

<https://doi.org/10.1038/s41524-025-01705-3>

# Enhancing transferability of machine learning-based polarizability models in condensed-phase systems via atomic polarizability constraint



Mandi Fang<sup>1,2</sup>, Yinqiao Zhang<sup>1</sup>, Zheyong Fan<sup>3</sup>, Daquan Tan<sup>1</sup>, Xiaoyong Cao<sup>2</sup>, Chunlei Wei<sup>2</sup>, Nan Xu<sup>1,2</sup>✉ & Yi He<sup>1,2,4</sup>✉

Accurate prediction of molecular polarizability is essential for understanding electrical, optical, and dielectric properties of materials. Traditional quantum mechanical (QM) methods, though precise for small systems, are computationally prohibitive for large-scale systems. In this work, we proposed an efficient approach for calculating molecular polarizability of condensed-phase systems by embedding atomic polarizability constraints into the tensorial neuroevolution potential (TNEP) framework. Using *n*-heneicosane as a benchmark, a training data set was constructed from molecular clusters truncated from the bulk systems. Atomic polarizabilities derived from semi-empirical QM calculations were integrated as training constraints for its balance of computational efficiency and physical interpretability. The constrained TNEP model demonstrated improved accuracy in predicting molecular polarizabilities for larger clusters and condensed-phase systems, attributed to the model's refined ability to properly partition molecular polarizabilities into atomic contributions across systems with diverse configurational features. Results highlight the potential of the TNEP model with atomic polarizability constraint as a generalizable strategy to enhance the scalability and transferability of other atom-centered machine learning-based polarizability models, offering a promising solution for simulating large-scale systems with high data efficiency.

Molecular polarizability  $\alpha$ , the measure of an electron cloud's response to an external electric field, plays a fundamental role in determining a material's dielectric and optical properties<sup>1–5</sup>. The widespread use of quantum mechanical (QM) methods, such as the density functional theory (DFT) allows the accurate calculations of polarizabilities for small molecules and solids. However, when it comes to large-scale systems, such as proteins and polymers, the calculation of their polarizabilities remains a daunting task, mainly because the computational cost scales superlinearly with the system size<sup>6</sup>. Traditional empirical methods like the bond polarizability model<sup>7</sup> and atom-dipole interaction models<sup>8</sup> are facing rigorous challenges in accuracy and reliability<sup>9</sup>. Fragment-based methods, which partition molecular systems into smaller subsystems for analysis<sup>10–14</sup>, improve scalability but require significant expertise in defining partitions<sup>15</sup> and still face resource constraints for large-scale applications<sup>16</sup>.

The emergence of machine learning (ML)-based polarizability models can potentially tackle this challenge as they achieve a great balance between accuracy and efficiency<sup>17–24</sup>. The efficacy of ML-based polarizability models for molecules and crystalline solids has been demonstrated in previous studies<sup>17,25–28</sup>. However, predicting polarizabilities for proteins and polymers remains challenging due to the significant computational cost and effort required to generate accurate polarizability training data using high-precision DFT calculations. Therefore, reducing the cost of data set preparation for model training is crucial for the efficient modeling of large-scale systems.

Utilizing small cluster structures extracted from large-scale systems as training data for ML-based polarizability models may provide a viable strategy. Previous studies have demonstrated the feasibility of simulating bulk systems with atom-centered machine learning force fields (MLFFs)

<sup>1</sup>College of Chemical and Biological Engineering, Zhejiang University, Hangzhou, China. <sup>2</sup>Institute of Zhejiang University–Quzhou, Quzhou, China. <sup>3</sup>College of Physical Science and Technology, Bohai University, Jinzhou, China. <sup>4</sup>Department of Chemical Engineering, University of Washington, Seattle, WA, USA.

✉ e-mail: [tamas@zju.edu.cn](mailto:tamas@zju.edu.cn); [yihe@zju.edu.cn](mailto:yihe@zju.edu.cn)

trained on small fragment data as MLFFs only rely on atomic energies with a local environment dependence<sup>29–32</sup>. In principle, this should also be applicable to atom-centered ML-based polarizability models as they infer molecular polarizability from individual atomic contributions in the same way. However, allocation of atomic polarizabilities is not unique and rigorous, since current ML-based polarizability models are typically trained for predictions of molecular polarizabilities. It has been reported that if only the global quantity is rigorously defined during training, the decomposition of the global quantity into local contributions by ML models can take place in numerous different ways<sup>29,33,34</sup>. For the ML-based polarizability models, the distributed atomic polarizabilities allocated by themselves can be flexible and arbitrary, and in some instances, incapable of characterizing the polarization of atoms correctly. This can inevitably introduce uncertainty into the model predictions, thus affecting the transferability of ML-based polarizability models. To this end, further research is essential to develop robust methodologies that address uncertainties in atomic polarizability predictions, thereby enhancing the reliability and transferability of ML-based polarizability models from small clusters to large systems without target data.

The tensorial neuroevolution potential (TNEP) models for molecular polarizability, proposed and implemented in our previous work<sup>28</sup>, have shown high accuracy and extraordinary efficiency and were successfully applied to liquid water and perovskite BaZrO<sub>3</sub>. In this work, the transferability of TNEP models trained on cluster data to condensed-phase systems was investigated, after which the atomic polarizability constraint was manually introduced into the TNEP framework. First, an original TNEP model was trained on cluster data truncated from bulk systems of *n*-heneicosane with a maximum cutoff radius of 7 Å, which was determined by the convergence test based on the QM method. Subsequently, a constrained TNEP model (referred to as the TNEP-C model) was trained on the same training data set augmented with atomic polarizabilities derived from the semi-empirical QM method (referred to as GFN2-computed atomic polarizabilities). Test data sets, including cluster data of varying sizes, were constructed to evaluate the extrapolative performance of these two models. Comparisons of schemes for partitioning molecular polarizability into atomic contributions from these two TNEP models and the QM method using the Hirshfeld partitioning scheme were conducted to further elucidate the key improvement by incorporating the atomic polarizability constraint into the TNEP model. Finally, the performance of these two models on bulk systems was also investigated using committee error estimates (CEEs) as the indicator.

## Results

### Performance of the original TNEP model

The extrapolative performance of the original TNEP model was evaluated on a series of test data sets, including structures of varying sizes (truncated from bulk systems of *n*-heneicosane with cutoff radii ranging from 6 to 13 Å in 1 Å increments). The test data sets were labeled as “R6–R13”, where “R”

stands for the “cutoff radius” used in data sets constructions. Performance metrics, including root mean square error (RMSE) and the coefficient of determination ( $R^2$ ), were calculated to evaluate the model’s accuracy in predicting per-atom diagonal and off-diagonal elements of molecular polarizabilities for configurations in test data sets, using DFT reference values as the benchmark standard.

While the original TNEP model achieved high consistency with DFT reference values in predicting molecular polarizabilities for small data sets such as R6 and R7, the prediction errors increased significantly when extrapolating to much larger clusters. As shown in Fig. 1, a systematic increase in RMSEs of the per-atom diagonal elements of molecular polarizability ( $\alpha_{xx}^{\text{TNEP,mol}}$ ,  $\alpha_{yy}^{\text{TNEP,mol}}$ ,  $\alpha_{zz}^{\text{TNEP,mol}}$ ) was observed for configurations exceeding the size of those in the training data set, accompanied by a corresponding decrease in  $R^2$  values. In contrast, the predictions for per-atom off-diagonal elements of molecular polarizabilities ( $\alpha_{xy}^{\text{TNEP,mol}}$ ,  $\alpha_{yz}^{\text{TNEP,mol}}$ ,  $\alpha_{xz}^{\text{TNEP,mol}}$ ) remained relatively stable, with slight increases in RMSEs and minor decreases in  $R^2$ .

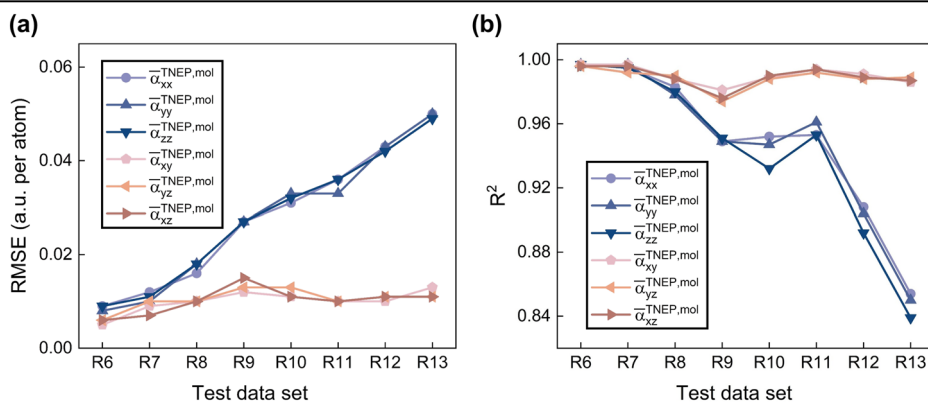
Parity plots of the diagonal elements of the predicted molecular polarizabilities versus the DFT reference values in Fig. 2 confirmed the large prediction errors by the original TNEP model for clusters much larger than those in the training data set. As the size of molecular clusters increased, the diagonal elements of molecular polarizabilities predicted by the TNEP model ( $\alpha_{\text{diag}}^{\text{TNEP,mol}}$ ) gradually deviated from the DFT reference values ( $\alpha_{\text{diag}}^{\text{ref,mol}}$ ), with significant overestimations for large clusters such as those in R12 and R13 test data sets. Conversely, the off-diagonal components predicted by the TNEP model were in good correlation with those calculated by the DFT method (Supplementary Fig. 1).

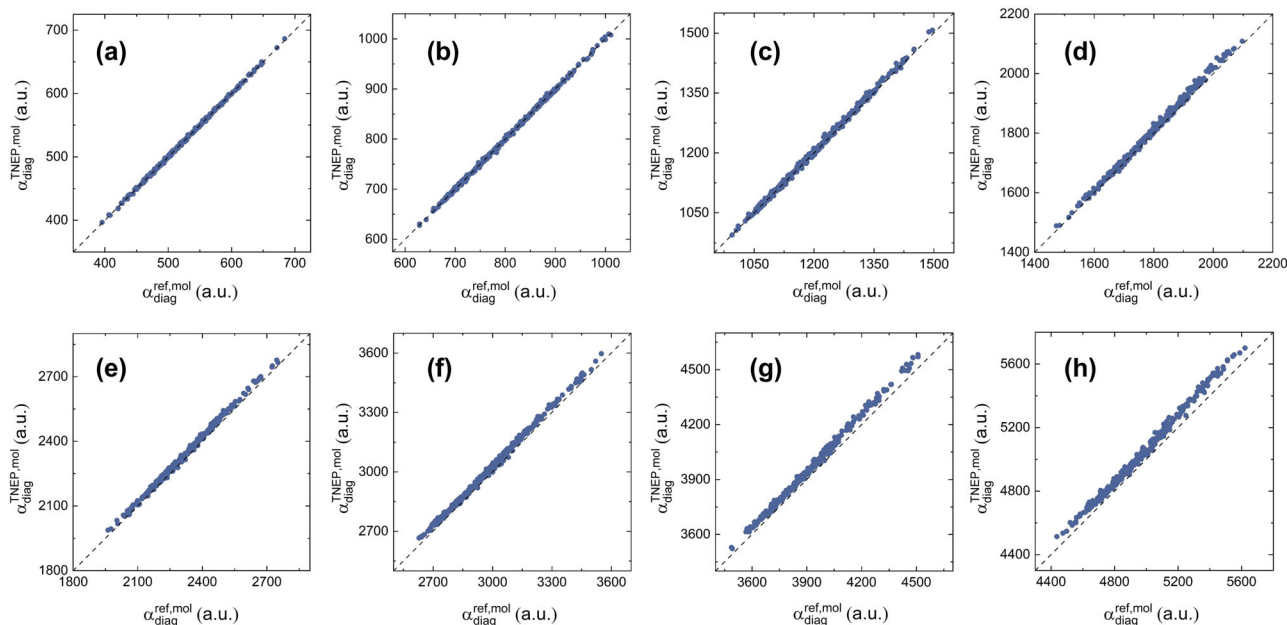
The opposite trends of prediction errors for diagonal and off-diagonal elements of molecular polarizability tensors may originate from their intrinsic characteristics. Diagonal elements of molecular polarizability tensors in ML-based polarizability models are typically divided into local atomic contributions modulated by the chemical environments, however, the schemes to assign atomic contributions by the TNEP model are closely related to the differences in the configurational features of the training and test data sets, and this will be discussed in the next section. On the contrary, for an isotropic system, the off-diagonal elements of polarizability tensors are mainly affected by molecular symmetry such as rotational operations<sup>35,36</sup>, weakening the influence of differences in configurational features between the training and test data sets.

### Analysis of key factors affecting the transferability of the original TNEP model

The significant discrepancy between predicted diagonal elements of molecular polarizabilities and reference values in Fig. 2 suggests that the original TNEP model tends to uniformly overestimate the polarizabilities of giant clusters unseen in the training data set. Due to the atom-centered structure of TNEP models, the discrepancy in molecular polarizabilities necessitates the inspection of the local atomic contributions to the total polarizability.

**Fig. 1 | Performance of the original TNEP model for predicting the per-atom molecular polarizabilities of configurations in test data sets.** Performance of the original TNEP model for predicting per-atom diagonal and off-diagonal elements of molecular polarizabilities of configurations in R6–R13 test data sets, evaluated by **a** RMSEs and **b**  $R^2$  values.

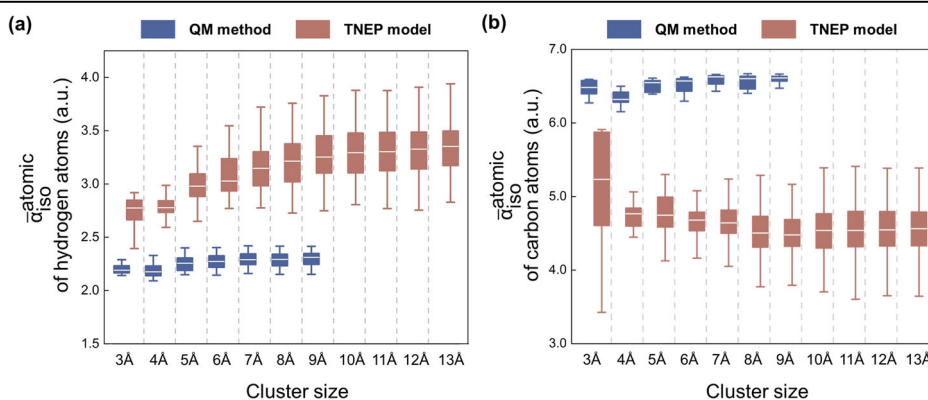




**Fig. 2 | Parity plots of the diagonal elements of the molecular polarizabilities predicted by the original TNEP model versus the DFT reference values for configurations in test data sets.** Parity plots of the diagonal elements of the molecular polarizabilities predicted by the original TNEP model versus the DFT

reference values for configurations in **a** R6 test data set, **b** R7 test data set, **c** R8 test data set, **d** R9 test data set, **e** R10 test data set, **f** R11 test data set, **g** R12 test data set, and **h** R13 test data set.

**Fig. 3 | Comparisons of atomic polarizability distributions calculated by the QM method using the Hirshfeld partitioning scheme and the original TNEP model for clusters of varying sizes centered on a certain carbon atom.** Comparisons of distributed atomic polarizabilities of **a** hydrogen and **b** carbon atoms in clusters of varying sizes centered on a certain carbon atom calculated by the QM method using the Hirshfeld partitioning scheme and the original TNEP model.



Atomic polarizability in TNEP models is calculated from individual artificial neural networks (ANNs) by using local chemical environments consisting of all atoms inside a cutoff sphere of radius as inputs. Therefore, the differences in atomic environments between the training and test data sets (input to the ANNs) and the distributions of atomic polarizabilities (output to the ANNs) were both analyzed to investigate the main factors that may affect the validity of the TNEP model trained on small molecules when applied to larger systems.

For the inspection of the inputs of ANNs, the similarity in atomic environments across different data sets was first compared by using the descriptor space analysis for carbon atoms and hydrogen atoms in the training and test data sets. Supplementary Figs. 2 and 3 suggested that the projections of the training data set almost entirely covered those of R6–R13 test data sets. This revealed that transferring from the training data set to test data sets should not introduce significant changes to the diversity of local atomic environments.

For the inspection of the outputs of ANNs, atomic polarizability distributions for hydrogen and carbon atoms in clusters of varying sizes centered on a certain carbon atom were calculated by the original TNEP model and the QM method using the Hirshfeld partitioning scheme,

respectively. Note that only averaged isotropic polarizability ( $\bar{\alpha}_{iso}^{atomic} = (\alpha_{xx}^{atomic} + \alpha_{yy}^{atomic} + \alpha_{zz}^{atomic})/3$ ) was considered here, and QM calculations were implemented on clusters with cutoff radii ranging from 3 to 9 Å due to the computational costs. The results from the QM method using the Hirshfeld partitioning scheme (referred to as QM-based atomic polarizabilities) were scaled to ensure consistency with the reference data for the TNEP model at the computational level. Fig. 3 shows that the original TNEP model tends to allocate polarizabilities to hydrogen and carbon atoms in a manner that differs markedly from the QM-based atomic polarizabilities. Specifically, hydrogen atoms were assigned excessively higher values, while carbon atoms were assigned substantially lower values. As the size of the clusters increased, the polarizabilities assigned to carbon and hydrogen atoms by the original TNEP model gradually approached a common value, indicating a diminishing capability of the model to differentiate between different elements. In addition, the instability of the predicted atomic polarizabilities was generally on the rise (the anomaly observed in clusters with a cutoff radius of 3 Å may be due to the limited number of atoms in the cluster). This indicated that for atoms of a given type, new atomic environments may have emerged while the model failed in representing them, and the proportion of atoms in such environments increased accordingly as the clusters expanded.

This limitation arose because the original TNEP model only accounted for the loss function of total polarizability during the training process, which led to internal flexibility of atomic polarizability contributions. The scheme for decomposing total polarizability into atomic contributions is sensitive to the configurational features of training data sets, including chemical compositions and proportions of atoms in various chemical environments. Table 1 shows that the stoichiometric ratios and the fractions of carbon atoms in bulk-like environments rose accordingly as the size of the configurations in data sets increased. First, the differences in the stoichiometric ratios between the training and the test data sets hinder the model's ability to distinguish between chemical elements in such transfer, similar to findings demonstrated in previous work on MLFF<sup>29</sup>. Second, the limited fraction of carbon atoms in bulk-like environments within the training data set contributes to the original TNEP model being undertrained in representing bulk-like atomic environments. To this end, when applied to much larger structures, the TNEP model insists on making predictions with physically unreasonable schemes, leading to an overall overestimation of molecular polarizabilities.

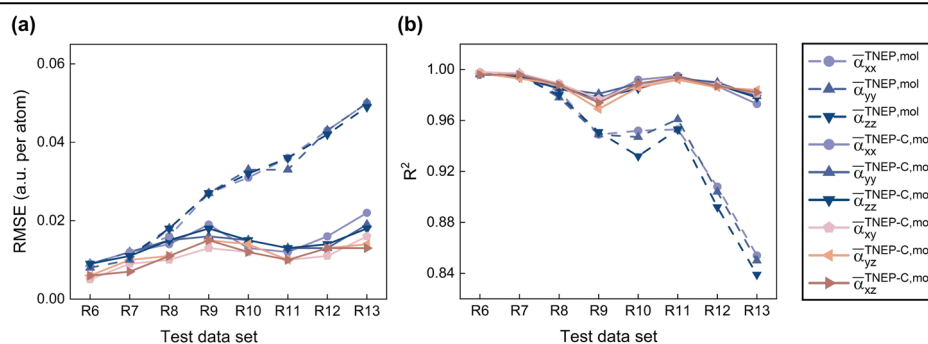
### Performance of the TNEP-C model

Since the original TNEP model failed to give a reliable decomposition of molecular polarizabilities among the constituent atoms, GFN2-computed atomic polarizabilities were introduced into the training process of the TNEP model as constraints. Before integrating, a systematic error correction of atomic polarizabilities was necessary according to the previous work<sup>37</sup>, and a quadratic fit with the zero intercept provided the best match between the GFN2-computed and DFT-computed polarizabilities of the labeled data (Supplementary Fig. 4).

**Table 1 | Results of configurational features including stoichiometric ratios and fractions of carbon atoms in bulk-like environments inside the cutoff radius of the symmetry functions for the training and test data sets, and the periodic system of *n*-heneicosane**

Data sets	Stoichiometric ratios of configurations		Fractions of carbon atoms in bulk-like environments for configurations (%)		
	Min	Max	Min	Max	Average
R6 test data set	0.330	0.444	0	0	0
R7 test data set	0.359	0.451	0	0.391	0.004
R8 test data set	0.370	0.450	0	1.61	0.085
R9 test data set	0.390	0.454	1.23	2.62	1.94
R10 test data set	0.401	0.460	2.03	4.51	3.01
R11 test data set	0.409	0.464	3.42	7.93	5.72
R12 test data set	0.418	0.466	7.75	12.2	10.4
R13 test data set	0.425	0.466	10.6	14.4	12.3
Training data set	0.273	0.450	0	0.439	0.002
Periodic system	0.477		100		

**Fig. 4 | Performance of the TNEP-C model for predicting the per-atom molecular polarizabilities of configurations in test data sets.** Performance of the TNEP-C model for predicting per-atom diagonal and off-diagonal elements of molecular polarizabilities of configurations in R6–R13 test data sets, evaluated by **a** RMSEs and **b**  $R^2$  values. Dashed lines represent the results from the original TNEP model and the solid lines represent the results from the TNEP-C model.



The extrapolative performance of the TNEP-C model demonstrated advantages over the original TNEP model. While maintaining the performance for training data set (Supplementary Fig. 5) and R6–R7 test data sets, the RMSEs of the per-atom diagonal elements of molecular polarizabilities predicted by the TNEP-C model decreased substantially for large data sets from R8 to R13 (Fig. 4 and Supplementary Table 1). The parity plots of the diagonal and off-diagonal elements of the predicted molecular polarizabilities versus the DFT reference values shown in Supplementary Figs. 6 and 7 also confirmed that by learning GFN2-computed atomic polarizabilities, the TNEP-C model exhibited improved accuracy on large clusters.

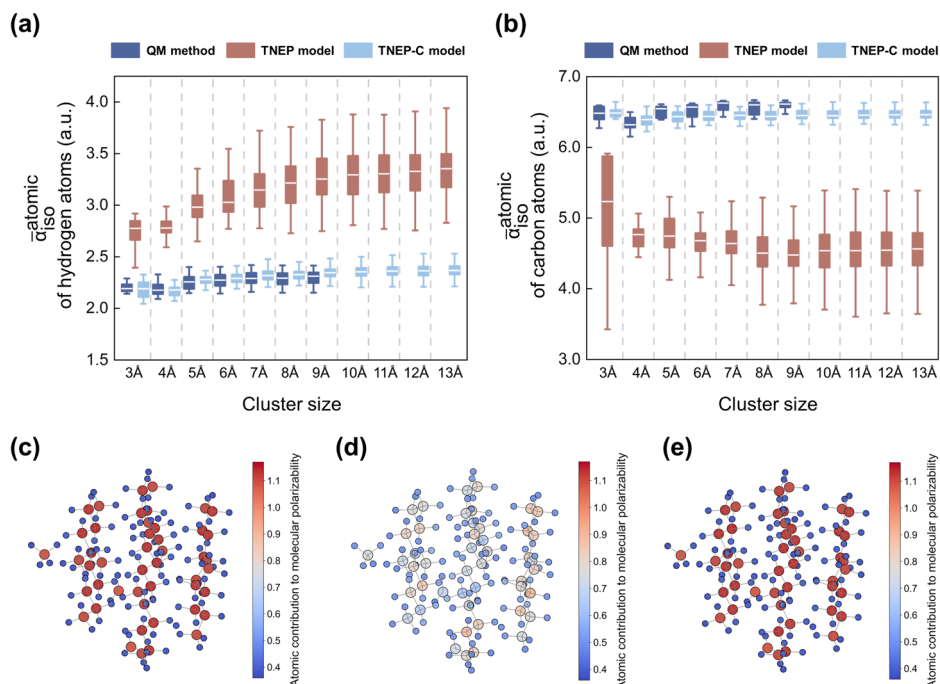
Atomic polarizability distributions for hydrogen and carbon atoms in clusters of varying sizes centered on a certain carbon atom were also calculated to investigate the improvement of the TNEP-C model. As shown in Fig. 5a, b, the results from the TNEP-C model were in good agreement with QM-based atomic polarizabilities and remained robust when extrapolating. Take a cluster with a cutoff radius of 6 Å as an example, distributed atomic polarizabilities were plotted as polarizability ellipsoids and atoms were colored based on their contributions to molecular polarizability. The values of the polarizability tensors for hydrogen atoms should be substantially smaller compared with those of carbon atoms, due to their smaller electronic population (Fig. 5c). However, the original TNEP model itself tended to allocate comparable values to carbon and hydrogen atoms (Fig. 5d). In contrast, the TNEP-C model can correctly differentiate between carbon and hydrogen atoms (Fig. 5e). This indicated that the TNEP-C model has embedded a physics-compliant partitioning strategy for partitioning total polarizabilities into atomic contributions. Consequently, the constrained model delivered enhanced robustness in extrapolation compared with the unconstrained counterpart, as evidenced by reduced errors when predicting diagonal elements of molecular polarizabilities for larger clusters beyond the training domain.

### Extrapolating the original TNEP and TNEP-C models to bulk systems

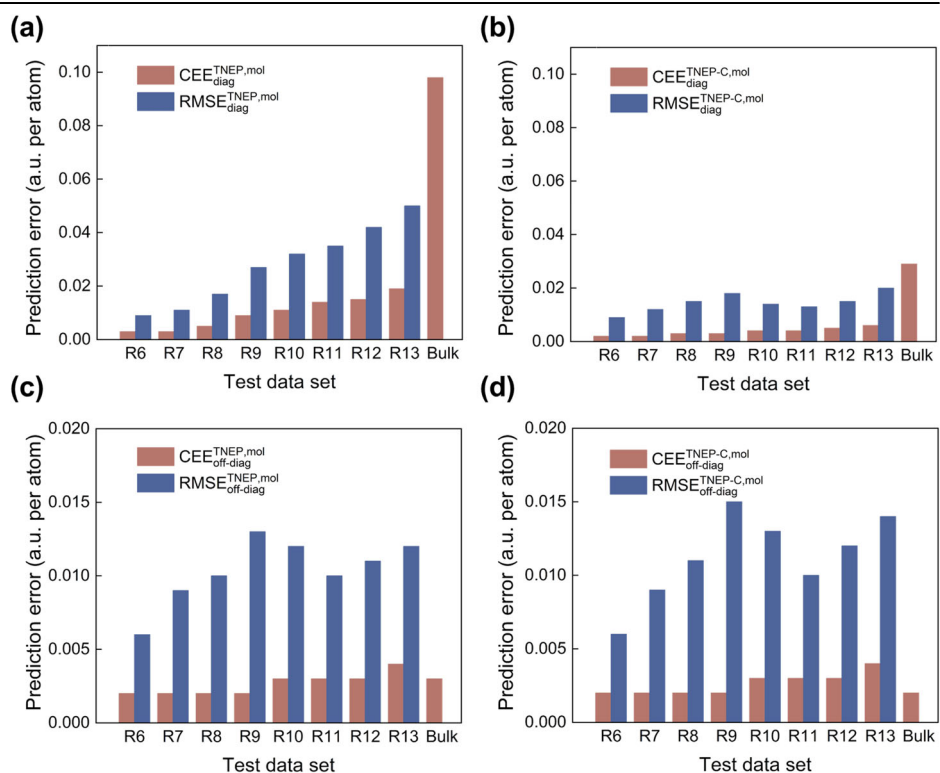
In addition to the evaluation on large molecular clusters, we also assessed the transferability of the original TNEP and TNEP-C models on bulk systems. The bulk data set contains 125,000 configurations sampled from the MD trajectories of *n*-heneicosane with an interval of 2 fs. Since calculations of polarizabilities at the DFT level for such a large system are unattainable, alternative uncertainty estimation metrics are required to evaluate the prediction errors instead of RMSE. Here, we employed the CEE algorithm<sup>38</sup>, which can provide a metric to quantify the generalization error in the form of the committee disagreement. CEE has been previously used to evaluate the performance of ML models<sup>39–42</sup>, including its application to TNEP models<sup>43</sup>. In this work, five instances of both the original TNEP and TNEP-C models were independently trained. The corresponding CEEs were computed on the test data set and compared with RMSEs for validation. The results indicated that while CEE tended to underestimate RMSE, the overall consistency between these two metrics suggested that CEE can serve as a reasonable approximation for evaluating prediction errors on systems that pose challenges to the DFT methods, and this correlation has also been reported in previous studies<sup>43</sup>.



**Fig. 5 | Comparisons of atomic polarizability distributions calculated by the QM method using the Hirshfeld partitioning scheme, the original TNEP model and the TNEP-C model for clusters of varying sizes centered on a certain carbon atom.** Comparisons of distributed atomic polarizabilities of **a** hydrogen and **b** carbon atoms in clusters of varying sizes centered on a certain carbon atom calculated by the QM method using the Hirshfeld partitioning scheme, the original TNEP model and the TNEP-C model; graphical representation of distributed atomic polarizabilities calculated by **c** the QM method using the Hirshfeld partitioning scheme, **d** the original TNEP model and **e** the TNEP-C model for a cluster with a cutoff radius of 6 Å.



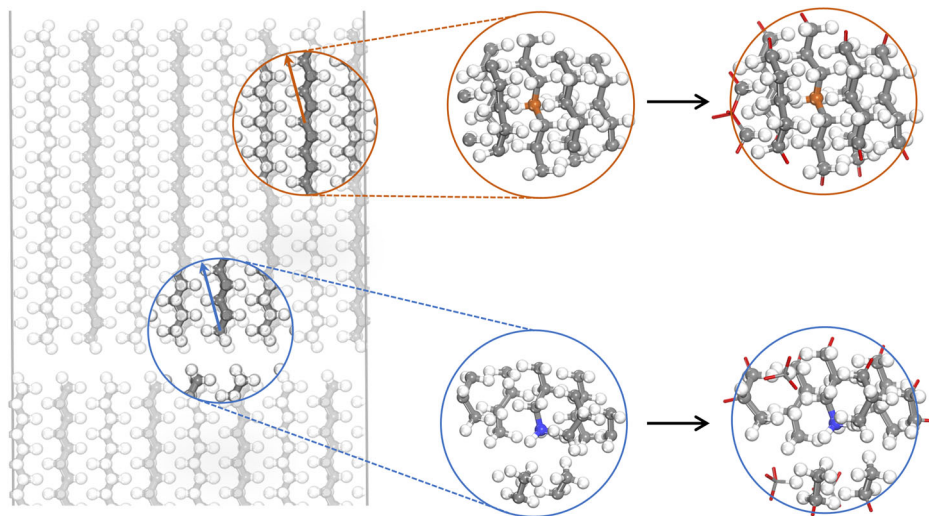
**Fig. 6 | Comparisons of prediction errors of per-atom molecular polarizabilities for the test data sets and the bulk data set of *n*-heneicosane obtained by the original TNEP model and the TNEP-C model.** Comparisons of prediction errors for per-atom (**a**, **b**) diagonal and (**c**, **d**) off-diagonal elements of molecular polarizabilities for the test data sets and the bulk data set of *n*-heneicosane, obtained by the original TNEP model and the TNEP-C model, respectively.



Significant enhancements were observed in the predictions for the per-atom diagonal elements of molecular polarizabilities. Figure 6a presents that the CEE can reach as high as 0.1 a.u. per atom when extrapolating to the bulk systems for the original TNEP model, while for the TNEP-C model, this error is reduced to 0.03 a.u. per atom (Fig. 6b). This result further demonstrates that, if only the molecular polarizabilities are fitted, TNEP models can achieve high accuracy on the training data set with several internal schemes to partition total polarizabilities into atomic contributions. While the atom-specific polarizabilities assigned by the original TNEP model itself can vary greatly with changes in configurational features of training data sets, and sometimes these atomic

predictions are physically inconsistent<sup>44</sup>. This uncertainty can lead to poor transferability across different test data sets, and can be substantially reduced via the implementation of constraints on atomic polarizabilities. In contrast, the original TNEP and TNEP-C models exhibited comparable accuracy in predicting the off-diagonal elements of molecular polarizabilities (Fig. 6c, d), even when transferring to the bulk systems. The reason for this lies in the fact that the isotropic GFN2-computed atomic polarizabilities incorporated into the TNEP-C model do not impose constraints on the off-diagonal components. Consequently, the predictions for the off-diagonal components remain primarily governed by the original TNEP model's inherent capability, which is less affected by the

**Fig. 7 | Schematic representation of molecular clusters derived from the bulk structure of *n*-heneicosane.** The central carbon atoms of the clusters are highlighted in blue and orange, respectively, and the red sticks represent the atoms artificially introduced to saturate the free valences.



variations in the system sizes and configurational features compared with the diagonal elements.

Rooms for further enhancement remain in this approach. For instance, while the GFN2-computed atomic polarizabilities are generally reasonable and readily obtainable, the current model may be limited when applied to systems with intense anisotropic effects due to the absence of contributions from off-diagonal components. Introducing atomic polarizabilities obtained from the partitioning of ground-state and field-perturbed electron densities of a molecular system such as quantum theory of atoms in molecules (QTAIM)<sup>45–47</sup> as training constraints in TNEP models may potentially yield better results, but it also poses a challenge in terms of the computational costs.

## Discussion

We demonstrated that incorporating atomic polarizability constraints into the TNEP model can significantly enhance its transferability, enabling accurate predictions of polarizabilities for condensed-phase systems based only on small molecular cluster data. By integrating atomic polarizabilities derived from semi-empirical QM calculations, the TNEP-C model learned a physically grounded partitioning scheme to divide atomic contributions, especially for configurations with increased stoichiometric ratios and proportions of atoms in bulk-like environments. Consequently, the TNEP-C model showed boosted performance in extrapolation, with largely reduced errors in predicting polarizabilities of large clusters and bulk systems.

In principle, this approach can be extended to organic systems with higher complexity at the chemical composition level, such as systems that include elements like oxygen and nitrogen. For systems involving metallic atoms, more sophisticated methods for assigning atomic polarizabilities (like QTAIM) will be essential and necessitate further testing and validation to ensure accuracy and reliability. Importantly, this methodology could also be applied to other atom-centered ML-based polarizability models, thereby providing a robust strategy for scalable, data-efficient predictions of molecular polarizability in complex condensed-phase materials. This would potentially pave the way for more insightful simulations of molecular properties, particularly in understanding electronic and spectroscopic characteristics of various materials.

## Methods

The training data set was composed of molecular clusters truncated from the configurations extracted from the MD trajectories of *n*-heneicosane (C<sub>21</sub>H<sub>44</sub>) with varying cutoff radii from 3 to 7 Å in 1 Å increments. To improve the efficiency of the data set construction, we started with the smallest clusters truncated with a cutoff radius of 3 Å, and continuously supplemented the clusters with larger cutoff radii through the farthest point

sampling (FPS) method iteratively. This section was organized as follows: First, the method to construct the clusters from the bulk system was introduced, followed by the computational details of calculating molecular polarizabilities for the initial training dataset. Subsequently, the principle of the original TNEP model was briefly outlined. The explorations of larger clusters through the FPS method were detailed afterwards, and this section finally ended with the calculations of atomic polarizabilities and refactoring of TNEP models for atomic polarizabilities.

### Construction of molecular clusters

An orthogonal structure of *n*-heneicosane with a bilayer of 6 × 8 unit cells (6240 molecules) was constructed as the initial configuration for MD simulations<sup>48–50</sup>. The temperature was held at 301 K using the Nosé–Hoover thermostat<sup>51,52</sup> with a time constant of 0.1 ps, and the pressure was maintained at 1 bar using the Parinello–Rahman barostat<sup>53</sup>. The MD simulation was performed using the COMPASS force field<sup>54</sup> in LAMMPS<sup>55</sup>. Snapshots were dumped every 25 ps from an MD production run of 250 ps, and a total of 11 snapshots were obtained.

Molecular clusters of different sizes were truncated from the bulk structures of *n*-heneicosane by extracting atoms within cutoff radii from 3 to 13 Å in 1 Å increments surrounding each carbon atom. Carbon atoms outside a certain radius will be kept if the valency was situated on a hydrogen or bonded to two carbon atoms within the radius, as illustrated in Fig. 7. Subsequently, free valencies were saturated with hydrogen atoms<sup>56</sup>. Constrained optimizations were performed for every cluster using the COMPASS force field to adjust the positions of hydrogen atoms while keeping the carbon skeleton fixed.

### Calculations of molecular polarizability

Molecular polarizabilities were calculated for each structure in the training data sets using the DFT method. Molecular polarizabilities were calculated by solving the coupled perturbed self-consistent field equations using the GTH-PBE pseudopotential and the DZVP-MOLOPT-SR-GTH basis set (400 Ry cutoff,  $\Gamma$  point)<sup>57–59</sup>. All DFT calculations were carried out using the Gaussian Plane Waves method (GPW) in CP2K<sup>60,61</sup>.

### Principle of the original TNEP model

The TNEP model for predicting tensorial properties is developed based on the NEP framework, which is implemented in the GPUMD package<sup>62</sup>. In the NEP-based potential energy surface (PES) model, the total energy of one system is given by the sum of atomic site energies  $U_i$ , which are computed using individual ANNs and depend on the local atomic chemical environments. Following the work of Behler and Parrinello<sup>63</sup>, the input layer

consists of descriptor vectors of high dimensions constructed from Chebyshev and Legendre's polynomials<sup>64,65</sup>. Explicit expressions of the descriptor vector and more detailed information on the NEP-based PES model are introduced in refs. 66–69.

The molecular polarizability tensor is a second-order symmetric tensor with nine components for a given structure with  $N$  atoms<sup>28</sup>. Components of  $\alpha$  can be expressed as:

$$\alpha_{\mu\nu} = \sum_i^N U_i \delta_{\mu\nu} - \sum_i^N \sum_{j \neq i}^N r_{ij}^\mu \frac{\partial U_i}{\partial r_{ij}^\nu} \quad (1)$$

where  $\nu$  refers to the direction of the applied external electric field (e.g.,  $x, y, z$  in Cartesian coordinates), while  $\mu$  denotes the direction of the induced dipole moment (e.g.,  $x, y, z$  in Cartesian coordinates). The polarizability tensor component  $\alpha_{\mu\nu}$  quantifies the linear response between the external electric field applied in the  $\nu$ -direction and the induced dipole moment in the  $\mu$ -direction. When  $\mu = \nu$ ,  $\alpha_{\mu\nu}$  corresponds to the diagonal element of the molecular polarizability tensor, while  $\mu \neq \nu$  represents the off-diagonal component.  $\delta_{\mu\nu}$  is the Kronecker delta.  $r_{ij}^\mu$  is the  $\mu$ -component of the vector  $\mathbf{r}_{ij} \equiv \mathbf{r}_j - \mathbf{r}_i$ , and  $\mathbf{r}_j$  is the position of neighboring atom  $j$  around atom  $i$ .  $U_i$  here has the dimension of polarizability.

The loss function of the original TNEP model is given by the weighted sum of the RMSEs of the molecular polarizability  $\mathcal{L}_{\text{mol}}(\mathbf{z})$  as well as two regularization terms, as:

$$\mathcal{L}(\mathbf{z}) = \mathcal{L}_{\text{mol}}(\mathbf{z}) + \lambda_1 \frac{1}{N_{\text{par}}} \sum_{n=1}^{N_{\text{par}}} |z_n| + \lambda_2 \sqrt{\frac{1}{N_{\text{par}}} \sum_{n=1}^{N_{\text{par}}} z_n^2} \quad (2)$$

where  $\mathbf{z}$  is a set of trainable parameters from the descriptors and the ANN model, and  $N_{\text{par}}$  is the total number of tunable parameters. The last two terms represent  $\mathcal{L}_1$  and  $\mathcal{L}_2$  regularizations. The weights  $\lambda_1$  and  $\lambda_2$  are tunable hyperparameters.

The loss term accounting for molecular polarizability is defined as:

$$\mathcal{L}_{\text{mol}}(\mathbf{z}) = \left\{ \frac{1}{6N_{\text{str}}} \sum_{n=1}^{N_{\text{str}}} \left[ \sum_{\mu=\nu} \left( \alpha_{\mu\nu}^{\text{TNEP,mol}}(n, \mathbf{z}) - \alpha_{\mu\nu}^{\text{ref,mol}}(n) \right)^2 + \lambda_s^{\text{mol}} \left( \sum_{\mu>\nu} \left( \alpha_{\mu\nu}^{\text{TNEP,mol}}(n, \mathbf{z}) - \alpha_{\mu\nu}^{\text{ref,mol}}(n) \right)^2 \right) \right] \right\}^{\frac{1}{2}} \quad (3)$$

where  $N_{\text{str}}$  is the number of structures in the whole training data set.  $\alpha_{\mu\nu}^{\text{TNEP,mol}}(n, \mathbf{z})$  is the molecular polarizability component predicted by the original TNEP model with parameters  $\mathbf{z}$  for the  $n^{\text{th}}$  structure while  $\alpha_{\mu\nu}^{\text{ref,mol}}(n)$  is the corresponding reference molecular polarizability component typically obtained by the DFT method. Since molecular polarizability is a symmetric second-order tensor ( $\alpha_{\mu\nu} = \alpha_{\nu\mu}$ ), we utilize the lower-triangular off-diagonal components ( $\mu > \nu$ ) of  $\alpha_{\mu\nu}^{\text{TNEP,mol}}(n, \mathbf{z})$  and  $\alpha_{\mu\nu}^{\text{ref,mol}}(n)$  for implementation.  $\lambda_s^{\text{mol}}$  is introduced as a regularization parameter to balance the contributions from the diagonal and off-diagonal components.

For the radial components of the original TNEP model, a cutoff radius of 7 Å and seven radial functions (each being a linear combination of 11 basis functions) were used in this work. For the angular components, a cutoff radius of 4 Å and seven radial functions (each being a linear combination of 11 basis functions) were used. The maximum expansion order for the three, four, and five-body terms of angular descriptor components is 4, 2, and 1, respectively. The fitting component is an ANN composed of one hidden layer with 30 neurons. For the regularization parameters,  $\lambda_s^{\text{mol}}$  was set to 1,  $\lambda_1$  and  $\lambda_2$  both were set to 0.03. The original TNEP model was trained for 300,000 generations using the SNES algorithm with a population size of 80.

## Iterative explorations of larger clusters through the FPS method

To improve computational efficiency, for the smallest cutoff radius (3 Å), only one structure for each stoichiometric ratio of hydrocarbon was randomly selected and labeled as the training data set to train the pre-TNEP model. On this basis, clusters left were labeled as the unselected data, and the completeness of the current training data set in relation to the unselected one was evaluated by descriptor space analysis with the pre-TNEP model<sup>70</sup>. New samples were added via the FPS method if needed to build the initial data set.

New samples were further added using the FPS method based on the initial data set, its corresponding TNEP model, and all cluster structures constructed with a cutoff radius of 4 Å represented as the unselected data set. This iteration continued until the cluster structures constructed with a cutoff radius of 7 Å were supplemented. The schematic representation of supplementing the training data set by the FPS method is demonstrated in Fig. 8. The final training data set contains 1980 configurations, with 380 configurations constructed with a cutoff radius of 3 Å, and every 400 configurations constructed with cutoff radii ranging from 4 to 7 Å in 1 Å increments. The reason for choosing a converged cutoff radius of 7 Å was discussed in Supplementary Information.

Every 100 structures were randomly selected from the unselected data sets with a certain cutoff radius from 6 to 13 Å in 1 Å increments, and were labeled as R6–R13 test data sets. The extrapolative performance of the TNEP models was evaluated on the test data sets.

## Calculations of atomic polarizability

Atomic polarizabilities of structures in the training data set were calculated by the GFN2-xTB method<sup>71</sup>. GFN2-computed atomic polarizabilities were selected as training constraints because they were derived based on atom types including the element number, hybridization state of carbon atoms, and some basic structural information, and were potentially physically more motivated to be transferred. GFN2-computed atomic polarizabilities depend on pre-computed atomic polarizabilities at a certain molecular geometry, i.e., with the atom having a GFN2-xTB computed atomic partial charge  $q_r$  and a covalent coordination number  $CN_{\text{cov}}^r$  (the index  $r$  indicates values for the reference structures)<sup>71–73</sup>.

In addition, a more accurate approach based on the QM determination and Hirshfeld partitioning scheme was employed. This approach served as an independent benchmark for evaluating the ability to partition isotropic molecular polarizability into atomic contributions of TNEP models for comparison. This method relies on the observation that atomic polarizability is proportional to the fuzzy atomic volume of the electron cloud<sup>74–77</sup>:

$$\alpha_{\text{eff}}(0) \equiv \alpha_{\text{free}}(0) \frac{V_{\text{eff}}}{V_{\text{free}}} \quad (4)$$

where  $\alpha_{\text{eff}}(0)$  and  $\alpha_{\text{free}}(0)$  are static polarizability for the atom-in-a-molecule (effective atomic polarizability) and the free-atom,  $V_{\text{eff}}$  and  $V_{\text{free}}$  are measures of the “volume” of the atom in a molecule and the free atom, respectively.

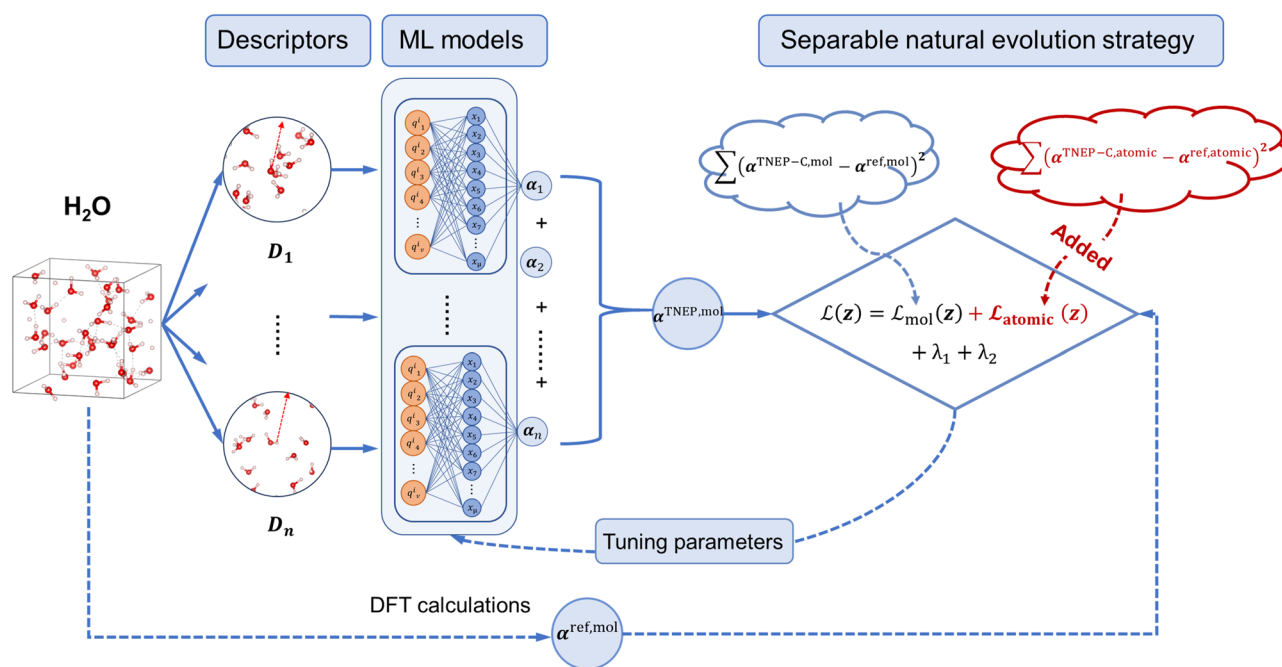
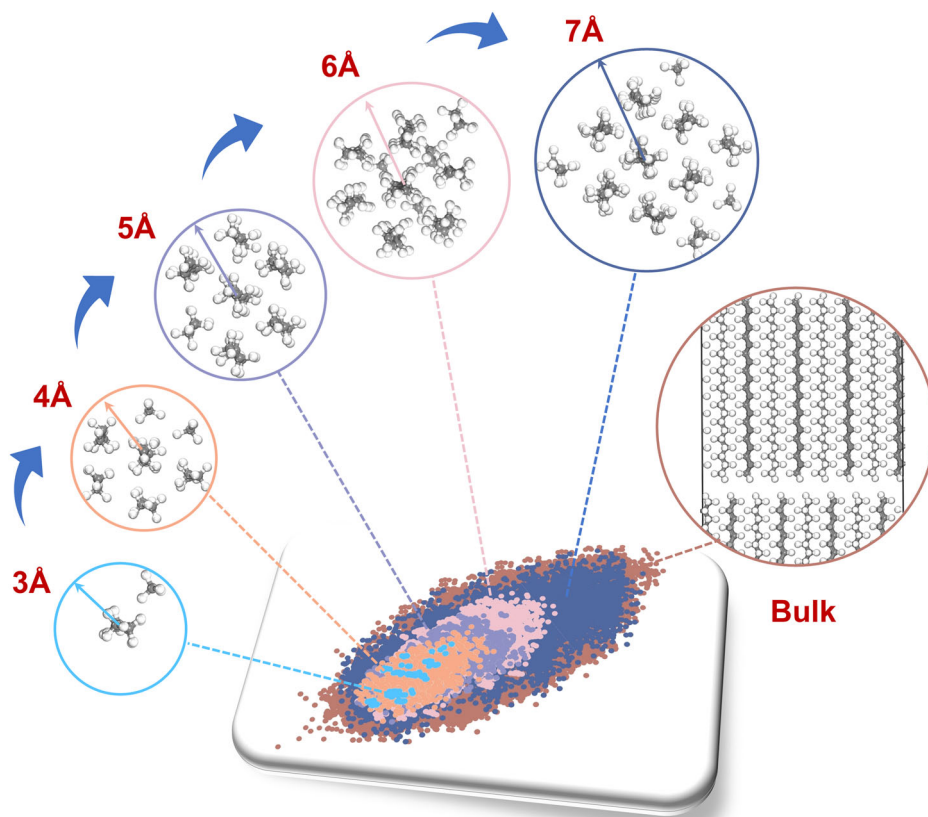
By employing the Hirshfeld partitioning scheme based on the electron density calculated from DFT calculations<sup>76,78,79</sup>, the ratio of the atom-in-a-molecule volume to the free-atom volume of each atom can be derived, and the QM-based atomic polarizability can be deduced by Eq. 4 subsequently.

## Principle of the TNEP-C model

The architecture of the TNEP-C model is shown in Fig. 9. To pose a constraint on atomic polarizability, an additional term is included in the loss function of the TNEP-C model as:

$$\mathcal{L}_{\text{atomic}}(\mathbf{z}) = \left\{ \frac{1}{6N_{\text{str}}} \sum_{n=1}^{N_{\text{str}}} \sum_{i=1}^{N_a} \left[ \sum_{\mu=\nu} \left( \alpha_{\mu\nu}^{\text{TNEP-C,atomic}}(n, i, \mathbf{z}) - \alpha_{\mu\nu}^{\text{ref,atomic}}(n, i) \right)^2 + \lambda_s^{\text{atomic}} \left( \sum_{\mu>\nu} \left( \alpha_{\mu\nu}^{\text{TNEP-C,atomic}}(n, i, \mathbf{z}) - \alpha_{\mu\nu}^{\text{ref,atomic}}(n, i) \right)^2 \right) \right] \right\}^{\frac{1}{2}} \quad (5)$$

**Fig. 8 | Schematic representation of supplementing the training data set by the FPS method.** Points in different colors in the projection represent configurations in different data sets.



**Fig. 9 | Schematic representation of the TNEP-C architecture.** The section highlighted in red represents the introduced loss function term enforcing the atomic polarizability constraint.

where  $N_{\text{str}}$  is the number of structures in the whole training data set and  $N_a$  is the number of atoms in the  $n^{\text{th}}$  structure.  $\alpha_{\mu\nu}^{\text{TNEP-C,atomic}}(n, i, \mathbf{z})$  is the atomic polarizability component predicted by the TNEP-C model with parameters  $\mathbf{z}$  for the  $i^{\text{th}}$  atom in the  $n^{\text{th}}$  structure.  $\alpha_{\mu\nu}^{\text{ref,atomic}}(n, i)$  is the corresponding reference atomic polarizability component, which is obtained by the GFN2-

xTB method in this work. Since the GFN2-computed atomic polarizabilities are inherently isotropic, the contributions from the off-diagonal components are zero in the training process. Consequently,  $\lambda_s^{\text{atomic}}$ , which is designed to balance the contributions from diagonal and off-diagonal elements, is set to a default value with no tuning required.



The total loss function for the TNEP-C model is thus defined as:

$$\mathcal{L}(\mathbf{z}) = \mathcal{L}_{\text{mol}}(\mathbf{z}) + \lambda_{\text{atomic}} \cdot \mathcal{L}_{\text{atomic}}(\mathbf{z}) + \lambda_1 \frac{1}{N_{\text{par}}} \sum_{n=1}^{N_{\text{par}}} |z_n| + \lambda_2 \sqrt{\frac{1}{N_{\text{par}}} \sum_{n=1}^{N_{\text{par}}} z_n^2} \quad (6)$$

where  $\lambda_{\text{atomic}}$  is the weight of the atomic polarizability term to balance the contributions from the molecular polarizability and atomic polarizability.

For the TNEP-C model,  $\lambda_{\text{atomic}}$  was set to 0.2 and other settings were kept identical to those implemented in the original TNEP model. The TNEP-C model was trained for 400,000 generations to ensure that the loss terms for molecular polarizability in the training and test data sets had largely converged.

## Data availability

The data supporting the findings of this study are available at [https://github.com/Daisy315/citable-data-yigroup/tree/main/npj\\_2024](https://github.com/Daisy315/citable-data-yigroup/tree/main/npj_2024).

Received: 4 June 2025; Accepted: 13 June 2025;

Published online: 05 July 2025

## References

- Chen, F. et al. Polarizability matters in enantio-selection. *Nat. Commun.* **15**, 3394 (2024).
- Qin, J., Liu, Z., Ma, M. & Li, Y. Optimizing and extending ion dielectric polarizability database for microwave frequencies using machine learning methods. *npj Comput. Mater.* **9**, 132 (2023).
- Wang, J., Xie, X. Q., Hou, T. & Xu, X. Fast approaches for molecular polarizability calculations. *J. Phys. Chem. A* **111**, 4443–4448 (2007).
- Berger, E. & Komsa, H.-P. Polarizability models for simulations of finite temperature Raman spectra from machine learning molecular dynamics. *Phys. Rev. Mater.* **8**, 043802 (2024).
- Thomas, M., Brehm, M., Fligg, R., Vöhringer, P. & Kirchner, B. Computing vibrational spectra from ab initio molecular dynamics. *Phys. Chem. Chem. Phys.* **15**, 6608 (2013).
- Nakata, H., Fedorov, D. G., Yokojima, S., Kitaura, K. & Nakamura, S. Simulations of Raman spectra using the fragment molecular orbital method. *J. Chem. Theory Comput.* **10**, 3689–3698 (2014).
- Guha, S., Menéndez, J., Page, J. B. & Adams, G. B. Empirical bond polarizability model for fullerenes. *Phys. Rev. B* **53**, 13106–13114 (1996).
- Applequist, J. An atom dipole interaction model for molecular optical properties. *Acc. Chem. Res.* **10**, 79–85 (1977).
- Bougeard, D. & Smirnov, K. S. Calculation of off-resonance Raman scattering intensities with parametric models. *J. Raman Spectrosc.* **40**, 1704–1719 (2009).
- Li, W., Dong, H., Ma, J. & Li, S. Structures and spectroscopic properties of large molecules and condensed-phase systems predicted by generalized energy-based fragmentation approach. *Acc. Chem. Res.* **54**, 169–181 (2021).
- Zhang, L., Cheng, Z., Li, W. & Li, S. Generalized energy-based fragmentation approach for accurate binding energies and Raman spectra of methane hydrate clusters. *Chin. J. Chem. Phys.* **35**, 167–176 (2022).
- Zhao, D. et al. Accurate and efficient prediction of post-Hartree-Fock polarizabilities of condensed-phase systems. *J. Chem. Theory Comput.* **19**, 6461–6470 (2023).
- Zhao, D. et al. Fragment-based deep learning for simultaneous prediction of polarizabilities and NMR shieldings of macromolecules and their aggregates. *J. Chem. Theory Comput.* **20**, 2655–2665 (2024).
- Wang, T. et al. Ab initio characterization of protein molecular dynamics with AI2BMD. *Nature* **635**, 1019–1027 (2024).
- Raghavachari, K. & Saha, A. Accurate composite and fragment-based quantum chemical models for large molecules. *Chem. Rev.* **115**, 5643–5677 (2015).
- Liu, J. & He, X. Fragment-based quantum mechanical approach to biomolecules, molecular clusters, molecular crystals and liquids. *Phys. Chem. Chem. Phys.* **22**, 12341–12367 (2020).
- Wilkins, D. M. et al. Accurate molecular polarizabilities with coupled cluster theory and machine learning. *Proc. Natl. Acad. Sci. USA* **116**, 3401–3406 (2019).
- Grisafi, A., Wilkins, D. M., Csányi, G. & Ceriotti, M. Symmetry-adapted machine learning for tensorial properties of atomistic systems. *Phys. Rev. Lett.* **120**, 036002 (2018).
- Zhang, Y. et al. Efficient and accurate simulations of vibrational and electronic spectra with symmetry-preserving neural network models for tensorial properties. *J. Phys. Chem. B* **124**, 7284–7290 (2020).
- Zhang, L. et al. Deep neural network for the dielectric response of insulators. *Phys. Rev. B* **102**, 041121 (2020).
- Sommers, G. M., Andrade, M. F. C., Zhang, L., Wang, H. & Car, R. Raman spectrum and polarizability of liquid water from deep neural networks. *Phys. Chem. Chem. Phys.* **22**, 10592–10602 (2020).
- Gastegger, M., Schütt, K. T. & Müller, K.-R. Machine learning of solvent effects on molecular spectra and reactions. *Chem. Sci.* **12**, 11473–11483 (2021).
- Schütt, K., Unke, O. & Gastegger, M. In *Equivariant Message Passing for the Prediction of Tensorial Properties and Molecular Spectra* 9377–9388 (PMLR, 2021).
- Zhang, Y., Jiang, J. & Jiang, B. Learning dipole moments and polarizabilities. In *Quantum Chemistry in the Age of Machine Learning* 453–465 (Elsevier, 2023).
- Raimbault, N., Grisafi, A., Ceriotti, M. & Rossi, M. Using Gaussian process regression to simulate the vibrational Raman spectra of molecular crystals. *N. J. Phys.* **21**, 105001 (2019).
- Zhang, Y. & Jiang, B. Universal machine learning for the response of atomistic systems to external fields. *Nat. Commun.* **14**, 6424 (2023).
- Kapil, V., Kovács, D. P., Csányi, G. & Michaelides, A. First-principles spectroscopy of aqueous interfaces using machine-learned electronic and quantum nuclear effects. *Faraday Discuss* **249**, 50–68 (2024).
- Xu, N. et al. Tensorial properties via the neuroevolution potential framework: fast simulation of infrared and Raman spectra. *J. Chem. Theory Comput.* **20**, 3273 (2024).
- Eckhoff, M. & Behler, J. From molecular fragments to the bulk: development of a neural network potential for MOF-5. *J. Chem. Theory Comput.* **15**, 3793–3809 (2019).
- Zaverkin, V., Holzmüller, D., Schuldt, R. & Kästner, J. Predicting properties of periodic systems from cluster data: a case study of liquid water. *J. Chem. Phys.* **156**, 114103 (2022).
- Schran, C., Briec, F. & Marx, D. Transferability of machine learning potentials: protonated water neural network potential applied to the protonated water hexamer. *J. Chem. Phys.* **154**, 051101 (2021).
- Herbold, M. & Behler, J. Machine learning transferable atomic forces for large systems from underconverged molecular fragments. *Phys. Chem. Chem. Phys.* **25**, 12979–12989 (2023).
- Chong, S. et al. Robustness of local predictions in atomistic machine learning models. *J. Chem. Theory Comput.* **19**, 8020–8031 (2023).
- El-Machachi, Z., Wilson, M. & Deringer, V. L. Exploring the configurational space of amorphous graphene with machine-learned atomic energies. *Chem. Sci.* **13**, 13720–13731 (2022).
- Elola, M. D. & Ladanyi, B. M. Molecular dynamics study of polarizability anisotropy relaxation in aromatic liquids and its connection with local structure. *J. Phys. Chem. B* **110**, 15525–15541 (2006).
- Tuschel, D. Raman crystallography and the effect of Raman polarizability tensor element values. *Spectroscopy* **35**, 5–12 (2020).

37. Hiener, D. C., Folmsbee, D. L., Langkamp, L. A. & Hutchison, G. R. Evaluating fast methods for static polarizabilities on extended conjugated oligomers. *Phys. Chem. Chem. Phys.* **24**, 23173–23181 (2022).
38. Schran, C., Brezina, K. & Marsalek, O. Committee neural network potentials control generalization errors and enable active learning. *J. Chem. Phys.* **153**, 104105 (2020).
39. Erlebach, A., Nachtigall, P. & Grajciar, L. Accurate large-scale simulations of siliceous zeolites by neural network potentials. *npj Comput. Mater.* **8**, 174 (2022).
40. Margraf, J. T. Science-driven atomistic machine learning. *Angew. Chem. Int. Ed.* **62**, e202219170 (2023).
41. Carrete, J., Montes-Campos, H., Wanzenböck, R., Heid, E. & Madsen, G. K. H. Deep ensembles vs committees for uncertainty estimation in neural-network force fields: comparison and application to active learning. *J. Chem. Phys.* **158**, 204801 (2023).
42. Zhang, H., Juraskova, V. & Duarte, F. Modelling chemical processes in explicit solvents with machine learning potentials. *Nat. Commun.* **15**, 6114 (2024).
43. Berger, E., Niemelä, J., Lampela, O., Juffer, A. H. & Komsa, H.-P. Raman spectra of amino acids and peptides from machine learning polarizabilities. *J. Chem. Inf. Model.* **64**, 4601–4612 (2024).
44. Feng, C., Xi, J., Zhang, Y., Jiang, B. & Zhou, Y. Accurate and interpretable dipole interaction model-based machine learning for molecular polarizability. *J. Chem. Theory Comput.* **19**, 1207–1217 (2023).
45. Dawes, R., Dwyer, J. R., Qu, W. & Gough, K. M. QTAIM investigation of the electronic structure and large Raman scattering intensity of bicyclo-[1.1.1]-pentane. *J. Phys. Chem. A* **115**, 13149–13157 (2011).
46. Ángyán, J. G., Jansen, G., Loos, M., Hättig, C. & Hess, B. A. Distributed polarizabilities using the topological theory of atoms in molecules. In *AIP Conference Proceedings* Vol. 330 67–67 (AIP, 1995).
47. Macchi, P. & Krawczuk, A. The polarizability of organometallic bonds. *Comput. Theor. Chem.* **1053**, 165–172 (2015).
48. Doherty, D. C. & Hopfinger, A. J. Molecular modeling of polymers: molecular dynamics simulation of the rotator phase of C<sub>21</sub>H<sub>44</sub>. *Phys. Rev. Lett.* **72**, 661–664 (1994).
49. Phillips, T. L. & Hanna, S. A comparison of computer models for the simulation of crystalline polyethylene and the long n-alkanes. *Polymer* **46**, 11003–11018 (2005).
50. Ryckaert, J.-P., McDonald, I. R. & Klein, M. L. Disorder in the pseudohexagonal rotator phase of n-alkanes: molecular-dynamics calculations for tricosane. *Mol. Phys.* **67**, 957–979 (1989).
51. Nosé, S. [Ubar] I. A molecular dynamics method for simulations in the canonical ensemble. *Mol. Phys.* **100**, 191–198 (2002).
52. Nosé, S. A molecular dynamics method for simulations in the canonical ensemble. *Mol. Phys.* **52**, 255–268 (1984).
53. Parrinello, M. & Rahman, A. Polymorphic transitions in single crystals: a new molecular dynamics method. *J. Appl. Phys.* **52**, 7182–7190 (1981).
54. Sun, H., Ren, P. & Fried, J. The COMPASS force field: parameterization and validation for phosphazenes. *Comput. Theor. Polym. Sci.* **8**, 229–246 (1998).
55. Thompson, A. P. et al. LAMMPS—a flexible simulation tool for particle-based materials modeling at the atomic, meso, and continuum scales. *Comput. Phys. Commun.* **271**, 108171 (2022).
56. Gastegger, M., Behler, J. & Marquetand, P. Machine learning molecular dynamics for the simulation of infrared spectra. *Chem. Sci.* **8**, 6924–6935 (2017).
57. Goedecker, S., Teter, M. & Hutter, J. Separable dual-space Gaussian pseudopotentials. *Phys. Rev. B* **54**, 1703 (1996).
58. Hartwigsen, C., Goedecker, S. & Hutter, J. Relativistic separable dual-space Gaussian pseudopotentials from H to Rn. *Phys. Rev. B* **58**, 3641 (1998).
59. Perdew, J. P., Burke, K. & Ernzerhof, M. Generalized gradient approximation made simple. *Phys. Rev. Lett.* **77**, 3865–3868 (1996).
60. Lippert, G., Hutter, J. & Parrinello, M. A hybrid Gaussian and plane wave density functional scheme. *Mol. Phys.* **92**, 477–488 (1997).
61. Kühne, T. D. et al. CP2K: an electronic structure and molecular dynamics software package - Quickstep: efficient and accurate electronic structure calculations. *J. Chem. Phys.* **152**, 194103 (2020).
62. Fan, Z., Chen, W., Vierimaa, V. & Harju, A. Efficient molecular dynamics simulations with many-body potentials on graphics processing units. *Comput. Phys. Commun.* **218**, 10–16 (2017).
63. Behler, J. & Parrinello, M. Generalized neural-network representation of high-dimensional potential-energy surfaces. *Phys. Rev. Lett.* **98**, 146401 (2007).
64. Musil, F. et al. Physics-inspired structural representations for molecules and materials. *Chem. Rev.* **121**, 9759–9815 (2021).
65. Langer, M. F., Goeßmann, A. & Rupp, M. Representations of molecules and materials for interpolation of quantum-mechanical simulations via machine learning. *npj Comput. Mater.* **8**, 41 (2022).
66. Fan, Z. et al. Neuroevolution machine learning potentials: combining high accuracy and low cost in atomistic simulations and application to heat transport. *Phys. Rev. B* **104**, 104309 (2021).
67. Fan, Z. Improving the accuracy of the neuroevolution machine learning potential for multi-component systems. *J. Phys.: Condens. Matter* **34**, 125902 (2022).
68. Fan, Z. et al. GPU-MD: A package for constructing accurate machine-learned potentials and performing highly efficient atomistic simulations. *J. Chem. Phys.* **157**, 114801 (2022).
69. Song, K. et al. General-purpose machine-learned potential for 16 elemental metals and their alloys. *Nat. Commun.* **15**, 10208 (2024).
70. Fang, M. et al. Transferability of machine learning models for predicting Raman spectra. *J. Phys. Chem. A* **128**, 2286–2294 (2024).
71. Bannwarth, C., Ehlert, S. & Grimme, S. GFN2-xTB—an accurate and broadly parametrized self-consistent tight-binding quantum chemical method with multipole electrostatics and density-dependent dispersion contributions. *J. Chem. Theory Comput.* **15**, 1652–1671 (2019).
72. Caldeweyher, E. et al. A generally applicable atomic-charge dependent London dispersion correction. *J. Chem. Phys.* **150**, 154122 (2019).
73. Caldeweyher, E., Bannwarth, C. & Grimme, S. Extension of the D3 dispersion coefficient model. *J. Chem. Phys.* **147**, 034112 (2017).
74. Tkatchenko, A. & Scheffler, M. Accurate molecular van der Waals interactions from ground-state electron density and free-atom reference data. *Phys. Rev. Lett.* **102**, 073005 (2009).
75. Blair, S. A. & Thakkar, A. J. Relating polarizability to volume, ionization energy, electronegativity, hardness, moments of momentum, and other molecular properties. *J. Chem. Phys.* **141**, 074306 (2014).
76. Hermann, J., DiStasio, R. A. & Tkatchenko, A. First-principles models for van der Waals interactions in molecules and materials: concepts, theory, and applications. *Chem. Rev.* **117**, 4714–4758 (2017).
77. Heid, E., Szabadi, A. & Schröder, C. Quantum mechanical determination of atomic polarizabilities of ionic liquids. *Phys. Chem. Chem. Phys.* **20**, 10992–10996 (2018).
78. Lu, T. & Chen, F.-W. Comparison of computational methods for atomic charges. *Acta Phys. Chim. Sin.* **28**, 1–18 (2012).
79. Lu, T. & Chen, F. Multiwfn: a multifunctional wavefunction analyzer. *J. Comput. Chem.* **33**, 580–592 (2012).

## Acknowledgements

This work is supported by “Pioneer” and “Leading Goose” R&D Program of Zhejiang (grant number: 2023C01182), and the National Natural Science Foundation of China (grant numbers: 22408314, 22178299, and 51933009). Nan Xu would like to thank the financial support provided by the Startup Funds of the Institute of Zhejiang University-Quzhou.

## Author contributions

Mandi Fang contributed to investigation, methodology, data analysis, writing—original draft, and writing—review and editing. Yinqiao Zhang contributed to software development. Zheyong Fan contributed to software development and writing—review and editing. Daquan Tan contributed to formal analysis. Xiaoyong Cao and Chunlei Wei contributed to data curation and validation. Nan Xu contributed to supervision, methodology, software development, funding acquisition, and writing—review and editing. Yi He contributed to supervision, funding acquisition, and writing—review and editing. All authors reviewed the manuscript.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s41524-025-01705-3>.

**Correspondence** and requests for materials should be addressed to Nan Xu or Yi He.

**Reprints and permissions information** is available at <http://www.nature.com/reprints>

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025