Digital Discovery

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal | View Issue

Check for updates

Cite this: *Digital Discovery*, 2025, **4**, 3389

Received 13th March 2025
Accepted 30th September 2025

DOI: 10.1039/d5dd00103j

rsc.li/digitaldiscovery

Assessing zero-shot generalisation behaviour in graph-neural-network interatomic potentials

Chiheb Ben Mahmoud, * Zakariya El-Machachi, Krystian A. Gierczak, John L. A. Gardner and Volker L. Deringer

With the rapidly growing availability of machine-learned interatomic potential (MLIP) models for chemistry, much current research focuses on the development of generally applicable and "foundational" MLIPs. An important question in this context is whether, and how well, such models can transfer from one application domain to another. Here, we assess this transferability for an MLIP model at the interface of materials and molecular chemistry. Specifically, we study GO-MACE-23, a model designed for the extended covalent network of graphene oxide, and quantify its zero-shot performance for small, isolated molecules outside its direct scope, as well as for examples of chemical reactions. Our work provides quantitative insight into the generalisation ability of graph-based MLIP models and, by exploring their limits, can help to inform future developments.

Introduction

Machine-learned interatomic potentials (MLIPs) for atomistic simulations, trained on quantum-mechanical energy and force data, have advanced remarkably in recent years[1–3] and now almost routinely allow researchers to address a wide range of questions in chemistry and materials science.[4–7] Recently, MLIPs incorporating graph-based representations, commonly referred to as graph neural networks (GNNs),[8–12] have emerged as cost-effective yet chemically rich models of atomic interactions. The favourable constant scaling of GNN-based MLIPs with the number of atomic species means that they are, in principle, able to cover many elements from across the Periodic Table all in a single model.[12–15]

The enhanced chemical versatility provided by GNNs has inspired the development of so-called "pre-trained",[12] "foundational",[13] or "universal"[15,16] interatomic potentials. These models have been trained on large, structurally and chemically diverse datasets; they show promising baseline performance for a range of systems[17,18] and thus provide a practical tool for starting computational projects, as well as a basis for fine-tuning.[19] In the long run, one might want to employ these pre-trained MLIPs "as is", in a zero-shot manner, without additional training or adaptation. Zero-shot performance also yields an important indication of how well the underlying model generalises to unseen tasks and chemistries. Understanding and improving the zero-shot behaviour of MLIPs is therefore an important challenge.

Herein, we study the zero-shot generalisation behaviour of GO-MACE-23 (ref. 20), an MLIP model that was initially developed specifically for graphene oxide (GO). Conceptually, GO bridges the gap between pristine graphene and organic chemistry: its structural landscape involves a variety of bonding motifs from sp^2 carbon sheets to oxygen-rich domains and reactive edge sites.[21] We test whether this structural and chemical complexity may serve as a basis for transferability (albeit initially we thought of GO-MACE-23 as a single-purpose MLIP!), subjecting GO-MACE-23 to a range of out-of-domain benchmarks, from energetics to high-temperature molecular-dynamics (MD) simulations of chemical reactions. In this way, our present study explores: (i) the role of a chemically rich training dataset in building robust and generalisable MLIPs;[22] (ii) the importance of GNN-based architectures in doing so; and (iii) the question whether GO-MACE-23 could form a starting point for foundational MLIPs bridging materials and molecular chemistry. Data and code supporting this work are publicly available (see "Data availability" statement below).

Methodology

The GO-MACE-23 and MACE-OFF24 models

We focus on the GO-MACE-23 model, which was built using the MACE architecture[10,11] together with a bespoke data-generation protocol.[20] Initial training data were generated "from scratch" using CASTEP + ML[23] (accelerating *ab initio* MD through on-the-fly fitting of GAP models[24]), and then largely augmented through subsequent iterative training from MD trajectories driven by intermediate versions of MACE models. Over multiple iterations, configurations with functionalised edges, involving hydroxyl (–OH), aldehyde (–CHO), and carboxylic acid (–CO_2H)

Inorganic Chemistry Laboratory, Department of Chemistry, University of Oxford, Oxford OX1 3QR, UK. E-mail: chiheb.benmahmoud@chem.ox.ac.uk

moieties, were added to ensure good coverage of the structural and chemical features that might be expected to appear in a "real-world" GO sheet. Training labels, *viz.* total energies and forces, were obtained from density-functional-theory (DFT) computations performed with the plane-wave software CASTEP[25] using on-the-fly generated pseudopotentials, the Perdew–Burke–Ernzerhof (PBE) exchange–correlation functional,[26] and a plane-wave energy cutoff of 550 eV. An overview of the GO dataset is available in the SI.

As a baseline for current practice in modelling organic molecules, we choose two variants of the MACE-OFF family of MLIPs:[27] the "large" version of MACE-OFF23 commonly referred to as MACE-OFF23(L), which is trained on the SPICE dataset of molecular data version 1,[28] and the "medium" version of MACE-OFF24 commonly referred to as MACE-OFF24(M), which is trained on the SPICE dataset version 2.[29] MACE-OFF24(M) is more similar to GO-MACE-23 in terms of architecture, with the exception of the radial cut-off: 3.7 Å for GO-MACE-23 and 6.0 Å for MACE-OFF24. More details about the hyperparameters of the GNNs used in this work are provided in the SI. In the remainder of this work, we refer to MACE-OFF23(L) simply as MACE-OFF23 and to MACE-OFF24(M) as MACE-OFF24. In using MACE-OFF24 models as benchmarks, it is important to note the different DFT levels of theory compared to GO-MACE-23 the SPICE labels were obtained using DFT with the ωB97M-D3(BJ) exchange–correlation functional[30,31] and the def2-TZVPPD basis set.[32,33]

Benchmark data

We carry out numerical experiments using the revised version of the MD17 dataset (rMD17)[34] as well as the QM7-X dataset.[35] We select the 6 molecules from rMD17 that only contain the elements C, H, and O—the only ones in the GO dataset, and thus the only ones that GO-MACE-23 and other models directly fitted to its dataset can handle. For each molecule, we randomly select 1000 configurations from the available trajectories. The rMD17 labels were obtained in the original work using the PBE functional and the def2-SVP basis set.[26,32] As for QM7-X, we randomly choose 100 configurations from each of the 6 most common chemical formulae that only include C, H, and O.

The other test sets used in the present study are generated either by running MD simulations in the *NVT* ensemble or by relaxing molecules. In both cases, we use GO-MACE-23 to perform these tasks. We compute reference data using DFT, matching the settings for GO-MACE-23 and MACE-OFF24, where applicable. For comparison to GO-MACE-23 labels are obtained from CASTEP by placing the molecules in large periodic cells (>20 Å). For MACE-OFF24 compatible labels are obtained using the Atomic Simulation Environment (ASE)[36] Python interface of Psi4,[37] version 1.4.

Data overlap between molecules and graphene oxide

Before benchmarking GO-MACE-23 it is important to set performance expectations based on the similarity of the various test sets and the GO training set. In Fig. 1, we present a two-dimensional embedding, from principal component analysis (PCA), of the average atomistic features per snapshot as learned by GO-MACE-23. The use of average features eliminates the system-size dependence of the descriptors. In the map of Fig. 1, static rMD17 molecules lie outside the scope of the training data (blue), but fall within the SPICE dataset domain (red), which constitutes the training data of MACE-OFF24. We should thus expect MACE-OFF24 to outperform GO-MACE-23 for static molecules. Fullerenes (magenta) and encapsulated molecular species ("M @ C$_{60}$", yellow) are located on the outskirts of the GO region of the map—this is unexpected at first glance, as fullerenes are not part of the GO training data. However, some of their key characteristics can be learned from the GO backbone.

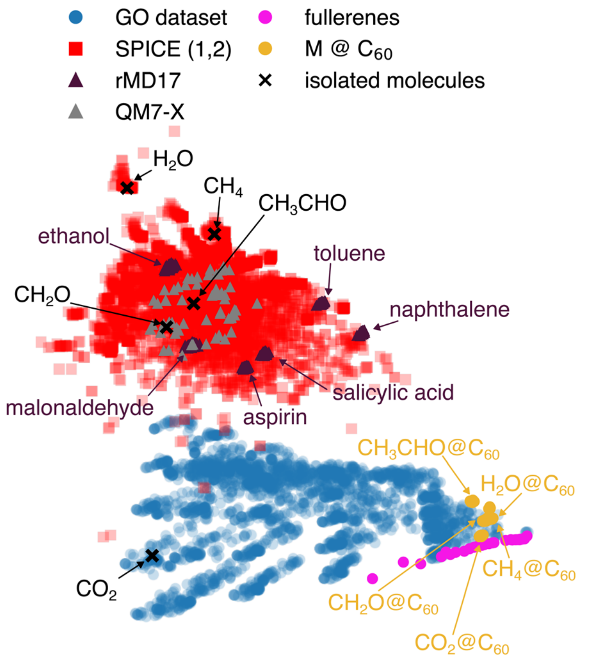

Fig. 1 Visualising the structural and chemical space explored in the present study. We show a two-dimensional embedding of the MACE descriptor trained on the GO dataset,[20] using principal component analysis. The points of the map correspond to the training set of GO-MACE-23 (blue), molecules containing C, H, and O atoms, representing ≈5% of the SPICE (version 1 and 2) datasets[28] (red), selected configurations from rMD17 trajectories[34] (purple) and the QM7-X dataset[35] (grey), a series of fullerenes with sizes ranging between 20 and 100 (magenta), five molecules encapsulated in C$_{60}$ fullerene cages (yellow), and the same molecules in vacuum (black crosses).

Zero-shot performance of GO-MACE-23

In this section, we evaluate the performance of GO-MACE-23 in predicting the energies and forces of small molecules, as well as vibrational spectra. Throughout this section, we use the terms "error" and "root mean square error" (RMSE) interchangeably.

Numerical performance for rMD17 and QM7-X

A common starting point in evaluating MLIP performance is to test prediction errors for energies and forces. These tests can be more complex than they look at first glance, because their outcome will strongly depend on the type of data used for testing (see, *e.g.*, ref. 38–44). In the present work, we are interested in zero-shot generalisability (without further modification of the model), which we here test by changing the application domain from extended GO structures to isolated small molecules.

We begin our series of zero-shot tests by evaluating the performance of GO-MACE-23 for the relevant trajectories from the rMD17 dataset. In Fig. 2, we summarise prediction errors on total energies and atomic forces relative to the recomputed QM targets using the same level of theory as that of GO-MACE-23. We obtain energy RMSE values below the often-quoted "chemical accuracy" of 1 kcal mol^{-1} or $\approx$40 meV $\mathrm{at.}^{-1}$. However, these errors can be significantly higher than the model's internal validation error for GO (1.8 meV atom^{-1} for energies and 109 meV $\mathrm{Å}^{-1}$ for forces, shown as dashed lines in Fig. 2). For aspirin, naphthalene, and salicylic acid from the rMD17 dataset, prediction errors of GO-MACE-23 for both energies and forces are compatible with the MLIP's validation errors on the GO dataset.

We next study the performance of GO-MACE-23 for molecules drawn from the more diverse QM7-X dataset (Fig. 3). Grouping the energy and force prediction errors according to the smallest ring size in any molecule (or the absence of any rings) reveals that the model's performance appears to correlate to some extent with the size and chemical nature of the smallest ring in the system. We select molecules containing 3-membered rings to illustrate this point: structure A has a cyclopropyl (C_3) ring, unlikely to be present in a well-annealed GO structure, and shows the highest energy RMSE of all selected structures containing any three-membered ring; by contrast, the 3-membered ring in B is an epoxy (C–O–C) moiety, a well-known structural motif in GO,[20,21] and this molecule has the lowest energy RMSE among those characterised in Fig. 3. Similar arguments can be made for the molecules with the highest and lowest force error, respectively: C contains a 3-membered cyclopropyl as well as a 4-membered oxetane ring, whereas D again shows an epoxy group as the 3-membered structural unit.

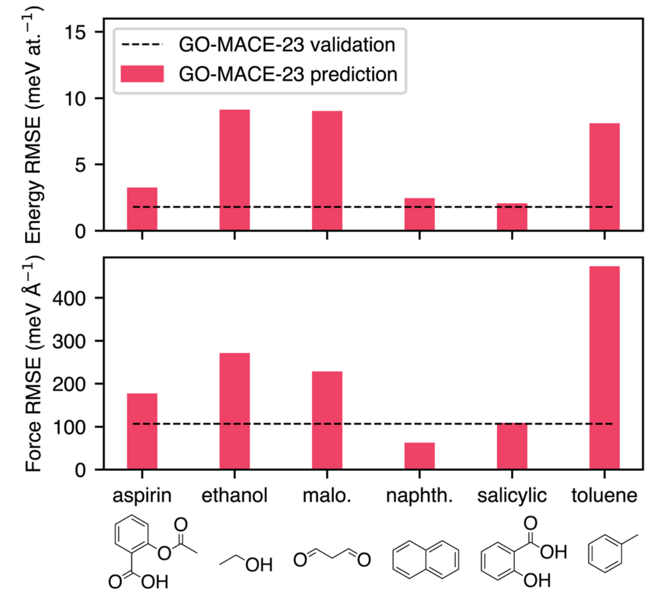

Fig. 2 Energy and force errors on six trajectories from the revised MD17 dataset for GO-MACE-23. The bars represent the RMSE of quantities between GO-MACE-23 predictions and its DFT level of theory. The dashed line is the internal validation error of GO-MACE-23.

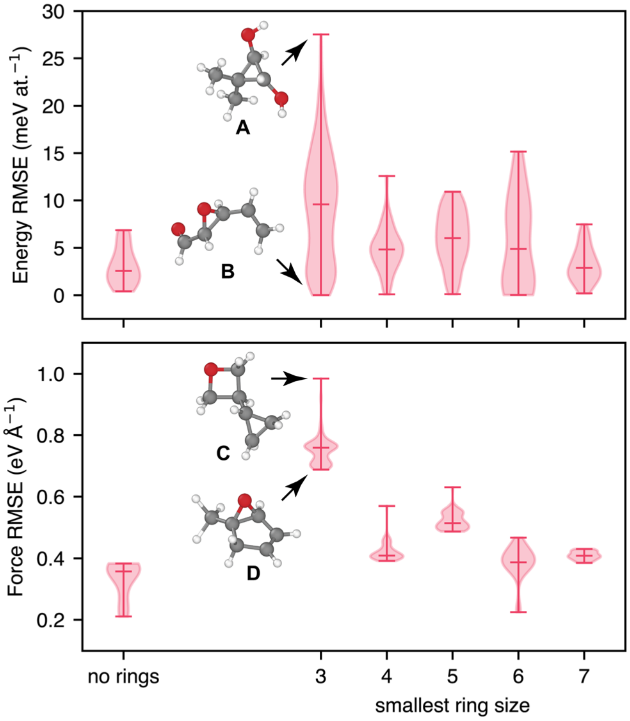

Fig. 3 As Fig. 2, but for configurations from the QM7-X dataset. We group the results based on the smallest ring found in the respective molecule. Four examples are highlighted: A and B, showing the highest and lowest energy error among all structures where the smallest ring is 3-membered, and C and D, showing the highest and lowest force error among those.

We note that despite these relatively large numerical errors, GO-MACE-23 is still robust: it yields stable MD trajectories of all molecules from rMD17 and QM7-X in the *NVT* ensemble at T = 500 K for 1 ns.

This evaluation highlights the importance of contextualising zero-shot performance of pretrained ML models across datasets. Most of the force prediction errors stem from the presence of under-represented geometries in the training set, as suggested by Fig. 2 where molecules with structural motifs resembling those in a GO sheet are better captured by GO-MACE-23 reinforcing the importance of dataset choice for generalisability. In the following subsection, we analyse one of these cases in detail: toluene from rMD17. It is worth noting that, although we were able to recompute DFT labels for all test molecules in our work, this is not a typical scenario. In many cases, comparisons are made across different levels of theory, meaning that systematic errors arising from the labels are entangled with, but distinct from, the model's own uncertainties. This underscores the importance of robust contextual analysis in ML model evaluation.

Toluene as a special case

To better understand the performance limits of GO-MACE-23 we analyse the errors for toluene in more detail, as it exhibits the highest force prediction RMSE among all 6 rMD17 molecules considered here. Fig. 4 summarises our approach to exploring possible sources of error. The toluene molecule contains an aromatic carbon atom directly bonded to an sp^3 carbon atom in a methyl group ($-CH_3$), coloured in red and blue in Fig. 4a, respectively. These two carbon atoms have the highest overall force errors exceeding 1.2 eV Å^{-1} (Fig. 4b). The high force errors on these specific atoms indicate that GO-MACE-23 cannot faithfully model their behaviour, due to the under-representation of similar atomic environments in the training set.

Most current MLIPs (including the MACE architecture) describe the total energy of a chemical system as a sum of atomic energies, following ref. 46 and 24. While this decomposition is useful for training and extrapolating ML models, it is not inherently physical and has no direct counterpart in a quantum-mechanical computation: so it is possible for the MLIP to reproduce the global behaviour without capturing the expected local energy distribution. This issue is evident in the present case of toluene (Fig. 4c): the combined error for the sum of the forces is only one-third of the individual force-component errors. The predicted atomic energies confirm this limitation (Fig. S2): the "red" atom of the aromatic ring has the lowest predicted atomic energy of all the carbon atoms, while the "blue" atom of the methyl group has the highest. When averaging the energies of these two atoms, the methyl carbon and its direct neighbour have the lowest local energy across the randomly selected 200 snapshots in the trajectory (Fig. S2). The atomic decomposition ansatz provides a partial explanation in this case. More generally, further work is necessary to fully understand the local predictions of MLIPs, and steps towards this goal have been made.[47–49]

Vibrational spectra

The vibrational spectrum—which provides information about bending, twisting, stretching, *etc.*, of individual bonds—is a fingerprint of a molecule (and experimentally accessible), and reproducing it accurately is therefore an important test for an MLIP. To assess the ability of GO-MACE-23 to predict vibrational spectra, we focus on three molecules from the rMD17 dataset: naphthalene and toluene representing the best and worst force predictions, respectively (*cf.* Fig. 2), and malonaldehyde as an example of a molecule without a 6-membered aromatic ring (the principal structural fragment of graphene). We also include one conformer each representing $C_5H_8O_2$, $C_6H_{12}O$, and $C_6H_{10}O$ from the QM7-X dataset. We start by selecting a random snapshot from the six subsets, then relax the molecules using GO-MACE-23. The force errors for the relaxed structures are 0.05, 0.32, and 0.22 eV Å^{-1} for naphthalene, toluene, and malonaldehyde, respectively, and 0.31, 0.66, and 0.33 eV Å^{-1} for the selected $C_5H_8O_2$, $C_6H_{12}O$, and $C_6H_{10}O$ structures, respectively. Then, we compute the vibrational spectra with the MLIP and DFT at the corresponding level, using finite displacements, with phonopy.[50,51] We present the resulting spectra in the upper panels of Fig. 5. The GO-MACE-23-predicted spectra agree qualitatively with their DFT counterparts, and the quality of the prediction correlates well with the model's force accuracy. The low-frequency modes, in particular, are well reproduced, while the accuracy decreases for the high-frequency modes. Additionally, we relaxed these molecules using DFT and report, in the SI, the root-mean-square displacement between geometries optimised with DFT and GO-MACE-23. We note that the discrepancies are relatively high, ranging between 0.17 Å and 0.28 Å. A recent study in ref. 52 suggests that these discrepancies may arise from a softened potential-energy surface near the relevant snapshots, which could explain the reduced accuracy for high-frequency modes.

We compare GO-MACE-23 to MACE-OFF23 and MACE-OFF24, two molecular MLIP models trained on different versions of the SPICE molecular dataset (see Methodology section). We compute the vibrational spectra on the GO-MACE-23-relaxed molecules using MACE-OFF24 and their corresponding DFT level of theory. The force errors of MACE-OFF23 are 0.003, 0.002, 0.016, 0.023, 0.015, and 0.008 eV Å^{-1} for naphthalene, toluene, malonaldehyde, $C_5H_8O_2$, $C_6H_{12}O$, and $C_6H_{10}O$ respectively. The force errors of MACE-OFF24 are 0.005, 0.003, and 0.005, 0.033, 0.03, 0.016 eV Å^{-1} for naphthalene, toluene, malonaldehyde, $C_5H_8O_2$, $C_6H_{12}O$, and $C_6H_{10}O$ respectively. We report the spectra in the lower panels of Fig. 5. As shown in Fig. 1, the rMD17

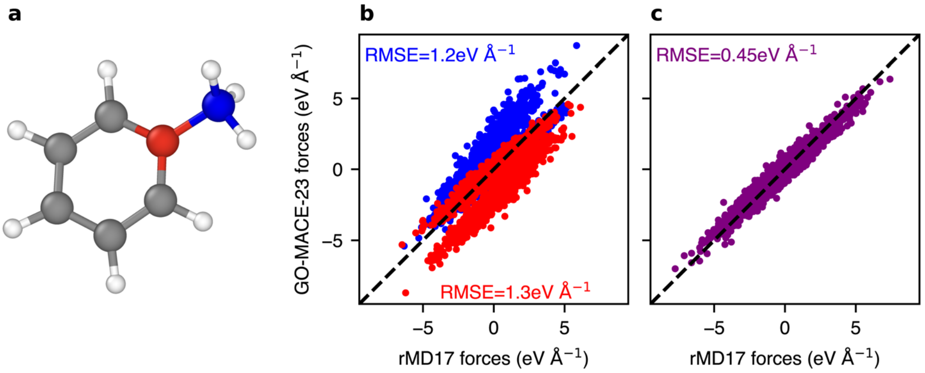

Fig. 4 (a) Visualisation of a toluene molecule obtained using OVITO.[45] Red- and blue-coloured atoms are carbon atoms part of the aromatic ring and the attached methyl group, respectively. (b) Force components parity plot of the DFT-computed and GO-MACE-23-predicted forces for the carbon atoms labelled red and blue in panel (a). (c) Force parity plot of the sum of forces of the red- and blue-labelled carbon atoms.

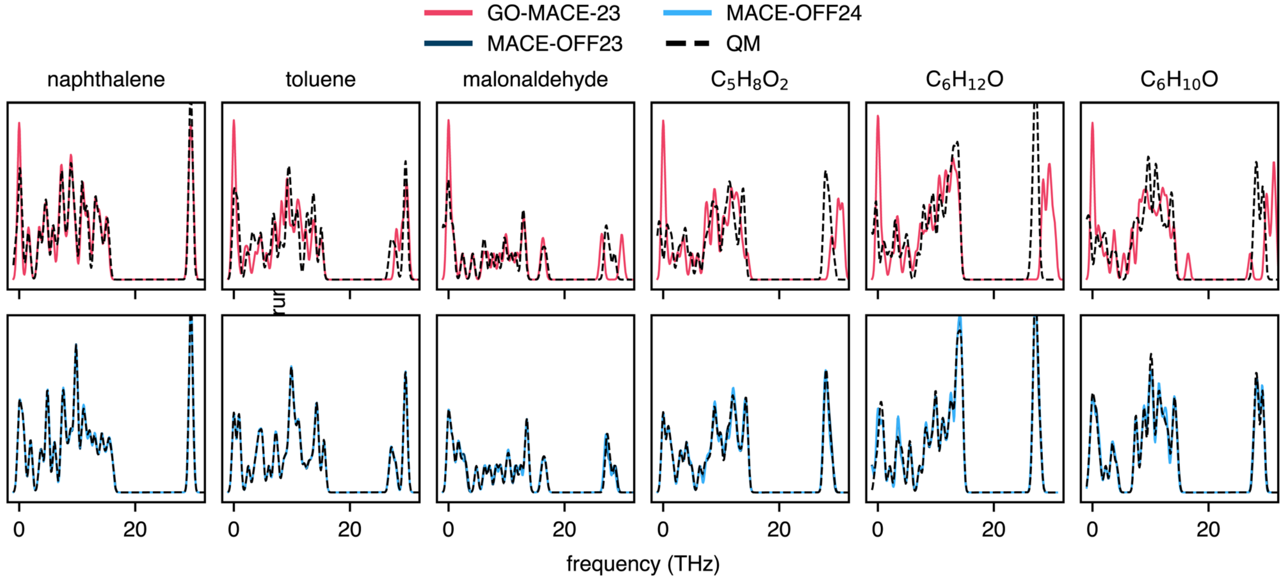

Fig. 5 Molecular vibrational spectra computed with MLIPs (solid lines) and DFT ("QM", dashed lines) for GO-MACE-23-relaxed naphthalene, toluene, malonaldehyde, $C_5H_8O_2$, $C_6H_{12}O$, and $C_6H_{10}O$ molecules. The upper row characterises the out-of-domain performance of GO-MACE-23 (red). The lower row shows the performance of MLIPs trained for molecules, *viz.* MACE-OFF[27] (dark and light blue, visually indistinguishable). Note that the DFT data have been computed at the level corresponding to the training data of the respective MLIP model; the DFT data in the upper and lower rows are therefore slightly different.

molecules are structurally similar to the training domain of the MACE-OFF24 models, which explains the models' high accuracy in predicting atomic forces. As a result, both MACE-OFF24 models produce more accurate vibrational spectra, reproducing both high- and low-frequency modes.

Fullerenes and encapsulated molecules

We use a series of fullerene molecules as another benchmark to quantify the transferability of GO-MACE-23 (and MACE-OFF24). The smallest fullerene is C_{20}, containing only 5-membered rings of carbon atoms and no 6-membered ones; consequently, its curvature is large. Yet, the fullerene was found to be the most stable C_{20} isomer using MP2 computations.[54] Larger fullerenes are structurally closer to graphene and graphite, and should therefore be closer to the training domain of GO-MACE-23 (*cf.* Fig. 1).

We first test the stability of GO-MACE-23 in generating MD trajectories for fullerenes. We run *NVT* simulations for C_{20}, C_{60}, and C_{100} for 1 ns, targeting $T = 500$ K. We find that GO-MACE-23 maintained structural integrity throughout the simulations, producing stable trajectories without signs of unphysical distortions. Both GO-MACE-23 and the MACE-OFF24 variants reproduce the general trend of growing stabilisation with fullerene size (Fig. 6). Prediction errors are highest for the smaller fullerenes, with energy errors of >100 meV at.$^{-1}$, and force RMSE >2 eV Å^{-1}, likely due to their high curvature. For C_{60}, the energy errors decrease to around 50 meV at.$^{-1}$ for all MLIPs, and force errors to around 250 meV Å^{-1} for GO-MACE-23 and MACE-OFF24. For small fullerenes (<60 carbon atoms), GO-MACE-23 performs better than both MACE-OFF24 models: we presume that this is due to the fact that it has encountered some curved graphene sheets (SI), including various odd-membered rings, during training. Note, however, that the latter are only a small fraction of the training data: the ring-size distribution in the GO–

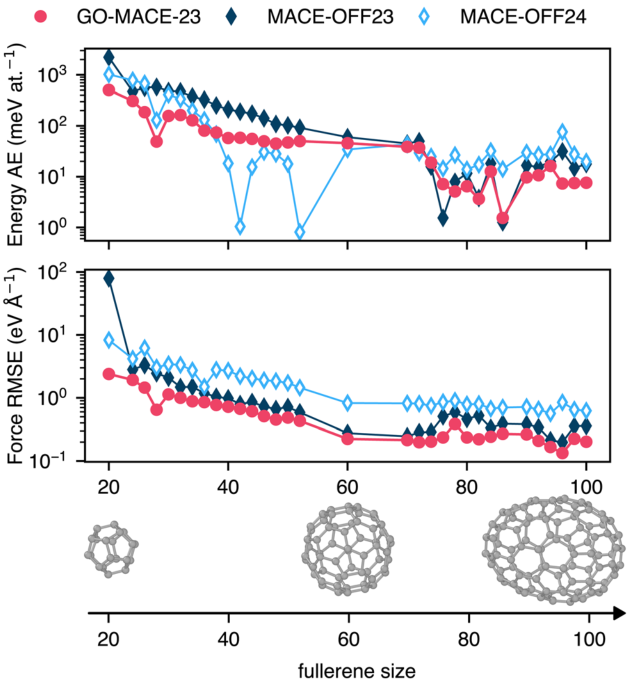

Fig. 6 Evolution of the prediction errors of the per-atom energy and forces of fullerenes, obtained from ref. 53, of sizes between 20 and 100 atoms computed with GO-MACE-23 and its corresponding DFT level of theory (red), and MACE-OFF and their corresponding DFT level of theory (dark and light blue). Similar to Fig. 5, lines represent the ML predictions, and the dashed lines represent the QM reference calculations. The rendered images show three fullerenes: C_{20}, C_{60}, and C_{100}.

MACE-23 dataset is 1 : 600 for 5 : 6-membered rings. MACE-OFF24 notably outperforms both GO-MACE-23 and MACE-OFF23 for fullerenes with 42 and 50 atoms. Analysis of the overlap between atomic environments in the SPICE datasets and the fullerene set (Fig. S4) shows that this overlap is limited to smaller fullerenes (<40 atoms for SPICE 1 and <50 atoms for SPICE 2), suggesting that the strong performance of MACE-OFF24 cannot be explained solely by training–test similarity. As a further test, we use GO-MACE-23 and both MACE-OFF24 models to calculate the vibrational spectra of two fullerenes, C_{20} and C_{60}, using the same protocol as for the rM17 and QM7-X molecules (Fig. S5). We find that GO-MACE-23 yields good accuracy compared to its DFT reference, while both MACE-OFF24 models fail to reproduce the spectrum of C_{20}.

In a recent study, Vyas *et al.* showed how formaldehyde (CH_2O) can be inserted into a C_{60} molecule by subsequent organic reaction steps,[55] expanding on existing work on endohedral fullerenes.[56,57] In the context of the present work, we show in Fig. 7 three case studies that have been discussed in the literature: encapsulated water (written as "$H_2O@C_{60}$"),[58] encapsulated methane ("$CH_4@C_{60}$"),[59] and encapsulated formaldehyde ("$CH_2O@C_{60}$").[55]

We use GO-MACE-23 to drive long MD trajectories of the three species in the *NVT* ensemble at $T = 500$ K, for 1 ns with a 0.5 fs timestep. Such simulations can be challenging test cases,[60] especially given the fusion temperature of C_{60} is estimated to be around 550 K.[61] We re-label snapshots from these MD trajectories with GO-MACE-23 and its corresponding DFT method, as well as MACE-OFF24 and its corresponding DFT method. In Fig. 7, we show the errors, expressed as absolute error (AE) values for energies and RMSE for forces. Both MLIPs exhibit similar energy prediction errors, with GO-MACE-23 performing better for the larger encapsulated molecules, and MACE-OFF23 for $H_2O@C_{60}$. However, GO-MACE-23 consistently yields lower force prediction errors across all of the test cases. This poorer performance of MACE-OFF23 and MACE-OFF24 may be attributed to the fact that fullerenes and encapsulated molecules are not present within the two versions of the SPICE training set. Additionally, GO-MACE-23 has encountered small molecules, such as CO and H_2O, near GO surfaces in its training data. Also, it is possible that GO-MACE-23 is accessing regions of configurational space that would be deemed unphysical by MACE-OFF24. To test this hypothesis, we run the same MD trajectories with MACE-OFF23 instead of GO-MACE-23 (SI). Of those simulations, only that for $CH_2O@C_{60}$ failed after the first timestep. We find that GO-MACE-23 more accurately reproduces the energies and forces for $H_2O@C_{60}$, whereas MACE-OFF23 performs better for $CH_4@C_{60}$. These results partially support the hypothesis that each MLIP explores regions of configurational space that are less well covered by the other MLIPs.

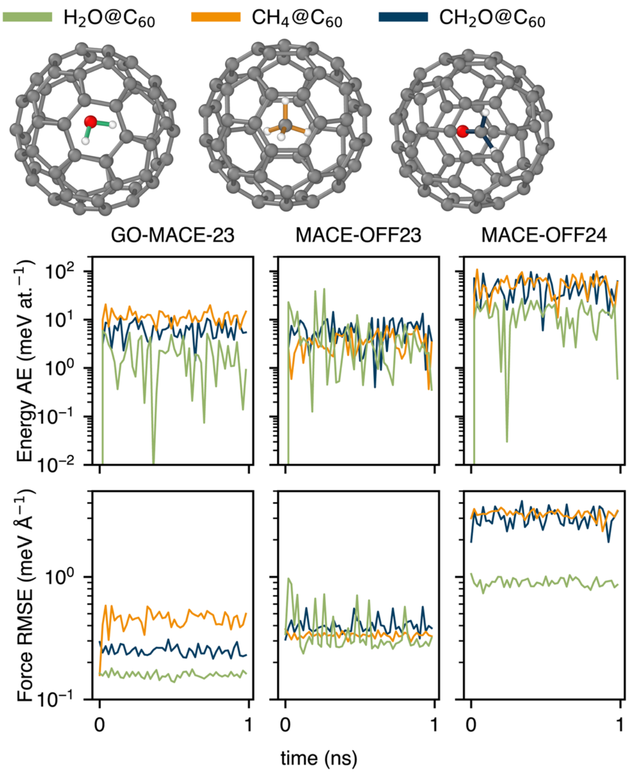

Fig. 7 Evolution of energy and force RMSE between GO-MACE-23 predictions and the corresponding DFT level of theory (left column), as well as between both MACE-OFF variants and their respective DFT levels of theory (middle and right columns). The errors are calculated from 1 ns trajectories at 500 K for H_2O, CH_4, and CH_2O enclosed in a C_{60} fullerene. The trajectories are driven by GO-MACE-23.

In the SI, we show two additional cases of encapsulated molecules, *viz.* CO_2 and acetaldehyde, the heavier homologue of CH_2O. Acetaldehyde is a challenging test case for GO-MACE-23, and has most likely not been seen during training (*cf.* Fig. 1). It is a thought experiment, of course, for the time being.

Experiments

Beyond the zero-shot performance evaluation so far, we carry out additional numerical experiments. These explore aspects of MLIP fitting methodology and provide an initial test for descriptions of gas-phase fragmentation reactions.

Model choice (I): effect of equivariant messages

The MACE architecture underlying GO-MACE-23 incorporates both invariant hidden features and equivariant hidden features of rank $L = 1$. In MACE, max L denotes the maximum degree of spherical harmonics used in the equivariant message-passing layers. It controls the complexity of the geometric information that the model can learn. For example, max $L = 0$ refers to an invariant model that can only capture isotropic features, and values of max $L = 0$ refer to an equivariant model encoding vectorial (and tensorial) information. To test the role of the equivariant features, we trained two modified versions of the model by varying MACE's internal symmetry rank. Specifically, we trained an invariant model by setting the highest rank of the internal features to max $L = 0$, and a higher-order equivariant model by setting max $L = 2$. This allows us to explore the possible correlation between the physical symmetries of an MLIP and its out-of-domain performance. Despite the fact that equivariant components can be included in MACE, the forces are computed by automatic differentiation of the total energy.

All MACE models are trained with the same protocol as GO-MACE-23, as detailed in the SI.

In Table 1, we compare the performance of MACE models with different maximum rank, *viz.* max $L \in \{0, 1, 2\}$. We train each model on 4 splits of the GO dataset, and compute prediction errors and uncertainty estimates, as standard deviation over the 4 splits, for all relevant rMD17 and QM7-X molecules. We notice that the original GO-MACE-23 model (max $L = 1$) does not systematically outperform its invariant counterpart (max $L = 0$). For example, the invariant model yields better energy predictions for toluene, aspirin, malonaldehyde, $C_5H_{10}O_2$, and C_6H_8O, as well as better force predictions for $C_5H_6O_2$, compared to GO-MACE-23. A similar trend is observed when comparing energies predicted by GO-MACE-23 and the max $L = 2$ MACE model. The force prediction errors are comparable within their uncertainties. Regardless of the benchmark reference calculation, we observe no clear correlation between max L and model performance, suggesting that equivariance and symmetry preservation play a limited role in generalisation for these domains. Particularly notable cases are toluene, $C_6H_{10}O$, and C_6H_8O, where GO-MACE-23 is the worst-performing model of the three, in terms of total-energy prediction.

Model choice (II): other GNN architectures

To further investigate the effect of design choices made for several popular GNNs on their generalisability, we trained multiple models on the GO-MACE-23 training dataset, using the universal interface graph-pes.[62] Particularly, we used the PaiNN,[63] TensorNet,[64] and NequIP[9] architectures. Details about hyperparameters, training protocol, and validation errors on the GO-MACE-23 dataset are provided in the SI.

Table 2 shows that GO-MACE-23 as well as re-fitted TensorNet and NequIP models generally yield low RMSE on most molecules considered. For instance, among the architectures in Table 2, NequIP achieves low energy errors on aspirin and malonaldehyde, whereas TensorNet performs best for toluene. Meanwhile, GO-MACE-23 has the lowest errors in force predictions for ethanol and naphthalene. These variations show that even closely related equivariant models can extract distinct mappings from the same data, influenced by subtle differences in model design and hyperparameters. We also compute the vibrational spectra of rMD17 and QM7-X molecules (*cf.* SI). We find that all of these GNNs reproduce the low-frequency spectrum with good accuracy, but the accuracy decreases substantially in the high-frequency regime.

These results highlight the importance of the MLIP architecture in capturing relevant atomistic information and transferring it beyond the training set. The extrapolation is not trivial and depends not only on the quality of the training data or the fit but also on the architecture itself. Notably, as shown in the SI, GO-MACE-23 has the lowest energy validation errors on the GO dataset, yet NequIP outperforms it for several rMD17 molecules in energy predictions. These results underscore the need for out-of-domain validation to fully assess model generalisation. Additionally, one could systematically investigate how the implementation choices of these GNNs, particularly in their atomic representations, influence their extrapolation capabilities, thereby enabling an *a priori* assessment of the performance of these MLIPs.[39,42,65]

Transferability to chemical reactions

The long-term goal of molecular interatomic potentials is to describe entire reaction mechanisms, rather than just the reactants and products. MLIPs are increasingly being used to describe transition states of reactions in vacuum[66,67] and in explicit solvent.[7] While GO-MACE-23 will have "seen" various rearrangements, decarbonylation reactions, *etc.*, during iterative training,[20] it has not been explicitly trained on molecular reaction mechanisms.

We use GO-MACE-23 to run a series of MD trajectories of an aspirin molecule in a periodic simulation cell of 30 Å length, using the *NVT* ensemble at $T = 1{,}500$ K. We re-label snapshots from the trajectories using the DFT reference method of GO-MACE-23, as well as using both MACE-OFF24 variants and their DFT reference method. In Fig. 8, we report two reaction pathways for the thermally driven decomposition of aspirin in

Table 1 Energy and force prediction RMSE as a function of the maximum rank of the equivariant hidden messages in the MACE architecture for trajectories from the rMD17 dataset and randomly selected structures from the QM7-X dataset (100 for each of the 6 given chemical formulae). "Malo." stands for malonaldehyde and "Naphth." for naphthalene. The lowest RMSE values for each molecules are highlighted in bold

	Energy RMSE (meV at.$^{-1}$)			Force RMSE (eV Å^{-1})		
max L	0	1	2	0	1	2
Aspirin	4.6 ± 0.4	4.7 ± 1.0	**3.2 ± 0.6**	0.27 ± 0.01	0.21 ± 0.02	**0.19 ± 0.01**
Ethanol	9.3 ± 0.9	**8.7 ± 1.3**	8.8 ± 0.5	0.33 ± 0.03	0.32 ± 0.03	**0.27 ± 0.02**
Malo.	9.6 ± 2.1	10.2 ± 3.0	**8.4 ± 0.7**	0.21 ± 0.03	0.20 ± 0.01	**0.17 ± 0.01**
Naphth.	2.4 ± 1.0	**1.3 ± 0.5**	**1.3 ± 0.3**	0.09 ± 0.01	**0.06 ± 0.01**	0.07 ± 0.01
Salicylic	2.5 ± 0.6	**2.1 ± 0.4**	2.8 ± 0.9	0.10 ± 0.01	0.10 ± 0.01	**0.09 ± 0.00**
Toluene	5.9 ± 3.3	8.3 ± 3.8	**4.3 ± 1.7**	0.22 ± 0.05	0.17 ± 0.01	**0.15 ± 0.02**
$C_6H_{10}O$	25.5 ± 4.8	29.0 ± 3.4	**24.3 ± 2.4**	0.63 ± 0.05	0.62 ± 0.05	**0.50 ± 0.06**
$C_5H_8O_2$	34.3 ± 7.6	34.5 ± 2.9	**34.0 ± 5.2**	0.46 ± 0.02	0.45 ± 0.04	**0.38 ± 0.04**
C_6H_8O	**44.8 ± 7.5**	56.5 ± 14.9	46.3 ± 6.7	0.56 ± 0.05	0.45 ± 0.05	**0.43 ± 0.06**
$C_6H_{12}O$	29.8 ± 12.0	33.7 ± 18.1	**15.7 ± 3.8**	0.49 ± 0.04	0.43 ± 0.06	**0.33 ± 0.04**
$C_5H_{10}O_2$	**22.0 ± 2.5**	24.2 ± 3.9	32.4 ± 6.2	0.43 ± 0.02	0.41 ± 0.03	**0.36 ± 0.02**
$C_5H_6O_2$	51.6 ± 6.2	**35.5 ± 10.0**	40.1 ± 7.0	0.43 ± 0.01	0.45 ± 0.02	**0.41 ± 0.05**

Table 2 Energy and force prediction RMSE different GNN architectures trained on the GO dataset, evaluated for structures from the revised MD17 dataset and QM7-X, as in Table 1. Errors are computed with respect to the DFT level of theory of the GO dataset. Malo." stands for malonaldehyde and "Naphth." for naphthalene. The lowest RMSE values for each molecules are highlighted in bold

	Energy RMSE (meV at.$^{-1}$)			
	GO-MACE-23	TensorNet	NequIP	PaiNN
Aspirin	4.7 ± 0.6	9.2 ± 1.2	**4.1 ± 0.3**	8.2 ± 1.0
Ethanol	**8.7 ± 0.7**	23.9 ± 4.1	12.3 ± 0.5	16.9 ± 1.9
Malo.	10.2 ± 1.7	21.0 ± 5.3	**8.7 ± 0.8**	19.9 ± 2.6
Naphth.	**1.3 ± 0.3**	4.9 ± 1.1	3.4 ± 0.9	10.1 ± 1.8
Salicylic	**2.1 ± 0.2**	10.4 ± 2.4	3.1 ± 0.6	19.4 ± 2.5
Toluene	8.3 ± 2.2	15.1 ± 1.5	**8.1 ± 1.8**	30.5 ± 6.4
$C_6H_{10}O$	**29.0 ± 3.4**	179.5 ± 162.2	51.4 ± 12.8	61.6 ± 10.8
$C_5H_8O_2$	**34.5 ± 2.9**	96.9 ± 24.1	34.9 ± 4.0	73.8 ± 14.1
C_6H_8O	56.5 ± 14.9	146.4 ± 20.5	**38.9 ± 9.0**	87.5 ± 16.1
$C_6H_{12}O$	**33.7 ± 18.1**	74.5 ± 38.7	34.5 ± 14.0	76.5 ± 13.3
$C_5H_{10}O_2$	**24.2 ± 3.9**	97.8 ± 46.2	44.1 ± 7.7	86.6 ± 7.3
$C_5H_6O_2$	**35.5 ± 10.0**	143.5 ± 38.8	40.9 ± 3.0	98.5 ± 16.6
	Force RMSE (eV Å^{-1})			
	GO-MACE-23	TensorNet	NequIP	PaiNN
Aspirin	**0.21 ± 0.03**	0.43 ± 0.09	0.24 ± 0.04	0.50 ± 0.12
Ethanol	**0.32 ± 0.05**	0.69 ± 0.24	0.39 ± 0.02	0.66 ± 0.13
Malo.	**0.20 ± 0.02**	0.52 ± 0.11	0.24 ± 0.03	0.46 ± 0.04
Naphth.	**0.06 ± 0.01**	0.21 ± 0.07	0.10 ± 0.02	0.30 ± 0.05
Salicylic	**0.10 ± 0.01**	0.26 ± 0.08	0.12 ± 0.01	0.36 ± 0.08
Toluene	**0.17 ± 0.03**	0.39 ± 0.12	0.20 ± 0.03	0.46 ± 0.13
$C_6H_{10}O$	**0.62 ± 0.05**	1.14 ± 0.12	0.70 ± 0.08	1.28 ± 0.09
$C_5H_8O_2$	**0.45 ± 0.04**	0.91 ± 0.16	0.52 ± 0.04	0.94 ± 0.05
C_6H_8O	**0.45 ± 0.05**	0.93 ± 0.18	0.56 ± 0.06	0.96 ± 0.10
$C_6H_{12}O$	**0.43 ± 0.06**	0.93 ± 0.19	0.53 ± 0.10	1.08 ± 0.07
$C_5H_{10}O_2$	**0.41 ± 0.03**	0.87 ± 0.19	0.45 ± 0.02	0.91 ± 0.05
$C_5H_6O_2$	**0.45 ± 0.02**	0.93 ± 0.18	0.47 ± 0.04	1.00 ± 0.08

vacuum into radical species which then recombine forming different molecules.

The upper panels of Fig. 8 depict the formation of reactive ketene and salicylic acid, a process involving the breaking of an ester bond. The reverse reaction was first described in ref. 68. Both GO-MACE-23 and the MACE-OFF24 variants accurately capture the energetics of the reactants and products. However, they significantly underestimate the energy of the intermediates. Despite this underestimation, the predicted average energy of the intermediates remains higher than that of the more stable reactants or products. In addition, these MLIPs were not able to reproduce the energy of the isolated radicals. We note that proper treatment of radicals requires open-shell methods, *e.g.* coupled-cluster theory,[69,70] or multireference approaches such as CASSCF,[71–74] particularly for modelling processes like *cis*-to *trans*-isomerisations. Stocker *et al.*[75] have previously discussed the limitations of MLIPs in accurately describing chemical reactions when radicals are not incorporated in the training data.

The lower panels of Fig. 8 illustrate the formation of an *o*-cresol molecule through a series of decarboxylation and decarbonylation steps. This reaction pathway shares the first set of radicals with the upper panel, with similar geometries, before developing into a different pathway. As with the previous pathway, all tested MLIPs underestimate the energy of the intermediate steps. The two models from the MACE-OFF24 family in particular overestimate the energy of the product system.

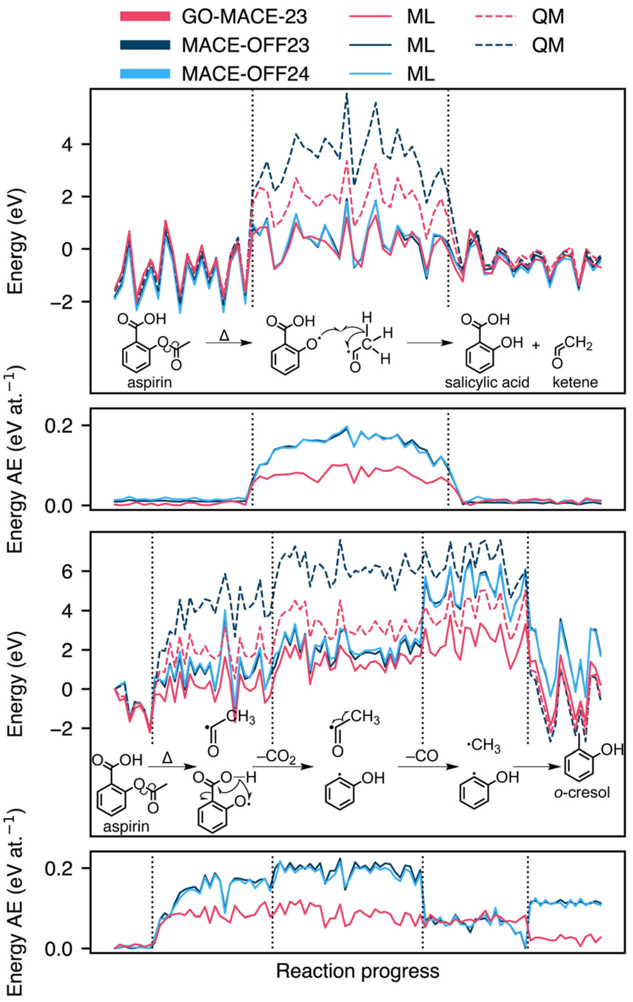

Fig. 8 Energy profiles of two exemplary high-temperature molecular-dynamics simulations computed with GO-MACE-23, MACE-OFF23, MACE-OFF24, and their respective QM references. The MD trajectories are driven by GO-MACE-23 and maintained at 1500 K. The first panel describes a reaction pathway to produce salicylic acid and ketene (H_2CCO) from aspirin. The third panel describes the decomposition of aspirin through a series of decarbonylations and decarboxylations to produce *o*-cresol. The second and fourth panels describe the difference between energies computed with ML and QM, for the first and second reactions, respectively, and expressed per atom.

Furthermore, we test an earlier model version from the iterative training of GO-MACE-23: this version, denoted "iter-8" in ref. 20, was not trained on edge structures. We find that GO-MACE-23 outperforms its simpler counterpart, especially in describing radicals and the products (SI). This indicates that some of the edge structures—with different chemical functionalisation—included in later iterations have likely contributed some information relevant to gas-phase molecular reactions to the GO-MACE-23 training data.

We emphasise that the present case study is not aimed at fully assessing performance in reaction modelling—but rather as a challenging test that deliberately takes the MLIP models away from their training domains. These (very-) high-temperature MD trajectories are not guaranteed to find the overall most favourable pathway, and yet they end in chemically sensible molecules. Following these trajectories as explored by the MLIPs themselves, we probe the potential-energy landscape for a range of configurations different from those in the rMD17 and QM7-X sets. This test completes our series of progressively more challenging "zero-shot" evaluations of GO-MACE-23 outside of its domain of training.

Conclusions

Located at the interface of materials and molecular modelling, graphene oxide offers an opportunity to connect these different domains of atomistic machine learning. In the present work, we have systematically assessed the zero-shot transferability of GO-MACE-23, an MLIP trained on data for GO, across relevant chemical benchmarks. We found good—perhaps surprisingly good—zero-shot performance compared to MACE-OFF24 a pre-trained model for molecular chemistry. The accuracy of both models decreases when describing reaction pathways, especially when radical species are involved.

Our study has tested the behaviour of recently proposed GNN MLIP models and their transferability, and we think that it can have implications for the future development of "foundational" models for atomistic simulations. Our results emphasise that including chemical reactivity in the training data is important in finding reaction pathways: in the process of building the GO-MACE-23 model,[20] we have sampled this reactivity in high-temperature MD simulations, and a similar approaches have been taken, *e.g.*, for the bulk carbon–hydrogen[76] and carbon–oxygen systems,[77] as well as organic reactions in the condensed phase.[78] We think that local-environment diversity will be as important as the chemical space coverage (*e.g.*, the number of chemical species) in defining future foundational models—this might include the addition of radical species (*cf.* Fig. 8) to the training data, either through very-high-temperature MD exploration or perhaps by explicitly involving "broken" bonds in the training protocol. Steps in this direction have been reported very recently.[79]

Despite its limitation to the three elements C, H, and O, the GO-MACE-23 model seems to provide a suitable starting point to study a wider range of chemistry-related questions than it was initially intended for, and we view this as a highly encouraging finding. We believe that together with improved data-generation strategies[22] as well as suitable workflows and automation approaches,[80–83] truly universal MLIPs for molecular systems, and for extended material structures built up from them, are coming within reach.

Author contributions

C. B. M., Z. E.-M., and V. L. D. designed the research. K. A. G. carried out pilot studies, and C. B. M. and Z. E.-M. carried out the final numerical experiments. J. L. A. G. provided code and methodology for MLIP fitting. All authors contributed to discussions. C. B. M. and V. L. D. wrote the manuscript, and all authors reviewed and approved the final version.

Conflicts of interest

There are no conflicts to declare.

Data availability

Data supporting the present study, including analysis code snippets, are available at https://github.com/cbenmahm/GO-Zero-Shot. A copy has been archived in Zenodo and is available at https://doi.org/10.5281/zenodo.17183916. An archived version of GraphPES is available at https://zenodo.org/records/14956211.

Supplementary information is available. See DOI: https://doi.org/10.1039/d5dd00103j.

Acknowledgements

We thank J. Holownia for useful discussions, and F. Duarte for useful discussions and comments on the manuscript. C.B.M. acknowledges funding from the Swiss National Science Foundation (SNSF) under grant number 217837. We are grateful for support from the EPSRC Centre for Doctoral Training in Theory and Modelling in Chemical Sciences (TMCS), under grant EP/L015722/1 (Z.E.M.). J.L.A.G. acknowledges a UKRI Linacre – The EPA Cephalosporin Scholarship, support from an EPSRC DTP award [grant number EP/T517811/1], and from the Department of Chemistry, University of Oxford. This work was supported by the Engineering and Physical Sciences Research Council [grant number EP/V049178/1] and UK Research and Innovation [grant number EP/X016188/1]. We are grateful for computational support from the UK national high performance computing service, ARCHER2,[84] for which access was obtained *via* the UKCP consortium and funded by EPSRC grant ref EP/X035891/1.

References

1 J. Behler, *Angew. Chem., Int. Ed.*, 2017, **56**, 12828–12840.

2 V. L. Deringer, M. A. Caro and G. Csányi, *Adv. Mater.*, 2019, **31**, 1902765.

3 O. T. Unke, S. Chmiela, H. E. Sauceda, M. Gastegger, I. Poltavsky, K. T. Schütt, A. Tkatchenko and K.-R. Müller, *Chem. Rev.*, 2021, **121**, 10142–10186.

4 A. P. Bartók, J. Kermode, N. Bernstein and G. Csányi, *Phys. Rev. X*, 2018, **8**, 041048.

5 B. Cheng, G. Mazzola, C. J. Pickard and M. Ceriotti, *Nature*, 2020, **585**, 217–220.

6 Y. Zhou, W. Zhang, E. Ma and V. L. Deringer, *Nat. Electron.*, 2023, **6**, 746–754.

7 H. Zhang, V. Juraskova and F. Duarte, *Nat. Commun.*, 2024, **15**, 6114.

8 K. T. Schütt, P. Kessel, M. Gastegger, K. A. Nicoli, A. Tkatchenko and K.-R. Müller, *J. Chem. Theory Comput.*, 2019, **15**, 448–455.

9 S. Batzner, A. Musaelian, L. Sun, M. Geiger, J. P. Mailoa, M. Kornbluth, N. Molinari, T. E. Smidt and B. Kozinsky, *Nat. Commun.*, 2022, **13**, 2453.

10 I. Batatia, D. P. Kovacs, G. Simm, C. Ortner and G. Csanyi, *Adv. Neural Inf. Process. Syst.*, 2022, 11423–11436.

11 I. Batatia, S. Batzner, D. P. Kovács, A. Musaelian, G. N. C. Simm, R. Drautz, C. Ortner, B. Kozinsky and G. Csányi, *Nat. Mach. Intell.*, 2025, **7**, 56–67.

12 B. Deng, P. Zhong, K. Jun, J. Riebesell, K. Han, C. J. Bartel and G. Ceder, *Nat. Mach. Intell.*, 2023, **5**, 1031–1041.

13 I. Batatia, P. Benner, Y. Chiang, A. M. Elena, D. P. Kovács, J. Riebesell, X. R. Advincula, M. Asta, W. J. Baldwin, N. Bernstein, A. Bhowmik, S. M. Blau, V. Cărare, J. P. Darby, S. De, F. Della Pia, V. L. Deringer, R. Elijošius, Z. El-Machachi, E. Fako, A. C. Ferrari, A. Genreith-Schriever, J. George, R. E. A. Goodall, C. P. Grey, 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, 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, E. Sivonxay, T. K. Stenczel, V. Svahn, C. Sutton, 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, *arXiv*, 2023, preprint, arXiv:2401.00096 [cond-mat, physics:physics], DOI: 10.48550/arXiv.2401.00096, http://arxiv.org/abs/2401.00096.

14 A. Merchant, S. Batzner, S. S. Schoenholz, M. Aykol, G. Cheon and E. D. Cubuk, *Nature*, 2023, **624**, 80–85.

15 H. Yang, C. Hu, Y. Zhou, X. Liu, Y. Shi, J. Li, G. Li, Z. Chen, S. Chen, C. Zeni, M. Horton, R. Pinsler, A. Fowler, D. Zügner, T. Xie, J. Smith, L. Sun, Q. Wang, L. Kong, C. Liu, H. Hao and Z. Lu, *MatterSim: A Deep Learning Atomistic Model Across Elements*, Temperatures and Pressures, 2024, https://arxiv.org/abs/2405.04967, Version Number: 2.

16 C. Chen and S. P. Ong, *Nat. Comput. Sci.*, 2022, **2**, 718–728.

17 B. Focassio, L. P. M. Freitas and G. R. Schleder, *ACS Appl. Mater. Interfaces*, 2024, 4c03815.

18 S. Ju, J. You, G. Kim, Y. Park, H. An and S. Han, Application of pretrained universal machine-learning interatomic potential for physicochemical simulation of liquid electrolytes in Li-ion battery, *Digital Discovery*, 2025, **4**, 1544–1559.

19 H. Kaur, F. Della Pia, I. Batatia, X. R. Advincula, B. X. Shi, J. Lan, G. Csányi, A. Michaelides and V. Kapil, *Faraday Discuss.*, 2025, **256**, 120–138.

20 Z. El-Machachi, D. Frantzov, A. Nijamudheen, T. Zarrouk, M. A. Caro and V. L. Deringer, *Angew. Chem., Int. Ed.*, 2024, e202410088.

21 D. R. Dreyer, S. Park, C. W. Bielawski and R. S. Ruoff, *Chem. Soc. Rev.*, 2010, **39**, 228–240.

22 C. Ben Mahmoud, J. L. A. Gardner and V. L. Deringer, *Nat. Comput. Sci.*, 2024, **4**, 384–387.

23 T. K. Stenczel, Z. El-Machachi, G. Liepuoniute, J. D. Morrow, A. P. Bartók, M. I. J. Probert, G. Csányi and V. L. Deringer, *J. Chem. Phys.*, 2023, **159**, 044803.

24 A. P. Bartók, M. C. Payne, R. Kondor and G. Csányi, *Phys. Rev. Lett.*, 2010, **104**, 136403.

25 S. J. Clark, M. D. Segall, C. J. Pickard, P. J. Hasnip, M. I. J. Probert, K. Refson and M. C. Payne, *Z. Kristallogr. Cryst. Mater.*, 2005, **220**, 567–570.

26 J. P. Perdew, K. Burke and M. Ernzerhof, *Phys. Rev. Lett.*, 1996, **77**, 3865–3868.

27 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.

28 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, *Sci. Data*, 2023, **10**, 11.

29 P. Eastman, B. P. Pritchard, J. D. Chodera and T. E. Markland, *J. Chem. Theory Comput.*, 2024, **20**, 8583–8593.

30 N. Mardirossian and M. Head-Gordon, *J. Chem. Phys.*, 2016, **144**, 214110.

31 A. Najibi and L. Goerigk, *J. Chem. Theory Comput.*, 2018, **14**, 5725–5738.

32 F. Weigend and R. Ahlrichs, *Phys. Chem. Chem. Phys.*, 2005, **7**, 3297.

33 D. Rappoport and F. Furche, *J. Chem. Phys.*, 2010, **133**, 134105.

34 A. S. Christensen and O. A. Von Lilienfeld, *Mach. Learn.: Sci. Technol.*, 2020, **1**, 045018.

35 J. Hoja, L. Medrano Sandonas, B. G. Ernst, A. Vazquez-Mayagoitia, R. A. DiStasio and A. Tkatchenko, *Sci. Data*, 2021, **8**, 43.

36 A. Hjorth Larsen, J. Jørgen Mortensen, J. Blomqvist, I. E. Castelli, R. Christensen, M. Dułak, J. Friis, M. N. Groves, B. Hammer, C. Hargus, E. D. Hermes, P. C. Jennings, P. Bjerre Jensen, J. Kermode, J. R. Kitchin, E. Leonhard Kolsbjerg, J. Kubal, K. Kaasbjerg, S. Lysgaard, J. Bergmann Maronsson, T. Maxson, T. Olsen, L. Pastewka, A. Peterson, C. Rostgaard, J. Schiøtz, O. Schütt, M. Strange, K. S. Thygesen, T. Vegge, L. Vilhelmsen, M. Walter, Z. Zeng and K. W. Jacobsen, *J. Phys.: Condens. Matter*, 2017, **29**, 273002.

37 D. G. A. Smith, L. A. Burns, A. C. Simmonett, R. M. Parrish, M. C. Schieber, R. Galvelis, P. Kraus, H. Kruse, R. Di Remigio, A. Alenaizan, A. M. James, S. Lehtola, J. P. Misiewicz, M. Scheurer, R. A. Shaw, J. B. Schriber, Y. Xie, Z. L. Glick, D. A. Sirianni, J. S. O'Brien, J. M. Waldrop, A. Kumar, E. G. Hohenstein, B. P. Pritchard, B. R. Brooks, H. F. Schaefer, A. Y. Sokolov, K. Patkowski, A. E. DePrince, U. Bozkaya, R. A. King, F. A. Evangelista, J. M. Turney, T. D. Crawford and C. D. Sherrill, *J. Chem. Phys.*, 2020, **152**, 184108.

38 L. Cheng, N. B. Kovachki, M. Welborn and T. F. Miller, *J. Chem. Theory Comput.*, 2019, **15**, 6668–6677.

39 C. Zeni, A. Anelli, A. Glielmo and K. Rossi, *Phys. Rev. B*, 2022, **105**, 165141.

40 D. Montes De Oca Zapiain, M. A. Wood, N. Lubbers, C. Z. Pereyra, A. P. Thompson and D. Perez, *npj Comput. Mater.*, 2022, **8**, 189.

41 B. Mazouin, A. A. Schöpfer and O. A. Von Lilienfeld, *Mater. Adv.*, 2022, **3**, 8306–8316.

42 A. K. A. Kandy, K. Rossi, A. Raulin-Foissac, G. Laurens and J. Lam, *Phys. Rev. B*, 2023, **107**, 174106.

43 D. F. Thomas Du Toit, Y. Zhou and V. L. Deringer, *J. Chem. Theory Comput.*, 2024, **20**, 10103–10113.

44 S. P. Niblett, P. Kourtis, I.-B. Magdău, C. P. Grey and G. Csányi, *J. Chem. Theory Comput.*, 2025, **21**, 6096–6112.

45 A. Stukowski, *Model. Simulat. Mater. Sci. Eng.*, 2010, **18**, 015012.

46 J. Behler and M. Parrinello, *Phys. Rev. Lett.*, 2007, **98**, 146401.

47 M. Eckhoff and J. Behler, *J. Chem. Theory Comput.*, 2019, **15**, 3793–3809.

48 S. Chong, F. Grasselli, C. Ben Mahmoud, J. D. Morrow, V. L. Deringer and M. Ceriotti, *J. Chem. Theory Comput.*, 2023, **19**, 8020–8031.

49 S. Chong, F. Bigi, F. Grasselli, P. Loche, M. Kellner and M. Ceriotti, *Faraday Discuss.*, 2025, **256**, 322–344.

50 A. Togo, L. Chaput, T. Tadano and I. Tanaka, *J. Phys.: Condens. Matter*, 2023, **35**, 353001.

51 A. Togo, *J. Phys. Soc. Jpn.*, 2023, **92**, 012001.

52 B. Deng, Y. Choi, P. Zhong, J. Riebesell, S. Anand, Z. Li, K. Jun, K. A. Persson and G. Ceder, *npj Comput. Mater.*, 2025, **11**, 9.

53 A. Barnard and G. Opletal, *Fullerene Data Set*, 2023, https://data.csiro.au/collection/csiro%3A59022v1.

54 V. Parasuk and J. Almlöf, *Chem. Phys. Lett.*, 1991, **184**, 187–190.

55 V. K. Vyas, G. R. Bacanu, M. Soundararajan, E. S. Marsden, T. Jafari, A. Shugai, M. E. Light, U. Nagel, T. Rõõm, M. H. Levitt and R. J. Whitby, *Nat. Commun.*, 2024, **15**, 2515.

56 A. A. Popov, S. Yang and L. Dunsch, *Chem. Rev.*, 2013, **113**, 5989–6113.

57 S. Bloodworth and R. J. Whitby, *Commun. Chem.*, 2022, **5**, 121.

58 O. Carrillo-Bohórquez, Á. Valdés and R. Prosmiti, *J. Chem. Theory Comput.*, 2021, **17**, 5839–5848.

59 S. Bloodworth, G. Sitinova, S. Alom, S. Vidal, G. R. Bacanu, S. J. Elliott, M. E. Light, J. M. Herniman, G. J. Langley, M. H. Levitt and R. J. Whitby, *Angew. Chem., Int. Ed.*, 2019, **58**, 5038–5043.

60 S. Stocker, J. Gasteiger, F. Becker, S. Günnemann and J. T. Margraf, *Mach. Learn.: Sci. Technol.*, 2022, **3**, 045010.

61 *Ullmann's encyclopedia of industrial chemistry*, ed. B. Elvers and G. Bellussi, Wiley-VCH, Weinheim, 7th edn, 2011.

62 J. L. A. Gardner, *Graph PES*, 2025, https://github.com/jla-gardner/graph-pes.

63 K. T. Schütt, O. T. Unke and M. Gastegger, Equivariant message passing for the prediction of tensorial properties and molecular spectra, *arXiv*, 2021, preprint, arXiv:2102.03150, 10.48550/arXiv.2102.03150, http://arxiv.org/abs/2102.03150.

64 G. Simeon and G. d. Fabritiis, TensorNet: Cartesian Tensor Representations for Efficient Learning of Molecular Potentials, *arXiv*, 2023, preprint, , arXiv:2306.06482 [cs], DOI: 10.48550/arXiv.2306.06482, http://arxiv.org/abs/2306.06482.

65 A. Glielmo, C. Zeni, B. Cheng, G. Csányi and A. Laio, *PNAS Nexus*, 2022, **1**, pgac039.

66 E. Komp and S. Valleau, *Chem. Sci.*, 2022, **13**, 7900–7906.

67 S. Choi, *Nat. Commun.*, 2023, **14**, 1168.

68 D. A. Nightingale, US Pat. 1604472, 1926.

69 P. J. Knowles, C. Hampel and H.-J. Werner, *J. Chem. Phys.*, 1993, **99**, 5219–5227.

70 J. D. Watts, J. Gauss and R. J. Bartlett, *J. Chem. Phys.*, 1993, **98**, 8718–8733.

71 B. O. Roos, P. R. Taylor and P. E. Sigbahn, *Chem. Phys.*, 1980, **48**, 157–173.

72 P. Siegbahn, A. Heiberg, B. Roos and B. Levy, *Phys. Scr.*, 1980, **21**, 323–327.

73 P. E. M. Siegbahn, J. Almlöf, A. Heiberg and B. O. Roos, *J. Chem. Phys.*, 1981, **74**, 2384–2396.

74 B. O. Roos, *Int. J. Quantum Chem.*, 2009, **18**, 175–189.

75 S. Stocker, G. Csányi, K. Reuter and J. T. Margraf, *Nat. Commun.*, 2020, **11**, 5505.

76 R. Ibragimova, M. S. Kuklin, T. Zarrouk and M. A. Caro, *Chem. Mater.*, 2025, **37**, 1094–1110.

77 T. Zarrouk, R. Ibragimova, A. P. Bartók and M. A. Caro, *J. Am. Chem. Soc.*, 2024, **146**, 14645–14659.

78 S. Zhang, M. Z. Makoś, R. B. Jadrich, E. Kraka, K. Barros, B. T. Nebgen, S. Tretiak, O. Isayev, N. Lubbers, R. A. Messerly and J. S. Smith, *Nat. Chem.*, 2024, **16**, 727–734.

79 S. Zhang, R. Zubatyuk, Y. Yang, A. Roitberg and O. Isayev, ANI-1xBB: an ANI based reactive potential, *J. Chem. Theory Comput.*, 2025, **21**(9), 4365–4374.

80 E. V. Podryabinkin and A. V. Shapeev, *Comput. Mater. Sci.*, 2017, **140**, 171–180.

81 T. A. Young, T. Johnston-Wood, V. L. Deringer and F. Duarte, *Chem. Sci.*, 2021, **12**, 10944–10955.

82 C. Van Der Oord, M. Sachs, D. P. Kovács, C. Ortner and G. Csányi, *npj Comput. Mater.*, 2023, **9**, 168.

83 Y. Liu, J. D. Morrow, C. Ertural, N. L. Fragapane, J. L. A. Gardner, A. A. Naik, Y. Zhou, J. George and V. L. Deringer, *Nat. Commun.*, 2025, **16**, 7666.

84 G. Beckett, J. Beech-Brandt, K. Leach, Z. Payne, A. Simpson, L. Smith, A. Turner and A. Whiting, *ARCHER2 Service Description*, Zenodo technical report, 2024, DOI: 10.5281/zenodo.14507040.