Preserving Geometric Symmetry in Uncertainty Estimation for Molecular Forces

Xinyu Li[1,2]
[1]Responsible AI Research Centre
Australian Institute for Machine Learning
Adelaide University, Australia
[2]CSIRO, Australia
henry.li@adelaide.edu.au

Zhen Zhang
Responsible AI Research Centre
Australian Institute for Machine Learning
Adelaide University, Australia
zhen@zzhang.org

Jiayu Huang
Australian Institute for Machine Learning
Adelaide University, Australia
jiayu.huang@adelaide.edu.au

Daniel M. Steinberg
CSIRO, Australia
dan.steinberg@csiro.au

Lina Yao[1,2]
[1]School of Computer Science and Engineering
University of New South Wales, Australia
[2]CSIRO, Australia
lina.yao@unsw.edu.au

Javen Qinfeng Shi
Responsible AI Research Centre
Australian Institute for Machine Learning
Adelaide University, Australia
javen.shi@adelaide.edu.au

Abstract

Accurate and reliable uncertainty estimation (UE) for vector-valued physical properties is crucial for scientific discovery in fields like drug and materials discovery. For example, atomic forces are central to finding optimal structures and identifying equilibrium systems, and a fundamental requirement of these vectors is that they must be equivariant under 3D rotations. However, existing UE methods often fail to respect these geometric constraints, leading to poorly calibrated uncertainty and degraded predictive performance. To address the problem, we introduce a novel framework for equivariant multivariate evidential regression. Our primary contribution is a theoretically grounded parameterization of the evidential prior's scale matrix, constructed via a decomposition into an invariant scalar component and an equivariant low-rank term. Furthermore, we address the lack of suitable evaluation standards for vector fields by introducing a rigorous calibration metric based on the Mahalanobis distance. Extensive experiments on molecular property prediction benchmarks, including MD17, QM7-X, and OC20, show that our framework consistently outperforms established baselines, achieving state-of-the-art single-model predictive accuracy with well-calibrated uncertainty. This work provides a principled and practical approach to uncertainty estimation for equivariant vector-valued properties, paving the way for more trustworthy machine learning applications in scientific discovery. Our code is publicly available at https://github.com/lixinyuu/EquivariantUE.

1 Introduction

The accurate prediction of molecular or crystal properties using their 3D structures is fundamental to modern computational science, enabling applications from drug discovery to materials design.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

The gold standard for these predictions, quantum mechanical (QM) calculations [17, 22], remains prohibitively expensive for large-scale simulations. To overcome this bottleneck, learning-based approaches, particularly Graph Neural Networks (GNNs), have emerged as powerful surrogates, capable of learning complex atomic interactions from QM data and accelerating simulations by orders of magnitude without sacrificing the solution quality [8, 13, 21, 36, 37].

Accurately predicting the properties of molecules and crystals is a crucial early step in materials and drug discovery. These predictions are typically followed by costly and time-consuming experimental validation of the most promising candidates identified by machine learning models. In this context, it has become essential to reliably quantify the uncertainty of such predictions, ensuring that resources are allocated efficiently and that confidence in the models is strengthened [1, 41, 55]. Beyond uncertainty quantification, it is equally important that predictive models respect the fundamental symmetries of molecular and crystalline systems. Equivariance [3, 4, 33, 43, 56] ensures that predictions remain consistent under rotations, translations, or permutations, thereby aligning the model with the underlying physics.

Current molecular/crystal property prediction systems mainly cover two categories of quantities: (1) scalar quantities such as total energy or log partition coefficients ($\log \mathrm{P}$); and (2) vector-valued quantities such as atomic forces. While scalar quantities can be used to describe many important properties of chemicals, vector-valued quantities often provide a more fine-grained characterization of molecular interactions. For example, accurate force predictions are essential for tasks such as geometry optimization and Molecular Dynamics (MD) simulations. Moreover, reliably estimating the uncertainty in these vectorized predictions is particularly important, since even small errors in predicted forces can accumulate and lead to qualitatively incorrect trajectories in long-timescale simulations. By quantifying the uncertainty of vectorized outputs, researchers can better assess the stability of simulated systems, identify regions of chemical space where models are less reliable, and guide experimental or high-fidelity computational validation more effectively [26, 40].

In current uncertainty estimation (UE), particularly deep learning–based approaches [2, 39, 46, 52], neural networks are trained to output not a single prediction but a parameterized distribution over the quantities of interest. For scalar quantities, these approaches naturally satisfy equivariance constraints, as the distribution remains invariant under transformations. However, for vector-valued quantities, existing methods often violate equivariance: the predicted distribution changes inconsistently when rotations are applied to the input. Most current UE techniques were originally designed for scalar outputs and can be broadly divided into two families. Sampling-based methods, such as ensembles and dropout [10, 51], approximate a predictive distribution by generating and aggregating multiple predictions. In contrast, sampling-free approaches, including evidential regression [2, 23] and stochastic variational Gaussian processes [15, 16, 35, 54], directly output the parameters of a distribution in a single forward pass, often with improved computational efficiency.

While sampling-based approaches can, in principle, be extended to respect equivariant constraints, such extensions are computationally expensive due to the need for large numbers of samples and symmetry operations. Sampling-free approaches, on the other hand, can be generalized from univariate to multivariate distributions by replacing scalar predictive distributions with their multivariate counterparts. However, a naïve extension such as Soleimany et al. [39] generally breaks equivariance, since the resulting distribution fails to transform consistently with the corresponding physical opera-

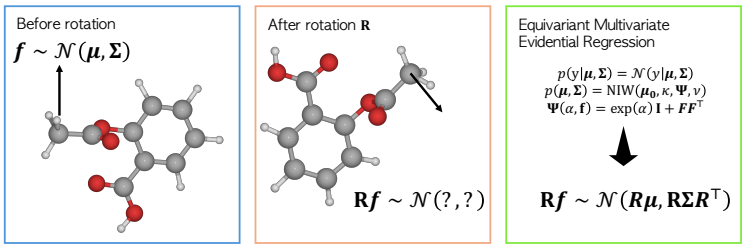


Figure 1: An illustration of equivariant uncertainty for a vector property $\boldsymbol{f}$, such as an atomic force. Before rotation, $\boldsymbol{f}$ follows a multivariate normal distribution. After rotation, the $\boldsymbol{f}$ should follow a specific multivariate normal distribution to reflect the equivariance. Our proposed method, Equivariant Multivariate Evidential Regression, is designed to directly learn this rotation-aware uncertainty.

tions. This highlights the need for new frameworks that can simultaneously achieve computational tractability and preserve equivariance in the UE of vector-valued quantities.

To bridge this critical gap, we introduce a novel framework called Equivariant Multivariate Evidential Regression (EMER). Our core contribution is a new, provably equivariant parameterization for the covariance matrix that allows a model to produce reliable and geometrically consistent UE for vector outputs. Furthermore, we propose a robust multivariate calibration metric based on the Mahalanobis distance to properly evaluate the quality of vector-based uncertainty. We demonstrate on molecular force and dipole moment prediction benchmarks that our method not only produces well-calibrated uncertainty but also achieves state-of-the-art single-model predictive accuracy, outperforming existing UE baselines.

To summarize, our contributions are:

- We propose a novel equivariant multivariate evidential regression framework featuring a new covariance parameterization that is provably equivariant under $\mathrm{SO}(3)$ rotations.
- We introduce a distance-based calibration score for equivariant multivariate uncertainty estimates, overcoming the inapplicability of standard scalar metrics in this setting.
- We empirically demonstrate that our method achieves state-of-the-art single-model accuracy and calibration.

2 Related Work

2.1 Graph Neural Networks

GNNs have emerged as a powerful framework for molecular property prediction [11, 36, 37]. The core of most GNNs is the message-passing paradigm, where node representations are iteratively updated by aggregating information from their local neighborhoods [13]. However, standard message-passing architectures often fail to capture the inherent geometric symmetries. For instance, in molecular modeling, some properties (e.g., forces) are dependent on their orientation in 3D space. A standard GNN, which might only use inter-atomic distances, discards crucial directional information.

To address this limitation, equivariant GNNs (EGNNs) were developed to explicitly incorporate geometric symmetries, particularly the rotational symmetries of 3D space (i.e., $\mathrm{SO}(3)$ symmetry) [3, 4, 33, 43, 56]. The principle of equivariance ensures that as the input coordinates are rotated, the intermediate feature vectors within the network transform in a corresponding, predictable manner. Mathematically, a function f (such as a neural network layer) is equivariant under a rotation $\mathcal{R} \in \mathrm{SO}(3)$ if rotating the input coordinates $\mathbf{x}$ produces an identically rotated output:

$$f(\mathcal{R}\mathbf{x}) = \mathcal{R}f(\mathbf{x})$$

This embeds a strong inductive bias for 3D geometry directly into the model's architecture. By preserving this geometric structure, equivariant models can predict both invariant scalar properties (e.g., energy) and equivariant vector properties (e.g., forces, dipoles), leading to significantly improved data efficiency and generalization performance.

2.2 Uncertainty Estimation

In the area of machine learning potentials, there is an urgent need for calibrated UE methods, as a high-error prediction during an MD simulation would lead to incorrect conclusions. While numerous UE strategies exist, several have become common in molecular modeling due to their differentiability and applicability to energy predictions.

Ensembles, often regarded as a gold standard for UE, train a diverse set of models on identical data [25, 34]. The prediction variance across the ensemble serves as an uncertainty metric. Despite their high accuracy, ensembles impose a significant computational burden, as the cost of both training and inference scales linearly with the number of models.

Monte Carlo (MC) Dropout offers a more computationally efficient approximation of Bayesian inference in deep neural networks [10]. This technique involves performing multiple stochastic forward passes during inference with dropout layers enabled. Depending on how dropout is performed, it can be categorized into Bernoulli [10], Gaussian [48] and Variational Dropout [20, 30]. Each pass samples a different sub-network, and the variance of the resulting predictions is used to estimate

model uncertainty. While this approach eliminates the need to train multiple models, the requirement for repeated forward passes can still be prohibitive.

Evidential Regression is a sampling-free UE method that learns the parameters of a higher-order evidential distribution (e.g., a Normal-Inverse-Gamma distribution) from which the predictive uncertainty can be derived analytically in a single forward pass [2, 39]. However, this method is formulated for independent scalar outputs. When naively applied to a vector by treating each component separately, the resulting uncertainty estimate fails to capture the covariance between components, thereby violating the crucial property of rotational equivariance.

Stochastic Variational Gaussian Process (SVGP) is a scalable Bayesian method that provides calibrated UE for molecular property prediction [16]. By using a sparse variational framework with inducing points, SVGP avoids the cubic complexity of exact Gaussian Processes [35] and can be trained efficiently on large molecular datasets with mini-batch optimization. SVGP is often combined with Deep Kernel Learning (DKL) [53], where a deep neural network maps molecular structures into a feature space for the SVGP. Building on this, Wollschläger et al. [54] introduced the Localizing Neural Kernel (LNK), which fits an SVGP to each atom's embedding and sums the per-atom contributions, consistent with the principle that total energy is additive over atoms. However, the nonlinear mappings used in LNK generally break geometric symmetries, meaning that equivariance is not preserved.

Post-hoc calibration methods aim to align predicted uncertainties with empirical errors without retraining the underlying model. Conformal prediction (CP) [47] provides distribution-free prediction intervals with finite-sample coverage guarantees. Hu et al. applied CP to machine-learned interatomic potentials for catalytic adsorption systems, calibrating energy uncertainties using residual-based nonconformity scores and latent-space distance heuristics [19]. In regression, CP typically uses a held-out calibration set to determine a quantile such that the true value lies within the predicted interval with probability $1 - \alpha$.

Calibration Score. To ensure the trustworthiness of a UE method, the quality of its uncertainty estimates must be evaluated. A primary means of this assessment is the calibration score [24]. A model is considered well-calibrated if its predicted uncertainty accurately reflects its true prediction error. This evaluation is performed by comparing the model's predicted Cumulative Distribution Function (CDF) against the empirically observed distribution. The discrepancy between the expected confidence and the actual observed frequency is often visualized in a reliability diagram, and this gap is quantified by metrics such as the expected calibration error [24] or the miscalibration area [45].

3 Proposed Methods

In this section, we first introduce the necessary preliminaries on equivariant learning and multivariate evidential regression. We then demonstrate that conventional UE methods fail to preserve rotational equivariance, which is essential for tasks involving 3D molecular or crystalline systems. To bridge this gap, we introduce a novel framework that provides robust uncertainty estimates while preserving geometric symmetries.

3.1 Preliminaries

Equivariant Predictor and Uncertainty Estimation Given a learning-based model $\hat{\boldsymbol{y}} = f(\boldsymbol{x}; \Theta)$, where $\boldsymbol{x}$ represents the input data and $\hat{\boldsymbol{y}}$ is the predicted vector property, equivariant methods provide a prediction $\hat{\boldsymbol{y}}$ that rotates consistently with the input $\boldsymbol{x}$, satisfying $\mathcal{R}(\hat{\boldsymbol{y}}) = f(\mathcal{R}(\boldsymbol{x}); \Theta)$ for any rotation operator $\mathcal{R}$. In UE, rather than providing a single deterministic output, the goal is to quantify the prediction uncertainty. Therefore, instead of simply a point estimate of $\hat{\boldsymbol{y}}$, the model must estimate a distribution. Particularly for an equivariant learning approach, this distribution must itself be equivariant. This imposes the constraint that for any rotation operator $\mathcal{R}$, the distribution must satisfy $p(\mathcal{R}(\boldsymbol{y})|\mathcal{R}(\boldsymbol{x}); \Theta) = p(\boldsymbol{y}|\boldsymbol{x}; \Theta)$.

Multivariate Evidential Regression To address the need for an efficient, single-pass and equivariant UE method for vectors, we turn to Multivariate Evidential Regression (MER). As an extension of evidential regression for scalar targets [2, 39], MER was introduced by Meinert and Lavin [31] to

handle multivariate outputs. The core innovation is to replace the Normal-Inverse-Gamma prior with a Normal-Inverse-Wishart (NIW) distribution, as

$$\begin{aligned} \boldsymbol{y} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma} &\sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \\ (\boldsymbol{\mu}, \boldsymbol{\Sigma}) &\sim \mathrm{NIW}(\boldsymbol{\mu}_0, \kappa, \boldsymbol{\Psi}, \nu). \end{aligned} \tag{1}$$

Here, $\boldsymbol{\mu}_0 \in \mathbb{R}^n$ is the location vector representing the estimated mean, with n denoting the dimension of the target vector ($n = 3$ for 3D forces). $\kappa \in \mathbb{R}_{>0}$ is a scalar scaling factor reflecting the confidence in the mean estimate $\boldsymbol{\mu}_0$. $\boldsymbol{\Psi} \in \mathbb{S}^n_{++}$ is an $n \times n$ symmetric positive-definite (SPD) scale matrix, and the degrees of freedom ν must satisfy $\nu > n + 1$ to ensure a valid mean and variance.

A neural network $f_{\boldsymbol{\theta}}$ learns a mapping from an input graph $\boldsymbol{x}$ to the four NIW parameters, i.e., $(\boldsymbol{\mu}_0, \kappa, \boldsymbol{\Psi}, \nu) = f_{\boldsymbol{\theta}}(\boldsymbol{x})$. Following Meinert and Lavin [31], we set $\kappa = r\nu$ with $r = 1$, the loss function can be written as

$$\begin{aligned} L_{\mathrm{NLL}_i} = -\log p(\boldsymbol{y}_i \mid \boldsymbol{\mu}_0, \kappa, \boldsymbol{\Psi}, \nu) &= \log\Gamma\left(\frac{\nu - n + 1}{2}\right) - \log\Gamma\left(\frac{\nu + 1}{2}\right) \\ + \frac{n}{2}\log(1+\nu) - \frac{\nu}{2}\log|\boldsymbol{\Psi}| \quad &+ \frac{\nu+1}{2}\log\left|\boldsymbol{\Psi} + \frac{1}{1+\nu}(\boldsymbol{y}_i - \boldsymbol{\mu}_0)(\boldsymbol{y}_i - \boldsymbol{\mu}_0)^\top\right| \end{aligned}$$

where $\Gamma(\cdot)$ denotes the Gamma function.

The estimation of predictions and uncertainties is

$$\hat{\boldsymbol{y}} = \mathbb{E}[\boldsymbol{\mu}] = \boldsymbol{\mu}_0 \qquad \mathbb{E}[\boldsymbol{\Sigma}] = \frac{\boldsymbol{\Psi}}{\nu - n - 1} \tag{2}$$

To ensure the matrix $\boldsymbol{\Psi}$ is symmetric and Positive Definite (PD), MER constructs it from its Cholesky factor $\mathbf{L}$, which is parameterized by the network outputs. That is, instead of predicting $\boldsymbol{\Psi}$ directly, the network yields the elements of the lower-triangular matrix $\mathbf{L}$. For the i-th data point, the components of $\mathbf{L}_i$ are defined as:

$$(L_i)_{jk} = \begin{cases} \exp\left(\boldsymbol{S}^{(i)}_j\right) & \text{if } j = k, \\ \boldsymbol{S}^{(i)}_{jk} & \text{if } j > k, \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

where $\boldsymbol{S}^{(i)} \in \mathbb{R}^{n \times n}$ is a neural network predicted matrix. The covariance matrix is then assembled via $\boldsymbol{\Psi}_i = \mathbf{L}_i \mathbf{L}_i^\top$.

However, directly applying the MER framework to equivariant UE is ill-posed. In MER, a network outputs the scalar components of a lower-triangular matrix $\mathbf{L}$. An equivariant network producing scalar outputs will necessarily yield rotationally invariant scalars, making the resulting matrices $\mathbf{L}$ and $\boldsymbol{\Psi} = \mathbf{L}\mathbf{L}^\top$ also invariant to input rotations, whereas in an equivariant system $\boldsymbol{\Psi}$ should transform according to Equation (4)[1]. Alternatively, constructing $\mathbf{L}$ from an equivariant vector output is challenging, since the use of the nonlinear exponentiation function for diagonal elements breaks the equivariance. In equivariant models like EGNN, all nonlinear functions must be specially designed to preserve equivariance, such as norm-based [12] and gated activations [49, 50]. This requirement for specialized nonlinearities also explains why other UE methods, such as SVGP, cannot be directly applied to vector-valued outputs.

3.2 Equivariant Multivariate Evidential Regression

While MER and other single-pass UE methods struggle to provide a valid equivariant covariance matrix, sampling-based approaches like ensembles or MC dropout natively satisfy this. In these cases, the covariance is estimated empirically from a set of K predictions $\{\mathbf{v}^{(k)}\}$ as $\hat{\boldsymbol{\Sigma}} = \frac{1}{K-1}\sum_{k=1}^{K}(\mathbf{v}^{(k)} - \bar{\mathbf{v}})(\mathbf{v}^{(k)} - \bar{\mathbf{v}})^\top$, where $\bar{\mathbf{v}}$ is the sample mean. This empirical covariance is guaranteed to be Positive Semidefinite (PSD) and to satisfy the equivariance of

$$\boldsymbol{\Sigma}(\boldsymbol{R}\boldsymbol{x}) = \boldsymbol{R}\boldsymbol{\Sigma}(\boldsymbol{x})\boldsymbol{R}^\top. \tag{4}$$

[1] By Equation (2) there is only a scale factor between Ψ and the expectation of Σ.

Now we focus on how to make MER equivariant. Naive equivariant alternatives, such as parameterizing the covariance via an outer product of an equivariant matrix ($\boldsymbol{\Sigma} = \boldsymbol{F}\boldsymbol{F}^\top$) satisfy this, but they create an ill-posed optimization problem due to overparameterization. An n-dimensional symmetric PSD matrix $\boldsymbol{\Sigma}$ has $M = n(n+1)/2$ unique, independent parameters. The general factor matrix $\boldsymbol{F}$, however, has $N = n^2$ parameters. For a 3D case (e.g., molecular forces, where $n = 3$), this results in $N - M = 3$ redundant parameters, making $\boldsymbol{F}$ overparameterized. Besides, the solution for $\boldsymbol{F}$ is also not unique. Let $\boldsymbol{Q}$ be any $n \times n$ orthogonal matrix, satisfying $\boldsymbol{Q}\boldsymbol{Q}^\top = \mathbf{I}$. If $\boldsymbol{F}$ is a factor that generates $\boldsymbol{\Sigma}$, then the rotated factor $\boldsymbol{F}' = \boldsymbol{F}\boldsymbol{Q}$ generates an identical covariance matrix:

$$\boldsymbol{\Sigma}' = \boldsymbol{F}'(\boldsymbol{F}')^\top = (\boldsymbol{F}\boldsymbol{Q})(\boldsymbol{F}\boldsymbol{Q})^\top = \boldsymbol{F}\boldsymbol{Q}\boldsymbol{Q}^\top\boldsymbol{F}^\top = \boldsymbol{F}\mathbf{I}\boldsymbol{F}^\top = \boldsymbol{\Sigma} \tag{5}$$

Since the set of orthogonal matrices $\boldsymbol{Q}$ is infinite, the loss function $L(\boldsymbol{\Sigma}) = L(\boldsymbol{F}\boldsymbol{F}^\top)$ possesses an infinite, continuous manifold of equivalent minima with respect to the predicted factor $\boldsymbol{F}$.

To mitigate these issues, we propose Equivariant Multivariate Evidential Regression, which introduces a novel parameterization for the covariance structure that is inherently equivariant by construction.

Proposition 3.1 *Let the scale matrix $\boldsymbol{\Psi}$ in the Normal-Inverse-Wishart (NIW) distribution be parameterized as*

$$\boldsymbol{\Psi}(\alpha, \boldsymbol{F}) = \exp(\alpha)\mathbf{I} + \boldsymbol{F}\boldsymbol{F}^\top$$

where $\alpha \in \mathbb{R}$ is an invariant scalar and $\boldsymbol{F} \in \mathbb{R}^{3\times m}$ ($m < 3$) is an equivariant matrix. In our main model, we use $m = 2$. Then this parameterization satisfies the following properties:

- *(i)* *$\boldsymbol{\Psi}$ is guaranteed to be Positive Definite (PD), as $\exp(\alpha) > 0$.*
- *(ii)* *The matrix $\boldsymbol{\Psi}$ is equivariant: $\boldsymbol{\Psi}' = \mathbf{R}\boldsymbol{\Psi}\mathbf{R}^\top$ for a rotation $\mathbf{R}$.*
- *(iii)* *For $m = 2$, the non-identifiability of the factor $\boldsymbol{F}$ is reduced from the full orthogonal freedom of a 3×3 factorization to the smaller class $\boldsymbol{Q} \in O(2)$ acting on the two columns of $\boldsymbol{F}$, i.e., $\boldsymbol{\Psi}(\alpha, \boldsymbol{F}) = \boldsymbol{\Psi}(\alpha, \boldsymbol{F}\boldsymbol{Q})$ for any $\boldsymbol{Q} \in O(2)$.*

Proof Sketch The proposed parameterization $\boldsymbol{\Psi} = \exp(\alpha)\mathbf{I} + \boldsymbol{F}\boldsymbol{F}^\top$ is guaranteed to be PD because it is the sum of a PD isotropic term ($\exp(\alpha)\mathbf{I}$) and a PSD low-rank term ($\boldsymbol{F}\boldsymbol{F}^\top$). Equivariance is inherently satisfied because the invariant scalar α is associated with $\mathbf{I}$, and the equivariant matrix $\boldsymbol{F}$ is used only in the outer product, ensuring $\boldsymbol{\Psi}$ transforms correctly under rotation ($\boldsymbol{\Psi}' = \mathbf{R}\boldsymbol{\Psi}\mathbf{R}^\top$). For identifiability, note that when $m = 2$, any right multiplication by $\boldsymbol{Q} \in O(2)$ leaves $\boldsymbol{F}\boldsymbol{F}^\top$ unchanged, as $(\boldsymbol{F}\boldsymbol{Q})(\boldsymbol{F}\boldsymbol{Q})^\top = \boldsymbol{F}\boldsymbol{Q}\boldsymbol{Q}^\top\boldsymbol{F}^\top = \boldsymbol{F}\boldsymbol{F}^\top$. Hence the ambiguity is not completely removed, but compared with a full 3×3 factorization, it is restricted to the lower-dimensional orthogonal group $O(2)$. In the appendix, we further provide a Gram–Schmidt-based variant that yields an identifiable parameterization up to sign flips under mild non-degeneracy conditions. However, in our experiments, this more constrained form did not lead to improved performance, so we adopt the simpler parameterization above in the main model.

Avoiding pitfalls To enhance the stability of our UE and address the challenges of combining a large EGNN with the MER and EMER heads, we integrated dropout into them (denoted as MER-D and EMER-D, respectively). Although EMER utilizes a NIW prior for evidential uncertainty, the inherent complexity and capacity of the deep GNN encoder can still risk learning an overly confident or poorly conditioned representation. This phenomenon has been observed previously in DKL and is referred to as feature collapse [32, 46]. By employing dropout, we effectively regularize the feature space, ensure the predicted covariance matrix remains well-conditioned and the overall UE is well-calibrated. Note that in comparison to MC Dropout [10], where dropout is applied during both training and inference, here dropout is used only during the training stage. As a result, dropout serves solely as a regularizer and does not provide UE. This regularization strategy has also been successfully employed in Wollschläger et al. [54] and Tran et al. [44].

3.3 A Principled Calibration Metric for Equivariant Vectors

Evaluating the calibration of UE for equivariant vectors presents a unique challenge. Standard univariate metrics like the probability integral transform are inapplicable. To address this, we introduce a calibration metric based on the Mahalanobis distance [5]. This distance provides a scalar

measure of the error between the true vector $\mathbf{v}_{\text{true}}$ and the predicted mean $\boldsymbol{\mu}_0$, normalized by the predicted covariance matrix $\boldsymbol{\Sigma}$:

$$d^2 = (\mathbf{v}_{\text{true}} - \boldsymbol{\mu}_0)^\top \boldsymbol{\Sigma}^{-1} (\mathbf{v}_{\text{true}} - \boldsymbol{\mu}_0). \tag{6}$$

This quantity is invariant under orthogonal transformations, so the calibration evaluation itself respects the rotational symmetry of the predictive distribution—a property that dimension-wise scalar metrics fail to satisfy. The reference distribution of d^2 depends on the predictive distribution. In the special case where the predictive distribution is multivariate Gaussian, d^2 follows a chi-squared distribution with n degrees of freedom, where n is the vector dimension, i.e., $d^2 \sim \chi^2_n$. However, in evidential regression frameworks such as MER and EMER, the predictive distribution is generally a multivariate Student's t distribution. In this case, the corresponding Mahalanobis statistic follows a scaled F distribution, with the scaling determined by the degrees of freedom of the predictive Student's t distribution. Therefore, Mahalanobis calibration is computed by comparing the empirical CDF of the observed d^2 values against the theoretical reference CDF implied by the model's predictive distribution. Any deviation from the identity line indicates multivariate miscalibration.

3.4 Overall Framework

Our framework is built upon a standard EGNN. As shown in Figure 2 in Appendix A.6, in addition to the head that predicts the target vector, our method requires two additional uncertainty heads. These heads output the parameters for an evidential distribution: two scalars, ν and α, and an equivariant matrix, $\boldsymbol{F}$. These parameters are then used to construct the NIW scale matrix $\boldsymbol{\Psi}$. The entire network is trained end-to-end by minimizing the NLL of the evidential distribution. To implement MER-D and EMER-D, we integrate dropout layers within the network, ensuring they are active only during the training phase. Finally, the model's predictive accuracy and uncertainty quality are evaluated during testing.

4 Experiments

We compare established UE techniques with our proposed approach, including sampling-based methods such as ensembles, MC Dropout [10], as well as two sampling-free methods, Classification And Regression Diffusion (CARD) [14] and MER. While the latter two methods do not satisfy uncertainty equivariance, they preserve target equivariance, making them suitable baselines for our setting. Calibration quality is measured using the distance-based Calibration Score defined in Section 3. In addition, we measure the Raw Calibration Score (RCal) defined in [24, 54]. Since there are 3 dimensions, we only report the maximum RCal. Additional details about the metrics are provided in Appendix A.4. Unless otherwise specified, for each backbone network we adopt the hyperparameter settings reported in the original papers.

We evaluate our method on two vector-valued quantum chemistry tasks using established benchmarks. For force prediction, we use MD17 [7] and OC20 [6] (due to page limitations, the OC20 results are reported in Appendix A.5). For dipole-moment prediction, we use QM7-X [18], which provides full vector dipole moments. Force models are trained with a joint energy-and-force loss, whereas dipole-moment models are trained with a single regression loss.

To further stress-test rotational robustness under distribution shift, we construct controlled OOD variants of MD17 and QM7-X, denoted as MD17-RotOOD and QM7X-RotOOD, respectively. In these settings, we rotate the training samples so that the dominant vector variation lies primarily in the xy-plane, while the test samples are additionally rotated by 90° about an in-plane axis (e.g. x-axis for MD17), shifting their dominant variation out of the xy-plane. This procedure induces a deliberate rotational mismatch between the training and test distributions while preserving the underlying molecular structures and target magnitudes. Such orientation-bias effects have been specifically discussed in the recent work by Lippmann et al. [29]; they also occur in the MD17 dataset used here, where Naphthalene and Uracil exhibit predominantly planar force-label distributions (see Figure 4 in Appendix A.6).

Table 1 reports accuracy and calibration across the seven MD17 molecules. Several consistent trends emerge. First, Dropout generally degrades energy accuracy while maintaining force accuracy close to the backbone. Its calibration behavior is mixed: although it achieves competitive dimension-wise

Mol	Metric	Backbone	Equivariant		Nonequivariant		Ours		
			Ens.	Drop.	CARD	MER	EMER	MER-D	EMER-D
Aspirin	Energy MAE	0.2266	0.2220	0.3531	0.2266	0.1887	**0.1855**	0.4233	0.3791
	Force MAE	0.5092	0.3744	0.4969	0.5092	0.4301	0.4027	0.4205	**0.3936**
	MCal	–	0.8407	1.0333	1.0607	1.3618	1.3614	0.2213	**0.1730**
	Max RCal	–	0.1504	0.4804	0.7727	1.0486	0.9471	0.3449	**0.1536**
	Max NLL	–	0.8494	2.2843	6.5071	11.8053	7.7751	1.2087	**0.7517**
Ethanol	Energy MAE	0.0766	0.0696	0.1741	0.0766	0.0692	0.0692	0.1123	**0.0840**
	Force MAE	0.3170	0.2454	0.3220	0.3170	0.2738	0.2771	0.2802	**0.2770**
	MCal	–	0.6687	0.9045	1.1064	1.4021	1.4759	**0.2442**	0.2652
	Max RCal	–	0.0830	0.3728	0.7435	1.0844	1.0611	0.2949	**0.2169**
	Max NLL	–	0.4043	1.6424	8.637	23.9746	30.9523	1.1015	**0.5716**
Nap	Energy MAE	0.1214	0.1161	0.2906	0.1214	0.1215	**0.1186**	0.2640	0.1938
	Force MAE	0.1301	0.1004	0.1547	0.1301	0.1060	**0.1008**	0.1561	0.1616
	MCal	–	0.8821	0.4167	0.4387	0.9354	0.7842	0.2013	**0.1472**
	Max RCal	–	0.2345	**0.0670**	0.3758	0.9636	0.5592	0.7058	0.1564
	Max NLL	–	-0.3437	0.1318	1.1694	5.4850	0.4728	0.6587	**-0.1473**
MDA	Energy MAE	0.1220	0.1092	0.2222	0.1220	0.1059	**0.1028**	0.1426	0.1526
	Force MAE	0.4862	0.3642	0.4760	0.4862	0.4119	0.4135	0.3981	**0.3969**
	MCal	–	0.7390	1.0709	0.8862	1.4306	1.4944	**0.3664**	0.4074
	Max RCal	–	0.0776	0.5195	0.9210	1.1459	1.0982	0.4195	**0.3096**
	Max NLL	–	0.8097	2.6737	11.6005	37.1668	39.7620	2.0800	**1.2271**
SAC	Energy MAE	0.1340	0.1258	0.2656	0.1340	0.1298	**0.1230**	0.2754	0.2158
	Force MAE	0.3325	0.2484	0.3479	0.3325	0.2696	**0.2586**	0.2893	0.2687
	MCal	–	0.8143	0.7579	0.9548	1.2625	1.2355	**0.1860**	0.2211
	Max RCal	–	0.1929	0.2993	0.7599	1.1451	0.8436	0.6172	**0.1197**
	Max NLL	–	0.5122	1.5292	8.4232	33.7121	10.3233	2.3769	**0.4758**
Uracil	Energy MAE	0.1123	0.1127	0.2435	0.1123	**0.1068**	0.1083	0.1785	0.1881
	Force MAE	0.2154	0.1676	0.2586	0.2154	0.1840	**0.1691**	0.2061	0.1932
	MCal	–	0.7853	0.4968	0.7516	1.2470	1.1297	**0.2250**	0.5193
	Max RCal	–	0.1298	**0.1000**	0.6029	1.1861	0.7644	0.8112	0.2832
	Max NLL	–	0.1527	0.7690	4.9905	49.0317	3.6239	1.6989	**0.1369**
Toluene	Energy MAE	0.1106	0.1022	0.2470	0.1106	**0.0941**	0.0984	0.2268	0.2518
	Force MAE	0.1517	0.1089	0.1719	0.1517	0.1276	**0.1233**	0.1553	0.1583
	MCal	–	0.8166	0.4451	0.5588	1.0514	1.0271	**0.2068**	0.2465
	Max RCal	–	0.1476	**0.0733**	0.4360	0.8808	0.7062	0.4495	0.1816
	Max NLL	–	-0.3579	0.2220	2.2630	5.5851	2.0166	0.0260	**-0.2297**

Table 1: Performance metrics on the MD17-RotOOD dataset with the PaiNN backbone. Calibration and NLL are evaluated on atomic forces only. Bold indicates the best single-model performance. EMER achieves strong predictive accuracy, while EMER-D provides substantially improved calibration and rotational robustness compared with non-equivariant evidential baselines.

Model	Metric	Backbone	Equivariant		Nonequivariant		Ours		
			Ens.	Drop.	CARD	MER	EMER	MER-D	EMER-D
PaiNN	DP MAE	0.0072	0.0073	0.0081	0.0072	**0.0055**	**0.0055**	0.0057	0.0060
	MCal	–	0.6437	0.7543	1.5090	1.5864	1.5317	0.3866	**0.2757**
	Max RCal	–	0.2293	0.2024	1.1807	1.3169	1.1432	0.3402	**0.2112**
	Max NLL	–	-2.9908	-2.701	-2.6671	76.356	34.048	-2.6695	**-3.2425**
NequIP	DP MAE	0.0076	0.0067	0.0079	0.0076	**0.0058**	0.0060	0.0068	0.0063
	MCal	–	0.6756	0.6377	1.4550	1.1585	1.2507	0.4578	**0.2990**
	Max RCal	–	0.3107	0.1309	1.1380	0.9929	0.7544	0.6103	**0.2520**
	Max NLL	–	-3.0630	-2.8805	-2.7313	18.235	0.1678	-2.8444	**-3.3469**
MACE	DP MAE	0.0082	0.0080	0.0081	0.0082	0.0044	0.0042	**0.0039**	0.0041
	MCal	–	0.7295	1.0801	1.3935	1.5088	1.3735	**0.2685**	0.2957
	Max RCal	–	0.3648	0.5419	1.0623	1.1855	0.9818	0.2725	**0.1484**
	Max NLL	–	-2.9172	-1.0919	-2.6682	43.885	9.8256	-3.3246	**-3.8785**

Table 2: Performance metrics for dipole moment prediction using QM7X-RotOOD dataset. Bold indicates the best single-model performance. EMER and EMER-D achieve competitive predictive accuracy, while EMER-D provides the strongest calibration robustness under rotational distribution shift.

RCal, it performs poorly under distance-based calibration (MCal) and likelihood-based evaluation (Max NLL). The diffusion-based CARD approach maintains the backbone's predictive accuracy because it reuses the point estimate for its mean prediction, but its calibration remains weak across both MCal and RCal. Our proposed EMER achieves the best overall predictive accuracy, with the lowest Energy MAE for five molecules and lowest Force MAE for four. However, its uncertainty metrics, including MCal and dimension-wise scores, remain poorly calibrated. In contrast, EMER-D greatly improves calibration by using dropout to avoid feature collapse, achieving the best Max NLL on all seven molecules and the best Max RCal on four. It is also worth noting that non-equivariant MER-D achieves very similar predictive accuracy and MCal compared to EMER-D.

Model	Metric	w/o $\alpha\mathbf{I}$	w/ $\alpha\mathbf{I}$			
		$m=3$	$m=0$	$m=1$	$m=2$	$m=3$
MD17-Naphthalene						
EMER	EMAE	–	$0.127_{\pm 0.010}$	$0.116_{\pm 0.002}$	$\mathbf{0.112}_{\pm 0.002}$	$0.116_{\pm 0.006}$
	FMAE	–	$0.108_{\pm 0.001}$	$0.103_{\pm 0.005}$	$0.101_{\pm 0.003}$	$\mathbf{0.100}_{\pm 0.001}$
	MCal	–	$0.804_{\pm 0.027}$	$\mathbf{0.771}_{\pm 0.039}$	$0.803_{\pm 0.038}$	$0.797_{\pm 0.011}$
	Max RCal	–	$0.657_{\pm 0.021}$	$0.625_{\pm 0.022}$	$0.581_{\pm 0.023}$	$\mathbf{0.549}_{\pm 0.013}$
	Max NLL	–	$0.728_{\pm 0.227}$	$0.715_{\pm 0.115}$	$0.497_{\pm 0.158}$	$\mathbf{0.307}_{\pm 0.150}$
EMER-D	EMAE	–	$0.305_{\pm 0.097}$	$0.240_{\pm 0.020}$	$\mathbf{0.205}_{\pm 0.078}$	–
	FMAE	–	$0.167_{\pm 0.005}$	$\mathbf{0.160}_{\pm 0.008}$	$0.166_{\pm 0.012}$	–
	MCal	–	$\mathbf{0.062}_{\pm 0.041}$	$0.158_{\pm 0.071}$	$0.157_{\pm 0.132}$	–
	Max RCal	–	$0.498_{\pm 0.020}$	$0.412_{\pm 0.041}$	$0.214_{\pm 0.060}$	–
	Max NLL	–	$-0.027_{\pm 0.046}$	$\mathbf{-0.151}_{\pm 0.044}$	$-0.127_{\pm 0.093}$	–
QM7X						
EMER	DPMAE	$\mathbf{0.005}_{\pm 0.0003}$	$\mathbf{0.005}_{\pm 0.0003}$	$\mathbf{0.005}_{\pm 0.0003}$	$0.006_{\pm 0.0004}$	$0.006_{\pm 0.0006}$
	MCal	$\mathbf{1.192}_{\pm 0.032}$	$1.482_{\pm 0.054}$	$1.545_{\pm 0.029}$	$1.532_{\pm 0.028}$	$1.536_{\pm 0.021}$
	Max RCal	$\mathbf{0.967}_{\pm 0.023}$	$1.202_{\pm 0.062}$	$1.177_{\pm 0.028}$	$1.143_{\pm 0.012}$	$1.189_{\pm 0.031}$
	Max NLL	$\mathbf{-3.218}_{\pm 0.092}$	$38.693_{\pm 10.218}$	$48.672_{\pm 22.505}$	$34.048_{\pm 8.669}$	$45.052_{\pm 16.886}$
EMER-D	DPMAE	–	$\mathbf{0.006}_{\pm 0.0006}$	$\mathbf{0.006}_{\pm 0.0006}$	$\mathbf{0.006}_{\pm 0.0005}$	–
	MCal	–	$0.325_{\pm 0.078}$	$0.285_{\pm 0.054}$	$\mathbf{0.276}_{\pm 0.097}$	–
	Max RCal	–	$0.311_{\pm 0.066}$	$0.227_{\pm 0.088}$	$\mathbf{0.211}_{\pm 0.086}$	–
	Max NLL	–	$-3.176_{\pm 0.296}$	$\mathbf{-3.251}_{\pm 0.437}$	$-3.243_{\pm 0.366}$	–

Table 3: Ablation study on the design of the matrix $\mathbf{\Psi} = \exp(\alpha)\mathbf{I} + \boldsymbol{F}\boldsymbol{F}^\top$. We compare configurations with and without the scalar term $\exp(\alpha)\mathbf{I}$ and vary the dimensionality $m \in \{0, 1, 2, 3\}$ of the matrix $\boldsymbol{F} \in \mathbb{R}^{3\times m}$ for EMER and EMER-D. The results demonstrate that the $\exp(\alpha)\mathbf{I}$ term is essential for stability. Moreover, a F with dimension of $m = 2$ provides the best balanced performance.

However, MER-D performs poorly on dimension-wise RCal and NLL. This is because MER-D is non-equivariant and provides invariant dimension-wise UE, and therefore fails under rotational distribution shift. In contrast, EMER-D preserves equivariance, leading to more reliable dimension-wise calibration under RotOOD. Indeed, equivariant methods, including Ensembles, Dropout, and EMER-D, consistently achieve much stronger Max RCal than non-equivariant methods, indicating that rotational equivariance is crucial for dimension-wise calibration in vector UE. Overall, EMER's high accuracy and EMER-D's calibration and robustness under rotation highlight the importance of introducing equivariance to UE. We further compare with a controlled Gaussian-NLL variant without the NIW prior (Gauss-D), as well as BLIP [9] and GMM [42] (Table 9 in Appendix A.5). The comparison with Gauss-D suggests that the NIW prior primarily improves force accuracy, while its calibration benefit is concentrated on systems with directional force uncertainty.

We further evaluate our method on QM7X-RotOOD for dipole-moment prediction using three backbones (PaiNN [38], NequIP [4], and MACE [3]), with results reported in Table 2. The performance is very similar to that of force prediction task – MER and EMER generally attain lower MAE than MER-D and EMER-D, and their MAEs are even better than Ensemble. Adding dropout can greatly improve the MCal, from >1.00 to ∼0.3. However, under the RotOOD setting, only EMER-D can maintain robustness – EMER-D achieves the best Max RCal and Max NLL for all three models. These results suggest that dropout improves the overall scale of uncertainty, but equivariance is crucial for ensuring that the uncertainty geometry rotates consistently with the predicted vector. This is particularly important for vector-valued molecular properties, where the target itself transforms under coordinate rotations.

Ablation Study An ablation study was conducted to evaluate the design of the matrix $\mathbf{\Psi} = \exp(\alpha)\mathbf{I} + \boldsymbol{F}\boldsymbol{F}^\top$. Using PaiNN as the backbone, we compared configurations with and without the $\exp(\alpha)\mathbf{I}$ term and varied the dimensionality m of the matrix $\boldsymbol{F}$. We tested these configurations with both EMER and EMER-D, with results summarized in Table 3. Variants without the $\exp(\alpha)\mathbf{I}$ term, as well as the $m = 3$ variant under EMER-D, were unstable during training, supporting our design rationale that the isotropic scalar term improves covariance stability, while a full-rank factor can be over-parameterized. Focusing on the better calibrated EMER-D variants, different choices of m achieve broadly similar MCal values, including the isotropic case $m = 0$. This is expected, as the distance-based calibration score is relatively insensitive to the orientation of an equivariant covariance. However, dimension-wise calibration shows a clearer advantage for $m = 2$: on MD17-Naphthalene, Max RCal decreases from 0.412 with $m = 1$ to 0.214 with $m = 2$, and on QM7X, $m = 2$ also gives the lowest MCal (0.276) and Max RCal (0.211) among stable EMER-D variants. Consequently, the

results support the adopted choice of $m = 2$, which provides the best balance between accuracy, distance-based calibration, dimension-wise calibration and training stability.

5 Conclusion

In this work, we addressed the critical and previously overlooked challenge of obtaining reliable, geometrically consistent uncertainty estimates for equivariant vector-valued properties in chemistry. We demonstrated that conventional UE methods fail to respect underlying physical symmetries, yielding uncertainty measures that are not only poorly calibrated but also physically meaningless. To resolve this, we introduced Equivariant Multivariate Evidential Regression (EMER), a novel framework built upon a new parameterization for the covariance matrix that is provably equivariant and positive definite. Furthermore, we proposed a Mahalanobis distance-based calibration score to enable the rigorous evaluation of multivariate uncertainty calibration.

Our extensive experiments on standard molecular property prediction benchmarks—MD17, OC20, and QM7-X—empirically validate our theoretical contributions. The results consistently show that EMER outperforms established baselines, achieving a new single-model state-of-the-art in predictive accuracy, vector distance-based calibration, and dimension-wise calibration. This work represents a significant step toward building more reliable and trustworthy machine learning models for scientific applications. By providing a principled and practical approach to equivariant UE, our framework paves the way for more robust active learning workflows, error detection in large-scale simulations, and adaptive multiscale modeling, thereby accelerating the pace of scientific discovery.

Acknowledgments and Disclosure of Funding

This work was supported by the Responsible AI Research Centre, an initiative of Adelaide University, CSIRO and the Government of South Australia.

References

[1] Jehad Abed, Jiheon Kim, Muhammed Shuaibi, Brook Wander, Boris Duijf, Suhas Mahesh, Hyeonseok Lee, Vahe Gharakhanyan, Sjoerd Hoogland, Erdem Irtem, Janice Lan, Niels Schouten, Anagha Usha Vijayakumar, Jason Hattrick-Simpers, John R. Kitchin, Zachary W. Ulissi, Aaike van Vugt, Edward H. Sargent, David Sinton, and C. Lawrence Zitnick. Open catalyst experiments 2024 (ocx24): Bridging experiments and computational models, 2024. URL `https://arxiv.org/abs/2411.11783`.

[2] Alexander Amini, Wilko Schwarting, Ava Soleimany, and Daniela Rus. Deep evidential regression. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL `https://proceedings.neurips.cc/paper/2020/hash/aab085461de182608ee9f607f3f7d18f-Abstract.html`.

[3] Ilyes Batatia, David Peter Kovacs, Gregor N. C. Simm, Christoph Ortner, and Gabor Csanyi. MACE: Higher order equivariant message passing neural networks for fast and accurate force fields. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL `https://openreview.net/forum?id=YPpSngE-ZU`.

[4] Simon Batzner, Albert Musaelian, Lixin Sun, Mario Geiger, Jonathan P. Mailoa, Mordechai Kornbluth, Nicola Molinari, Tess E. Smidt, and Boris Kozinsky. E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nature Communications*, 13(1), May 2022. ISSN 2041-1723. doi: 10.1038/s41467-022-29939-5. URL `http://dx.doi.org/10.1038/s41467-022-29939-5`.

[5] Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*, volume 4. Springer, 2006.

[6] Lowik Chanussot, Abhishek Das, Siddharth Goyal, Thibaut Lavril, Muhammed Shuaibi, Morgane Riviere, Kevin Tran, Javier Heras-Domingo, Caleb Ho, Weihua Hu, Aini Palizhati, Anuroop Sriram, Brandon Wood, Junwoong Yoon, Devi Parikh, C. Lawrence Zitnick, and Zachary Ulissi. Open catalyst 2020 (OC20) dataset and community challenges. *ACS Catalysis*, 11(10):6059–6072, May 2021. ISSN 2155-5435, 2155-5435. doi: 10.1021/acscatal.0c04525.

[7] Stefan Chmiela, Alexandre Tkatchenko, Huziel E. Sauceda, Igor Poltavsky, Kristof T. Schütt, and Klaus-Robert Müller. Machine learning of accurate energy-conserving molecular force fields. *Science Advances*, 3(5), May 2017. ISSN 2375-2548. doi: 10.1126/sciadv.1603015. URL http://dx.doi.org/10.1126/sciadv.1603015.

[8] Junyoung Choi, Gunwook Nam, Jaesik Choi, and Yousung Jung. A perspective on foundation models in chemistry. *JACS Au*, 5(4):1499–1518, 2025.

[9] Dario Coscia, Pim de Haan, and Max Welling. Blips: Bayesian learned interatomic potentials, 2026. URL https://arxiv.org/abs/2508.14022.

[10] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In Maria-Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, volume 48 of *JMLR Workshop and Conference Proceedings*, pages 1050–1059. JMLR.org, 2016. URL http://proceedings.mlr.press/v48/gal16.html.

[11] Johannes Gasteiger, Florian Becker, and Stephan Günnemann. GemNet: Universal directional graph neural networks for molecules. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan, editors, *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, Virtual*, pages 6790–6802, 2021.

[12] Mario Geiger and Tess Smidt. e3nn: Euclidean neural networks, 2022. URL https://arxiv.org/abs/2207.09453.

[13] Justin Gilmer, Samuel S. Schoenholz, Patrick F. Riley, Oriol Vinyals, and George E. Dahl. Neural message passing for quantum chemistry. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning, ICML*, volume 70 of *Proceedings of Machine Learning Research*, pages 1263–1272. PMLR, 2017.

[14] Xizewen Han, Huangjie Zheng, and Mingyuan Zhou. Card: Classification and regression diffusion models. In *Thirty-Sixth Conference on Neural Information Processing Systems*, 2022.

[15] James Hensman, Nicoló Fusi, and Neil D. Lawrence. Gaussian processes for big data. In Ann E. Nicholson and Padhraic Smyth, editors, *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence, UAI 2013, Bellevue, WA, USA, August 11-15, 2013*. AUAI Press, 2013. URL https://dslpitt.org/uai/displayArticleDetails.jsp?mmnu=1&smnu=2&article_id=2389&proceeding_id=29.

[16] James Hensman, Alexander G. de G. Matthews, and Zoubin Ghahramani. Scalable variational gaussian process classification. In Guy Lebanon and S. V. N. Vishwanathan, editors, *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2015, San Diego, California, USA, May 9-12, 2015*, volume 38 of *JMLR Workshop and Conference Proceedings*. JMLR.org, 2015. URL http://proceedings.mlr.press/v38/hensman15.html.

[17] P. Hohenberg and W. Kohn. Inhomogeneous electron gas. *Physical Review*, 136(3B):B864–B871, November 1964. ISSN 0031-899X. doi: 10.1103/physrev.136.b864. URL http://dx.doi.org/10.1103/PhysRev.136.B864.

[18] Johannes Hoja, Leonardo Medrano Sandonas, Brian G. Ernst, Alvaro Vazquez-Mayagoitia, Robert A. DiStasio, and Alexandre Tkatchenko. Qm7-x, a comprehensive dataset of quantum-mechanical properties spanning the chemical space of small organic molecules. *Scientific Data*, 8(1), February 2021. ISSN 2052-4463. doi: 10.1038/s41597-021-00812-2. URL http://dx.doi.org/10.1038/s41597-021-00812-2.

[19] Yuge Hu, Joseph Musielewicz, Zachary W Ulissi, and Andrew J Medford. Robust and scalable uncertainty estimation with conformal prediction for machine-learned interatomic potentials. *Machine Learning: Science and Technology*, 3(4):045028, December 2022. ISSN 2632-2153. doi: 10.1088/2632-2153/aca7b1. URL http://dx.doi.org/10.1088/2632-2153/aca7b1.

[20] Diederik P. Kingma, Tim Salimans, and Max Welling. Variational dropout and the local reparameterization trick. In Corinna Cortes, Neil D. Lawrence, Daniel D. Lee, Masashi Sugiyama, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 2575–2583, 2015. URL https://proceedings.neurips.cc/paper/2015/hash/bc7316929fe1545bf0b98d114ee3ecb8-Abstract.html.

[21] Johannes Klicpera, Janek Groß, and Stephan Günnemann. Directional message passing for molecular graphs. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.

[22] W. Kohn and L. J. Sham. Self-consistent equations including exchange and correlation effects. *Physical Review*, 140(4A):A1133–A1138, November 1965. ISSN 0031-899X. doi: 10.1103/physrev.140.a1133. URL http://dx.doi.org/10.1103/PhysRev.140.A1133.

[23] Anna-Kathrin Kopetzki, Bertrand Charpentier, Daniel Zügner, Sandhya Giri, and Stephan Günnemann. Evaluating robustness of predictive uncertainty estimation: Are dirichlet-based models reliable? In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 5707–5718. PMLR, 2021. URL http://proceedings.mlr.press/v139/kopetzki21a.html.

[24] Volodymyr Kuleshov, Nathan Fenner, and Stefano Ermon. Accurate uncertainties for deep learning using calibrated regression. In Jennifer G. Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 2801–2809. PMLR, 2018. URL http://proceedings.mlr.press/v80/kuleshov18a.html.

[25] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017.

[26] Haoyang Li, Xin Wang, Ziwei Zhang, and Wenwu Zhu. Out-of-distribution generalization on graphs: A survey, 2022. URL https://arxiv.org/abs/2202.07987.

[27] Yi-Lun Liao and Tess E. Smidt. Equiformer: Equivariant graph attention transformer for 3D atomistic graphs. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023.

[28] Yi-Lun Liao, Brandon Wood, Abhishek Das*, and Tess Smidt*. EquiformerV2: Improved Equivariant Transformer for Scaling to Higher-Degree Representations. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/forum?id=mCOBKZmrzD.

[29] Peter Lippmann, Roman Remme, Manuel Viktor Klockow, and Fred A. Hamprecht. Take note: Your molecular dataset is probably aligned. In Carl Vondrick, Bharath Hariharan, Colin Raffel, Lerrel Pinto, Diyi Yang, and Aleksandra Faust, editors, *The Fourteenth International Conference on Learning Representations, ICLR 2026, Rio de Janeiro, Brazil, April 23-27, 2026*. proceedings.iclr.cc, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/783b0aae591adfccad4a325abdd00807-Abstract-Conference.html.

[30] Yuhang Liu, Wenyong Dong, Lei Zhang, Dong Gong, and Qinfeng Shi. Variational bayesian dropout with a hierarchical prior. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 7124–7133. Computer Vision Foundation / IEEE, 2019. doi: 10.1109/CVPR.2019.00729. URL http://openaccess.thecvf.com/content_CVPR_2019/html/Liu_Variational_Bayesian_Dropout_With_a_Hierarchical_Prior_CVPR_2019_paper.html.

[31] Nis Meinert and Alexander Lavin. Multivariate deep evidential regression. *arXiv preprint arXiv:2104.06135*, 2021.

[32] Sebastian W. Ober, Carl E. Rasmussen, and Mark van der Wilk. The promises and pitfalls of deep kernel learning. In Cassio de Campos and Marloes H. Maathuis, editors, *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence*, volume 161 of *Proceedings of Machine Learning Research*, pages 1206–1216. PMLR, 27–30 Jul 2021. URL https://proceedings.mlr.press/v161/ober21a.html.

[33] Saro Passaro and C. Lawrence Zitnick. Reducing SO(3) convolutions to SO(2) for efficient equivariant gnns. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 27420–27438. PMLR, 2023.

[34] Tim Pearce, Felix Leibfried, and Alexandra Brintrup. Uncertainty in neural networks: Approximately bayesian ensembling. In *International conference on artificial intelligence and statistics*, pages 234–244. PMLR, 2020.

[35] Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian processes for machine learning*. Adaptive computation and machine learning. MIT Press, 2006. ISBN 026218253X. URL https://www.worldcat.org/oclc/61285753.

[36] K. T. Schütt, F. Arbabzadah, S. Chmiela, K. R. Müller, and A. Tkatchenko. Quantum-chemical insights from deep tensor neural networks. *Nat Commun*, 8:13890, January 2017. ISSN 2041-1723 (Electronic) 2041-1723 (Linking). doi: 10.1038/ncomms13890.

[37] Kristof Schütt, Pieter-Jan Kindermans, Huziel Enoc Sauceda Felix, Stefan Chmiela, Alexandre Tkatchenko, and Klaus-Robert Müller. SchNet: A continuous-filter convolutional neural network for modeling quantum interactions. In *Advances in Neural Information Processing Systems*, pages 992–1002, 2017.

[38] Kristof Schütt, Oliver T. Unke, and Michael Gastegger. Equivariant message passing for the prediction of tensorial properties and molecular spectra. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 9377–9388. PMLR, 2021. URL http://proceedings.mlr.press/v139/schutt21a.html.

[39] Ava P. Soleimany, Alexander Amini, Samuel Goldman, Daniela Rus, Sangeeta N. Bhatia, and Connor W. Coley. Evidential deep learning for guided molecular property prediction and discovery. *ACS Central Science*, 7(8):1356–1367, July 2021. ISSN 2374-7951. doi: 10.1021/acscentsci.1c00546. URL http://dx.doi.org/10.1021/acscentsci.1c00546.

[40] Sina Stocker, Johannes Gasteiger, Florian Becker, Stephan Günnemann, and Johannes T Margraf. How robust are modern graph neural network potentials in long and hot molecular dynamics simulations? *Machine Learning: Science and Technology*, 3(4):045010, November 2022. ISSN 2632-2153. doi: 10.1088/2632-2153/ac9955. URL http://dx.doi.org/10.1088/2632-2153/ac9955.

[41] Emma Svensson, Hannah Rosa Friesacher, Susanne Winiwarter, Lewis Mervin, Adam Arany, and Ola Engkvist. Enhancing uncertainty quantification in drug discovery with censored regression labels. *Artificial Intelligence in the Life Sciences*, 7:100128, 2025.

[42] Aik Rui Tan, Shingo Urata, Samuel Goldman, Johannes C. B. Dietschreit, and Rafael Gómez-Bombarelli. Single-model uncertainty quantification in neural network potentials does not consistently outperform model ensembles. *npj Computational Materials*, 9(1), December 2023. ISSN 2057-3960. doi: 10.1038/s41524-023-01180-8. URL http://dx.doi.org/10.1038/s41524-023-01180-8.

[43] 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, 2018. URL https://arxiv.org/abs/1802.08219.

[44] Gia-Lac Tran, Edwin V. Bonilla, John Cunningham, Pietro Michiardi, and Maurizio Filippone. Calibrating deep convolutional gaussian processes. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 1554–1563. PMLR, 16–18 Apr 2019. URL https://proceedings.mlr.press/v89/tran19a.html.

[45] Kevin Tran, Willie Neiswanger, Junwoong Yoon, Qingyang Zhang, Eric Xing, and Zachary W Ulissi. Methods for comparing uncertainty quantifications for material property predictions. *Machine Learning: Science and Technology*, 1(2):025006, May 2020. ISSN 2632-2153. doi: 10.1088/2632-2153/ab7e1a. URL http://dx.doi.org/10.1088/2632-2153/ab7e1a.

[46] Joost van Amersfoort, Lewis Smith, Andrew Jesson, Oscar Key, and Yarin Gal. On feature collapse and deep kernel learning for single forward pass uncertainty, 2022. URL https://arxiv.org/abs/2102.11409.

[47] Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*. Springer, Boston, MA, 2005. ISBN 978-0-387-25061-8.

[48] Sida Wang and Christopher D. Manning. Fast dropout training. In *Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013*, volume 28 of *JMLR Workshop and Conference Proceedings*, pages 118–126. JMLR.org, 2013. URL http://proceedings.mlr.press/v28/wang13a.html.

[49] Maurice Weiler, Mario Geiger, Max Welling, Wouter Boomsma, and Taco Cohen. 3d steerable cnns: Learning rotationally equivariant features in volumetric data. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 10402–10413, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/488e4104520c6aab692863cc1dba45af-Abstract.html.

[50] Maurice Weiler, Fred A. Hamprecht, and Martin Storath. Learning steerable filters for rotation equivariant cnns. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 849–858. Computer Vision Foundation / IEEE Computer Society, 2018. doi: 10.1109/CVPR.2018.00095. URL http://openaccess.thecvf.com/content_cvpr_2018/html/Weiler_Learning_Steerable_Filters_CVPR_2018_paper.html.

[51] Yeming Wen, Dustin Tran, and Jimmy Ba. Batchensemble: an alternative approach to efficient ensemble and lifelong learning. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL https://openreview.net/forum?id=Sklf1yrYDr.

[52] Florian Wenzel, Jasper Snoek, Dustin Tran, and Rodolphe Jenatton. Hyperparameter ensembles for robustness and uncertainty quantification. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/481fbfa59da2581098e841b7afc122f1-Abstract.html.

[53] Andrew Gordon Wilson, Zhiting Hu, Ruslan Salakhutdinov, and Eric P. Xing. Deep kernel learning. In Arthur Gretton and Christian C. Robert, editors, *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics, AISTATS 2016, Cadiz, Spain, May 9-11, 2016*, volume 51 of *JMLR Workshop and Conference Proceedings*, pages 370–378. JMLR.org, 2016. URL http://proceedings.mlr.press/v51/wilson16.html.

[54] Tom Wollschläger, Nicholas Gao, Bertrand Charpentier, Mohamed Amine Ketata, and Stephan Günnemann. Uncertainty estimation for molecules: Desiderata and methods. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 37133–37156. PMLR, 2023. URL https://proceedings.mlr.press/v202/wollschlager23a.html.

[55] Miao Zhong, Kevin Tran, Yimeng Min, Chuanhao Wang, Ziyun Wang, Cao-Thang Dinh, Phil De Luna, Zongqian Yu, Armin Sedighian Rasouli, Peter Brodersen, Song Sun, Oleksandr Voznyy, Chih-Shan Tan, Mikhail Askerka, Fanglin Che, Min Liu, Ali Seifitokaldani, Yuanjie Pang, Shen-Chuan Lo, Alexander Ip, Zachary Ulissi, and Edward H. Sargent. Accelerated discovery of CO2 electrocatalysts using active machine learning. *Nature*, 581(7807):178–183, May 2020. ISSN 0028-0836, 1476-4687. doi: 10.1038/s41586-020-2242-8.

[56] Larry Zitnick, Abhishek Das, Adeesh Kolluru, Janice Lan, Muhammed Shuaibi, Anuroop Sriram, Zachary W. Ulissi, and Brandon M. Wood. Spherical channels for modeling atomic interactions. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022.

A Appendix

A.1 Proof of Proposition 3.1

We prove the three properties in turn.

Positive Definiteness (PD): Since $\alpha \in \mathbb{R}$, $\exp(\alpha) > 0$. The first term, $\exp(\alpha)\mathbf{I}$, is a PD matrix. The second term, $\boldsymbol{F}\boldsymbol{F}^\top$, is a PSD matrix. As the sum of a PD matrix and a PSD matrix is always PD, $\boldsymbol{\Psi}$ is guaranteed to be PD.

Equivariance: The equivariant nature of $\boldsymbol{F}$ means an input rotation $\mathbf{R}$ generates a new matrix $\boldsymbol{F}' = \mathbf{R}\boldsymbol{F}$, while the invariant scalar α remains unchanged. The resulting scale matrix $\boldsymbol{\Psi}'$ then correctly transforms under rotation:

$$\boldsymbol{\Psi}' = \exp(\alpha)\mathbf{I} + \boldsymbol{F}'(\boldsymbol{F}')^\top = \exp(\alpha)\mathbf{I} + (\mathbf{R}\boldsymbol{F})(\mathbf{R}\boldsymbol{F})^\top = \mathbf{R}\left(\exp(\alpha)\mathbf{I} + \boldsymbol{F}\boldsymbol{F}^\top\right)\mathbf{R}^\top = \mathbf{R}\boldsymbol{\Psi}\mathbf{R}^\top$$

This demonstrates that the proposed parameterization satisfies the equivariance property.

Residual non-identifiability. For $k = 2$, the factor $\boldsymbol{F} \in \mathbb{R}^{3\times 2}$ is not identifiable from $\boldsymbol{F}\boldsymbol{F}^\top$. Specifically, for any orthogonal matrix $\boldsymbol{Q} \in O(2)$ satisfying $\boldsymbol{Q}\boldsymbol{Q}^\top = \boldsymbol{I}_2$, we have

$$(\boldsymbol{F}\boldsymbol{Q})(\boldsymbol{F}\boldsymbol{Q})^\top = \boldsymbol{F}\boldsymbol{Q}\boldsymbol{Q}^\top\boldsymbol{F}^\top = \boldsymbol{F}\boldsymbol{F}^\top.$$

Therefore,

$$\boldsymbol{\Psi}(\alpha, \boldsymbol{F}\boldsymbol{Q}) = \exp(\alpha)\boldsymbol{I} + (\boldsymbol{F}\boldsymbol{Q})(\boldsymbol{F}\boldsymbol{Q})^\top = \boldsymbol{\Psi}(\alpha, \boldsymbol{F}).$$

This shows that the parameterization is invariant to right multiplication of $\boldsymbol{F}$ by any $\boldsymbol{Q} \in O(2)$. Hence the ambiguity is not limited to independent sign flips; rather, it corresponds to a two-dimensional orthogonal change of basis within the column space of $\boldsymbol{F}$.

This residual ambiguity is nevertheless smaller than the full 3×3 factorization ambiguity. If one used a full factor $\boldsymbol{F} \in \mathbb{R}^{3\times 3}$, then $\boldsymbol{F}\boldsymbol{F}^\top$ would be invariant under the full group $O(3)$. In our main setting with $k = 2$, the remaining non-identifiability is reduced to $O(2)$, while the isotropic term $\exp(\alpha)\boldsymbol{I}$ guarantees positive definiteness and the low-rank term captures anisotropic equivariant uncertainty.

A.2 An Identifiable Gram–Schmidt Variant

The main EMER parameterization used in the paper is deliberately simple:

$$\boldsymbol{\Psi}(\alpha, \boldsymbol{F}) = \exp(\alpha)\boldsymbol{I} + \boldsymbol{F}\boldsymbol{F}^\top, \qquad \boldsymbol{F} \in \mathbb{R}^{3\times 2}.$$

As shown above, this form is equivariant and PD, but it is not fully identifiable because $\boldsymbol{F}\boldsymbol{F}^\top$ is invariant under $\boldsymbol{F} \mapsto \boldsymbol{F}\boldsymbol{Q}$ for any $\boldsymbol{Q} \in O(2)$. We also considered a more constrained variant based on Gram–Schmidt orthogonalization (denoted as GSMER), which removes this continuous $O(2)$ ambiguity under mild non-degeneracy assumptions.

Proposition A.1. *Let the NIW scale matrix be parameterized as*

$$\boldsymbol{\Psi}(\alpha, \boldsymbol{h}_1, \boldsymbol{h}_2) = \exp(\alpha)\boldsymbol{I} + \beta_1\boldsymbol{u}_1\boldsymbol{u}_1^\top + \beta_2\boldsymbol{u}_2\boldsymbol{u}_2^\top,$$

where $\alpha \in \mathbb{R}$ and $\boldsymbol{h}_1, \boldsymbol{h}_2 \in \mathbb{R}^3$ are equivariant vectors (EGNN outputs). Define

$$\boldsymbol{u}_1 = \frac{\boldsymbol{h}_1}{\|\boldsymbol{h}_1\|}, \qquad \widetilde{\boldsymbol{h}}_2 = \boldsymbol{h}_2 - (\boldsymbol{u}_1^\top\boldsymbol{h}_2)\boldsymbol{u}_1, \qquad \boldsymbol{u}_2 = \frac{\widetilde{\boldsymbol{h}}_2}{\|\widetilde{\boldsymbol{h}}_2\|},$$

and

$$\beta_2 = \|\widetilde{\boldsymbol{h}}_2\|^2, \qquad \beta_1 = \|\boldsymbol{h}_1\|^2 + \beta_2.$$

Assume $\boldsymbol{h}_1 \neq \mathbf{0}$ and $\widetilde{\boldsymbol{h}}_2 \neq \mathbf{0}$. Then:

- *(i)* $\boldsymbol{\Psi}$ *is symmetric PD.*
- *(ii)* $\boldsymbol{\Psi}$ *is equivariant: for any $\boldsymbol{R} \in \mathrm{SO}(3)$,*

$$\boldsymbol{\Psi}(\alpha, \boldsymbol{R}\boldsymbol{h}_1, \boldsymbol{R}\boldsymbol{h}_2) = \boldsymbol{R}\boldsymbol{\Psi}(\alpha, \boldsymbol{h}_1, \boldsymbol{h}_2)\boldsymbol{R}^\top.$$

(iii) The representation is identifiable up to sign flips of the orthonormal directions when the anisotropic eigenvalues are distinct. More precisely, if

$$\exp(\alpha)\boldsymbol{I} + \beta_1 \boldsymbol{u}_1\boldsymbol{u}_1^\top + \beta_2 \boldsymbol{u}_2\boldsymbol{u}_2^\top = \exp(\alpha')\boldsymbol{I} + \beta_1' \boldsymbol{u}_1'(\boldsymbol{u}_1')^\top + \beta_2' \boldsymbol{u}_2'(\boldsymbol{u}_2')^\top,$$

where $(\boldsymbol{u}_1, \boldsymbol{u}_2)$ and $(\boldsymbol{u}_1', \boldsymbol{u}_2')$ are orthonormal pairs and the coefficients are ordered as

$$\beta_1 \geq \beta_2 \geq 0, \qquad \beta_1' \geq \beta_2' \geq 0,$$

then, whenever $\beta_1 > \beta_2$, we have

$$\alpha = \alpha', \qquad \beta_1 = \beta_1', \qquad \beta_2 = \beta_2',$$

and

$$\boldsymbol{u}_1' = \pm \boldsymbol{u}_1, \qquad \boldsymbol{u}_2' = \pm \boldsymbol{u}_2.$$

If $\beta_1 = \beta_2$, only the two-dimensional eigenspace $\operatorname{span}\{\boldsymbol{u}_1, \boldsymbol{u}_2\}$ is identifiable.

Proof. We prove the three claims in turn.

Positive definiteness. Since $\alpha \in \mathbb{R}$, $\exp(\alpha) > 0$, so $\exp(\alpha)\boldsymbol{I}$ is PD. By construction,

$$\beta_2 = \|\widetilde{\boldsymbol{h}}_2\|^2 \geq 0, \qquad \beta_1 = \|\boldsymbol{h}_1\|^2 + \beta_2 \geq 0.$$

For any $\boldsymbol{x} \in \mathbb{R}^3$, the anisotropic term satisfies

$$\begin{aligned}\boldsymbol{x}^\top \left(\beta_1 \boldsymbol{u}_1\boldsymbol{u}_1^\top + \beta_2 \boldsymbol{u}_2\boldsymbol{u}_2^\top\right) \boldsymbol{x} &= \beta_1(\boldsymbol{u}_1^\top \boldsymbol{x})^2 + \beta_2(\boldsymbol{u}_2^\top \boldsymbol{x})^2 \\ &\geq 0.\end{aligned}$$

Thus the anisotropic term is PSD. Therefore $\boldsymbol{\Psi}$ is the sum of a PD matrix and a PSD matrix, and is positive definite.

Equivariance. Let $\boldsymbol{R} \in \mathrm{SO}(3)$. Since rotations preserve norms and inner products,

$$\|\boldsymbol{R}\boldsymbol{h}_1\| = \|\boldsymbol{h}_1\|, \qquad (\boldsymbol{R}\boldsymbol{a})^\top(\boldsymbol{R}\boldsymbol{b}) = \boldsymbol{a}^\top \boldsymbol{b}.$$

The first normalized direction transforms as

$$\boldsymbol{u}_1' = \frac{\boldsymbol{R}\boldsymbol{h}_1}{\|\boldsymbol{R}\boldsymbol{h}_1\|} = \boldsymbol{R}\frac{\boldsymbol{h}_1}{\|\boldsymbol{h}_1\|} = \boldsymbol{R}\boldsymbol{u}_1.$$

For the second Gram–Schmidt residual, we have

$$\begin{aligned}\widetilde{\boldsymbol{h}}_2' &= \boldsymbol{R}\boldsymbol{h}_2 - ((\boldsymbol{u}_1')^\top \boldsymbol{R}\boldsymbol{h}_2)\boldsymbol{u}_1' \\ &= \boldsymbol{R}\boldsymbol{h}_2 - ((\boldsymbol{R}\boldsymbol{u}_1)^\top \boldsymbol{R}\boldsymbol{h}_2)\boldsymbol{R}\boldsymbol{u}_1 \\ &= \boldsymbol{R}\boldsymbol{h}_2 - (\boldsymbol{u}_1^\top \boldsymbol{h}_2)\boldsymbol{R}\boldsymbol{u}_1 \\ &= \boldsymbol{R}\left(\boldsymbol{h}_2 - (\boldsymbol{u}_1^\top \boldsymbol{h}_2)\boldsymbol{u}_1\right) \\ &= \boldsymbol{R}\widetilde{\boldsymbol{h}}_2.\end{aligned}$$

Hence

$$\boldsymbol{u}_2' = \frac{\widetilde{\boldsymbol{h}}_2'}{\|\widetilde{\boldsymbol{h}}_2'\|} = \frac{\boldsymbol{R}\widetilde{\boldsymbol{h}}_2}{\|\widetilde{\boldsymbol{h}}_2\|} = \boldsymbol{R}\boldsymbol{u}_2.$$

The coefficients are invariant:

$$\beta_2' = \|\widetilde{\boldsymbol{h}}_2'\|^2 = \|\boldsymbol{R}\widetilde{\boldsymbol{h}}_2\|^2 = \beta_2,$$

and

$$\beta_1' = \|\boldsymbol{R}\boldsymbol{h}_1\|^2 + \beta_2' = \|\boldsymbol{h}_1\|^2 + \beta_2 = \beta_1.$$

Therefore,

$$\begin{aligned}\boldsymbol{\Psi}(\alpha, \boldsymbol{R}\boldsymbol{h}_1, \boldsymbol{R}\boldsymbol{h}_2) &= \exp(\alpha)\boldsymbol{I} + \beta_1(\boldsymbol{R}\boldsymbol{u}_1)(\boldsymbol{R}\boldsymbol{u}_1)^\top + \beta_2(\boldsymbol{R}\boldsymbol{u}_2)(\boldsymbol{R}\boldsymbol{u}_2)^\top \\ &= \boldsymbol{R}\left(\exp(\alpha)\boldsymbol{I} + \beta_1 \boldsymbol{u}_1\boldsymbol{u}_1^\top + \beta_2 \boldsymbol{u}_2\boldsymbol{u}_2^\top\right)\boldsymbol{R}^\top \\ &= \boldsymbol{R}\boldsymbol{\Psi}(\alpha, \boldsymbol{h}_1, \boldsymbol{h}_2)\boldsymbol{R}^\top.\end{aligned}$$

Thus the parameterization is equivariant.

Identifiability. Assume two parameterizations give the same matrix:

$$\exp(\alpha)\boldsymbol{I} + \beta_1 \boldsymbol{u}_1 \boldsymbol{u}_1^\top + \beta_2 \boldsymbol{u}_2 \boldsymbol{u}_2^\top = \exp(\alpha')\boldsymbol{I} + \beta_1' \boldsymbol{u}_1' (\boldsymbol{u}_1')^\top + \beta_2' \boldsymbol{u}_2' (\boldsymbol{u}_2')^\top.$$

Because $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$ are orthonormal, the anisotropic matrix

$$\boldsymbol{A} = \beta_1 \boldsymbol{u}_1 \boldsymbol{u}_1^\top + \beta_2 \boldsymbol{u}_2 \boldsymbol{u}_2^\top$$

has eigenvalues β_1, β_2, and 0, with eigenspaces spanned by $\boldsymbol{u}_1$, $\boldsymbol{u}_2$, and the one-dimensional orthogonal complement, respectively. Therefore $\boldsymbol{\Psi}$ has eigenvalues

$$\exp(\alpha) + \beta_1, \qquad \exp(\alpha) + \beta_2, \qquad \exp(\alpha).$$

The second representation has eigenvalues

$$\exp(\alpha') + \beta_1', \qquad \exp(\alpha') + \beta_2', \qquad \exp(\alpha').$$

Since the two matrices are equal, their eigenvalues coincide. The smallest eigenvalue is the isotropic floor, so

$$\exp(\alpha) = \exp(\alpha'), \qquad \text{and hence} \qquad \alpha = \alpha'.$$

Subtracting $\exp(\alpha)\boldsymbol{I}$ from both sides gives equality of the anisotropic parts. Since their ordered eigenvalues coincide, we obtain

$$\beta_1 = \beta_1', \qquad \beta_2 = \beta_2'.$$

If $\beta_1 > \beta_2$, the eigenspaces associated with β_1 and β_2 are one-dimensional and therefore unique. Hence the corresponding unit eigenvectors are unique up to sign:

$$\boldsymbol{u}_1' = \pm \boldsymbol{u}_1, \qquad \boldsymbol{u}_2' = \pm \boldsymbol{u}_2.$$

Thus, under distinct anisotropic eigenvalues, the representation is identifiable up to independent sign flips.

If $\beta_1 = \beta_2$, the anisotropic term reduces to

$$\beta_1 \left(\boldsymbol{u}_1 \boldsymbol{u}_1^\top + \boldsymbol{u}_2 \boldsymbol{u}_2^\top \right),$$

which depends only on the projector onto the subspace $\mathrm{span}\{\boldsymbol{u}_1, \boldsymbol{u}_2\}$, not on the particular orthonormal basis chosen inside this subspace. Therefore, in this degenerate case, only the two-dimensional eigenspace is identifiable.

A.3 Implementation details

Backbones Equivariant graph neural networks were employed due to their vector-prediction capabilities. Specifically, this study utilizes PaiNN [38], NequIP [4], and MACE [3]. All models were trained using the default hyperparameters from their respective original works, with two exceptions. First, for PaiNN on OC20, the number of interaction blocks was set to 4 and the hidden dimension to 512. Second, to improve baseline accuracy on QM7X, the hidden irreps for NequIP and MACE were increased to `64x0o+64x0e+64x1o+64x1e+64x2o+64x2e` and `128x0e+128x1o+128x2e`, respectively.

The backbones are trained using a standard combined energy and force loss function:

$$\mathcal{L} = (1 - \rho)\, \mathcal{L}_{\text{energy}} + \rho \mathcal{L}_{\text{force}} \tag{7}$$

where $\mathcal{L}_{\text{energy}}$ is the energy loss for the total energy (or adsorption energy for OC20), $\mathcal{L}_{\text{force}}$ is the force loss, and $\rho \in [0, 1]$ is a weighting coefficient that balances the contribution of the two terms. Specifically, the weighting coefficient ρ and the choice of loss functions for the energy ($\mathcal{L}_{\text{energy}}$) and force ($\mathcal{L}_{\text{force}}$) terms are detailed in Table 4 for the MD17 and OC20 datasets. For the task of dipole moment prediction on the QM7X dataset, the MER loss was employed for the MER and EMER families, L2 loss was used for PaiNN and L1 loss was utilized for NequIP and MACE, as detailed in Table 4. For QM7-X, MACE achieves its best dipole accuracy with the MER loss (Table 2); L1 is used only for the deterministic NequIP/MACE backbones.

Dataset & Method	Weighting ρ	$\mathcal{L}_{\text{energy}}$	$\mathcal{L}_{\text{force}}$	$\mathcal{L}_{\text{DP}}$
MD17	0.99	L1	L2	–
MD17+EMER	0.99	L1	MER	–
OC20	0.96	L1	L1	–
OC20+EMER	0.7	L1	MER	–
QM7X-PaiNN Backbones	–	–	–	L2
QM7X-NequIP/MACE Backbones	–	–	–	L1
QM7X-PaiNN/NequIP/MACE EMER	–	–	–	MER

Table 4: Loss function hyperparameters for different benchmarks

	MD17	OC20	QM7X
batch size	10	32	32
num of train	1000	90000	5584
num of test	1000	10000	698

Table 5: Hyperparameters of datasets used with all models

Vector Output Head Details For both the MD17 and OC20 datasets, forces are computed as the negative gradient of the predicted energy.

To predict dipole moments, we follow the approach of Schütt et al. [38], given by

$$\vec{\mu} = \sum_{i=1}^{N} \left(\vec{\mu}_{\text{atom}}(\vec{v}_i) + q_{\text{atom}}(s_i)\vec{r}_i \right),$$

where $\vec{\mu}_{\text{atom}}$ is the local atomic dipole, q_{atom} is the learned partial atomic charge, and $\vec{r}_i$ is the atomic position vector. This formulation constructs the total dipole moment by summing contributions from both induced atomic dipoles and the dipoles generated by the learned atomic charges. For the dipole covariance, we mean-pool the per-atom α_i and $\boldsymbol{L}_i$ and form $\boldsymbol{\Psi} = \exp(\alpha)\mathbf{I} + \boldsymbol{L}\boldsymbol{L}^\top$, which preserves equivariance. The dipole mean is therefore size-extensive while its covariance is size-intensive, which suffices for the in-distribution molecular sizes of QM7-X; how the covariance should scale under size shift is left open.

Monte Carlo Dropout Bernoulli Dropout is employed for MC Dropout uncertainty estimation. It is exclusively applied to the features following each message-passing interaction block. Crucially, no dropout is utilized within the interaction blocks or the final readout layers. To maintain $SO(3)$-equivariance, we apply Equivariant Dropout [27, 28] to all tensorial features ($l \geq 0$), ensuring that the transformation properties of the representations are preserved. The hyperparameters used for MC Dropout and the backbone are listed in Table 6 and Table 7.

Model / Hyperparameter	MD17	OC20	OC20 MER/EMER family
PaiNN			
learning rate	0.001	0.00025	0.00025
warmup steps	500	20000	20000
decay steps	40000	2000000	2000000
decay rate	0.01	0.01	0.01
EMA decay	0.99	0.99	0.99
Epochs	400	120	120
Batch size	10	32	32
force weighting factor ρ	0.99	0.96	0.70
PaiNN Dropout			
learning rate	0.001	0.00025	0.00025
warmup steps	500	20000	20000
decay steps	40000	2000000	2000000
decay rate	0.01	0.01	0.01
EMA decay	0.99	0.99	0.99
Epochs	400	120	120
Batch size	10	32	32
force weighting factor ρ	0.99	0.96	0.70
dropout ratio	0.01	0.1	0.1
dropout locations	after every interaction block		

Table 6: Hyperparameters of PaiNN

Hyperparameter	PaiNN	NequIP	MACE
learning rate	0.001	0.01	0.01
warmup epochs	1	1	1
decay epochs	200	250	300
decay rate	0.01	0.01	0.01
EMA decay	0.99	0.99	0.99
Epochs	200	250	300
Batch size	10	32	32
Dropout			
learning rate	0.001	0.01	0.01
warmup epochs	1	1	1
decay epochs	200	250	300
decay rate	0.01	0.01	0.01
EMA decay	0.99	0.99	0.99
Epochs	200	250	300
Batch size	10	32	32
dropout ratio ρ (Drop./-D)	0.1	0.1	0.1/0.4
dropout locations	after every interaction block		

Table 7: Hyperparameters of PaiNN, NequIP, MACE for QM7X dipole moment prediction. For MACE, the dropout ratio is 0.1 for Drop. and 0.4 for MER-D/EMER-D.

Classification and Regression Diffusion We implemented the Classification and Regression Diffusion (CARD) model [14] as a conditional generative head attached to pre-trained equivariant backbones (e.g., PaiNN, MACE). Since CARD is originally designed for scalar regression, we adapted it to vector-valued prediction by flattening the $(N, 3, m)$ tensor of $l = 1$ irreps into a 2D tensor of shape $(N \times 3, m)$, where each of the x, y, and z force components is treated as an independent scalar variable. This adaptation effectively removes the coupling between vector components, meaning CARD receives no information about rotational structure beyond the backbone's atomic embeddings and its deterministic force prediction.

The diffusion process uses 1,000 timesteps with a linear beta schedule from $\beta_{\text{start}} = 10^{-4}$ to $\beta_{\text{end}} = 0.02$. The model is trained for 100 epochs using Adam (learning rate 10^{-3}, batch size 256), minimizing the mean-squared error between the predicted and true noise. Gradient norms are clipped to 1.0.

For uncertainty estimation, inference is performed via Monte Carlo reverse diffusion with $K = 100$ samples per atom, starting from standard Gaussian noise. The predictive mean is computed as the sample average, and uncertainty is quantified by the sample variance. Because the model predicts each coordinate independently, the estimated covariance matrix is strictly diagonal. Consequently, this scalarized CARD adaptation cannot model cross-coordinate correlations and breaks rotational equivariance.

Compute Resources Most experiments were conducted on a local workstation equipped with a 12th Gen Intel® Core™ i9-12900KF CPU with 24 threads, one NVIDIA RTX 4090 GPU, and 128 GB RAM. For the larger-scale OC20 experiments, we additionally used a workstation equipped with four NVIDIA H100 GPUs with 80 GB memory each. These resources were sufficient to train the reported models and reproduce the main experimental results.

A.4 Evaluation Metrics

To rigorously assess the reliability of our proposed framework, we employ three complementary metrics: the Mahalanobis Calibration Score (MCal) for vector-valued consistency proposed in this work, the dimension-wise Raw Calibration Score (RCal) and dimension-wise Negative Log Likelihood (NLL) for marginal consistency.

Mahalanobis Calibration Score (MCal) Standard calibration metrics designed for scalar regression are ill-suited for equivariant vector quantities, as they fail to account for correlations between components and rotational symmetries. The Mahalanobis Calibration Score is designed to address this problem by evaluating calibration through a coordinate-independent distance statistic. For a predicted mean vector $\boldsymbol{\mu}_i$ and covariance matrix $\boldsymbol{\Sigma}_i$ corresponding to the ground truth $\mathbf{v}_{\text{true},i}$, the squared Mahalanobis distance is defined as:

$$d_i^2 = (\mathbf{v}_{\text{true},i} - \boldsymbol{\mu}_i)^\top \boldsymbol{\Sigma}_i^{-1} (\mathbf{v}_{\text{true},i} - \boldsymbol{\mu}_i).$$

The reference distribution of $\{d_i^2\}_{i=1}^N$ depends on the predictive distribution. If the model is perfectly calibrated and the predictive distribution is multivariate normal, then d_i^2 follows a Chi-squared distribution with n degrees of freedom (χ_n^2), where n is the dimensionality of the vector (here, $n = 3$). In contrast, for evidential regression models such as MER and EMER, the NIW prior induces a multivariate Student's t predictive distribution, for which the corresponding Mahalanobis statistic follows a scaled F distribution. We therefore define MCal by comparing the empirical cumulative distribution of the observed distances against the theoretical reference CDF implied by the predictive family. Let $F_{\text{ref}}(z)$ denote this reference CDF evaluated at threshold z. We compute the empirical frequency of observed distances falling below the theoretical quantile associated with probability level p:

$$\hat{p}(p) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}\left[d_i^2 \leq F_{\text{ref}}^{-1}(p)\right].$$

The MCal score is then calculated as the root sum of squared error between the empirical frequency and the expected probability levels $p \in \{0.1, \ldots, 0.9\}$:

$$\text{MCal} = \sqrt{\sum_{j=1}^{M} (\hat{p}(p_j) - p_j)^2}.$$

A lower MCal indicates that the model's multivariate uncertainty ellipsoids encapsulate the ground truth with the expected frequency under the appropriate predictive reference distribution. MCal is the primary coordinate-independent calibration metric for vector uncertainty, but it should be read together with Max RCal and NLL to verify that the uncertainty ellipsoid is not only correctly scaled but also correctly oriented.

Dimension-wise Raw Calibration Score (RCal) To facilitate comparison with methods that treat vector components independently, we also report the Raw Calibration Score (RCal) in Wollschläger et al. [54], originally defined by Kuleshov et al. [24]. The empirical probability is computed as the fraction of test samples where the true scalar value falls within this interval:

$$\hat{p}_{\text{marg}}(p) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}\left[y_{i,d} \in I_p^{(i,d)} \right]$$

where $y_{i,d}$ is the d-th component of the i-th target vector. The scalar calibration score is defined as:

$$\text{RCal}_d = \sqrt{\sum_{p \in \{0.1,\ldots,0.9\}} (\hat{p}_{\text{marg}}(p) - p)^2}$$

Since an equivariant vector model may exhibit different calibration behaviors along different axes (though ideally, it should not), reporting the average RCal can be misleading. Therefore, we report the Max RCal, which is the maximum calibration error observed across the three spatial dimensions:

$$\text{Max RCal} = \max_{d \in \{x,y,z\}} \text{RCal}_d$$

This serves as a worst-case estimate of marginal calibration.

Negative Log-Likelihood (NLL) While calibration scores assess the alignment of uncertainty estimates, they do not fully capture the sharpness of the predictive uncertainty. We therefore also report a dimension-wise likelihood-based score using the predicted marginal mean and variance. For sampling-based baselines, these marginal statistics are estimated from predictive samples. For MER and EMER, they are obtained from the predictive mean and covariance implied by the NIW parameters. Similar to RCal, we compute the score independently for each spatial dimension and report the maximum value over dimensions:

$$\text{Max NLL} = \max_{d \in \{x,y,z\}} \text{NLL}_d.$$

This provides a worst-case measure of marginal uncertainty quality across spatial components.

A.5 Additional Results

MD17 Table 8 reports the results on the standard MD17 setting, without the RotOOD shift. EMER achieves the strongest predictive accuracy, obtaining the lowest Energy MAE on six out of seven molecules and the lowest Force MAE on four out of seven molecules. For calibration, the three dropout-based variants, MER-D, EMER-D, and GSMER-D, show broadly comparable performance, with EMER-D providing a slightly better overall balance across MCal, Max RCal, and Max NLL. However, the strong calibration of MER-D in this standard setting should be interpreted with caution. Since the training and test sets share similar marginal distributions across spatial dimensions, a non-equivariant uncertainty model can appear well calibrated. As shown in the RotOOD results in Table 1, this apparent advantage collapses once the vector-label distribution is rotated, indicating a potentially serious failure mode in practical deployment.

GSMER-D, which uses the Gram–Schmidt identifiable parameterization, also achieves competitive calibration. In some cases, it improves over EMER-D, such as Aspirin, where MCal decreases from 0.1295 to 0.0742. However, its performance is not uniformly better; for example, on SAC, its MCal increases from 0.1837 to 0.3824. Given that GSMER-D introduces additional implementation complexity through Gram–Schmidt orthogonalization, while offering no consistent empirical gain, we adopt EMER-D as the primary model. Overall, these results suggest that EMER-D provides the best trade-off between predictive accuracy, calibration robustness, implementation simplicity, and rotational consistency.

Mol	Metric	Backbone	Equivariant		Nonequivariant		Ours			
			Ens.	Drop.	CARD	MER	EMER	MER-D	EMER-D	GSMER-D
Aspirin	EMAE	0.2197	0.1873	0.3549	0.2197	0.1943	**0.1897**	0.4059	0.3796	0.4901
	FMAE	0.5184	0.3799	0.4952	0.5184	0.4324	0.4124	0.4145	**0.3898**	0.4000
	MCal	–	0.8167	0.9933	1.0810	1.3582	1.3630	0.1190	0.1295	**0.0742**
	Max RCal	–	0.0801	0.4476	0.7790	0.9652	0.9500	0.1324	0.1305	**0.0701**
	Max NLL	–	0.8120	2.2683	6.9817	7.9218	8.3517	0.8137	0.7556	**0.7388**
Ethanol	EMAE	0.0872	0.0670	0.1801	0.0872	0.0713	**0.0690**	0.1075	0.0971	0.1060
	FMAE	0.3210	0.2364	0.3227	0.3210	0.2837	0.2852	0.2743	0.2795	**0.2723**
	MCal	–	0.6743	0.8974	1.1185	1.4507	1.4741	**0.1391**	0.2927	0.1777
	Max RCal	–	0.0692	0.3698	0.7430	1.0649	1.0593	**0.1370**	0.2339	0.1572
	Max NLL	–	0.2507	1.7626	10.140	16.6358	21.5976	0.5609	**0.5136**	0.5663
Nap	EMAE	0.1249	0.1168	0.3017	0.1249	0.1181	**0.1144**	0.2184	0.2523	0.2642
	FMAE	0.1296	0.0904	0.1520	0.1296	0.1026	**0.1014**	0.1467	0.1537	0.1860
	MCal	–	0.7972	0.3846	0.4395	0.7581	0.7884	0.3564	**0.2380**	0.5376
	Max RCal	–	0.0779	**0.0768**	0.3752	0.5404	0.5923	0.2216	0.1930	0.3347
	Max NLL	–	-0.6304	0.1107	1.0555	0.1782	0.5168	**-0.2100**	-0.1973	0.0394
MDA	EMAE	0.1183	0.1127	0.2214	0.1183	0.1079	**0.1063**	0.1354	0.1562	0.1361
	FMAE	0.4900	0.3790	0.4698	0.4900	0.4348	0.4207	0.4041	0.3997	**0.3948**
	MCal	–	0.7626	1.0735	1.3320	1.4618	1.4919	**0.3527**	0.4289	0.3682
	Max RCal	–	0.0714	0.5027	0.9373	1.0967	1.1125	**0.2740**	0.3217	0.2796
	Max NLL	–	0.8031	2.7720	10.945	26.6849	44.5812	1.2263	**1.0905**	1.1935
SAC	EMAE	0.1477	0.1262	0.2727	0.1477	0.1310	**0.1165**	0.2552	0.2246	0.1841
	FMAE	0.3286	0.2511	0.3455	0.3286	0.2603	**0.2503**	0.2701	0.2628	0.2829
	MCal	–	0.7943	0.7863	0.9283	1.2343	1.2441	0.2441	**0.1837**	0.3824
	Max RCal	–	0.0817	0.3207	0.7549	0.8406	0.8533	0.1493	**0.1352**	0.2201
	Max NLL	–	0.3554	1.4537	6.4255	3.8882	5.7351	0.4785	**0.4615**	0.5561
Uracil	EMAE	0.1195	0.1057	0.2517	0.1195	**0.1063**	0.1117	0.1878	0.1889	0.1838
	FMAE	0.2251	0.1660	0.2589	0.2251	0.1801	**0.1694**	0.2120	0.2056	0.2020
	MCal	–	0.7527	0.4784	0.7958	1.1124	1.1331	0.4507	**0.4094**	0.5192
	Max RCal	–	0.0661	**0.1012**	0.6422	0.8005	0.7762	0.2408	0.2112	0.2959
	Max NLL	–	-0.0920	0.7718	4.6010	2.6979	2.6519	0.2419	**0.1890**	0.2073
Toluene	EMAE	0.1018	0.0929	0.2586	0.1018	0.0976	**0.0938**	0.2548	0.2604	0.1787
	FMAE	0.1511	0.1084	0.1681	0.1511	0.1229	**0.1217**	0.1629	0.1633	0.1722
	MCal	–	0.8090	0.4373	0.5489	0.9405	1.0002	0.1898	0.2405	**0.1651**
	Max RCal	–	0.0749	0.0726	0.4406	0.6190	0.6546	**0.0702**	0.1372	0.1880
	Max NLL	–	-0.4513	0.2003	1.9618	1.0650	1.6419	-0.1779	**-0.1925**	-0.0949

Table 8: Performance metrics on the MD17 dataset. Calibration and NLL are evaluated on atomic forces only. Bold indicates the best single-model performance. Our proposed methods (EMER, MER-D, and EMER-D) achieve the best accuracy and calibration across the majority of tasks.

Additional Baselines Table 9 compares EMER and EMER-D with additional baselines on MD17-RotOOD. Gauss and Gauss-D use the same PaiNN backbone and equivariant covariance $\boldsymbol{\Psi} = \exp(\alpha)\mathbf{I} + \boldsymbol{F}\boldsymbol{F}^\top$ as EMER and EMER-D, but are trained with the Gaussian NLL without the NIW prior, and Gauss-D uses the same dropout ratio as EMER-D. For BLIP [9], we use the prior probability $p = 0.5$ and KL weight $\beta = 0.1$. GMM fits a Gaussian mixture model on the learned features and uses its density as the uncertainty [42]. EMER-D attains a lower force MAE than Gauss-D on all seven molecules, indicating that the NIW prior mainly improves force accuracy. Its calibration benefit is structure-dependent: EMER-D is better calibrated than Gauss-D on the rigid and aromatic molecules (Nap, SAC, Toluene, and Uracil), while Gauss-D is better on Aspirin, Ethanol, and MDA. Replacing $\exp(\alpha)\mathbf{I} + \boldsymbol{F}\boldsymbol{F}^\top$ with $\boldsymbol{L}\boldsymbol{L}^\top$ or $\boldsymbol{L}\boldsymbol{L}^\top + 10^{-3}\mathbf{I}$ was unstable during training even without the NIW prior, so the learned isotropic term is necessary. BLIP and GMM do not provide a consistent advantage over EMER-D.

Mol	Metric	Equivariant			Nonequivariant	Ours	
		Gauss	Gauss-D	BLIP	GMM	EMER	EMER-D
Aspirin	Force MAE	$0.431_{\pm 0.010}$	$0.423_{\pm 0.004}$	$0.477_{\pm 0.004}$	$0.509_{\pm 0.006}$	$0.403_{\pm 0.008}$	$\mathbf{0.394}_{\pm 0.007}$
	MCal	$1.200_{\pm 0.013}$	$\mathbf{0.142}_{\pm 0.038}$	$1.319_{\pm 0.008}$	$0.405_{\pm 0.073}$	$1.361_{\pm 0.024}$	$0.173_{\pm 0.024}$
	Max RCal	$0.755_{\pm 0.007}$	$\mathbf{0.072}_{\pm 0.021}$	$0.875_{\pm 0.009}$	$0.378_{\pm 0.043}$	$0.947_{\pm 0.025}$	$0.154_{\pm 0.019}$
Ethanol	Force MAE	$\mathbf{0.273}_{\pm 0.008}$	$0.279_{\pm 0.013}$	$0.304_{\pm 0.007}$	$0.317_{\pm 0.000}$	$0.277_{\pm 0.012}$	$0.277_{\pm 0.007}$
	MCal	$1.394_{\pm 0.011}$	$\mathbf{0.149}_{\pm 0.003}$	$0.748_{\pm 0.018}$	$0.590_{\pm 0.117}$	$1.476_{\pm 0.010}$	$0.265_{\pm 0.029}$
	Max RCal	$0.926_{\pm 0.022}$	$\mathbf{0.083}_{\pm 0.008}$	$0.448_{\pm 0.009}$	$0.417_{\pm 0.084}$	$1.061_{\pm 0.007}$	$0.217_{\pm 0.023}$
Nap	Force MAE	$0.118_{\pm 0.004}$	$0.171_{\pm 0.012}$	$0.124_{\pm 0.002}$	$0.130_{\pm 0.001}$	$\mathbf{0.101}_{\pm 0.003}$	$0.162_{\pm 0.006}$
	MCal	$0.563_{\pm 0.089}$	$0.428_{\pm 0.091}$	$0.334_{\pm 0.051}$	$0.310_{\pm 0.045}$	$0.784_{\pm 0.049}$	$\mathbf{0.147}_{\pm 0.028}$
	Max RCal	$0.384_{\pm 0.055}$	$0.355_{\pm 0.069}$	$0.272_{\pm 0.022}$	$0.543_{\pm 0.021}$	$0.559_{\pm 0.030}$	$\mathbf{0.156}_{\pm 0.041}$
MDA	Force MAE	$0.420_{\pm 0.015}$	$0.403_{\pm 0.017}$	$0.461_{\pm 0.010}$	$0.486_{\pm 0.031}$	$0.413_{\pm 0.022}$	$\mathbf{0.397}_{\pm 0.010}$
	MCal	$1.382_{\pm 0.048}$	$\mathbf{0.191}_{\pm 0.059}$	$1.211_{\pm 0.014}$	$0.511_{\pm 0.035}$	$1.494_{\pm 0.011}$	$0.407_{\pm 0.044}$
	Max RCal	$0.925_{\pm 0.060}$	$\mathbf{0.126}_{\pm 0.050}$	$0.770_{\pm 0.004}$	$0.401_{\pm 0.030}$	$1.098_{\pm 0.010}$	$0.310_{\pm 0.036}$
SAC	Force MAE	$0.280_{\pm 0.004}$	$0.305_{\pm 0.013}$	$0.312_{\pm 0.010}$	$0.333_{\pm 0.008}$	$\mathbf{0.259}_{\pm 0.005}$	$0.269_{\pm 0.005}$
	MCal	$1.034_{\pm 0.028}$	$0.385_{\pm 0.123}$	$0.932_{\pm 0.029}$	$0.450_{\pm 0.035}$	$1.235_{\pm 0.015}$	$\mathbf{0.221}_{\pm 0.063}$
	Max RCal	$0.637_{\pm 0.029}$	$0.259_{\pm 0.082}$	$0.646_{\pm 0.029}$	$0.635_{\pm 0.025}$	$0.844_{\pm 0.023}$	$\mathbf{0.120}_{\pm 0.040}$
Uracil	Force MAE	$0.195_{\pm 0.012}$	$0.233_{\pm 0.004}$	$0.219_{\pm 0.010}$	$0.215_{\pm 0.005}$	$\mathbf{0.169}_{\pm 0.003}$	$0.193_{\pm 0.004}$
	MCal	$0.939_{\pm 0.023}$	$0.636_{\pm 0.084}$	$\mathbf{0.484}_{\pm 0.035}$	$0.518_{\pm 0.041}$	$1.130_{\pm 0.019}$	$0.519_{\pm 0.054}$
	Max RCal	$0.562_{\pm 0.008}$	$0.385_{\pm 0.069}$	$0.383_{\pm 0.021}$	$0.681_{\pm 0.027}$	$0.764_{\pm 0.014}$	$\mathbf{0.283}_{\pm 0.046}$
Toluene	Force MAE	$0.139_{\pm 0.006}$	$0.170_{\pm 0.018}$	$0.147_{\pm 0.002}$	$0.152_{\pm 0.005}$	$\mathbf{0.123}_{\pm 0.001}$	$0.158_{\pm 0.006}$
	MCal	$0.900_{\pm 0.064}$	$0.440_{\pm 0.126}$	$\mathbf{0.187}_{\pm 0.037}$	$0.502_{\pm 0.005}$	$1.027_{\pm 0.019}$	$0.247_{\pm 0.103}$
	Max RCal	$0.557_{\pm 0.049}$	$0.309_{\pm 0.070}$	$\mathbf{0.144}_{\pm 0.014}$	$0.498_{\pm 0.009}$	$0.706_{\pm 0.015}$	$0.182_{\pm 0.045}$

Table 9: Additional baselines on MD17-RotOOD with the PaiNN backbone, mean$_{\pm \mathrm{SD}}$ over three independent runs. Gauss and Gauss-D are trained with the Gaussian NLL, without the NIW prior. Bold indicates the best performance.

QM7X Table 10 reports the standard QM7X dipole-moment prediction results without the RotOOD shift. Consistent with the main results, evidential models substantially improve predictive accuracy over the backbone models. In particular, MER and EMER reduce DP MAE across all three backbones, with the improvement being especially clear for MACE, where EMER decreases the error from 0.0079 to 0.0041. However, the raw MER and EMER variants remain poorly calibrated, yielding large MCal and Max NLL values despite their strong accuracy. Introducing dropout as a training-time regularizer greatly improves calibration. Among non-ensemble methods, EMER-D achieves the best MCal for two and Max NLL for all three backbones, reducing MCal to 0.2925, 0.2558, and 0.3053 for PaiNN, NequIP, and MACE, respectively. Although MC Dropout or ensembles can obtain competitive marginal RCal in some cases, their distance-based calibration and likelihood scores are generally weaker. Overall, these results show that EMER-D provides the most balanced performance on QM7X, maintaining competitive dipole-moment accuracy while producing substantially better calibrated multivariate UE.

OC20 For OC20, forces are computed as the negative energy gradient (conservative forces), with $\rho = 0.7$ for the evidential variants. The performance on OC20 is summarized in Table 11 across four validation splits: In-Domain (ID), Out-of-Domain Adsorbate (OOD Ads), OOD Catalyst (OOD Cat), and OOD Both. All evidential variants reach a lower force MAE than the backbone on all four splits, with EMER lowest on ID, OOD Ads, and OOD Cat, and EMER-D lowest on OOD Both. Regarding calibration, MER-D has the lowest MCal on all splits, EMER-D has the best Max RCal on ID and OOD Cat, and the best Max NLL on all splits. Yet, as demonstrated in Table 1, the apparent calibration success of non-equivariant methods stems from the similarity in marginal force distributions between the training and test sets, whereas EMER-D ensures rotational robustness,

Model	Metric	Backbone	Equivariant		Nonequivariant		Ours		
			Ens.	Drop.	CARD	MER	EMER	MER-D	EMER-D
PaiNN	DP MAE	0.0073	0.0054	0.0068	0.0073	0.0056	**0.0055**	0.0059	0.0059
	MCal	–	0.4394	1.0030	1.6881	1.5328	1.5328	0.3443	**0.2925**
	Max RCal	–	0.1160	**0.1723**	1.2034	1.1723	1.1319	0.2079	0.2103
	Max NLL	–	-3.5406	-2.7610	-2.6909	31.2264	40.4600	-3.2890	**-3.4716**
NequIP	DP MAE	0.0069	0.0060	0.0079	0.0069	0.0061	**0.0060**	0.0070	0.0063
	MCal	–	0.5122	0.6117	1.4434	1.1098	1.1817	0.5090	**0.2558**
	Max RCal	–	0.1866	**0.1155**	1.1938	0.8276	0.6161	0.4921	0.2098
	Max NLL	–	-3.1946	-2.9163	-2.7945	2.8172	-0.6902	-3.1633	**-3.6026**
MACE	DP MAE	0.0079	0.0074	0.0084	0.0079	0.0044	0.0041	**0.0040**	0.0042
	MCal	–	0.6728	1.1019	0.7507	1.4913	1.3586	**0.2902**	0.3053
	Max RCal	–	0.2816	0.5373	0.6256	1.0889	0.9749	0.1724	**0.1681**
	Max NLL	–	-2.9139	-1.3933	-1.1121	21.4033	10.3550	-3.6113	**-3.8851**

Table 10: Performance metrics for dipole moment prediction using QM7X dataset. Bold indicates the best single-model performance. Our proposed methods (EMER, MER-D, and EMER-D) achieve the best accuracy and calibration across the majority of tasks.

which is essential for physical consistency. The energy MAE of the evidential variants remains above that of the backbone and the ensemble.

Split	Metric	Backbone	Equivariant		Nonequivariant	Ours		
			Ens.	Drop.	MER	EMER	MER-D	EMER-D
Val ID	Energy MAE	**0.5894**	0.5410	0.6173	0.7416	0.7078	0.8215	0.7397
	Force MAE	0.0834	0.0728	0.0858	0.0727	**0.0699**	0.0818	0.0717
	MCal	–	1.2819	1.4897	0.3449	0.4551	**0.0727**	0.1123
	Max RCal	–	0.5925	0.9009	0.3193	0.3470	0.1901	**0.1754**
	Max NLL	–	1.4838	7.0786	0.8829	-0.4838	-0.2020	**-0.7101**
Val OOD Ads	Energy MAE	0.6579	0.6031	**0.6551**	0.7990	0.7914	0.8507	0.7858
	Force MAE	0.0921	0.0809	0.0910	0.0810	**0.0786**	0.0864	0.0787
	MCal	–	1.3032	1.5136	0.5715	0.7053	**0.2069**	0.3305
	Max RCal	–	0.6084	0.9526	0.4702	0.5132	**0.2817**	0.3095
	Max NLL	–	1.5825	7.9118	3.5241	0.2614	-0.0198	**-0.3724**
Val OOD Cat	Energy MAE	**0.6452**	0.5955	0.6758	0.7634	0.7390	0.8365	0.7616
	Force MAE	0.0874	0.0755	0.0890	0.0777	**0.0741**	0.0852	0.0752
	MCal	–	1.2749	1.4947	0.3908	0.4973	**0.0974**	0.1417
	Max RCal	–	0.5858	0.9149	0.3510	0.3774	0.2162	**0.1998**
	Max NLL	–	1.4579	7.1594	7.3281	-0.3976	-0.2194	**-0.6729**
Val OOD Both	Energy MAE	**0.8596**	0.8079	0.8650	0.9891	0.9812	1.0471	0.9813
	Force MAE	0.1068	0.0931	0.1044	0.0954	0.0921	0.1001	**0.0915**
	MCal	–	1.2909	1.5265	0.6436	0.7734	**0.2828**	0.3954
	Max RCal	–	0.6002	0.9804	0.5140	0.5534	**0.3295**	0.3493
	Max NLL	–	1.7218	8.8972	12.3247	0.8802	0.4709	**-0.0028**

Table 11: Performance metrics for different splits with the OC20 dataset, where forces are computed as the negative energy gradient. Mean over three independent runs. Calibration and NLL are evaluated on atomic forces only. Bold indicates the best single-model performance.

Standard Deviations Tables 12, 13, and 14 report the standard deviation (SD) across random seeds for the results in Tables 1, 2, and 11; the means are given in the corresponding tables.

Mol	Metric	Backbone	Equivariant		Nonequivariant		Ours		
			Ens.	Drop.	CARD	MER	EMER	MER-D	EMER-D
Aspirin	Energy MAE (10^{-3})	18.6	11.2	6.5	18.6	5.9	13.6	92.9	62.0
	Force MAE (10^{-3})	7.8	9.1	6.5	7.8	5.9	9.5	3.1	8.8
	MCal (10^{-3})	–	8.7	5.9	18.6	10.7	29.3	18.2	28.9
	Max RCal (10^{-3})	–	2.1	2.5	16.6	15.4	30.2	18.2	23.3
	Max NLL (10^{-3})	–	60.5	56.7	792	915	270	37.1	67.0
Ethanol	Energy MAE (10^{-3})	6.2	3.2	3.9	6.2	2.1	1.5	16.6	12.7
	Force MAE (10^{-3})	0.6	2.5	6.0	0.6	5.0	14.2	16.8	8.2
	MCal (10^{-3})	–	54.7	32.3	35.9	36.6	12.0	113	35.5
	Max RCal (10^{-3})	–	14.8	25.5	27.9	35.7	8.4	74.1	27.8
	Max NLL (10^{-3})	–	45.0	95.8	1131	7055	17684	211	41.2
Nap	Energy MAE (10^{-3})	10.0	3.1	9.9	10.0	8.1	8.5	1.7	48.6
	Force MAE (10^{-3})	1.1	2.7	3.0	1.1	3.1	3.2	17.2	6.9
	MCal (10^{-3})	–	37.7	10.6	12.6	62.6	59.7	80.1	33.9
	Max RCal (10^{-3})	–	27.0	5.2	17.5	41.0	37.1	61.0	50.6
	Max NLL (10^{-3})	–	66.7	18.3	218	794	131	408	16.1
MDA	Energy MAE (10^{-3})	11.7	3.4	9.9	11.7	6.4	10.8	1.3	18.7
	Force MAE (10^{-3})	37.7	9.8	33.0	38.1	20.5	26.5	25.2	12.8
	MCal (10^{-3})	–	37.6	57.3	56.8	94.4	13.5	76.8	54.2
	Max RCal (10^{-3})	–	5.1	42.1	54.5	92.7	12.0	47.0	44.6
	Max NLL (10^{-3})	–	50.9	495	3121	16876	13751	368	226
SAC	Energy MAE (10^{-3})	5.3	4.7	2.3	5.3	7.5	7.9	8.8	109
	Force MAE (10^{-3})	10.3	1.8	12.5	10.3	16.4	6.1	8.3	6.4
	MCal (10^{-3})	–	32.6	16.4	29.8	45.7	18.2	53.4	77.3
	Max RCal (10^{-3})	–	26.5	20.1	31.6	25.7	27.8	37.7	48.6
	Max NLL (10^{-3})	–	9.7	226	2259	10102	4989	726	66.3
Uracil	Energy MAE (10^{-3})	5.0	1.5	11.4	5.0	2.4	4.6	60.2	47.0
	Force MAE (10^{-3})	5.9	3.7	6.7	5.9	1.6	3.2	4.1	5.3
	MCal (10^{-3})	–	31.2	21.2	37.1	31.9	23.7	16.2	65.9
	Max RCal (10^{-3})	–	10.8	5.0	28.4	14.5	17.5	17.5	56.6
	Max NLL (10^{-3})	–	53.6	71.5	1455	31748	658	191	30.7
Toluene	Energy MAE (10^{-3})	20.1	7.1	2.6	20.1	1.8	9.2	47.1	87.5
	Force MAE (10^{-3})	6.4	2.2	1.3	6.4	3.4	1.7	18.0	6.9
	MCal (10^{-3})	–	3.5	18.0	28.6	79.8	23.1	217	127
	Max RCal (10^{-3})	–	3.7	4.2	7.1	55.3	18.8	68.7	55.5
	Max NLL (10^{-3})	–	28.2	25.9	181	1644	473	236	42.9

Table 12: Standard deviation (SD) of the results reported in Table 1. The unit of each metric is given in the Metric column.

Model	Metric	Backbone	Equivariant		Nonequivariant		Ours		
			Ens.	Drop.	CARD	MER	EMER	MER-D	EMER-D
PaiNN	DP MAE (10^{-4})	5.0	4.0	10.0	5.0	7.0	4.0	5.0	5.0
	MCal (10^{-3})	–	39.4	30.1	18.2	23.8	27.7	73.2	97.4
	Max RCal (10^{-3})	–	42.2	39.2	13.3	49.5	11.9	80.7	85.8
	Max NLL (10^{-3})	–	105	157	62.3	15369	8669	263	366
NequIP	DP MAE (10^{-4})	3.0	1.0	4.0	3.0	<1	1.0	9.0	4.0
	MCal (10^{-3})	–	29.6	50.8	33.1	69.7	46.3	101	58.5
	Max RCal (10^{-3})	–	19.5	22.8	38.0	5.2	37.3	48.4	32.7
	Max NLL (10^{-3})	–	203	183	22.0	7884	986	643	699
MACE	DP MAE (10^{-4})	5.0	3.0	3.0	5.0	1.0	3.0	4.0	4.0
	MCal (10^{-3})	–	52.0	32.3	20.0	20.9	30.5	45.5	67.2
	Max RCal (10^{-3})	–	61.0	41.1	27.7	56.4	28.1	15.7	59.0
	Max NLL (10^{-3})	–	177	1290	18.9	30301	2994	333	88.4

Table 13: Standard deviation (SD) of the results reported in Table 2. The unit of each metric is given in the Metric column.

Split	Metric	Backbone	Equivariant		Nonequivariant	Ours		
			Ens.	Drop.	MER	EMER	MER-D	EMER-D
Val ID	Energy MAE (10^{-3})	6.9	3.4	9.9	3.5	10.6	36.9	5.7
	Force MAE (10^{-4})	8.0	1.0	5.0	3.0	5.0	29.0	5.0
	MCal (10^{-3})	–	3.2	4.5	82.5	19.3	29.2	17.4
	Max RCal (10^{-3})	–	2.4	12.3	45.7	9.7	37.9	9.7
	Max NLL (10^{-3})	–	19.6	322	829	77.3	308	58.2
Val OOD Ads	Energy MAE (10^{-3})	11.1	8.5	14.0	1.2	6.3	37.4	12.1
	Force MAE (10^{-4})	7.0	1.0	7.0	1.0	3.0	25.0	4.0
	MCal (10^{-3})	–	4.5	7.3	80.2	12.5	32.7	28.1
	Max RCal (10^{-3})	–	4.2	17.1	47.4	6.8	31.6	19.7
	Max NLL (10^{-3})	–	42.3	463	2001	53.8	258	31.8
Val OOD Cat	Energy MAE (10^{-3})	30.5	17.3	24.0	13.6	18.9	27.8	16.9
	Force MAE (10^{-4})	12.0	3.0	9.0	3.0	7.0	27.0	3.0
	MCal (10^{-3})	–	1.9	4.7	74.2	19.5	36.8	11.6
	Max RCal (10^{-3})	–	1.0	11.3	39.5	11.0	38.6	7.2
	Max NLL (10^{-3})	–	21.5	482	5610	75.7	230	35.3
Val OOD Both	Energy MAE (10^{-3})	24.5	9.7	40.5	6.1	14.7	54.1	16.1
	Force MAE (10^{-4})	9.0	3.0	10.0	5.0	4.0	24.0	5.0
	MCal (10^{-3})	–	2.4	3.8	63.8	10.1	22.7	16.7
	Max RCal (10^{-3})	–	4.0	10.1	37.6	6.0	27.0	7.8
	Max NLL (10^{-3})	–	28.3	366	8710	68.5	332	86.0

Table 14: Standard deviation (SD) of the results reported in Table 11. The unit of each metric is given in the Metric column.

Dropout Ratio Ablation Study We investigated the effect of the dropout ratio, a key hyperparameter in our model, by evaluating values of 0.01, 0.05, 0.1, and 0.2 using the PaiNN backbone. The results, summarized in Table 15, show differing trends across the evaluation datasets. For the MD17 benchmark, a clear inverse relationship is observed, where increasing the dropout ratio leads to a degradation in performance across energy MAE, force MAE, and the force calibration score. Conversely, on the QM7X dataset, the trend is less uniform; while the lowest dropout ratio (0.01) yields the best dipole moment MAE, the dipole calibration improves as the dropout ratio increases. Prioritizing the performance on energy and force prediction, which are central to MD simulations, we selected a dropout ratio of 0.01 for our final model, as it provides the best overall accuracy on the MD17 tasks. For QM7X, we selected a dropout ratio of 0.1 for accuracy/calibration balance.

Dropout Ratio	0.01	0.05	0.1	0.2
MD17 Average				
Energy MAE	0.2069	0.4404	0.6029	0.5720
Force MAE	0.2709	0.4244	0.5767	0.8758
Force Calibration Score	0.2989	0.5025	0.8201	1.1753
QM7X Dipole Moment				
DP MAE	0.0056	0.0069	0.0063	0.0075
DP Calibration	0.5844	0.8068	0.4544	0.1757

Table 15: Ablation study on the dropout ratio using PaiNN backbone

Force Loss Weight For machine learning force fields that jointly predict energy and forces, the weight assigned to the force loss is a critical hyperparameter. We evaluated the energy and force MAE across varying weights, as detailed in Table 16. The results indicate that a coefficient in the range of 0.99–0.999 minimizes force MAE, while a range of 0.8–0.99 yields the lowest energy MAE. This suggests that our choice of 0.99 provides a reasonable balance between energy and force accuracy.

		Force Loss Weight					
Mol		0.01	0.2	0.5	0.8	0.99	0.999
Aspirin	Energy MAE	3.65	1.60	0.40	0.39	0.36	0.46
	Force MAE	19.48	4.74	0.69	0.51	0.42	0.40
Ethanol	Energy MAE	0.98	0.14	0.11	0.09	0.12	0.11
	Force MAE	7.60	0.65	0.43	0.33	0.29	0.28

Table 16: Ablation study of force weight for aspirin and ethanol on the MD17 dataset.

Active Learning. We evaluated the performance of our proposed method in an active learning setting for QM7X dipole moment prediction. Motivated by the high cost of molecular verification in wet-lab environments, we start from 500 labeled molecules and query 500 samples per round based on the trace of the uncertainty covariance matrix, and compare with random acquisition from the same starting model. We benchmarked against MC Dropout with an MSE loss, and an MSE model without dropout. As shown in Table 17, the EMER-based models reach a lower MAE than the MSE-based baselines at essentially every budget. For EMER-D, uncertainty-based acquisition yields a lower MAE than random acquisition at every budget from 1000 to 4500 labels, whereas for EMER the two acquisition rules perform similarly.

Model	**Acquisition**	**Labeled samples**								
		500	1000	1500	2000	2500	3000	3500	4000	4500
EMER-D	Uncertainty	$\mathbf{25.5}_{\pm 3.7}$	$\mathbf{16.1}_{\pm 0.4}$	$\mathbf{12.5}_{\pm 0.4}$	$\mathbf{12.2}_{\pm 0.7}$	$\mathbf{10.2}_{\pm 0.5}$	$\mathbf{9.3}_{\pm 0.4}$	$\mathbf{9.2}_{\pm 0.7}$	$\mathbf{7.9}_{\pm 0.2}$	$\mathbf{7.3}_{\pm 0.4}$
	Random	$\mathbf{25.5}_{\pm 3.7}$	$18.1_{\pm 1.2}$	$15.1_{\pm 2.3}$	$13.1_{\pm 1.4}$	$10.7_{\pm 0.3}$	$10.4_{\pm 0.6}$	$9.9_{\pm 1.3}$	$9.4_{\pm 1.3}$	$8.4_{\pm 1.2}$
EMER	Uncertainty	$\mathbf{22.5}_{\pm 1.6}$	$\mathbf{17.0}_{\pm 0.8}$	$\mathbf{14.7}_{\pm 1.9}$	$\mathbf{10.7}_{\pm 0.5}$	$11.1_{\pm 1.2}$	$\mathbf{8.9}_{\pm 0.7}$	$9.4_{\pm 0.7}$	$8.1_{\pm 0.6}$	$8.1_{\pm 0.0}$
	Random	$\mathbf{22.5}_{\pm 1.6}$	$17.2_{\pm 1.7}$	$14.8_{\pm 1.9}$	$11.4_{\pm 0.6}$	$\mathbf{10.5}_{\pm 0.2}$	$9.6_{\pm 0.4}$	$\mathbf{8.7}_{\pm 0.6}$	$\mathbf{7.7}_{\pm 0.8}$	$\mathbf{7.6}_{\pm 0.4}$
MC Dropout	Uncertainty	$\mathbf{35.0}_{\pm 0.8}$	$\mathbf{20.0}_{\pm 0.2}$	$36.1_{\pm 29.2}$	$19.8_{\pm 10.2}$	$\mathbf{11.1}_{\pm 0.5}$	$39.7_{\pm 41.9}$	$\mathbf{9.8}_{\pm 0.1}$	$\mathbf{9.1}_{\pm 0.4}$	$\mathbf{8.7}_{\pm 0.7}$
	Random	$35.0_{\pm 0.8}$	$21.7_{\pm 0.8}$	$\mathbf{28.0}_{\pm 10.8}$	$\mathbf{13.6}_{\pm 0.8}$	$12.4_{\pm 0.9}$	$\mathbf{10.8}_{\pm 0.6}$	$11.8_{\pm 1.7}$	$9.8_{\pm 0.9}$	$8.9_{\pm 0.6}$
MSE	Random	$29.8_{\pm 0.4}$	$24.2_{\pm 5.4}$	$18.5_{\pm 2.4}$	$17.0_{\pm 2.2}$	$15.1_{\pm 3.5}$	$11.8_{\pm 1.3}$	$14.3_{\pm 2.8}$	$11.3_{\pm 0.7}$	$10.1_{\pm 0.8}$

Table 17: Test MAE of dipole moment prediction (10^{-3}, mean$_{\pm \mathrm{SD}}$ over three independent runs) versus the number of labeled samples, with uncertainty-based and random acquisition. The two acquisition rules of each model share the same first round. Bold indicates the lower value among uncertainty-based and random acquisition for each method. The MC Dropout rows contain single diverged runs at some budgets, which inflate the mean and SD.

A.6 Additional Visualization

Figure 2 shows the overview of our proposed framework. Except for the general vector output head μ_0, two additional heads are needed to output the parameters of Multivariate Evidential Loss. One is a scalar head that outputs scalars ν and α, and another one is a vector head that outputs a $3 \times n$ matrix $\boldsymbol{F}$.

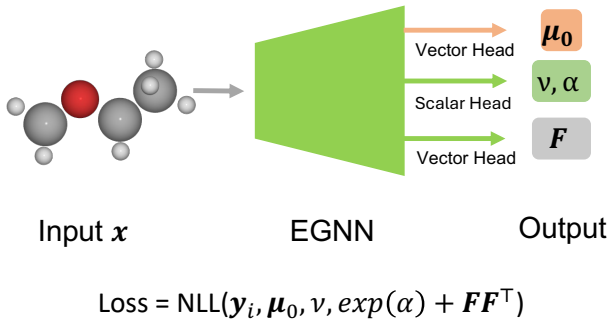


Figure 2: Overview of the proposed framework. Two additional uncertainty heads are incorporated to produce the evidential distribution parameters.

We performed a singular value decomposition of the learned matrix, as shown in Figure 3. It shows that the sum of the first two singular values accounts for $> 90\%$ of the variance, confirming that a rank-2 approximation is sufficient to capture physical symmetries without the computational cost of non-parametric methods.

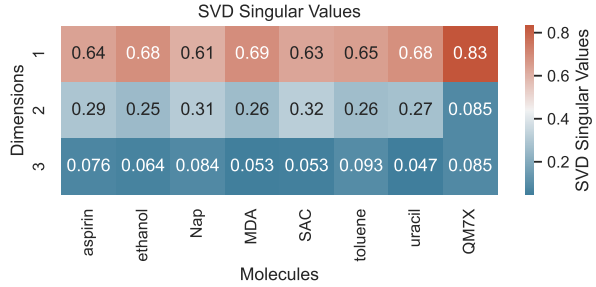


Figure 3: SVD singular values for molecules in MD17 and QM7X. For all molecules, the first two singular values are dominant, indicating that the learned matrix is well-approximated by a rank-2 component.

Figure 4 illustrates the distribution of force labels for Aspirin and Naphthalene (Nap). For Aspirin, the forces are uniformly distributed across the sphere. In contrast, for Nap, the forces are predominantly concentrated near the xy-plane.

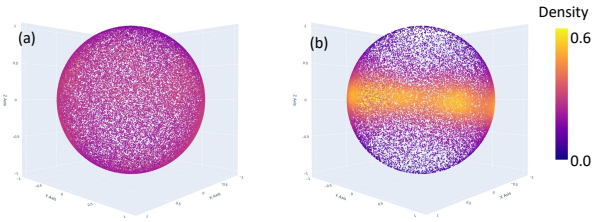


Figure 4: Distribution of force labels for (a) aspirin and (b) Nap in MD17. The forces are uniformly distributed across the $x - z$ directions for aspirin, but predominantly distributed within the xy plane for Nap.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state the main contributions: equivariant multivariate evidential regression, an equivariant covariance parameterization, and distance-based calibration for vector-valued uncertainty. The claims are supported by the theoretical discussion in Section 3 and the experiments on MD17, QM7-X, and OC20.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer:[Yes]

   Justification: The paper discusses practical limitations through the ablation study and implementation details, including instability without the isotropic $\exp(\alpha)\mathbf{I}$ term and instability of some higher-rank variants. The scope is also limited to molecular vector-valued prediction benchmarks.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: The main theoretical result states the assumptions on the invariant scalar, equivariant matrix, and rotation operation. A proof sketch is provided in the main text, with complete proofs and an identifiable Gram–Schmidt variant provided in Appendix A.1 and Appendix A.2.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper describes the model design, covariance parameterization, training objectives, datasets, evaluation metrics, baselines, and implementation details needed to reproduce the main results. Additional hyperparameters and training settings are provided in Appendix A.3.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We provide open access to an anonymized code repository at https://github.com/lixinyuu/EquivariantUE. The repository includes the updated implementation of EMER and EMER-D, baseline implementations, configuration files, and scripts for reproducing the main experimental results. It also provides environment setup instructions, data access and preprocessing guidance for the public benchmark datasets, and example commands for training and evaluation. The released scripts support reproduction of the reported accuracy and calibration metrics, including Energy MAE, Force MAE, MCal, Max RCal, and Max NLL.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper specifies the datasets, data sizes, train/test settings, loss functions, backbones, dropout locations, learning rates, batch sizes, number of epochs, and other hyperparameters needed to understand the reported results. These details are reported in the experimental section and Appendix A.3. In addition, we provide an GitHub repository with configuration files and training/evaluation scripts, which document the exact experimental settings used to reproduce the main results for EMER, EMER-D, and the baseline methods.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: All reported results are averaged over independent runs with different random seeds. The standard deviations (SD) across seeds are reported for the main results in Appendix A.5 (Tables 12, 13, 9, and 14). In these molecular property prediction benchmarks, lower averaged error, calibration score, and NLL values are generally interpreted as better performance under the same experimental setting.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The paper reports the main compute resources used in Appendix A.3, including the CPU, GPU, and memory of the primary workstation and the four-H100 workstation used for OC20 training. We do not provide exact wall-clock time for every individual run, but the reported hardware gives sufficient context for reproducing the experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

Justification: The research uses standard public molecular benchmark datasets and does not involve human subjects, private data, surveillance, or safety-critical deployment. We have reviewed the NeurIPS Code of Ethics and believe the work conforms to it.

Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: The paper discusses positive impacts through more reliable uncertainty estimation for scientific discovery, including molecular simulation and materials or drug discovery. Potential negative impacts are limited because the work is foundational and evaluated on public molecular benchmarks, but unreliable deployment could still misguide downstream scientific decisions if uncertainty estimates are misinterpreted.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The paper does not release high-risk models, scraped datasets, generative media models, or pretrained language models. The proposed method is an uncertainty estimation framework for molecular property prediction on standard scientific benchmarks.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.

- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The paper cites the original sources for the datasets and model architectures used, including MD17, OC20, QM7-X, PaiNN, NequIP, and MACE. We use these assets only for standard academic benchmarking and follow their intended research use.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not introduce a new dataset or pretrained model asset. The main contribution is a new method and evaluation framework built on existing public benchmarks.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: The paper does not involve crowdsourcing, user studies, participant data collection, or research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: The paper does not involve human subjects, crowdsourcing, or participant data, so IRB approval or equivalent review is not applicable.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: The core method development, experiments, and evaluation do not use LLMs as an important, original, or non-standard component. Any use of writing or editing tools does not affect the scientific methodology, results, or originality of the work.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.