variety in the reference database. I_{perc} is a user-provided input and depends on the size of the reference data set. Next, this small database is split into a training set (70% of the data) that will be used to train the GAP model and a test set (30% of the data) that will serve as a measure of the accuracy of the achieved MLIP through the calculation of the root mean square error (RMSE) on the predicted energies compared to the reference FPMD data.

Within GAP methodology, the potential energy surface is decomposed into individual energy points, where each point is expressed as a sum of a local atomic energy functions, denoted ϵ_i. The functional form of ϵ_i depends on the geometry of the local atomic environment surrounding the central atom i, within a cutoff radius of r_{cut}. Subsequently, the Gaussian regression procedure is used to establish a model for the total energy as a linear combination of nonlinear kernel functions. GAP has been effectively used in the modeling of glasses, liquids, and crystals.[18,19,55] In the present study, we resort to standard GAP MLIP approach, as implemented in QUIP,[2] where the structure is described through a combination of two descriptors: SOAP descriptor and a nonparametric two-body distance descriptor. This approach is implemented to mitigate the occurrence of nonphysical clusters of atoms. It should be noted that the Coulomb matrix is not used when fitting GAP, as it is solely used to build the hybrid kernel that will allow the clustering of the amorphous configurations.

The GAP model, has intrinsically several hyperparameters that need fine-tuning in order to determine the optimal model for a given training data set. BO is an effective methodology for the automatic optimization of hyperparameters in ML models that are computationally expensive.[75,76] The BO framework comprises a surrogate model that represents the objective function and an acquisition function that facilitates the selection of the next set of hyperparameters to be sampled. In practice, the BO algorithm conducts a series of explorations over the hyperparameter space in order to construct a surrogate model based on an error metric. Subsequently, a sequence of exploitation is conducted using the insights acquired during the first step, with the objective of enhancing the surrogate model and obtaining more refined samples of hyperparameters that have the potential to minimize the error measure. The error is measured on the test set as the deviation of the predicted energies from the reference FPMD data. The energy convergence threshold (E_{tol}) is user-defined and is set to 2 meV/atom in this work. In the case the optimal GAP model refined using BO fails to meet the desired level of accuracy, another loop of AL is initiated where the size of the database initially set to I_{perc} of the reference data set, will be increased by I_{pers} (or any other user defined fraction) (Figure 1, part C). In general, for the nth AL iteration, the database size should correspond to $n \times I_{perc}$, until reaching 100% of the reference data set size.[18] The AL workflow stops when the RMSE in energy prediction for the optimal GAP model reaches or falls below E_{tol}, or when the maximum database size is reached. As such, upon convergence, we obtain the smallest possible training database together with the optimal GAP hyperparameters that are required to achieve an accurate GAP MLIP potential.

2.5 | ML modeling of glassy systems

The MLIP resulting from the AL fit is subsequently used to run MD simulations by resorting to the LAMMPS classical molecular dynamics simulation code.[77] In particular, we produce $AsTe_3$ and TeO_2 glassy models by quenching from the melt and consider for each system four system sizes (for $AsTe_3$: 240, 1920, 6480, and 30 000 atoms and for TeO_2: 480, 3840, 12 960, and 30 720 atoms), with starting configurations built randomly. The MLIP MD are performed in the NVT or NPT ensemble using a Nosé–Hoover thermostat to ensure thermal control[78,79] and adopting a time step of 0.5 fs.

In the case of $AsTe_3$ system, the glass is produced through a thermal annealing cycle as follows: 5 ps at T = 300 K, 5 ps at T = 500 K, 100 ps at T = 650 K, 100 ps at T = 500 K, and 100 ps at T = 300 K. For the TeO_2 system, the following thermal annealing cycle was implemented: 5 ps at T = 300 K, 5 ps at T = 500 K, 5 ps at T = 750 K, 100 ps at T = 1000 K, 100 ps at T = 750 K, 100 ps at T = 500 K, and 100 ps at T = 300 K. Subsequently, all the studied systems

undergo further annealing at T = 300 K for a duration of 100 ps in the NVT ensemble at lattice constants that correspond to a pressure of 0 GPa, which were obtained from Refs. 55, 56.

3 | RESULTS

3.1 | Clustering

The first focus of our study is to investigate the outcome of the AL clustering workflow applied to FPMD data sets. In order to evaluate the contributions from both long-range and local atomic environments descriptors, the value of δ in Equation (11) is varied from 0 to 1 with a step of 0.1. For $\delta = 0$ and $\delta = 1$, the hybrid kernels correspond to exactly the average SOAP and Coulomb kernels, respectively. For each δ value, the optimal number of clusters is obtained by the AL procedure that achieves a change in the average distance between all the clusters below d_{tol} which is set to be very small due to the nature of the studied systems. Indeed, the similarity between amorphous configurations is quite high, especially those belonging to the same temperature plateau, which requires very low values in order to distinguish them. In the case of $AsTe_3$ (A data set), d_{tol} was set to 10^{-4}, while for TeO_2 systems (B and C data sets), d_{tol} was set to 10^{-3}.

As the choice of δ is arbitrary, one have to decide which value is the more appropriate to achieve a good classification of the studied amorphous configurations. This can be achieved as follows: for a given δ value, we calculate for each cluster (s) the standard deviation of the energies (σ_s) of the configurations belonging to that cluster and then take the average standard deviation over all the clusters (N_c): $E = \frac{1}{N_c}\sum_{s=1}^{N_c}\sigma_s$. This metric allows to have a direct access to the level of dispersion of the achieved classification. In the case clusters contain very different configurations, the average energy standard deviation will be high, and vice versa. Therefore, the best δ value is the one corresponding to the lowest average standard deviation. The results of the evolution of the average energy standard deviation are shown in Figure 2.

In the case of TeO_2^G system (data set B), the clustering results are shown in Figure S2 with various δ values. When $\delta = 0$, the clustering is only dependent on the local atomic environment similarity (SOAP) matrix. In this case, besides configurations belonging to the temperature plateau at T = 300 K with 0 GPa, we observe a significant number of widely dispersed clusters with substantial overlap over several temperature plateaus resulting in a high $< \sigma >_E$ as shown in the bottom panel of Figure 2. As such, sampling from these clusters containing a wide energy dispersion might lead to an ill representation of

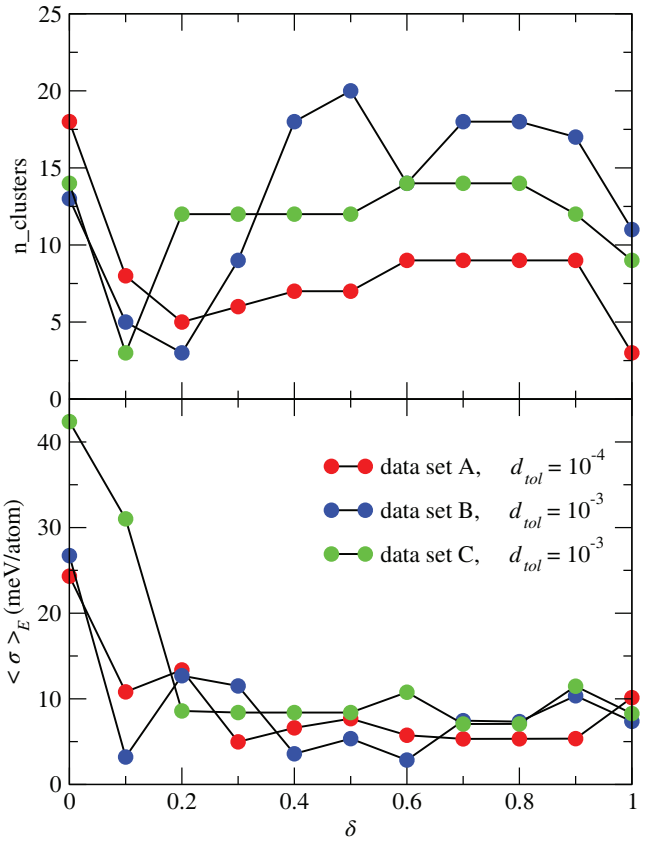

FIGURE 2 The evolution of number of clusters (top panel) and the average standard deviation on the energy $< \sigma >_E$ (bottom panel) as a function of increased the value of δ for $AsTe_3$ (data set A, red line), TeO_2^G (data set B, blue line), and $TeO_2^{G+\gamma}$ (data set C, green line). The considered tolerance of the change in the average distance between obtained clusters for each system is also displayed.

the overall data. These results indicate the limitations of using SOAP as a global descriptor for classification of disordered systems.[80] For $\delta < 0.3$, the inclusion of a small fraction of long-range similarity matrix leads to a substantial reduction of the number of identified clusters (see Figure 2). These identified clusters show a well-organized temperature-dependent structure, as shown in Figures S2 and S3. Specifically, each temperature plateau form one cluster (or a very small number) which can be explained by the insufficient information regarding the long-range similarity between configurations. The number of clusters exhibits a maximum value around $\delta = 0.5$ before decreasing for larger δ values.

In the case of the $TeO_2^{G+\gamma}$ systems (data set C), similar trends to those seen in the case of pure TeO_2^G are observed in Figures 2, S3, and S4. We note that the inclusion of the gamma-crystal configurations in the TeO_2 database results in the formation of well-distinguished clusters within each temperature plateau when $\delta \geq 0.2$. This effect is primarily driven by the dissimilarity between the gamma-crystal

and the glass system, leading to a decrease in the overall average similarity between configurations. In addition, a decrease in the number of clusters is noticed when the value of δ exceeds 0.8, similar to that observed in the case of the pure TeO_2 data set.

Finally, in the case of $AsTe_3$ systems (data set A), a very high similarity value is observed between the amorphous configurations, yielding a scattered and small number of clusters when $d_{tol} = 10^{-3}$ is used, as shown in Figure S5. To overcome this limitation, d_{tol} of 10^{-4} is instead used, thereby allowing for a better classification of configurations belonging to the same temperature plateau as displayed in Figure S6. Figure 2 shows also a similar evolution of the $< \sigma >_E$ and the number of clusters as a function of δ compared to TeO_2^G and $TeO_2^{G+\gamma}$ data sets.

Based on these results, δ values ($0 \leq \delta \leq 1$) leading to the smallest average standard deviation of energy were identified as optimum values. Consequently, the best values of δ were determined to be 0.6, 0.7, and 0.3 for data sets B, C, and A, respectively.

Figure 3 displays the final classification of clusters used to train the MLIP model for each of the studied data sets, together with their respective clustering hyperparameters. In addition, the organization of these clusters can be graphically visualized in a 2D plot using the MDS method, as shown in Figure 4.

3.2 | AL MLIP

After achieving clustering of the FPMD data sets, one needs to select the most effective learning configurations for fitting stable and accurate MLIP. This can be accomplished through an AL process, with the aim of attaining a predefined energy convergence E_{tol} set to 2 meV/atom and a stable MLIP. In order to assess the performances of our scheme against common MLIP fitting practice, we define the following MLIP potentials: MLIP(0) is obtained with data selection based on clusters achieved with an optimized δ value. MLIP(1) is optimized based on a clustering of the amorphous snapshots with the hybrid similarity kernel with $\delta = 0$. This choice corresponds to a clustering solely based on SOAP descriptor as a global similarity metric. Finally, we fit a potential, hereafter MLIP(2), using all available data for $AsTe_3$ (data set A) and TeO_2 (data set B) glassy systems with the same hyperparameters obtained for MLIP(0). Comparing MLIP(0) and MLIP(2) results will be informative about the completeness of the minimal database construction approach.

Focusing on the case of MLIP(0) (produced with data selected using an optimal δ value), the AL yields excellent training configurations with a minimal number of AL iterations, namely, three iterations for both A and B data sets. While for data set C, four AL iterations were required to achieve a training database and GAP hyperparameters that satisfy the convergence criteria. Correspondingly, the MLIP(0) converged database sizes (N_{train}) contain 97, 127, and 175 configurations and correspond to 10%, 10%, and 14% of the reference data set size for A, B, and C data sets, respectively. These results demonstrate the efficiency of the AL procedure that goes beyond the bias and limitations that a user can induce in manually building a data set based on a trial-and-error approach.[18,63] Coming to MLIP(1), the AL builds a minimal data set and GAP hyperparameters that achieve the energy convergence criteria after six iterations in the case of $AsTe_3$ (data set A) and two iterations in the case of TeO_2 (data set B). The obtained database sizes amount to 21% and 7% for $AsTe_3$ and TeO_2, respectively. Details of the converged hyperparameters for MLIP(0) and MLIP(1) are provided in Table S1. Moreover, Table S2 shows the computational costs and memory requirement associated with the AL clustering, and AL MLIP optimization processes.

The validation plots of the active learned MLIP(0) are displayed in Figure S7 and show that it yields typical linear correlations between the predicted GAP energies and the corresponding DFT energies. For MLIP(0), MLIP(1), and MLIP(2), the RMSE for the predicted energies are found to be less than 2.0 meV/atom as set by the AL energy convergence criteria. In addition, the convergence of GAP forces compared to DFT falls below an RMSE value of 0.26 eV/Å (see Figure S7). We note that a GAP MLIP for glassy $AsTe_3$ is already available in the literature,[55] however, it was produced by training on a data set made of about 900 configurations. Overall, we here show that a properly built clustering and AL procedure can outperform manual fitting of MLIP leading to minimalist data sets as well as well-converged MLIPs.

3.3 | MLIP-MD

We now access the performances of our MLIP(0) in producing amorphous systems and compare it to those achieved with MLIP(1) and MLIP(2). To this end, MD simulations are performed to produce glassy systems using the LAMMPS package with each of the obtained MLIPs (MLIP(0), MLIP(1), and MLIP(2)). The initial configurations are all made of atoms randomly placed in a cubic simulation cell corresponding to the measured experimental density.[55,56]

In the case of pure TeO_2 glass models, we consider the MLIPs built on data set B (TeO_2^G, $MLIP^B(0)$) or data set C ($TeO_2^{G+\gamma}$, $MLIP^C(0)$) and generate systems with sizes 480, 3840, 12 960, and 30 720 atoms. For the sake of

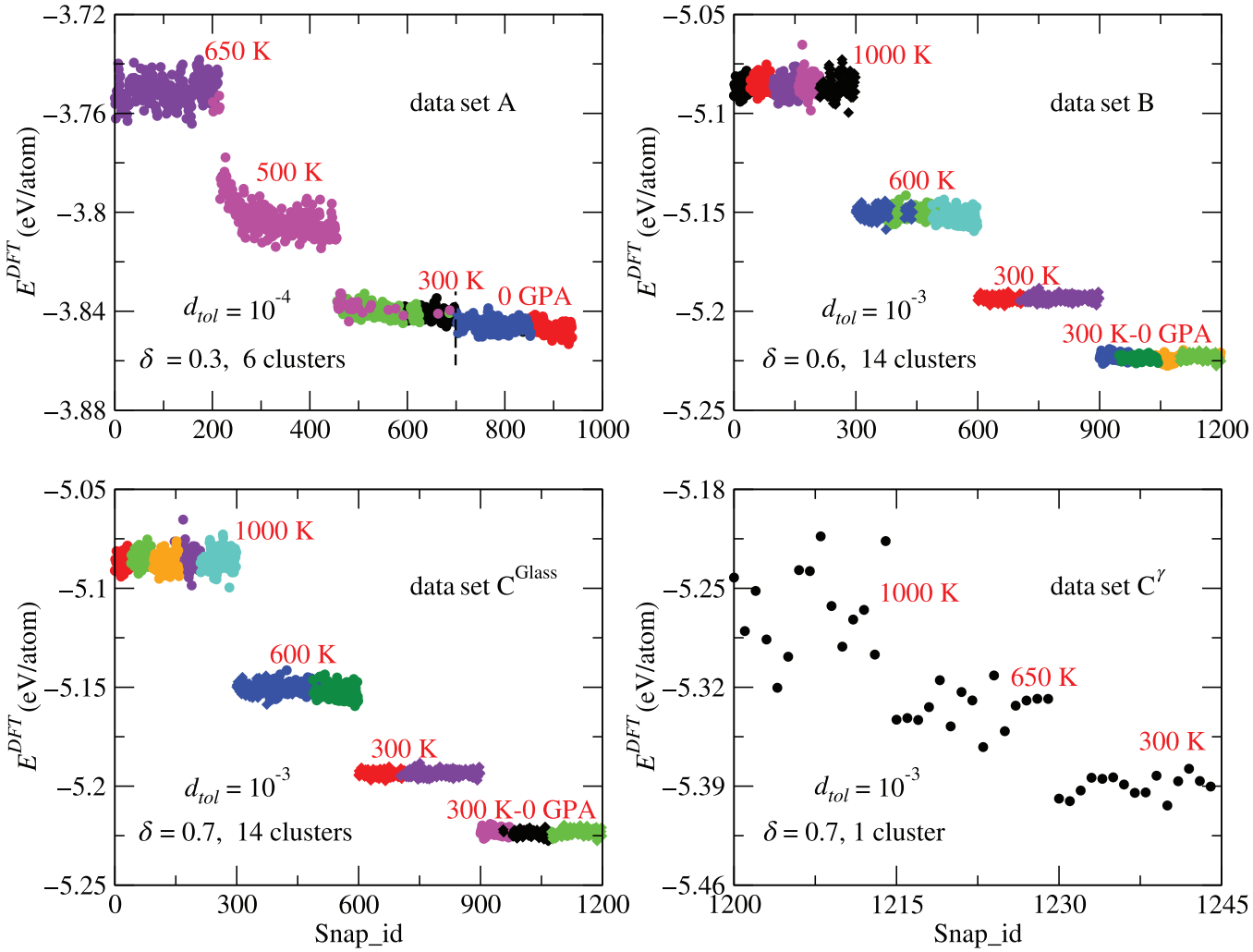

FIGURE 3 Clusters used to generate a machine learning interatomic potential (MLIP) for each of the studied systems with their corresponding value of δ. Point distributions for data sets of 937, 1200, and 1245 points, for $AsTe_3$ (data set A), TeO_2^G (data set B), and $TeO_2^{G+\gamma}$ (data set C), respectively. Dataset C is slit into the bottom two panels for a better visibility of the configurations of $\gamma - TeO_2$. Points are colored according to the cluster to which they are assigned.

comparison, systems with sizes 480 and 3840 are also generated considering $MLIP^B(1)$ and $MLIP^B(2)$. In the case of $AsTe_3$ systems, we consider $MLIP^A(0)$ and produce models with 240, 1920, 6480, and 30 000 atoms. Similarly, $AsTe_3$ models of 240 and 1920 atoms were produced with $MLIP^A(1)$ and $MLIP^A(2)$. All the initial configurations were subject to a thermal annealing cycle as described in Subsection 2.5. Details of the computational cost associated with MD simulations via MLIP (MLIP-MD) as implemented for each system size are available in Table S2.

3.3.1 | Total and partial structure factors

Focusing on obtained results from MLIP(0), Figure 5 depicts the computed total X-ray structure factor obtained from the original FPMD models and the large MLIP(0) models with ≈30 000 atoms, compared to experimental data of pure TeO_2 and $AsTe_3$ glassy systems. Similarly, Figure 6 shows the results of the partial Faber–Ziman (FZ) structural factors. The results of total $S(q)$ and FZ partial structure factors for other system sizes obtained from our MLIP(0)-GAP models are provided in Figures S8 and S9, respectively. Within typical statistical fluctuations, our findings indicate excellent agreement between the FPMD, the MLIP(0)-GAP models, and the experimental data over the whole range of reciprocal space. When comparing MLIP(0) to FPMD results, the positions and intensities of the peaks in the total $S(q)$ for $AsTe_3$ glassy system are well reproduce. Furthermore, as similar results are obtained for all systems with >240 atoms, one can infer the stability of the produced $MLIP^A(0)$. Modest changes in intensity and position of the first, second, and third peaks

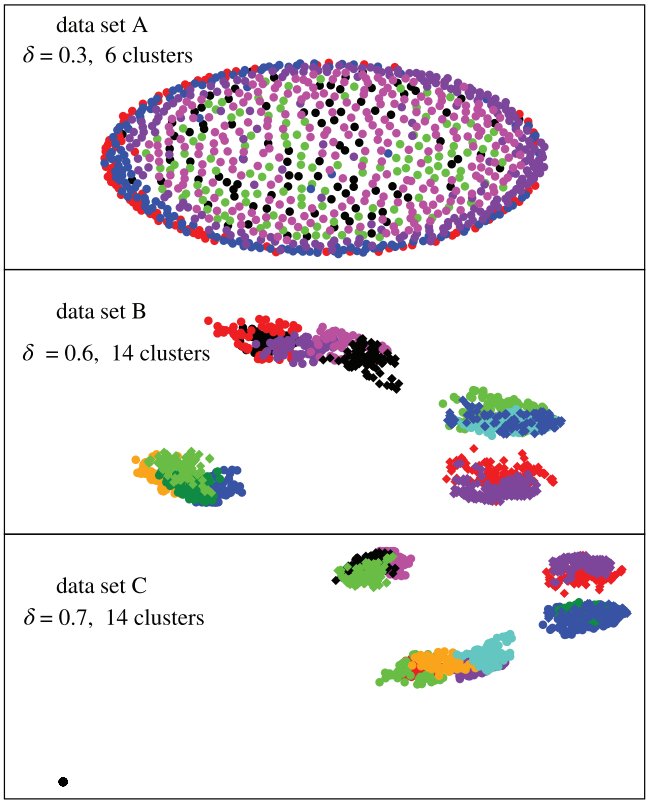

FIGURE 4 Two-dimensional (2D) plot of the hybrid similarity matrix corresponds to best number of clusters found in each $AsTe_3$ (data set A), TeO_2^G (data set B), and $TeO_2^{G+\gamma}$ (data set C). The color of each point depends on the cluster to which it belongs.

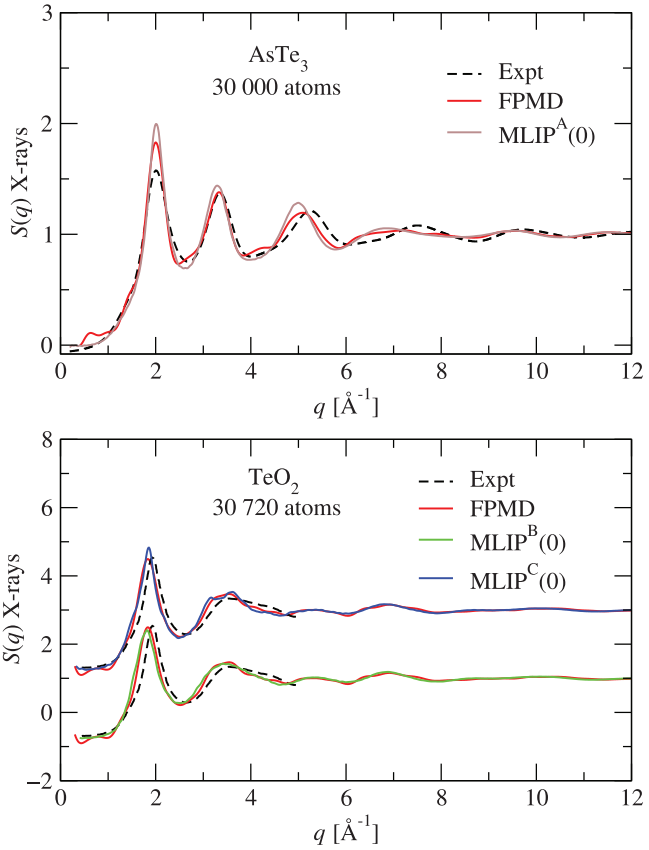

FIGURE 5 Total X-ray structure factor for amorphous TeO_2 and $AsTe_3$ glassy systems at T = 300 K. The experimental results (black dashed lines) are compared to the calculated first-principles molecular dynamics (FPMD) $S(q)$ (red lines) and to the $S(q)$ for the large machine learning interatomic potential (MLIP) systems with ≈30 000 atoms (brown line for data set A, green line for data set B, and blue line for data set C). The curves are shifted vertically for clarity.

in the calculated total $S(q)$ are observed for systems bigger than 240 atoms, hinting toward a very minor size effect in the description of the atomic structure of the $AsTe_3$ glassy system. For $q > 5$ Å^{-1}, one can notice slight discrepancies between the modeled $S(q)$ and the experimental data. These discrepancies were attributed to Te–Te correlation interactions and call for the use of a higher level of theory for describing the van der Waals (vdW) interactions within the DFT framework.[55] Generally, the fine details of structure of Te-rich amorphous chalcogenides have shown significant sensitivity to the adopted DFT exchange and correlation functional and the used vdW scheme.[81,82]

When considering the network topology of $AsTe_3$, the analysis of the partial FZ structure factors (see Figures 6 and S9) gives access to the respective contributions of different chemical species. Across all system sizes, there is a notable concurrence between our MLIP(0) $S^{FZ}_{Te\text{-}Te}(q)$ and $S^{FZ}_{As\text{-}Te}(q)$ structural factor and those obtained from FPMD. However, the MLIP(0) $S^{FZ}_{As\text{-}As}(q)$ structure factor show a dependence on the system size where a decrease in the fluctuations for $q < 4$ Å^{-1} is observed for models with 6480 and 30 000 atoms. This behavior is mainly due to the low number of As atoms in the small models that lead to strong statistical fluctuations. Therefore, system sizes larger than those studied in this work might lead a better convergence of the low q range in the $S^{FZ}_{As\text{-}As}(q)$.

Coming to TeO_2^G (MLIP^B(0), data set B) and $TeO_2^{G+\gamma}$ (MLIP^C(0), data set C) glassy models, the small ML-GAP system with 480 atoms shows a good reproduction of the FPMD reference data in terms of position and intensity. In addition, for models with more than 480 atoms, the intensity of the first $S(q)$ peak shows a slight increase for both MLIP^B(0) and MLIP^C(0) models which can be attributed to an increase in $S^{FZ}_{Te\text{-}Te}(q)$ first peak intensity as shown in Figure 6. We note that, in the case of the largest MLIP^B(0) model (30 720 atoms), the total $S(q)$ show a very good agreement with the FPMD obtained results. Furthermore, we note a very good agreement between calculated FPMD and MLIP(0) FZ structural factors. Overall, the good agreement between the MLIP(0) and the FPMD

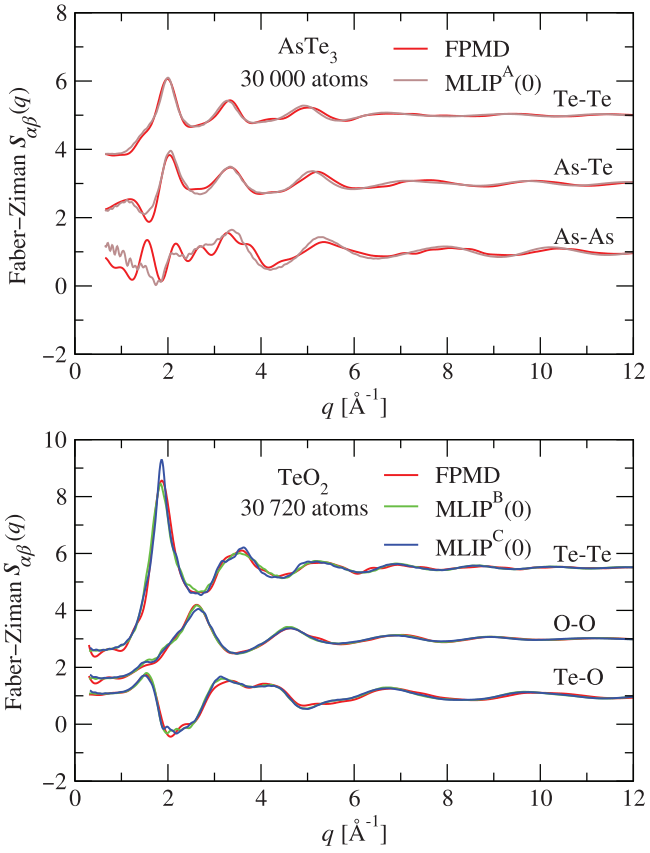

FIGURE 6 The Faber–Ziman partial structure factors $S_{\alpha\beta}(q)$ for glassy TeO_2 and $AsTe_3$ systems at T = 300 K obtained from first-principles molecular dynamics (FPMD) (red lines) and machine learning interatomic potentials (MLIP) for the systems with ≈30 000 atoms (brown line for data set A, green line for data set B, and blue line for data set C). The curves are shifted vertically for clarity.

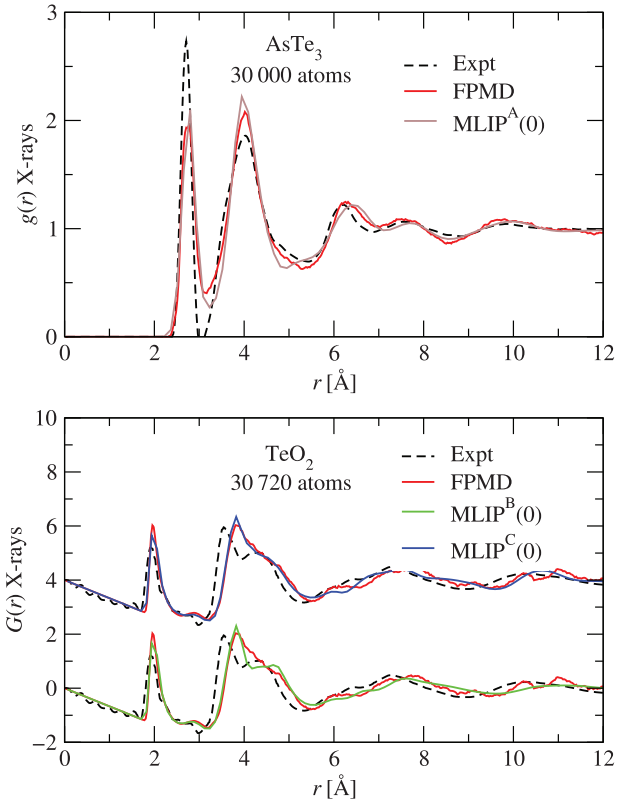

FIGURE 7 Total pair correlation function for TeO_2 and $AsTe_3$ systems at T = 300 K. The experimental results (black dashed lines) are compared to the first-principles molecular dynamics (FPMD) calculated $G(r)$ (red lines) and the large machine learning interatomic potential (MLIP) glassy system with 30 000 atoms (brown line for data set A, green line for data set B, and blue line for data set C). The curves are shifted vertically for clarity.

data for various system sizes reflects a very limited size effect in these systems.

3.3.2 | Total and partial pair distribution functions (PDFs)

The calculated total and partial X-ray PDFs on our MLIP(0) models are depicted in Figure 7 (and Figure S10), and Figure 8 (and Figure S11), respectively, and are compared to the reference FPMD data as well as experimental results. Overall, the measured and calculated total PDFs ($G(r)$) exhibit a pattern characteristic of amorphous materials. The computed $G(r)$ from the FPMD and MLIP(0) models exhibit a good level of agreement over the entire real space range for all the studied systems, within typical statistical fluctuations. When compared to the experimental results, the computed $G(r)$'s reproduce the main experimental features up to small discrepancies in the peak positions and intensities. Specifically, the level of agreement between ML-GAP and FPMD PDFs is determined by computing the Wright parameter (R_X)[83] as defined by Equation (15).

$$R_X = \left\{ \frac{\sum_i \left[G^{\mathrm{FPMD}}(r_i) - G^{\mathrm{ML\text{-}GAP}}(r_i) \right]^2}{\sum_i \left[G^{\mathrm{FPMD}}(r_i) \right]^2} \right\}^{\frac{1}{2}}. \quad (15)$$

In this formula, $G^{\mathrm{FPMD}}(r_i)$ and $G^{\mathrm{ML\text{-}GAP}}(r_i)$ represent the FPMD and ML-GAP calculated total X-ray PDF at a given distance r_i, respectively.

In the case of TeO_2 systems, the obtained R_X values for MLIP^B(0) and MLIP^C(0) models are equal to 44.4% and 44.6%, respectively, for systems with 480 atoms. For all systems with more than 480 atoms, the R_X value decreases to ∼42% and ∼40% in the case of MLIP^B(0) and MLIP^C(0), respectively (see Table S3). This minor reduction in R_X obtained from MLIP(0) large models, hints toward a very slight improvement in describing the overall topology of TeO_2 glass as we increase the system size.

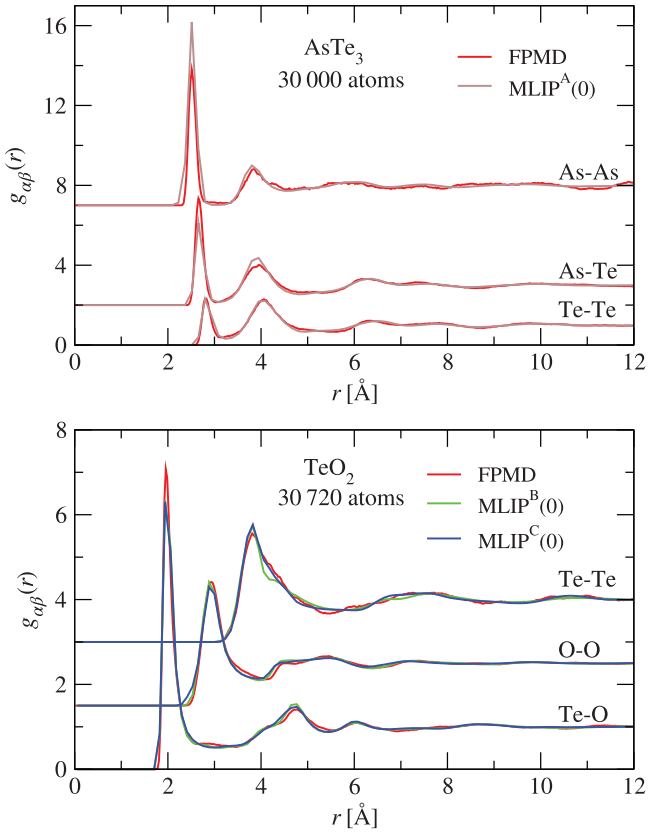

FIGURE 8 The partial pair correlation functions $g_{\alpha\beta}(r)$ for TeO_2 and $AsTe_3$ glassy systems at T = 300 K obtained from first-principles molecular dynamics (FPMD) (red lines) and large machine learning interatomic potential (MLIP) models with ≈30 000 atoms (brown line for data set A, green line for data set B, and blue line for data set C). The curves are shifted vertically for clarity.

Focusing on comparing the calculated total X-ray $G(r)$ obtained from MLIP^B(1) (trained using configurations selected assuming $\delta = 0$) and MLIP^B(2) (trained using the whole reference data set) models with sizes of 480 and 3840 atoms to that obtained from FPMD simulation (refer to Figure S12). The computed R_X for MLIP^B(2) model shows very similar values for both system sizes (480 and 3840 atoms) as shown in Table S3. Interestingly, these values are also very close to those achieved with MLIP^B(0) indicating that this later, although trained with much less training configurations, is able to correctly reproduce the amorphous structure of amorphous TeO_2. As for MLIP^B(1), the obtained R_X values show higher values compared to those achieved with MLIP(0) models. This result, shows that MLIP^B(1) models achieved with $\delta = 0$ are less accurate compared to those achieved with the optimized δ value (MLIP(0)). Furthermore, besides being high, the R_X value obtained with the MLIP^B(1) model increases from 51.2% to 61.6% (refer to Table S3) when the system size goes from 480 to 3840 atoms, reflecting the occurrence of size effect. This increase is mainly due to an overestimation of the intensity of the peaks at the medium-range distances.

Considering glassy TeO_2 system based on B and C data set (MLIP^B(0) and MLIP^C(0)) models, for r values <5.0 Å, the first and second $G(r)$ peaks exhibit little changes in intensity compared to the FPMD reference $G(r)$. It is worth noting that the experimental total $G(r)$ shows a shoulder at around 4.3 Å, which is not accurately captured by the reference FPMD model. However, this shoulder is reproduced in MLIP^B(0) and MLIP^B(2) models (see Figure S12). Here, MLIP^B(1) fails to reproduce this feature and at the same time leads to an overestimation of peak intensities for $r > 5$ Å as observed in the 3840 atoms model compared to MLIP^B(0) and MLIP^B(2) models. By looking at the partial pair correlation functions (Figures 8, S11, and S13), we observe a good concordance between the MLIP(0), MLIP(2), and FPMD results. Regardless of the system size, we notice that the MLIP^B(0), MLIP^C(0), and MLIP^B(2) models show a Te–Te partial PDFs with a shoulder at around 4.3 Å, corresponding to the shoulder observed at the same position in the measured total $G(r)$, while this shoulder is not reproduced in MLIP^B(1) result.

Coming the $g(r)$ of $AsTe_3$ systems, the first peak observed for the MLIP^A(0) models show a very good agreement with the reference FPMD data, except for systems with more than 6480 atoms, where a slight intensity overestimation is noticeable. The second $G(r)$ peak from the MLIP(0) models, exhibits a slight sharpening of the peak compared to the FPMD results for all the studied systems. For distances larger than 5.5 Å, FPMD and MLIP(0) models show a good agreement with the reference data and accurately capture the positions and intensities of the experimental peaks that occur within this distance range.

The computed R_X value on MLIP^A(0) and MLIP^A(2) simulations are around ∼7% as presented in Table S3 for the two considered system sizes (240 and 1920 atoms). This can be correlated, again, to the efficiency of the presented AL scheme in obtaining simple and accurate ML-GAP models fitted on small databases that are representative of the whole reference data set. Considering MLIP^A(1), we find R_X equal ∼6.7% for model size with 240 atoms, showing a close value to those achieved with MLIP(0). However, by looking at the partial PDFs of amorphous $AsTe_3$ in Figure S13, we find that the MLIP^A(1) shows a good estimation of the intensity and the first peak position of the As–As correlations, while it overestimates the intensity of the second peak and underestimates its position. As for the Te–Te and As–Te correlations, they are found to be in a good agreement with those achieved with the MLIP^A(0) and reference FPMD data. Nevertheless, when producing models with 1920 atoms and larger sizes using MLIP^A(1), the potential turned out to be unstable as the amorphous models undergo severe segregation that

leads to wrong structures. This instability MLIP^{A}(1) might be explained by the fact that the training database was constructed by sampling clusters (achieved solely using SOAP as a global descriptor) that are widely dispersed and have a significant overlap of configuration across multiple temperature plateaus.

Overall, these results demonstrate that clustering amorphous configurations using a global similarity descriptor that includes both local and global descriptors with an appropriate weight δ, together with an AL approach leads to accurate and stable MLIPs that outperform those fitted using configurations extracted from clusters obtained by only considering local descriptors. In addition, this minimal size database approach leads to the same results when a large size database is used to fit the MLIP, therefore highlighting the high potential of data-distillation strategies within the field of computational materials science.

3.3.3 | Coordination number and local environment analysis

The coordination numbers obtained through integration of the partial X-ray PDFs ($g_{\alpha\beta}(r)$) can provide insights into the network topology and connectivity. The average coordination number $n_{\alpha}(r)$, where α refers to the considered chemical species is shown in Table 1. Furthermore, the decomposition of the of atomic local environments into different l-fold ($l = 1, 2, 3, 4,$ or 5) is presented in Tables S4 and S5.

Looking at TeO_2 systems, n_{Te} and n_O coordination numbers found from MLIP^{C}(0) and MLIP^{B}(0) models show a very good agreement with the FPMD reference data for all the considered systems sizes, indicating the very minor size effects on the local environments. It is worth noting that MLIP^{B}(0) and MLIP^{B}(2) show very similar values that are in excellent agreement with the FPMD results as presented in Table 1. Furthermore, Table S5 shows that irrespective of the model size the various l-fold environments of models MLIP^{B}(0) and MLIP^{B}(2) do not exhibit significant evolution as a function of the model size within typical statistical fluctuations. In the contrary, we find that MLIP^{B}(1) models overestimate the average Te and O coordination numbers compared to the FPMD reference data. This overestimation is due to the reduction of the fractions of onefold O and threefold Te atoms, while those of threefold O and 5-Te exhibit a considerable increase compared to the FPMD results.

Coming to $AsTe_3$ MLIP^{A}(0) systems, we find that n_{As} is in very good agreement with that obtained from the FPMD data for systems with sizes up to 12 960 atoms. For larger size systems a slight increase of n_{As} is observed and can be correlated to the slight increase observed in the $g_{As-As}(r)$. Irrespective of the system size, n_{Te} is found slightly larger than the reference FPMD results as shown in Table 1. The atomic local environments of As and Te atoms (see Table S4) indicate a few changes that could be attributed to the improved description of the As–As correlations in the MLIP^{A}(0) model. Remarkably, it is observed that the fraction of threefold Te increases by approximately 10% in the MLIP(0) model, while the proportion of twofold Te decreases compared to the FPMD results. This result is in agreement with the occurrence of threefold Te in amorphous $AsTe_3$, which were longly addressed in the literature.[84,85] Unsurprisingly, the MLIP^{A}(2) reproduce close result to MLIP^{A}(0) and FPMD ones, while an overestimation of the FPMD values is observed in the case of MLIP^{A}(1). We recall that this latter, MLIP^{A}(1), turned out to be unstable when producing systems larger than 240 atoms.

4 | CONCLUSION

In this work, we present an automated workflow able to deliver accurate and stable MLIP by efficiently exploiting the FPMD data. Our strategy relies on the exploitation of a local atomic environment descriptor based on SOAP descriptor and a long-range descriptor based on the Coulomb matrix to build a hybrid similarity matrix able to compare amorphous snapshots. By efficiently tailoring the mixing between the local and the long-range parts of the hybrid similarity kernel, we show that one can achieve a good clustering of the amorphous configurations for both $AsTe_3$ and TeO_2 glassy systems. The clustering procedure is embedded within an AL loop that finds the best number of clusters based on a threshold distance cutoff. Subsequently, the achieved clusters are sampled to build a training set for MLIP potential fitting in an AL fashion. The MLIP hyperparameters are optimized through a BO cycle that ensures an efficient convergence toward an optimal setup that satisfies a user-defined energy convergence threshold with respect to the reference data. In this manner, we build minimum size databases, on top of which MLIP can be achieved with ab initio accuracy.

Our workflow is tested on various data sets of glassy $AsTe_3$ and TeO_2 systems, as well as pure TeO_2 glass mixed with TeO_2 gamma polymorph, where we show that the AL clustering procedure based on the output of the hybrid similarity kernel leads to an efficient clustering of the amorphous configurations of our systems. Consequently, less than 200 configurations are generally required to achieve an MLIP with ab initio accuracy. These MLIP were used to generate atomistic models with sizes as large as ≈30 000 atoms exhibiting excellent quantitative agreement with both FPMD reference data and

TABLE 1 Coordination numbers each chemical species α as obtained from the active learning (AL) fit procedure with optimal δ value (MLIP(0)), using configurations selected assuming δ=0 (MLIP(1)) and one trained using all available data sets (MLIP(2)). Results computed by integrating the pair distribution function. For glassy $AsTe_3$ bond radius cutoff between As–As, As–Te, and Te–Te is 3.15 Å, 3.03 Å, and 3.2 Å, respectively. While for TeO_2 and TeO_2 + γ-crystal glassy systems, the used bond length cutoff between Te-O is 2.46 Å. Values between parenthesis correspond to the first-principles molecular dynamics (FPMD) reference model.

		n_α			
Model	**System size [atoms]**	α	**MLIP(0)**	**MLIP(1)**	**MLIP(2)**
$AsTe_3$(data set A)	240	As	3.06 (3.04)	3.08	3.06
		Te	2.13 (2.08)	2.17	2.04
	1920	As	3.05	—	3.05
		Te	2.14	—	2.04
	6840	As	3.06	—	—
		Te	2.14	—	—
	30 000	As	3.12	—	—
		Te	2.14	—	—
TeO_2(data set B)	480	Te	3.70 (3.65)	3.98	3.72
		O	1.85 (1.83)	2.00	1.86
	3840	Te	3.71	4.04	3.67
		O	1.85	2.02	1.84
	12 960	Te	3.72	—	—
		O	1.86	—	—
	30 720	Te	3.70	—	—
		O	1.86	—	—
TeO_2 + γ-crystal(data set C)	480	Te	3.75 (3.65)	—	—
		O	1.87 (1.83)	—	—
	3840	Te	3.74	—	—
		O	1.87	—	—
	12 960	Te	3.74	—	—
		O	1.87	—	—
	30 720	Te	3.74	—	—
		O	1.87	—	—

experimental measurements, as shown by the consistency seen in the structure factors and the PDFs. Furthermore, our achieved MLIPs are compared to those achieved: (i) by fitting GAP on all the available data sets (≈1000 configurations). The obtained structures indicate excellent agreement to those obtained from the AL fit procedure, which demonstrates that the reduced training data set is sufficient for training the MLIP, (ii) by fitting GAP on a database achieved using configurations from clusters that were created based on only the local atomic environment similarity matrix (SOAP) as a global similarity metric. As a consequence of sampling over widely dispersed clusters with substantial overlap of the amorphous configurations over several temperature plateaus, the obtained structural models show strong discrepancy with the FPMD results and limited stability in the case of $AsTe_3$. Thereby showing that accounting for long-range correlations can be useful in building minimal data sets that capture the intricacies of the network connectivity in the case of amorphous systems. Overall, our work demonstrates the powerfulness of AL approaches in efficiently exploiting the FPMD data and producing accurate atomistic potentials using ML techniques. Our workflow can be readily applied to various classes of systems, in particular, disordered systems.[86]

ACKNOWLEDGMENTS

This work was supported by the French ANR via the AMSES project (ANR-20-CE08-0021) and by région Nouvelle Aquitaine via the CANaMIAS project AAPR2021-2020-11779110. Calculations were performed by using resources from Grand Equipement National de Calcul Intensif (GENCI, projects No. 0910832, 0913426, and 0914978). We used computational resources provided by the computing facilities Mésocentre de Calcul Intensif Aquitain (MCIA) of the Université de Bordeaux and of the Université de Pau et des Pays de l'Adour.

ORCID

Guido Ori https://orcid.org/0000-0003-3336-8277
Assil Bouzid https://orcid.org/0000-0002-9363-7240

REFERENCES

1. Behler J, Parrinello M. Generalized neural-network representation of high-dimensional potential-energy surfaces. Phys Rev Lett. 2007;98(14):146401.
2. Bartók AP, Payne MC, Kondor R, Csányi G. Gaussian approximation potentials: the accuracy of quantum mechanics, without the electrons. Phys Rev Lett. 2010;104(13):136403.
3. Thompson AP, Swiler LP, Trott CR, Foiles SM, Tucker GJ. Spectral neighbor analysis method for automated generation of quantum-accurate interatomic potentials. J Comput Phys. 2015;285:316–30.
4. Gilmer J, Schoenholz SS, Riley PF, Vinyals O, Dahl GE. Neural message passing for quantum chemistry. In: PMLR; 2017. p. 1263–72.
5. Schütt KT, Arbabzadah F, Chmiela S, Müller KR, Tkatchenko A. Quantum-chemical insights from deep tensor neural networks. Nat Commun. 2017;8(1):13890.
6. Chen C, Ye W, Zuo Y, Zheng C, Ong SP. Graph networks as a universal machine learning framework for molecules and crystals. Chem Mater. 2019;31(9):3564–72.
7. Chmiela S, Sauceda HE, Poltavsky I, Müller KR, Tkatchenko A. sGDML: constructing accurate and data efficient molecular force fields using machine learning. Comput Phys Commun. 2019;240:38–45.
8. Levchenko EV, Dappe YJ, Ori G. Theory and simulation in physics for materials applications: cutting-edge techniques in theoretical and computational materials science. Springer Cham; 2020. https://doi.org/10.1007/978-3-030-37790-8
9. Zuo Y, Chen C, Li X, Deng Z, Chen Y, Behler J, et al. Performance and cost assessment of machine learning interatomic potentials. J Phys Chem. 2020;124(4):731–45.
10. Massobrio C, Du J, Bernasconi M, Salmon PS. Molecular dynamics simulations of disordered materials. Vol. 215. Springer; 2015.
11. Massobrio C. The structure of amorphous materials using molecular dynamics. IOP Publishing; 2022.
12. Duong TQ, Bouzid A, Massobrio C, Ori G, Boero M, Martin E. First-principles thermal transport in amorphous $Ge_2Sb_2Te_5$ at the nanoscale. RSC Adv. 2021;11(18):10747–52.
13. Lorenz S, Groß A, Scheffler M. Representing high-dimensional potential-energy surfaces for reactions at surfaces by neural networks. Chem Phys Lett. 2004;395(4-6):210–15.
14. Artrith N, Urban A. An implementation of artificial neural-network potentials for atomistic materials simulations: performance for TiO_2. Comput Mater Sci. 2016;114:135–50.
15. Szlachta WJ, Bartók AP, Csányi G. Accuracy and transferability of Gaussian approximation potential models for tungsten. Phys Rev B. 2014;90(10):104108.
16. Deringer VL, Csányi G. Machine learning based interatomic potential for amorphous carbon. Phys Rev B. 2017;95(9):094203.
17. Bartók AP, Kermode J, Bernstein N, Csányi G. Machine learning a general-purpose interatomic potential for silicon. Phys Rev X. 2018;8(4):041048.
18. Sivaraman G, Krishnamoorthy AN, Baur M, Holm C, Stan M, Csányi G, et al. Machine-learned interatomic potentials by active learning: amorphous and liquid Hafnium dioxide. npj Comput Mater. 2020;6(1):104.
19. Mocanu FC, Konstantinou K, Lee TH, Bernstein N, Deringer VL, Csányi G, et al. Modeling the phase-change memory material, $Ge_2Sb_2Te_5$, with a machine-learned interatomic potential. J Phys Chem B. 2018;122(38):8998–9006.
20. Veit M, Jain SK, Bonakala S, Rudra I, Hohl D, Csányi G. Equation of state of fluid methane from first principles with machine learning potentials. J Chem Theory Comput. 2019;15(4):2574–86.
21. Deringer VL, Caro MA, Csányi G. A general-purpose machine-learning force field for bulk and nanostructured phosphorus. Nat Commun. 2020;11(1):5461.
22. Smith JS, Isayev O, Roitberg AE. ANI-1: an extensible neural network potential with DFT accuracy at force field computational cost. Chem Sci. 2017;8(4):3192–203.
23. Cheng B, Engel EA, Behler J, Dellago C, Ceriotti M. Ab initio thermodynamics of liquid and solid water. Proc Natl Acad Sci. 2019;116(4):1110–15.
24. De S, Bartók AP, Csányi G, Ceriotti M. Comparing molecules and solids across structural and alchemical space. Phys Chem Chem Phys. 2016;18(20):13754–69.
25. Glielmo A, Macocco I, Doimo D, Carli M, Zeni C, Wild R, et al. DADApy: distance-based analysis of data-manifolds in Python. Patterns. 2022;3(10):100589.
26. Settles B. Active learning literature survey. 2009.
27. Finkbeiner J, Tovey S, Holm C. Generating minimal training sets for machine learned potentials. Phys Rev Lett. 2024;132(16):167301.
28. Ben Mahmoud C, Gardner JL, Deringer VL. Data as the next challenge in atomistic machine learning. Nat Comput Sci. 2024:1–4.
29. Speckhard DT, Bechtel T, Ghiringhelli LM, Kuban M, Rigamonti S, Draxl C. How big is Big Data? arXiv preprint arXiv:240511404. 2024.
30. Kaur H, Pia FD, Batatia I, Advincula XR, Shi BX, Lan J, et al. Data-efficient fine-tuning of foundational models for first-principles quality sublimation enthalpies. arXiv preprint arXiv:240520217. 2024.
31. Sadeghi A, Ghasemi SA, Schaefer B, Mohr S, Lill MA, Goedecker S. Metrics for measuring distances in configuration spaces. J Chem Phys. 2013;139(18):184118.
32. Goryaeva AM, Maillet JB, Marinica MC. Towards better efficiency of interatomic linear machine learning potentials. Comput Mater Sci. 2019;166:200–209.
33. Zhu L, Amsler M, Fuhrer T, Schaefer B, Faraji S, Rostami S, et al. A fingerprint based metric for measuring similarities of crystalline structures. J Chem Phys. 2016;144(3):034203.
34. Bartók AP, Kondor R, Csányi G. On representing chemical environments. Phys Rev B. 2013;87(18):184115.
35. Prodan E, Kohn W. Nearsightedness of electronic matter. Proc Natl Acad Sci. 2005;102(33):11635–38.
36. Bernstein N, Bhattarai B, Csányi G, Drabold DA, Elliott SR, Deringer VL. Quantifying chemical structure and machine-learned atomic energies in amorphous and liquid silicon. Angew Chem. 2019;131(21):7131–35.
37. Pham TL, Guerboub M, Bouzid A, Boero M, Massobrio C, Shin YH, et al. Unveiling the structure and ion dynamics of amorphous $Na_{3-x}OH_xCl$ antiperovskite electrolytes

by first-principles molecular dynamics. J Mater Chem A. 2023;11(42):22922–40.
38. Kjellander R. Focus Article: Oscillatory and long-range monotonic exponential decays of electrostatic interactions in ionic liquids and other electrolytes: the significance of dielectric permittivity and renormalized charges. J Chem Phys. 2018;148(19):193701.
39. Guo Z, Ambrosio F, Chen W, Gono P, Pasquarello A. Alignment of redox levels at semiconductor–water interfaces. Chem Mater. 2018;30(1):94–111.
40. Jorn R, Kumar R, Abraham DP, Voth GA. Atomistic modeling of the electrode–electrolyte interface in Li-ion energy storage systems: electrolyte structuring. J Phys Chem C. 2013;117(8):3747–61.
41. French RH, Parsegian VA, Podgornik R, Rajter RF, Jagota A, Luo J, et al. Long range interactions in nanoscale science. Rev Mod Phys. 2010;82(2):1887.
42. Rupp M, Tkatchenko A, Müller KR, von Lilienfeld OA. Fast and accurate modeling of molecular atomization energies with machine learning. Phys Rev Lett. 2012;108(5):058301.
43. Huo H, Rupp M. Unified representation of molecules and crystals for machine learning. arXiv preprint arXiv:170406439. 2017.
44. Hirn M, Mallat S, Poilvert N. Wavelet scattering regression of quantum chemical energies. Multiscale Model Simul. 2017;15(2):827–63.
45. Grisafi A, Ceriotti M. Incorporating long-range physics in atomic-scale machine learning. J Chem Phys. 2019;151(20):204105.
46. Rosenbrock CW, Gubaev K, Shapeev AV, Pártay LB, Bernstein N, Csányi G, et al. Machine-learned interatomic potentials for alloys and alloy phase diagrams. npj Comput Mater. 2021;7(1):24.
47. Young TA, Johnston-Wood T, Deringer VL, Duarte F. A transferable active-learning strategy for reactive molecular force fields. Chem Sci. 2021;12(32):10944–55.
48. Podryabinkin EV, Tikhonov EV, Shapeev AV, Oganov AR. Accelerating crystal structure prediction by machine-learning interatomic potentials with active learning. Phys Rev B. 2019;99(6):064114.
49. Gubaev K, Podryabinkin EV, Hart GL, Shapeev AV. Accelerating high-throughput searches for new alloys with active learning of interatomic potentials. Comput Mater Sci. 2019;156:148–56.
50. Smith JS, Nebgen B, Lubbers N, Isayev O, Roitberg AE. Less is more: sampling chemical space with active learning. J Chem Phys. 2018;148(24):241733.
51. Vandermause J, Torrisi SB, Batzner S, Xie Y, Sun L, Kolpak AM, et al. On-the-fly active learning of interpretable Bayesian force fields for atomistic rare events. Npj Comput Mater. 2020;6(1):20.
52. Zeng J, Zhang D, Lu D, Mo P, Li Z, Chen Y, et al. DeePMD-kit v2: a software package for deep potential models. arXiv preprint arXiv:230409409. 2023.
53. Jinnouchi R, Lahnsteiner J, Karsai F, Kresse G, Bokdam M. Phase transitions of hybrid perovskites simulated by machine-learning force fields trained on the fly with Bayesian inference. Phys Rev Lett. 2019;122(22):225701.
54. Jinnouchi R, Miwa K, Karsai F, Kresse G, Asahi R. On-the-fly active learning of interatomic potentials for large-scale atomistic simulations. J Phys Chem Lett. 2020;11(17):6946–55.
55. Delaizir G, Piarristeguy A, Pradel A, Masson O, Bouzid A. Short range order and network connectivity in amorphous $AsTe_3$: a first principles, machine learning, and XRD study. Phys Chem Chem Phys. 2020;22(43):24895–906.
56. Raghvender R, Bouzid A, Cadars S, Hamani D, Thomas P, Masson O. Structure of amorphous TeO_2 revisited: a hybrid functional ab initio molecular dynamics study. Phys Rev B. 2022;106(17):174201.
57. Becke AD. Density-functional exchange-energy approximation with correct asymptotic behavior. Phys Rev A. 1988;38(6):3098.
58. Bouzid A, Massobrio C. Note: Accounting for pressure effects on the calculated equilibrium structure of glassy $GeSe_2$. J Chem Phys. 2012;137(4):046101.
59. Himanen L, Jäger MO, Morooka EV, Canova FF, Ranawat YS, Gao DZ, et al. DScribe: library of descriptors for machine learning in materials science. Comput Phys Commun. 2020;247:106949.
60. Ramakrishnan R, Hartmann M, Tapavicza E, von Lilienfeld OA. Electronic spectra from TDDFT and machine learning in chemical space. J Chem Phys. 2015;143(8):084111.
61. Rupp M. Machine learning for quantum mechanics in a nutshell. Int J Quantum Chem. 2015;115(16):1058–73.
62. Berg C, Christensen JPR, Ressel P. Harmonic analysis on semigroups: theory of positive definite and related functions. Association for Computing Machinery. Vol. 100. Springer; New York, NY, United States, 1984. https://doi.org/10.1145/1390156.1390183
63. Dasgupta S, Hsu D. Hierarchical sampling for active learning. In: Proceedings of the 25th International Conference on Machine Learning, UCSD; 2008. p. 208–15.
64. Dasgupta S. Two faces of active learning. Theor Comput Sci. 2011;412(19):1767–81.
65. von Luxburg U. A tutorial on spectral clustering. Stat Comput. 2007;17:395–416.
66. Bach F, Jordan M. Learning spectral clustering. In: Advances in Neural Information Processing Systems 16. 2003.
67. Jia H, Ding S, Xu X, Nie R. The latest research progress on spectral clustering. Neural Comput Appl. 2014;24:1477–186.
68. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. J Mach Learn Res. 2011;12:2825–30.
69. Nascimento MC, De Carvalho AC. Spectral methods for graph clustering—a survey. Eur J Oper. 2011;211(2):221–31.
70. Ding S, Jia H, Zhang L, Jin F. Research of semi-supervised spectral clustering algorithm based on pairwise constraints. Neural Comput Appl. 2014;24:211–19.
71. Cai XY, Dai GZ, Yang LB. Survey on spectral clustering algorithms. Comput Sci. 2008;35(7):14–18.
72. Borg I, Groenen PJ. Modern multidimensional scaling: theory and applications. Springer Science & Business Media; Springer New York, NY, 2005. https://doi.org/10.1007/0-387-28981-X
73. Kruskal JB. Nonmetric multidimensional scaling: a numerical method. Psychometrika. 1964;29(2):115–29.
74. Kruskal JB. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. Psychometrika. 1964;29(1):1–27.
75. Snoek J, Larochelle H, Adams RP. Practical Bayesian optimization of machine learning algorithms. In: Advances in Neural Information Processing Systems 25. 2012.

76. Shahriari B, Swersky K, Wang Z, Adams RP, De Freitas N. Taking the human out of the loop: A review of Bayesian optimization. Proc IEEE. 2016;104(1):148–75.
77. Plimpton S. Fast parallel algorithms for short-range molecular dynamics. J Comput Phys. 1995;117(1):1–19.
78. Nosé S. A unified formulation of the constant temperature molecular dynamics methods. J Chem Phys. 1984;81(1):511–19.
79. Hoover WG. Canonical dynamics: equilibrium phase-space distributions. Phys Rev A. 1985;31(3):1695.
80. Pham TL, Guerboub M, Wendj S, Bouzid A, Tugène C, Boero M, et al. Structural properties of amorphous Na_3OCl electrolyte by first-principles and machine learning molecular dynamics. arXiv preprint arXiv:240411442. 2024.
81. Bouzid A, Massobrio C, Boero M, Ori G, Sykina K, Eric F. Role of the van der Waals interactions and impact of the exchange-correlation functional in determining the structure of glassy $GeTe_4$. Phys Rev B. 2015;92(13):134208.
82. Silvestrelli PL, Martin E, Boero M, Bouzid A, Ori G, Massobrio C. Atomic structure of glassy $GeTe_4$ as a playground to assess the performances of density functional schemes accounting for dispersion forces. J Phys Chem B. 2020;124(49):11273–79.
83. Wright AC. The comparison of molecular dynamics simulations with diffraction experiments. J Non Cryst Solids. 1993;159(3):264–68.
84. Tverjanovich A, Rodionov K, Bychkov E. Raman spectroscopy of glasses in the As–Te system. J Solid State Chem. 2012;190:271–76.
85. Tenhover M, Boolchand P, Bresser W. Atomic structure and crystallization of As_xTe_{1-x} glasses. Phys Rev B. 1983;27(12):7533.
86. Klawohn S, Darby JP, Kermode JR, Csányi G, Caro MA, Bartók AP. Gaussian approximation potentials: theory, software implementation and application examples. J Chem Phys. 2023;159(17):174108.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Shuaib F, Ori G, Thomas P, Masson O, Bouzid A. Multikernel similarity-based clustering of amorphous systems and machine-learned interatomic potentials by active learning. J Am Ceram Soc. 2025;108:e20128. https://doi.org/10.1111/jace.20128