

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

Chiheb Ben Mahmoud<sup>\*1</sup>, Zakariya El-Machachi<sup>1</sup>, Krystian A. Gierczak<sup>1</sup>,  
John L. A. Gardner<sup>1</sup>, and Volker L. Deringer<sup>1</sup>

<sup>1</sup>Inorganic Chemistry Laboratory, Department of Chemistry, University of Oxford, Oxford  
OX1 3QR, UK

---

<sup>\*</sup>chiheb.benmahmoud@chem.ox.ac.uk

## Abstract

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 and chemical reactions outside its direct scope—in direct comparison with a state-of-the-art model which has been trained in-domain. Our work provides quantitative insight into the transfer and generalisation ability of graph-neural-network potentials and, more generally, makes a step towards the more widespread applicability of MLIPs in chemistry.

## Introduction

Machine-learned interatomic potentials (MLIPs) for atomistic simulations, trained on quantum-mechanical energy and force data, have advanced remarkably in recent years<sup>1–3</sup> and now almost routinely allow researchers to address a wide range of questions in chemistry and materials science<sup>4–7</sup>. Recently, MLIPs incorporating graph-based representations, commonly referred to as graph neural networks (GNNs)<sup>8–11</sup>, have emerged as cost-effective yet chemically rich models of atomic interactions. The favourable scaling of GNN-based MLIPs with the number of atomic species means that they are, in principle, able to cover elements from across the Periodic Table all in a single model<sup>11–14</sup>.

The enhanced chemical versatility provided by GNNs has inspired the development of so-called “pre-trained”<sup>11</sup>, “foundational”<sup>12</sup>, or “universal”<sup>14,15</sup> interatomic potentials. These models have been trained on large, structurally and chemically diverse datasets; they show promising baseline performance for a range of systems<sup>16,17</sup> and thus provide a practical tool for starting computational projects, as well as a basis for fine-tuning<sup>18</sup>. 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. 19), 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<sup>20</sup>. 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- rather than general-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<sup>21</sup>; (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.

## Methodology

### The GO-MACE-23 and MACE-OFF models

We focus on the GO-MACE-23 model, which was built using the MACE architecture<sup>9,10</sup> together with a bespoke data-generation protocol<sup>19</sup>. Initial training data were generated “from scratch” using CASTEP+ML<sup>22</sup> (accelerating *ab initio* MD through on-the-fly fitting of GAP models<sup>23</sup>), and then largely augmented through subsequent iterative training from MD trajectories driven by intermediate versions of MACE models. Over time, configurations with functionalised edges, involving hydroxyl ( $-OH$ ), aldehyde ( $-CHO$ ), and carboxylic acid ( $-CO_2H$ ) 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

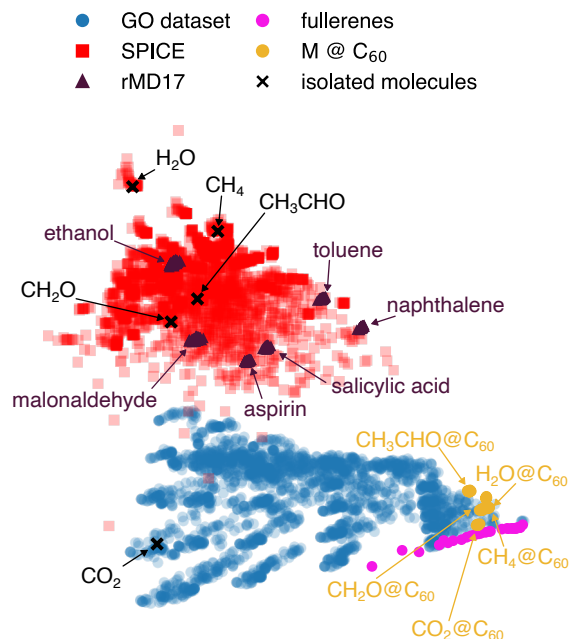
energies and forces, were obtained from density-functional-theory (DFT) computations performed with the plane-wave software CASTEP<sup>24</sup> using on-the-fly generated pseudopotentials and the Perdew–Burke–Ernzerhof (PBE) exchange–correlation functional<sup>25</sup>.

As a baseline for the current state-of-the-art (SOTA) in molecular modelling, we choose two variants of the MACE-OFF family of MLIPs<sup>26</sup>: 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<sup>27</sup>, and MACE-OFF24, which is trained on the SPICE dataset version 2<sup>28</sup>. MACE-OFF24 is more similar to G0-MACE-23 in terms of architecture, with the exception of the radial cut-off: 3.7 Å for G0-MACE-23 and 6.0 Å for MACE-OFF24. More details about the hyperparameters of all the GNNs used in this work are provided in the Supplementary Information. In the remainder of this work, we refer to MACE-OFF23(L) simply as MACE-OFF23. In using MACE-OFF models as benchmarks, it is important to note the different DFT levels of theory compared to G0-MACE-23: the SPICE labels were obtained from DFT computations with the  $\omega$ B97M-D3(BJ) exchange–correlation functional<sup>29,30</sup> and the def2-TZVPPD basis set<sup>31,32</sup>.

## Benchmark data

We carry out initial tests using the revised version of the MD17 dataset (rMD17)<sup>33</sup>. 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 G0-MACE-23 and other models directly fitted to its dataset can handle. For each molecule, we randomly select 1,000 configurations from the available trajectories. The rMD17 labels were obtained in the original work using the PBE functional and the def2-SVP basis set<sup>25,31</sup>.

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 G0-MACE-23 to perform these tasks. We compute reference data using DFT, matching the settings for G0-MACE-23 and MACE-OFF, where applicable. For comparison to G0-MACE-23, labels are obtained from CASTEP by placing the molecules in large periodic cells ( $> 20$  Å). For MACE-OFF, compatible labels are obtained using the Atomic Simulation Environment (ASE)<sup>34</sup> Python interface



**Figure 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<sup>19</sup>, 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  $\approx 5\%$  of the SPICE (version 1) dataset<sup>27</sup> (red), configurations from rMD17 trajectories<sup>33</sup> (purple), 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).

of Psi4<sup>35</sup>, 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. We observe that 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-OFF. We should thus expect MACE-OFF to outperform GO-MACE-23 for static molecules. “Acyclic” molecules such as ethanol seem to

be farther from the GO domain compared to cyclic molecules, such as aspirin. As a result, we expect GO-MACE-23 to provide more accurate predictions for cyclic molecules compared to acyclic configurations. Fullerenes (*magenta*) and encapsulated molecular species ("M @ C<sub>60</sub>", *yellow*) are located on the outskirts of the GO region of the map in Fig. 1—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.

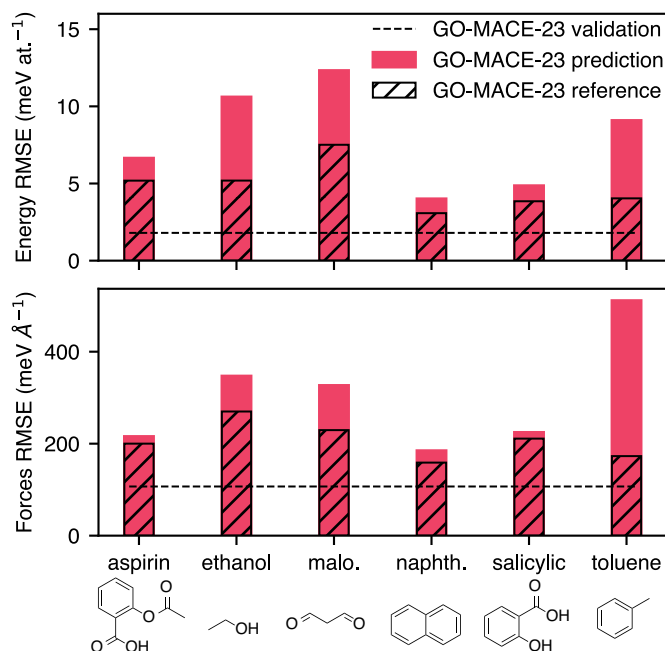
## Zero-shot performance of GO-MACE-23

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 MD17

A common starting point in evaluating MLIP performance is in testing 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. 36). 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.

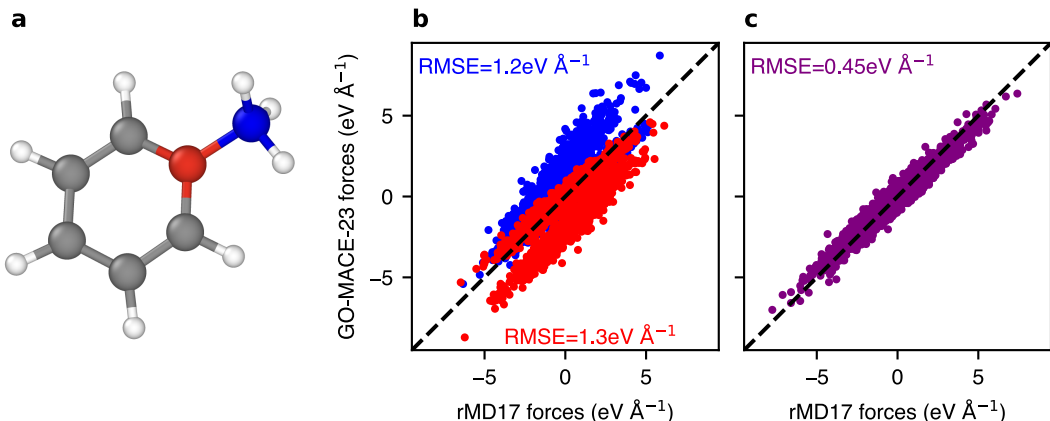
We begin our series of zero-shot tests by evaluating the performance of GO-MACE-23 for the relevant trajectories from the rMD17 dataset. In Fig. 2, we summarise the prediction errors on total energies and atomic forces relative to the QM targets of the rMD17 molecules. Despite the differences in the levels of theory between GO-MACE-23 and rMD17, we observe RMSE values below the often-quoted "chemical accuracy" of 1 kcal mol<sup>-1</sup> or  $\approx 40$  meV at.<sup>-1</sup>. However, these errors can be significantly higher than the model's *internal* internal validation error for GO (1.8 meV atom<sup>-1</sup> for energies and 109 meV Å<sup>-1</sup> for forces, shown as dashed lines in Fig. 2), which is the case for malonaldehyde. The latter is an example



**Figure 2:** Energy and force errors on six trajectories from the revised MD17 dataset using GO-MACE-23. The bars represent the RMSE of quantities between GO-MACE-23 predictions and rMD17 labels. The dashed area represents the errors between the DFT levels of theory used to label the GO dataset and the rMD17 dataset. The dashed line is the internal validation error of GO-MACE-23.

of an acyclic molecule (not containing an aromatic ring) that is not well represented in the GO-MACE-23 dataset. To explore the origin of these errors, we performed DFT calculations, using the same parameters as used for training GO-MACE-23, on the different test snapshots, and we report the RMSE between the levels of theory, as shown by hatched bars in Fig. 2. We find that GO-MACE-23 is primarily constrained by its own training QM labels, as systematic discrepancy errors account for approximately 30% to 90% of the errors. For aspirin, naphthalene, and salicylic acid, GO-MACE-23 introduces almost no additional errors beyond those inherent to its DFT training labels, and its prediction errors are comparable to its internal validation errors. GO-MACE-23 introduces almost no additional errors to the predictions made on cyclic molecules, although the magnitude of the errors varies notably.

This evaluation highlights the importance of contextualising zero-shot performance of pre-trained ML models across datasets. Most of the force prediction errors for the rMD17 molecules stem from discrepancies in the underlying DFT, with the exception of toluene



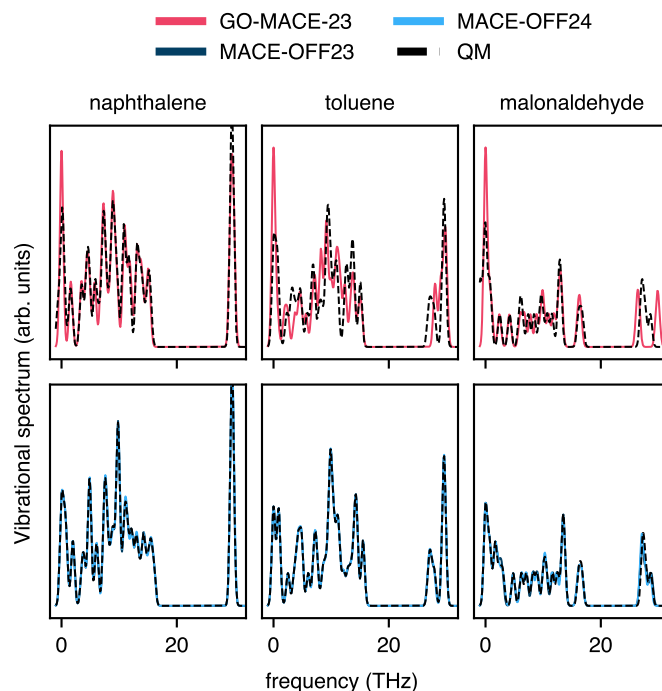
**Figure 3:** (a) Visualisation of a toluene molecule obtained using OVITO<sup>37</sup>. Red- and blue-coloured atoms are carbon atoms part of the aromatic ring and the attached methyl group, respectively. (b) Force components parity plot of the DFT-computed and GO-MACE-23-predicted forces for the carbon 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.

(which we address in the following). Figure 2 suggests that 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. While the ideal situation is to always compare data coming from uniform sources, we understand that this might not always be computationally feasible, underscoring the need for robust contextual analysis in ML model evaluations.

### 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. 3 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. 3a, respectively. These two carbon atoms have the highest overall force errors exceeding  $1.2 \text{ eV } \text{\AA}^{-1}$  (Fig. 3b). The high force errors on these specific atoms indicate that GO-MACE-23 is incapable of faithfully modelling their behaviour, due to the under-representation of similar atomic environments in the GO





**Figure 4:** Molecular vibrational spectra computed with MLIPs (*solid lines*) and DFT (“QM”, *dashed lines*) for G0-MACE-23-relaxed naphthalene, toluene, and malonaldehyde molecules. The upper row characterises the out-of-domain performance of G0-MACE-23 (*red*). The lower row shows the performance of SOTA MLIPs for molecules, viz. MACE-OFF<sup>26</sup> (*dark and light blue*). 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.

training set.

Most current MLIPs (including the MACE architecture) describe the total energy of a chemical system as a sum of atomic energies, following Refs. 38 and 23. 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. 3c): 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. S1): 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. S1). More generally, further work is necessary to fully understand the local predictions of MLIPs, and steps towards this goal have been made<sup>39,40</sup>.

## Vibrational spectra

The vibrational spectrum—which provides information about bending, twisting, and stretching of individual bonds—is a fingerprint of a molecule (and experimentally accessible), and is therefore an important test for an MLIP to accurately reproduce. To assess the ability of G0-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 start by selecting a random snapshot from the three trajectories, then relax the molecules using G0-MACE-23. The force errors for the relaxed structures are 0.05 eV Å<sup>-1</sup> for naphthalene, 0.32 eV Å<sup>-1</sup> for toluene, and 0.22 eV Å<sup>-1</sup> for malonaldehyde. Then, we compute the spectra with the MLIP and DFT at the corresponding level, using finite displacements, from phonopy<sup>41,42</sup>. We present the resulting vibrational spectra in the upper panels of Fig. 4. The G0-MACE-23-predicted spectra

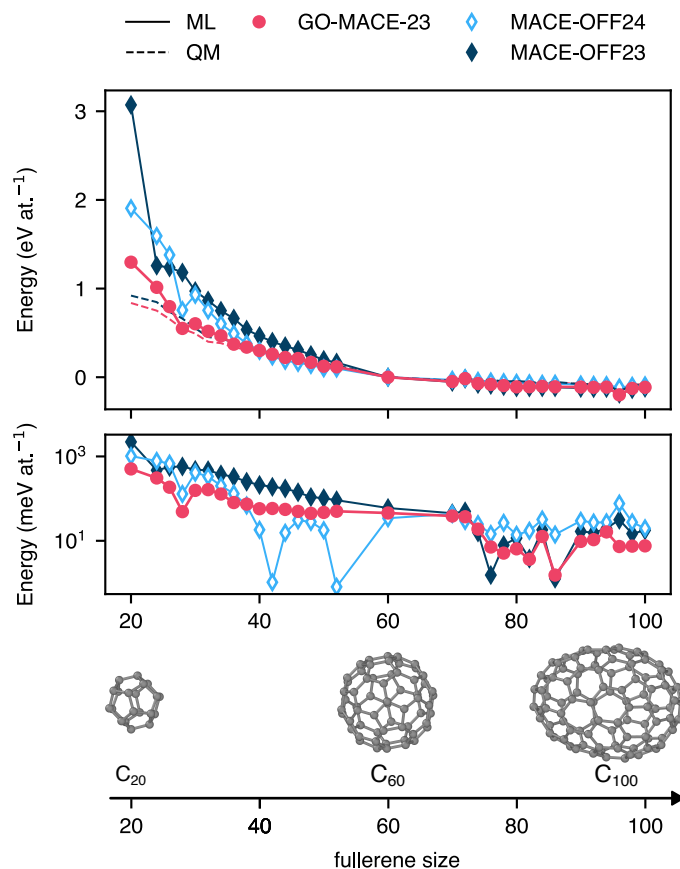
agree qualitatively with their DFT counterparts, and the quality of the prediction correlates well with the model’s force accuracy. The low-frequency modes, in particular, are well reproduced, while the accuracy decreases for the high-frequency modes. A recent study in Ref. 43 suggests that these discrepancies 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 SOTA 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-OFF and their corresponding DFT level of theory. The force errors of MACE-OFF23 are 0.003, 0.002, and 0.016 eV Å<sup>-1</sup> for naphthalene, toluene, and malonaldehyde, respectively. The force errors of MACE-OFF24 are 0.005, 0.003, and 0.005 eV Å<sup>-1</sup> for naphthalene, toluene, and malonaldehyde, respectively. We report the spectra in the lower panels of Fig. 4. As shown in Fig. 1, the rMD17 molecules fall within the training domain of the MACE-OFF models, which explains the models’ high accuracy in predicting atomic forces. As a result, both MACE-OFF models produce more accurate vibrational spectra, reproducing both high- and low-frequency modes.

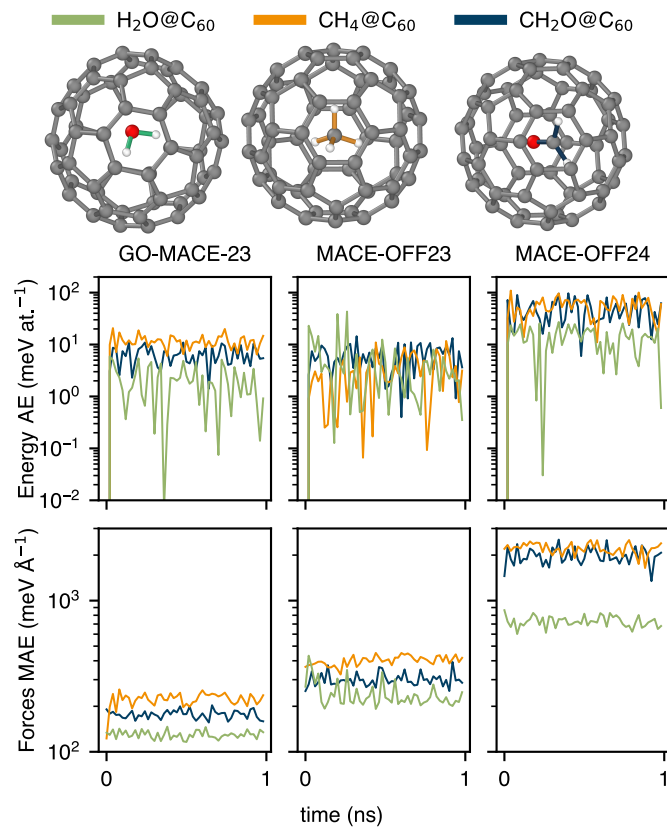
## Fullerenes and encapsulated molecules

We use a series of fullerene molecules as another benchmark to quantify the transferability of GO-MACE-23 (and MACE-OFF). The smallest fullerene is C<sub>20</sub>, containing only five-membered rings of carbon atoms and no six-membered ones. Consequently, its curvature is large, and the fullerene is found to be the most stable C<sub>20</sub> conformer using MP2 calculations<sup>45</sup>. Larger fullerenes are energetically and structurally closer to graphene and graphite, and should therefore be closer to the training domain of GO-MACE-23 (cf. Fig. 1).

Both GO-MACE-23 and MACE-OFF variants reproduce the general trend of growing stabilization with fullerene size, as shown in Fig. 5. Prediction errors are highest for the smaller fullerenes, with RMSE values higher than  $> 100$  meV at.<sup>-1</sup>, likely due to their high cur-



**Figure 5:** Evolution of the per-atom energy of fullerenes, obtained from Ref. 44, 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. 4, lines represent the ML predictions, and the dashed lines represent the QM reference calculations. All energies are referenced to C<sub>60</sub>. The lower panel describes the difference between energies computed with ML and QM, and expressed per atom. The rendered images show three fullerenes: C<sub>20</sub>, C<sub>60</sub>, and C<sub>100</sub>.



**Figure 6:** 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  $\text{H}_2\text{O}$ ,  $\text{CH}_4$ , and  $\text{CH}_2\text{O}$  enclosed in a  $C_{60}$  fullerene. The trajectories are driven by GO-MACE-23.

vature. For  $C_{60}$ , the RMSE decreases to around 50 meV  $\text{at.}^{-1}$  for all MLIPs. For small fullerenes ( $< 60$  carbon atoms), G0-MACE-23 performs better than both MACE-OFFmodels: we presume that this is due to the fact that it has encountered some curved graphene sheets, 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 G0-MACE-23 dataset is 1:600 for 5:6-membered rings. MACE-OFF24 significantly outperforms both G0-MACE-23 and MACE-OFF23 for fullerenes with the sizes of 42 and 50 atoms, hinting towards the existence of relevant motifs within the updated version of the SPICE dataset. This requires further investigation.

In a recent study, Vyas et al. showed how formaldehyde ( $\text{CH}_2\text{O}$ ) can be inserted into a  $C_{60}$  molecule by subsequent organic reaction steps<sup>46</sup>, expanding on existing work on endohedral fullerenes<sup>47,48</sup>. In the context of the present work, we show in Fig. 6 three case studies that have been discussed in the literature: encapsulated water (written as “ $\text{H}_2\text{O}@C_{60}$ ”) <sup>49</sup>, encapsulated methane (“ $\text{CH}_4@C_{60}$ ”) <sup>50</sup>, and encapsulated formaldehyde (“ $\text{CH}_2\text{O}@C_{60}$ ”) <sup>46</sup>.

We use G0-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<sup>51</sup>, especially given the fusion temperature of  $C_{60}$  is estimated to be around 550 K<sup>52</sup>. We re-label snapshots from these MD trajectories with G0-MACE-23 and its corresponding DFT method, as well as MACE-OFF and its corresponding DFT method. In Fig. 6, we show the errors, expressed as absolute errors (AE) for energies and mean absolute errors (MAE) for forces rather than our usual RMSE, for snapshots sampled every 20 ps. Both MLIPs exhibit similar energy prediction errors, with G0-MACE-23 performing better for the larger encapsulated molecules, and MACE-OFF23 for  $\text{H}_2\text{O}@C_{60}$ . However, G0-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, G0-MACE-23 has encountered small molecules, such as CO and  $\text{H}_2\text{O}$ , near GO surfaces in its training data. Also, it is possible that G0-MACE-23 is accessing regions of configurational

**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. Errors are computed with respect to the DFT level of theory of rMD17. The lowest RMSE values for each molecules are highlighted in bold

max L	Energy RMSE (meV at. <sup>-1</sup> )			Force RMSE (eV Å <sup>-1</sup> )		
	0	1	2	0	1	2
aspirin	6.2	6.6	<b>4.9</b>	0.25	<b>0.22</b>	0.28
ethanol	12.3	<b>10.6</b>	12.2	0.49	<b>0.35</b>	0.48
malonaldehyde	<b>7.7</b>	12.3	9.2	0.28	0.33	<b>0.25</b>
naphthalene	<b>3.3</b>	4.0	3.6	0.18	0.18	<b>0.17</b>
salicylic acid	5.3	<b>4.9</b>	6.8	<b>0.22</b>	<b>0.22</b>	0.26
toluene	<b>5.6</b>	9.1	6.9	0.32	0.51	<b>0.25</b>

space that would be deemed unphysical by MACE-OFF. Uncertainty estimation of predictions made by these models could provide an answer, even partially, to this question.

In the Supplementary Information, we show two additional cases of encapsulated molecules, carbon dioxide and acetaldehyde, the heavier homologue of CH<sub>2</sub>O. 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  $L = 1$ . To test the role of equivariance, 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.

In Table 1, we compare the performance of MACE models using invariant *vs* equivariant messages with different maximum rank  $\max L \in \{0, 1, 2\}$ . We calculate the prediction errors for all relevant rMD17 molecules, using MACE models re-fitted with different  $\max L$  values. We notice that the original G0-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, and naphthalene, as well as better force predictions for salicylic acid, compared to G0-MACE-23. A similar trend is observed when comparing G0-MACE-23 to the  $\max L = 2$  MACE model. 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. A particularly notable case is the toluene trajectory, where G0-MACE-23 is the *worst*-performing model of the three, in terms of total energy and force predictions (cf. Fig. 3).

## 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 G0-MACE-23 training dataset, using the universal interface graph-pes<sup>53</sup>. Particularly, we used the SchNet<sup>54</sup>, PaiNN<sup>55</sup>, TensorNet<sup>56</sup>, and NequIP<sup>8</sup> architectures. Details about hyperparameters and validation errors on the G0-MACE-23 dataset are provided in the Supplementary Information.

Table 2 shows that G0-MACE-23, TensorNet, and NequIP generally yield low RMSE on most molecules for both energy and force predictions. For instance, NequIP achieves low energy errors on aspirin and malonaldehyde, whereas TensorNet performs best for toluene. Meanwhile, G0-MACE-23 has the best errors in force predictions for ethanol and naphthalene. These variations demonstrate that even closely related equivariant models can extract distinct mappings from the same data, influenced by subtle differences in model design and



**Table 2:** Energy and force prediction RMSE different GNN architectures trained on the GO dataset for trajectories from the revised MD17 dataset. Errors are computed with respect to the DFT level of theory of rMD17. The lowest RMSE values for each molecules are highlighted in bold

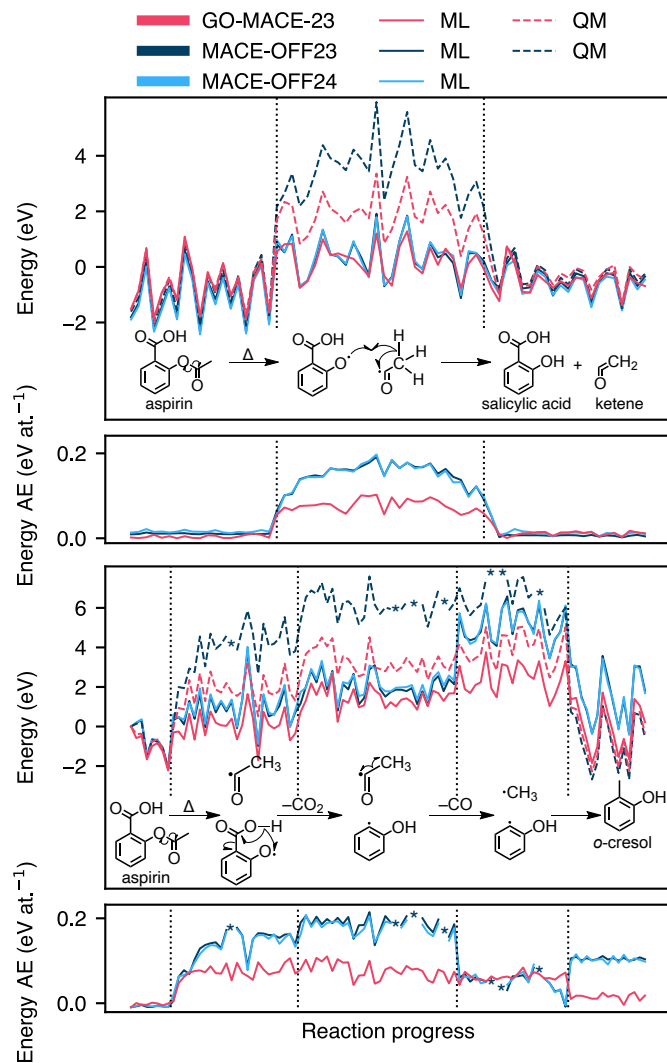
Energy RMSE (meV at. <sup>-1</sup> )					
	GO-MACE-23	SchNet	TensorNet	NequIP	PaiNN
aspirin	6.6	22.4	6.6	<b>5.7</b>	11.3
ethanol	<b>10.6</b>	33.4	17.4	17.2	27.6
malonaldehyde	12.3	38.5	10.8	<b>8.8</b>	17.2
naphthalene	4.0	9.9	5.0	<b>3.9</b>	5.8
salicylic	4.9	19.7	5.6	<b>3.9</b>	7.4
toluene	9.1	16.8	<b>8.7</b>	24.0	14.0
Force RMSE (eV Å <sup>-1</sup> )					
	GO-MACE-23	SchNet	TensorNet	NequIP	PaiNN
aspirin	<b>0.22</b>	0.86	0.38	0.31	0.57
ethanol	<b>0.35</b>	1.13	0.61	0.47	1.01
malonaldehyde	<b>0.33</b>	0.98	0.34	<b>0.33</b>	0.38
naphthalene	<b>0.18</b>	0.54	0.21	0.21	0.30
salicylic	0.22	0.42	0.24	<b>0.19</b>	0.25
toluene	0.51	0.59	<b>0.28</b>	0.38	0.32

hyperparameters.

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 Supplementary Information, GO-MACE-23 has the lowest energy validation errors on the GO dataset, yet NequIP outperforms it for several rMD17 molecules. These results underscore the need for systematic out-of-domain validation to fully assess model generalisation.

## 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<sup>57,58</sup> and in explicit solvent<sup>7</sup>. While GO-MACE-23 has been trained on various rearrangements, decarbonylation reactions, etc., it has not been explicitly trained on molecular reaction mechanisms. This makes it a partic-



**Figure 7:** 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 1,500 K. The first panel describes a reaction pathway to produce salicylic acid and ketene ( $\text{H}_2\text{CCO}$ ) 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. The asterisks correspond to failed DFT calculations after 30 self-consistent cycles.

ularly challenging and relevant “real-world” benchmark for complex chemical transformations.

We use G0-MACE-23 to run a series of MD trajectories of an aspirin molecule in a periodic simulation cell of  $30 \text{ \AA}^3$ , using the *NVT* ensemble at  $T = 1,500 \text{ K}$ . We also re-label the trajectories using the DFT reference method of G0-MACE-23, as well as using both MACE-OFF variants and their DFT reference method. In Fig. 7, we report two reaction pathways demonstrating the thermally driven decomposition of aspirin in vacuum into radical species which then recombine forming different molecules.

The upper panels of Fig. 7 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. 59. Both G0-MACE-23 and the MACE-OFF 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. This poor performance of both MLIPs is expected, as they are not explicitly trained on reaction pathways, and their underlying datasets do not include radicals or ions. In addition, these MLIPs were not able to reproduce the energy of the isolated radicals. Stocker et al.<sup>60</sup> have previously discussed the limitations of MLIPs in accurately describing chemical reactions when radicals are not explicitly incorporated in the training data.

The lower panels of Fig. 7 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-OFF family in particular overestimate the energy of the product system.

## 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-OFF, 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<sup>19</sup>, we have sampled this reactivity in high-temperature MD simulations, and a similar approach has been taken, e.g., for the bulk carbon–hydrogen<sup>61</sup> and carbon–oxygen<sup>62</sup> systems. We think that local-environment diversity will be as important as the chemical space coverage (e.g., the number of chemical species) in defining foundational models – this might include the addition of radical species (cf. Fig. 7) to the training data, either through very-high-temperature MD exploration or perhaps by explicitly involving “broken” bonds in the training protocol.

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<sup>21</sup> as well as suitable workflows and automation approaches<sup>63–66</sup>, 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.

## Data availability

Data supporting the present study are available at [GitHub link].

## 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. V.L.D. acknowledges support from 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, for which access was obtained via the UKCP consortium and funded by EPSRC grant ref EP/X035891/1.

## References

1. J. Behler, *Angewandte Chemie International Edition*, 2017, **56**, 12828–12840.

2. V. L. Deringer, M. A. Caro and G. Csányi, *Advanced Materials*, 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, *Chemical Reviews*, 2021, **121**, 10142–10186.
4. A. P. Bartók, J. Kermode, N. Bernstein and G. Csányi, *Physical Review 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, *Nature Electronics*, 2023, **6**, 746–754.
7. H. Zhang, V. Juraskova and F. Duarte, *Nature Communications*, 2024, **15**, 6114.
8. 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.
9. I. Batatia, D. P. Kovacs, G. Simm, C. Ortner and G. Csany, Advances in neural information processing systems, 2022, pp. 11423–11436.
10. I. Batatia, S. Batzner, D. P. Kovács, A. Musaelian, G. N. C. Simm, R. Drautz, C. Ortner, B. Kozinsky and G. Csányi, *Nature Machine Intelligence*, 2025, **7**, 56–67.
11. B. Deng, P. Zhong, K. Jun, J. Riebesell, K. Han, C. J. Bartel and G. Ceder, *Nature Machine Intelligence*, 2023, **5**, 1031–1041.
12. 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*, 2023, <http://arxiv.org/abs/2401.00096>, arXiv:2401.00096 [cond-mat, physics:physics].
13. A. Merchant, S. Batzner, S. S. Schoenholz, M. Aykol, G. Cheon and E. D. Cubuk, *Nature*, 2023, **624**, 80–85.
  14. 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.
  15. C. Chen and S. P. Ong, *Nature Computational Science*, 2022, **2**, 718–728.
  16. B. Focassio, L. P. M. Freitas and G. R. Schleder, *ACS Applied Materials & Interfaces*, 2024, acsami.4c03815.
  17. 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*, 2025, <http://arxiv.org/abs/2501.05211>, arXiv:2501.05211 [cond-mat].
  18. H. Kaur, F. Della Pia, I. Batatia, X. R. Advincula, B. X. Shi, J. Lan, G. Csányi, A. Michaelides and V. Kapil, *Faraday Discussions*, 2025, **256**, 120–138.
  19. Z. El-Machachi, D. Frantzov, A. Nijamudheen, T. Zarrouk, M. A. Caro and V. L. Deringer, *Angewandte Chemie International Edition*, 2024, e202410088.
  20. D. R. Dreyer, S. Park, C. W. Bielawski and R. S. Ruoff, *Chem. Soc. Rev.*, 2010, **39**, 228–240.
  21. C. Ben Mahmoud, J. L. A. Gardner and V. L. Deringer, *Nature Computational Science*, 2024, **4**, 384–387.

22. T. K. Stenczel, Z. El-Machachi, G. Liepuoniute, J. D. Morrow, A. P. Bartók, M. I. J. Probert, G. Csányi and V. L. Deringer, *The Journal of Chemical Physics*, 2023, **159**, 044803.
23. A. P. Bartók, M. C. Payne, R. Kondor and G. Csányi, *Physical Review Letters*, 2010, **104**, 136403.
24. S. J. Clark, M. D. Segall, C. J. Pickard, P. J. Hasnip, M. I. J. Probert, K. Refson and M. C. Payne, *Zeitschrift für Kristallographie - Crystalline Materials*, 2005, **220**, 567–570.
25. J. P. Perdew, K. Burke and M. Ernzerhof, *Physical Review Letters*, 1996, **77**, 3865–3868.
26. D. P. Kovács, J. H. Moore, N. J. Browning, I. Batatia, J. T. Horton, V. Kapil, W. C. Witt, I.-B. Magdău, D. J. Cole and G. Csányi, *MACE-OFF23: Transferable Machine Learning Force Fields for Organic Molecules*, 2023, <http://arxiv.org/abs/2312.15211>, arXiv:2312.15211 [physics].
27. 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.
28. P. Eastman, B. P. Pritchard, J. D. Chodera and T. E. Markland, *Journal of Chemical Theory and Computation*, 2024, **20**, 8583–8593.
29. N. Mardirossian and M. Head-Gordon, *The Journal of Chemical Physics*, 2016, **144**, 214110.
30. A. Najibi and L. Goerigk, *Journal of Chemical Theory and Computation*, 2018, **14**, 5725–5738.
31. F. Weigend and R. Ahlrichs, *Physical Chemistry Chemical Physics*, 2005, **7**, 3297.
32. D. Rappoport and F. Furche, *The Journal of Chemical Physics*, 2010, **133**, 134105.
33. A. S. Christensen and O. A. Von Lilienfeld, *Machine Learning: Science and Technology*, 2020, **1**, 045018.



34. A. Hjorth Larsen, J. Jørgen Mortensen, J. Blomqvist, I. E. Castelli, R. Christensen, M. Duřak, 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, *Journal of Physics: Condensed Matter*, 2017, **29**, 273002.
35. 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, *The Journal of Chemical Physics*, 2020, **152**, 184108.
36. D. F. Thomas Du Toit, Y. Zhou and V. L. Deringer, *Journal of Chemical Theory and Computation*, 2024, **20**, 10103–10113.
37. A. Stukowski, *Modelling and Simulation in Materials Science and Engineering*, 2010, **18**, 015012.
38. J. Behler and M. Parrinello, *Physical Review Letters*, 2007, **98**, 146401.
39. S. Chong, F. Grasselli, C. Ben Mahmoud, J. D. Morrow, V. L. Deringer and M. Ceriotti, *Journal of Chemical Theory and Computation*, 2023, **19**, 8020–8031.
40. S. Chong, F. Bigi, F. Grasselli, P. Loche, M. Kellner and M. Ceriotti, *Faraday Discussions*, 2025, **256**, 322–344.
41. A. Togo, L. Chaput, T. Tadano and I. Tanaka, *Journal of Physics: Condensed Matter*, 2023, **35**, 353001.
42. A. Togo, *Journal of the Physical Society of Japan*, 2023, **92**, 012001.

43. B. Deng, Y. Choi, P. Zhong, J. Riebesell, S. Anand, Z. Li, K. Jun, K. A. Persson and G. Ceder, *npj Computational Materials*, 2025, **11**, 9.
44. A. Barnard and G. Opletal, *Fullerene Data Set*, 2023, <https://data.csiro.au/collection/csiro%3A59022v1>.
45. V. Parasuk and J. Almlöf, *Chemical Physics Letters*, 1991, **184**, 187–190.
46. V. K. Vyas, G. R. Bacanu, M. Soundararajan, E. S. Marsden, T. Jafari, A. Shugai, M. E. Light, U. Nagel, T. Rööm, M. H. Levitt and R. J. Whitby, *Nature Communications*, 2024, **15**, 2515.
47. A. A. Popov, S. Yang and L. Dunsch, *Chemical Reviews*, 2013, **113**, 5989–6113.
48. S. Bloodworth and R. J. Whitby, *Communications Chemistry*, 2022, **5**, 121.
49. O. Carrillo-Bohorquez, A. Valdes and R. Prosmiti, *Journal of Chemical Theory and Computation*, 2021, **17**, 5839–5848.
50. 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, *Angewandte Chemie International Edition*, 2019, **58**, 5038–5043.
51. S. Stocker, J. Gasteiger, F. Becker, S. Günnemann and J. T. Margraf, *Machine Learning: Science and Technology*, 2022, **3**, 045010.
52. *Ullmann’s encyclopedia of industrial chemistry*, ed. B. Elvers and G. Bellussi, Wiley-VCH, Weinheim, 7th edn, 2011.
53. J. L. A. Gardner, *Graph PES*, 2025, <https://github.com/jla-gardner/graph-pes>.
54. K. T. Schütt, P. Kessel, M. Gastegger, K. A. Nicoli, A. Tkatchenko and K.-R. Müller, *Journal of Chemical Theory and Computation*, 2019, **15**, 448–455.
55. K. T. Schütt, O. T. Unke and M. Gastegger, *Equivariant message passing for the prediction of tensorial properties and molecular spectra*, 2021, <http://arxiv.org/abs/2102.03150>, arXiv:2102.03150 [cs].

56. G. Simeon and G. d. Fabritiis, *TensorNet: Cartesian Tensor Representations for Efficient Learning of Molecular Potentials*, 2023, <http://arxiv.org/abs/2306.06482>, arXiv:2306.06482 [cs].
57. E. Komp and S. Valteau, *Chemical Science*, 2022, **13**, 7900–7906.
58. S. Choi, *Nature Communications*, 2023, **14**, 1168.
59. D. A. Nightingale, US Pat., 1604472, 1926.
60. S. Stocker, G. Csányi, K. Reuter and J. T. Margraf, *Nature Communications*, 2020, **11**, 5505.
61. R. Ibragimova, M. S. Kuklin, T. Zarrouk and M. A. Caro, *Chemistry of Materials*, 2025, **37**, 1094–1110.
62. T. Zarrouk, R. Ibragimova, A. P. Bartók and M. A. Caro, *Journal of the American Chemical Society*, 2024, **146**, 14645–14659.
63. E. V. Podryabinkin and A. V. Shapeev, *Computational Materials Science*, 2017, **140**, 171–180.
64. T. A. Young, T. Johnston-Wood, V. L. Deringer and F. Duarte, *Chemical Science*, 2021, **12**, 10944–10955.
65. C. Van Der Oord, M. Sachs, D. P. Kovács, C. Ortner and G. Csányi, *npj Computational Materials*, 2023, **9**, 168.
66. 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, *An automated framework for exploring and learning potential-energy surfaces*, 2024, <http://arxiv.org/abs/2412.16736>, arXiv:2412.16736 [physics].