Check for updates

View Article Online
View Journal

Digital Discovery

Accepted Manuscript

This article can be cited before page numbers have been issued, to do this please use: M. Sanocki and J. Zavadlav, *Digital Discovery*, 2025, DOI: 10.1039/D5DD00570A.

This is an Accepted Manuscript, which has been through the Royal Society of Chemistry peer review process and has been accepted for publication.

Accepted Manuscripts are published online shortly after acceptance, before technical editing, formatting and proof reading. Using this free service, authors can make their results available to the community, in citable form, before we publish the edited article. We will replace this Accepted Manuscript with the edited and formatted Advance Article as soon as it is available.

You can find more information about Accepted Manuscripts in the Information for Authors.

Please note that technical editing may introduce minor changes to the text and/or graphics, which may alter content. The journal's standard Terms & Conditions and the Ethical guidelines still apply. In no event shall the Royal Society of Chemistry be held responsible for any errors or omissions in this Accepted Manuscript or any consequences arising from the use of any information it contains.

ROYAL SOCIETY OF CHEMISTRY

rsc.li/digitaldiscovery

View Article Online
DOI: 10.1039/D5DD00570A

Journal Name

ARTICLE TYPE

Cite this: DOI: 00.0000/xxxxxxxxxx

Generalization of Long-Range Machine Learning Potentials in Complex Chemical Spaces

Michał Sanocki,[a] and Julija Zavadlav * [a]

Received Date
Accepted Date

DOI: 00.0000/xxxxxxxxxx

The vastness of chemical space makes generalization a central challenge in the development of machine learning interatomic potentials (MLIPs). While MLIPs could enable large-scale atomistic simulations with near-quantum accuracy, their usefulness is often limited by poor transferability to out-of-distribution samples. Here, we systematically evaluate different MLIP architectures with long-range corrections across diverse chemical spaces and show that such schemes are essential, not only for improving in-distribution performance but, more importantly, for enabling significant gains in transferability to unseen regions of chemical space. To enable a more rigorous benchmarking, we introduce biased train–test splitting strategies, which explicitly test the model performance in significantly different regions of chemical space. Together, our findings highlight the importance of long-range modeling for achieving generalizable MLIPs and provide a framework for diagnosing systematic failures across chemical space. While this study focuses on metal–organic frameworks and related systems, the proposed methodology is not limited to this class of materials and may inform the design of more robust and transferable MLIPs in other systems."

Open Access Article. Published on 13 April 2026. Downloaded on 4/14/2026 6:36:52 PM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

Digital Discovery Accepted Manuscript

Introduction

One of the largest issues in data-driven chemical modeling is the generalizability of the developed methods over the vast chemical space[1–3]. In fact, even the chemical space of only small organic molecules has been estimated to encapsulate around 10^{60} possibilities[4]. This challenge is not limited to computational methods, as experimental methods also struggle with the sheer enormity of the possible molecular space[5]. Paradoxically, the difficulty in exploring chemical space has often been the reason for adopting data-driven approaches[6,7]. Navigating such a vast chemical space poses a fundamental challenge for predictive modeling: no matter how large a dataset is, it will inevitably cover only a minute fraction of possible chemistries. As a result, the core problem for universal models is not whether they can interpolate within known regions of chemical space, but whether they can generalize to unseen regions.

This is especially problematic for the development of machine learning interatomic potentials (MLIPs), which have become increasingly popular in recent years due to their ability to deliver near-DFT accuracy at a fraction of the computational cost[8], enabling simulations of larger systems and longer timescales that would otherwise be unfeasible[9,10]. The increasing popularity of MLIPs has also led to a surge in new architectures, with graph neural networks (GNN) emerging as a particularly promising approach[11]. Further advances, such as equivariant GNNs (ensuring the preservation of physical symmetries) and message-passing, have significantly improved the viability of these models[12]. This has made MLIPs a practical alternative to both classical force fields and quantum methods. However, their usefulness is ultimately constrained by the issue of generalization to out-of-distribution samples[13]. This limitation arises from not only the vastness of chemical space but also the conformational diversity of individual molecules. At the same time, there has been a growing interest in developing universal or foundational MLIPs[14–17]. This trend is motivated by the desire to cover broader regions of chemical space with a single model, making generalizability even more important.

Achieving a truly generalizable MLIP would require capturing different types of interactions without overestimating any single one. Typically the total energy of a system is decomposed into E_{SR} (short-range) and E_{LR} (long-range) contributions:

$$E_{\mathrm{total}} = E_{\mathrm{SR}} + E_{\mathrm{LR}}. \tag{1}$$

Due to computational constraints, MLIPs can only access interactions within a finite effective cut-off radius. Thus, they are likely to overestimate short-range interactions to compensate for the missing or incorrectly modeled long-range contributions, especially in strictly local models such as Allegro[12], which only exchange information within a short cut-off[11] (see Figure 1). This would suggest that such models may overfit to the training data,

[a] Multiscale Modeling of Fluid Materials, Department of Engineering Physics and Computation, TUM School of Engineering and Design, Technical University of Munich, Germany; E-mail: julija.zavadlav@tum.de

View Article Online
DOI: 10.1039/D5DD00570A

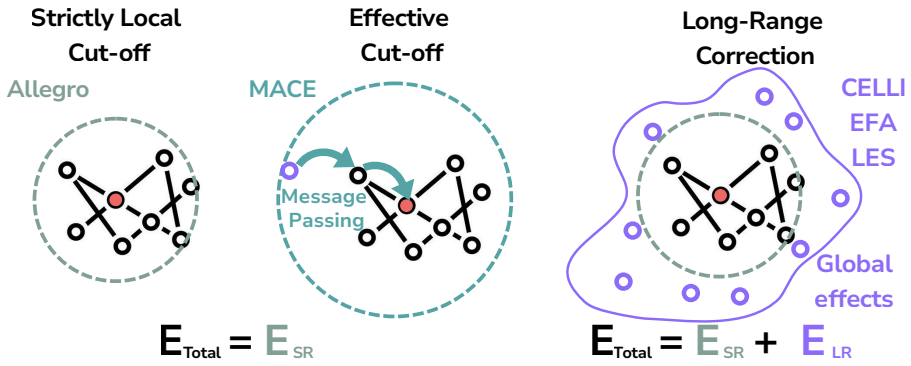

Fig. 1 Difference between strictly local, message passing, and long-range methods in MLIPs. We refer to the cutoff extended by message passing layers as the effective cutoff to differentiate it from the cutoff present in strictly local models.

thereby reducing their generalizability.

The issues with strictly local cutoffs can also be partially offset by message-passing neural networks (MPNNs), which gained significant popularity partially due to their longer effective receptive fields. However, they face challenges related to parallelization and scalability[8], particularly in the context of molecular dynamics simulations of large systems, due to the growth of their receptive fields[12], bottlenecks in per-layer communication[18], and layer-wise synchronization barriers[19]. These limitations have motivated the development of strictly local models, which are well suited for incorporating long-range correction schemes due to their lower computational cost[11]. Additionally, increasing the number of propagation layers beyond a certain point leads to diminishing returns and eventually feature collapse[20]. We hypothesize that separate modeling of short- and long-range energy contributions may improve performance by encouraging more balanced representations of both interaction types. While such separation does not by itself guarantee improved generalization, it may mitigate overfitting arising from compensatory adjustments within the short-range component.

Recently, many different approaches to long-range modeling have been proposed with solutions ranging from charge equilibration schemes[21–23] to charge-independent methods such as self-consistent neural networks[24], reciprocal space transformations[25], Gaussian multipliers[26], attention-based architectures[27], and methods based on Ewald summation[28]. It is worth noting that, despite their proven viability, many foundational models do not incorporate long-range schemes[15,29], which may explain why they often struggle to predict experimental measurements[30].

The challenges with chemical diversity and long-range effect modeling outlined above are particularly pronounced in the context of metal–organic frameworks (MOFs). As their modular architecture allows for precise control over porosity, surface area, and chemical functionality, making them highly tunable for a wide range of applications, including gas storage, catalysis, separations, and sensing[31,32]. Unlike traditional porous materials such as zeolites, MOFs offer exceptional structural diversity; tens of thousands of variants have already been synthesized[33], and computational design opens the door to virtually limitless hypothetical structures[34,35]. This vast chemical and configurational space makes it essentially impossible to identify optimal materials for specific applications through empirical methods alone[36–38]. Computational methods are therefore essential to accelerate MOF discovery and to probe their behavior under working conditions, using methods such as classical MD[39], DFT[40], and grand canonical Monte Carlo[41,42].

However, classical modeling techniques face several limitations in their applicability to MOF modeling, as they face trade-offs between accuracy and computational efficiency, making accurate large-scale MOF simulations unfeasible[38,43]. Therefore, MOFs present an ideal application for MLIPs, as classical force fields often lack the accuracy required to capture the complex interactions and are very difficult to parameterize, while DFT is too computationally expensive to model structures with large unit cells and at longer time scales[9,10,44], thus making MLIPs a perfect candidate for large scale MOF simulations. As a result, a wide range of studies have employed MLIPs to investigate the chemical and mechanical properties of MOFs[9,43,45–48]. However, two issues limit the usefulness of MLIPs. First, they often struggle to generalize to out-of-distribution samples, which confines their applicability to a narrow region of chemical space[13]. Second, long-range effects and electrostatics, which are particularly challenging to model for MLIPs[10], can significantly affect MOF behavior[49,50]. Additionally, foundational models also struggle with modeling MOFs, often failing to outperform simple classical force fields[51], suggesting that large training sets might not be enough to achieve truly universal MLIPs. All of these factors make MOFs an ideal target for our study, as they are highly relevant to experimental and computational chemists, have been previously investigated using MLIPs, and exhibit a vast and complex chemical space, providing an ideal test case for evaluating model generalizability across chemical space.

In this work, we test the generalizability of three widely used baseline architectures: DimeNet++[52], MACE[53], and Allegro[12], on diverse chemical spaces defined by subsets of QMOF[34], ODAC25[54], and OMOL25[55] datasets. We perform a direct comparison of different long-range correction schemes; thus, we test two recently introduced frameworks in combination with different baseline models: the Charge Equilibration Layer for Long-range Interactions (CELLI), recently introduced by the authors[23],

Open Access Article. Published on 13 April 2026. Downloaded on 4/14/2026 6:36:52 PM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

Digital Discovery Accepted Manuscript

View Article Online
DOI: 10.1039/D5DD00570A

Open Access Article. Published on 13 April 2026. Downloaded on 4/14/2026 6:36:52 PM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

and Euclidean Fast Attention (EFA)[27]. We show that such corrections not only allow cheaper models to achieve state-of-the-art accuracy but are also essential for improving generalizability across chemical space, even in MPNNs. Unlike most prior works[56,57], which have mainly focused on conformational generalization, we center our analysis on chemical diversity. Furthermore, we demonstrate that partial charges cannot be inferred without training on reference partial charge labels in the case of challenging datasets such as the ones tested in this work. This is also the case for the Latent Ewald Summation (LES)[28] framework, which recently claimed the opposite[28,58–60]. In these challenging environments, models trained with CELLI based on reference charges consistently produce physically meaningful results, highlighting that leveraging accurate charge information remains critical for developing truly generalizable long-range MLIPs.

Methods

Experiment Design

To evaluate generalizability, we use three datasets: QMOF[34], ODAC25[54], and a metal-complex subsplit of OMOL25[55]. The QMOF dataset contains MOFs with up to 500 atoms per unit cell in their ground state, which is particularly suitable here since our focus is on generalizability across chemical space. By contrast, ODAC25 and OMOL25 include non-ground-state structures. To ensure methodological consistency, maintain acceptable computational cost for training a large number of MLIPs, and focus on chemical space generalization, we constructed subsplits containing only the lowest-energy molecules. Although OMOL25 does not include MOF structures but only metal-organic complexes, we incorporated it due to the lack of alternative MOF datasets. To our knowledge, OMOL25 also lacks ground-state structures, so we selected the lowest-energy conformer for each molecule (see Supporting Information section 2 for details on subsplit construction). Likewise, for ODAC25, we used up to the 10 lowest-energy conformations per molecule, otherwise, the resulting subsets would be too small to train or evaluate models reliably. The resulting subsets comprise 76,525 unique molecules from OMOL25 (up to 350 atoms), 20869 MOFs from ODAC25 (up to 616 atoms), and the full QMOF dataset with 20,407 MOFs (up to 500 atoms).

To evaluate how well different MLIPs generalize to out-of-distribution samples, we consider four evaluation strategies. First, we use a previously introduced method, where the model is trained on a subset of structures with 100 or fewer atoms and then tested on a subset of larger molecules[49]. Since larger molecules are likely to be significantly different from those in the smaller subset. Secondly, we introduce two additional biased train-test split methods: cluster and maximal separation, to systematically investigate differences in model generalization to unseen regions of chemical space. Lastly, a regular random split was used as a comparison to biased split methods.

To create our biased split methods, we employ SOAP (Smooth Overlap of Atomic Positions) descriptors[61], which combine radial and angular information into rotationally invariant atomic features. First, we calculate SOAP descriptors of each atom and average them to create a global similarity measure[62]. Unfortunately, this vector is uninterpretable; however, it allows us to define an architecture-independent descriptor space for molecules, enabling the investigation of MLIP's performance in out-of-distribution samples. The SOAP average similarity kernel has also been previously utilized by Rosen *et al.* on the QMOF dataset to identify local trends in feature space and for band gap predictions, achieving good results[34], and demonstrating SOAP's ability to produce a meaningful representation of the MOF's atomic environment.

The **cluster** method groups molecules into structurally similar clusters using K-Means clustering on their SOAP descriptors. A subset of clusters is then randomly selected for the training set, while the remainder forms the test set. This enforces a structural distinction between splits.

The **maximal separation** method computes pairwise cosine similarities between descriptor vectors $\mathbf{A}$ and $\mathbf{B}$

$$\text{Similarity}(\mathbf{A},\mathbf{B}) = \frac{\mathbf{A}\cdot\mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|}. \tag{2}$$

Starting with a random training data sample, the maximal separation method iteratively adds a candidate to the test dataset. The selected candidate $\hat{\mathbf{A}}$ has minimal similarity to training data samples $\mathbf{B}$

$$\hat{\mathbf{A}} = \arg\min_{\mathbf{A}\in\text{Remaining}} \left(\max_{\mathbf{B}\in\text{Train}} \text{Similarity}(\mathbf{A},\mathbf{B}) \right). \tag{3}$$

The proposed biased split methods evaluate distinct aspects of generalizability: the small–large split assesses performance on cells of varying sizes; the maximal separation method measures performance on a subset maximally different from the training set to stress-test MLIP; and the cluster method examines generalization to distinct structural families by partitioning chemical space into structurally coherent clusters. This addresses a critical limitation of MLIPs: available datasets likely under-sample the vast chemical space, creating biases that may lead to overoptimistic performance evaluation and, as a result, limit their applicability to only its narrow part.

To ensure a sufficiently large test set for meaningful generalizability analysis, we allocated 50% of the OMOL25 and QMOF datasets, and 30% of the ODAC25 dataset, to testing. In all cases, the validation set was drawn from the training data, comprising 10% of its datapoints. To account for random effects introduced by the splitting process, we generated each split three times with different seeds for the ODAC25 and QMOF datasets, and only once for OMOL25 due to higher computational cost. Then, to visualize how different split methods divide given chemical space, we applied Uniform Manifold Approximation and Projection (UMAP)[63], a nonlinear dimensionality reduction technique that preserves both local and global structure, and has been often used to visualize chemical space[64–66]. Figure 2 shows that the proposed biased splitting methods, and especially the maximal separation method, produce splits that are significantly different (at least in the SOAP descriptor space) and can serve as a challenging benchmark for evaluating the generalizability of MLIPs.

Digital Discovery Accepted Manuscript

View Article Online
DOI: 10.1039/D5DD00570A

Open Access Article. Published on 13 April 2026. Downloaded on 4/14/2026 6:36:52 PM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

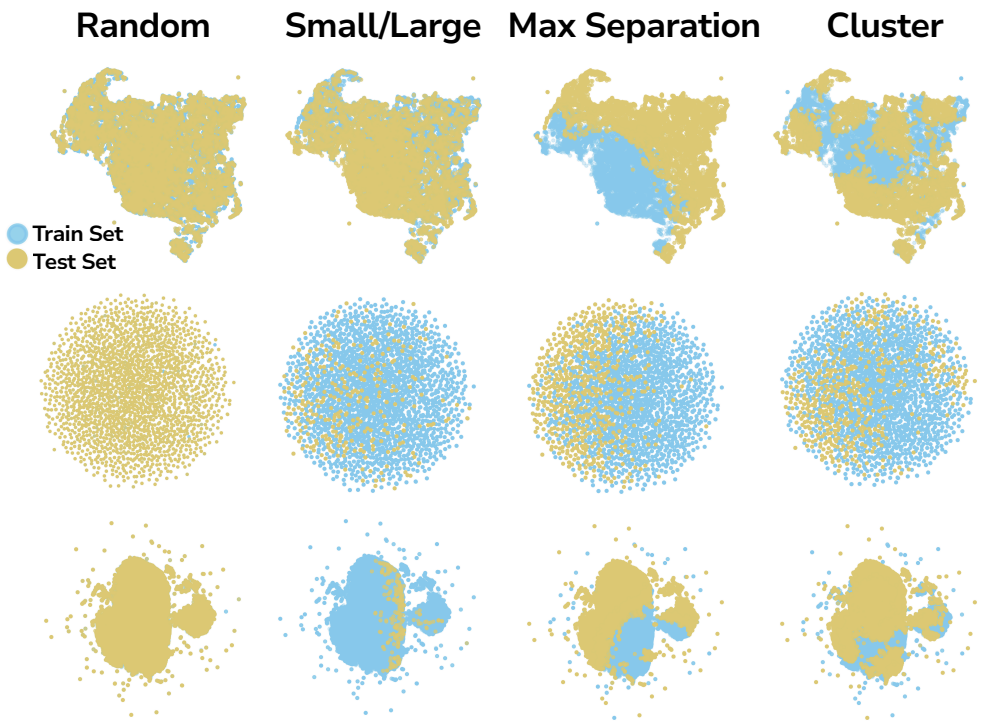

Fig. 2 Overview of the max separation, cluster, size-based splits, and random for QMOF (top), ODAC25 (middle), and OMOL25 (bottom). Each split is visualized on the UMAP dimensionally reduced chemical space. In both cases, the goal is to create a train–test split that introduces significant differences between the subsets to account for the inherent bias in the available chemical space and to establish a stress test–like scenario for evaluating the generalizability of MLIPs. Here we show one split for each dataset, while the remaining splits are reported in the Supplementary Information section 2.

The SOAP descriptors were generated using the Dscribe package[62] and parameters provided by Rosen *et al.*[34]. The maximal separation method is computationally expensive and scales poorly with dataset size, as it requires computing similarity between all members of the training set and remaining molecules. For this reason, in the OMOL25 dataset, a simplified version was used, where similarity was computed only between subsets of 1000 molecules in the training set and the remaining molecules. For the cluster method, the number of clusters was set to 20, and 50% (70% in ODAC25) were used for training.

Architectures and Long-Range Corrections

As baseline models, we consider DimeNet++[52], MACE[53], and Allegro[12], which represent distinct design paradigms: Allegro is an equivariant, strictly local, edge-centric model; MACE is an equivariant, message-passing, node-centric model; and DimeNet++ is an invariant message-passing architecture that incorporates explicit three-body interactions. In order to show that the improvements resulting from the introduction of long-range methods cannot be achieved by simply increasing the effective cutoff via additional message passing steps, we trained an additional baseline model with two additional (four in total) message passing layers, which we refer to as the MACE-MP4 model. MACE and Allegro models also included CELLI[23] and EFA[27] long-range corrections schemes. Although both EFA and CELLI have shown effectiveness at modeling long-range interactions[23,27], they are based on fundamentally different design principles. Specifically, EFA relies on attention mechanisms to learn global representations of chemical systems, whereas CELLI is grounded in physics-based design and dynamically redistributes charge to account for long-range interactions and charge transfer.

These differences make them ideal candidates for this study, as they allow for a comparison of the advantages and disadvantages of physics-inspired versus purely AI-driven approaches. In principle, physics-based models are expected to generalize better due to their grounding in physical theory[67]. However, imposing such constraints may also limit the expressiveness of the model[68]. On the other hand, data-driven methods often offer greater flexibility but are prone to overfitting and may generalize poorly to out-of-distribution samples[13]. In addition, when fitted to partial charges, CELLI can only model electrostatic interactions and charge transfer, whereas EFA can represent all long-range effects, including van der Waals interactions. To ensure comparability, the same hyperparameters were used for each architecture across different splits.

CELLI

The Charge Equilibration Layer for Long-range Interactions (CELLI) is a model-agnostic component that enables the integration of non-local electrostatics into local equivariant GNN architectures. CELLI follows the charge equilibration (Qeq) framework, computing partial charges based on environment-

Digital Discovery Accepted Manuscript

View Article Online
DOI: 10.1039/D5DD00570A

Open Access Article. Published on 13 April 2026. Downloaded on 4/14/2026 6:36:52 PM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

dependent electronegativities, species-dependent hardnesses, and charge radius[23]. Electronegativity values are predicted per node by aggregating local scalar edge features using an MLP-based mechanism. Hardness and atomic radius values are derived from species embeddings, with radii initialized from covalent radii and scaled by a learned, positive factor.

Those parameters are then used to redistribute charge based on the following energy minimization problem. The total Qeq energy is defined as

$$\alpha_{ij} = \frac{1}{\sqrt{2}}(\gamma_i^2 + \gamma_j^2)^{-1/2} \tag{4}$$

$$U_{\text{Coul}}(\boldsymbol{R}, \boldsymbol{Q}) = \sum_i^N \sum_{j>i}^N \frac{\text{erf}(\alpha_{ij} r_{ij})}{r_{ij}} Q_i Q_j + \sum_{i=1}^N \frac{2\alpha_{ii}}{\sqrt{\pi}} Q_i^2 \tag{5}$$

$$U_{\text{Qeq}}(\boldsymbol{R}, \boldsymbol{Q}) = U_{\text{Coul}}(\boldsymbol{R}, \boldsymbol{Q}) + \sum_{i=1}^N \left[\chi_i Q_i + \frac{J_{ii}}{2} Q_i^2 \right] \tag{6}$$

where the final term captures the atomic contributions through previously computed electronegativities χ_i, atomic radii γ_i, and chemical hardnesses J_{ii}[69]. The minimum of U_{Qeq} is the solution of the linear system

$$\left[\frac{\partial^2 U_{\text{Coul}}}{\partial Q_i \partial Q_j} \middle| \boldsymbol{R} + Jii \right] Q_j = -\chi_i \tag{7}$$

under the charge conservation constraint.

Once charges are obtained, they are embedded into the GNN by generating charge-dependent feature vectors via a second MLP, which are combined with existing latent scalar features through a residual update. This allows downstream layers to incorporate non-local information while maintaining the local structure of the network. Although CELLI is model agnostic and can be utilized with multiple frameworks, it's best suited for strictly local models such as Allegro[12], as the benefits of long-range scheme correction are larger compared to message passing architectures. However, here we also consider an MACE implementation. Regardless of the MLIP architecture used, the long-range electrostatic interactions could be partially accounted for by the long-range contribution and partially by the short-range contribution to the total energy. However, when CELLI is trained on partial charges, electrostatic interactions are assumed to be entirely accounted for by the long-range contribution. In other words, the electrostatic interactions are not double-counted.

EFA

Euclidean Fast Attention (EFA) encodes atomic spatial information into feature representations by combining distance-aware modulation with rotational invariance through integration over the unit sphere. EFA builds on Euclidean Rotary Positional Encoding (ERoPE)[70], which maps 3D positions into complex-valued feature vectors using frequency-based phase shifts[27].

Specifically, given a position vector $\mathbf{r} \in \mathbb{R}^3$, a feature vector $\mathbf{x}$, a frequency $\omega \in \mathbb{R}$, and a unit vector $\mathbf{u} \in S^2$, ERoPE modulates the feature as

$$\phi_{\mathbf{u}}(\mathbf{x}, \mathbf{r}) = \mathbf{x} \cdot e^{i\omega \mathbf{u} \cdot \mathbf{r}}, \tag{8}$$

where the dot product $\mathbf{u} \cdot \mathbf{r}$ projects the position onto the direction $\mathbf{u}$. To ensure rotational invariance, EFA averages over all directions $\mathbf{u}$ on the unit sphere S^2, resulting in the attention mechanism

$$\text{EFA}(\mathbf{X}, \mathbf{R})_m = \frac{1}{4\pi} \int_{S^2} \phi_{\mathbf{u}}(\mathbf{q}_m, \mathbf{r}_m)^\top \sum_{n=1}^N \phi_{\mathbf{u}}(\mathbf{k}_n, \mathbf{r}_n) \mathbf{v}_n^\top \, d\mathbf{u}, \tag{9}$$

where $\mathbf{q}_m$, $\mathbf{k}_n$, and $\mathbf{v}_n$ are the query, key, and value vectors for atoms m and n, respectively, and $\mathbf{r}_m$, $\mathbf{r}_n$ are their 3D coordinates.

By integrating over all directions on the unit sphere, the resulting attention weights depend only on interatomic distances, ensuring rotational invariance. To support directional reasoning, EFA incorporates a tensor product with spherical harmonics, allowing the mechanism to operate on equivariant inputs. This enables the network to model both long-range interactions and geometric relationships without requiring explicit neighbor cutoffs, making it well-suited for large atomistic systems.

Originally, the EFA scheme was proposed for a generic MPNN architecture, which consisted only of node features and a message passing step, which strongly limits its applicability. Thus, there is a need to adapt this scheme to popular frameworks. Unlike in the network utilized by Frank *et al.*, there are several ways in which EFA can be integrated into MACE or Allegro. However, the detailed analysis of the optimal EFA implementation is outside of the scope of this study, and in this section, we provide a basic description of our integration.

The EFA block takes as input the current node embeddings and atomic positions. To integrate it with MACE, the output is then fused with the node features before being passed into subsequent MACE interaction layers. By inserting EFA at configurable points in the message passing stack, the architecture captures long-range geometric dependencies prior to or between equivariant tensor-product updates. This design enables the network to blend fast, global spatial awareness with deep local equivariant reasoning, allowing for flexible integration of EFA alongside or in place of conventional MACE message passing without disrupting equivariance or scalability.

In the Allegro integration, EFA features are combined with the original species embeddings and linearly projected into a refined feature space just before the Allegro tensor product layer (initial invariant scalar latent features in the first MLP only utilize original species embeddings). This insertion point ensures that EFA-enhanced node-level information flows into the pairwise feature prediction pipeline of Allegro. EFA was initially implemented for MPNNs, however, message passing is not fully required and can be combined with a strictly-local Allegro architecture.

LES

Unlike charge-equilibration approaches, LES does not solve a redistribution problem. Instead, LES learns latent electrostatic charges directly from local atomic descriptors and uses them to compute a long-range energy contribution through an Ewald formulation[28]. These charges are not constrained to match reference partial charges and are optimized solely through the global energy and force loss terms. As a result, they do not conserve

Digital Discovery Accepted Manuscript

View Article Online
DOI: 10.1039/D5DD00570A

Open Access Article. Published on 13 April 2026. Downloaded on 4/14/2026 6:36:52 PM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

the total charge. For periodic systems, the reciprocal-space Ewald term is used:

$$U_{\mathrm{lr}}(\mathbf{R},\mathbf{q}) = \frac{1}{2\varepsilon_0 V} \sum_{0<|\mathbf{k}|<k_c} \frac{e^{-\sigma^2|\mathbf{k}|^2/2}}{|\mathbf{k}|^2} |S(\mathbf{k})|^2, \tag{10}$$

$$S(\mathbf{k}) = \sum_{i=1}^{N} q_i e^{i\mathbf{k}\cdot\mathbf{r}_i}, \tag{11}$$

where V is the simulation cell volume and σ is a Gaussian smearing parameter.

For isolated systems, LES uses a screened direct Coulomb sum:

$$U_{\mathrm{lr}}(\mathbf{R},\mathbf{q}) = \frac{1}{8\pi\varepsilon_0} \sum_{i=1}^{N} \sum_{j>i}^{N} \frac{q_i q_j}{r_{ij}} \left[1 - \mathrm{erf}\left(\frac{r_{ij}}{\sqrt{2}\sigma}\right)\right]. \tag{12}$$

LES also enables computation of Born effective charge tensors, which can be used for simulations of systems under an electric field[59] (for details, see Supplementary Information section 9). After computing the long-range energy, LES includes electrostatic information into the model only through its total energy contribution. No additional solvers, constraints, or charge-conservation steps are utilized. In this work, we utilized the MACE[53,71] integrations of LES due to the high computational cost of CACE (for details on reproducibility of previous results, see Supplementary Information section 8).

Total Charge Embeddings

First, we extend the standard MACE node-embedding layer to incorporate a global conditioning on total charge alongside atomic species. In addition to the species embedding, a separate embedding is learned and projected into the same scalar feature space and broadcast across all nodes. The species and charge embeddings are fused and refined through an MLP. This modification injects global charge information into every node before message passing, enhancing expressivity without affecting the symmetry guarantees of the core MACE architecture.

For Allegro, we combine learned embedding of total charge with the species embeddings and linearly project into a refined feature space just before the Allegro tensor product layer (similarly to the EFA integration initial invariant scalar latent features in the first MLP, only utilize original species embeddings). Although Allegro is strictly local, the global charge embedding provides a simple mechanism to modulate all node features consistently based on system-level charge, enhancing expressivity without modifying the underlying equivariant structure.

Model Training

All baseline, CELLI and EFA models were trained in JAX using chemtrain[72], adapted JAX-MD[73] packages together with JAX-compatible implementations of Allegro[74], DimeNet++[52,75], and MACE[53,76] via the Force Matching method[72,77]. For the OMOL25 and QMOF datasets, models with CELLI were first pretrained using only charges, and subsequently trained on charges, forces, and energies. Since the ODAC25 dataset does not include force information, it was excluded from this training procedure. For details on training, data preprocessing, and hyperparameters, see Supplementary Information (sections 3-5). We restrict our evaluation to molecules containing only species that appear at least 10 times in the training set, as it is unrealistic to expect accurate modeling of rare elements. Species that occurred only in the test set were also discarded. Models with LES were trained using MACE[53,78] and LES[79] PyTorch implementation.

Results

Increased Generalization of Long-Range Models

Our results show that the integration of long-range correction schemes has a significant effect on both MACE and Allegro and is necessary for the latter to achieve SOTA performance on the QMOF dataset (see Figure 3). For Allegro, this improvement in Root Mean Squared Error (RMSE) is noticeable even in the simplest benchmark (random split), which aligns with our expectations, as Allegro is strictly local and therefore struggles to model systems where long-range interactions can be significant. However, incorporating EFA into Allegro yields a smaller performance gain compared to CELLI, especially for the cluster and maximal separation splits, suggesting that EFA does not generalize as well to out-of-distribution samples as CELLI. This difference may be influenced by two factors: (1) the inherent generalizability of physics-based models, and (2) the large chemical diversity of the QMOF dataset relative to its size (only 20,387 samples), as EFA may outperform CELLI in larger datasets.

For MACE trained on the random split, the effect of adding long-range schemes is minimal and comparable to statistical noise; however, in more challenging test cases (maximum separation split), the improvement becomes evident even for MACE, as it fails to generalize to out-of-distribution test sets. Although CELLI and EFA models also struggle with biased splits, they perform significantly better than baseline models, suggesting that long-range corrections are crucial for achieving robust generalization. Our results also show that increasing the number of message passing steps can lead to overfitting and a decrease in performance compared to the baseline model. For example, in the cluster and maximum separation splits, an increase in the number of message passing steps more than doubles the RMSE, showing that the addition of separate long-range corrections is advantageous.

Interestingly, DimeNet++, which obtains reasonable results in the random, size, and maximum separation splits, fails to generalize under cluster split. This shows that the proposed benchmarks probe different aspects of generalizability and can be used in future studies focusing on the development of long-range correction schemes. Our results also support our earlier hypothesis on the necessity of inclusion of a dedicated mechanism for modeling long-range interactions. The issues with generalizability are particularly significant for MACE, as foundational and universal models have been developed using this architecture[14,16]. However, our results show that MACE without any long-range correction does not generalize well to out-of-distribution samples, suggesting it may not be suitable for this task.

Digital Discovery Accepted Manuscript

View Article Online
DOI: 10.1039/D5DD00570A

Open Access Article. Published on 13 April 2026. Downloaded on 4/14/2026 6:36:52 PM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

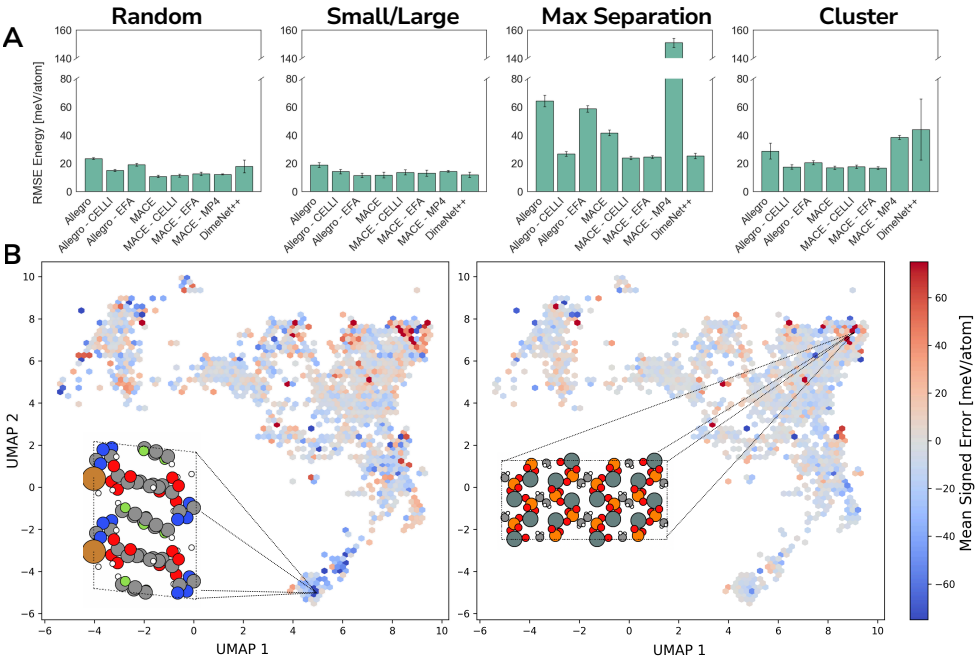

Fig. 3 Long-Range schemes increase generalizability of MLIPs. A Average model performance across maximum separation, cluster, size, and random data splits on the QMOF dataset measured by Root Mean Squared Error (RMSE). Each model was trained three times with different initializations. The QMOF dataset provides only energies and no force labels. B Average error on the QMOF test space of the max separation split for (short-range) Allegro (left) and (long-range) Allegro-CELLI (right). The highlighted molecules represent areas of chemical space with similar biases. Detailed results and plots for all models are available in the Supporting Information section 7.

Modeling Charged Systems

The biased dataset splits introduced in this study also allow us to visually investigate the generalizability of MLIPs across chemical space. To illustrate this, we visualized the prediction error of each model on the test portion of the max separation split of the QMOF dataset (see Figure 3 **B**). Interestingly, although the Allegro–CELLI model achieves significantly better overall accuracy, both the baseline Allegro and Allegro–CELLI exhibit similar regions of bias across chemical space. The magnitude of the error is substantially larger for the baseline Allegro model, and only a few regions show qualitatively different types of error (i.e., overestimation versus underestimation). This observation suggests that certain areas of chemical space are inherently challenging for both models. Possible explanations include intrinsic difficulty in modeling complex structures, inaccuracies within the QMOF dataset, or undersampling of specific chemical motifs. Nevertheless, the model incorporating long-range corrections performs uniformly better, highlighting that such corrections are essential for improving the generalizability of MLIPs. It should also be noted that near-zero average error does not necessarily mean that the error is smaller, only that it does not exhibit bias.

In our next benchmark, we utilized the OMOL25 dataset to evaluate the performance of different schemes on charged molecules. First, we trained MACE and Allegro on the neutral molecules of our OMOL25 subset, with and without long-range schemes. While performance on the neutral subset is not strongly affected by CELLI, the baseline versions of both MACE and Allegro are degenerate with respect to total charge, meaning they cannot distinguish between molecules with different charge states and assign identical energies to species that differ only in charge. This limitation becomes critical once multiple charge states are present (see Figure 4); therefore, we limited our subsequent analysis to the charge-dependent models.

To enable a fair comparison for CELLI, we added total charge embeddings to baseline Allegro and MACE (for details, see Methods). Interestingly, models with total charge embeddings perform slightly better compared to CELLI in the majority of cases. This demonstrates that CELLI can effectively act as a total charge embedding scheme, while also offering two key practical advantages over explicit embeddings: 1) it provides long-range capabilities, which are likely to be relevant in simulations of larger systems, and 2) although accounting for nonzero total charges is necessary for training, MD simulations are typically conducted under neutral conditions, meaning those embeddings would not be used outside of training. Additionally, total charge embeddings can also be combined with all long-range methods, including CELLI; hence, they should not be viewed as a replacement, but rather as a potential extension. The integration of these embeddings into long-range schemes is, however, beyond the scope of this study.

Inferring Charges from Forces and Energy

In contrast to the other two datasets, ODAC25 does not provide reference partial charges, meaning that CELLI has no target charge values to fit. Although it is technically possible to use CELLI without reference charges, there is no guarantee that the inferred charges would accurately represent the underlying electrostatics of the system. Importantly, in this dataset, EFA and CELLI do not lead to improvements in either forces or energies

Digital Discovery Accepted Manuscript

View Article Online
DOI: 10.1039/D5DD00570A

Open Access Article. Published on 13 April 2026. Downloaded on 4/14/2026 6:36:52 PM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

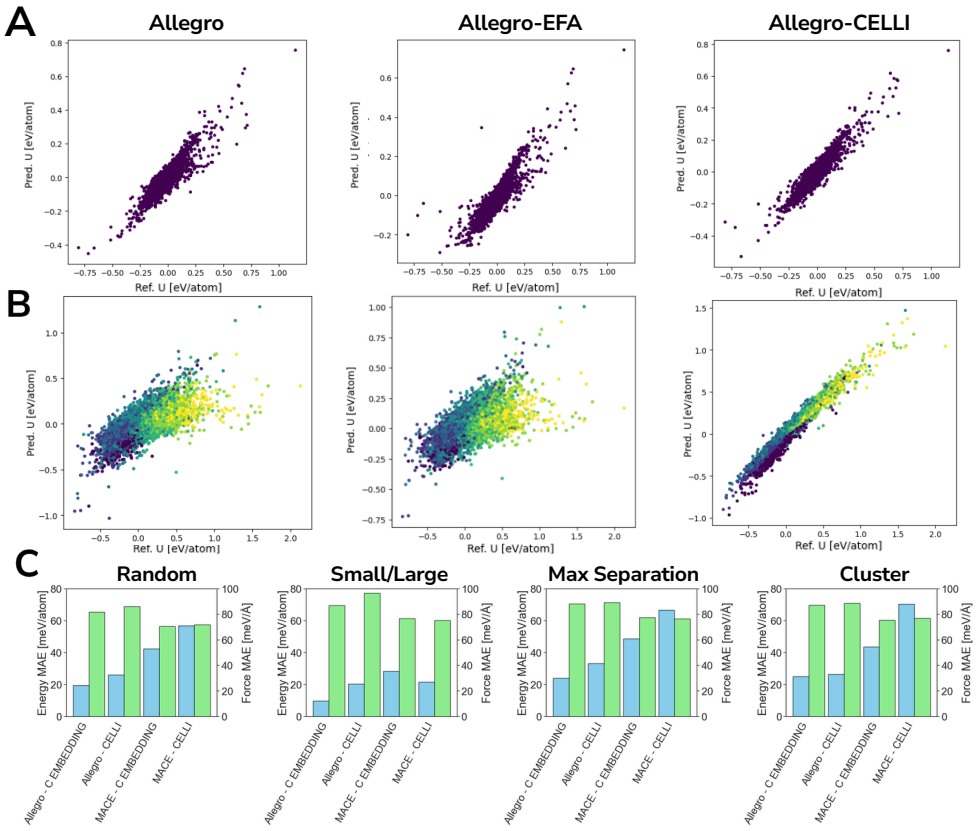

Fig. 4 Accuracy on Charged vs. Neutral Systems Parity plots for three Allegro models with and without long-range corrections trained on only neutral (A) and both neutral and charged molecules (B) from the OMOL25 dataset. Colors correspond to the total charge. Baseline Allegro and Allegro-EFA models exhibit a significant increase in error when trained on charged samples. C Model performance across maximal separation, cluster, size, and random data splits on the OMOL25 dataset measured by Mean Absolute Error (MAE). Blue bars correspond to energy MAE, while green bars correspond to force MAE. "C - EMBEDDING" corresponds to models with an additional embedding for total charge; see Methods for details.

(see Figure 5). A pronounced increase in error is observed for MACE-based models in the size split, where their performance is significantly poorer than in the other benchmarks. A closer inspection reveals that CELLI-based models were unable to infer meaningful charges in the absence of suitable reference data and effectively predicted zero charge for the majority of atoms (see Figure 6 **A**), even on a random split. Consequently, unlike in earlier benchmarks, Allegro with EFA outperforms CELLI across most benchmarking subsplits. Therefore, we recommend using CELLI only when reference charges are available, at least for complex systems such as MOFs.

The issue of relying on partial charges to model electrostatic interactions is well known[68], as charge-based schemes require predefined charges and their accuracy depends strongly on the chosen charge partitioning method, which often yields significantly different results[80]. To address this, several approaches have been proposed to bypass the need for reference charges by inferring them directly from forces and energies, such as the LES framework[28,58,59]. King *et al.* demonstrated that LES can successfully infer charges for small systems (e.g., polar dipeptides from the SPICE dataset[59,81]); however, to our knowledge, it has never been tested on MOFs, which exhibit significantly more complex electrostatic environments than small biomolecules.

The MACE-LES model trained on the random split achieved MAE of 8.7 meV/atom and a force MAE of 24.6 meV/Å, meaning that the force error is around 50% higher than that of the baseline MACE model, whereas energy error is 50% lower (we also tested whether we can reproduce results reported before by King *et al.* to ensure that this is due to performance of the MACE framework rather than technical issues on our side for details see Supporting Information section 8). Further investigation shows that LES is also unable to infer correct charges without reference, and similarly to CELLI, it predicts most partial charges to be almost zero. To further validate this result, we trained an additional MACE-LES model on the QMOF dataset, which includes DFT-derived partial charges and allows direct comparison to reference values. As shown in Figure 6, the same behavior was observed: LES again failed to infer meaningful charges and collapsed to predicting near-zero values. Although LES frequently assigns charges opposite in sign to the reference values, this is not inherently problematic-Coulomb's law is symmetric under global charge inversion, meaning forces and energies would remain unchanged if all charges were flipped. What is problematic is the lack of consistency, as LES sometimes predicts correct signs and sometimes their opposites. Based on this, we attribute better performance on energies of MACE-LES compared to baseline MACE are likely

Digital Discovery Accepted Manuscript

View Article Online
DOI: 10.1039/D5DD00570A

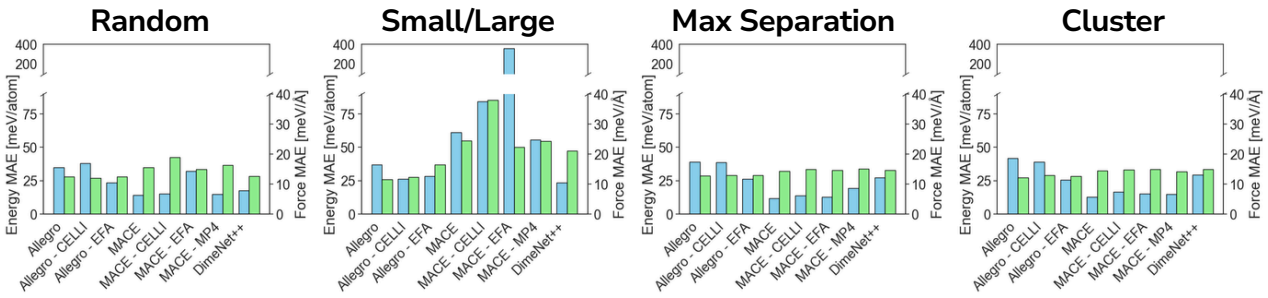

Fig. 5 Long-range modeling without reference charges. Model performance across maximal separation, cluster, size, and random data splits on the ODAC25 dataset. Blue bars correspond to energy MAE, while green bars correspond to force MAE.

due to discrepancies in MACE implementations rather than properties of LES. Namely, baseline MACE was trained in JAX[76] and MACE-LES in PyTorch[78]. However, given that the ODAC25 subsets used in this study are relatively small, we cannot rule out the possibility that CELLI or LES may recover correct charge distributions when trained on sufficiently large datasets. Nevertheless, these results indicate that when no partial charges are available, other approaches are necessary, as none of the schemes achieved a significant improvement, except for the Allegro–EFA model. We would also like to highlight that in this work, accurate partial charge prediction is not treated as a goal in itself, but rather as a diagnostic indicating whether a model is able to infer meaningful long-range electrostatics, with the near-zero charges predicted by LES and CELLI in the absence of reference data signaling a failure to capture such effects in MOFs.

Conclusions

Overall, our results establish a clear framework for evaluating long-range MLIPs on complex, chemically diverse systems and demonstrate that physically grounded treatments of electrostatics are essential for robust generalization. The presented biased split methods provide a robust test for generalizability to out-of-distribution samples, which is necessary for practical MLIP applications. They are also more broadly applicable than alternative functional group–based biased splits (for example, training on all structures except aliphatic hydrocarbons and testing on those), as they do not rely on predefined chemical heuristics. The proposed methods could capture differences arising from coordination environments, pore geometries, or topology features that are especially important in the context of MOFs, where structural diversity extends far beyond simple chemical fragments. Thus, they provide a more flexible and less assumption-driven assessment of MLIP generalizability and can be easily applied to other datasets with minimal modifications.

Nevertheless, the reliance on SOAP descriptors may partly explain the limited decrease in performance observed for the ODAC25 subset, as this representation likely restricts the effectiveness of our splitting strategy. SOAP may fail to provide a sufficiently meaningful representation of certain atomic environments, resulting in a biased split that, in practice, is not very different from a random one. Additionally, SOAP descriptors are inherently short-ranged, and therefore may not fully distinguish structures that differ primarily in long-range electrostatics, potentially limiting how strictly these splits probe out-of-distribution behavior with respect to long-range physics. Nevertheless, if long-range interactions are not properly captured, models may compensate by overfitting short-range contributions, which can degrade short-range generalizability across chemically diverse environments. Future work could explore alternative representations, such as learned embeddings, electronic descriptors[83], or graph similarity measures[84], to construct more reliable biased splits. However, results from the QMOF dataset suggest that this limitation is not universally detrimental, and that SOAP-based splits can still expose meaningful performance differences when applied to sufficiently diverse datasets.

The performed benchmark tests clearly show that the incorporation of long-range schemes is necessary to achieve accurate and generalizable MLIPs. Especially in the QMOF benchmark, both Allegro and MACE benefited from the introduction of either EFA or CELLI, and only models based on CELLI exhibited good results in all three biased split benchmarks. We also show that the same cannot be achieved by increasing message passing layers. Furthermore, all variants share identical hyperparameters and nearly identical parameter counts, hence the observed improvements cannot be attributed to increased model capacity and instead arise from explicit long-range corrections. Although total charge embeddings can recover much of CELLI's performance on the OMOL25 dataset, they lack true long-range capabilities and generalize poorly in size-based extrapolation, indicating that embeddings alone cannot substitute for physically grounded treatments. In contrast, charge-based methods such as CELLI are inherently suited to modeling charged systems without relying on auxiliary embeddings, making them more robust and transferable, while embeddings should be viewed as complementary enhancements rather than replacements.

Interestingly, our results reveal some discrepancies with earlier studies that introduced several of the long-range methods evaluated here. For instance, Fuchs *et al.* reported that adding CELLI to MACE does not substantially improve performance over the baseline model, whereas in our out-of-distribution test cases, we observe clear gains. Similarly, our results indicate that LES is not able to infer charges from forces and energies alone, in contrast to conclusions drawn from smaller and less complex systems[59]. These differences can likely be attributed to two main

Open Access Article. Published on 13 April 2026. Downloaded on 4/14/2026 6:36:52 PM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

Digital Discovery Accepted Manuscript

View Article Online
DOI: 10.1039/D5DD00570A

Open Access Article. Published on 13 April 2026. Downloaded on 4/14/2026 6:36:52 PM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

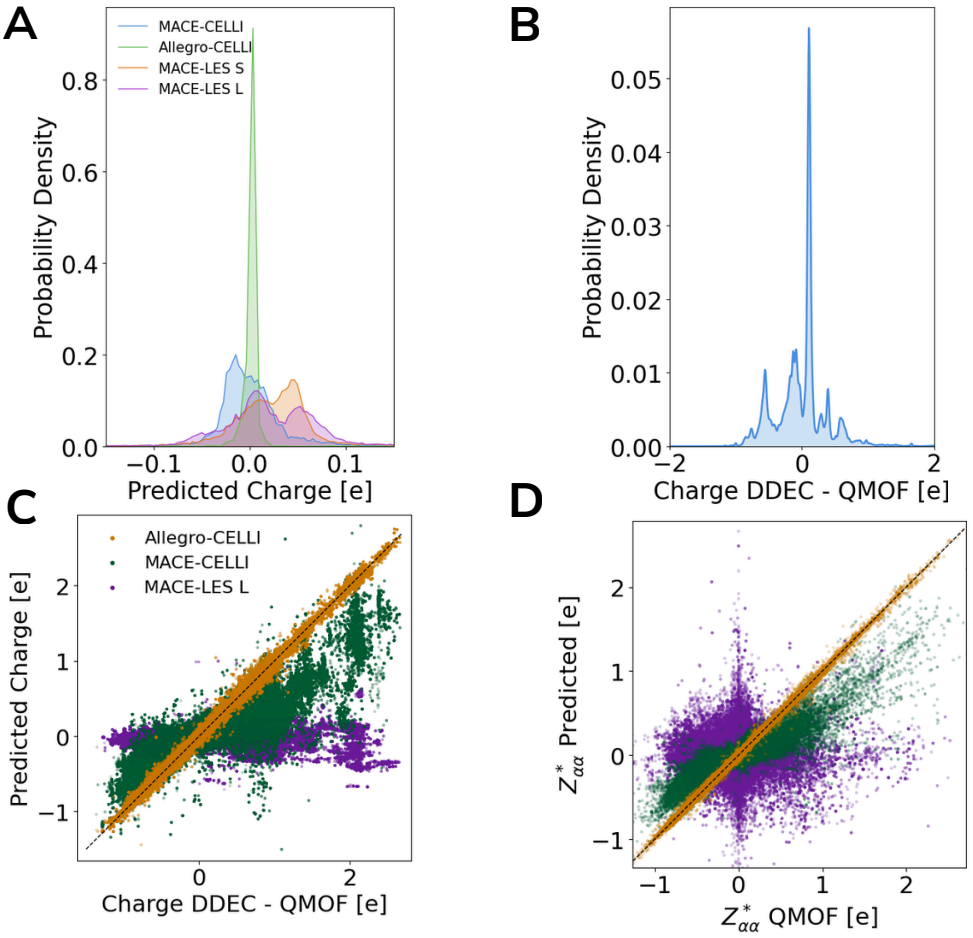

Fig. 6 Inferring Charge distribution in MOFs. Probability density of charges inferred by different models on the ODAC25 dataset (A) and reference DDEC6 charges from the QMOF dataset (B). MACE-LES L and MACE-LES S differ in hyperparameters, with MACE-LES L being more robust. Probability densities were obtained using Gaussian kernel density estimation in SciPy[82]. Parity plot of charges (C) and born effective charges (D) for models trained on a random split of the QMOF dataset (LES was excluded from this plot for clarity and moved to Supporting Information). As in ODAC25, LES was not able to recover the charge distribution. Despite MACE-CELLI being fitted to reference charges, a sufficiently low weight for the charge in the loss function allowed it to deviate from the reference charges. For details and additional plots, see Supporting Information section 9.

factors: our use of out-of-distribution splits and the fact that we apply these models to more challenging systems than those commonly considered in similar studies. In addition, LES relies on a fixed smearing parameter[85], which may further restrict transferability across chemically diverse environments[86]. This also highlights a broader issue in the development of long-range methods: they are often benchmarked on datasets containing relatively small molecules or systems in which long-range interactions are weak or negligible. Our results indicate that such evaluations can be misleading. Long-range schemes should instead be tested on more challenging systems, such as MOFs, where the electrostatic environment is highly complex and long-range contributions play an important role.

We should also note that long-range interactions can extend beyond pairwise Coulomb terms due to polarization and higher-order multipole effects[26]. Those effects are not explicitly modeled in the present CELLI and LES implementations. However, they can, in principle, capture polarization through charge prediction at each timestep. Future work could explore replacing QEq with polarizable charge equilibration schemes such as PQEq[85,87,88], or employing more expressive long-range kernels, including sum-of-Gaussians approaches (SOG-Net)[26], to better capture these contributions. In addition, EFA may also capture higher-order and polarization effects. While detailed inference-time benchmarking is beyond the scope of this work, existing studies indicate that the considered long-range approaches do not significantly increase evaluation time[23,27], further suggesting that they can be routinely integrated into MLIPs.

Our results also provide a cautionary perspective on the limits of inferring charges solely from energy and force. While CELLI delivers substantial improvements when reliable reference charges are available, its performance collapses in their absence, as it fails to recover meaningful electrostatic information and degrades overall accuracy. LES, which was designed with charge inference in mind, exhibits a similar breakdown in MOFs, as it consistently converges to near-zero or inconsistently signed charges across both ODAC25 and QMOF, suggesting that charge inference becomes unreliable once the electrostatic environment becomes

Digital Discovery Accepted Manuscript

View Article Online
DOI: 10.1039/D5DD00570A

Open Access Article. Published on 13 April 2026. Downloaded on 4/14/2026 6:36:52 PM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

too complex. Collectively, these findings suggest that charge-inference schemes, while appealing, are not yet robust enough for systems with complex long-range physics such as MOFs. Although none of the tested models were able to correctly infer charges without reference data, it is possible that more robust architectures or optimized hyperparameter choices could achieve this; however, a systematic investigation of such configurations is beyond the scope of this study. A potential solution to this issue may involve pretraining on reference charges followed by training solely on forces and energies, resulting in a scheme at least partially decoupled from the chosen charge partitioning method. Other promising directions could include modifying the loss function, introducing additional constraints into the learning process, predicting the whole electron density, or using methods that skip charges altogether, like EFA. Ultimately, the most appropriate long-range strategy depends on whether access to accurate partial charges is beneficial. CELLI is preferred when reliable electrostatics and charge-transfer are important and difficult to infer directly from data, whereas charge-free approaches such as EFA can be used for systems where explicit electrostatic information is not needed.

Author contributions

M.S. performed the experiments, analyzed the results, and wrote the manuscript. J.Z. supervised the project, reviewed the manuscript, provided resources, and acquired funding.

Conflicts of interest

The authors declare no financial or non-financial conflicts of interest. The authors have contributed to the development of the CELLI method used in this study.

Data Availability

The datasets used in this study are publicly available to download. Our experiments utilized QMOF v14[34] (DOI: https://doi.org/10.6084/m9.figshare.13147324.v14), OMOL25[55] (DOI: https://doi.org/10.48550/arXiv.2505.08762, accessed June 20, 2025), and ODAC25[54] (DOI: https://doi.org/10.48550/arXiv.2508.03162, accessed September 3, 2025). Details on the creation of utilized subsets of ODAC25 and OMOL25 datasets can be found in the Supplementary Information section 1.

Code Availability

The software `chemtrain`[72] used to train MLIPs is publicly available at `https://github.com/tummfm/chemtrain` (DOI: 10.5281/zenodo.15438477, version 0.1.0). The CACE and MACE models[71] with LES were trained using code available at `https://github.com/ChengUCB/les/tree/main`, `https://github.com/BingqingCheng/cace`, and `https://github.com/ChengUCB/mace` (accessed 09.2025). All code utilized in this work, including requirements, is available at `https://github.com/msan9908/MLIP_generalizability` (DOI: 10.5281/zenodo.19486082).

Acknowledgements

Funded by the European Union. Views and opinions expressed are, however, those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them. This work was funded by the ERC (StG SupraModel) - 101077842.

References

1 E. O. Pyzer-Knapp and T. Laino, in *Machine Learning in Chemistry: Data-Driven Algorithms, Learning Systems, and Predictions*, American Chemical Society, 2019, vol. 1326, pp. ix–x.
2 L. C. Gallegos, G. Luchini, P. C. St John, S. Kim and R. S. Paton, *Acc Chem Res*, 2021, **54**, 827–836.
3 N. Segal, A. Netanyahu, K. P. Greenman, P. Agrawal and R. Gómez-Bombarelli, *npj Computational Materials*, 2025, **11**, 345.
4 R. S. Bohacek, C. McMartin and W. C. Guida, *Medicinal Research Reviews*, 1996, **16**, 3–50.
5 A. Lavecchia, *Drug Discovery Today*, 2024, **29**, 104133.
6 I. O. Betinol, J. Lai, S. Thakur and J. P. Reid, *J. Am. Chem. Soc.*, 2023, **145**, 12870–12883.
7 P. Schwaller and T. Laino, in *Data-Driven Learning Systems for Chemical Reaction Prediction: An Analysis of Recent Approaches*, American Chemical Society, 2019, vol. 1326, pp. 61–79.
8 D. M. Anstine and O. Isayev, *The Journal of Physical Chemistry A*, 2023, **127**, 2417–2431.
9 M. Eckhoff and J. Behler, *J. Chem. Theory Comput.*, 2019, **15**, 3793–3809.
10 O. T. Unke, S. Chmiela, H. E. Sauceda, M. Gastegger, I. Poltavsky, K. T. Schütt, A. Tkatchenko and K.-R. Müller, *Chemical Reviews*, 2021, **121**, 10142–10186.
11 S. Batzner, A. Musaelian, L. Sun, M. Geiger, J. P. Mailoa, M. Kornbluth, N. Molinari, T. E. Smidt and B. Kozinsky, *Nature Communications*, 2022, **13**, 2453.
12 A. Musaelian, S. Batzner, A. Johansson, L. Sun, C. J. Owen, M. Kornbluth and B. Kozinsky, *Nature Communications*, 2023, **14**, 579.
13 T. Cui, C. Tang, D. Zhou, Y. Li, X. Gong, W. Ouyang, M. Su and S. Zhang, *Nature Communications*, 2025, **16**, 1891.
14 I. Batatia, P. Benner, Y. Chiang, A. M. Elena, D. P. Kovács, J. Riebesell, X. R. Advincula, M. Asta, M. Avaylon, W. J. Baldwin, F. Berger, N. Bernstein, A. Bhowmik, S. M. Blau, V. Cărare, J. P. Darby, S. De, F. D. Pia, V. L. Deringer, R. Elijošius, Z. El-Machachi, F. Falcioni, E. Fako, A. C. Ferrari, A. Genreith-Schriever, J. George, R. E. A. Goodall, C. P. Grey, P. Grigorev, S. Han, W. Handley, H. H. Heenen, K. Hermansson, C. Holm, J. Jaafar, S. Hofmann, K. S. Jakob, H. Jung, V. Kapil, A. D. Kaplan, N. Karimitari, J. R. Kermode, N. Kroupa, J. Kullgren, M. C. Kuner, D. Kuryla, G. Liepuoniute, J. T. Margraf, I.-B. Magdău, A. Michaelides, J. H. Moore, A. A. Naik, S. P. Niblett, S. W. Norwood, N. O'Neill, C. Ortner, K. A. Persson, K. Reuter, A. S. Rosen, L. L. Schaaf, C. Schran,

Digital Discovery Accepted Manuscript

View Article Online
DOI: 10.1039/D5DD00570A

Open Access Article. Published on 13 April 2026. Downloaded on 4/14/2026 6:36:52 PM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

B. X. Shi, E. Sivonxay, T. K. Stenczel, V. Svahn, C. Sutton, T. D. Swinburne, J. Tilly, C. van der Oord, E. Varga-Umbrich, T. Vegge, M. Vondrák, Y. Wang, W. C. Witt, F. Zills and G. Csányi, *A foundation model for atomistic materials chemistry*, 2024, https://arxiv.org/abs/2401.00096.

15 B. M. Wood, M. Dzamba, X. Fu, M. Gao, M. Shuaibi, L. Barroso-Luque, K. Abdelmaqsoud, V. Gharakhanyan, J. R. Kitchin, D. S. Levine, K. Michel, A. Sriram, T. Cohen, A. Das, A. Rizvi, S. J. Sahoo, Z. W. Ulissi and C. L. Zitnick, *UMA: A Family of Universal Models for Atoms*, 2025, https://arxiv.org/abs/2506.23971.

16 D. P. Kovács, J. H. Moore, N. J. Browning, I. Batatia, J. T. Horton, Y. Pu, V. Kapil, W. C. Witt, I.-B. Magdău, D. J. Cole and G. Csányi, *J. Am. Chem. Soc.*, 2025, **147**, 17598–17611.

17 G. Benedini, A. Loew, M. Hellstrom, S. Botti and M. A. L. Marques, *Universal Machine Learning Potential for Systems with Reduced Dimensionality*, 2025, https://arxiv.org/abs/2508.15614.

18 J. Xia and B. Jiang, *Efficient Parallelization of Message Passing Neural Network Potentials for Large-scale Molecular Dynamics*, 2025, https://arxiv.org/abs/2505.06711.

19 M. Besta and T. Hoefler, *Parallel and Distributed Graph Neural Networks: An In-Depth Concurrency Analysis*, 2023, https://arxiv.org/abs/2205.09702.

20 N. Keriven, Proceedings of the 36th International Conference on Neural Information Processing Systems, Red Hook, NY, USA, 2022.

21 S. A. Ghasemi, A. Hofstetter, S. Saha and S. Goedecker, *Phys. Rev. B*, 2015, **92**, 045131.

22 T. W. Ko, J. A. Finkler, S. Goedecker and J. Behler, *Nature Communications*, 2021, **12**, 398.

23 P. Fuchs, M. Sanocki and J. Zavadlav, *Learning non-local molecular interactions via equivariant local representations and charge equilibration*, 2025.

24 A. Gao and R. C. Remsing, *Nature Communications*, 2022, **13**, 1572.

25 A. Kosmala, J. Gasteiger, N. Gao and S. Günnemann, International Conference on Machine Learning (ICML), 2023.

26 Y. Ji, J. Liang and Z. Xu, *Phys. Rev. Lett.*, 2025, **135**, 178001.

27 J. T. Frank, S. Chmiela, K.-R. Müller and O. T. Unke, *Euclidean Fast Attention: Machine Learning Global Atomic Representations at Linear Cost*, 2024.

28 B. Cheng, *npj Computational Materials*, 2025, **11**, 80.

29 D. P. Kovács, J. H. Moore, N. J. Browning, I. Batatia, J. T. Horton, Y. Pu, V. Kapil, W. C. Witt, I.-B. Magdău, D. J. Cole and G. Csányi, *J. Am. Chem. Soc.*, 2025, **147**, 17598–17611.

30 S. Mannan, V. Bihani, C. Gonzales, K. L. K. Lee, N. N. Gosvami, S. Ranu, S. Miret and N. M. A. Krishnan, *Evaluating Universal Machine Learning Force Fields Against Experimental Measurements*, 2025, https://arxiv.org/abs/2508.05762.

31 S. Yuan, J. Peng, B. Cai, Z. Huang, A. T. Garcia-Esparza, D. Sokaras, Y. Zhang, L. Giordano, K. Akkiraju, Y. G. Zhu, R. Hübner, X. Zou, Y. Román-Leshkov and Y. Shao-Horn, *Nature Materials*, 2022, **21**, 673–680.

32 J. Burner, L. Schwiedrzik, M. Krykunov, J. Luo, P. G. Boyd and T. K. Woo, *J. Phys. Chem. C*, 2020, **124**, 27996–28005.

33 M. J. Kalmutzki, N. Hanikel and O. M. Yaghi, *Science Advances*, 2018, **4**, eaat9180.

34 A. S. Rosen, S. M. Iyer, D. Ray, Z. Yao, A. Aspuru-Guzik, L. Gagliardi, J. M. Notestein and R. Q. Snurr, *Matter*, 2021, **4**, 1578–1597.

35 Y. Kang, H. Park, B. Smit and J. Kim, *Nature Machine Intelligence*, 2023, **5**, 309–318.

36 J. Burner, J. Luo, A. White, A. Mirmiran, O. Kwon, P. G. Boyd, S. Maley, M. Gibaldi, S. Simrod, V. Ogden and T. K. Woo, *Chemistry of Materials*, 2023, **35**, 900–916.

37 S. Liu, R. Dupuis, D. Fan, S. Benzaria, M. Bonneau, P. Bhatt, M. Eddaoudi and G. Maurin, *Chem. Sci.*, 2024, **15**, 5294–5302.

38 A. Sriram, S. Choi, X. Yu, L. M. Brabson, A. Das, Z. Ulissi, M. Uyttendaele, A. J. Medford and D. S. Sholl, *ACS Cent. Sci.*, 2024, **10**, 923–941.

39 M. Islamov, H. Babaei, R. Anderson, K. B. Sezginel, J. R. Long, A. J. H. McGaughey, D. A. Gomez-Gualdron and C. E. Wilmer, *npj Computational Materials*, 2023, **9**, 11.

40 S. Davis, E. Athira and V. K. Rajan, *Computational Materials Science*, 2025, **247**, 113537.

41 Y. Tao, G. Zhang and H. Xu, *Sustainable Materials and Technologies*, 2022, **32**, e00383.

42 L. Li, Y. Zhu, Z. Qi, X. Li, H. Pan, B. Liu and Y. Liu, *Applied Organometallic Chemistry*, 2023, **37**, e7199.

43 S. Vandenhaute, M. Cools-Ceuppens, S. DeKeyser, T. Verstraelen and V. Van Speybroeck, *npj Computational Materials*, 2023, **9**, 19.

44 J. Zhang, D. Chen, Y. Xia, Y.-P. Huang, X. Lin, X. Han, N. Ni, Z. Wang, F. Yu, L. Yang, Y. I. Yang and Y. Q. Gao, *J. Chem. Theory Comput.*, 2023, **19**, 4338–4350.

45 D. Fan, A. Ozcan, P. Lyu and G. Maurin, *Nanoscale*, 2024, **16**, 3438–3447.

46 A. Sharma and S. Sanvito, *npj Computational Materials*, 2024, **10**, 237.

47 J. Luo, O. B. Said, P. Xie, M. Gibaldi, J. Burner, C. Pereira and T. K. Woo, *npj Computational Materials*, 2024, **10**, 224.

48 S. Kancharlapalli, A. Gopalan, M. Haranczyk and R. Q. Snurr, *J. Chem. Theory Comput.*, 2021, **17**, 3052–3064.

49 S. Thaler, F. Mayr, S. Thomas, A. Gagliardi and J. Zavadlav, *npj Computational Materials*, 2024, **10**, 86.

50 G. B. Damas, L. T. Costa, R. Ahuja and C. M. Araujo, *The Journal of Chemical Physics*, 2021, **155**, 024701.

51 H. Kraß, J. Huang and S. M. Moosavi, *MOFSimBench: Evaluating Universal Machine Learning Interatomic Potentials In Metal–Organic Framework Molecular Modeling*, 2025, https://arxiv.org/abs/2507.11806.

52 J. Gasteiger, S. Giri, J. T. Margraf and S. Günnemann, *Fast and Uncertainty-Aware Directional Message Passing for Non-Equilibrium Molecules*, 2022.

53 I. Batatia, D. P. Kovács, G. N. C. Simm, C. Ortner and G. Csányi, *MACE: Higher Order Equivariant Message Pass-*

Digital Discovery Accepted Manuscript

View Article Online
DOI: 10.1039/D5DD00570A

Open Access Article. Published on 13 April 2026. Downloaded on 4/14/2026 6:36:52 PM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

ing Neural Networks for Fast and Accurate Force Fields, 2023, https://arxiv.org/abs/2206.07697.

54 A. Sriram, L. M. Brabson, X. Yu, S. Choi, K. Abdelmaqsoud, E. Moubarak, P. de Haan, S. Löwe, J. Brehmer, J. R. Kitchin, M. Welling, C. L. Zitnick, Z. Ulissi, A. J. Medford and D. S. Sholl, *The Open DAC 2025 Dataset for Sorbent Discovery in Direct Air Capture*, 2025, https://arxiv.org/abs/2508.03162.

55 D. S. Levine, M. Shuaibi, E. W. C. Spotte-Smith, M. G. Taylor, M. R. Hasyim, K. Michel, I. Batatia, G. Csányi, M. Dzamba, P. Eastman, N. C. Frey, X. Fu, V. Gharakhanyan, A. S. Krishnapriyan, J. A. Rackers, S. Raja, A. Rizvi, A. S. Rosen, Z. Ulissi, S. Vargas, C. L. Zitnick, S. M. Blau and B. M. Wood, *The Open Molecules 2025 (OMol25) Dataset, Evaluations, and Models*, 2025, https://arxiv.org/abs/2505.08762.

56 D. P. Kovács, C. v. d. Oord, J. Kucera, A. E. A. Allen, D. J. Cole, C. Ortner and G. Csányi, *J. Chem. Theory Comput.*, 2021, **17**, 7696–7711.

57 Y. Liu, X. He and Y. Mo, *npj Computational Materials*, 2023, **9**, 174.

58 D. Kim, X. Wang, P. Zhong, D. S. King, T. J. Inizan and B. Cheng, *A universal augmentation framework for long-range electrostatics in machine learning interatomic potentials*, 2025, https://arxiv.org/abs/2507.14302.

59 D. S. King, D. Kim, P. Zhong and B. Cheng, *Nature Communications*, 2025, **16**, 8763.

60 P. Zhong, D. Kim, D. S. King and B. Cheng, *arXiv preprint arXiv:2504.05169*, 2025.

61 A. P. Bartók, R. Kondor and G. Csányi, *Phys. Rev. B*, 2013, **87**, 184115.

62 L. Himanen, M. O. Jäger, E. V. Morooka, F. Federici Canova, Y. S. Ranawat, D. Z. Gao, P. Rinke and A. S. Foster, *Computer Physics Communications*, 2020, **247**, 106949.

63 T. Sainburg, L. McInnes and T. Q. Gentner, *Neural Computation*, 2021, **33**, 2881–2907.

64 A. A. Orlov, T. N. Akhmetshin, D. Horvath, G. Marcou and A. Varnek, *Mol Inform*, 2024, **44**, e202400265.

65 S. Sosnin, *Drug Discovery Today*, 2025, **30**, 104392.

66 D. Boldini, D. Ballabio, V. Consonni, R. Todeschini, F. Grisoni and S. A. Sieber, *Journal of Cheminformatics*, 2024, **16**, 35.

67 M. Thürlemann, L. Böselt and S. Riniker, *Journal of Chemical Theory and Computation*, 2023, **19**, 562–579.

68 Y. Shaidu, F. Pellegrini, E. Küçükbenli, R. Lot and S. de Gironcoli, *npj Computational Materials*, 2024, **10**, 47.

69 A. K. Rappe and W. A. Goddard III, *The Journal of Physical Chemistry*, 1991, **95**, 3358–3363.

70 J. Su, M. Ahmed, Y. Lu, S. Pan, W. Bo and Y. Liu, *Neurocomputing*, 2024, **568**, 127063.

71 B. Cheng, *npj Computational Materials*, 2024, **10**, 157.

72 P. Fuchs, S. Thaler, S. Röcken and J. Zavadlav, *Computer Physics Communications*, 2025, **310**, 109512.

73 S. S. Schoenholz and E. D. Cubuk, *Journal of Statistical Mechanics: Theory and Experiment*, 2021, **2021**, 124016.

74 G. Mario and A. Daigavane, *allegro-jax*, https://github.com/mariogeiger/allegro-jax.

75 S. Thaler and J. Zavadlav, *Nature Communications*, 2021, **12**, 6884.

76 G. Mario and A. Daigavane, *mace-jax*, https://github.com/ACEsuit/mace-jax.

77 F. Ercolessi and J. B. Adams, *Europhysics Letters (EPL)*, 1994, **26**, 583–588.

78 I. Batatia, davkovacs, bernstei, ttompa, WillBaldwin0, J. Riebesell, H. Helal, M. Avaylon, R. Elijosius, V. Bharadwaj, wcwitt, EszterVU, A. M. Elena, R. Goodall, ThomasWarford, ElliottKasoar, A. S. Rosen, C. H. Ho, F. Musil, A. Spears, H. Beck, E. Sivonxay, N. Goennheimer, L. Schaaf, C. Joshi, S. De, H. Moore, T. Stenczel, samwaltonnorwood and Leo, *ACEsuit/mace: v0.3.14*, 2025, https://doi.org/10.5281/zenodo.16748079.

79 D. S. King, D. Kim, P. Zhong and B. Cheng, *LES*, https://github.com/ChengUCB/les_fit.

80 Y. Mei, A. C. Simmonett, F. C. Pickard, 4th, R. A. DiStasio, Jr, B. R. Brooks and Y. Shao, *J Phys Chem A*, 2015, **119**, 5865–5882.

81 P. Eastman, P. K. Behara, D. L. Dotson, R. Galvelis, J. E. Herr, J. T. Horton, Y. Mao, J. D. Chodera, B. P. Pritchard, Y. Wang, G. De Fabritiis and T. E. Markland, *Scientific Data*, 2023, **10**, 11.

82 P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt and SciPy 1.0 Contributors, *Nature Methods*, 2020, **17**, 261–272.

83 C. M. de Armas-Morejón, L. A. Montero-Cabrera, A. Rubio and J. Jornet-Somoza, *J. Chem. Theory Comput.*, 2023, **19**, 1818–1826.

84 H. Shiokawa, S. Ishida and K. Terayama, *Journal of Cheminformatics*, 2025, **17**, 57.

85 F. Grasselli, K. Rossi, S. de Gironcoli and A. Grisafi, *Long-range electrostatics in atomistic machine learning: a physical perspective*, 2026, https://arxiv.org/abs/2602.11071.

86 V. Zaverkin, M. Ferraz, F. Alesiani, H. Christiansen, M. Takamoto, F. Errica and M. Niepert, EurIPS 2025 Workshop on SIMBIOCHEM, 2025.

87 R. Gao, C. Yam, J. Mao, S. Chen, G. Chen and Z. Hu, *Nature Communications*, 2025, **16**, 10484.

88 S. Naserifar, D. J. Brooks, I. Goddard, William A. and V. Cvicek, *The Journal of Chemical Physics*, 2017, **146**, 124117.

Digital Discovery Accepted Manuscript

View Article Online
DOI: 10.1039/D5DD00570A

Data Availability Statement

Data Availability

The datasets used in this study are publicly available to download. Our experiments utilized QMOF v14 (DOI: https://doi.org/10.6084/m9.figshare.13147324.v14), OMOL25 (DOI: https://doi.org/10.48550/arXiv.2505.08762, accessed June 20, 2025), and ODAC25 (DOI: https://doi.org/10.48550/arXiv.2508.03162, accessed September 3, 2025). Details on the creation of utilized subsets of ODAC25 and OMOL25 datasets can be found in the Supplementary Information section 1.

Code Availability

The software chemtrain used to train MLIPs is publicly available at https://github.com/tummfm/chemtrain (DOI: 10.5281/zenodo.15438477, version 0.1.0). The CACE and MACE models with LES were trained using code available at https://github.com/ChengUCB/les/tree/main, https://github.com/BingqingCheng/cace, and https://github.com/ChengUCB/mace (accessed 09.2025). All code utilized in this work, including requirements, is available at https://github.com/msan9908/MLIP_generalizability (DOI: 10.5281/zenodo.19486082).

Open Access Article. Published on 13 April 2026. Downloaded on 4/14/2026 6:36:52 PM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

Digital Discovery Accepted Manuscript