A predictive machine learning force-field framework for liquid electrolyte development

Sheng Gong [1,3], Yumin Zhang[1,3], Zhenliang Mu [2,3], Zhichen Pu[2], Hongyi Wang[2], Xu Han[2], Zhiao Yu[1], Mengyi Chen[2], Tianze Zheng [2], Zhi Wang [1], Lifei Chen[2], Zhenze Yang[1], Xiaojie Wu[1], Shaochen Shi [2], Weihao Gao[2] ✉, Wen Yan [1] ✉ & Liang Xiang[2]

Despite the widespread applications of machine learning force fields (MLFFs) in solids and small molecules, there is a notable gap in applying MLFFs to simulate liquid electrolytes—a critical component of current commercial lithium-ion batteries. Here we introduce ByteDance Artificial intelligence Molecular simulation Booster (BAMBOO), a predictive framework for molecular dynamics simulations, with a demonstration of its capability in the context of liquid electrolytes for lithium batteries. We design a physics-inspired graph equivariant transformer architecture as the backbone of BAMBOO to learn from quantum mechanical simulations. Additionally, we introduce an ensemble knowledge distillation approach and apply it to MLFFs to reduce the fluctuation of observations from molecular dynamics simulations. Finally, we propose a density alignment algorithm to align BAMBOO with experimental measurements. BAMBOO demonstrates state-of-the-art accuracy in predicting key electrolyte properties such as density, viscosity and ionic conductivity across various solvents and salt combinations. The current model, trained on more than 15 chemical species, achieves an average density error of 0.01 g cm^{-3} on various compositions compared with experiment.

Liquid electrolyte is an indispensable component in most of the current commercial lithium-ion batteries[1–3]. The existing commercial electrolytes are mostly carbonate-based, and often composed of more than five components to meet different requirements. Experimentally exploring molecular interactions for the rational design of electrolytes is expensive, time consuming and heavily reliant on the chemists' intuition and experience. Such limitations pose challenges in transitioning from proof-of-concept experiments in laboratories to products in markets, particularly due to the exponentially increasing complexity involved in optimizing the properties of multicomponent liquid electrolytes.

Atomistic simulation is an accessible, efficient and flexible alternative to exhaust experimentation. However, to achieve a reliable simulation, requirements such as high accuracy and sufficient time-scale and lengthscale are necessary. Quantum mechanical simulation offers high accuracy in describing electronic properties, yet they are computationally intensive and impractical for studying large-scale and complex systems such as liquid electrolytes. By contrast, classical force fields, although computationally efficient, often sacrifice accuracy in capturing the intricate solvation structures and dynamic behaviour of electrolytes. Hence, there exists a pressing need for a balanced and general approach that can reconcile the trade-off between accuracy

[1]ByteDance Research, Bellevue, WA, USA. [2]ByteDance Research, Beijing, China. [3]These authors contributed equally: Sheng Gong, Yumin Zhang, Zhenliang Mu. ✉e-mail: liverlover001@gmail.com; wen.yan@bytedance.com

and speed in modelling liquid electrolytes with different molecular solvents and varying salt types and concentrations.

In recent years, there has been a growing utilization of machine learning force fields (MLFFs)[4] to perform molecular dynamics (MD) simulations[5]. This trend is primarily attributed to the ability of MLFFs to deliver results faster than ab initio quantum mechanical simulations, as well as fit quantum mechanical data with higher accuracy compared with classical force fields. The development of MLFFs has seen two prominent trends. On one hand, there has been a gradual integration of concepts from the field of machine learning into designing the force field. This evolution has seen a shift from local-descriptor-based rotation-invariant MLFFs[6–8] towards graph neural network (GNN)-based models[9] with rotation-equivariant representations[10] and the transformer architecture[11,12]. On the other hand, there has been an emphasis on incorporating interactions grounded in clear physical foundation into MLFFs. These include electrostatic[13,14], dispersion[15,16] and spin–spin interactions[17,18].

With advancements in the model architecture of MLFFs, the concept of 'universal MLFF', which aims to use a single MLFF to simulate a wide range of systems, has garnered increasing attention within the realms of solid-state materials[19–23] and bio-organic molecules[15,24]. However, when it comes to liquids, particularly liquid electrolytes containing solvents and ions, a universal MLFF that can accurately predict multiple properties across various solvents and salts is still lacking. This limitation may arise from the complex local structures inherent in liquid electrolytes, such as the coexistence of different structural motifs like solvent-separated ion pairs (SSIP), contacted ion pairs (CIP) and aggregates (AGG). As a result, despite some studies utilizing MLFFs to investigate aqueous systems[25,26], molecular liquids[27] and ionic liquids[28], there remains a notable scarcity of research specifically focused on MLFFs for liquid electrolytes.

To the best of our knowledge, only two notable previous attempts have been made to simulate liquid electrolytes using MLFFs: ref. [29] utilized deep potential[7] to calculate the density and solvation structure of LiFSI in triglyme, and ref. [30] used a charge-recursive neural network[31] to compute the density, viscosity and diffusivity of LiPF$_6$ in carbonate solvents. Despite achieving some success, these studies lack conclusive evidence regarding the generalizability of their findings across a wide range of liquid electrolytes. For instance, ref. [30] shows that a charge-recursive neural network struggled to achieve high accuracy across both linear and cyclic carbonates simultaneously, as well as solutions with low and high concentrations of LiPF$_6$.

In addition to the scarcity of studies on MLFFs for liquid electrolytes, there are two overarching challenges associated with MLFFs. As highlighted in another work[32], MD simulations using MLFFs often encounter issues of collapsed simulations and large fluctuation of results, which stem from the inherent randomness in machine learning[33] and limit the practical usability of MLFFs. Moreover, most deep learning-based MLFFs solely rely on learning from quantum mechanical simulations, which do not necessarily guarantee the accurate reproduction of experimental measurements across diverse atomistic systems. Although the concept of differentiable molecular simulation[34,35] has been introduced to optimize classical force fields based on experimentally measured macroscopic observables, currently only a few deep learning-based MLFFs are available whose parameters are directly optimized using experimental results. One underlying reason for this absence could be that current optimization methods based on differentiable molecular simulations rely on backpropagating the gradients of MD trajectories to train the parameters of force fields. This process is computationally expensive and has the possibility to induce overfitting of deep neural network-based MLFFs to the limited amount of experimental data.

In this work, we introduce the ByteDance Artificial intelligence Molecular simulation Booster (BAMBOO) workflow, specifically designed for constructing MLFFs for the MD simulations of organic liquids, with a particular emphasis on liquid electrolytes. The main methodological contributions of this paper are summarized as follows. (1) We propose an MLFF model that integrates a graph equivariant transformer (GET) architecture with physics-based separation of semi-local, electrostatic and dispersion interactions. (2) We introduce the application of an ensemble knowledge distillation algorithm to lower the fluctuation of results obtained from MLFF-based MD simulations. (3) We propose a physics-inspired density alignment algorithm aiming at aligning MLFF-based MD simulations with experimental data. This approach requires only a minimal amount of experimental data and demonstrates considerable transferability to liquids not initially included in the alignment process. As a result, BAMBOO achieves state-of-the-art accuracy in predicting the density, viscosity and ionic conductivity of various liquid electrolytes. The high predictive ability of BAMBOO makes it a valuable tool for electrolyte design driven by molecular structure engineering.

Results

We illustrate the workflow of BAMBOO in Fig. 1a. Initially, we sample local atomic environments within liquid electrolytes as gas-phase clusters and subsequently use density functional theory (DFT) to compute their energies, atomic forces and charges. In this study, for showcasing the broad applicability of BAMBOO, we include diverse molecules and salts in the DFT dataset (Supplementary Fig. 1). In particular, our focus encompasses components commonly found in liquid electrolytes utilized in lithium-ion batteries, such as cyclic carbonates, linear carbonates, Li$^+$ cations, and PF$_6^-$, FSI$^-$ and TFSI$^-$ anions. Additionally, we incorporate two frequently used organic liquids—ethanol (EO) and acetone (ACT)—along with an engineering fluid Novec 7000 to showcase the generalizability of the trained model. Subsequently, the quantities calculated using DFT are used to train a group of GNNs with different random seeds. To lower the fluctuation of results from MD simulations, these independently trained GNNs are fused into a single, unified GNN via ensemble knowledge distillation[36]. Finally, we use experimentally measured data, specifically density in this context, to align the MLFF model with experimental observations. Further details about the DFT dataset, training methodology, ensemble knowledge distillation and density alignment algorithm are provided in the subsequent sections.

In Fig. 1b, we describe the separation of semi-local, electrostatic and dispersion interactions in BAMBOO. Given the significance of long-range electrostatic and dispersion interactions in simulating liquid electrolytes[37,38], we explicitly compute their corresponding energies in BAMBOO. BAMBOO takes the types of atom $\{Z_i\}$ and the three-dimensional coordinates of atoms $\{r_i\}$ as input, and computes the energy as follows. (1) Semi-local energy is modelled by a GNN comprising GET layers whose detailed structure will be introduced in the following subsection. GET takes the atom types $\{Z_i\}$ and the relative coordinates $\{\mathbf{r}_{ij} = \mathbf{r}_i - \mathbf{r}_j\}$ of edges (i,j) as the input, and outputs atom representations $\{x_i\}$ that encode the local environments of atoms. We further input $\{x_i\}$ into a multilayer perceptron to obtain the neural network-predicted energy $\{E_i^{NN}\}$. (2) Electrostatic energy is computed as follows. The atom representations $\{x_i\}$ are fed into another multilayer perceptron to predict the atomic partial charge $\{q_i\}$. We then compute the electrostatic energy $\{E_i^{elec}\}$ based on the predicted partial charge under the framework of charge equilibrium[39]. (3) Dispersion energy $\{E_i^{disp}\}$ is directly computed based on DFT-D3(CSO) correction[40].

Finally, the total atomic energy E_i^{total} is obtained by summing E_i^{NN}, E_i^{elec} and E_i^{disp}, and the total energy of the system E is computed by summing all the atomic energies. Further details of the energy model are provided in the Methods. Regarding the forces, as the energy is computed based on relative coordinates $\mathbf{r}_{ij}$ rather than absolute coordinates $\mathbf{r}_i$, the pairwise force can be naturally defined and computed as $\mathbf{f}_{ij} = -\frac{\partial E}{\partial r_{ij}} + \frac{\partial E}{\partial r_{ji}}$, satisfying Newton's third law ($\mathbf{f}_{ij} = -\mathbf{f}_{ji}$) for computing
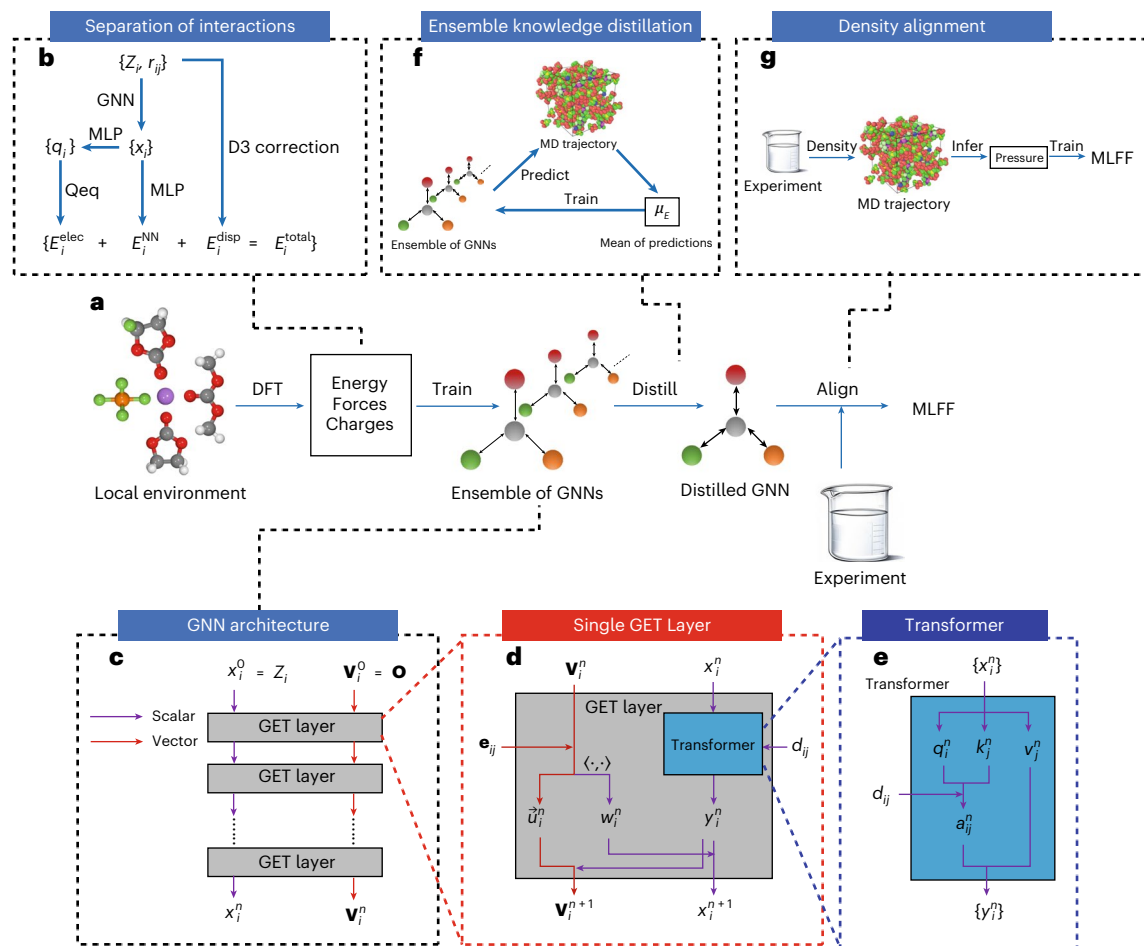
Fig. 1 | Overview of BAMBOO. a, Schematic of the training process of BAMBOO. **b**, Separation of interactions in BAMBOO. MLP, multilayer perceptron; Qeq, charge equilibrium model. **c–e**, Schematic of the GNN (**c**), the GET layer inside the GNN (**d**) and the transformer inside the GET layer (**e**). **f,g**, Schematic of ensemble knowledge distillation (**f**) and density alignment of BAMBOO (**g**).

the microscopic stress[41]. The atomic force is then computed by summing the pairwise forces: $\mathbf{f}_i = \sum_{j \neq i} \mathbf{f}_{ji}$.

Architecture of GET layers

In Fig. 1c–e, we present the GET architecture utilized in BAMBOO, which draws inspiration mainly from the architecture of TorchMD-NET[12]. Beginning with the atom types $\{Z_i\}$ and the displacement vectors between atoms $\{\mathbf{r}_{ij}\}$, we first initialize the atom scalar representation x_i^0 and the atom vector representation $\mathbf{v}_i^0$, as well as the edge scalar representation d_{ij} and the edge vector representation $\mathbf{e}_{ij}$. Subsequently, we feed x_i^0 and $\mathbf{v}_i^0$ into the GET layers to update them iteratively (Fig. 1c). As depicted in Fig. 1d, within each GET layer, scalar representation x_i^n undergoes a transformer layer to exchange information with its neighbours within a specified cut-off radius (5 Å in this study). Figure 1e illustrates the transformer block designed as the attention mechanism[42] on edges. After the transformer block, on the scalar side, we obtain the intermediate atom scalar representation y_i^n. On the vector side, along with $\mathbf{e}_{ij}$, the atom vector representation $\mathbf{v}_i^n$ is transformed into an intermediate vector representation $\mathbf{u}_i^n$ and another scalar representation w_i^n. Here the transformation from vector to scalar is achieved through the inner-product operation to maintain rotational invariance. Finally, we combine y_i^n and w_i^n to obtain the next-layer atom scalar representation x_i^{n+1}, and combine y_i^n, $\mathbf{u}_i^n$ and $\mathbf{v}_i^n$ to obtain the next-layer atom vector representation $\mathbf{v}_i^{n+1}$. In this process, scalar y_i^n is multiplied to a vector to preserve the equivariance of the vector representation. Overall, within each GET layer, all the neighbouring atoms

exchange information with one another through the transformer, and scalar and vector representations also exchange information with each other. Details of the GET architecture are provided in the Methods.

In Fig. 2a–c, we elucidate the roles of the equivariant feature, the transformer block and the prediction of partial charges by ablation studies. For predicting the energy and forces from DFT, GET demonstrates superior performance compared with the graph equivariant network (GE) and the graph invariant transformer (GIT). In particular, GE lacks the transformer layer, whereas GIT does not incorporate equivariant features. This comparison highlights the crucial role of both equivariant features and transformer within the GET, as evidenced by the smaller errors observed. Furthermore, we observe that equivariance may hold greater importance than the transformer, as GE demonstrates smaller errors than GIT, suggesting avenues for accelerating inference in future applications. In addition, we conduct a comparison with a model termed GET-no-charge, which excludes the prediction of partial charges. Interestingly, we observed that GET-no-charge outperforms GET in terms of force errors, but it suffers from larger errors in energy due to the long-range dependence of electrostatic energy (proportional to $1/r$) and the reduced capacity of GNN to capture long-range interactions as opposed to the short range[43]. As illustrated in Fig. 2c, GET exhibits the smallest error in density from MD simulations compared with GE, GIT and GET-no-charge, probably attributable to its overall superior performance in predicting the DFT quantities. These observations collectively emphasize the suitability of the GET architecture in BAMBOO for simulating liquid electrolytes compared
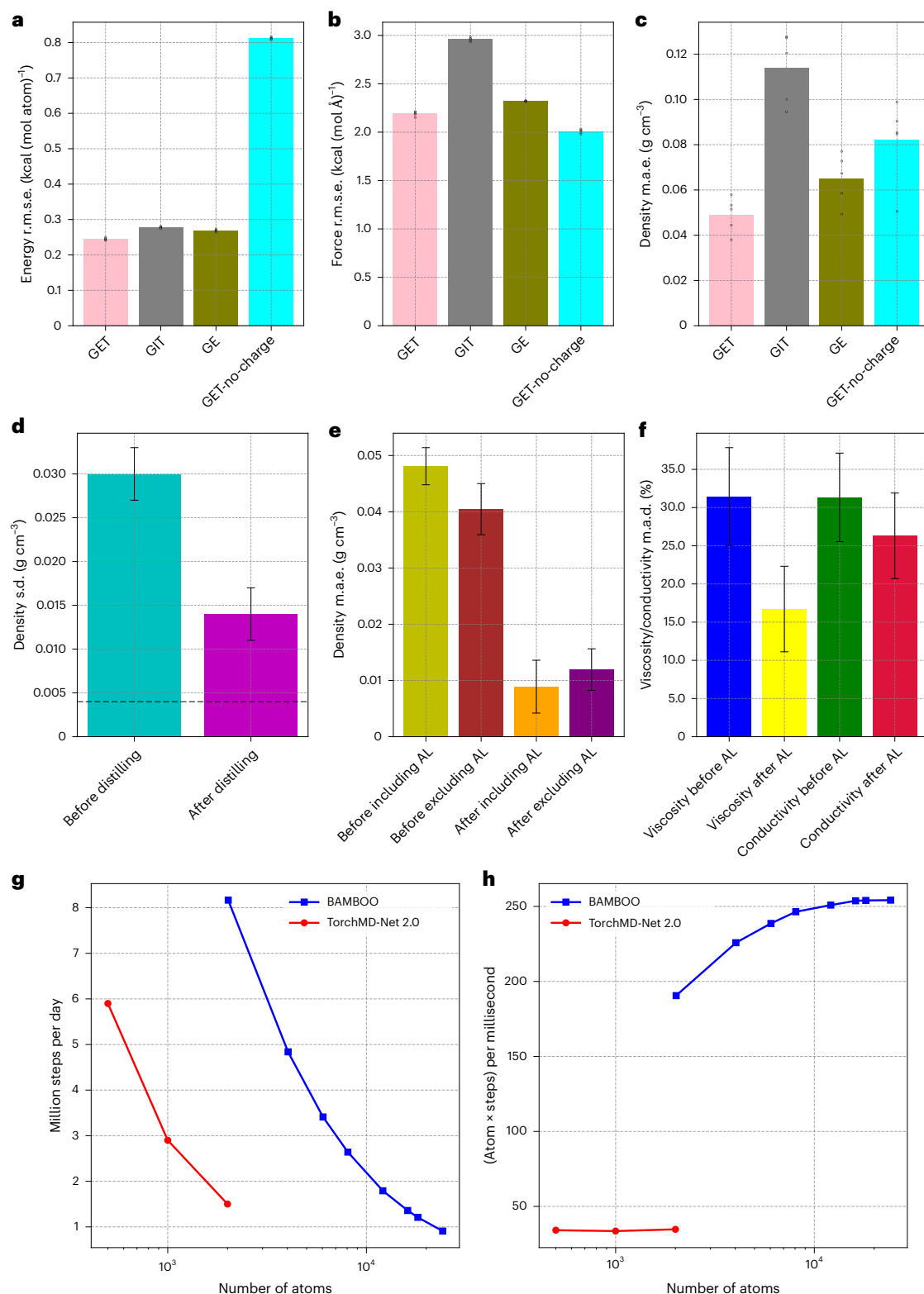
Fig. 2 | Effects of GET layers, ensemble knowledge distillation and density alignment. a–c, Effects of equivariance, transformer and charge in BAMBOO for predicting energy (**a**), force (**b**) and density (**c**). All the models are trained using the same DFT training set, excluding ensemble knowledge distillation and density alignment. The evaluation metrics include the root mean squared error (r.m.s.e.) and mean absolute error (m.a.e.). Here the bar plots reflect the mean values, and the overlaid dots are values from five models trained using different random seeds. **d**, Effect of ensemble knowledge distillation on reducing the standard deviation (s.d.) of density from the MD simulations by different randomly trained MLFFs. Data are presented as means ± s.d. of 13 samples shown

in Supplementary Fig. 7. **e,f**, Effect of density alignment (AL) on the predictions of density, viscosity and ionic conductivity from MD. The mean absolute deviation (m.a.d.) is used as a metric in **f**, expressed in percentage. Data are presented as means ± s.d. of 13, 15, 27 and 21 samples shown in Supplementary Fig. 10 and Supplementary Tables 12–15. **g,h**, Inference speed comparison between BAMBOO and TorchMD-Net 2.0 (ref. 48). We compare the three-layer BAMBOO with the two-layer TensorNet[74] module implemented in TorchMD-Net 2.0 (ref. 48). We delineate the inference speed in million steps per day in **g** and in atom times step per millisecond in **h**. Speeds are tested on a single NVIDIA A100 GPU.

Table 1 | Comparison of inference speed of BAMBOO with other GNN MLFFs on a gas-phase single molecule 3-(benzyloxy)pyridin-2-amine, focusing on fair comparison

Model	MACE (L=1)[47]	Allegro (L=1)[45]	VisNet (L=1)[46]	BAMBOO
Speed (ms)	17.5	13.0	12.2	**6.9**

BAMBOO is compared with other MLFFs using a spherical harmonic order of $L=1$, as BAMBOO exclusively utilizes scalar and vector embeddings. The experiments are conducted on an NVIDIA A100 GPU.

with previous MLFFs without equivariance, transformer and explicit computation of electrostatic interactions. Further results of the ablation study, along with details of the ablated models, are provided in Supplementary Table 5.

Compared with TorchMD-NET[12], our method enhances efficiency through two key strategies. On one hand, we utilize attention mechanisms on graphs, lowering the computational complexity from $\mathcal{O}(N^2)$ to $\mathcal{O}(N)$. On the other hand, we eliminate certain connections and neural network parameters that do not significantly contribute to the capacity of the model. As a result, our model benefits from a noticeable speed enhancement (Fig. 2g,h). In this work, we use the large-scale atomic/molecular massively parallel simulator (LAMMPS)[44] as the engine to run the MD simulations, and we design the interface between BAMBOO and LAMMPS by taking inspiration from Allegro[45]. As shown in Table 1 and Fig. 2g,h, relative to other GNN-based MLFFs like ViSNet[46], Allegro[45], MACE[47] or TorchMD-Net 2.0 (ref. 48), BAMBOO achieves a higher inference speed (two million steps per day for a system with 10,000 atoms on a single NVIDIA A100 graphics processing unit (GPU)). The inference of BAMBOO can be further accelerated by using multiple GPUs in parallel, which will be introduced in a future release.

Ensemble knowledge distillation and density alignment

In Fig. 1f, we elucidate the concept of ensemble knowledge distillation, which stems from the recognition that the inherent randomness in machine learning can introduce various challenges into MD simulations with MLFFs. Particularly in the case of liquid electrolytes, we observe that GNNs trained using different random seeds, despite exhibiting similar validation errors, may yield divergent macroscopic properties such as density. This discrepancy arises from two main factors: first, an MD simulation is inherently a random process[33], which, in liquid electrolytes, can manifest as density fluctuations of approximately 0.004 g cm^{-3} across different random seeds of MD by the same MLFF; second, during MD simulations, MLFFs often extrapolate to out-of-domain structures, particularly evident in liquid electrolytes in which the training data comprise gas-phase clusters and the inference domain is bulk liquids. In particular, for neural networks, the degree of randomness is more pronounced for out-of-domain inference compared with in-domain inference[49], leading to varying behaviours of MLFFs during MD simulations despite being trained on the same dataset.

To address this issue, our approach aims to mitigate the discrepancy among MLFFs in MD simulations by using an ensemble of MLFFs to predict the energy and forces of MD trajectories. Subsequently, we aggregate the mean predictions and utilize this mean value to further optimize the MLFFs. In Fig. 2d and Supplementary Fig. 7c, we demonstrate that ensemble knowledge distillation effectively reduces the standard deviation of density predictions from five models by more than 50%, from 0.030 g cm^{-3} to 0.014 g cm^{-3}. In particular, ensemble knowledge distillation does not require new DFT labels. Beyond BAMBOO and liquid electrolytes, this concept can be applied to any MLFF for any system. In Supplementary Fig. 8, we provide another example of utilizing ensemble knowledge distillation to lower the fluctuation of results from M3GNet[19] in simulating solid-state phase

transformation, further illustrating the generalizability of ensemble knowledge distillation.

As the final step of training BAMBOO, we introduce the concept of density alignment in Fig. 1g. For MLFFs targeting liquid electrolytes, we identify two potential sources of systematic bias between MLFF predictions and experimental data. On one hand, the choice of DFT functional, basis set and dispersion correction may lead to systematic bias on intermolecular interactions. On the other hand, the deviation between the training data composed of small gas-phase clusters and the application scenario of large bulk liquid structures may induce additional bias.

To address this, we propose aligning BAMBOO with experimental data. Due to the limited availability of experimental data and the high dimensionality of MLFFs, such alignment must be grounded in physics to ensure its transferability. Hence, we use density as the macroscopic observable for alignment, leveraging pressure as the physics-based link between the macroscopic and microscopic realms. Experimental density can be used to deduce the pressure adjustments required to align MD simulations with experimental density (Fig. 1g). Moreover, we can correlate these pressure adjustments with intermolecular forces, followed by utilizing the adjusted force to refine the parameters of BAMBOO.

In Fig. 2e, we show that density alignment effectively reduces density errors from around 0.05 g cm^{-3} to 0.01 g cm^{-3}. Importantly, this reduction in density error, achieved using only 13 experimental data points included in the alignment, transfers to liquids not originally part of the alignment, particularly different solvents and solutions with higher salt concentrations (Supplementary Figs. 10a and 12a). For example, the densities of molecules with bonding characteristics not included in the training set of the alignment, such as VC with the C=C double bond, EO with the −OH group and ACT with the ketone group, can be better predicted after alignment. Another notable example is that the alignment process can improve the prediction of density for electrolytes with salt concentrations (such as 3.70 mol kg^{-1}, 3.74 mol kg^{-1} and 3.78 mol kg^{-1}) much higher than those in the training set in which the highest salt concentration is just 2.22 mol kg^{-1}. This improvement suggests that density alignment, which adjusts the strength of intermolecular interactions, has some degree of transferability to molecules with similar structures or solutions with varying salt concentrations. Moreover, in Fig. 2f, we illustrate that density alignment decreases errors in the predictions of other properties beyond density, such as viscosity and ionic conductivity. Further theoretical analysis, training specifics of the density alignment, and the relationship between density, viscosity and ionic conductivity are detailed in the Methods.

From Fig. 2e,f and Supplementary Fig. 10, it is evident that BAMBOO demonstrates an average density error of 0.01 g cm^{-3}, viscosity deviation of 17% and ionic conductivity deviation of 26% across a diverse range of molecular liquids and solutions with varying salt concentrations. The level of error exhibited by BAMBOO represents the state-of-the-art compared with existing simulation studies (Supplementary Tables 6–8). Moreover, the error magnitude of BAMBOO is close to the degree of variation observed in experimental measurements. For instance, density variation from the same research group typically hovers around 0.01 g cm^{-3}, whereas viscosity and conductivity exhibit variations of approximately 2% and 1% (ref. 50), respectively. Across different research groups, the viscosity can vary up to approximately 20% in some cases[51–54], and approximately 5% for conductivity[50,55,56]. In Supplementary Fig. 12, we delve into the MD-simulated bulk and microscopic properties of solvents and simple electrolytes using BAMBOO, providing comprehensive insights into the predictions.

Beyond pure solvents and simple electrolytes, we also provide the results of BAMBOO simulations on more practical electrolytes with multiple (from four to eight) components[57] in Supplementary Tables 9 and 10. Although the density alignment process is mostly
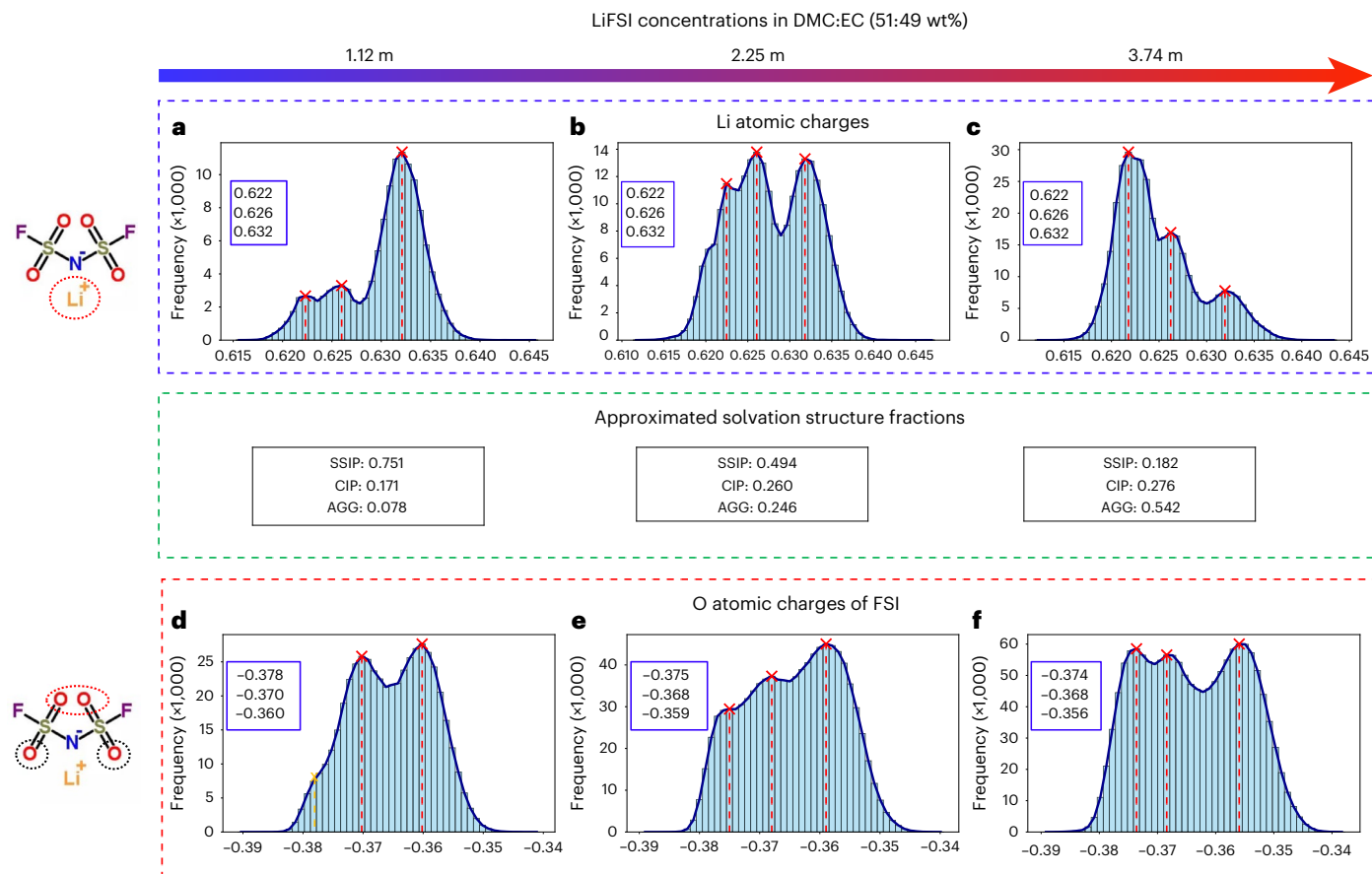
Fig. 3 | Atomic charge distributions and solvation structure fractions of three simulated LiFSI electrolytes using BAMBOO. a–c, Left to right: the Li charge histograms of 1.12 m (**a**), 2.25 m (**b**) and 3.74 m (**c**) LiFSI in a DMC:EC mixture of 51:49 wt%. The dotted lines within each panel, progressing from left to right, denote the Li charge distributions correlated with the SSIP, CIP and AGG solvation structures. **d–f**, Left to right: O_{FSI}– charge histograms illustrating the behaviour within the three LiFSI electrolytes mentioned above. In particular, the first dotted line may correlate to the distributions of paired O_{FSI}–. The second and third lines represent the two distinguished O_{FSI}– charges linked to the distributions of unpaired intramolecular interactions. The charge values associated with the peaks from left to right are tabulated in each panel.

based on simple systems such as pure solvents and electrolytes with fewer than four components, for practical multicomponent electrolytes, BAMBOO achieves a similar degree of prediction accuracy for density, viscosity and ionic conductivity compared with simple systems (Supplementary Fig. 10), which further demonstrates the transferability of density alignment in terms of the number of components in the system. More importantly, we show that for these multicomponent systems, BAMBOO achieves a higher prediction accuracy than the current off-the-shelf Optimized Potentials for Liquid Simulations-All Atoms (OPLS-AA) classical force field[58] for simulating the liquid electrolyte, which highlights the potential of BAMBOO on simulating and designing practical liquid electrolytes.

As for transferability to molecules not included in the DFT dataset, in Supplementary Table 11, we analyse the trend and degree of transferability of BAMBOO to unseen chemical species. We find that BAMBOO can stably simulate unseen molecules with bonding types similar to those included in the DFT training set (Supplementary Fig. 1). However, we find that for species with unseen elements and unseen bond types (such as the C≡N triple bond), BAMBOO cannot achieve long-time stable MD simulations. For the molecules that can be stably simulated by BAMBOO, we can see that BAMBOO has a density m.a.e. of 0.0495 g cm⁻³ and the mean viscosity deviation of 52.4%, which is higher than that of systems included in the DFT training set (0.011 g cm⁻³ and 16.7%, respectively). To further reveal the trend of density prediction on unseen molecules, we conduct MD simulations on fluorinated solvent molecules (Supplementary Fig. 14). We find that

all the MD simulations of the fluorinated solvents (ethylene carbonate (EC), dimethyl carbonate (DMC) and ethyl acetate (EA)) remain stable within a broad temperature range of 283–343 K. For density prediction, we find that the more structurally alike the unseen fluorinated molecule is compared with the base molecule, the more accurate is the prediction of density from BAMBOO.

To make MLFF more transferable to out-of-sample systems, we need constant effort to improve the MLFFs. In general, there are two approaches to elevate the transferability of MLFFs. The first approach is to conduct large-scale pretraining on datasets with diverse molecules, such as MACE-OFF23 (ref. 24), which is trained on nearly a million conformers. Although the strategy of large-scale pretraining is very popular in the recent wave of universal force fields[19–23], it is not proved to guarantee stable MD simulations for all molecules. The other approach is to combine the machine learning model with strong physical restrictions, such as our latest work that uses machine learning models to parameterize force fields with classical functional forms[59], which ensures that the force field can stably simulate any molecule without a spurious reaction during the MD simulation. However, such restrictions also limit the usage of MLFF on cases where bond breaking happens. Therefore, we think that it is important to further develop MLFF that can stably simulate as many molecules as possible and preserve the ability to model bond-breaking events when necessary. In a very recent work[60], it is shown that the GNN-based model is more transferable to new molecules than a local-descriptor-based model, which we hope can inspire further

development of machine learning architecture for higher transferability of force fields.

Solvation structures and atomic partial charges

In addition to the accurate predictions of macroscopic properties, BAMBOO's ability to explicitly predict atomic partial charges enables it to provide additional insights about solvation structures in liquid electrolytes. In Fig. 3, we present the Li^+ and O_{FSI}^- charge histograms derived from the simulations of three distinct LiFSI electrolytes spanning from 1.12 m to 3.74 m (m is the molality denoting the moles of salt per kilogram of solvent). The weight ratio of DMC to EC, set at 51:49 wt%, remains consistent across all these electrolytes. The histograms of all the other atoms of FSI^- are presented in Supplementary Fig. 15.

Three distinct peaks can be observed in Fig. 3a, each intersecting the x axis (Li^+ charges) at approximately 0.622, 0.626 and 0.632, respectively. As we transit from Fig. 3a to 3c, the peak locations roughly persist, although with varying heights. Specifically, the highest peak undergoes a shift from approximately 0.632 to 0.622 as the LiFSI concentration increases from 1.12 m to 3.74 m.

This shift in Li^+ charge distributions evinces the evolving solvation structure populations across different salt concentrations. To demonstrate this interpretation, we analyse the frames of the final 3 ns of MD simulations, and examine the solvation structures around the Li^+ ions. The types of solvation structure and their fractions are tabulated and shown in each Li^+ charge histogram. These fractions are derived by first analysing the radial distribution of the simulation trajectory for the radius of the first solvation shell around Li^+, followed by extracting the central Li^+ ions along with the surrounding solvated molecules. We find that the first solvation shell has a radius of around 2.2 Å and the coordination numbers in all the three systems are around 4 (Supplementary Fig. 16).

In general, the solvation structures are classified as SSIP, CIP and AGG. Here SSIP denotes the structure solely composed of solvent molecules; CIP, the structure containing one anion; and AGG, the collective structure composed of more than one anion in the first solvation shell. Our analysis reveals SSIP ratios of 0.751, 0.494 and 0.182 for LiFSI concentrations of 1.12 m, 2.25 m and 3.74 m, respectively. The CIP ratios are 0.171, 0.260 and 0.276 and the AGG ratios are 0.078, 0.246 and 0.542. The difference in ratios between SSIP, CIP and AGG in each panel and their changes as a function of salt concentrations roughly mirror the peak height evolution (Fig. 3a,c).

It is well known that solvation interactions between Li^+ ions and the surrounding molecules, including solvents and anions, primarily occur through Li^+–O interactions in LiFSI-based carbonate electrolytes. These intermolecular interactions may lead to polarization-induced changes in the partial charge within the local solvation environment. To demonstrate this, in addition to analysing Li^+, we also examine the atomic charge distributions across all types of atom within the FSI^- anion. Interestingly, we observe that O_{FSI}^- exhibits three overlapping distributions (Fig. 3d–f), whereas the other atom types display a single distribution for all the concentrations (Supplementary Fig. 15).

As depicted in Fig. 3d, two peaks (right and middle) are distinctly visible, which are highlighted using the red dotted lines. We posit that these peaks originate from the two types of oxygen atom within the FSI^- molecule. This hypothesis stems from our observation that at the 1.12 m LiFSI concentration, most of the anions should remain unpaired due to low CIP and AGG fractions. In addition, these unpaired anions are probably not solvated by solvent molecules. In lithium-ion battery electrolytes, solvents such as EC and DMC are designed to primarily solvate positive ions, resulting in anions often becoming free from forming solvent–anion pairs[2]. Consequently, the two prominent distributions with peak locations at around −0.370 and −0.360 are indicative of potential intramolecular differences in O_{FSI} values of free FSI^- in

the simulated 1.12 m LiFSI electrolyte. An additional peak (Fig. 3d, yellow dotted line) is visually identified based on its similarity in charge location compared with those shown in Fig. 3e,f. We observe that this peak progressively becomes more prominent as the concentration increases from 1.12 m to 3.74 m. We hypothesize that this peak is associated with ion pairing, as the growth in peak height correlates the increasing total fractions of CIP and AGG. More discussion about the solvation structure and charge prediction from BAMBOO is provided in Supplementary Figs. 15 and 16.

Discussions

In summary, this study presents an MLFF framework tailored for MD simulations of liquid electrolytes. First, we devise a GET architecture that segregates semi-local, electrostatic and dispersion interactions, leveraging knowledge from DFT calculations. Second, we introduce an ensemble knowledge distillation on MLFF, suppressing the fluctuation of results from MD simulations based on MLFF. Moreover, we propose a physics-based alignment approach to reconcile the simulated density with experimental data, thereby establishing a connection between macroscopic and microscopic scales. Our results demonstrate the effectiveness of the density alignment process in reducing disparities between simulated and experimental outcomes, extending its benefits to properties beyond density. To further enhance the model's performance in predicting properties beyond density, such as conductivity, a simultaneous and direct alignment of these additional properties is essential. Hence, one of our future endeavours will be dedicated towards expanding the alignment process to encompass a wider spectrum of properties.

We conduct a comprehensive assessment for the performance of BAMBOO on solvents and liquid electrolytes. Our simulated results demonstrate that a single BAMBOO model can predict densities, viscosities and ionic conductivities for a broad range of chemistries with high accuracy. The current BAMBOO model is able to simulate up to 15 species in various compositions. This robustness underscores BAMBOO's value in facilitating the design and optimization of practical liquid electrolytes, which may contain up to ten components. In addition to the state-of-the-art accuracy for predicting bulk properties, we show the quantification of the relationship between solvation structures and atomic partial charges as a function of electrolyte compositions, providing insights into solvation engineering that are unreachable by classical force fields or DFT. Compared with current classical (polarizable) force fields, since BAMBOO is an MLFF without an explicit restriction of functional forms, it has the potential to be extended to other areas beyond simulating bulk properties, such as simulating chemical reactions in liquids[61]. We envision this work as laying the foundation for the development of universal MLFFs capable of accurately simulating properties and behaviours of most organic liquids.

Methods
GET

Given the atomic structures, we first initialize node scalar representation x_i^0 and node vector representation $\mathbf{V}_i^0$ for atom i based on the atom type z_i, and initialize edge scalar representation d_{ij} and edge vector representation $\mathbf{e}_{ij}$ based on the relative vector between atom i and atom j ($\mathbf{r}_{ij}$) within a certain cut-off radius r_{cut} (5 Å in this work). After initialization, we input x_i^0, $\mathbf{V}_i^0$, d_{ij} and $\mathbf{e}_{ij}$ into the GET layers. In the nth layer, we first input x_i^n into a LayerNorm[62] layer; then, we input x_i^n, $||r_{ij}||$ and d_{ij} into the classical qkv transformer shown in Fig. 1e. In the transformer block, we compute the attention weight a_{ij}^n and the intermediate node scalar representation y_i^n. Furthermore, we build an intermediate vector representation $\mathbf{u}_i^n$ to incorporate edge vector representation $\mathbf{e}_{ij}$ into the GET layer and sum the contributions of neighbouring atoms into the vector representation. Finally, we output the updated node scalar representation x_i^{n+1} and node vector representation $\mathbf{V}_i^{n+1}$ for the (n+1)th layer by combining all the scalar and vector information on the

node level and edge level. More details of the GET model are provided in Supplementary equations (1)–(21).

Separation of interactions

Once we reach the final GET layer, we compute neural network-based per atom energy E_i^{NN} and atomic partial charges q_i based on x_i^n using two fully connected networks. Then, we define the electrostatic energy model based on the charge equilibrium theory[39]. In this work, we incorporate the charge equilibrium model into our MLFF for efficiency. We further use a regularization strategy to make our GNN model directly output the charges from the charge equilibrium model such that an expensive iterative equilibrium solver per MD timestep is completely removed. In addition, given the atomic structures, we can directly compute the dispersion interaction (E_i^{disp} and $\mathbf{f}_i^{disp}$) using the D3 correction[40]. Finally, we combine the neural network, electrostatic and dispersion interactions to compute the total energy and force of each atom. More details of the separation of interactions are provided in Supplementary equations (22)–(45), Supplementary Table 1 and Supplementary Figs. 2 and 3.

Initial training of BAMBOO

Given the design of the GET architecture and the scheme to separate different interactions, we use a loss function combining the energy, force, virial and charge information to train the neural networks in BAMBOO. Specifically, the charge learned by BAMBOO is determined by four aspects: the first is to fit the energy, force and virial together with the contribution from the neural network; the second is to fit the electrostatic potential by fitting the charge from electrostatic potentials[63]; the third is to fit the dipole moment, a macroscopic observable derived from charge densities by DFT; and the fourth is to approximate the charge equilibrium.

In this work, we build the neural networks in BAMBOO using PyTorch[64], and we use the AdaMax optimizer[65] at a learning rate of 10^{-2}. In the training process, the weight decay is set to 0.001, whereas the learning rate decreases at a rate of 0.99. The batch size is 128 for both training and validation. The training process is composed of 750 training epochs. The model consists of three layers with an embedding dimension of 64. It features 16 attention heads, uses a cut-off radius of 5.0 Å for each message passing layer and includes the fully connected networks with two layers for predicting the energy and charge.

Here we use 720,000 clusters with the DFT-calculated energy, atomic forces, virial tensor, dipole moment and partial charge to train BAMBOO. The calculations are performed utilizing the B3LYP[66] functional and the def2-SVPD[67] basis set, supplemented by density fitting techniques based on the def2-universal-jkfit auxiliary basis[68]. Throughout the computational process, the convergence threshold for the self-consistent field is set to 1.0×10^{-10} a.u., with a maximum allowance of 100 iterations. Here we use the open-source GPU4PySCF package[69] developed by ByteDance. Compared with Q-Chem[70] on 32 CPUs, GPU4PySCF on a single V100 GPU can save up to 97% computation time and 95% cost. We randomly choose 90% data as the training set and 10% as the validation set to evaluate the performance of the training process. More details of the initial training of BAMBOO are provided in Supplementary equation (46), Supplementary Table 2 and Supplementary Figs. 4–6.

Ensemble knowledge distillation for MLFF

In this study, we first train five models purely using the DFT data with different random seeds. Subsequently, we randomly select one model to run 600-ps isothermal–isobaric (NPT) MD simulations and 400-ps canonical (NVT) MD simulations for all the systems (Supplementary Fig. 7c). With the obtained trajectories, we use all five models to predict the energy, force and virial tensor of 100 randomly selected frames from the trajectory of each system. Finally, we use the mean predictions

as the labels to fine-tune the randomly selected model by the loss function (Supplementary equation (52)). More details of the ensemble knowledge distillation of BAMBOO are provided in Supplementary equations (47)–(53), Supplementary Tables 3 and 4 and Supplementary Figs. 7 and 8.

Density alignment

We use the ensemble distilled model as the foundational base model for performing density alignment. To initiate the alignment, we first calculate the compressibility (β) of each liquid. This is achieved by training the distilled BAMBOO model to adapt to uniformly assigned pressure changes (ΔP). Utilizing these differently aligned MLFFs, we conduct 600-ps NPT simulations to gather the resulting densities (ρ). Consequently, as depicted in Supplementary Fig. 9a–m, we establish a linear correlation between $\Delta\ln[\rho]$ and the assigned ΔP for each liquid. Leveraging this linearity, we can determine the β value for each liquid. This enables us to identify the specific ΔP value that should be incorporated into the base MLFF for each liquid to accurately align the $\Delta\ln[\rho]$. Finally, we use the loss function specified in Supplementary equation (58)—applying the unique ΔP value identified for each liquid—to align the base MLFF with the experimental data. More details on the density alignment of BAMBOO are provided in Supplementary equations (54)–(58) and Supplementary Figs. 9–11.

MD

We use LAMMPS[44] to conduct the MD simulations for BAMBOO. For each system, the initial step is energy minimization, which is constrained to a maximum of 1,000 iterations and 100,000 evaluations. Following this, the liquid systems are subjected to equilibrium under the NPT ensemble for 1 ns, with a step size maintained at 1 fs. The density values presented in this paper represent the mean value during the final 50 ps of the NPT simulation. Subsequently, we carry out a 4-ns production run under the NVT ensemble to determine the viscosity, diffusivity and ionic conductivity of the systems. Two approaches are used to calculate the ionic conductivities: the Nernst–Einstein method and a method described in another work[71] built on Stefan–Maxwell diffusivities. More details of the MD simulations are provided in Supplementary equations (64)–(68).

Data availability

DFT datasets of clusters are available at https://huggingface.co/datasets/mzl/bamboo (ref. 72). The input parameters, template of atomic structures for LAMMPS MD simulations and the final trained, ensemble knowledge distilled and density-aligned model of BAMBOO to reproduce the results in the paper are available via Zenodo at https://doi.org/10.5281/zenodo.14603020 (ref. 73).

Code availability

The source codes, including the GET model, the training module and the LAMMPS interface for the MD simulations, are available via GitHub at https://github.com/bytedance/bamboo and via Zenodo at https://zenodo.org/records/14603020 (ref. 73).

References

1. Xu, K. Nonaqueous liquid electrolytes for lithium-based rechargeable batteries. *Chem. Rev.* **104**, 4303–4418 (2004).
2. Xu, K. *Electrolytes, Interfaces and Interphases: Fundamentals and Applications in Batteries* (Royal Society of Chemistry, 2023).
3. Meng, Y. S., Srinivasan, V. & Xu, K. Designing better electrolytes. *Science* **378**, eabq3750 (2022).
4. Unke, O. T. et al. Machine learning force fields. *Chem. Rev.* **121**, 10142–10186 (2021).
5. Frenkel, D. & Smit, B. *Understanding Molecular Simulation: From Algorithms To Applications* (Elsevier, 2023).

6. Behler, J. & Parrinello, M. Generalized neural-network representation of high-dimensional potential-energy surfaces. *Phys. Rev. Lett.* **98**, 146401 (2007).

7. Zhang, L., Han, J., Wang, H., Car, R. & E, W. Deep potential molecular dynamics: a scalable model with the accuracy of quantum mechanics. *Phys. Rev. Lett.* **120**, 143001 (2018).

8. Lysogorskiy, Y. et al. Performant implementation of the atomic cluster expansion (PACE) and application to copper and silicon. *npj Comput. Mater.* **7**, 97 (2021).

9. Schütt, K. T., Sauceda, H. E., Kindermans, P.-J., Tkatchenko, A. & Müller, K.-R. SchNet—a deep learning architecture for molecules and materials. *J. Chem. Phys.* **148**, 241722 (2018).

10. Batzner, S. et al. E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nat. Commun.* **13**, 2453 (2022).

11. Liao, Y.-L. & Smidt, T. Equiformer: equivariant graph attention transformer for 3D atomistic graphs. In *Proc. Eleventh International Conference on Learning Representations* https://openreview.net/forum?id=KwmPfARgOTD (ICLR, 2023).

12. Thölke, P. & De Fabritiis, G. TorchMD-NET: equivariant transformers for neural network based molecular potentials. In *Proc. Tenth International Conference on Learning Representations* https://openreview.net/forum?id=zNHzqZ9wrRB (ICLR, 2022).

13. Ko, T. W., Finkler, J. A., Goedecker, S. & Behler, J. A fourth-generation high-dimensional neural network potential with accurate electrostatics including non-local charge transfer. *Nat. Commun.* **12**, 398 (2021).

14. Zhang, L. et al. A deep potential model with long-range electrostatic interactions. *J. Chem. Phys.* **156**, 124107 (2022).

15. Anstine, D., Zubatyuk, R. & Isayev, O. AIMNet2: a neural network potential to meet your neutral, charged, organic, and elemental-organic needs. Preprint at *ChemRxiv* https://doi.org/10.26434/chemrxiv-2023-296ch-v3 (2023).

16. Unke, O. & Meuwly, M. PhysNet: a neural network for predicting energies, forces, dipole moments and partial charges. *J. Chem. Theory Comput.* **15**, 3678–3693 (2019).

17. Unke, O. et al. SpookyNet: learning force fields with electronic degrees of freedom and nonlocal effects. *Nat. Commun.* **12**, 12 (2021).

18. Yu, H. et al. Spin-dependent graph neural network potential for magnetic materials. *Phys. Rev. B* **109**, 144426 (2023).

19. Chen, C. & Ong, S. A universal graph deep learning interatomic potential for the periodic table. *Nat. Comput. Sci.* **2**, 718–728 (2022).

20. Deng, B. et al. CHGNet as a pretrained universal neural network potential for charge-informed atomistic modelling. *Nat. Mach. Intell.* **5**, 1031–1041 (2023).

21. Merchant, A. et al. Scaling deep learning for materials discovery. *Nature* **624**, 80–85 (2023).

22. Zhang, D. et al. DPA-2: towards a universal large atomic model for molecular and material simulation. *npj Comput. Mater.* https://doi.org/10.1038/s41524-024-01493-2 (2024).

23. Batatia, I. et al. A foundation model for atomistic materials chemistry. Preprint at https://arxiv.org/abs/2401.00096 (2024).

24. Kovács, D. P. et al. MACE-OFF23: transferable machine learning force fields for organic molecules. Preprint at https://arxiv.org/abs/2312.15211v1 (2023).

25. Wang, H. & Yang, W. Force field for water based on neural network. *J. Phys. Chem. Lett.* **9**, 3232–3240 (2018).

26. Zhang, J., Pagotto, J., Gould, T. & Duignan, T. T. Accurate, fast and generalisable first principles simulation of aqueous lithium chloride. Preprint at https://arxiv.org/abs/2310.12535v1 (2023).

27. Magdău, I. B. et al. Machine learning force fields for molecular liquids: ethylene carbonate/ethyl methyl carbonate binary solvent. *npj Comput. Mater.* **9**, 146 (2023).

28. Montes-Campos, H., Carrete, J., Bichelmaier, S., Varela, L. M. & Madsen, G. K. H. A differentiable neural-network force field for ionic liquids. *J. Chem. Inf. Model* **62**, 88–101 (2022).

29. Wang, F. & Cheng, J. Understanding the solvation structures of glyme-based electrolytes by machine learning molecular dynamics. *Chinese J. Struct. Chem.* **42**, 100061 (2023).

30. Dajnowicz, S. et al. High-dimensional neural network potential for liquid electrolyte simulations. *J. Phys. Chem. B* **126**, 08 (2022).

31. Jacobson, L. et al. Transferable neural network potential energy surfaces for closed-shell organic molecules: extension to ions. *J. Chem. Theory Comput.* **18**, 03 (2022).

32. Fu, X. et al. Forces are not enough: benchmark and critical evaluation for machine learning force fields with molecular simulations. *Trans. Mach. Learn. Res.* https://openreview.net/pdf?id=A8pqQipwkt (2023).

33. Ormeño, F. & General, I. Convergence and equilibrium in molecular dynamics simulations. *Commun. Chem.* **7**, 02 (2024).

34. Wang, X. et al. DMFF: an open-source automatic differentiable platform for molecular force field development and molecular dynamics simulation. *J. Chem. Theory Comput.* **19**, 5897–5909 (2023).

35. Greener, J. G. & Jones, D. T. Differentiable molecular simulation can learn all the parameters in a coarse-grained force field for proteins. *PLoS ONE* **16**, e0256990 (2021).

36. Asif, U., Tang, J. & Harrer, S. Ensemble knowledge distillation for learning improved and efficient network. In *24th European Conference on Artificial Intelligence* http://ecai2020.eu/papers/521_paper.pdf (ECAI, 2020).

37. Chandler, D., Weeks, J. D. & Andersen, H. C. Van der Waals picture of liquids, solids, and phase transformations. *Science* **220**, 787–794 (1983).

38. Kontogeorgis, G. M., Maribo-Mogensen, B. & Thomsen, K. The Debye-Hückel theory and its importance in modeling electrolyte solutions. *Fluid Ph. Equilib.* **462**, 130–152 (2018).

39. Poier, P., Lagardère, L., Piquemal, J.-P. & Jensen, F. Molecular dynamics using nonvariational polarizable force fields: theory, periodic boundary conditions implementation, and application to the bond capacity model. *J. Chem. Theory Comput.* **2019**, 09 (2019).

40. Schröder, H., Creon, A. & Schwabe, T. Reformulation of the D3(Becke–Johnson) dispersion correction without resorting to higher than C_6 dispersion coefficients. *J. Chem. Theory Comput.* **11**, 3163–3170 (2015).

41. Torres-Sánchez, A., Vanegas, J. M. & Arroyo, M. Geometric derivation of the microscopic stress: a covariant central force decomposition. *J. Mech. Phys. Solids* **93**, 224–239 (2016).

42. Vaswani, A. et al. Attention is all you need. In *Advances in Neural Information Processing Systems* Vol. 30 (eds Guyon, I. et al.) https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf (Curran, 2017).

43. Gong, S. et al. Examining graph neural networks for crystal structures: limitations and opportunities for capturing periodicity. *Sci. Adv.* **9**, eadi3245 (2023).

44. Thompson, A. P. et al. LAMMPS—a flexible simulation tool for particle-based materials modeling at the atomic, meso, and continuum scales. *Comp. Phys. Comm.* **271**, 108171 (2022).

45. Musaelian, A. et al. Learning local equivariant representations for large-scale atomistic dynamics. *Nat. Commun.* **14**, 579 (2023).

46. Wang, Y. et al. Enhancing geometric representations for molecules with equivariant vector-scalar interactive message passing. *Nat. Commun.* https://doi.org/10.1038/s41467-023-43720-2 (2024).

47. Batatia, I., Kovacs, D. P., Simm, G., Ortner, C. & Csányi, G. MACE: higher order equivariant message passing neural networks for fast and accurate force fields. *Adv. Neural Inf. Process. Syst.* **35**, 11423–11436 (2022).

48. Pelaez, R. P. et al. TorchMD-Net 2.0: fast neural network potentials for molecular simulations. Preprint at https://arxiv.org/abs/2402.17660 (2024).

49. Martius, G. & Lampert, C. H. Extrapolation and learning equations. In *Proc. 35th International Conference on Machine Learning* https://proceedings.mlr.press/v80/sahoo18a.html (PMLR, 2018).

50. Dave, A. R. *Automated Design and Discovery of Liquid Electrolytes for Lithium-Ion Batteries*. PhD thesis, Carnegie Mellon Univ. (2023).

51. Hagiyama, K. et al. Physical properties of substituted 1,3-dioxolan-2-ones. *Chem. Lett.* **37**, 210–211 (2008).

52. Sasaki, Y. in *Fluorinated Materials for Energy Conversion* (eds Nakajima, T. & Groult, H.) 285–304 (Elsevier Science, 2005).

53. Jänes, A., Thomberg, T., Eskusson, J. & Lust, E. Fluoroethylene carbonate and propylene carbonate mixtures based electrolytes for supercapacitors. *ECS Trans.* **58**, 71 (2014).

54. Gores, H. J. et al. in *Handbook of Battery Materials* 525–626 (John Wiley & Sons, 2011).

55. Dave, A. et al. Autonomous optimization of non-aqueous Li-ion battery electrolytes via robotic experimentation and machine learning coupling. *Nat. Commun.* **13**, 5454 (2022).

56. Logan, E. R. et al. A study of the transport properties of ethylene carbonate-free Li electrolytes. *J. Electrochem. Soc.* **165**, A705–A716 (2018).

57. Zhu, S. et al. Differentiable modeling and optimization of non-aqueous Li-based battery electrolyte solutions using geometric deep learning. *Nat. Commun.* **15**, 8649 (2024).

58. Jorgensen, W. L., Maxwell, D. S. & Tirado-Rives, J. Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. *J. Am. Chem. Soc.* **118**, 11225–11236 (1996).

59. Zheng, T. et al. Data-driven parametrization of molecular mechanics force fields for expansive chemical space coverage. *Chem. Sci.* **16**, 2730–2740 (2025).

60. Niblett, S. P., Kourtis, P., Magdău, I. B., Grey, C. P. & Csányi, G. Transferability of datasets between machine-learning interaction potentials. Preprint at https://arxiv.org/abs/2409.05590 (2024).

61. Zhang, H., Juraskova, V. & Duarte, F. Modelling chemical processes in explicit solvents with machine learning potentials. *Nat. Commun.* **15**, 6114 (2024).

62. Lei Ba, J., Kiros, J. R. & Hinton, G. E. Layer normalization. Preprint at https://arxiv.org/abs/1607.06450 (2016).

63. Breneman, C. M. & Wiberg, K. B. Determining atom-centered monopoles from molecular electrostatic potentials. The need for high sampling density in formamide conformational analysis. *J. Comput. Chem.* **11**, 361–373 (1990).

64. Paszke, A. et al. PyTorch: an imperative style, high-performance deep learning library. *Adv. Neural Inf. Process. Syst.* 32 (2019).

65. Kingma, D. P. & Ba, J. Adam: a method for stochastic optimization. Preprint at https://arxiv.org/abs/1412.6980 (2017).

66. Becke, A. D. Density functional thermochemistry. III. The role of exact exchange. *J. Chem. Phys.* **98**, 5648–5652 (1993).

67. Hellweg, A. & Rappoport, D. Development of new auxiliary basis functions of the Karlsruhe segmented contracted basis sets including diffuse basis functions (def2-SVPD, def2-TZVPPD, and def2-QVPPD) for RI-MP2 and RI-CC calculations. *Phys. Chem. Chem. Phys.* **17**, 1010–1017 (2015).

68. Weigend, F. Hartree–Fock exchange fitting basis sets for H to Rn. *J. Comput. Chem.* **29**, 167–175 (2008).

69. Wu, X. et al. Python-based quantum chemistry calculations with GPU acceleration. Preprint at https://arxiv.org/abs/2404.09452v1 (2024).

70. Shao, Y. et al. Advances in molecular quantum chemistry contained in the Q-Chem 4 program package. *Mol. Phys.* **113**, 184–215 (2015).

71. Mistry, A., Yu, Z., Cheng, L. & Srinivasan, V. On relative importance of vehicular and structural motions in defining electrolyte transport. *J. Electrochem. Soc.* **170**, 110536 (2023).

72. Mu, Z. mzl/bamboo. *Hugging Face* https://doi.org/10.57967/hf/3971 (2025).

73. Mu, Z. muzhenliang/bamboo: v0.1. *Zenodo* https://doi.org/10.5281/zenodo.14603020 (2025).

74. Simeon, G. & De Fabritiis, G. TensorNet: Cartesian tensor representations for efficient learning of molecular potentials. *Adv. Neural Inf. Process. Syst.* **36**, 37334–37353 (2024).

Author contributions

Conceptualization: Y.Z., W.Y., W.G., Z.M., Z. Yu, S.G., T.Z., X.H., Z. Yang, Z.W., L.C., X.W., S.S. and L.X. Methodology: S.G., Y.Z., Z.M., Z.P., H.W., M.C., W.Y. and W.G. Investigation: S.G., Y.Z., Z.M., Z.P., H.W., M.C., X.H., Z. Yu, W.Y. and W.G. Supervision: W.G., W.Y. and L.X. Writing: S.G., Y.Z., Z.M., Z.P., H.W., W.Y. and W.G.

Competing interests

All authors were employees of ByteDance when conducting this research project. ByteDance holds intellectual property rights pertinent to the research presented here. Furthermore, the innovations described here have resulted in the filing of a patent application in China (application no. 202311322469.2), which is currently pending.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42256-025-01009-7.

Correspondence and requests for materials should be addressed to Weihao Gao or Wen Yan.

Peer review information *Nature Machine Intelligence* thanks Jiayu Peng and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

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