TAKE NOTE: YOUR MOLECULAR DATASET IS PROBABLY ALIGNED

Anonymous authors
Paper under double-blind review

ABSTRACT

Massive training datasets are fueling the astounding progress in molecular machine learning. Since these datasets are typically generated with computational chemistry codes which do not randomize pose, the resulting geometries are usually not randomly oriented. While cheminformaticians are well aware of this fact, it can be a real pitfall for machine learners entering the burgeoning field of molecular machine learning. We demonstrate that molecular poses in the popular datasets QM9, QMugs and OMol25 are indeed biased. While the fact can easily be overseen by visual inspection alone, we show that a simple classifier can separate original data samples from randomly rotated ones with high accuracy. Second, we validate empirically that neural networks can and do exploit the orientedness in these datasets by successfully training a model on chemical property regression using the molecular orientation as *sole* input. Third, we present visualizations of all molecular orientations and confirm that chemically similar molecules tend to have similar canonical poses. In summary, we recall and document orientational bias in the prevalent datasets that machine learners should be aware of.

1 INTRODUCTION

Machine learning has become a well established tool for designing, discovering and studying molecular systems, for instance in drug-discovery, materials science and physical chemistry. Much of the progress in the field is enabled by the curation of large-scale datasets that provide accurate molecular properties. Computational chemistry codes used in the underlying data generating processes usually do not generate molecular geometries in random orientations. At the same time, handling the arbitrariness of coordinate systems by incorporating symmetries into machine learning models has become a central theme in geometric deep learning. $\mathrm{SO}(3)$-equivariant neural networks guarantee well-defined transformation behavior under rotations. In other words, equivariant models produce consistent predictions for inputs that differ only by rotation (or equivalently, by choice of reference frame). Consequently, equivariant architectures are agnostic to the orientation of molecular geometries in ML datasets. However, most existing equivariant architectures rely on non-standard building blocks, such as specialized normalization layers, nonlinearities and tensor operations, which often are computationally demanding (Passaro & Zitnick, 2023) and can be challenging to tune in practice (Abramson et al., 2024). As a consequence, while exact equivariance is desirable in principle, softening the constraint by learning approximate symmetries or breaking built-in equivariance is a (re-)emerging pattern in molecular machine learning (Langer et al., 2024; Eissler et al., 2025). In that case, possible orientational bias in molecular datasets might affect machine learning workflows. In this paper, we show that molecules in some of the most popular molecular datasets, including QM9, QMugs and OMol25, are by default not presented in random orientations and discuss implications for machine learning practitioners.

To date, many architectures for molecular machine learning are benchmarked on the widely used QM9 dataset (Ruddigkeit et al., 2012; Ramakrishnan et al., 2014), a collection of around 134,000 small organic molecules with up to nine heavy atoms. While QM9 is the gold standard for property prediction on smaller molecules, the QMugs collection (Isert et al., 2022) comprises quantum mechanical properties for almost 2M larger drug-like molecules of up to 100 heavy atoms extracted from the ChEMBL database (Mendez et al., 2019). The OMol25 dataset (Levine et al., 2025) combines broad chemical diversity with a high level of accuracy at an unprecedented scale (100M sys-

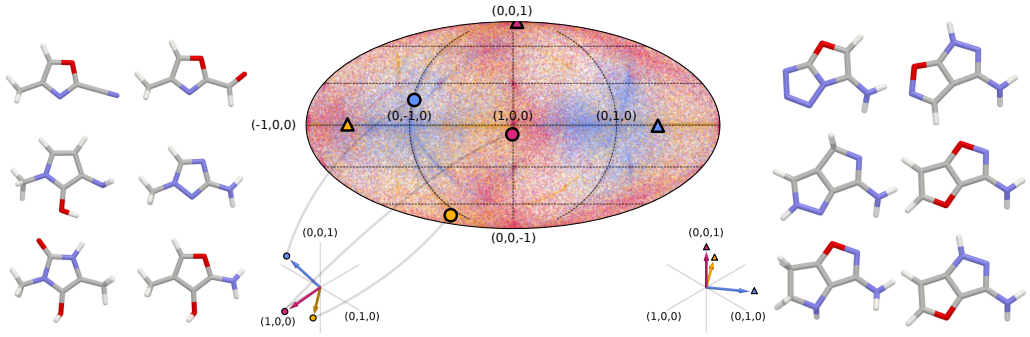


Figure 1: **Molecules in many popular molecular datasets (here: QM9) are not randomly oriented. Structurally similar molecules are oriented similarly.** 2D Visualization of normalized principal components (PCs) of all QM9 molecules reveals a clear non-uniform structure. Orientation reference frames consist of a triplet of points in the 2D area-preserving (Mollweide) projection (blue $\sim$ PC1, yellow $\sim$ PC2, magenta $\sim$ PC3). Left and right: Two groups of structurally similar molecules that share practically the same PC-orientation (left orientation $\bigcirc$, right orientation $\triangle$).

tems), comprising not only neutral organics but also biomolecules, metal complexes and electrolytes consisting of 83 different elements.

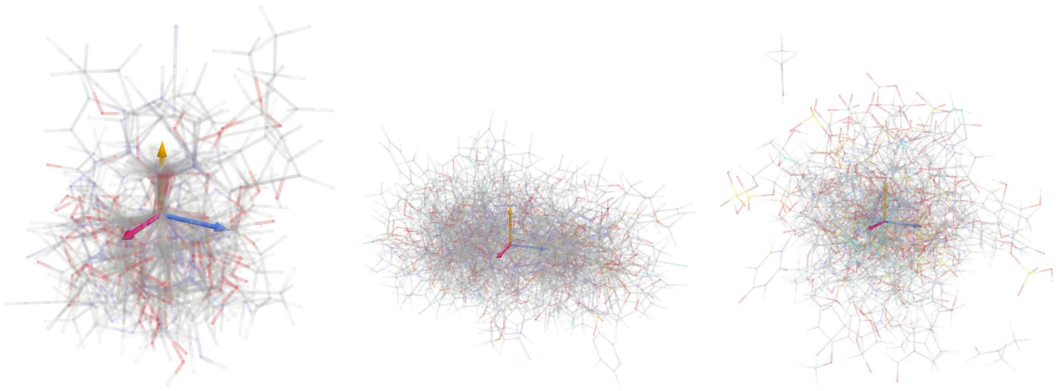
We want to raise awareness of the fact that molecules in ML datasets are not oriented randomly for two crucial reasons. First, non-equivariant architectures trained on these datasets must employ explicit data augmentation; otherwise, their test performance will degrade significantly when evaluated on randomly oriented molecules. Second, architectures that introduce symmetry breaking or are only approximately equivariant (Liao et al., 2024; Wang et al., 2023; Pertigkiozoglou et al., 2024; Wang et al., 2022; Kaba & Ravanbakhsh, 2024; Lawrence et al., 2025) might artificially inflate their performance by exploiting the extrinsic canonicalization of molecule poses. In addition, when molecular properties such as the ground state electron density are evaluated on a grid that is not spherically symmetric (Brockherde et al., 2017; Bogojeski et al., 2020; Jørgensen & Bhowmik, 2022; Li et al., 2025), the canonical orientation may introduce a systematic bias in the grid orientation even when equivariant architectures are used later in the prediction.

In this paper, we make the following contributions: we demonstrate, using QM9, QMugs and OMol25 as prominent examples, that molecules in many popular ML datasets are not randomly oriented by training a simple classifier that distinguishes between rotated and unrotated samples with very high accuracy. We show that the accuracy remains high even when the default atom positions are perturbed with substantial noise and random rotations up to 90°. Further, we demonstrate that neural networks can leverage the canonical orientation to achieve artificially high accuracy in an extreme scenario: using only the normalized principal components of atom positions as input, we regress molecular properties and observe performance on the three standard datasets that exceeds the best possible accuracy expected for randomly oriented data. Lastly, we visualize the orientations of all molecules in these datasets and show that chemically similar molecules tend to be oriented similarly (see Fig. 1).

2 Background and Related Work

Formally, a function $\varphi : V \to W$, mapping between vector spaces V and W, is *equivariant* under a group G if $\rho_{\text{out}}(g)\varphi(x) = \varphi(\rho_{\text{in}}(g)x)$ for all $g \in G$ and $x \in V$. Here, $\rho_{\text{in}}, \rho_{\text{out}}$ are group representations on V and W, defining how elements of the group G act on elements from the vector spaces respectively. In diagrammatic form, equivariance means that the following commutes:

$$\begin{array}{ccc} x & \xrightarrow{\varphi} & \varphi(x) \\ {\scriptstyle \rho_{\text{in}}(g)}\downarrow & & \downarrow{\scriptstyle \rho_{\text{out}}(g)} \\ x' & \xrightarrow{\varphi} & \varphi(x') \end{array}$$



(a) QM9 (b) QMugs (c) OMol25 (centered)

Figure 2: **Layering of 100 random molecular geometries from QM9, QMugs and OMol25 respectively.** QM9 shows strong alignment of the edge adjacent to the origin with the Cartesian y-axis (yellow). Bulges of QMugs and OMol25 molecules exhibit less structure, but are not spherically symmetric.

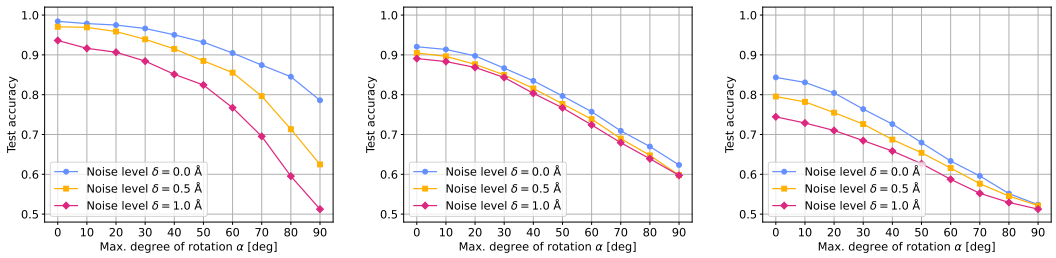
Several approaches for incorporating exact $\mathrm{SO}(3)$-equivariance into neural networks exist. Most famously, tensor field networks use tensorial features in all hidden layers to maintain a well-defined transformation behavior throughout the network (Thomas et al., 2018; Geiger & Smidt, 2022; Batatia et al., 2022; Liao et al., 2024; Aykent & Xia, 2025). Similarly, specialized architectures exist to use elements of the projective geometric (or Clifford) algebra to achieve Euclidean rotation and translation equivariance (Brehmer et al., 2023; Ruhe et al., 2023). On the contrary, canonicalization approaches avoid the need for specialized architectural building blocks and use canonical reference frames to guarantee exact equivariance (Puny et al., 2021; Pozdnyakov & Ceriotti, 2024; Spinner et al., 2025; Lippmann et al., 2025). Data augmentation offers a simple and practical alternative where equivariance is learned approximately by presenting the network with randomly rotated inputs (and targets). It is an open research question when built-in equivariance is favorable over data augmentation and often a fair comparison is non-trivial (Brehmer et al., 2024). Some evidence points to the superiority of non-equivariant architectures (Langer et al., 2024; Lippmann et al., 2025), possibly due to their less limited design space.

3 Investigating orientations in molecular datasets: methods, results, and implications

Clearly, the first step that comes to mind when investigating the orientations of molecular geometries is to visually inspect the 3D geometries for obvious alignment. Figure 2 shows 100 randomly sampled molecular geometries from each dataset. For QM9 clear structure is visible. Most strikingly, the first edge (adjacent to the origin) almost perfectly aligns with the Cartesian y-axis. The original QM9 paper (Ramakrishnan et al., 2014) invoked the cheminormatics tool Corina (Version 3.491 2013) (Sadowski & Gasteiger, 1993) to generate 3D structures from a SMILES string representation. The geometries were then relaxed using Kohn-Sham DFT calculations at the B3LYP/6-31G(2df,p) level. The Corina algorithm (closed-source) most likely introduces the alignment with the y-axis, while the subsequent geometry relaxation softens the strict alignment. For QMugs and OMol25 no similarly distinct structure is visible regarding the orientation, but the central bulge of molecules is clearly not spherically symmetric, hinting at a systematic orientation bias.

3.1 Learned detection of default orientations

To validate empirically that molecules in ML datasets exhibit pose bias, we have trained a simple message passing network to distinguish between molecules in their original ("canonical") orientation and ones that have been randomly rotated. If the dataset were truly orientation invariant, such a classification task would be impossible beyond random guessing.



(a) QM9 (b) QMugs (c) OMol25

Figure 3: **Canonical orientations in QM9, QMugs and OMol25 are highly consistent and detectable.** A simple geometric message passing network accurately discerns canonical samples from randomly rotated ones, even if the atom positions are previously perturbed and the molecule randomly rotated by an angle up to α.

For each sample in the dataset[1], we randomly decide whether to apply a global rotation, sampling a rotation matrix uniformly from $\mathrm{SO}(3)$. To ensure that the learned detection is not just based on a simple geometric pattern (such as a particular edge being aligned with a coordinate axis, cf. Fig. 2), we apply Gaussian noise (with standard deviation δ up to 1 Å) to the default atomic positions as well as random rotations up to a maximum angle α (prior to the uniformly sampled rotation that should be detected by the network). A noise level of $\delta = 1$ Å is quite substantial, considering that the length of a carbon-carbon single bond is 1.5 Å, and shorter for a double or triple bond. The network is then trained to minimize a binary cross-entropy loss, predicting whether the input molecule has been randomly rotated or is in its (perturbed) default pose. We employ a straightforward point cloud architecture, consisting of three layers of message passing:

$$f_i^{(k+1)} = \bigoplus_{j \in \mathcal{N}(i)} \mathrm{MLP}(f_j^{(k)}, \mathrm{emb}(x_i - x_j)), \tag{1}$$

where $f_i^{(k)}$ denotes the feature vector of atom i in layer k, x_i its position and $\mathcal{N}(i)$ its neighborhood (defined by a radial cutoff of 10 Å). This network is neither rotation equivariant, nor invariant, but rotation dependent by design. During message passing the angular and radial part of the relative distance vectors $x_i - x_j$ are embedded using Gaussian radial basis functions. Prior to these message passing layers, we combine a learned embedding of the neighbor geometry of every atom with an embedding of the atom type to create the initial node features $f^{(0)}$. We use the same architecture for all three datasets, see App. C for details.

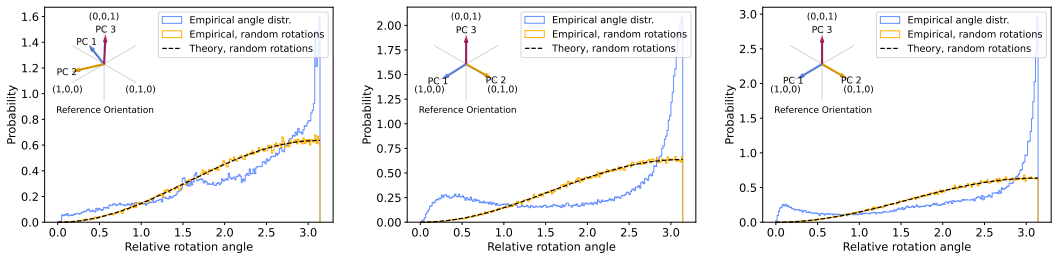
The results, summarized in Fig. 3, reveal that even a simple classifier can discern random and canonical poses with very test high accuracy, even when the default atom positions are substantially perturbed. This provides clear evidence that molecules in the considered datasets are not randomly oriented, and that the canonical poses are highly consistent and detectable.

3.2 Quantitative orientation analysis

To systematically study the orientation of molecules in the respective datasets, we devise a mapping Ω from the set of molecular geometries $\mathcal{M} = \{\{(z_a, x_a)\}_{a \in A}\}$ to the set of orientations $\mathrm{SO}(3)$. For each molecule, comprised of atoms A with charges z_a and positions x_a, it outputs an orientation. Clearly, for general molecules no canonical orientation function exists. However, it is very reasonable to require that orientations predicted for two molecules that differ just by a rotation must be consistent. Intuitively, the orientation $\Omega(M)$ can be thought of as a coordinate frame attached to the molecule that should rotate along when the molecule is rotated. Formally, we thus demand that the mapping $\Omega : \mathcal{M} \to \mathrm{SO}(3)$ is equivariant under rotations R applied to the molecular geometry:

$$\Omega(R\mathcal{M}) := \Omega(\{(z_a, Rx_a)\}_{a \in A}) = \Omega(\{(z_a, x_a)\}_{a \in A})R^{\mathrm{T}} = \Omega(\mathcal{M})R^{\mathrm{T}}. \tag{2}$$

[1]For all our experiments we use the QM9 dataset readily available in PytorchGeometric (Fey & Lenssen, 2019), QMugs from `https://doi.org/10.3929/ethz-b-000482129` and OMol25 available through the `fairchem` repository (`https://github.com/facebookresearch/fairchem`).



(a) QM9 (b) QMugs (c) OMol25

Figure 4: **Quantitative comparison of canonical orientations vs. uniform random orientations.** The empirical distribution of relative angles between all orientations in the respective dataset and the most common reference orientation differs significantly from the same distribution for a randomly rotated dataset and from the theoretical expectation for uniform random orientations. Histograms for QMugs and OMol25 are based on 130,000 randomly sampled molecules.

To understand the constraint imposed by Eq. (2), we view $\Omega(M) \in \mathrm{SO}(3)$ as a collection of three row vectors $e_i \in \mathbb{R}^3, i = 1, 2, 3$, i.e. $\Omega(M) = (e_1, e_2, e_3)^{\mathrm{T}}$. The e_i are precisely the basis vectors of the reference frame that rotates along with the molecular geometry M, i.e.

$$M' = RM \quad \rightarrow \quad e'_i = Re_i \quad \text{or equivalently:}$$
$$\Omega(RM) = (Re_1, Re_2, Re_3)^{\mathrm{T}} = (e_1, e_2, e_3)^{\mathrm{T}} R^T = \Omega(M) R^T,$$

as demanded per Eq. (2). A simple such orientation function is obtained by choosing the basis vectors e_i to be the normalized principal components of the centered atom position. To account for the fact that the sign of eigenvectors is ambiguous, we choose the sign of the first two principal components such that they point in the direction of the largest absolute projection, that is, we orient them to satisfy

$$\max_{a \in A} |x_a \cdot e_i| = \max_{a \in A} x_a \cdot e_i \quad \text{for } i = 1, 2. \tag{3}$$

The orientation of the third principal component e_3 is fixed by the constraint that $\det((e_1, e_2, e_3)^{\mathrm{T}}) = 1$. While one may also use a weighted covariance matrix based on atomic masses (yielding the same eigenvectors as the moment of inertia tensor), we opted for the simpler unweighted version which is robust against exchange of atom types.

To compare the orientations of different molecules we define a distance measure based on the following fact: Any rotation can be described by a rotation axis (vectors pointing along this axis are left invariant) and the rotation angle, specifying how much to rotate around that axis. This angle is given by $\mathrm{acos}((\mathrm{tr}(R) - 1)/2)$, see App. A. Then, for two rotations $R_1, R_2 \in \mathrm{SO}(3)$, we use the rotation angle of the relative rotation matrix $R_1^T R_2$ as a measure of distance:

$$\theta(R_1, R_2) = \mathrm{acos}\left(\frac{\mathrm{tr}(R_1^{\mathrm{T}} R_2) - 1}{2}\right). \tag{4}$$

Intuitively, $\theta(R_1, R_2)$ is the angle of the rotation that changes from reference frame R_1 to R_2 and vice versa. The properties of the trace imply that $\theta(R_1, R_2) = \theta(R_2, R_1) = \theta(R_1^{\mathrm{T}}, R_2^{\mathrm{T}})$. If the orientations of molecules in a dataset $\mathcal{D}$ were truly random, the set of all orientations would sample $\mathrm{SO}(3)$ uniformly. In that case, for any given reference orientation $\tilde{R}$ the empirical distribution of distances $\{\theta(\tilde{R}, \Omega(M_i)) \mid i \in \mathcal{D}\}$ should approximate the following distribution (see App. B):

$$p(\theta) = \frac{2}{\pi} \sin(\theta/2)^2, \quad \theta \in [0, \pi]. \tag{5}$$

To characterize the deviation from this distribution, for each dataset, we propose the following: First, we compute the full distance matrix $\Theta_{ij} = \theta(\Omega(M_i), \Omega(M_j))$ for all $M_i, M_j \in \mathcal{D}$. Afterwards, we apply a 1D Gaussian kernel[2] $k(\theta|\mu, \sigma^2)$ centered at $\theta = 0$ to each entry in Θ and sum over the rows

[2]Given that we work with small σ, we approximate the von Mises distribution on the circle with a Gaussian kernel.

Table 1: **Molecular property prediction from a molecule's orientation alone.** A simple MLP trained on the canonical datasets to regress molecular properties given *only* normalized principal components of atom positions as input significantly outperforms the test performance achievable with uninformative input features (averages over 5 runs each).

Dataset	Property	Random orientation	MSE of mean (test)	MSE of MLP (test)
QM9	ϵ_{LUMO} [eV]	no	1.6355	**1.4237 ± 0.0048**
		yes	1.6355	1.6367 ± 0.0001
QM9	ZPVE [eV]	no	0.8107	**0.6204 ± 0.0011**
		yes	0.8107	0.8111 ± 0.0001
QM9	c_V $\left[\frac{\text{cal}}{\text{mol K}}\right]$	no	16.169	**13.814 ± 0.083**
		yes	16.169	16.173 ± 0.001
QMugs	U_{RT} $[E_h]$	no	890.54	**843.48 ± 0.09**
		yes	890.54	890.55 ± 0.02
QMugs	$\hat{V}_{ee}$ $[E_h]$	no	$4{,}894.0 \times 10^3$	$(\mathbf{4{,}596.6 \pm 0.7}) \times \mathbf{10^3}$
		yes	$4{,}894.0 \times 10^3$	$(4{,}894.6 \pm 0.2) \times 10^3$
OMol25	E_{tot} [eV]	no	$14{,}394.3 \times 10^6$	$(\mathbf{13{,}689.1 \pm 1.7}) \times \mathbf{10^6}$
		yes	$14{,}394.3 \times 10^6$	$(14{,}398.7 \pm 0.3) \times 10^6$

(or columns) of the resulting matrix to obtain the following kernel density estimate at $\theta = 0$:

$$\text{KDE}(\theta = 0|M_i) = \frac{1}{|\mathcal{D}|} \sum_{j \in \mathcal{D}} k(0|\Theta_{ij}, \sigma^2) = \frac{1}{|\mathcal{D}|} \sum_{j \in \mathcal{D}} k(\Theta_{ij}|0, \sigma^2), \tag{6}$$

$$\text{with kernel } k(\theta|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(\theta - \mu)^2}{2\sigma^2}\right). \tag{7}$$

The most prominent orientation in the dataset is then given by

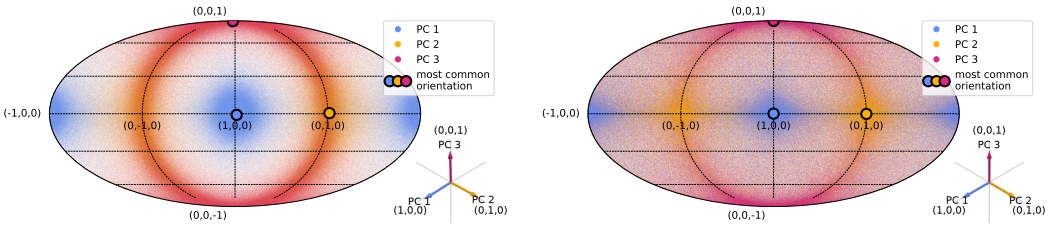
$$\Omega(M_{i^*}) \text{ with } i^* = \arg\max_{i \in \mathcal{D}} \text{KDE}(\theta = 0|M_i). \tag{8}$$

Empirically, we find that $\Omega(M_{i^*})$ is fairly robust against the choice of σ. Using $\Omega(M_{i^*})$ as reference rotation, the distribution of relative rotation angles in row Θ_{i^*} differs strongly from the theoretically expected distribution for uniformly sampled orientations given by Eq. (5), see Fig. 4. For both QMugs and OMol25 the most common principal component directions (using $\sigma = 0.5$) indeed align closely with the standard Cartesian coordinate frame ($e_1 = (1, 0, 0)^{\text{T}}$, $e_2 = (0, 1, 0)^{\text{T}}$, $e_3 = (0, 0, 1)^{\text{T}}$), see also Fig. 5. For QM9 the most common principal component direction is more ambiguous (compare Figs. 1 and 5). Our analysis has identified the most common principal component orientation in QM9 to lie around $e_1 = (-0.34, -0.94, 0)^{\text{T}}, e_2 = (0.94, -0.34, 0)^{\text{T}}, e_3 = (0, 0, 1)^{\text{T}}$. All three datasets contain significantly more molecules in the most common orientation than expected for a uniform distribution of orientations as well as significantly more orientations that differ by a rotation of 180°, forming the peak at $\theta = \pi$. The latter likely correspond to the same principal components, but with orientations flipped relative to the reference.

3.3 Exploiting the canonical orientation for property prediction

While our previous investigations demonstrate that molecules in QM9, QMugs and OMol25 are not randomly oriented, it does not yet address possible impact on the performance of machine learning models. In the following, we will investigate whether neural networks can exploit the "canonical" orientation of molecules in typical machine learning tasks such as molecular property prediction.

We here consider an extreme scenario: Can a network learn anything about molecular properties when presented *with the orientation of the molecule alone*, without any information about the molecule's constitution or geometry? In the following, we investigate the regression performance of a model which receives *only* normalized principal components of the atom positions as input features (cf. Sec. 3.2), once in canonical pose and once after random rotation. For a fair comparison,



(a) QMugs (b) OMol25

Figure 5: **2D visualization of normalized principal components of all molecules in QMugs and 10^6 random samples from OMol25.** The equal-area Mollweide projection of the three principal axes reveals that the most common principal axes orientation aligns with the standard Cartesian coordinate system. Orientation frames consist of a triplet of points in the 2D projection (blue $\sim$ PC1, yellow $\sim$ PC2, magenta $\sim$ PC3).

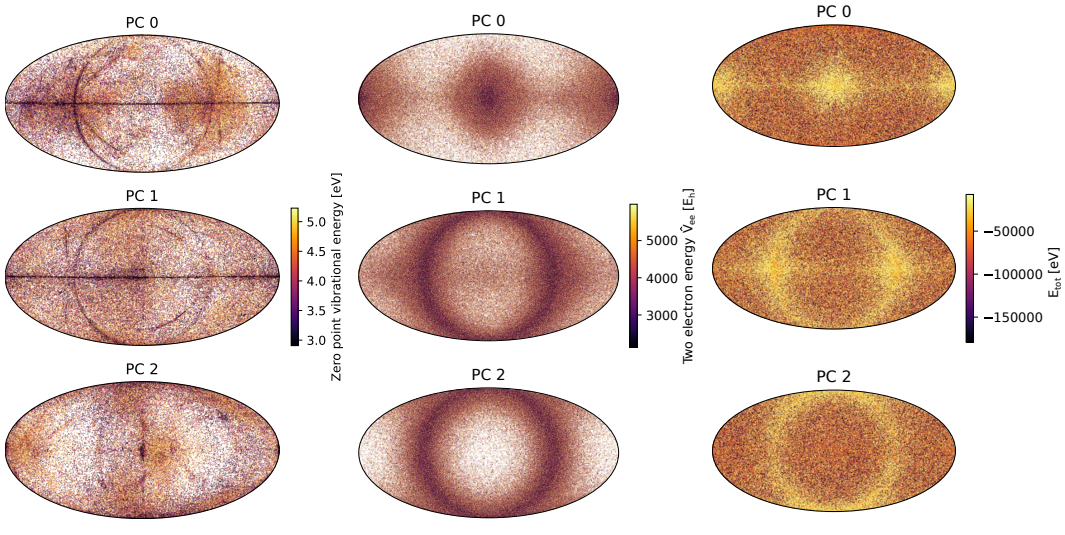
we here use a deterministic rotation conditioned on the molecule index so that the same sample will be transformed with the same rotation when revisited during training. This is equivalent to using a version of the dataset that has been rotated once prior to training. For molecules in random orientations the normalized principal components do not contain any chemically relevant information. Using MSE (mean squared error) loss, the best performance any model can achieve in this case is by approximating the mean of the target feature, since

$$\text{mean}(\{y_i\}) = \arg\min_x \sum_i (x - y_i)^2. \tag{9}$$

Therefore, if the trained model achieves a significantly better MSE on the test set than the mean of the test targets does, the model has learned a non-trivial pattern from the normalized principal components. This indicates that chemically similar molecules, by default, tend to have similar orientations. The results, presented in Tab. 1, reveal precisely that. Indeed, simple MLPs trained and tested on the canonical datasets significantly outperform the theoretically best possible results, while models trained on the transformed datasets do not, as expected. The chemical properties used for regression were chosen based on the amount of structure in the visualization of all molecular orientations when using the respective features as heat map (see Sec. 3.4). This is empirical proof that the canonical orientation alone holds information about a molecule's properties, which could be exploited by non-equivariant models, potentially leading to artificially inflated performance metrics in the absence of a randomly oriented test set.

3.4 Visualization of molecular orientations

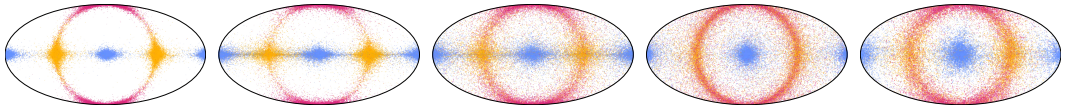
To further investigate whether chemically similar molecules in molecular datasets tend to be similarly oriented, we have devised the following visualization strategy: for every molecule M, we compute the normalized principal components $e_1, e_2, e_3 \in \mathbb{R}^3$ and combine them into a rotation matrix $\Omega(M) = (e_1, e_2, e_3)^{\mathrm{T}}$, as described in Sec. 3.2. The $e_1, e_2, e_3 \in \mathbb{S}^2$ are projected using the equal-area Mollweide projection, and are each visualized in a different color (blue $\sim$ PC1, yellow $\sim$ PC2, magenta $\sim$ PC3), such that an orientation frame $\Omega(M)$ consist of a triplet of orthogonal points in the projection, see Figs. 1 and 5. Notably, using an equal-area projection, a perfectly uniform distribution would also be perceived as uniform distribution. Contrarily, the visualizations show a clear pattern and illustrate the previous finding (Sec. 3.2) that the most common principal component orientation in the QMugs and OMol25 dataset aligns with the standard Cartesian coordinate system. Similarly, Fig. 1 reveals the structure in all principal component orientations for QM9. Based on the distance measure defined in Eq. (4), we have cherry-picked two groups of QM9 molecules of practically identical orientation that arguably have very similar chemical constitution and geometry. These examples nicely illustrate that, as one of the signatures of the cheminformatics codes used in the data generating process, chemically similar molecules tend to have similar canonical orientations. In Fig. 6 we show one selected chemical property for each dataset as heat map in the projections instead of using colors to differentiate the principal components. The three visualizations visibly confirm the correlation between chemical properties of molecules and their default orientations.



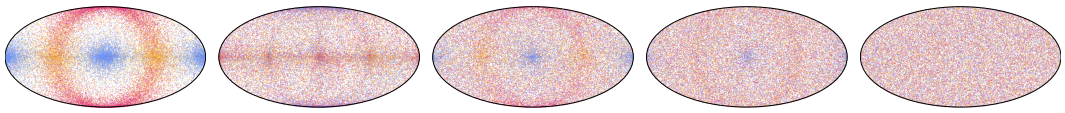
(a) QM9 (b) QMugs (c) OMol25

Figure 6: **Default orientations of molecules are correlated with chemical properties.** Plots are colored by chemical properties in 2D projections of the three normalized principal components to reveal substructure within the distributions. Non-equivariant architectures may exploit such correlation to artificially inflate performance.

Lastly, we show that the orientation distribution in the large collection of OMol25 differ strongly between different subsets of the dataset (subsets are based on the "data_id" field of OMol25 samples), see Fig. 7. The plots demonstrate that while molecules in some subsets are visibly strongly aligned with the standard Cartesian axes, for other subsets the distribution of orientation is (almost) perceptually uniform (e.g. for the SPICE2 dataset (Eastman et al., 2023), Fig. 7(j)). In particular small biases as in the Biomolecules subset (Fig. 7(i)) may be easily overlooked, highlighting the need for random rotations even in the absence of an obvious alignment.



(a) GEOM (b) ANI-2X (c) Reactivity (d) Orbnet Denali (e) Transition-1X



(f) RGD1 (g) Metal Complexes (h) Electrolytes (i) Biomolecules (j) SPICE2

Figure 7: **Visualization of molecular orientation in OMol25 subsets.** Molecules in some subsets display a strong alignment of principal components (PCs) with the standard Cartesian coordinate system (a,b,c,d,e,f). For others, orientations are more uniformally distributed (g,h,i,j). PCs are projected using the equal-area Mollweide projection and colored as in Fig. 5.

4 Conclusion

We have demonstrated in various ways that the default orientations of molecules in some of the most popular molecular datsets (QM9, QMugs, OMol25) are far from random, and that the alignment of chemically similar molecules can in principle be exploited by machine learning models. We are not aware of prior work that describes the orientation of molecules in these datasets and believe that these orientation biases can be easily overlooked. Our experiments highlight the degree

of orientational bias across different datasets and estimate the dominant orientations. Given that generally agreed-upon canonical orientations do not—and probably cannot—exist, the presence of orientational bias is not a deficit of these important community resources. However, for researchers entering the field of molecular machine learning the assumption that molecule poses are fully random can be a significant source of error. Based on our findings, we thus recommend the following best practices for rigorous evaluation and development of molecular machine learning models: First, it is essential to evaluate equivariant models on randomly oriented test sets as a sanity check for true equivariance. This ensures that any claimed equivariant behavior is genuine and bug-free. Secondly, for non-equivariant or only approximately equivariant models, it is crucial to use data augmentation during training to prevent overfitting to any canonical orientations and to provide a more realistic assessment of model generalization. It is quite likely that other data generating processes too introduce preferred orientations for geometric data. Therefore, we recommend, when in doubt, to follow the same best practices also for other geometric datasets.

At the same time, our results highlight the potential benefits of leveraging a well-defined canonical pose, as explored in recent work (e.g. by Baker et al. (2024)). In scenarios where a meaningful canonicalization is available and justified by the application, it can be advantageous to incorporate this information explicitly.

REFERENCES

Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J Ballard, Joshua Bambrick, et al. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, 630(8016):493–500, 2024.

Sarp Aykent and Tian Xia. Gotennet: Rethinking efficient 3d equivariant graph neural networks. In *The Thirteenth International Conference on Learning Representations*, 2025.

Justin Baker, Shih-Hsin Wang, Tommaso de Fernex, and Bao Wang. An explicit frame construction for normalizing 3D point clouds. In *Forty-first International Conference on Machine Learning*, 2024.

Ilyes Batatia, David P Kovacs, Gregor Simm, Christoph Ortner, and Gábor Csányi. MACE: Higher order equivariant message passing neural networks for fast and accurate force fields. *Advances in Neural Information Processing Systems*, 35:11423–11436, 2022.

Mihail Bogojeski, Leslie Vogt-Maranto, Mark E. Tuckerman, Klaus-Robert Müller, and Kieron Burke. Quantum chemical accuracy from density functional approximations via machine learning. *Nature Communications*, 11(1):5223, October 2020. ISSN 2041-1723. doi: 10.1038/s41467-020-19093-1. URL https://www.nature.com/articles/s41467-020-19093-1. Publisher: Nature Publishing Group.

Johann Brehmer, Pim de Haan, Sönke Behrends, and Taco S Cohen. Geometric algebra transformer. In *Advances in Neural Information Processing Systems*, volume 36, pp. 35472–35496, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/6f6dd92b03ff9be7468a6104611c9187-Paper-Conference.pdf.

Johann Brehmer, Sönke Behrends, Pim De Haan, and Taco Cohen. Does equivariance matter at scale? *arXiv preprint arXiv:2410.23179*, 2024.

Felix Brockherde, Leslie Vogt, Li Li, Mark E. Tuckerman, Kieron Burke, and Klaus-Robert Müller. Bypassing the Kohn-Sham equations with machine learning. *Nature Communications*, 8(1):872, October 2017. ISSN 2041-1723. doi: 10.1038/s41467-017-00839-3. URL https://www.nature.com/articles/s41467-017-00839-3. Publisher: Nature Publishing Group.

Peter Eastman, Pavan Kumar Behara, David L Dotson, Raimondas Galvelis, John E Herr, Josh T Horton, Yuezhi Mao, John D Chodera, Benjamin P Pritchard, Yuanqing Wang, et al. Spice, a dataset of drug-like molecules and peptides for training machine learning potentials. *Scientific Data*, 10(1):11, 2023.

Max Eissler, Tim Korjakow, Stefan Ganscha, Oliver T Unke, Klaus-Robert MÃžller, and Stefan Gugler. How simple can you go? an off-the-shelf transformer approach to molecular dynamics. *arXiv preprint arXiv:2503.01431*, 2025.

Matthias Fey and Jan E. Lenssen. Fast graph representation learning with PyTorch Geometric. In *ICLR Workshop on Representation Learning on Graphs and Manifolds*, 2019.

Mario Geiger and Tess Smidt. e3nn: Euclidean neural networks. *arXiv preprint arXiv:2207.09453*, 2022.

Clemens Isert, Kenneth Atz, José Jiménez-Luna, and Gisbert Schneider. Qmugs, quantum mechanical properties of drug-like molecules. *Scientific Data*, 9(1):273, 2022.

Peter Bjørn Jørgensen and Arghya Bhowmik. Equivariant graph neural networks for fast electron density estimation of molecules, liquids, and solids. *npj Computational Materials*, 8(1):183, 2022.

Sékou-Oumar Kaba and Siamak Ravanbakhsh. Symmetry breaking and equivariant neural networks, 2024. URL https://arxiv.org/abs/2312.09016.

Marcel F Langer, Sergey N Pozdnyakov, and Michele Ceriotti. Probing the effects of broken symmetries in machine learning. *arXiv preprint arXiv:2406.17747*, 2024.

Hannah Lawrence, Vasco Portilheiro, Yan Zhang, and Sékou-Oumar Kaba. Improving equivariant networks with probabilistic symmetry breaking, 2025. URL https://arxiv.org/abs/2503.21985.

Daniel S Levine, Muhammed Shuaibi, Evan Walter Clark Spotte-Smith, Michael G Taylor, Muhammad R Hasyim, Kyle Michel, Ilyes Batatia, Gábor Csányi, Misko Dzamba, Peter Eastman, et al. The open molecules 2025 (omol25) dataset, evaluations, and models. *arXiv preprint arXiv:2505.08762*, 2025.

Chenghan Li, Or Sharir, Shunyue Yuan, and Garnet Kin-Lic Chan. Image super-resolution inspired electron density prediction. *Nature Communications*, 16(1):4811, 2025.

Yi-Lun Liao, Brandon M Wood, Abhishek Das, and Tess Smidt. EquiformerV2: Improved equivariant transformer for scaling to higher-degree representations. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=mCOBKZmrzD.

Peter Lippmann, Gerrit Gerhartz, Roman Remme, and Fred A Hamprecht. Beyond canonicalization: How tensorial messages improve equivariant message passing. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Representation Learning*, volume 2025, pp. 88067–88087, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/db7534a06ace69f4ec95bc89e91d5dbb-Paper-Conference.pdf.

David Mendez, Anna Gaulton, A Patrícia Bento, Jon Chambers, Marleen De Veij, Eloy Félix, María Paula Magariños, Juan F Mosquera, Prudence Mutowo, Michał Nowotka, et al. Chembl: towards direct deposition of bioassay data. *Nucleic acids research*, 47(D1):D930–D940, 2019.

Saro Passaro and C Lawrence Zitnick. Reducing SO(3) convolutions to SO(2) for efficient equivariant gnns. In *International Conference on Machine Learning*, pp. 27420–27438. PMLR, 2023.

Stefanos Pertigkiozoglou, Evangelos Chatzipantazis, Shubhendu Trivedi, and Kostas Daniilidis. Improving equivariant model training via constraint relaxation. *Advances in Neural Information Processing Systems*, 37:83497–83520, 2024.

Sergey Pozdnyakov and Michele Ceriotti. Smooth, exact rotational symmetrization for deep learning on point clouds. *Advances in Neural Information Processing Systems*, 36, 2024.

Omri Puny, Matan Atzmon, Heli Ben-Hamu, Ishan Misra, Aditya Grover, Edward J Smith, and Yaron Lipman. Frame averaging for invariant and equivariant network design. *arXiv preprint arXiv:2110.03336*, 2021.

Raghunathan Ramakrishnan, Pavlo O Dral, Matthias Rupp, and O Anatole von Lilienfeld. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific Data*, 1, 2014.

Lars Ruddigkeit, Ruud Van Deursen, Lorenz C Blum, and Jean-Louis Reymond. Enumeration of 166 billion organic small molecules in the chemical universe database gdb-17. *Journal of chemical information and modeling*, 52(11):2864–2875, 2012.

David Ruhe, Johannes Brandstetter, and Patrick Forré. Clifford group equivariant neural networks. In *Advances in Neural Information Processing Systems*, volume 36, pp. 62922–62990, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/c6e0125e14ea3d1a3de3c33fd2d49fc4-Paper-Conference.pdf.

Jens Sadowski and Johann Gasteiger. From atoms and bonds to three-dimensional atomic coordinates: automatic model builders. *Chemical Reviews*, 93(7):2567–2581, 1993.

Jonas Spinner, Luigi Favaro, Peter Lippmann, Sebastian Pitz, Gerrit Gerhartz, Tilman Plehn, and Fred A Hamprecht. Lorentz local canonicalization: How to make any network lorentz-equivariant. *arXiv preprint arXiv:2505.20280*, 2025.

Nathaniel Thomas, Tess Smidt, Steven Kearnes, Lusann Yang, Li Li, Kai Kohlhoff, and Patrick Riley. Tensor field networks: Rotation-and translation-equivariant neural networks for 3d point clouds. *arXiv preprint arXiv:1802.08219*, 2018.

Rui Wang, Robin Walters, and Rose Yu. Approximately equivariant networks for imperfectly symmetric dynamics. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 23078–23091. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/wang22aa.html.

Rui Wang, Elyssa Hofgard, Han Gao, Robin Walters, and Tess E Smidt. Discovering symmetry breaking in physical systems with relaxed group convolution. *arXiv preprint arXiv:2310.02299*, 2023.

A Derivation of the rotation angle of a given rotation matrix

Let $R \in SO(3)$ be a rotation matrix. Since R is real orthogonal, all its eigenvalues lie on the unit circle and complex ones occur in conjugate pairs. With $\det R = 1$, the three eigenvalues must be $\{1, e^{i\varphi}, e^{-i\varphi}\}$ for some $\varphi \in [0, \pi]$. Let u be a unit eigenvector with $Ru = u$ (the rotation axis). Now, let us extend u to an orthonormal basis $\{e_1, e_2, u\}$ with appropriate basis vectors e_1 and e_2. In this basis R leaves $\mathrm{span}\{e_1, e_2\}$ invariant and acts on the corresponding subspace as a 2×2 planar rotation by θ. Hence R is (by change of basis) orthogonally similar to

$$\begin{pmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Since the trace is invariant under similarity, we have

$$\mathrm{tr}(R) = \cos\theta + \cos\theta + 1 = 1 + 2\cos\theta, \tag{10}$$

which yields the rotation angle of R

$$\theta = \arccos\left(\frac{\mathrm{tr}(R) - 1}{2}\right) \in [0, \pi]. \tag{11}$$

Furthermore, the fact that the trace of R is also given by the sum its eigenvalues $\mathrm{tr}(R) = 1 + e^{i\varphi} + e^{-i\varphi} = 1 + 2\cos(\varphi)$ reveals that $\varphi = \theta$.

B Derivation of angle distribution for uniformly sampled rotations

Let $R \in \mathrm{SO}(3)$ be Haar–uniform, i.e. drawn from the unique probability measure on the rotation group that is invariant under multiplying by any fixed rotation on the left or right. Further, we identify each $R \in \mathrm{SO}(3)$ with a unit quaternion $q = (w, \vec{v}) \in \mathbb{S}^3 \subset \mathbb{R}^4$ modulo the antipodal map $q \sim -q$. The rotation angle $\theta \in [0, \pi]$ of R is related to q by

$$\theta \;=\; 2\arccos\big(|w|\big). \tag{12}$$

Further, the rotation axis $\vec{n}$ of R is related to $\vec{v}$ by $\vec{n} = \vec{v}/\|\vec{v}\|_2$. Now, if we parametrize the 3-sphere by $q = (\cos\chi,\ \sin\chi\,\vec{n})$ with $\vec{n} \in \mathbb{S}^2$ and $\chi \in [0, \pi]$ the polar angle as measured from the "north pole" in the w-direction, the uniform surface element on $\mathbb{S}^3$ factorizes as

$$\mathrm{d}\sigma_{\mathbb{S}^3} \;=\; \sin^2\chi\,\mathrm{d}\chi\,\mathrm{d}\Omega_2, \tag{13}$$

where $\mathrm{d}\Omega_2$ is the uniform measure on $\mathbb{S}^2$. Since w is directly related to the rotation angle θ by Eq. (12), we are interested in the marginal distribution of w. To get the marginal of w, we compute the area (hence the probability mass for the uniform measure) of the "spherical band" between χ and $\chi + \mathrm{d}\chi$. This area is proportional to

$$\sin^2\chi\,\mathrm{d}\chi = \sin^2\chi\,\left|\frac{\mathrm{d}\chi}{\mathrm{d}w}\right|\mathrm{d}w = \sin^2\chi\,\frac{1}{\sin\chi}\,\mathrm{d}w = \sin\chi\,\mathrm{d}w = \sqrt{1-w^2}\,dw, \tag{14}$$

where we have used that $w = \cos\chi$ and that $\mathrm{d}w = -\sin\chi\,\mathrm{d}\chi$, the marginal density of w is given by

$$f_w(w) \;\propto\; \sqrt{1-w^2}, \qquad w \in [-1, 1]. \tag{15}$$

Normalizing with $\int_{-1}^{1} \sqrt{1-w^2}\,\mathrm{d}w = \pi/2$ gives

$$f_w(w) \;=\; \frac{2}{\pi}\sqrt{1-w^2}, \qquad w \in [-1,1]. \tag{16}$$

Now , since q and $-q$ represent the same rotation, let us consider the density of $|w|$:

$$f_{|w|}(u) \;=\; 2f_W(u) \;=\; \frac{4}{\pi}\sqrt{1-u^2}, \qquad u \in [0,1]. \tag{17}$$

Using that $u = |w| = \cos(\theta/2)$ (cf. Eq. (12)) and thus $\mathrm{d}u/\mathrm{d}\theta = -\frac{1}{2}\sin(\theta/2)$, the density of $\theta \in [0,\pi]$ follows by change of variables:

$$\begin{aligned} p(\theta) &= f_{|w|}\big(\cos(\theta/2)\big)\,\Big|\frac{\mathrm{d}}{\mathrm{d}\theta}\cos(\theta/2)\Big| \\ &= \frac{4}{\pi}\sqrt{1-\cos^2(\theta/2)}\;\cdot\;\frac{1}{2}\sin(\theta/2) \\ &= \frac{2}{\pi}\sin^2\!\big(\theta/2\big)\,, \qquad \theta \in [0,\pi]. \end{aligned} \tag{18}$$

Hence, the principal rotation angle of a Haar–uniform $R \in SO(3)$ has density $p(\theta) = \frac{2}{\pi}\sin^2(\theta/2)$ on $[0,\pi]$.

C Details regarding model training and architectures

Details on the message passing architecture used for the detection of default orientations. In Sec. 3.1, we demonstrate that a simple geometric message passing network accurately discerns canonical samples from randomly rotated ones. For the model, we employ a straightforward point cloud architecture, consisting of three layers of message passing:

$$f_i^{(k+1)} = \bigoplus_{j\in\mathcal{N}(i)} \mathrm{MLP}(f_j^{(k)}, \mathrm{emb}(x_i - x_j)), \tag{19}$$

where $f_i^{(k)}$ denotes the feature vector of atom i in layer k, x_i its position and $\mathcal{N}(i)$ its neighborhood (defined by a radial cutoff of 10 Å). As aggregation function $\bigoplus_{j\in\mathcal{N}(i)}$ we use the component-wise max operation. The input to these message passing layers consists of a learned embedding of the neighbor geometry combined with an embedding of the atom type. More specifically, for the radial embedding of the relative distance $r_{ij} = \|x_i - x_j\|$ we use Bessel functions of the first kind with 32 learnable frequencies. Similarly, we use Bessel functions with 20 learnable frequencies to separately embed each component of the normalized relative distance vector as angular embedding. The aggregated (summed) angular and radial embeddings from the local neighborhood are combined with a one-hot embedding of the atom type to form the input node features $f_i^{(0)}$. During message passing the angular and radial part of the relative distance vectors $x_i - x_j$ are embedded using 64 Gaussian radial basis functions (spaced equidistantly between 0 and 10 Å). The network is then trained to minimize a binary cross-entropy loss, predicting whether the input molecule has been randomly rotated or is in its (perturbed) default pose. All hyperparameters are summarized in Tab. 2.

MLP used for the molecular property prediction from molecular orientations alone. In Sec. 3.3, we demonstrate that a simple MLP receiving as input only normalized principal components (PCs) of atom positions can successfully regress molecular properties. The MLP receives the first two normalized PCs as input, uses SiLU activations and four hidden channels with 256 features each. It is trained with MSE loss without weight decay. We have trained one separate model of the same architecture for each property from the different datasets reported in Tab. 1.

Dealing with the dataset size. For QM9 we train all models for 100 epochs on the full dataset using a train-val-test split of (11000, 10000, ∼20000). For the larger QMugs dataset we use a train-val-test split of (80%, 10%, 10%). However, in order to keep the train time and learning rate scheduling comparable to the one in QM9, we train all QMUGS models for 100 epochs with a different random train subsets of size 110000 for each epoch. For the massive OMol25 dataset we use a train-val-test split of (99,8%, 0.1%, 0.1%) and train for a total of 200 epochs again on different random subsets of the training set of size 110000.

LLM usage. Large Language Models (LLMs) were used in the preparation of this submission to polish the writing regarding formulations and wording. In addition, we have used LLM based auto-completion in the development of our research code.

Table 2: **Hyperparameters for training our simple message passing network on default orientation detection.**

	Architecture hyperparameter
Bessel frequencies (radial)	32
Bessel frequencies (angular)	20
Num. message passing layers (Eq. (19))	3
Aggregation operation in message passing	max
Node feature dimension	512
Hidden layers in message MLP	[128]
Activation function	SiLU
Radial cutoff for message passing	10 Å
Hidden layers in readout MLP	[512, 128, 32]

	Training hyperparameter
Optimizer	AdamW
Weight decay	5e-3
Learning rate	5e-4
Scheduler	Cosine-LR
Epochs	200 for OMol25, 100 otherwise
Warm up epochs	5
Gradient clip	0.5
Loss function	BCE-loss