On Uncertainty Calibration for Equivariant Functions

Edward Berman *berman.ed@northeastern.edu*
Department of Mathematics and Department of Physics
Northeastern University

Jacob Ginesin *ginesin.j@northeastern.edu*
Department of Mathematics and Khoury College of Computer Science
Northeastern University

Consultant: Prof. Robin Walters *r.walters@northeastern.edu*
Khoury College of Computer Science
Northeastern University

Abstract

Uncertainty estimation shows pronounced importance in the data-sparse settings where equivariant models tend to excel, including pick-and-place robotics tasks, galaxy morphology classification, and chemical physics. This work studies the relationship between equivariance, model calibration, and model confidence. We present bounds calibration error as well as on over and underconfidence for classifiers under the assumption of $G-$invariance. Next, we define a generalization of calibration error for regression tasks beyond scalar valued predictions for mean and variance before presenting its upper bound in the cases of $G-$invariance and $G-$equivariance. We show how this upper bound can be realized on real data, and we explore how sensitive our metric is to different binning approximations. In parallel, this paper provides a discussion of how model calibration relates to notions of aleatoric and epistemic uncertainties. In particular, we define the aleatoric bleed, a metric of assessing how well a model can disentangle aleatoric and epistemic uncertainties, and show its relationship to equivariance. We explore the consequences of these bounds experimentally, examining trends with correct/incorrect/extrinsic equivariance, group order, and aleatoric bleed. This work marks a first step toward a unifying framework for how equivariance relates to uncertainty estimation.

1 Introduction

A once puzzling result found that equivariant models can still be effective even in cases of apparent symmetry mismatch between the model and the data (Wang et al., 2023a). This in turn motivated the work of Wang et al. (2024), which explored how equivariance can affect model *accurancy*, both positively and negatively. However, it is not yet understood how equivariance impacts model *calibration*, loosely defined as the disagreement between a model's accuracy and predicted *confidence*. Understanding both model calibration and confidence is particularly useful in data-sparse settings where equivariant neural networks tend to thrive, such as pick-and-place robotics tasks (Kalashnikov et al., 2018; Wang et al., 2022b;a; Fu et al., 2023; Huang et al., 2023; 2024b;a), galaxy morphology classification (Pandya et al., 2023; 2025), and molecular physics (Zou et al., 2023; Ramakrishnan et al., 2014). However, equivariance is not a golden bullet: the benefits of equivariance are fairly limited at scale (Brehmer et al., 2024; Wang et al., 2023b; Abramson et al., 2024; Klee et al., 2023; Gruver et al., 2023), and equivariance has provable degradation on model performance in cases of symmetry mismatch (Wang et al., 2024). These works, to some extent, suggest that equivariant models can be victims of *The Bitter Lesson* (Sutton, 2019). That is, equivariant models, which are constrained to abide by what we might intuit as a reasonable inductive bias, are outperformed by more generalized approaches that better leverage computation (Finzi et al., 2021; Wang et al., 2022c; Romero & Lohit, 2022; van der

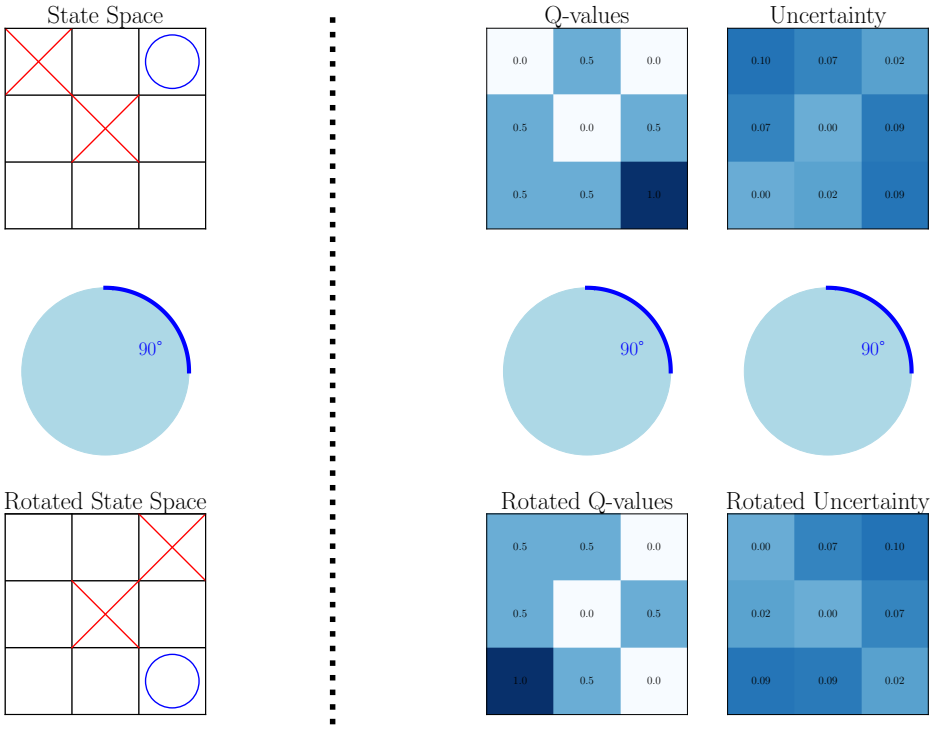

Figure 1: A motivating example for equivariant regression where the uncertainty estimate abides by the same symmetry as the Q-value prediction. When the input state space rotates by 90° so do the output Q-values and uncertainty estimates. In this case, the uncertainty is specifically the model's own uncertainty on the estimated Q-value, since the solution to the Bellman equation is unique. This figure is inspired by Q-learning in spatial action spaces, see Wang et al. (2022b). For ease of understanding, this particular Q-learning task is tic-tac-toe.

Ouderaa et al., 2022; Petrache & Trivedi, 2023; McNeela, 2023; Veefkind & Cesa, 2024; Park et al., 2024; Hofgard et al., 2024; Samudre et al., 2024; Mitchel et al., 2024), motivating additional work investigating the general viability of equivariant models. Particularly, several questions remain unanswered on the subject of calibration and confidence of equivariant models. Namely, when does equivariance help a model predict its own confidence? How do notions of correct/incorrect/extrinsic equivariance (the equivariance taxonomy hereafter) affect model calibration? What is the general relationship between equivariance and uncertainty estimation? For the purposes of claiming a scientific discovery or avoiding disaster when handling expensive and brittle equipment, model calibration is very often more important than model performance, and the effect of equivariance on such calibration error remains unclear.

The purpose of this work is to address both the lack of a unified theory relating equivariance to uncertainty estimation, and the absence of experiments exploring this relationship in practice. To accomplish this, we extend the error bounds given by Wang et al. (2024) to a broader class of calibration losses. In this way, we can quantify the effect of equivariance not just on accuracy, but also on calibration. To the best of our knowledge, the only works that really explore the relationship between equivariance and uncertainty are Sun et al. (2023); Cherif et al. (2024). Accordingly, we supplement our bounds with experiments on a wide variety of real and simulated datasets that indicate how these results manifest. Importantly, we show that our bounds can be realized in a true experimental setting. We summarize our contributions as follows:

1. We provide general bounds on how equivariance affects calibration error for classification tasks. We further examine the limiting cases of over and under confidence in equivariant models (§3.3).
2. We generalize the expected normalized calibration error (Equation 4) beyond scalar values for mean and variance predictions and derive its theoretical upper bound on equivariant models (§3.4-3.5).

We further study our proposed metric in the limiting case of minimized regression error using the bounds from Wang et al. (2024). Additionally, we coin another metric, the aleatoric bleed, in order to study miscalibration in terms of aleatoric and epistemic uncertainty (§3.6).

3. In the effort to better understand the impact of our bounds, we explore the effect of equivariance on model calibration, confidence, and aleatoric bleed using a wide variety of real and simulated datasets (§4). We examine trends in group order, performance on data within different regimes of the equivariance taxonomy, and the ability to disaggregate aleatoric and epistemic uncertainties (§4.1 - 4.4).
4. We compare the effect of binning approximations on our generalized expected normalized calibration error with the results of Pernot (2023) on QM9s data (Ramakrishnan et al., 2014) (§4.5).
5. Following the discussion of §4.5, we show that regardless of the number of bins used to approximate the true calibration error, the upper bound on the expected normalized calibration error can be realized on real datasets. Specifically, we do this on augmented versions of QM9s (§4.6).
6. We publicly release all code and datasets used in this work at 

2 Related Works

Equivariant Learning. Our work is closest in flavor to Petrache & Trivedi (2023) and Wang et al. (2024), which both establish bounds on function generalization under various assumptions of symmetry mismatch. We adopt many of the same formalisms as Wang et al. (2024), which was motivated by Wang et al. (2023a) and takes theoretical inspirations from Finzi et al. (2020). Our work makes use of the assumption that equivariant functions are universal approximators (i.e. they can approximate any arbitrary function), which was shown to be true for $G-$equivariant functions in Maron et al. (2019); Yarotsky (2022). Our work also uses the strategy of decomposing the input and output spaces in order to prove some bounds, which we accomplish through taking the quotient by a group. This is a strategy similarly employed in Sannai et al. (2021); Lawrence (2022); Petrache & Trivedi (2023); Wang et al. (2024), with the formulation of Petrache & Trivedi (2023) sharing the most commonalities with our work. Theory on equivariant learning is partly addressed for probabilistic symmetries in Bloem-Reddy et al. (2020), but no bounds similar to ours were presented. Reasoning about distributions in terms of invariants also has a rich history in arguing why one prior is better than another for Bayesian analysis – see Jaynes (1968).

Aleatoric and Epistemic Uncertainties. A longstanding goal in the computational sciences is to disaggregate model (epistemic) uncertainties from (aleatoric) uncertainties inherent to the data (Hüllermeier & Waegeman, 2021; Osband et al., 2023; Fuchsgruber et al., 2024; Ulmer et al., 2021). A key distinction is that epistemic uncertainties can be reconciled with more data, whereas aleatoric uncertainties can not.

Learning Parameterized Distributions. For many learning tasks, it is natural to design a neural network that approximates a probability distribution rather than a single point estimation. There are many ways of doing this, such as with Bayesian Neural Networks (Kononenko, 1989), Epistemic Neural Networks (Osband et al., 2023), Normalizing Flows (Kobyzev et al., 2020; Papamakarios et al., 2021), or simply employing the softmax (Goodfellow et al., 2016). These approaches are often computationally expensive to compute in practice. Thus, we often employ parameterization techniques to constrain neural networks to output simplified probability distributions and train them using a negative log likelihood loss derived from said distribution. The parameters that define the output distributions are easy to interpret and are often easier to learn than aforementioned alternative approaches. Mean variance estimators (MVE) are the simplest type of neural network that predicts a parameterized distribution. Instead of predicting a single output, they predict a mean μ and a variance σ^2 (Nix & Weigend, 1994; Seitzer et al., 2022). There is also work extending MVEs to learn covariances for multivariate distributions (Tomczak et al., 2020) and linear combinations of Gaussians (Diakonikolas et al., 2020). Amini et al. (2020) extend MVEs by imposing a prior on μ and σ^2, which in turn provides enough parameters to disentangle aleatoric and epistemic uncertainties. Hüttel et al. (2023) further extended Amini et al. (2020) so that one can also learn different quantiles of the distribution as a way of learning aleatoric uncertainty.

Calibration Error. It is common in classification tasks to choose the target label with the highest probability. However, even if a model gives a label y the highest probability p, that does not mean the model will necessarily be correct with probability p. This mismatch is often quantified with the expected calibration error (ECE) (Guo et al., 2017). Miscalibration can analogously be measured for regression tasks (Pernot, 2023; Levi et al., 2022b). The idea is to compare true labels y with a predicted mean μ and variance σ^2 of a MVE. One should expect the squared errors $(y-\mu)^2$ to average out to the variance σ^2. This idea was made precise by Levi et al. (2022a), who proposed the expected normalized calibration error (ENCE) to quantify that exact discrepancy. A key limitation of their work is that it is formulated in terms of binning approximations (i.e. discretizing our output space, by approximating our continuous function into probabalistic "bins") rather than in terms of a continuous probability density. Another key limitation of their work is that they assume mean and variance are scalar values. The work of Sun et al. (2023) suggests that equivariance can improve model calibration, but a theoretical justification for this is not present in the literature. Beyond ECE and ENCE, we also explore calibration in terms of coverage, for which we refer the reader to Gneiting & Raftery (2007); Sun et al. (2023).

3 Theoretical Results

3.1 General Notation

For clarity, we summarize here the principal notation and conventions used in our work.

Spaces and Functions. We denote by X the input space and by Y the output space. In classification tasks, Y is a finite set of labels and X is an arbitrary domain, whereas for regression tasks we assume $Y \subset \mathbb{R}^n$. The ground truth function is given by

$$f\colon X \to Y.$$

Our model class consists of functions that are arbitrarily expressive (that is, our model class contains all possible f) but subject to an equivariance constraint with respect to a group G acting on X (and, possibly, on Y).

Classification. In the classification setting, each model h produces both a label prediction and an associated confidence:

$$h(x) = \big(h_Y(x),\, h_P(x)\big), \quad h_Y(x) \in Y, \quad h_P(x) \in \hat{P} = [0,1].$$

We denote by $q\colon X \to \mathbb{R}_+$ the probability density over X, and by $r\colon [0,1] \to \mathbb{R}_+$ the density of confidence values, so that for any interval $[p_1, p_2] \subseteq [0,1]$,

$$\mathbb{P}\big(p_1 \le h_P(x) \le p_2\big) = \int_{p_1}^{p_2} r(p)\, dp.$$

The expected calibration error (ECE) (where p denotes our predicted confidence) is defined as

$$\mathrm{ECE}(h) = \mathbb{E}_{h_P}\Big[\Big|\mathbb{P}\big(f_Y(x) = h_Y(x) \mid h_P(x) = p\big) - p\Big|\Big].$$

Regression. For regression tasks, each model h outputs both a mean prediction and a variance estimate:

$$h(x) = \big(h_\mu(x),\, h_{\sigma^2}(x)\big), \quad h_\mu(x) \in \mathbb{R}^n, \quad h_{\sigma^2}(x) \in \mathbb{R}^n_+.$$

When the model predicts a specific variance σ^2, we define the corresponding domain as

$$X' = \{x \in X \mid h_{\sigma^2}(x) = \sigma^2\},$$

and the domain-restricted regression error as

$$\mathrm{err}_{\mathrm{reg}}(h, \sigma^2) = \int_{X'} q(x)\, \big\|h_\mu(x) - f_Y(x)\big\|_2^2\, dx.$$

We generalize the expected normalized calibration error (ENCE) for regression by

$$\text{ENCE} = \int_{\sigma^2} r(\sigma^2) \frac{\left| \mathbb{E}_{x,y}\left[\left\| \sigma - |h_\mu(x) - f_Y(x)| \right\|_2^2 \mid h_{\sigma^2}(x) = \sigma^2 \right] \right|}{\|\sigma\|_2^2} \, d\sigma^2.$$

Equivariance. A central assumption is that the model class is (approximately) equivariant with respect to a group G. In particular, we assume that for all $g \in G$ and $x \in X$, either

$$h(gx) = h(x) \quad \text{(invariance)}$$

or more generally

$$h(gx) = \rho_Y(g)\, h(x),$$

where $\rho_Y(g)$ denotes the representation of G acting on Y. In the regression setting, for an orbit $Gx \subset X$ we define its probability mass as

$$q(Gx) = \int_{z \in Gx} q(z)\, dz,$$

and the normalized density on the orbit by

$$q'(z) = \frac{q(z)}{q(Gx)}.$$

The mean and variance of a function f on the orbit are then given by

$$\mathbb{E}_{Gx}[f] = \int_{Gx} q'(z)\, f(z)\, dz, \qquad \mathbb{V}_{Gx}[f] = \int_{Gx} q'(z)\, \|f(z) - \mathbb{E}_{Gx}[f]\|_2^2 \, dz.$$

When h is equivariant (but not necessarily invariant), we further introduce matrices Q_{Gx} and $Q^\dagger(gx)$ satisfying

$$\int_G Q^\dagger(gx)\, dg = \text{Id},$$

and define the group-averaged function

$$\mathbf{E}_G[f, x] = \int_G Q^\dagger(gx)\, g^{-1} f(gx)\, dg.$$

Miscellaneous. Throughout, $\|\cdot\|_2$ denotes the Euclidean norm and $\mathbb{I}(\cdot)$ is the indicator function (equal to 1 when its argument holds and 0 otherwise).

This notation is used consistently in our derivations of calibration bounds and in the discussion of equivariant uncertainty estimation.

For more thorough definitions of Equivariance, Evidential Regression, Aleatoric and Epistemic Uncertainty, and Iterated integration, we refer the reader to Appendix Sections **??**, B, C.

3.2 Problem Statements

In this section, we lay out the relevant problem statements as well as preliminary notation necessary to prove our main results. We review the definition of equivariance in Appendix A.

Throughout the problem statements, we introduce additional notation to make certain things explicit. However, we will drop these in the proofs as they become clear. Throughout the problem statements, we will denote vectors with a vector hat, such as $\vec{e}$. Additionally, throughout the problem setups when we introduce a set, vector space, or group, we will give it a boldface, such as $\mathbf{E}$.

Where we use $g\vec{x}$ is it implicit that we are considering the representation $\rho(g)$, an $n \times n$ matrix, for $g \in \mathbf{G}$. This means that $g\vec{x}$, the result of the group action acting on $\vec{x}$, is a vector.

Classification Problem Setup. Consider a function $f : \mathbf{X} \to \mathbf{Y}$ where $\mathbf{Y}$ is a finite set of labels and $\mathbf{X}$ is an arbitrary domain.. Let $q : \mathbf{X} \to \mathbb{R}$ be the probability density of the domain $\mathbf{X}$. Now, we define a model class as the set $\{h : \mathbf{X} \to \mathbf{Y} \times \hat{\mathbf{P}}\}$ where $\hat{\mathbf{P}} = [0, 1]$ with elements p representing the normalized confidence. We will denote the two outputs by h_Y and h_P. The goal for our model class is to fit the function f *and* to properly predict its own confidence by minimizing the expected calibration error (Equation 1, and Equation 1 in Guo et al. (2017)). Following Wang et al. (2024), we assume that the model class $\{h\}$ is arbitrarily expressive and contains all possible f, except that it is constrained to be equivariant with respect to a group G. For this classification setting, we specifically choose h to be $G-$invariant. Let $\mathbb{I}$ be an indicator function that equals 1 if the condition is satisfied and 0 otherwise. Let $r(p)$ be the probability density such that $\mathbb{P}(p_1 \leq h_P(\vec{x}) \leq p_2) = \int_{p_1}^{p_2} r(p)dp$.

$$\text{ECE}(h) = \mathbb{E}_{h_P}\left[\left|\, \mathbb{P}\left(f = h_Y | h_P = p\right) - p \,\right|\right]. \tag{1}$$

Additionally, we will use the concepts of iterated integration and fundamental domains, which we review in Appendix C.2. We will also use the equivariance taxonomy, which we review in Appendix C.1.

Invariant Regression Problem Setup. Consider a function $f : \mathbf{X} \to \mathbf{Y}$ where $\mathbf{Y} = \mathbb{R}^n$ and $\mathbf{X}$ is an arbitrary domain. Now, we define a model class as the set $\{h : \mathbf{X} \to \boldsymbol{\mu} \times \boldsymbol{\sigma^2}\}$ where $\boldsymbol{\mu} = \mathbb{R}^n$ represents the space of all mean-vectors and $\boldsymbol{\sigma^2} = \mathbb{R}^n_+$ represents the space of all variance-vectors. We will denote the two outputs by h_μ and h_{σ^2}. Let $p : \mathbf{X} \to \mathbb{R}$ be the probability density over the domain $\mathbf{X}$. Denote the subdomain of $\mathbf{X}$ given by the constraint $h_{\sigma^2}(\vec{x}) = \vec{\sigma^2}$ as $\mathbf{X}' = \mathbf{X}|_{h_{\sigma^2}(\vec{x}) = \vec{\sigma^2}}$. We will denote the fundamental domain of $\mathbf{G}$ in $\mathbf{X}'$ as $\mathbf{F}'$. We point out that this construction is reminiscent of orbit averaging as defined in §3.1 of Petrache & Trivedi (2023). We assume that each fundamental domain $\mathbf{F}'$ satisfies the condition that the union of all pairwise intersections $\cup_{g_1 \neq g_2}(\mathbf{g_1}\mathbf{F}' \cap \mathbf{g_2}\mathbf{F}')$ have measure 0. We also assume that $\mathbf{F}'$ and $\mathbf{G}\mathbf{x}'$ are differentiable manifolds for any $\mathbf{X}'$ and for all $\vec{x'} \in \mathbf{X}'$.

Next, we define a family of probability densities for all of the possible domain restrictions induced by $\boldsymbol{\sigma^2}$. Specifically, we define a density over $\mathbf{X}'$ by $q : \mathbf{X} \times \boldsymbol{\sigma^2} \to \mathbb{R}$ via $q(\vec{x}) = p(\vec{x} | h_{\sigma^2}(\vec{x}) = \vec{\sigma^2})$. For $\vec{x} \notin \mathbf{X}'$, $q(\vec{x}) = 0$. This allows us to define the domain restricted regression error

$$\text{err}_{\text{reg}}(h, \vec{\sigma^2}) = \int_{\mathbf{X}'} q(\vec{x'}) \left\| h_\mu(\vec{x'}) - f(\vec{x'}) \right\|_2^2 d\vec{x'} \tag{2}$$

where it is implicit that $\mathbf{X}'$ is defined with respect to a specific $\vec{\sigma^2}$. Denote by $q(\mathbf{G}\mathbf{x}') = \int_{\vec{z} \in \mathbf{G}\mathbf{x}'} q(\vec{z})d\vec{z}$ the probability of the orbit $\mathbf{G}\mathbf{x}'$ on $\mathbf{X}'$. Denote by $q_{\text{norm}}(\vec{z}) = \frac{q(\vec{z})}{q(\mathbf{G}\mathbf{x}')}$ the normalized probability density on the orbit $\mathbf{G}\mathbf{x}'$ such that $\int_{\mathbf{G}\mathbf{x}'} q_{\text{norm}}(\vec{z})d\vec{z} = 1$. Let $\mathbb{E}_{\mathbf{G}\mathbf{x}'}[f]$ be the mean of the function f on the orbit $\mathbf{G}\mathbf{x}'$ defined, and let $\mathbb{V}_{\mathbf{G}\mathbf{x}'}[f]$ be the variance of f on the orbit $\mathbf{G}\mathbf{x}'$,

$$\mathbb{E}_{\mathbf{G}\mathbf{x}'}[f] = \int_{\mathbf{G}\mathbf{x}'} q_{\text{norm}}(\vec{z}) f(\vec{z}) d\vec{z} = \frac{\int_{\mathbf{G}\mathbf{x}'} q(\vec{z}) f(\vec{z}) d\vec{z}}{\int_{\mathbf{G}\mathbf{x}'} q(\vec{z}) d\vec{z}}, \qquad \mathbb{V}_{\mathbf{G}\mathbf{x}'}[f] = \int_{\mathbf{G}\mathbf{x}'} q_{\text{norm}}(\vec{x}) \left\| \mathbb{E}_{\mathbf{G}\mathbf{x}'}[f] - f(\vec{z}) \right\|_2^2. \tag{3}$$

These definitions are discussed in Appendix C.3 and arise from bounds in Wang et al. (2024).

Finally, we define a generalization of the expected normalized calibration error (ENCE), as given by Equation 8 in Levi et al. (2022a). The main drawbacks of their ENCE metric are two-fold: their metric is defined in terms of binning approximations and assumes that $h_\mu(\vec{x})$ and $h_{\sigma^2}(\vec{x})$ output scalar values. Our formalism not only works for output vectors with dimension greater than 1 but also avoids binning approximations, allowing for a discussion of continuous group symmetries. While binning approximations are still necessary to compute ENCE in practice, our theory supports the more generalized continuous case. We take the absolute value function applied to a vector $|\cdot|$ to be applied element-wise. Similarly, if we use $\vec{\sigma}$ as a vector, that means the square root function was applied element-wise to the vector $\vec{\sigma^2}$. We will also note that vectors

can be described in terms of a partial ordering, where $\vec{a} \leq \vec{b}$ if $a_i \leq b_i$ for all $a_i \in \vec{a}, b_i \in \vec{b}$. Accordingly, let $\mathbf{D}$ be the region containing all vectors $\vec{d}$ in between two variance vectors such that $\vec{\sigma_1^2} \leq \vec{d} \leq \vec{\sigma_2^2}$. We define a probability density $r : \mathbf{X} \times \boldsymbol{\sigma^2} \to \mathbb{R}$ such that $\mathbb{P}(h_{\sigma^2}(\vec{x}) \in \mathbf{D}) = \int_{\mathbf{D}} r(\vec{\sigma^2}) d\vec{\sigma^2}$. The goals for our model class are to fit the function f *and* and to properly predict the its own confidence by minimizing our generalized ENCE metric (Equation 4):

$$ENCE = \int_{\vec{\sigma^2}} r(\vec{\sigma^2}) \frac{\mathbb{E}_{x,y}\left[\left\|\sqrt{\frac{2}{\pi}}\vec{\sigma} - |h_\mu(\vec{x}) - f(\vec{x})|\right\|_2^2 \;\middle|\; h_{\sigma^2}(\vec{x}) = \vec{\sigma^2}\right]}{\left\|\sqrt{\frac{2}{\pi}}\vec{\sigma}\right\|_2^2} d\vec{\sigma^2}. \tag{4}$$

We assume that the model class $\{h\}$ is arbitrarily expressive except that it is constrained to be invariant with respect to a group $\mathbf{G}$.

Remark 1. *If errors are normally distributed and a model is well calibrated (i.e. a model reports high confidence in correct results), then* $\sqrt{\frac{2}{\pi}}\sigma = \overline{|\mu - f(x)|}$ *is the same as* $\sigma^2 = \overline{(\mu - f(x))^2}$*, where the factor of* $\sqrt{\frac{2}{\pi}}$ *comes from Geary (1935). We choose the former as a componenet in our metric because it allows us to later apply a theorem from Wang et al. (2024). Even though it appears slightly different than the binned approximate form in Levi et al. (2022a) it is functionally the same as it penalizes miscalibration in the same way.*

We also discuss a variation of this problem where mean and variance predictions obey different $G-$equivariances in Appendix D.

Equivariant Regression Problem Setup. Our setup is the same as for invariant regression, with the addition of the following: We define a matrix $Q_{\mathbf{Gx}} \in \mathbb{R}^{n \times n}$, $Q^\dagger(\vec{gx}) \in \mathbb{R}^{n \times n}$ such that $\int_{\mathbf{G}} Q^\dagger(\vec{gx}) dg = \text{Id}$ via

$$Q_{\mathbf{Gx}} = \int_{\mathbf{G}} q(\vec{gx}) \rho_Y(g)^T \rho_Y(g) \alpha(x, g) dg \tag{5}$$

$$Q^\dagger(\vec{gx}) = Q_{\mathbf{Gx}}^{-1} q(\vec{gx}) \rho_Y(g)^T \rho_Y(g) \alpha(x, g). \tag{6}$$

We also define $\mathcal{E}_{\mathbf{G}}[f, \vec{x}]$ by

$$\mathcal{E}_{\mathbf{G}}[f, \vec{x}] = \int_G Q^\dagger(\vec{gx}) g^{-1} f(\vec{gx}) dg. \tag{7}$$

This definition is also discussed in Appendix C.3 and comes from a bound in Wang et al. (2024).

3.3 Invariant Classification Upper Bounds

Theorem 1. *Equation 1 is a normalized metric, it is lower bounded by* 0 *and is upper bounded by* 1*. When a model is consistently overconfident* ($\mathbb{P}(f = h_Y | h_P = p) < p$ *for all* p)*, then the upper bound on ECE is given by* $\mathbb{E}_{x \sim q}[h_P]$*. Similarly, when a model is consistently underconfident* ($\mathbb{P}(f = h_Y | h_P = p) > p$ *for all* p)*, then the upper bound on ECE is given by* $1 - \mathbb{E}_{x \sim q}[h_P]$*.*

Proof. We begin by rewriting Equation 1 as

$$\mathbb{E}_{h_P}\left[\left|\mathbb{P}\left(f = h_Y \;\middle|\; h_P = p\right) - p\right|\right] = \int_{p=0}^{p=1} r(p) \left|\mathbb{P}\left(f(x) = h_Y(x) \;\middle|\; h_P(x) = p\right) - p\right| dp. \tag{8}$$

For constants A and B, the triangle inequality gives us $0 \leq |A - B| \leq |A| + |B|$. If both A and B are nonnegative then $|A - B| \leq \max\{A, B\}$ with equality if and only if one of A or B is zero. Since $\mathbb{P}\left(f = h_Y \;\middle|\; h_P = p\right)$ and p are non-negative, we can further restrict the inequality to be

$$0 \leq \int_{p=0}^{p=1} r(p) \left| \mathbb{P}\left(f(x) = h_Y(x) \;\middle|\; h_P(x) = p\right) - p \right| dp \tag{9}$$

$$\leq \int_{p=0}^{p=1} r(p) \max\{\mathbb{P}\left(f(x) = h_Y(x) | h_P(x) = p\right), p\} dp. \tag{10}$$

Recall that we said $\{h\}$ is arbitarily expressive except for that it is invariant with respect to a group G. Thus, we can always choose h_Y and h_P such that

$$\left| \mathbb{P}\left(f(x) = h_Y(x) \;\middle|\; h_P(x) = p\right) - p \right| = 0. \tag{11}$$

This gives us our lower bound. We will now proceed to show the upper bound. We will denote by c the original integral,

$$c := \int_{p=0}^{p=1} r(p) \left| \mathbb{P}\left(f(x) = h_Y(x) \;\middle|\; h_P(x) = p\right) - p \right| dp. \tag{12}$$

We may abbreivate $\max\{\mathbb{P}\left(f(x) = h_Y(x) | h_P(x) = p\right), p\}$ as $\max(p)$. Now consider the following three cases:

1. **The model h is always overconfident:** In this case, $\max(p) = p$ for all p.

$$c \leq \int_{p=0}^{p=1} r(p) \max(p) dp \tag{13}$$

$$= \int_{p=0}^{p=1} r(p) p dp \tag{14}$$

$$= \mathbb{E}[h_P] \tag{15}$$

   Notice that we get this result exactly if we take $\mathbb{P}\left(f(x) = h_Y(x) \;\middle|\; h_P(x) = p\right) = 0$ in Equation 1.

2. **The model h is always underconfident:** In this case, $\max(p) = \mathbb{P}\left(f(x) = h_Y(x) | h_P(x) = p\right)$ for all p. It was revealing that the upper bound in the prior case came about from $\mathbb{P}\left(f(x) = h_Y(x)\right) = 0$. Similarly, we will find that the upper bound in this case comes about from $\mathbb{P}\left(f(x) = h_Y(x)\right) = 1$.

   Since we assumed that there are finitely many labels in the codomain Y, we can assume that $\mathbb{P}\left(h_Y(x) | h_P(x) = p\right)$ is a discrete probability distribution for each value of p. Therefore, each outcome in the distribution has a probability less than one. Accordingly, the outcome $\mathbb{P}\left(f(x) = h_Y(x) | h_P(x) = p\right)$ also has a probability less than one. So, $|\mathbb{P}\left(f(x) = h_Y(x) | h_P(x) = p\right) - p|$ is maximized when $\mathbb{P}\left(f(x) = h_Y(x) | h_P(x) = p\right)$ attains its upper bound of one for each p on $[0, 1]$.

   We compute

$$c = \int_{p=0}^{p=1} r(p) \left| 1 - p \right| dp \tag{16}$$

$$= \int_{p=0}^{p=1} r(p)(1 - p) dp \tag{17}$$

$$= 1 - \mathbb{E}[h_P]. \tag{18}$$

   Notice the parallel with Equation 15. This symmetry can be attributed to the nature of the absolute value and the fact that the limiting values occur when $\text{err}_{\text{cls}}(h)$ is zero or one, see Figure 2.

   In Appendix L, we show how the bound can be related to traditional classification error.

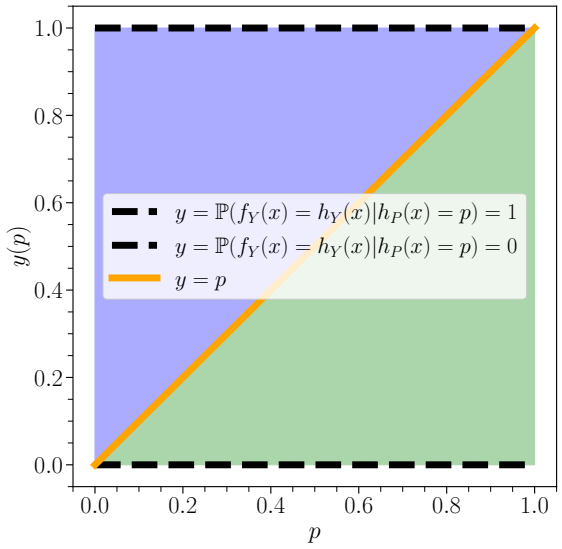

Figure 2: The relationship between $\mathbb{P}\left(f(x) = h_Y(x)|h_P(x) = p\right)$ and p under different limiting values of accuracy.

3. **The model h is sometimes overconfident and sometimes underconfident:** In this case, $\max(p)$ will vary with p. Let p_1 be the domain $[0, 1]$ restricted to regions where the model h is overconfident, and similarly let p_2 be the domain $[0, 1]$ restricted to regions where the model h is underconfident.

$$c \leq \int_{p=0}^{p=1} r(p) \max(p) dp \tag{19}$$

$$= \int_{\substack{p \in p_1 \\ \leq \mathbb{E}[h_P]}} r(p) p dp + \int_{p \in p_2} r(p) \mathbb{P}\left(f(x) = h_Y(x) \,\middle|\, h_P(x) = p \right) dp \tag{20}$$

$$\leq \mathbb{E}[h_P] + (1 - \mathbb{E}[h_P]) \tag{21}$$

$$= 1 \tag{22}$$

We conclude that c, and therefore $\mathrm{ECE}(h)$, is upper bounded by 1. Under the assumptions of over and under confidence, the upper bounds become $\mathbb{E}[h_P]$ and $1 - \mathbb{E}[h_P]$ respectively.

□

Remark 2. *Note that case* 2 *would not necessarily work if we assumed infinitely many labels in the codomain* Y*.*

Corollary 1. *We consider the case where* h *is incorrectly invariant under a finite group* G *on our input domain* X*. Let* $\mathcal{F}_p$ *be a fiber given by* $h_P^{-1}(\{p\})$ *with elements* x_p*, let* $b_n(p)$ *be the probability* $\mathbb{P}(x_p \in \{f(x_p) = y_n\})$*, where* y_n *is a label in* Y*. Let* k *be given by* $\min \sum_{n=1}^{N} b_n(p) \frac{n-1}{|G|}$ *on* p_2 *where* p_2 *is as defined in Theorem 1. With the incorrect invariance constraint on* h*, ECE becomes upper bounded by* $\mathbb{E}[h_p] + 1 - k$*.*

Proof. We start by the revisiting the bound on c from Theorem 1:

$$c \leq \int_{p=0}^{p=1} r(p) \max(p) dp \tag{23}$$

$$= \int_{\substack{p_1 \\ \leq \mathbb{E}[h_P]}} r(p) p dp + \int_{p_2} r(p) \mathbb{P}\left(f(x) = h_Y(x) \;\middle|\; h_P(x) = p \right) dp. \tag{24}$$

From proposition $A1$ in Wang et al. (2023a), the accuracy on any fiber induced from $p \in p_2$ is upper bounded by $1 - \sum_{n=1}^{N} b_n(p) \frac{n-1}{|G|}$. See that

$$\int_{p \in p_2} r(p) \mathbb{P}\left(f(x) = h_Y(x) \;\middle|\; h_P(x) = p \right) dp \leq \int_{p \in p_2} r(p) \left(1 - \sum_{n=1}^{N} b_n(p) \frac{n-1}{|G|} \right) dp \tag{25}$$

$$\leq \int_{p \in p_2} r(p)(1-k) dp \tag{26}$$

$$= (1-k) \int_{p \in p_2} r(p) dp \tag{27}$$

$$\leq (1-k)\,. \tag{28}$$

Therefore, $c \leq E[h_P] + 1 - k$. This completes the proof.

□

Remark 3. *If $k > \mathbb{E}[h_P]$, then this bound becomes tighter than the case of an unconstrained model as in Theorem 1.*

Corollary 2. *Let us restrict Equation 1 to be invariant under G, with no restrictions on the type of equivariance. Recall that the region of overconfidence has a calibration error bounded from above by $\mathbb{E}[h_P]$ and the region of underconfidence has a calibration error bounded from above by $1 - \mathbb{E}[h_P]$. Denote by $k(Gx)$ the expression $k(Gx) = \max_{p' \in \hat{P}} \int_{z \in Gx} q(z) p' dz$. Then $\mathbb{E}[h_P]$ attains an upper bound of $\int_{x \in F} k(Gx) dz$. Accordingly, the lower bound on $1 - \mathbb{E}[h_P]$ is given by $1 - \int_{x \in F} k(Gx) dz$. The calibration error on the regions of over and under confidence satisfy these respective bounds.*

Proof. Let p' be the probability output of h_P on a given orbit Gx. This proof will now use iterated iteration over a group, and we remind the reader to see Appendix C.2 for a review of the iterated integration formalism. We can express the expected value of h_P as

$$\mathbb{E}[h_p] = \int_X q(x) h_P(x) dx. \tag{29}$$

Then, we can use iterated integration to show

$$\int_X q(x) h_P(x) dx = \int_{x \in F} \int_{z \in Gx} q(z) h_P(z) dz dx. \tag{30}$$

We can bound the inner integral by looking at the label $h_Y(x) = y_n$ that corresponds to the highest probability $h_P(z) = p'$ on each orbit. In other words, we take the max p' and corresponding label argmax y_n. This gives us

$$\int_{z \in Gx} q(z) h_P(z) dz \leq \max_{p' \in \hat{P}} \int_{z \in Gx} q(z) p' dz. \tag{31}$$

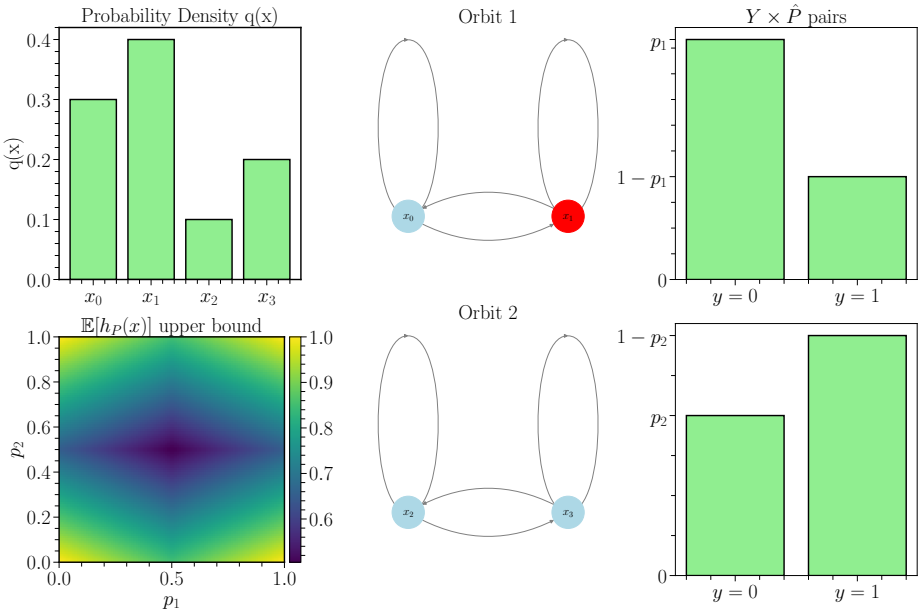

Figure 3: ***Top Left:*** The distribution over the input domain x. ***Bottom Left:*** The upper bound on $\mathbb{E}[h_P(x)]$ as a function of different output distributions. ***Middle Column:*** The permutation group action on X. ***Right Column:*** Output distributions over labels $y = 0$ and $y = 1$.

Assuming $|Y| = n$ labels and softmax normalization (conversion of real-valued vectors into probability distributions) given by

$$\sigma(\mathbf{z})_i = \frac{e^{z_i}}{\Sigma_{j=1}^{n} e^{z_j}}, \tag{32}$$

p' is given by $\max \sigma(\mathbf{z})$. We prove the corollary by integrating over F.

□

Example 1. *Consider Figure 3. In the top left, we have a density over the input domain $q(x)$. In the middle, we have four points $\{x_0, x_1, x_2, x_3\} = X$. We have the cyclic group C_2, which acts on G. We see that the cyclic group divides X onto two distinct orbits, $\{x_0, x_1\}$ and $\{x_2, x_3\}$, thereby giving us a fundamental domain of $\{x_0, x_2\}$. We assume our classifier h is invariant under the group action. Thus, we get the same predictions across each orbit, which we assumed to be softmax normalized. On the $\{x_0, x_1\}$ orbit, the model will always predict $\max\{p_1, 1-p_1\}$ and corresponding label $y = 0$ or $y = 1$. A similar statement holds for the $\{x_2, x_3\}$ orbit. These distributions make up the right most column of the figure. The bottom left plot indicates the upper bound on $\mathbb{E}[h_P(x)]$ for different combinations of output distributions. Since each distribution is over two labels, the distributions are uniquely describes by just one of the label probabilities, p_1 for the first distribution and p_2 for the second.*

3.4 Invariant Regression Upper Bounds

Theorem 2. *Let us further impose that the model $\{h\}$ has G invariance. Then, ENCE given by Equation 4 is lower bounded by 0 and upper bounded by $1 + \mathop{\mathbb{E}}_{h_{\sigma^2}}\left[\frac{err_{reg}(h,\sigma^2)}{\|\sqrt{\frac{2}{\pi}}\sigma\|_2^2}\right]$. If $err_{reg}(h, \sigma^2)$ is minimized, then the upper bound becomes $1 + \mathop{\mathbb{E}}_{h_{\sigma^2}}\left[\frac{\int_{F'} q(Gx')\mathbb{V}_{Gx'}[f]dx'}{\|\sqrt{\frac{2}{\pi}}\sigma\|_2^2}\right]$.*

Proof. We begin by observing that the entries in $\sqrt{\frac{2}{\pi}}\sigma$ and $|h_\mu(x) - f(x)|$ are nonnegative. Therefore, we have that

$$0 \leq \mathbb{E}_{x,y}\left[\left\|\sqrt{\frac{2}{\pi}}\sigma - |h_\mu(x) - f(x)|\right\|_2^2 \;\middle|\; h_{\sigma^2}(x) = \sigma^2\right] \leq \mathbb{E}_{x,y}\left[\left\|\sqrt{\frac{2}{\pi}}\sigma\right\|_2^2 + \left\||h_\mu(x) - f(x)|\right\|_2^2 \;\middle|\; h_{\sigma^2}(x) = \sigma^2\right]. \tag{33}$$

Consequently,

$$0 \leq ENCE \leq \int_{\sigma^2} r(\sigma^2) \frac{\mathbb{E}_{x,y}\left[\left\|\sqrt{\frac{2}{\pi}}\sigma\right\|_2^2 + \left\||h_\mu(x) - f(x)|\right\|_2^2 \;\middle|\; h_{\sigma^2}(x) = \sigma^2\right]}{\left\|\sqrt{\frac{2}{\pi}}\sigma\right\|_2^2} d\sigma^2 \tag{34}$$

$$= \int_{\sigma^2} r(\sigma^2) \frac{\left\|\sqrt{\frac{2}{\pi}}\sigma\right\|_2^2 + \mathbb{E}_{x,y}\left[\left\||h_\mu(x) - f(x)|\right\|_2^2 \;\middle|\; h_{\sigma^2}(x) = \sigma^2\right]}{\left\|\sqrt{\frac{2}{\pi}}\sigma\right\|_2^2} \tag{35}$$

$$= \int_{\sigma^2} r(\sigma^2) \frac{\left\|\sqrt{\frac{2}{\pi}}\sigma\right\|_2^2 + \int_{X|_{h_{\sigma^2}(x)=\sigma^2}} p(x|h_{\sigma^2}(x) = \sigma^2)\left\|h_\mu(x) - f(x)\right\|_2^2 dx}{\left\|\sqrt{\frac{2}{\pi}}\sigma\right\|_2^2} d\sigma^2 \tag{36}$$

$$. \tag{37}$$

We can appeal to the definitions in our problem setup to conclude that

$$\int_{X|_{h_{\sigma^2}(x)=\sigma^2}} p(x|h_{\sigma^2}(x) = \sigma^2)\left\|h_\mu(x) - f(x)\right\|_2^2 dx = \int_{X'} q(x')\left\|h_\mu(x') - f(x')\right\|_2^2 dx' = \mathrm{err}_{\mathrm{reg}}(h, \sigma^2). \tag{38}$$

This gives us

$$ENCE \leq \int_{\sigma^2} r(\sigma^2)\left[1 + \frac{\mathrm{err}_{\mathrm{reg}}(h, \sigma^2)}{\|\sqrt{\frac{2}{\pi}}\sigma\|_2^2}\right] d\sigma^2 \tag{39}$$

$$= 1 + \mathop{\mathbb{E}}_{h_{\sigma^2}}\left[\frac{\mathrm{err}_{\mathrm{reg}}(h, \sigma^2)}{\|\sqrt{\frac{2}{\pi}}\sigma\|_2^2}\right]. \tag{40}$$

Now, if the domain restricted regression error is minimized, then by Theorem 4 we have that

$$\mathrm{err}_{\mathrm{reg}}(h, \sigma^2) = \int_{F'} q(Gx')\mathbb{V}_{Gx'}[f]dx'. \tag{41}$$

This completes the proof.

□

Remark 4. *An important reason why this works is that when you impose a domain restriction on X such that $h_{\sigma^2}(x) = \sigma^2$, $x' \in X' \implies gx' \in X'$ by assumption of invariance on $h_{\sigma^2}(x')$.*

Remark 5. *The upper bound in Equation 40 is not guaranteed to converge.*

In Appendix E, we discuss a class of functions where the approximation error for $G-$invariant functions can be made arbitarily small on each orbit.

Example 2. *In this example, we consider a set X of five chemical compounds and a codomain Y of spectra. We will calculate the ENCE upper bound assuming a minimized regression error. Each molecule x has information related to position and atomic number. We assume that our model class h has $E(2)$-invariance to the atomic positions, where $E(2)$ is the Euclidean group for $\mathbb{R}^2$. The atomic positions are shown in Figure 4. Note that for the purposes of this example we are not giving the exact atomic positions as you would find in Ramakrishnan et al. (2014) or a similar dataset. The dataset contains some duplicates of the chemical compounds up to transformations in $E(2)$. Different orientations are notated with $+$ and $\times$.*

X contains molecules $\{CH_4(+), CH_4(\times), H_20, SO_2, NH_3\}$. $f : X \to Y$ produces four unique spectra, one for each of the unique molecules. The model class h predicts three unique spectra, as it produces the same spectra for both H_20 and SO_2 due to their equivalence up to $e(3)$ in our simplified example. [1] *The model class h produces two distinct vectors $\vec{\sigma_1^2}$ and $\vec{\sigma_2^2}$. The probability of obtaining each of the molecules is $\{0.125, 0.125, 0.125, 0.125, 0.5\}$ for $\{CH_4(+), CH_4(\times), H_20, SO_2, NH_3\}$ respectively.*

Along each row of Figure 4 shows an orbit in the fundamental domain of X, the spectral lines from f and h_μ, and the σ line from h_{σ^2}. The bottom left shows the probability distribution over X. We start by considering the first row. Since h is $E(2)$ invariant, h is able to fully fit the function f on the orbit containing two rotated version of CH_4. Thus, the regression error is zero. Along the second row, h produces the same spectral lines for H_2O and SO_2 since they are the same up to $E(2)$ in our example. Since H_2O and SO_2 are equally probable, the output of h that minimizes the regression error is just the average of the two spectral lines.

We now refer to Equation 41 to compute the domain restricted regression error for $\vec{\sigma_1^2}$. The first orbit gave us a regression error of zero. As such, we must have $\mathbb{V}_{Gx'}[f] = 0$ on this orbit. This is certainly true, as f outputs the same spectra for what is the same molecule, methane. For the second orbit, we compute

$$q(Gx') = \frac{p(H_2O) + p(SO_2)}{p(CH_4(\times)) + p(CH_4(+)) + p(H_2O) + p(SO_2)} = \frac{0.125 + 0.125}{0.125 + 0.125 + 0.125 + 0.125} = \frac{1}{2}. \tag{42}$$

In other words, when restricted to the domain that outputs a variance of $\vec{\sigma^2}_1$, we have a 50% chance of being on the orbit containing H_2O and SO_2. Now, the mean of the function f on the orbit Gx' is just the average of the two spectral lines, since the probabilities of H_2O and SO_2 are equal. Now, the variance can be computed as the average distance of the real spectral lines from the prediction h, which is itself just the average of the two lines. Obtaining exact values for the spectra from Ramakrishnan et al. (2014) and Coblentz Society (1977), we can calculate a value of approximately 38.5 for the variance of f on the second orbit. Thus, we can compute $err_{reg}(h, \vec{\sigma_1^2}) = (0.5 \times 0) + (0.5 \times 38.5) = 19.25$.

As with the first orbit, the minimizing regression error on the third orbit is zero. Since there is only one element in X we can fit, the mapping between the molecule and the spectra is an isomorphism. This corresponds to no variance of f on Gx', the orbit corresponding to ammonia. So, $err_{reg}(h, \vec{\sigma_2^2}) = 0$.

Notice that the probability of obtaining $\vec{\sigma_1^2}$ is the same as the probability of obtaining $\vec{\sigma_2^2}$. Putting it all together, we get

$$ENCE \leq 1 + \left(0.5 \times \frac{0}{||\sqrt{\frac{2}{\pi}}\sigma_2^2||_2^2}\right) + \left(0.5 \times \frac{19.5}{||\sqrt{\frac{2}{\pi}}\sigma_1^2||_2^2}\right) \tag{43}$$

$$= 1 + \frac{9.625}{||\sqrt{\frac{2}{\pi}}\sigma_1^2||_2^2}. \tag{44}$$

[1] H_20 and SO_2 are both polarized molecules, hence the same apparent shape with the Mickey Mouse ears.

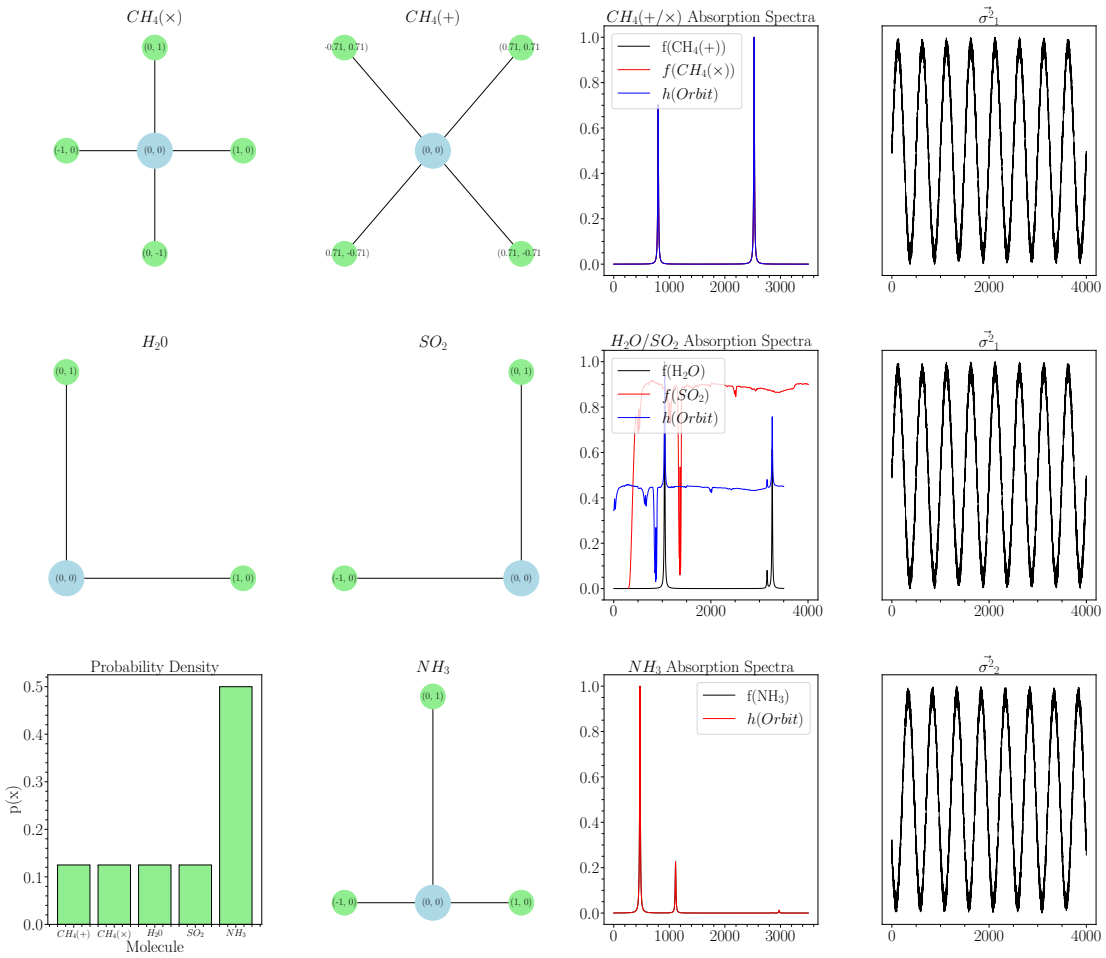

Figure 4: An example on how the ENCE upper bound behaves for an $e(3)$ invariant model class h on a set of molecules producing transmission spectra. Each row contains an orbit, the transmission spectra specified by f and the error minimizing h, the predicted variance vector. The bottom left of the plot additionally contains the PDF over the molecules in the domain X. The top two orbits have the same variance vector σ_1^2 white the final orbit has a distinct variance vector σ_2^2.

We conclude that the upper bound for ENCE in our example is $1 + \frac{9.625}{||\sqrt{\frac{2}{\pi}}\sigma_1^2||_2^2}$. *This is plotted as a function of* $||\sigma_1^2||_2^2$ *in Figure 5. We can see that in the limit as* $||\sqrt{\frac{2}{\pi}}\sigma_1^2||_2^2$ *goes to* ∞ *the upper bound on ENCE becomes unity. Alternatively, in the limit as* $||\sqrt{\frac{2}{\pi}}\sigma_1^2||_2^2$ *goes to* 0*, the upper bound on ENCE diverges.*

3.5 Equivariant Regression Upper Bounds

Theorem 3. *We assume the model class* $\{h\}$ *is equivariant but not necessarily invariant. ENCE given by Equation 4 is lower bounded by* 0 *and upper bounded by* $1 + \mathop{\mathbb{E}}_{h_{\sigma^2}}\left[\frac{err_{reg}(h,\sigma^2)}{\|\sqrt{\frac{2}{\pi}}\sigma\|_2^2}\right]$. *If* $err_{reg}(h,\sigma^2)$ *is minimized then the upper bound becomes* $1 + \mathop{\mathbb{E}}_{h_{\sigma^2}}\left[\frac{\int_F \int_G q(gx)\|f(gx) - g\mathbf{E}_G[f,x]\|_2^2 \alpha(x,g) dg dx}{\|\sqrt{\frac{2}{\pi}}\sigma\|_2^2}\right]$.

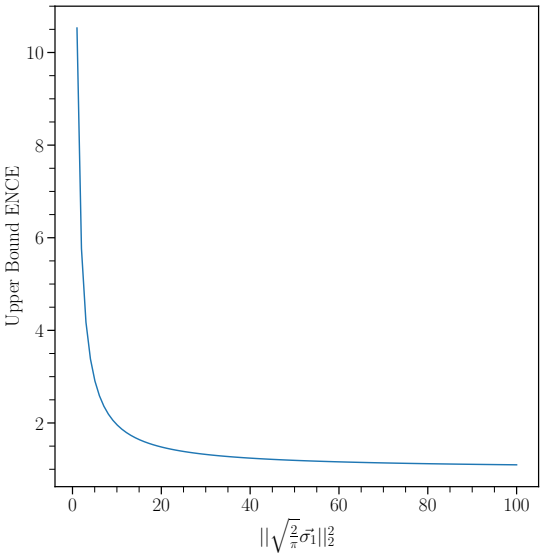

Figure 5: Upper bound on ENCE as a function of $||\sqrt{\frac{2}{\pi}}\sigma_1^2||_2^2$ in Example 2.

Proof. Unlike before, we can not decompose an integral over X' into an iterated integral over F' and Gx'. This is because without the assumption of invariance, the restriction $h_{\sigma^2}(x) = \sigma^2$ no longer preserves entire orbits. Therefore, we instead note that we can write the domain restricted regression error in terms of the original domain X. Consider the integral in Equation 2. Since $q(x) = 0$ for $h_{\sigma^2}(x) \neq \sigma^2$, we may write

$$\text{err}_{\text{reg}}(h, \sigma^2) = \int_X q(x) \Big\| h_\mu(x) - f(x) \Big\|_2^2 dx. \tag{45}$$

Said differently, the integrand takes on a value of 0 outside of X', so without loss of generality, we can replace X' with X and x' with x everywhere. Using the same argument as we did with invariant regression, we arrive at

$$ENCE \leq \int_{\sigma^2} r(\sigma^2) \left[1 + \frac{\text{err}_{\text{reg}}(h, \sigma^2)}{\| \sqrt{\frac{2}{\pi}} \sigma \|_2^2} \right] d\sigma^2 \tag{46}$$

$$= 1 + \mathop{\mathbb{E}}_{h_{\sigma^2}} \left[\frac{\text{err}_{\text{reg}}(h, \sigma^2)}{\| \sqrt{\frac{2}{\pi}} \sigma \|_2^2} \right]. \tag{47}$$

Applying Theorem 5, if $\text{err}_{\text{reg}}(h, \sigma^2)$ is a minimizer then we have

$$ENCE \leq 1 + \mathop{\mathbb{E}}_{h_{\sigma^2}} \left[\frac{\int_F \int_G q(gx) \| f(gx) - g\mathbf{E}_G[f, x] \|_2^2 \alpha(x, g) dg dx}{\| \sqrt{\frac{2}{\pi}} \sigma \|_2^2} \right]. \tag{48}$$

This completes the proof. □

3.6 Relation to Aleatoric Uncertainty

Our goal for this section is to motivate the definition of *aleatoric bleed* and examine how it relates to equivariance. We do this using the classic definition for err_{reg}. Again, see Appendix C for a summary of the formalism put forth in Wang et al. (2024). Consider a model that attempts to disentangle aleatoric and epistemic uncertainty using evidential priors, which we review in Appendix B. While we may not always have access to such in practice, denote the ground truth aleatoric uncertainty at a point x by $f(x)$. We do not consider a ground truth epistemic uncertainty, as such a thing is not well defined. Now consider a model class $\{h : X \to \sigma^2_{\text{aleatoric}}\}$. As before, h is constrained to be equivariant. We insist that both f and h are nonnegative, but are otherwise arbitrarily expressive.

Definition 1. *We define the aleatoric bleed as the regression error $err_{reg}(h)$ for a model h that is attempting to predict the aleatoric uncertainty $\sigma^2_{aleatoric}$.*

We note that by Theorem 5, the aleatoric bleed has a known lower bound. Moreover, the aleatoric bleed can be easily computed in cases where the ground truth aleatoric uncertainty is identically the zero vector, which we will explore in §4.

4 Experiments

In this section, we conduct exploratory data analysis to better understand the relationship between equivariance and uncertainty. We study miscalibration on real and simulated datasets for both calibration and regression tasks. In particular, we study the effect of different levels of symmetry mismatch using the framework of correct, incorrect, and extrinsic equivariance (§4.1). Next, we examine the effects of equivariance under different levels of ground truth uncertainty (§4.2). We then study how different types of equivariance contribute to aleatoric bleed (§4.3). We follow by studying the disagregation of uncertainties into epistemic and aleatoric uncertainty, and dissect how equivariance plays a role in such (§4.4). Finally, we show experimentally that the bound on ENCE from §3.5 can be realized in a true experimental setting (§4.5-4.6). All experiments in this work were ran on a NVIDIA A100 GPU (Choquette et al., 2021) and all terminated in under 8 hours.

4.1 Swiss Roll and the Equivariance Taxonomy

Here, we study the effect of different notions of equivariance on ECE. Specifically, we use the taxonomy of correct, incorrect, and extrinsic equivariance, which we review in Appendix C.1. Our study takes place on the vertically separated Swiss Roll distributions from (Wang et al., 2024). These distributions contain a $3D$ point cloud arranged in a spiral-like fashion, and binary labels are assigned to each point — see Figure 14 and Figures 7 and 11 in Wang et al. (2024). Specifically, we consider two different Swiss Roll distributions containing correct and incorrect equivariance. We build our dataset by taking samples from these distributions in different ratios. We then train an unconstrained MLP and a $z-$invariant network to predict the label. We provide further experimental details in Appendix G. Our aim is to realize a similar trend to accuracy in Wang et al. (2024) for ECE as a function of the correct vs. incorrect equivariance ratio used to build our dataset.

Results. While it may seem obvious that ECE is inversely correlated with model accuracy, we note this need not be the case. In fact, we will later see an example where it is not. Our experiment illustrates a circumstance wherein incorrect equivariance not only harms model accuracy, but also model calibration. When the correct ratio of equivariance is low, Figure 6 shows that not only will the model be correct less than about 70% of the time, but also its ECE will reach as high as 25%.

4.2 Galaxy Zoo Morphology and Trends in Group Order

The previous experiment suggested that a model's ECE improves as the amount of correct equivariance increases. Here, we seek to push the limits of just how far correct equivariance can help you. Specifically, we consider the task of galaxy morphology classification, which naturally has $E(2)-$invariance. We can

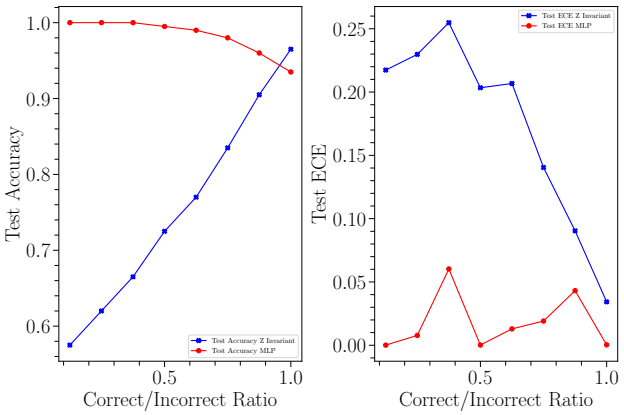

Figure 6: The left plot shows test accuracy for the z-invariant network (blue) and baseline unconstrained MLP (red) under different ratios of correct/incorrect ratios, ranging from 0% to 100% correctness. The right hand plot is the same but for ECE instead of accuracy.

approximate $E(2)-$invariance in CNNs with C_n and D_n group convolution layers, where C_n denotes the cyclic group of order n and D_n denotes the dihedral group of order n. We examine trends in group order n. While any C_n or D_n has correct equivariance, an increase in n certainly captures more of the underlying symmetry as you better approximate the $E(2)-$invariance, where E(2) is the euclidean group (with the caveat of aliasing, see Zhang (2019); Karras et al. (2021); Gruver et al. (2023)). We look at trends in both accuracy and ECE under different levels of point-spread function (PSF) convolution. This allows us to look at performance under different levels of ground truth noise, which should ideally affect both the model confidence and accuracy. PSF blurring is also an extremely relevant source of noise for the astrophysics community, specifically weak lensing analysis and exoplanet imaging, see Appendix H.1. Our data comes from Galaxy Zoo images (Walmsley et al., 2024) of galaxies from the DESI and SDSS surveys. Further experimental details are recorded in Appendix H.

Results. Figure 7 shows that ECE does not follow the same trends in cyclic group order as accuracy. The result is in accordance with the fact that ECE has no absolute lower bound, but accuracy *does* have a lower bound that explicitly depends on the type of equivariance in your model. This is echoed in Appendix J, which shows similar trends for both cyclic and dihedral group order for both DESI and SDSS surveys. We note that the accuracy curves are reminiscent of Weiler & Cesa (2019); Pandya et al. (2023).

4.3 Vector Field Regression, Equivariance Taxonomy, and Aleatoric Bleed

Vector Field Regression. Our goal for this section is to demonstrate the relationship between the equivariance taxonomy and aleatoric bleed in a regression setting. We consider a model $h : \mathbb{R}^3 \to \mathbb{R}^3 \times \mathbb{R}^3$ that predicts two vector fields representing a mean and a variance prediction. That is, we predict vectors attributed to any given point in $\mathbb{R}^3$. We denote the ground truth vector field at a given point x by $f(x)$, and h is constrained to be equivariant with respect to $E(3)$. Figure 8 presents two examples of how the equivariance taxonomy can result in different levels of aleatoric bleed.

We consider two different ground truth functions f designed to produce both correct and incorrect $E(3)-$equivariance such that we can explore how the equivariance taxonomy can affect downstream aleatoric bleed. ① ***Spiral.*** Let Q be a 90 degree rotation matrix in $\mathbb{R}^3$. We define $f(x) := Qx$. ② ***Sinusoidal.*** $f(x) := -\sin^2(||x||)x$. See that the spiral creates incorrect and extrinsic equivariance, since in general rotations in $\mathbb{R}^3$ do not commute. Therefore $f(gx) = Qgx$ and $g(f(x)) = gQx$ and $f(gx) \neq gf(x)$. For the sinusoidal case, we note that rotations, translations, and reflections in $\mathbb{R}^3$ preserve the norm of a vector x. We have that $f(gx) = -\sin^2(||gx||)gx = -\sin^2(||x||)gx$ and $g(f(x)) = g(-\sin^2(||x||)x) = -\sin^2(||x||)gx$. Therefore $gf(x) = f(gx)$, and we have correct equivariance. In both cases, f is completely deterministic, meaning any nonzero norm of a variance vector is indicative of aleatoric bleed.

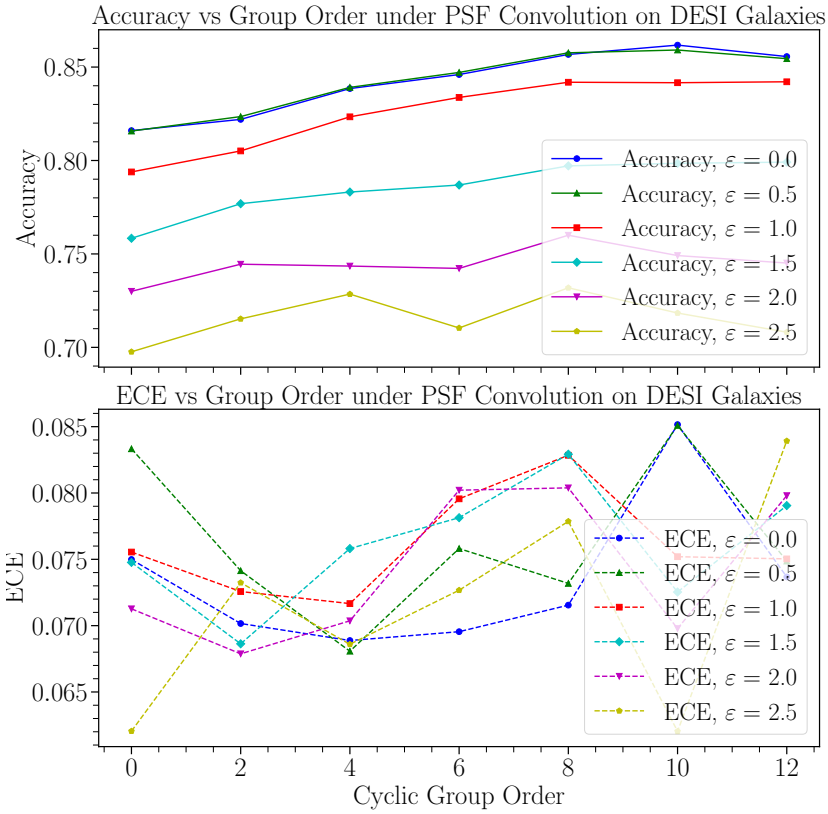

Figure 7: Accuracy and ECE vs Cyclic Group Order under PSF Convolution on DESI Galaxies.

For simplicity and visualization purposes, we construct our dataset with vectors in $\mathbb{R}^3$ with a z component of 0 and we choose rotation matrices Q that keep the vectors in the $xy-$plane for the spiral setup. We provide relevant training details in Appendix F.

Results. As expected, Figure 8 shows that the incorrect and extrinsic equivariance makes the $E(3)-$equivariant model unable to fit the data appropriately with its mean predictions. Consequently, it predicts extremely high variance vectors, as our $\beta-$NLL loss function can reach a local minimum when the variance prediction is significantly larger than the mean squared error. For details on the loss function and training setup, we remind the reader to see Appendix F. See in Figure 9 that despite the mean vector field failing to appropriately fit the data, the $\beta-$NLL loss is still fairly close to the MLP for vectors at any given angle θ. However, in the case of the correct equivariance with the sinusoidal dataset, the correctly applied $E(3)-$equivariance helps the model both in terms of MSE and $\beta-$NLL, accurately fitting the data and minimizng aleatoric bleed.

4.4 Chemical Properties and Aleatoric Bleed

The goal of this experiment is to assess whether a model's learned variance predictions are themselves reliable in a setting that is more realistic than the vector fields in Figure 8. That is, we ask if the model's confidence predictions are consistent with what the ground truth variance should be, and how equivariance can affect this. Our experiment is motivated by the analyses of Valdenegro-Toro & Mori (2022); Wimmer et al. (2023); Nevin et al. (2024); Jürgens et al. (2024), which illustrate that a model can conflate the uncertainty in its weights with the real uncertainty of the data.

Scalar-Valued Predictions. Our scalar-valued property prediction tasks take as input chemical compounds sourced from QM9s (Ramakrishnan et al., 2014). We predict various chemical properties with two dif-

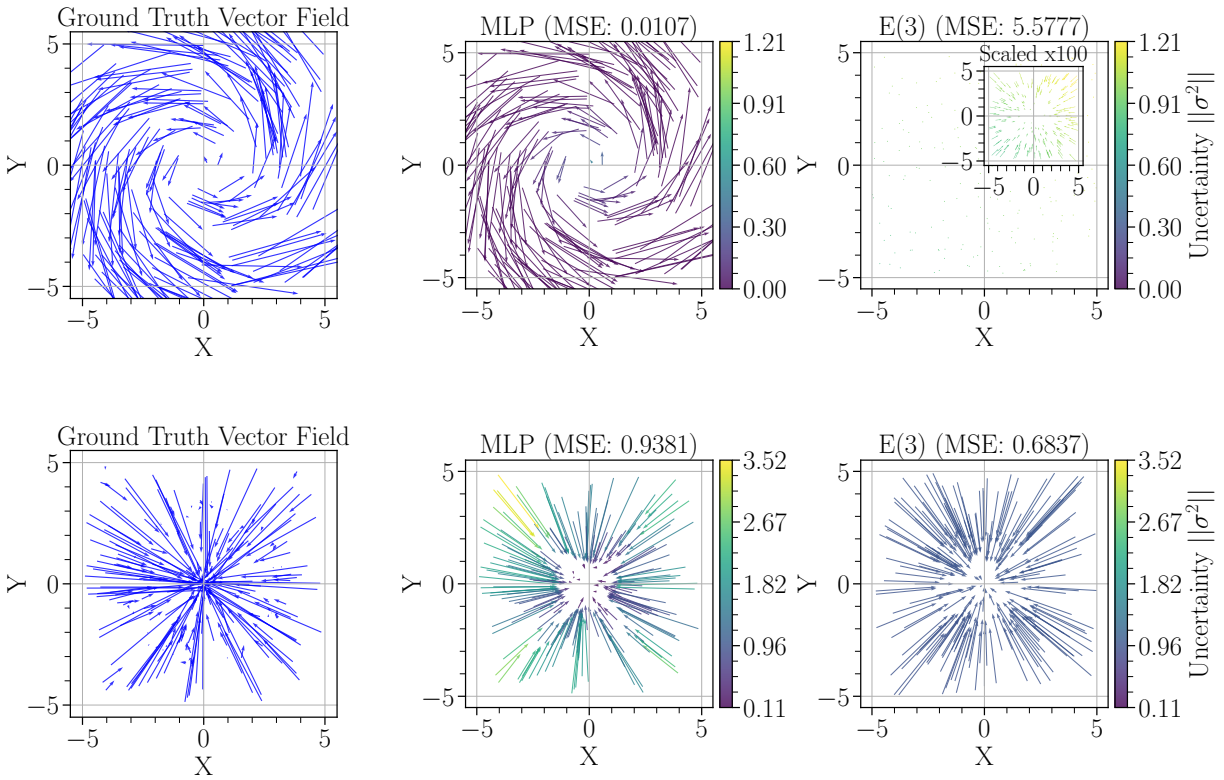

Figure 8: Vector regression results for the rotational and sinusoidal datasets (top and bottom respectively). For the model predictions in the middle and right columns, the color of the vector indicates the norm of the variance. The mean predictions for the $E(3)$ equivariant model on the rotational dataset are very small. The window in that image provides a 100 scaled up version of the image in order to see the behavior of the vectors.

ferent message-passing graph neural networks, one non-equivariant baseline and one with $E(3)-$equivariance. Specifically, our non-equivariant baseline is the GIN model (Xu et al., 2018) and we compare with an $E(3)-$invariant model (Batzner et al., 2022), with implementations based on (Backenköhler et al., 2023). Both models are equipped with independent feed forward neural network decoder heads which are used to learn a four parameter family that characterizes a Student t distribution. This in turn gives us enough degrees of freedom to reconcile epistemic and aleatoric uncertainties. We review the formalism that describes the relationship between these parameters and uncertainties in Appendix B. Further experimental details are provided in Appendix I. Physically, the relationship between these scalar values and chemical compounds should be a deterministic process, and accordingly the ground truth aleatoric uncertainty should always be zero. Since we are dealing with scalar values, aleatoric bleed reduces from a norm to a simple average over the square of predicted uncertainties. Note that the goal with this experiment is not to train the models to optimal performance; in fact, having models that cannot perfectly generalize is useful for us to study models with non-trivial uncertainties. We strive instead to compare models with similar accuracy but potentially varying levels of aleatoric bleed.

Scalar-Valued Results. In contrast to Figure 8, which showed the dangers of incorrect and extrinsic equivariance on aleatoric bleeding, we find that correct equivariance has little impact on aleatoric bleeding. As a specific case study, consider the dipole moment prediction task shown in Figure 10. We see that the aleatoric bleed is nearly identical between the GIN and $E(3)-$invariant models, with no significant deviations

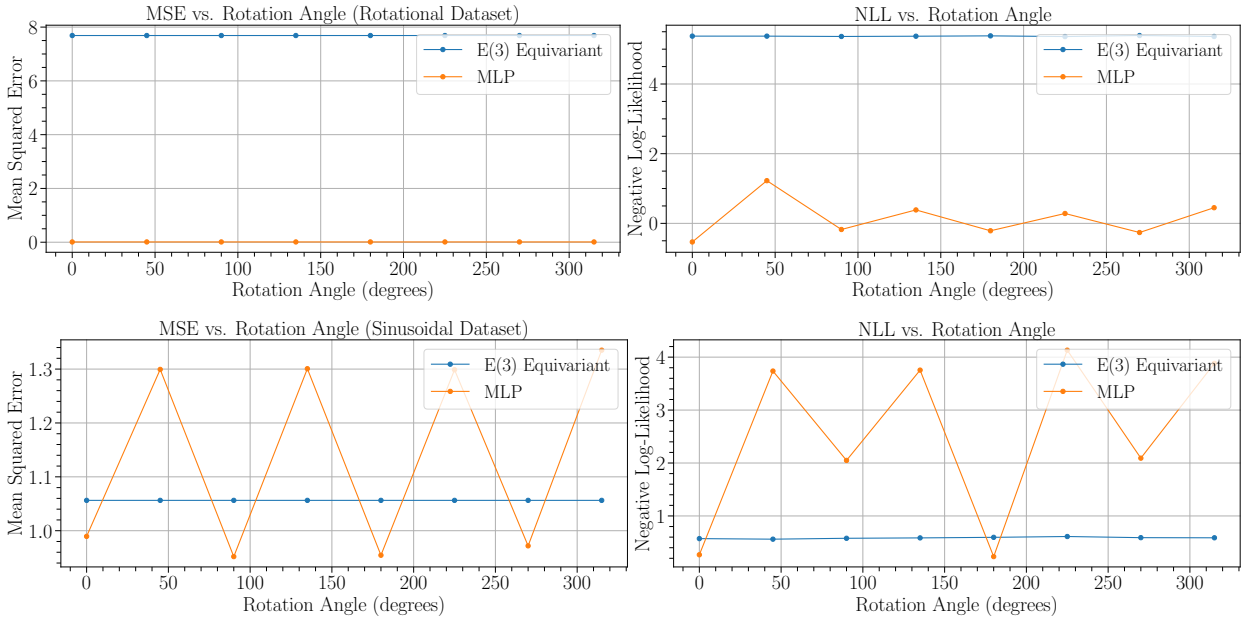

Figure 9: MSE and β-NLL losses for different rotation angle in the $xy-$plane for the rotational and sinusoidal datasets (top and bottom respectively).

in how errors are distributed. This result mirrors what we found for classification: correct equivariance does little to improve model calibration, but incorrect equivariance can punish a model all the same.

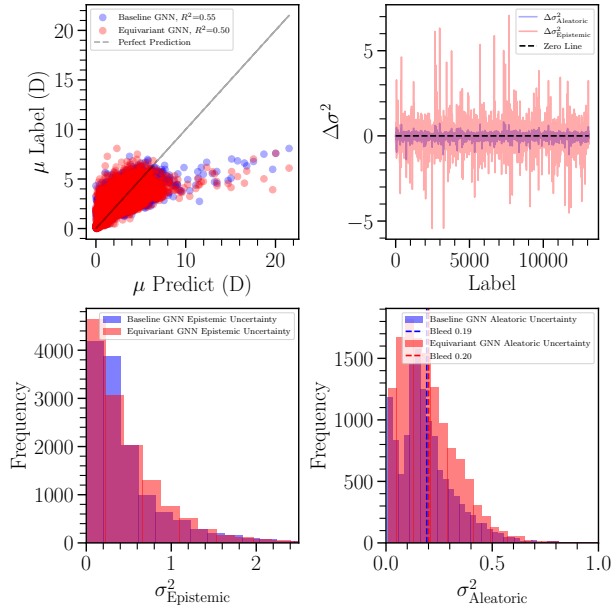

Figure 10: ***Top Left:*** Prediction versus label for both GIN and $E(3)-$Invariant models. ***Top Right:*** The difference between aleatoric and epistemic uncertainties between the two models for each label. ***Bottom Left:*** A distribution of the epistemic uncertainty predictions for the two models. ***Bottom Right:*** A distribution of the aleatoric uncertainty predictions for the two models.

Chemical Property	Unit	GIN MAE	$E(3)-$Invariant MAE	GIN AB	$E(3)-$Invariant AB
ε_{LUMO}	eV	0.4404	0.6710	**0.0081**	3.0288
$\Delta\varepsilon$	eV	0.6877	0.7283	**0.0013**	3.0287
U_0	eV	0.2563	0.0563	0.0333	**0.0053**
U	eV	0.2558	0.0563	0.0323	**0.0053**
U_0^{ATOM}	eV	0.1908	0.1458	0.0193	**0.0052**
G^{ATOM}	eV	0.7954	0.7706	**0.0000**	0.0014
A	GHz	0.0667	0.2504	**0.0000**	0.0014
B	GHz	0.1625	0.0992	0.0066	**0.0040**
C	GHz	0.0780	0.0493	**0.0008**	0.0013
$\langle R^2 \rangle$	$(a_0)^2$	0.8621	0.8956	**0.0005**	0.0013

Table 1: Accuracy and Aleatoric Bleed (AB) for various scalar properties in QM9s for baseline and equivariant graph neural network models. Predictions and error estimates are given for the $z-$scored scalar values.

In Table 1, we compare the aleatoric bleed between the baseline and equivariant models when their accuracy is comparable, which we quantify as having a mean absolute error (MAE) within 0.25. The model with the lower aleatoric bleed seems to depend neither on the performance of the model nor the inclusion of equivariance. Our findings support the same conclusion that correct equivariance does little to help prevent aleatoric bleed.

In §3.6, we discussed how the aleatoric bleed has a known lower bound. The results from this experiment suggest that the lower bound is not tight enough to be meaningful to practitioners working on scalar-valued properties in the QM9 dataset. This is likely because our invariance constraint here is correct, so the lower bound on aleatoric bleed is zero. This is in contrast to the vector regression spiral experiment where incorrect equivariance clearly caused an increase of aleatoric bleed.

Vector-Valued Predictions. This experiment highlights a need for qualitative analysis to work in tandem with our aleatoric bleeding metric for high-dimensional outputs. In particular, aleatoric bleed fails to describe the individual coordinates in which the predicted variance vector suffers the most in terms of bleeding.

This experiment again uses QM9s, however, this time we instead predict spectral lines emitted from the chemical compounds using a network with steerable $E(3)-$vectors (Brandstetter et al., 2021). As before, we use independent feed forward neural network decoder heads in order to learn a four parameter family that characterizes a Student t distribution. Again, physics tells us that the ground truth aleatoric uncertainty should ideally be exactly zero. Any non-zero uncertainties are indicative of epistemic uncertainties bleeding into the aleatoric uncertainty prediction. As such, the aleatoric bleed becomes the mean squared norm of the predicted aleatoric variance vectors.

Vector-Valued Results. We compute an aleatoric bleed of ≈ 17.613. However, as seen in Figure 11, the model tends to conflate high frequency signals with noise. The model's aleatoric uncertainty follows the epistemic uncertainty quite closely, indicating that the model can not tell them apart. This is echoed with 31 examples in the Appendix K. Thus, we conclude that the aleatoric bleed is only vivid enough to tell a practitioner that the model is confusing different sources of uncertainty. Evidently, the metric can not tell a practitioner *where* the model tends to mess up the uncertainty estimate without additional diagnostics. That is not to say the aleatoric bleed metric is not useful; rather, we suggest practitioners should proceed with caution when examining aleatoric bleed.

Additionally, while not our main contribution, we point out that our model's raw performance is comparable to the state-of-the-art DetaNet (Zou et al., 2023), with both models achieving R^2 scores above 0.9 and often close to 1.0 on QM9s test molecules. Our steerable $E(3)$ vector-based model also provides initial steps towards quantifying its own uncertainty. We further discuss the merits and limitations of our approach compared to DetaNet in Appendix I. We leave further model development and evaluation as an oppurtunity for future work.

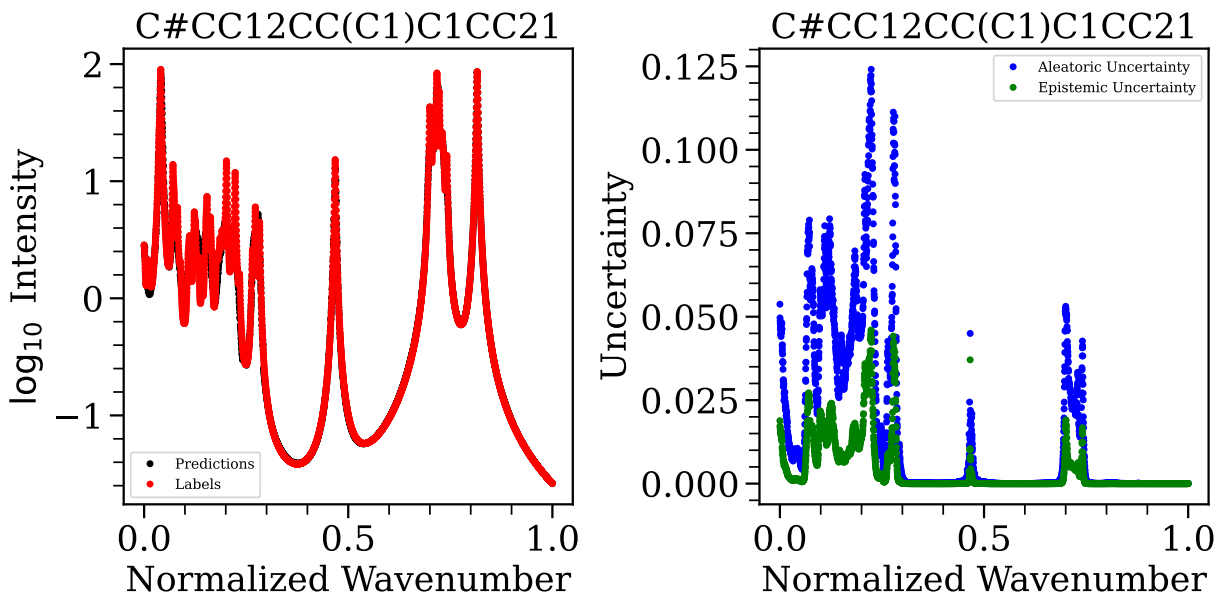

Figure 11: ***Left:*** Sample prediction vs ground truth spectra for the molecule given by SMILES (Weininger, 1988) string $C\#CC12CC(C1)C1CC21$. ***Right:*** The model's predicted aleatoric and epistemic uncertainties for each of the normalized wavenumbers.

4.5 Chemical Properties and Binning Approximations for ENCE

Binning Approximations. A necessary prerequisite for attempting to realize our upper bound on ENCE as derived in §3.5 is to understand if we can well approximate Equation 4 with binning approximations. Specifically, we must create discrete bins for our output variance vectors so that we can approximate the continuous density $r(\sigma^2)$. While we need not use approximations to compute the theoretical upper bound on the calibration error term in Equation 4, the binning approximations for computing the density term are unavoidable (Guo et al., 2017). We motivate our binning approximations for vector-valued regression talks with the established formalism for scalar-valued variance prediction binning approximations. Levi et al. (2022a) employ binning approximations as follows: Assume that your model outputs a mean $\hat{y}$ and a standard deviation σ. First, divide samples σ into N equally spaced bins B_j between the minimum and maximum σ predictions. Group the model outputs according to the σ bins. The following per bin quantities are then computed:

$$RMV(j) = \sqrt{\frac{1}{|B_j|}\sum_{t\in B_j}\sigma_t^2} \tag{49}$$

$$RMSE(j) = \sqrt{\frac{1}{|B_j|}\sum_{t\in B_j}(y_t-\hat{y}_t)} \tag{50}$$

$$ENCE = \frac{1}{N}\sum_{j=1}^{N}\frac{|RMV(j)-RMSE(j)|}{RMV(J))}. \tag{51}$$

For our generalized ENCE metric, we adopt a similar procedure for approximating with bins. If our variance vector is in $\mathbb{R}^m$, then we divide *each dimension* m into N bins $B_{n,m}$. For each dimension, the bins are equally sized and bounded by the minimum and maximum variance prediction in that coordinate. Each coordinate of a model's variance output will be in one of N bins in each dimension m. There will therefore also be sequences of m bins $B_j := \{B_i\}_{i=1}^m \subseteq B_{n,m}$ that contain entire vectors. If a variance vector falls into a bin

sequence B_j, then we may index that vector with a subscript t along with the corresponding error vector. Our approximation is thus as follows:

$$h_\sigma = \sqrt{h_{\sigma^2}} \tag{52}$$

$$e = |h_\mu - f| \tag{53}$$

$$MNS(j) = \frac{1}{|B_j|} \sum_{t \in B_j} \| \sqrt{\frac{2}{\pi}} h_{\sigma,t} \|_2^2 \tag{54}$$

$$MNCE(j) = \frac{1}{|B_j|} \sum_{t \in B_j} \| \sqrt{\frac{2}{\pi}} h_{\sigma,t} - e_t \|_2^2 \tag{55}$$

$$ENCE = \frac{1}{N} \sum_j |B_j| \frac{MNCE(j)}{MNS(j)} \tag{56}$$

where MNS abbreviates mean norm standard deviation and $MNCE$ abbreviates mean norm calibration error. The fraction $\frac{|B_j|}{N}$ can be intuited as the term approximating the true probability density $r(\sigma^2)$.

The task of using bins to approximate ENCE is already difficult in the case of scalar values. As Pernot (2023) demonstrates, the ENCE metric for scalar valued predictions (Levi et al., 2022a) already exhibits an "annoying" correlation with the number of bins used. Specifically, for N bins used, the ENCE grows linearly with $N^{1/2}$ for scalar-valued outputs.

We specifically compute calibration error in terms of the predicted epistemic uncertainty. Calibration error is intended to compare a model's own confidence with its error, making epistemic uncertainty the more suitable thing to use. Additionally, given our discussions of aleatoric bleed, we find the epistemic uncertainty to be a more reliable metric here.

Binning Approximation Results. Intriguingly, we find that our generalized ENCE instead increases linearly with N when applied to our chemical spectra results, as shown in Figure 12 . We study this for up to 100 bins, in accordance with Scalia et al. (2020), even though most other works use $\sim 10 - 15$ bins (e.g., Palmer et al., 2022; Busk et al., 2021; Levi et al., 2022a; Scalia et al., 2020). Our difference in scaling law can potentially be attributed to two factors. First, our metric penalizes differences in $\sim 0.8\times$ the standard deviation with mean absolute error instead of differences of variance with mean squared error. Second, this is a possible result of the curse of dimensionality. The probability that any two vectors occupy the same region in $\mathbb{R}^n$ is very small for large n, this in turn makes it hard to approximate the true density $r(\sigma^2)$ in Equation 4. In this specific case we are in $\mathbb{R}^{3501}$, the resolution of the spectral line.

The consequence of this dependence on bin number N has the following sequence: we can only use ENCE to inform us of relative model miscalibration. That is, we can use the metric to tell us if one model is more miscalibrated than another. However, due to this dependence on the number of bins N, we can not reliably conclude that a model is well calibrated (i.e. a model is confident in correct results) simply because it has a low ENCE score.

4.6 Chemical Properties and Realizing the Upper Bound on ENCE

Realizing the Bound. In this experiment, we revisit the spectral line predictions from §4.4 and show that the computed ENCE respects the upper bound given in §3.5 *regardless* of the choice of bin number. In §4.5 , the number of bins considered was based on a range from previous works. Accordingly, the following calculations will show that no matter how liberal we are with binning, the upper bound still holds.

In Equation 40, we showed that the upper bound on ENCE is related to the domain restricted regression error. Since the model we are studying is specifically invariant, we are interested in the error on entire orbits that are in the pre-image of $h_{\sigma^2}^{-1}(\sigma^2)$. The QM9 dataset is canonicalized such that molecules are unique up to rotation. Therefore, we include 90, 180, and 270 degree rotations of each molecule in the test set. In this way, we can compute the upper bound on regression error on non-trivial orbits and in turn compute an upper

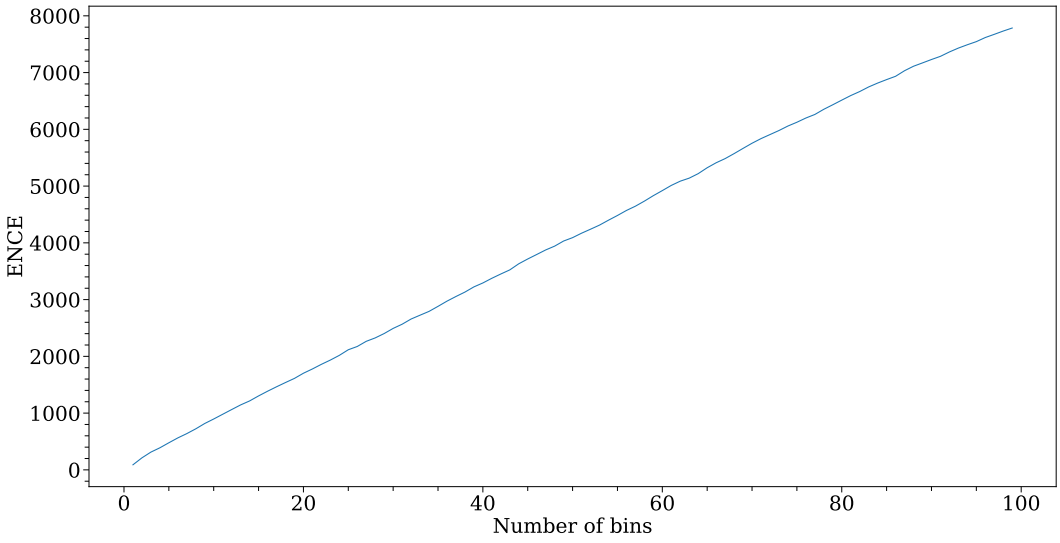

Figure 12: ENCE as a function of bins N.

bound on ENCE. [2] For the true label corresponding to the rotated molecules in the test set, we consider two variations, each corresponding to rows (1) and (2) in Figure 4. ① ***Correct Augmentation.*** Here, the labels for the rotated molecules are the same as the original. In this circumstance, our $E(3)-$invariance is correct, and we expect our ENCE to be well within the bound. Additionally, we expect that this data augmentation matches the underlying physics of our world. ② ***Incorrect Augmentation.*** Here, the spectral line has an explicit dependence on the rotation angle of the molecule. Specifically, each 90 rotation corresponds to a reflection over the $y-$axis. This augmentation is not true to real physics, and is done to make our data have incorrect $E(3)-$invariance. In this way, we can illustrate how having incorrect equivariance takes us closer to our theoretical upper bound.

To upper bound the regression error on each orbit, we put two clamps on our $E(3)-$invariant model, which we view as part of the function approximator h. Since our upper bound is given by $\mathop{\mathbb{E}}_{h_{\sigma^2}}\left[\frac{\text{err}_{\text{reg}}(h,\sigma^2)}{||\sqrt{\frac{2}{\pi}}\sigma||_2^2}\right]$, these clamps ensure that $\text{err}_{\text{reg}}(h,\sigma^2) \not\to \infty$ and $||\sqrt{\frac{2}{\pi}}\sigma||_2^2 \not\to 0$. In turn, we can compute an upperbound $ENCE < \infty$. The first clamp bounds the output to be between -3.3 and 3.3 on each coordinate, which is roughly the minimum and maximum of the $\log_{10}$ intensities of the true spectra and is in accordance with the fact that we trained on the spectra normalized by a factor of $1/3.3$. This clamp is after the pooling layer that ensures invariance and thus does not cause the $E(3)-$invariance to break. With this clamp, we are able to determine the max error for each label. We also impose a clamp on the variance prediction. If the variance prediction goes towards the zero vector then the upper bound on ENCE will also grow without bound. Thus, we clamp the variance predictions such that the minimum the norm $||\sqrt{\frac{2}{\pi}}\sigma||_2^2$ can be is ≈ 7.0, which was roughly the minimum of the computed samples before we introduced a clamp. With these clamps and the approximation of the density $r(\sigma^2)$, we are able to get compute the true upper bound on h subject to the binning approximations needed to approximate $r(\sigma^2)$.

Realizing the Bound Results. As seen in Figure 13, our computed ENCE is less than the predicted upper bound no matter how many bins are used to approximate the density $r(\sigma^2)$. Since our model seems to converge fairly well to the true absorption spectra, it is to be expected that our result is well within the

[2] An important reason why this works is that the spherical harmonic embedding of coordinates in $\mathbb{R}^3$ is equivariant (Appendix A4. Brandstetter et al., 2021).

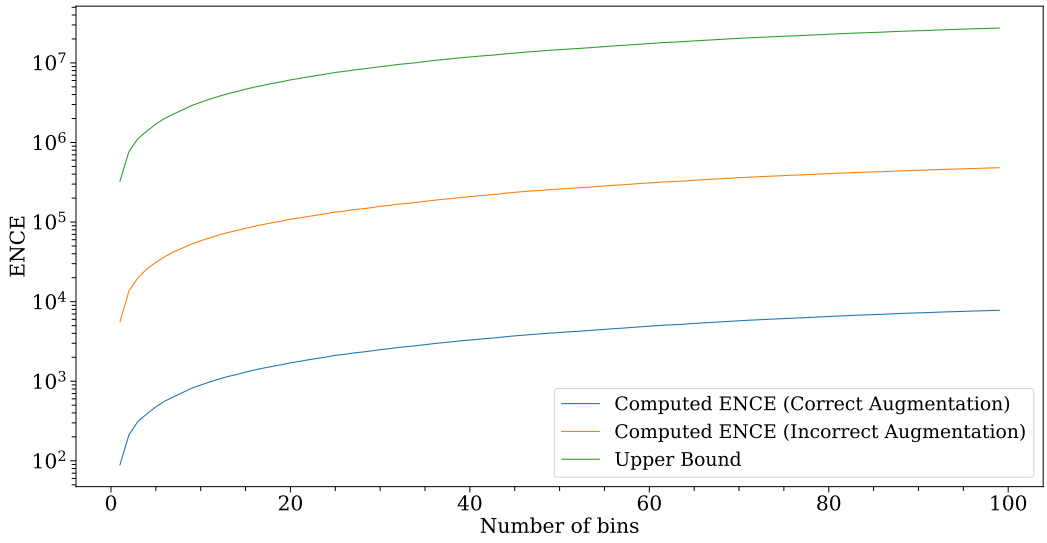

Figure 13: The Computed ENCE versus the theoretical upper bound.

upper bound for correct augmentation. Similarly, incorrect augmentation introduces error into our results, causing bringing our computed ENCE closer to its upper bound.

5 Limitations

The Potential for Biased Estimators. ECE and ENCE are difficult to compute in practice due to discrete binning approximations (Pernot, 2023), as the lack of an unbiased estimator adds uncertainty as to how reliable computed ECE and ENCE scores are. In this work, we attempted to circumnavigate the lack of a known unbiased estimator in the following ways. First, in classification settings where the model is consistently high in its confidence, we increase the granularity at which we bin, see Appendix G and H. Second, we carefully studied the effect of binning approximations in §4.5 and used that as a form of ablation in §4.6.

Single Label Metric. ECE only computes miscalibration with respect to a single label, but does not consider secondary or tertiary outputs that may be useful to practitioners as done in (Nixon et al., 2019).

Upper Bound Computation. Our computation of the upper bound in §4.6 depended upon experiment-specific clamps on our model output that correspond with an understanding of the underlying physics, relevant inductive biases, and the dataset. As such, these clamps are not generalizable outside of this specific case study. We recognize that different experimental settings may require different constraints to realize the proven upper bounds.

6 Discussion and Conclusions

Experiments in the natural sciences, especially in data-sparse settings, ***demand*** both equivariance and uncertainty estimation, and yet, no general theory for explaining how equivariance relates to uncertainty exists in the literature. We fill this gap, presenting the first unified theory explaining how equivariance relates to uncertainty estimation. We proved upper bounds on model calibration error for a class of functions that are arbitrarily expressive except that they are constrained to be equivariant. We do this in both classification

and regression settings. This leads onto a set of experiments illustrating how the proven bounds manifest in practice.

Our core takeaways are best explained in light of the equivariance taxonomy. We highlight how incorrect and extrinsic equivariance can provably degrade model calibration for both classification and regression tasks. Our experiments support this. In the cases of the swiss roll and vector field regression experiments, both incorrect and extrinsic equivariance not only make the model less accurate, but also more poorly calibrated. In the case of the vector field regression experiment we also saw that incorrect and extrinsic equivariance contributed to aleatoric bleed. By contrast, we have shown that a model with correct equivariance is not necessarily better calibrated than a similarly sized non-equivariant baseline. As illustrated by the galaxy morphology classification and scalar-valued chemical property prediction experiments, equivariance is not strong enough to help a model become well calibrated nor is it strong enough to prevent aleatoric bleeding. It was only for the vector regression experiment on the sinusoidal dataset that the introduction of correct equivariance was able to significantly improve the raw performance and prevent aleatoric bleeding. The vector regression result is especially interesting in light of the galaxy morphology classification experiment, which showed that correct equivariance can help a model perform better without necessarily making it better calibrated.

7 Future Work

Having laid the groundwork for a first unified theory for equivariance and uncertainty, there are several interesting and exciting potential avenues for future work:

- Our upper bound for ENCE presented in §3.5 is not necessarily the supremum, and we seek to try and tighten the bound such that it more closely follows the observed ENCE.
- Our theory on regression could potentially be extended to circumstances where the output mean and variance predictions obey different $G-$equivariances. We touch on this only briefly in Appendix D.
- We seek to better understand the relationship between miscalibration, incorrect and extrinsic equivariance, and model scale. We intend to study these scaling laws with a study similar to Brehmer et al. (2024).
- An unbiased estimator for ENCE that does not depend on binning approximations would be an extremely useful contribution, and we will study such estimators in forthcoming work.
- One experimental domain where our work is closely applicable that we will further investigate is robotics. While our work can be difficult to frame in typical Q-learning setups since the optimal solution to the Bellman update function is unique (c.f. Figure 1), a discussion of both equivariance and uncertainty quantification lends itself naturally to behavior cloning tasks such as in Jia et al. (2023). Accordingly, future work will examine calibration error for equivariant models in an imitation learning environment.
- Another closely related experimental domain is cosmological large-scale structure. In particular, we will benchmark calibration error with symmetry-preserving models using the benchmark released in Balla et al. (2024). The appeal of this benchmark is that it includes graph-level predictions on ΛCDM (Ryden, 2016; Carroll, 2019) cosmological parameters Ω_m and σ_8. In cosmology, these predictions are more commonly phrased as constraints on posterior distributions rather than point estimates (e.g., Collaboration: et al., 2016; Abbott et al., 2022), and so this is a natural playground for our work.

8 Reproducibility Statement

Our codebase containing all of our experiments, as well as the instructions to reproduce our results, is publicly available at .

We provide further experimental details and sources for the datasets used in this work throughout Appendices F, G, H, and I. We'd like to highlight the work of Wang et al. (2024) and Pandya et al. (2025); the artifacts associated with these two works were simple to reproduce and aided us in our study.

9 Ethics Statement

Authors have no conflicts of interest to disclose.

References

Timothy MC Abbott, Michel Aguena, Alex Alarcon, S Allam, O Alves, A Amon, F Andrade-Oliveira, James Annis, S Avila, D Bacon, et al. Dark energy survey year 3 results: Cosmological constraints from galaxy clustering and weak lensing. *Physical Review D*, 105(2):023520, 2022.

Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J Ballard, Joshua Bambrick, et al. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, pp. 1–3, 2024.

Alexander Amini, Wilko Schwarting, Ava Soleimany, and Daniela Rus. Deep evidential regression. *Advances in neural information processing systems*, 33:14927–14937, 2020.

Michael Artin. *Algebra*. Birkhäuser, 1998.

Michael Backenköhler, Paula Linh Kramer, Joschka Groß, Gerrit Großmann, Roman Joeres, Azat Tagirdzhanov, Dominique Sydow, Hamza Ibrahim, Floriane Odje, Verena Wolf, and Andrea Volkamer. TeachOpenCADD goes Deep Learning: Open-source Teaching Platform Exploring Molecular DL Applications. *ChemRxiv preprint*, 2023. doi: 10.26434/chemrxiv-2023-kz1pb.

Julia Balla, Siddharth Mishra-Sharma, Carolina Cuesta-Lazaro, Tommi Jaakkola, and Tess Smidt. A cosmic-scale benchmark for symmetry-preserving data processing. *arXiv preprint arXiv:2410.20516*, 2024.

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):2453, 2022.

Eric Beh. Exploring how to simply approximate the p-value of a chi-squared statistic. *Austrian Journal of Statistics*, 47(3):63–75, 2018.

Edward Berman and Jacqueline McCleary. Shopt.jl: A julia package for empirical point spread function characterization of jwst nircam data. *Journal of Open Source Software*, 9(100):6144, 2024. doi: 10.21105/joss.06144. URL https://doi.org/10.21105/joss.06144.

Edward Berman and Jacqueline McCleary. On Differentiable Correlation Functions. In *American Astronomical Society Meeting Abstracts*, volume 245 of *American Astronomical Society Meeting Abstracts*, pp. 424.01, January 2025.

Edward Berman, Sneh Pandya, Jacqueline McCleary, Marko Shuntov, Caitlin Casey, Nicole Drakos, Andreas Faisst, Steven Gillman, Ghassem Gozaliasl, Natalie Hogg, Jeyhan Kartaltepe, Anton Koekemoer, Wilfried Mercier, Diana Scognamiglio, COSMOS-Web, :, and The JWST Cosmic Origins Survey. On soft clustering for correlation estimators: Model uncertainty, differentiability, and surrogates, 2025. URL https://arxiv.org/abs/2504.06174.

Edward M. Berman, Jacqueline E. McCleary, Anton M. Koekemoer, Maximilien Franco, Nicole E. Drakos, Daizhong Liu, James W. Nightingale, Marko Shuntov, Diana Scognamiglio, Richard Massey, Guillaume Mahler, Henry Joy McCracken, Brant E. Robertson, Andreas L. Faisst, Caitlin M. Casey, Jeyhan S. Kartaltepe, and COSMOS-Web: The JWST Cosmic Origins Survey. Efficient point-spread function modeling with shopt.jl: A point-spread function benchmarking study with jwst nircam imaging. *The Astronomical Journal*, 168(4):174, sep 2024. doi: 10.3847/1538-3881/ad6a0f. URL https://dx.doi.org/10.3847/1538-3881/ad6a0f.

Simon Birrer, Anowar J. Shajib, Daniel Gilman, Aymeric Galan, Jelle Aalbers, Martin Millon, Robert Morgan, Giulia Pagano, Ji Won Park, Luca Teodori, Nicolas Tessore, Madison Ueland, Lyne Van de Vyvere, Sebastian Wagner-Carena, Ewoud Wempe, Lilan Yang, Xuheng Ding, Thomas Schmidt, Dominique Sluse, Ming Zhang, and Adam Amara. lenstronomy ii: A gravitational lensing software ecosystem. *Journal of Open Source Software*, 6(62):3283, 2021. doi: 10.21105/joss.03283. URL https://doi.org/10.21105/joss.03283.

Benjamin Bloem-Reddy, Yee Whye, et al. Probabilistic symmetries and invariant neural networks. *Journal of Machine Learning Research*, 21(90):1–61, 2020.

Olivier Bousquet, Stéphane Boucheron, and Gábor Lugosi. Introduction to statistical learning theory. *Advanced Lectures on Machine Learning: ML Summer Schools 2003, Canberra, Australia, February 2-14, 2003, Tübingen, Germany, August 4-16, 2003, Revised Lectures*, pp. 169–207, 2004.

Johannes Brandstetter, Rob Hesselink, Elise van der Pol, Erik J Bekkers, and Max Welling. Geometric and physical quantities improve e (3) equivariant message passing. *arXiv preprint arXiv:2110.02905*, 2021.

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

Michael M Bronstein, Joan Bruna, Taco Cohen, and Petar Veličković. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. *arXiv preprint arXiv:2104.13478*, 2021.

Jonas Busk, Peter Bjørn Jørgensen, Arghya Bhowmik, Mikkel N Schmidt, Ole Winther, and Tejs Vegge. Calibrated uncertainty for molecular property prediction using ensembles of message passing neural networks. *Machine Learning: Science and Technology*, 3(1):015012, 2021.

Sean M Carroll. *Spacetime and geometry*. Cambridge University Press, 2019.

Caitlin M Casey, Jeyhan S Kartaltepe, Nicole E Drakos, Maximilien Franco, Santosh Harish, Louise Paquereau, Olivier Ilbert, Caitlin Rose, Isabella G Cox, James W Nightingale, et al. Cosmos-web: an overview of the jwst cosmic origins survey. *The Astrophysical Journal*, 954(1):31, 2023.

Mostafa Cherif, Tobías I Liaudat, Jonathan Kern, Christophe Kervazo, and Jérôme Bobin. Uncertainty quantification for fast reconstruction methods using augmented equivariant bootstrap: Application to radio interferometry. *arXiv preprint arXiv:2410.23178*, 2024.

Jack Choquette, Wishwesh Gandhi, Olivier Giroux, Nick Stam, and Ronny Krashinsky. Nvidia a100 tensor core gpu: Performance and innovation. *IEEE Micro*, 41(2):29–35, 2021.

Coblentz Society. Sulfur dioxide (so_2) infrared spectrum. https://webbook.nist.gov/cgi/cbook.cgi?ID=C7446095&Index=0&Type=IR-SPEC, 1977. URL https://webbook.nist.gov/cgi/cbook.cgi?ID=C7446095&Index=0&Type=IR-SPEC. Spectrum recorded by Dow Chemical Company, digitized by NIST.

Dark Energy Survey Collaboration:, T Abbott, FB Abdalla, J Aleksić, S Allam, A Amara, D Bacon, E Balbinot, M Banerji, K Bechtol, et al. The dark energy survey: more than dark energy–an overview. *Monthly Notices of the Royal Astronomical Society*, 460(2):1270–1299, 2016.

Ilias Diakonikolas, Samuel B Hopkins, Daniel Kane, and Sushrut Karmalkar. Robustly learning any clusterable mixture of gaussians. *arXiv preprint arXiv:2005.06417*, 2020.

Carlos Esteves. Theoretical aspects of group equivariant neural networks. *arXiv preprint arXiv:2004.05154*, 2020.

Brandon Y Feng, Rodrigo Ferrer-Chávez, Aviad Levis, Jason J Wang, Katherine L Bouman, and William T Freeman. Exoplanet detection via differentiable rendering. *arXiv preprint arXiv:2501.01912*, 2025.

Marc Finzi, Samuel Stanton, Pavel Izmailov, and Andrew Gordon Wilson. Generalizing convolutional neural networks for equivariance to lie groups on arbitrary continuous data. In *International Conference on Machine Learning*, pp. 3165–3176. PMLR, 2020.

Marc Finzi, Gregory Benton, and Andrew G Wilson. Residual pathway priors for soft equivariance constraints. *Advances in Neural Information Processing Systems*, 34:30037–30049, 2021.

Wendy L Freedman, Barry F Madore, In Sung Jang, Taylor J Hoyt, Abigail J Lee, and Kayla A Owens. Status report on the chicago-carnegie hubble program (cchp): three independent astrophysical determinations of the hubble constant using the james webb space telescope. *arXiv preprint arXiv:2408.06153*, 2024.

Jiahui Fu, Yilun Du, Kurran Singh, Joshua B Tenenbaum, and John J Leonard. Neuse: Neural se (3)-equivariant embedding for consistent spatial understanding with objects. *arXiv preprint arXiv:2303.07308*, 2023.

Dominik Fuchsgruber, Tom Wollschläger, and Stephan Günnemann. Energy-based epistemic uncertainty for graph neural networks. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=6vNPPtWH1Q.

Roy C Geary. The ratio of the mean deviation to the standard deviation as a test of normality. *Biometrika*, 27(3/4):310–332, 1935.

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

Mario Geiger, Tess Smidt, Alby M., Benjamin Kurt Miller, Wouter Boomsma, Bradley Dice, Kostiantyn Lapchevskyi, Maurice Weiler, Michał Tyszkiewicz, Simon Batzner, Dylan Madisetti, Martin Uhrin, Jes Frellsen, Nuri Jung, Sophia Sanborn, Mingjian Wen, Josh Rackers, Marcel Rød, and Michael Bailey. Euclidean neural networks: e3nn, April 2022. URL https://doi.org/10.5281/zenodo.6459381.

Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, 102(477):359–378, 2007.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 6.2. 2.3 softmax units for multinoulli output distributions. *Deep learning*, 180, 2016.

Nate Gruver, Marc Anton Finzi, Micah Goldblum, and Andrew Gordon Wilson. The lie derivative for measuring learned equivariance. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=JL7Va5Vy15J.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pp. 1321–1330. PMLR, 2017.

Brian C Hall. *Lie groups, Lie algebras, and representations*. Springer, 2013.

Jonathan Heek, Anselm Levskaya, Avital Oliver, Marvin Ritter, Bertrand Rondepierre, Andreas Steiner, and Marc van Zee. Flax: A neural network library and ecosystem for JAX, 2024. URL http://github.com/google/flax.

Christopher Hirata and Uroš Seljak. Shear calibration biases in weak-lensing surveys. *Monthly Notices of the Royal Astronomical Society*, 343(2):459–480, 2003.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *The collected works of Wassily Hoeffding*, pp. 409–426, 1994.

Elyssa Hofgard, Rui Wang, Robin Walters, and Tess Smidt. Relaxed equivariant graph neural networks. *arXiv preprint arXiv:2407.20471*, 2024.

Haojie Huang, Dian Wang, Xupeng Zhu, Robin Walters, and Robert Platt. Edge grasp network: A graph-based se (3)-invariant approach to grasp detection. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 3882–3888. IEEE, 2023.

Haojie Huang, Owen Howell, Dian Wang, Xupeng Zhu, Robin Walters, and Robert Platt. Fourier transporter: Bi-equivariant robotic manipulation in 3d. *arXiv preprint arXiv:2401.12046*, 2024a.

Haojie Huang, Dian Wang, Arsh Tangri, Robin Walters, and Robert Platt. Leveraging symmetries in pick and place. *The International Journal of Robotics Research*, 43(4):550–571, 2024b.

Eyke Hüllermeier and Willem Waegeman. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine learning*, 110(3):457–506, 2021.

Frederik Boe Hüttel, Filipe Rodrigues, and Francisco Câmara Pereira. Deep evidential learning for bayesian quantile regression. *arXiv preprint arXiv:2308.10650*, 2023.

Mike Jarvis, GM Bernstein, A Amon, C Davis, PF Léget, K Bechtol, I Harrison, M Gatti, A Roodman, C Chang, et al. Dark energy survey year 3 results: point spread function modelling. *Monthly Notices of the Royal Astronomical Society*, 501(1):1282–1299, 2021.

Edwin T Jaynes. Prior probabilities. *IEEE Transactions on systems science and cybernetics*, 4(3):227–241, 1968.

Mingxi Jia, Dian Wang, Guanang Su, David Klee, Xupeng Zhu, Robin Walters, and Robert Platt. Seil: Simulation-augmented equivariant imitation learning. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1845–1851. IEEE, 2023.

Mira Jürgens, Nis Meinert, Viktor Bengs, Eyke Hüllermeier, and Willem Waegeman. Is epistemic uncertainty faithfully represented by evidential deep learning methods? *arXiv preprint arXiv:2402.09056*, 2024.

Dmitry Kalashnikov, Alex Irpan, Peter Pastor, Julian Ibarz, Alexander Herzog, Eric Jang, Deirdre Quillen, Ethan Holly, Mrinal Kalakrishnan, Vincent Vanhoucke, et al. Qt-opt: scalable deep reinforcement learning for vision-based robotic manipulation. corr abs/1806.10293 (2018). *arXiv preprint arXiv:1806.10293*, 2018.

Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. *Advances in neural information processing systems*, 34:852–863, 2021.

David Klee, Jung Yeon Park, Robert Platt, and Robin Walters. A comparison of equivariant vision models with imagenet pre-training. In *NeurIPS 2023 Workshop on Symmetry and Geometry in Neural Representations*, 2023.

Ivan Kobyzev, Simon JD Prince, and Marcus A Brubaker. Normalizing flows: An introduction and review of current methods. *IEEE transactions on pattern analysis and machine intelligence*, 43(11):3964–3979, 2020.

Risi Kondor, Zhen Lin, and Shubhendu Trivedi. Clebsch-gordan nets: a fully fourier space spherical convolutional neural network, 2018. URL https://arxiv.org/abs/1806.09231.

Igor Kononenko. Bayesian neural networks. *Biological Cybernetics*, 61(5):361–370, 1989.

Hannah Lawrence. Barron's theorem for equivariant networks. In *NeurIPS 2022 Workshop on Symmetry and Geometry in Neural Representations*, 2022.

Dan Levi, Liran Gispan, Niv Giladi, and Ethan Fetaya. Evaluating and calibrating uncertainty prediction in regression tasks. *Sensors*, 22(15):5540, 2022a.

Dan Levi, Liran Gispan, Niv Giladi, and Ethan Fetaya. Evaluating and calibrating uncertainty prediction in regression tasks. *Sensors*, 22(15), 2022b. ISSN 1424-8220. doi: 10.3390/s22155540. URL https://www.mdpi.com/1424-8220/22/15/5540.

Yi-Lun Liao and Tess Smidt. Equiformer: Equivariant graph attention transformer for 3d atomistic graphs. *arXiv preprint arXiv:2206.11990*, 2022.

Yi-Lun Liao, Brandon Wood, Abhishek Das, and Tess Smidt. Equiformerv2: Improved equivariant transformer for scaling to higher-degree representations. *arXiv preprint arXiv:2306.12059*, 2023.

Tobias Liaudat, Jean-Luc Starck, Martin Kilbinger, and Pierre-Antoine Frugier. Rethinking data-driven point spread function modeling with a differentiable optical model. *Inverse Problems*, 39(3):035008, 2023.

Rachel Mandelbaum, Christopher M Hirata, Uroš Seljak, Jacek Guzik, Nikhil Padmanabhan, Cullen Blake, Michael R Blanton, Robert Lupton, and Jonathan Brinkmann. Systematic errors in weak lensing: application to sdss galaxy-galaxy weak lensing. *Monthly Notices of the Royal Astronomical Society*, 361(4): 1287–1322, 2005.

Haggai Maron, Ethan Fetaya, Nimrod Segol, and Yaron Lipman. On the universality of invariant networks. In *International conference on machine learning*, pp. 4363–4371. PMLR, 2019.

Haggai Maron, Or Litany, Gal Chechik, and Ethan Fetaya. On learning sets of symmetric elements. In *International conference on machine learning*, pp. 6734–6744. PMLR, 2020.

Jacqueline McCleary, I Dell'Antonio, and P Huwe. Mass substructure in abell 3128. *The Astrophysical Journal*, 805(1):40, 2015.

Jacqueline McCleary, Ian dell'Antonio, and Anja von der Linden. Dark matter distribution of four low-z clusters of galaxies. *The Astrophysical Journal*, 893(1):8, 2020.

Daniel McNeela. Almost equivariance via lie algebra convolutions. *arXiv preprint arXiv:2310.13164*, 2023.

Kevin Michalewicz, Martin Millon, Frédéric Dux, and Frédéric Courbin. Starred: a two-channel deconvolution method with starlet regularization. *Journal of Open Source Software*, 8(85):5340, 2023. doi: 10.21105/joss.05340. URL https://doi.org/10.21105/joss.05340.

Thomas W Mitchel, Michael Taylor, and Vincent Sitzmann. Neural isometries: Taming transformations for equivariant ml. *arXiv preprint arXiv:2405.19296*, 2024.

Rebecca Nevin, Aleksandra Ćiprijanović, and Brian D Nord. Deepuq: Assessing the aleatoric uncertainties from two deep learning methods. *arXiv preprint arXiv:2411.08587*, 2024.

David A Nix and Andreas S Weigend. Estimating the mean and variance of the target probability distribution. In *Proceedings of 1994 ieee international conference on neural networks (ICNN'94)*, volume 1, pp. 55–60. IEEE, 1994.

Jeremy Nixon, Michael W Dusenberry, Linchuan Zhang, Ghassen Jerfel, and Dustin Tran. Measuring calibration in deep learning. In *CVPR workshops*, volume 2, 2019.

Ian Osband, Zheng Wen, Seyed Mohammad Asghari, Vikranth Dwaracherla, Morteza Ibrahimi, Xiuyuan Lu, and Benjamin Van Roy. Epistemic neural networks. *Advances in Neural Information Processing Systems*, 36:2795–2823, 2023.

Glenn Palmer, Siqi Du, Alexander Politowicz, Joshua Paul Emory, Xiyu Yang, Anupraas Gautam, Grishma Gupta, Zhelong Li, Ryan Jacobs, and Dane Morgan. Calibration after bootstrap for accurate uncertainty quantification in regression models. *npj Computational Materials*, 8(1):115, 2022.

Sneh Pandya, Purvik Patel, Jonathan Blazek, et al. E (2) equivariant neural networks for robust galaxy morphology classification. *arXiv preprint arXiv:2311.01500*, 2023.

Sneh Pandya, Purvik Patel, Brian D. Nord, Mike Walmsley, and Aleksandra Ćiprijanović. Sidda: Sinkhorn dynamic domain adaptation for image classification with equivariant neural networks, 2025. URL https://arxiv.org/abs/2501.14048.

George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *Journal of Machine Learning Research*, 22(57):1–64, 2021.

Jung Yeon Park, Sujay Bhatt, Sihan Zeng, Lawson LS Wong, Alec Koppel, Sumitra Ganesh, and Robin Walters. Approximate equivariance in reinforcement learning. *arXiv preprint arXiv:2411.04225*, 2024.

Pascal Pernot. Properties of the ence and other mad-based calibration metrics. *arXiv preprint arXiv:2305.11905*, 2023.

Marshall D Perrin, Anand Sivaramakrishnan, Charles-Philippe Lajoie, Erin Elliott, Laurent Pueyo, Swara Ravindranath, and Loïc Albert. Updated point spread function simulations for jwst with webbpsf. In *Space telescopes and instrumentation 2014: optical, infrared, and millimeter wave*, volume 9143, pp. 1174–1184. SPIE, 2014.

Mircea Petrache and Shubhendu Trivedi. Approximation-generalization trade-offs under (approximate) group equivariance. *Advances in Neural Information Processing Systems*, 36:61936–61959, 2023.

Ava Polzin. spike: A tool to drizzle hst, jwst, and roman psfs for improved analyses, 2025. URL https://arxiv.org/abs/2503.02288.

Raghunathan Ramakrishnan, Pavlo O Dral, Matthias Rupp, and O Anatole Von Lilienfeld. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific data*, 1(1):1–7, 2014.

David W Romero and Suhas Lohit. Learning partial equivariances from data. *Advances in Neural Information Processing Systems*, 35:36466–36478, 2022.

Barbara Ryden. *Introduction to cosmology*. Cambridge University Press, 2016.

Ashwin Samudre, Mircea Petrache, Brian D Nord, and Shubhendu Trivedi. Symmetry-based structured matrices for efficient approximately equivariant networks. *arXiv preprint arXiv:2409.11772*, 2024.

Akiyoshi Sannai, Masaaki Imaizumi, and Makoto Kawano. Improved generalization bounds of group invariant/equivariant deep networks via quotient feature spaces. In *Uncertainty in artificial intelligence*, pp. 771–780. PMLR, 2021.

Gabriele Scalia, Colin A Grambow, Barbara Pernici, Yi-Pei Li, and William H Green. Evaluating scalable uncertainty estimation methods for deep learning-based molecular property prediction. *Journal of chemical information and modeling*, 60(6):2697–2717, 2020.

Diana Scognamiglio. Exploring Cosmic Matter with Weak Gravitational Lensing in COSMOS-Web. *Bulletin of the AAS*, 56(2), feb 7 2024. https://baas.aas.org/pub/2024n2i149p06.

Maximilian Seitzer, Arash Tavakoli, Dimitrije Antic, and Georg Martius. On the pitfalls of heteroscedastic uncertainty estimation with probabilistic neural networks. *arXiv preprint arXiv:2203.09168*, 2022.

Yu Shi, Shuxin Zheng, Guolin Ke, Yifei Shen, Jiacheng You, Jiyan He, Shengjie Luo, Chang Liu, Di He, and Tie-Yan Liu. Benchmarking graphormer on large-scale molecular modeling datasets. *arXiv preprint arXiv:2203.04810*, 2022.

Sophia Huiwen Sun, Robin Walters, Jinxi Li, and Rose Yu. Probabilistic symmetry for multi-agent dynamics. In *Learning for Dynamics and Control Conference*, pp. 1231–1244. PMLR, 2023.

Richard Sutton. The bitter lesson. *Incomplete Ideas (blog)*, 13(1):38, 2019.

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.

Marcin Tomczak, Siddharth Swaroop, and Richard Turner. Efficient low rank gaussian variational inference for neural networks. *Advances in Neural Information Processing Systems*, 33:4610–4622, 2020.

Dennis Ulmer, Christian Hardmeier, and Jes Frellsen. Prior and posterior networks: A survey on evidential deep learning methods for uncertainty estimation. *arXiv preprint arXiv:2110.03051*, 2021.

Matias Valdenegro-Toro and Daniel Saromo Mori. A deeper look into aleatoric and epistemic uncertainty disentanglement. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1508–1516. IEEE, 2022.

Tycho van der Ouderaa, David W Romero, and Mark van der Wilk. Relaxing equivariance constraints with non-stationary continuous filters. *Advances in Neural Information Processing Systems*, 35:33818–33830, 2022.

Lars Veefkind and Gabriele Cesa. A probabilistic approach to learning the degree of equivariance in steerable cnns. *arXiv preprint arXiv:2406.03946*, 2024.

Mike Walmsley, Micah Bowles, Anna MM Scaife, Jason Shingirai Makechemu, Alexander J Gordon, Annette Ferguson, Robert G Mann, James Pearson, Jürgen J Popp, Jo Bovy, et al. Scaling laws for galaxy images. *arXiv preprint arXiv:2404.02973*, 2024.

Dian Wang, Mingxi Jia, Xupeng Zhu, Robin Walters, and Robert Platt. On-robot learning with equivariant models. *arXiv preprint arXiv:2203.04923*, 2022a.

Dian Wang, Robin Walters, Xupeng Zhu, and Robert Platt. Equivariant q learning in spatial action spaces. In *Conference on Robot Learning*, pp. 1713–1723. PMLR, 2022b.

Dian Wang, Jung Yeon Park, Neel Sortur, Lawson L.S. Wong, Robin Walters, and Robert Platt. The surprising effectiveness of equivariant models in domains with latent symmetry. In *The Eleventh International Conference on Learning Representations*, 2023a. URL https://openreview.net/forum?id=P4MUGRM4Acu.

Dian Wang, Xupeng Zhu, Jung Yeon Park, Mingxi Jia, Guanang Su, Robert Platt, and Robin Walters. A general theory of correct, incorrect, and extrinsic equivariance. *Advances in Neural Information Processing Systems*, 36, 2024.

Rui Wang, Robin Walters, and Rose Yu. Approximately equivariant networks for imperfectly symmetric dynamics. In *International Conference on Machine Learning*, pp. 23078–23091. PMLR, 2022c.

Yuyang Wang, Ahmed AA Elhag, Navdeep Jaitly, Joshua M Susskind, and Miguel Ángel Bautista. Swallowing the bitter pill: Simplified scalable conformer generation. In *Forty-first International Conference on Machine Learning*, 2023b.

Maurice Weiler and Gabriele Cesa. General e (2)-equivariant steerable cnns. *Advances in neural information processing systems*, 32, 2019.

Maurice Weiler, Mario Geiger, Max Welling, Wouter Boomsma, and Taco Cohen. 3d steerable cnns: Learning rotationally equivariant features in volumetric data, 2018. URL https://arxiv.org/abs/1807.02547.

David Weininger. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of chemical information and computer sciences*, 28(1):31–36, 1988.

Lisa Wimmer, Yusuf Sale, Paul Hofman, Bernd Bischl, and Eyke Hüllermeier. Quantifying aleatoric and epistemic uncertainty in machine learning: Are conditional entropy and mutual information appropriate measures? In *Uncertainty in Artificial Intelligence*, pp. 2282–2292. PMLR, 2023.

Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? *arXiv preprint arXiv:1810.00826*, 2018.

Dmitry Yarotsky. Universal approximations of invariant maps by neural networks. *Constructive Approximation*, 55(1):407–474, 2022.

Richard Zhang. Making convolutional networks shift-invariant again. In *International conference on machine learning*, pp. 7324–7334. PMLR, 2019.

Zihan Zou, Yujin Zhang, Lijun Liang, Mingzhi Wei, Jiancai Leng, Jun Jiang, Yi Luo, and Wei Hu. A deep learning model for predicting selected organic molecular spectra. *Nature Computational Science*, 3(11): 957–964, 2023.

A Equivariance

In this section, we give precise definitions of equivariance and invariance, following §A.2 in Brandstetter et al. (2021). For a general review of abstract algebra and representation theory, we direct the reader towards Artin (1998); Esteves (2020); Hall (2013).

Invariance and Equivariance. Equivariance is a property of an operator $\phi : \mathcal{X} \to \mathcal{Y}$ that maps between input and output vector spaces $\mathcal{X}$ and $\mathcal{Y}$. Given a group G and its representations $\rho^{\mathcal{X}}$ and $\rho^{\mathcal{Y}}$ which transform vectors in $\mathcal{X}$ and $\mathcal{Y}$ respectively, an *operator $\phi : \mathcal{X} \to \mathcal{Y}$ is said to be equivariant if it satisfies the following constraint*

$$\rho^{\mathcal{Y}}(g)[\phi(x)] = \phi(\rho^{\mathcal{X}}(g)[x]) \ , \text{ for all } g \in G, x \in \mathcal{X} \ . \tag{57}$$

Invariance is a special case of equivariance in which $\rho^{\mathcal{Y}} = \mathcal{I}^{\mathcal{Y}}$ for all $g \in G$. I.e., *an operator $\phi : \mathcal{X} \to \mathcal{Y}$ is said to be invariant if it satisfies the following constraint*

$$\phi(x) = \phi(\rho^{\mathcal{X}}(g)[x]) \ , \text{ for all } g \in G, x \in \mathcal{X} \ . \tag{58}$$

Thus, with an invariant operator, the output of ϕ is unaffected by transformations applied to the input.

B Evidential Regression

The appeal of (deep) evidential regression for our work is that it allows us to precisely define epistemic and aleatoric uncertainties. Here, we review the relevant theory and notation from Amini et al. (2020).

Given $(y_1, \ldots, y_n) \sim \mathcal{N}(\mu, \sigma^2)$, we may impose priors

$$\mu \sim \mathcal{N}(\gamma, \sigma^2 \nu^{-1}) \tag{59}$$
$$\sigma^2 \sim \Gamma^{-1}(\alpha, \beta) \tag{60}$$

where $\Gamma(\cdot)$ is the gamma function. Let $m = (\gamma, \nu, \alpha, \beta)$, and $\gamma \in \mathbb{R}$, $\nu > 0$, $\alpha > 1$, $\beta > 0$. One can then show that $p(y_i|m) = St(y_i; \gamma, \frac{\beta(1+\nu)}{\alpha\nu}, 2\alpha)$, where the St distribution given by

$$St(t; \mu, \sigma, \nu) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\pi\nu}\sigma\Gamma\left(\frac{\nu}{2}\right)} \left(1 + \frac{1}{\nu}\left(\frac{t-\mu}{\sigma}\right)^2\right)^{-(\nu+1)/2} . \tag{61}$$

Parameterizing the St distribution as a four parameter family is useful because it allows us to define our prediction, aleatoric uncertainty, and epistemic uncertainty in a rigorous way:

$$\mathbb{E}[\mu] = \gamma \quad \text{(Prediction)} \tag{62}$$
$$\mathbb{E}[\sigma^2] = \frac{\beta}{\alpha - 1} \quad \text{(Aleatoric Uncertainty)} \tag{63}$$
$$\text{Var}[\mu] = \frac{\beta}{\nu(\alpha - 1)} \quad \text{(Epistemic Uncertainty).} \tag{64}$$

With some slight abuse of notation, we may abbreivate these as $\sigma^2_{\text{aleatoric}} := \mathbb{E}[\sigma^2]$ and $\sigma^2_{\text{epistemic}} := \text{Var}[\mu]$. Some works take these uncertainties to be addititive. That is, the model will predict $\gamma \pm \sigma^2_{\text{aleatoric}} \pm \sigma^2_{\text{epistemic}}$. See Freedman et al. (2024) and Berman & McCleary (2025); Berman et al. (2025) for examples where this is done in the sciences. We do not assume these uncertainties to be additive in this work.

C Notation and Main Results from Wang et al. 2024

Given our expressed goal of generalizing the bounds from Wang et al. (2024) and our heavy reliance on their formalism, we review the relevant notation and main results here. We start by restating the main problem statement.

Problem Statement. Consider a function $f : X \to Y$. Let $p : X \to \mathbb{R}$ be the probability density function of the domain X. We assume that there is no distribution shift during testing, i.e., p is always the underlying distribution during training and testing. The goal for a model class $\{h : X \to Y\}$ is to fit the function f by minimizing an error function $\text{err}(h)$. We assume the model class $\{h\}$ is arbitrarily expressive except that it is constrained to be equivariant with respect to a group G. Let $\mathbb{I}$ be an indicator function that equals to 1 if the condition is satisfied and 0 otherwise. In classification, $\text{err}(h)$ is the classification error rate; for regression tasks, the error function is a L2 norm function,

$$\text{err}_{\text{cls}} = \mathbb{E}_{x\sim q}\left[\mathbb{I}(f(x) \neq h(x))\right] \tag{65}$$

$$\text{err}_{\text{reg}} = \mathbb{E}_{x\sim q}\left[\left\|h(x) - f(x)\right\|_2^2\right]. \tag{66}$$

C.1 A taxonomy of equivariance

Definition 2 (**Correct Equivariance, Definition 3.1 in Wang et al. (2024)**). *For all $x \in X$, $g \in G$ where $p(x) > 0$, if $p(gx) > 0$ and $f(gx) = gf(x)$, h has correct equivariance with respect to f.*

Definition 3 (**Incorrect Equivariance, Definition 3.2 in Wang et al. (2024)**). *For all $x \in X$, $g \in G$ where $p(x) > 0$, if $p(gx) > 0$ and $f(gx) \neq gf(x)$, h has incorrect equivariance with respect to f.*

Definition 4 (**Extrinsic Equivariance, Definition 3.3 in Wang et al. (2024)**). *For all $x \in X$, $g \in G$ where $p(x) > 0$, if $p(gx) = 0$, h has extrinsic equivariance with respect to f.*

Definition 5 (**Pointwise Correct Equivariance, Definition 3.5 in Wang et al. (2024)**). *For $g \in G$ and $x \in X$ where $p(x) \neq 0$, if $p(gx) \neq 0$ and $f(gx) = gf(x)$, h has correct equivariance with respect to f at x under transformation g.*

Definition 6 (**Pointwise Incorrect Equivariance, Definition 3.6 in Wang et al. (2024)**). *For $g \in G$ and $x \in X$ where $p(x) \neq 0$, if $p(gx) \neq 0$ and $f(gx) \neq gf(x)$, h has incorrect equivariance with respect to f at x under transformation g.*

Definition 7 (**Pointwise Extrinsic Equivariance, Definition 3.7 in Wang et al. (2024)**). *For $g \in G$ and $x \in X$ where $p(x) \neq 0$, if $p(gx) = 0$, h has extrinsic equivariance with respect to f at x under transformation g.*

C.2 Iterated Integration

Definition 8 (**Definition 4.1 in Wang et al. (2024)**). *Let d be the dimension of a generic orbit of G in X and n the dimension of X. Let ν be the $(n-d)$ dimensional Hausdorff measure in X. A closed subset F of X is called a fundamental domain of G in X if X is the union of conjugates of F, i.e., $X = \cup_{g\in G} gF$, and the intersection of any two conjugates has 0 measure under ν.*

If we assume that $\cup_{g_1\neq g_2}(g_1F \cap g_2F)$ has measure 0 and F and Gx are differentiable manifolds, then we may lift an integral Gx to itself. Denote the identification of the orbit Gx and coset space G/G_x with respect to the stabilizer $G_x = \{g : gx = x\}$ by $a_x : G/G_x \to Gx$. Then we have

$$\int_{Gx} f(z)dz = \int_G f(gx)\alpha(g,x)dg \tag{67}$$

where

$$\alpha(g,x) = \left(\int_{Gx} dh\right)^{-1} \left|\frac{\partial a_x(\bar{g})}{\partial \bar{g}}\right|. \tag{68}$$

C.3 Regression Bounds

Theorem 4 (**Theorem 4.8 in Wang et al. (2024)**). *We assume h is G invariant so that $h(gx) = h(x)$ for all $g \in G$. Assume $Y = \mathbb{R}^n$. Denote by $p(Gx) = \int_{z \in Gx} p(z)dz$ the probability of the orbit Gx. Denote by $q(z) = \frac{p(z)}{p(Gx)}$ the normalized probability density of the orbit Gx such that $\int_{Gx} q(z)dz = 1$. Let $\mathbb{E}_{Gx}[f]$ be the mean of the function f on the orbit Gx defined, and let $\mathbb{V}_{Gx}[f]$ be the variance of f on the orbit Gx,*

$$\mathbb{E}_{Gx}[f] = \int_{Gx} q(z)f(z)dz = \frac{\int_{Gx} p(z)f(z)dz}{\int_{Gx} p(z)dz} \tag{69}$$

$$\mathbb{V}_{Gx}[f] = \int_{Gx} q(x)\|\mathbb{E}_{Gx}[f] - f(z)\|_2^2. \tag{70}$$

We have $err(h) \geq \int_F p(Gx)\mathbb{V}_{Gx}[f]$.

Theorem 5 (**Theorem 4.9 in Wang et al. (2024)**). *We now only assume equivariance on h, that is, $h(\rho_X(g)x) = \rho_Y(g)h(x)$ where $g \in G$, ρ_X and ρ_Y are group representations associated with X and Y. We will denote $\rho_X(g)x$ and $\rho_Y(g)y$ by gx and gy, leaving the representation implicit. Assume $Y = \mathbb{R}^n$. Let Id be the identity. Define a matrix $Q_{Gx} \in \mathbb{R}^{n \times n}$ and $q(gx) \in \mathbb{R}^{n \times n}$ so that $\int_G q(gx)dg = Id$ by*

$$Q_{Gx} = \int_G p(gx)\rho_Y(g)^T\rho_Y(g)\alpha(x,g)dg \tag{71}$$

$$q(gx) = Q_{Gx}^{-1}p(qx)\rho_Y(g)^T\rho_Y(g)\alpha(x,g). \tag{72}$$

If f is equivariant, $g^{-1}f(gx)$ is a constant for all $g \in G$. Define $\mathcal{E}_G[f,x]$

$$\mathcal{E}_G[f,x] = \int_G q(gx)g^{-1}f(gx)dg. \tag{73}$$

The error of h has lower bound $err(h) \geq \int_F \int_G p(gx)\|f(gx) - g\mathcal{E}_G[f,x]\|_2^2\alpha(x,g)dgdx$.

D Mean-Invariant Coverage Problem Setup

In many practical circumstances, we will be working with a model h that has different equivariance constraints for the prediction μ and variance σ^2. Here, we consider a class of models $\{h : X \to \mu \times \sigma^2\}$ where h_μ is $G-$invariant but h_{σ^2} is a constant function. Let $\hat{C}^{1-\alpha}$ be a $1-\alpha$ confidence interval for $1-\alpha \in [0,1]$, where α is our significance level for a given prediction. Our goal for our model is to maximize the expected coverage, where coverage is defined as in Sun et al. (2023):

$$\text{Coverage} = \mathbb{I}\left(\mathbb{P}(f \in \hat{C}^{1-\alpha}) \geq 1-\alpha\right). \tag{74}$$

Accordingly the expected coverage is

$$\mathbb{E}_{x\sim p}\left[\mathbb{I}\left(\mathbb{P}(f(x) \in \hat{C}^{1-\alpha}) \geq 1-\alpha\right)\right] = \int_F \int_{Gx} p(z)\left(\mathbb{I}\left(\mathbb{P}(f(x) \in \hat{C}^{1-\alpha}) \geq 1-\alpha\right)\right)dzdx. \tag{75}$$

The expression $\mathbb{P}(f \in \hat{C}^{1-\alpha})$ can be formulated in terms of the Mahalanobis distance:

$$d^2(h_\mu(x), f(x)) = -\left[h_\mu(x) - f(x)\right]^T \Sigma^{-1}\left[h_\mu(x) - f(x)\right] \tag{76}$$

where $\Sigma^{-1} = \text{diag}(h_{\sigma^2})$. For brevity, we will sometimes just abbreviate d^2 or $d^2(x)$ instead of $d^2(h_\mu(x), f(x))$. Additionally, we will note that the distance is symmetric in that

$$d^2(h_\mu(x), f(x)) = d^2(f(x), h_\mu(x)). \tag{77}$$

It is known that d^2 follows a χ^2 distribution, which has Cumulative Distribution Function describing the continuous probabilty distribution [3] given by

$$\gamma(s, w) = \int_0^w t^{s-1} e^{-t} dt \tag{78}$$

where s is the degrees of freedom. The degrees of freedom s is taken to be $\dim(Y)$. If $\gamma(s, w^*) = 1 - \alpha$, then the $1-\alpha$ confidence interval is given by $\gamma^{-1}(1-\alpha)$. We know that γ is invertible because it is monotonically increasing in w.

Now, we can rewrite coverage via

$$\mathbb{I}\left(\mathbb{P}(f \in \hat{C}^{1-\alpha}) \geq 1 - \alpha\right) = \mathbb{I}(d^2 \leq \gamma^{-1}(1-\alpha)). \tag{79}$$

The expected coverage then becomes

$$\mathbb{E}_{x \sim p}[\text{Coverage}] = \int_F \int_{Gx} p(z) \left[\mathbb{I}(d^2 \leq \gamma^{-1}(1-\alpha))\right] dz dx. \tag{80}$$

One may try to take advantage of invariance to bound the Coverage metric by placing a bound on the distance. However, on a given orbit Gx, the best fit function h_μ is the average of the function f on the orbit defined. Therefore, h_μ is usually equal to f on at least one input x, making the lower bound on d^2 become 0, and the assumption of invariance less than useful.

E Special Case for Invariant Approximation Error

Corollary 3. *A function f is said to be uniformly continuous if for some $\delta > 0$ and for any x, $y \in \mathbb{R}$ we have that $|x - y| < \delta \implies |f(x) - f(y)| < \varepsilon$ for all $\varepsilon > 0$. Let G be a group and let h be a $G-$invariant function. Let $\Pi : [a, b] \to [a, b]$ be a function that rearranges the elements $x \in [a, b]$ so that the orbits Gx are arranged sequentially, let $\widetilde{f}$ be given by $f(\Pi^{-1}(x))$, and assume the following conditions:*

1. *The domain of $\widetilde{f}$ is $\Pi([a, b])$ for some a, $b \in \mathbb{R}$. The co-domain is $\mathbb{R}$.*
2. *$\widetilde{f}$ is continuous on $\Pi([a, b])$.*
3. *For each orbit Gx, $\sup Gx \in Gx$, $\inf Gx \in Gx$.*
4. *For each orbit Gx, $|\sup Gx - \inf Gx| < \delta$.*

It follows that we can choose the $G-$invariant function h such that $|\widetilde{f}(x) - h(x)| < \varepsilon$ for all $x \in \Pi([a, b])$.

Proof. It is known that continuous functions on closed a closed interval are uniformaly continuous. Now, we partition the domain $[a, b]$ into each of the orbits Gx after applying Π. By intermediate value theorem, $\widetilde{f}$ attains its average on each orbit. On each orbit, we choose y to be the real number that corresponds to $\widetilde{f}(y) = \frac{1}{\sup Gx - \inf Gx} \int_{z \in Gx} \widetilde{f}(z) dz$ on the orbit. Therefore, we have that for x, $y \in Gx$, $|x - y| < \delta \implies$

[3] For more on the χ^2 CDF, see Beh (2018)

$|\widetilde{f}(x) - \widetilde{f}(y)| < \varepsilon$, where $|x - y| < \delta$ is true by the fourth assumption. We set $h(x) \equiv \widetilde{f}(y)$ which gives us that $|\widetilde{f}(x) - h(x)| < \varepsilon$ for all $x \in \Pi([a, b])$. This completes the proof.

□

Remark 6. *This proof is a special case of the more general fact that if f is continuous on a closed interval $[a, b]$, then we can choose a piecewise constant function $h_\varepsilon(x)$ such that $|f(x) - h_\varepsilon(x)| < \varepsilon$. The significance of this corollary is that it tells us that if our orbits are sufficiently small in $\mathbb{R}$ then the $G-$invariant regression error can be made arbitrarily small too. If we partition $[a, b]$ into the orbits induced by σ^2, then we may use this to understand if the domain restricted regression error can be made arbitrarily small. We caution that the assumptions for this to work are quite strong. In particular, we assume that f is a function $[a, b] \to \mathbb{R}$. The language of representation theory still applies, since fields are trivially vectors spaces over themselves (e.g. real numbers are vector spaces over themselves), however, we are still heavily constrained by this choice.*

F Vector Regression Setup

Our training of the $E(3)-$equivariant neural network uses e3nn_jax (Geiger et al., 2022; Geiger & Smidt, 2022; Kondor et al., 2018; Weiler et al., 2018; Thomas et al., 2018). The MLP baseline is built entirely with Flax (Heek et al., 2024). The models are trained for a minimum of 10 epochs for a maximum number of 100, with early stopping if the validation loss stops improving after 5 epochs. We train on 2000 generated samples. We train using both a $\beta-$NLL loss (Seitzer et al., 2022) and an MSE loss, equally weighted, with $\beta = 1$. The $\beta-$NLL loss is given by

$$\mathcal{L}_{\beta-NLL} := \mathbb{E}_{X,Y}\left[\lfloor\hat{\sigma}^{2\beta}\rfloor\left(\frac{1}{2}\log\hat{\sigma}^2 + \frac{(Y - \hat{\mu}(X))^2}{2\hat{\sigma}^2}\right) + C\right] \tag{81}$$

where $\lfloor\cdot\rfloor$ represents a stop-gradient. For consistency with the figures, the reported metrics are calculated on the $xy-$coordinates. The MSE and β-NLL scores are average over all vectors and $xy-$coordinates.

G Swiss Roll Experiment Details

The Swiss Roll distributions are created by generating points in polar coordinates using some r as a function of θ. Additionally, the points are given a $z-$coordinate of 0 or 1. An example of a spiral distribution with extrinsic equivariance seen from a $z-$invariant point-of-view is given in Figure 14. See also Figures 7 and 11 in Wang et al. (2024). The correct and incorrect Swiss Roll Distributions are similar. For correct equivariance, the color labels are the same for each spiral at $z = 0$ and $z = 1$. For incorrect, the labels are the opposite. For extrinsic, the spirals do not overlap. For details further, see Wang et al. (2024).

Binning Approximations. We compute ECE using the following binning approximations:

$$\text{acc}(B_m) = \frac{1}{|B_m|}\sum_{i\in B_m}\mathbb{I}(f = h_Y) \tag{82}$$

$$\text{conf}(B_m) = \frac{1}{|B_m|}\sum_{i\in B_m} h_P \tag{83}$$

$$\text{ECE} = \sum_{m=1}^{M}\frac{|B_m|}{n}\Big|\,\text{acc}(B_m) - \text{conf}(B_m)\,\Big|\,. \tag{84}$$

We use 100 bins. We adapt models, data generation, and training materials from Wang et al. (2024) and `https://github.com/pointW/ext_theory/`. The $z-$invariant network is implemented using DSS layers (Maron et al., 2020).

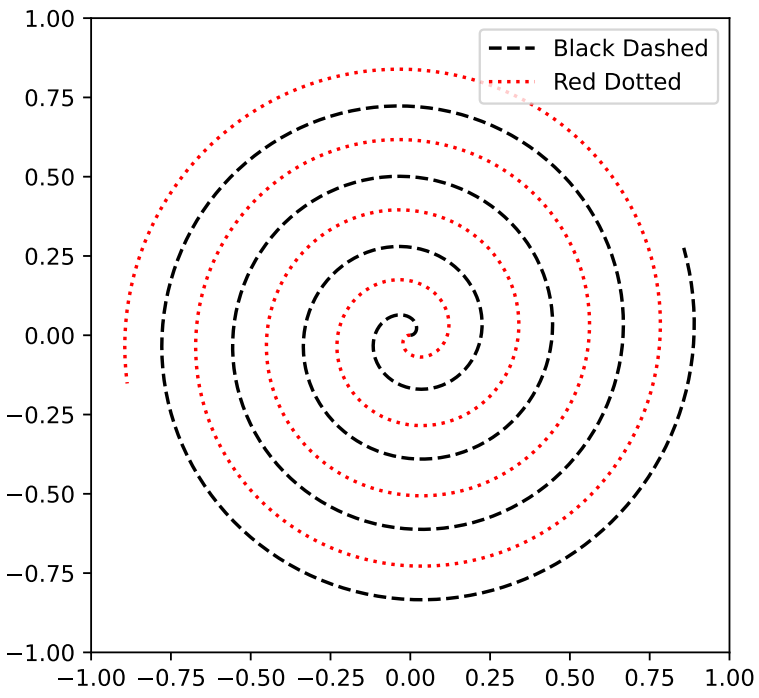

Figure 14: The extrinsic Swiss Roll Distribution seen from a $z-$invariant point of view.

Sample Calibration Approximation Error. Theorem 1 tells us that ECE is bounded on a closed interval (regardless of any assumption of invariance). This allows us to say something about how many samples we need to approximate the true ECE. In particular, since ECE is bounded on $[0, 1]$, we may apply Hoeffding's Inequality (Hoeffding, 1994; 1963). We adopt notation similar to Bousquet et al. (2004); Petrache & Trivedi (2023). We sample x according to $q(x)$ and obtain h_Y and h_p. Define the calibration CE as the term inside the integrand of Equation 1, $|\mathbb{P}(f = h_Y | h_p = p) - p|$. For brevity, we will use $h(x) := (h_Y(x), h_P(x))$. The CE term still uses $h_Y(x)$ and $h_P(x)$ terms as distinct inputs. Imagine we sample n times, giving us the set $\{(x_1, h(x_1)), \ldots, (x_n, h(x_n))\}$. We will abbreviate each i.i.d. $(x_i, h(x_i))$ pair as Z_i.

Corollary 4.

$$\mathbb{P}\left[\left| \frac{1}{n}\sum_{i=1}^{n} CE(Z_i) - ECE(Z) \right| > \varepsilon\right] \leq 2\exp(-2n\varepsilon^2) \tag{85}$$

for all $\varepsilon > 0$.

Proof. Since Theorem 1 tells us that ECE is bounded on $[0, 1]$, the result follows immediately from Hoeffding's Inequality. □

H Galaxy Experiment Details

H.1 Motivation and Implementation of PSF Blurring

Motivation. A point-spread function (PSF) is an impulse response of an optical system to light. PSFs occur all throughout medical and astronomical imaging. The science case we explore in this work is the distortion of galaxy images. With next generation imagers like JWST and large astronomical surveys like

COSMOS-Web (Casey et al., 2023), there are renewed efforts to characterize the effect of the PSF and its effects on downstream scientific analysis (Perrin et al., 2014; Birrer et al., 2021; Jarvis et al., 2021; Michalewicz et al., 2023; Liaudat et al., 2023; Berman et al., 2024; Berman & McCleary, 2024; Feng et al., 2025; Polzin, 2025). Understanding how the PSF harms a model's ability to identify a galaxy's morphology class can hint at the effect of the PSF on measured ellipticity moments (Hirata & Seljak, 2003; Mandelbaum et al., 2005), which is a crucial ingredient for maps of large scale structure (Scognamiglio, 2024). See, for example, McCleary et al. (2015; 2020).

Implementation. The way we implement PSF blurring follows Pandya et al. (2025). Consider an image grid I with values (ζ, ξ) and channels c. PSF blurring with a Gaussian kernel of width ϵ via

$$I_{\text{PSF}}(\zeta, \xi) = (I * G)(\zeta, \xi), \tag{86}$$

where

$$G(\zeta, \xi) = \frac{1}{2\pi\epsilon^2} \exp\left(-\frac{\zeta^2 + \xi^2}{2\epsilon^2}\right). \tag{87}$$

We apply this convolution on each channel c.

H.2 Training and Evaluating

Our models and training scripts are adapted from Pandya et al. (2025) and `https://github.com/deepskies/SIDDA`. The galaxy datasets are initially sourced from `https://zenodo.org/records/14583107`, and there is a script to produce the datasets with PSF blurring in our GitHub artifact. We compute ECE using the same approximations as in Equations 82 - 84. We summarize the number of parameters for each model in Table 2 below:

Model	CNN	C_2	C_4	C_6	C_8	C_{10}	C_{12}
Parameters	$1,188,486$	$1,190,070$	$1,197,750$	$1,205,430$	$1,213,110$	$1,220,790$	$1,228,470$
Model	–	D_2	D_4	D_6	D_8	D_{10}	D_{12}
Parameters	–	$1,197,750$	$1,213,110$	$1,228,470$	$1,243,830$	$1,259,190$	$1,274,550$

Table 2: Number of model parameters for Galaxy CNN and GCNN group order experiment.

For further guidance on how many hidden units are needed to approximate the ground truth as a function of group order, we direct the reader to Theorem 16 in Lawrence (2022).

I Chemical Property Experiment Details

Our experiment for the chemical properties used a modified version of Backenköhler et al. (2023) for the data preprocessing and main training loop. While their analysis uses on feed forward network head for the prediction task, we use four independent feed forward heads that predict the quantities $m = (\gamma, \nu, \alpha, \beta)$. We train with a negative log likelihood loss function with an added regression loss regularizer,

$$\Omega = 2\beta(1+\nu) \tag{88}$$

$$\mathcal{L}_i^{\mathrm{NLL}}(w) = \frac{1}{2}\log\left(\frac{\pi}{\nu}\right) - \alpha\log(\Omega) \tag{89}$$

$$+ \left(\alpha + \frac{1}{2}\right)\log((y_i-\gamma)^2\nu + \Omega) + \log\left(\frac{\Gamma(\alpha)}{\Gamma(\alpha+\frac{1}{2})}\right) \tag{90}$$

$$\mathcal{L}_i^{\mathrm{R}}(w) = |y_i - \mathbb{E}[\mu_i]| \cdot \Phi \tag{91}$$

$$= |y_i - \gamma| \cdot (2\nu + \alpha) \tag{92}$$

$$\mathcal{L}_i(w) = \mathcal{L}_i^{\mathrm{NLL}}(w) + \lambda\mathcal{L}_i^{\mathrm{R}}(w). \tag{93}$$

The GIN model has $52,417$ paramaters and the $E(3)-$invariant model has $51,969$. Through ablation study, we found that training stability is sensitive to a choice of λ, which we choose to be either $\lambda = 0.1$ or $\lambda = 1$. This instability is consistent with §S2.1.3 in Amini et al. (2020). Additionally, we found $z-$scoring the training, validation, and testing sets was necessary for ensuring stability during training for all molecular properties outside of the dipole moment.

Our model for emulating spectral lines is trained in the same way, partially taking inspiration from Zou et al. (2023). We note the following tradeoffs between our approach and DetaNet:

1. Our adoption of the message-passing framework is more general than the attentional one used in their work (Bronstein et al., 2021).
2. DetaNet has arbitrary resolution, relying on a sum of basis functions.
3. DetaNet is trained not to produce the spectral line directly, but to produce the dipole moment, polarizability, and the inter-atomic and atomic hessians, which in turn gives the spectral line.
4. DetaNet can be limited by its usage of the Quantum Harmonic Oscillator approximation in some cases.

We leave further comparison and model development as an opportunity for future work. Other potential baselines could include Equiformer (Liao & Smidt, 2022), EquiformerV2 (Liao et al., 2023), Graphormer (Shi et al., 2022), or Graphormer with data augmentation. The model we use in this work has $36,997,125$ parameters.

J Galaxy Experiment Additional Results

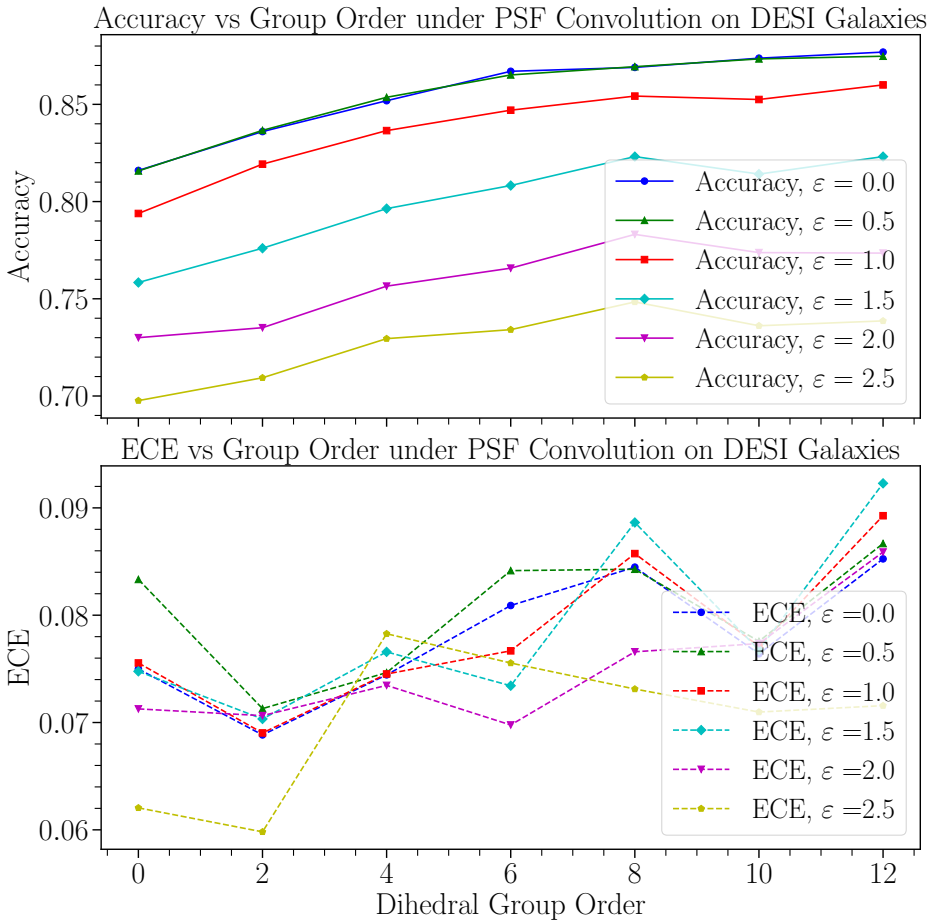

Figure 15: Accuracy and ECE vs Dihedral Group Order under PSF Convolution on DESI Galaxies.

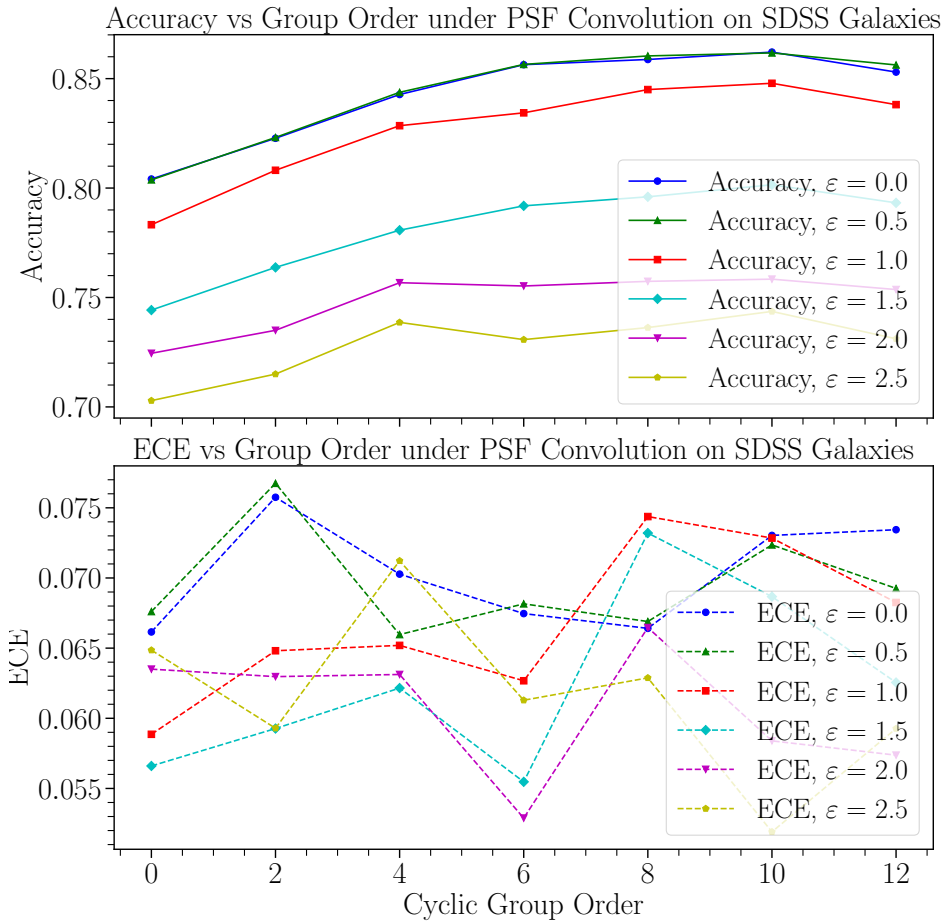

Figure 16: Accuracy and ECE vs Cyclic Group Order under PSF Convolution on SDSS Galaxies.

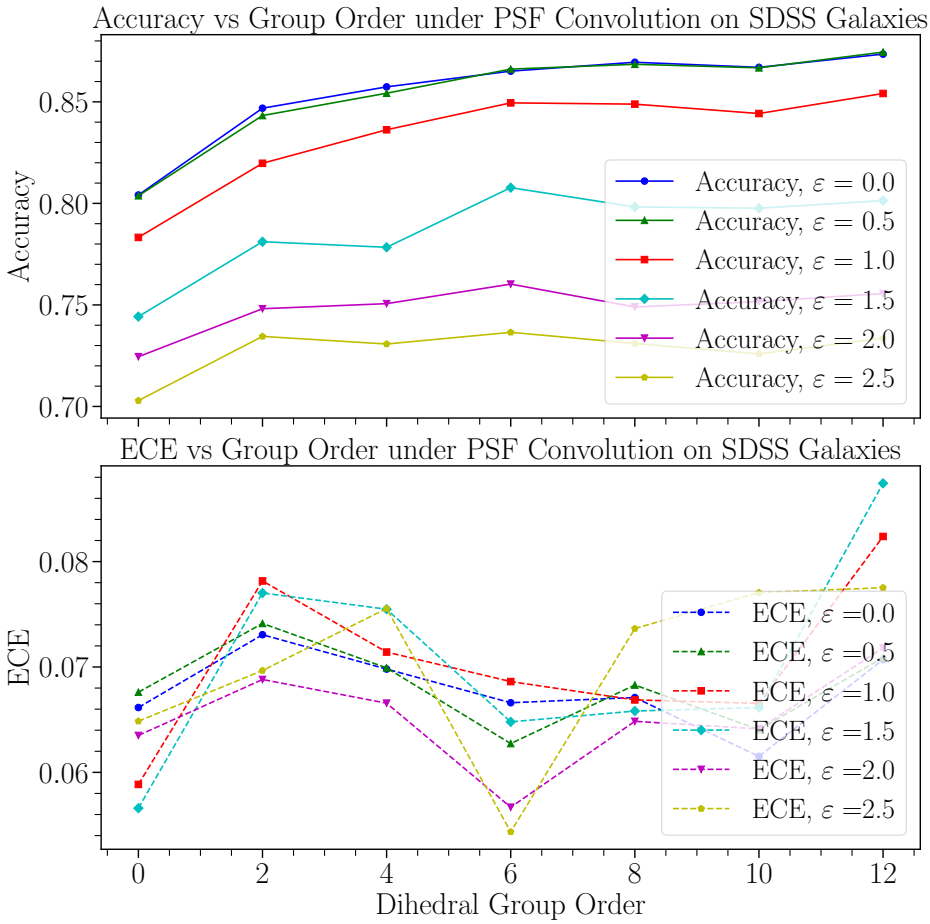

Figure 17: Accuracy and ECE vs Dihedral Group Order under PSF Convolution on SDSS Galaxies.

K Additional Spectra Results

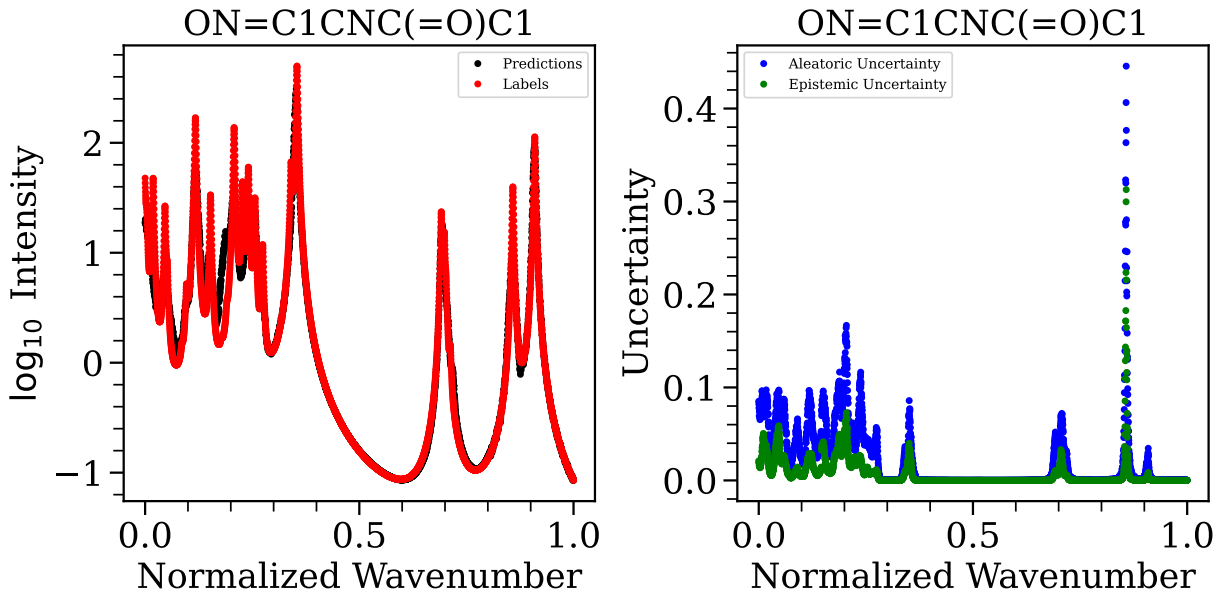

Figure 18: Same as Figure 11 for the molecule given by SMILES string $ON = C1CNC(= O)C1$.

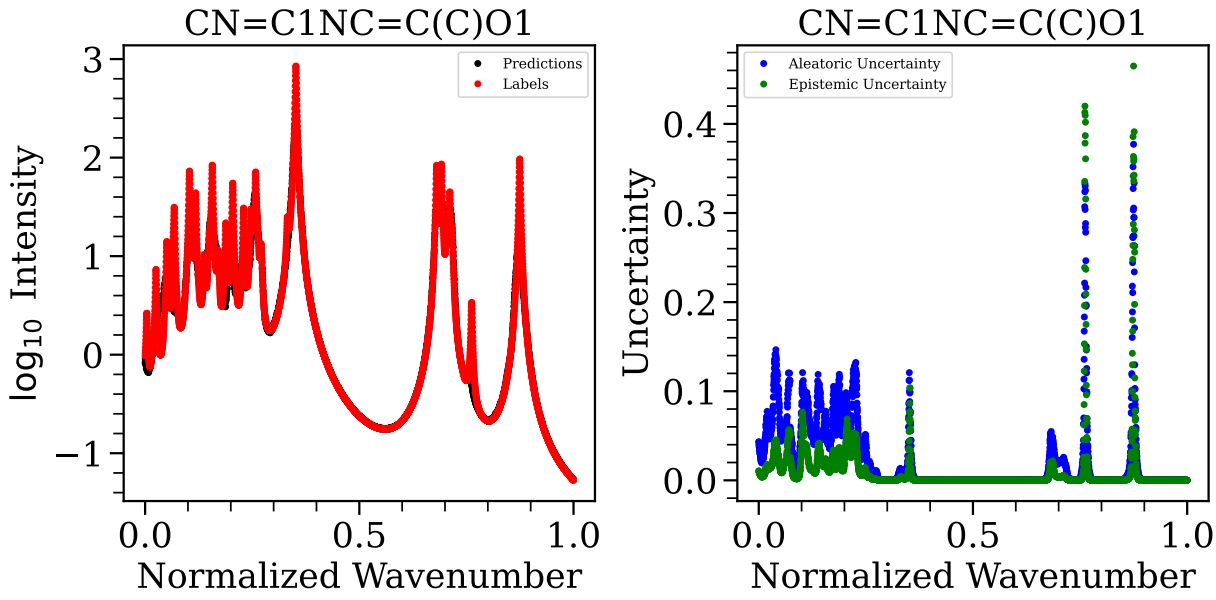

Figure 19: Same as Figure 11 for the molecule given by SMILES string $CN = C1NC = C(C)O1$.

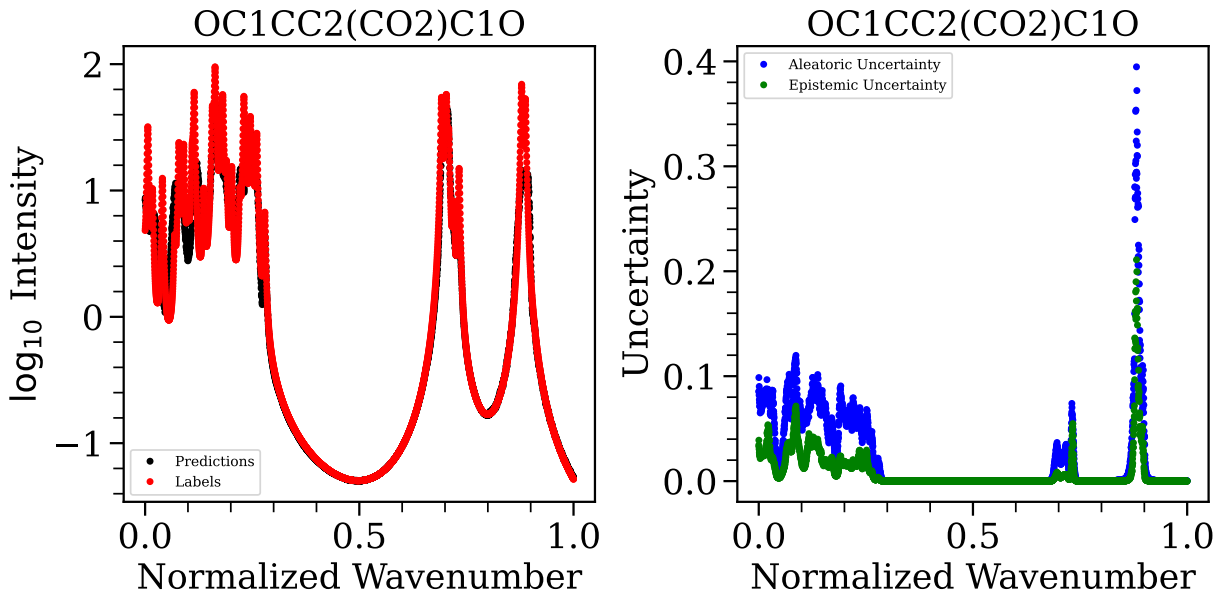

Figure 20: Same as Figure 11 for the molecule given by SMILES string $OC1CC2(CO2)C10$.

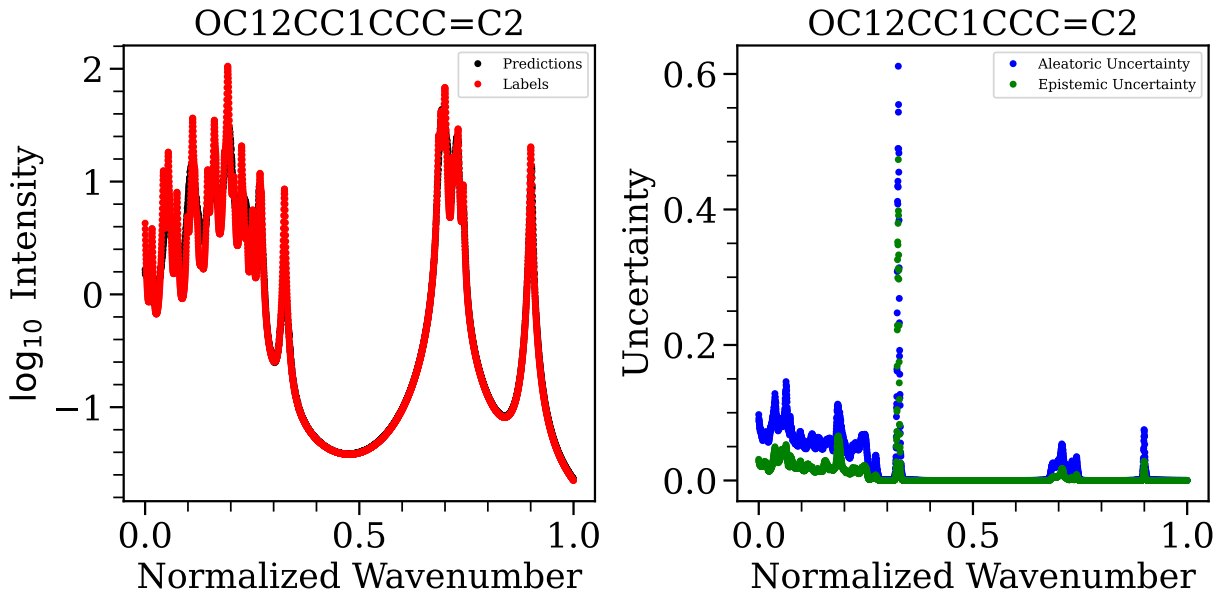

Figure 21: Same as Figure 11 for the molecule given by SMILES string $OC12CC1CCC=C2$.

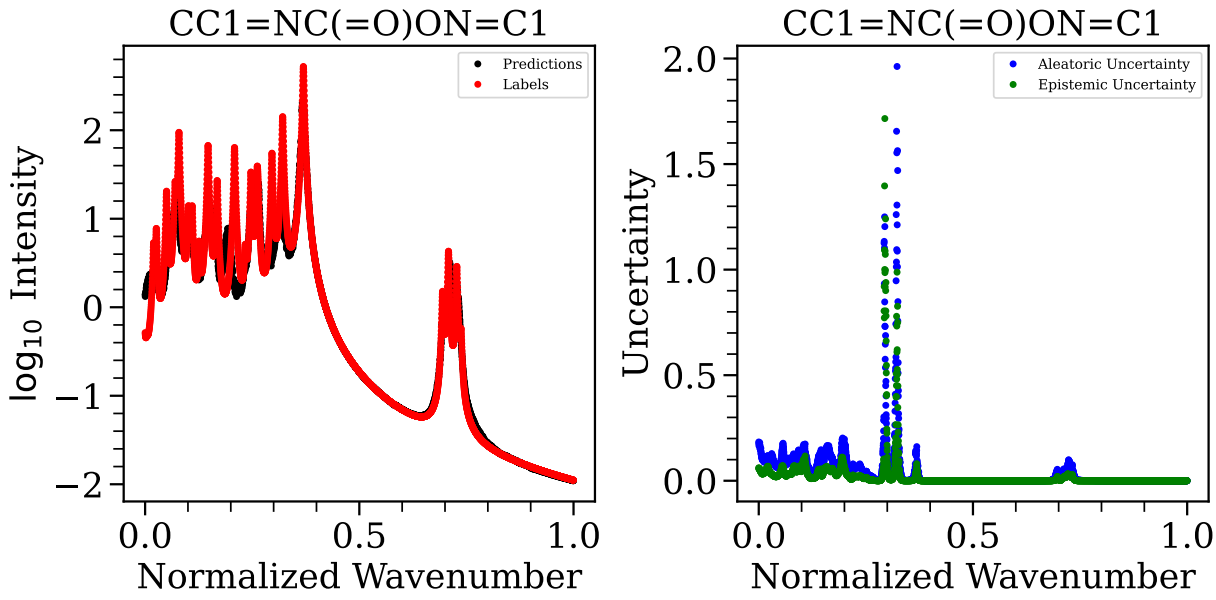

Figure 22: Same as Figure 11 for the molecule given by SMILES string $CC1 = NC(= O)ON = C1$.

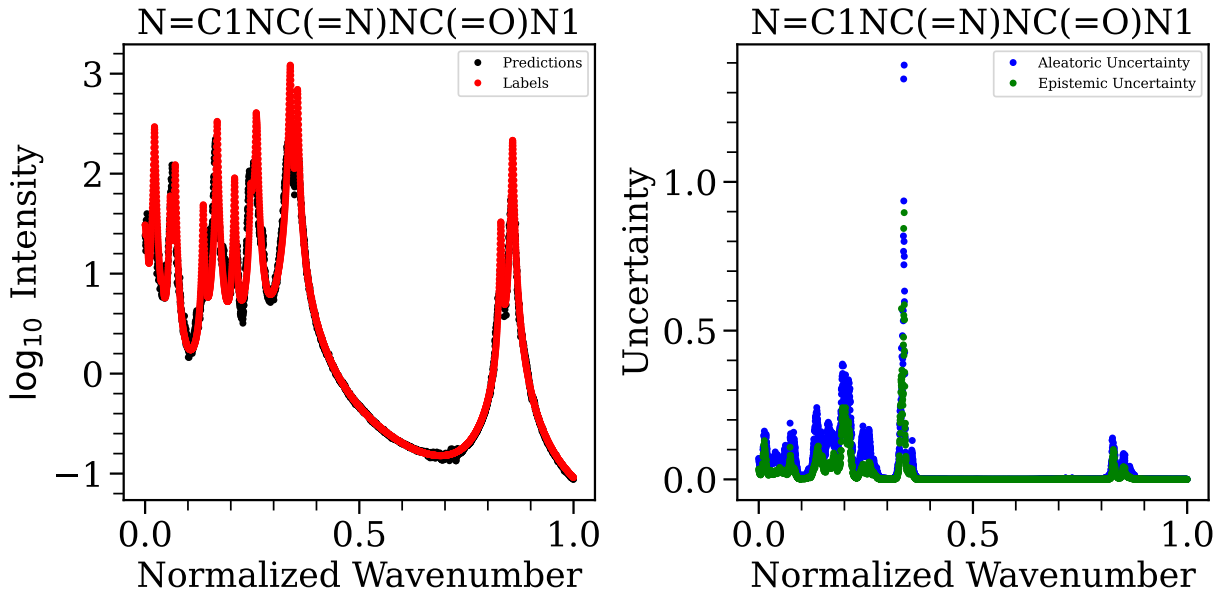

Figure 23: Same as Figure 11 for the molecule given by SMILES string $N = C1NC(= N)NC(= O)N1$.

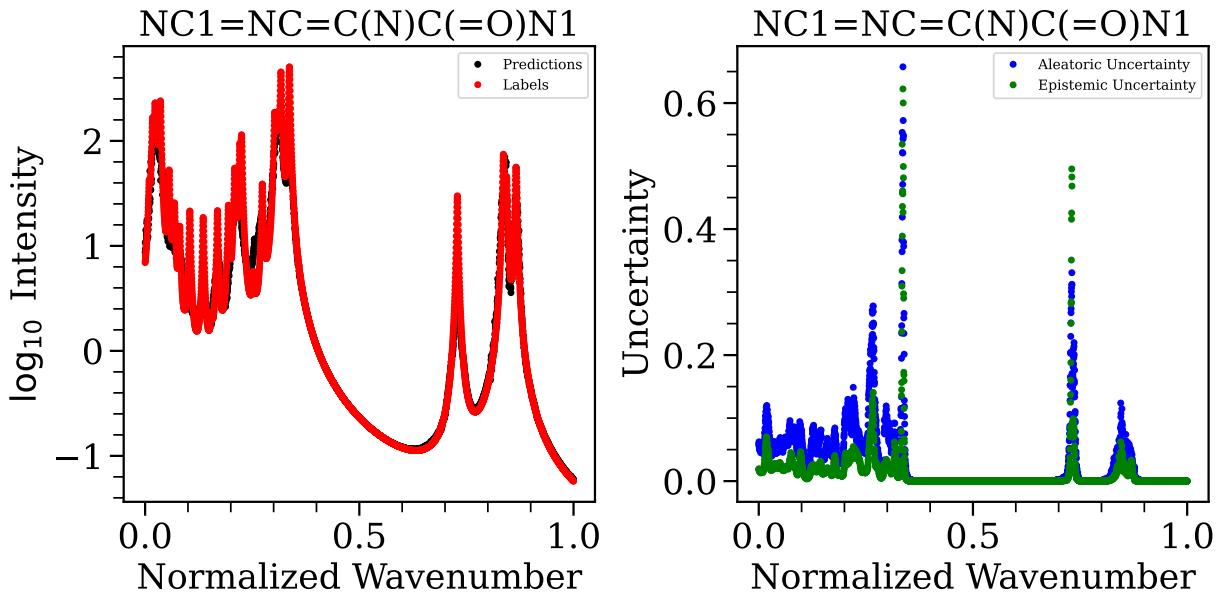

Figure 24: Same as Figure 11 for the molecule given by SMILES string $NC1 = NC = C(N)C(= O)N1$.

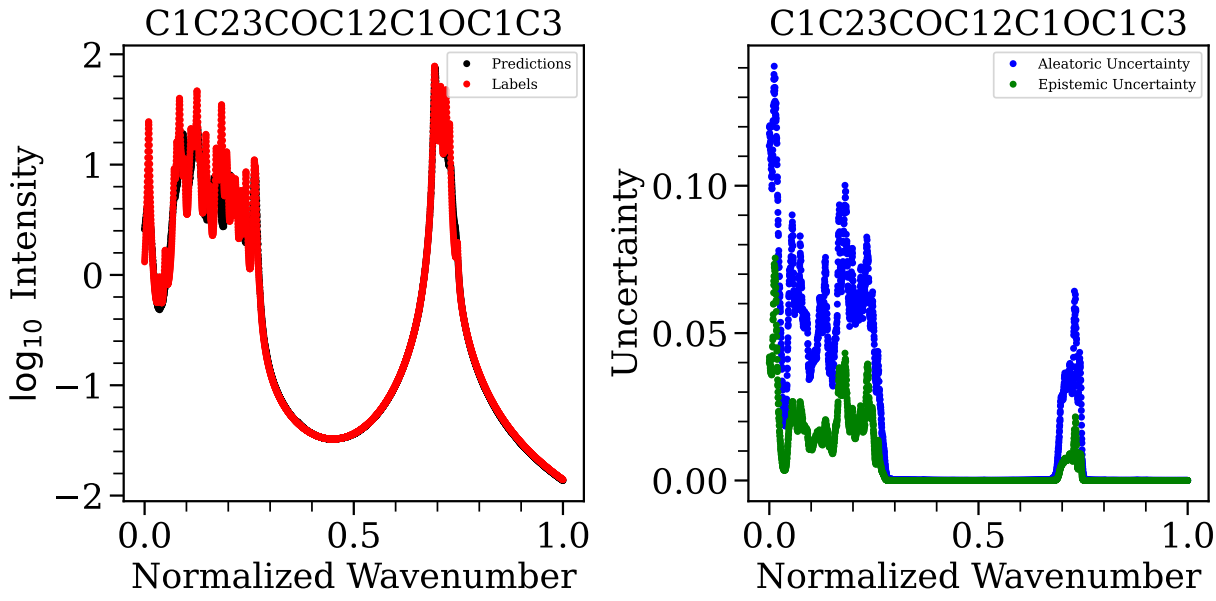

Figure 25: Same as Figure 11 for the molecule given by SMILES string $C1C23COC12C1OC1C3$.

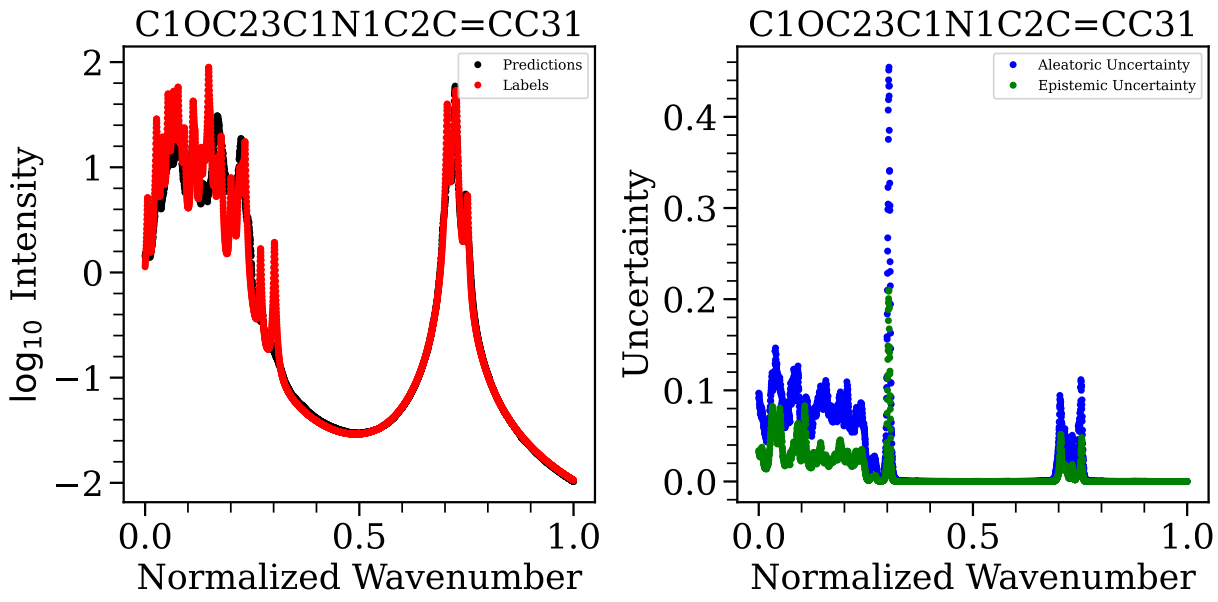

Figure 26: Same as Figure 11 for the molecule given by SMILES string $C1OC23C1N1C2C = CC31$.

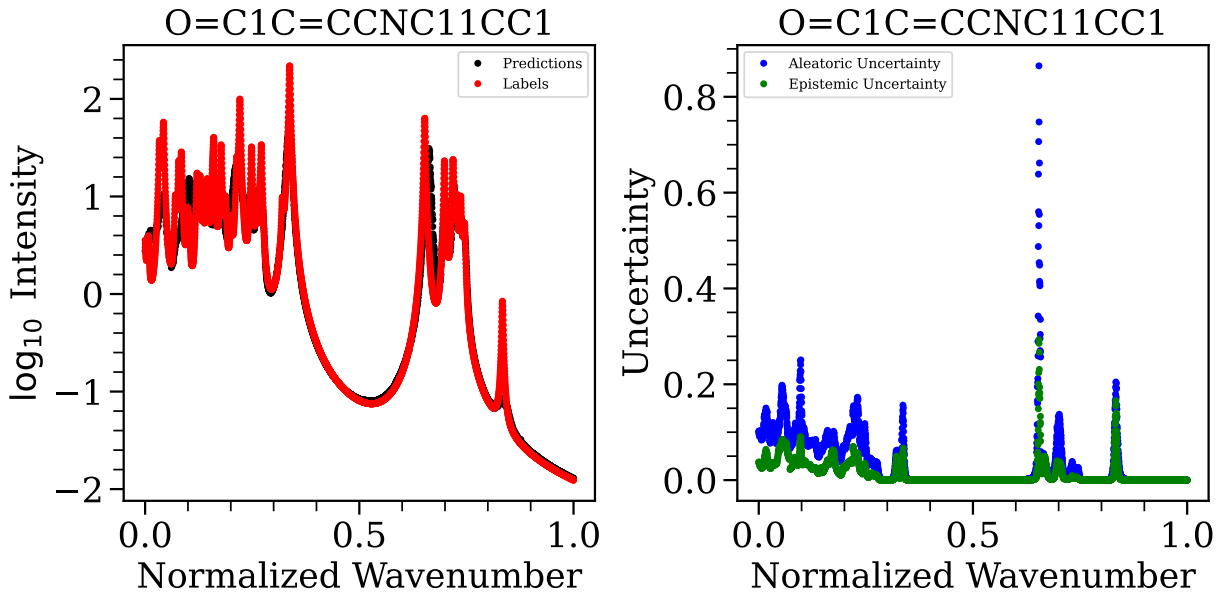

Figure 27: Same as Figure 11 for the molecule given by SMILES string $O = C1C = CCNC11CC1$.

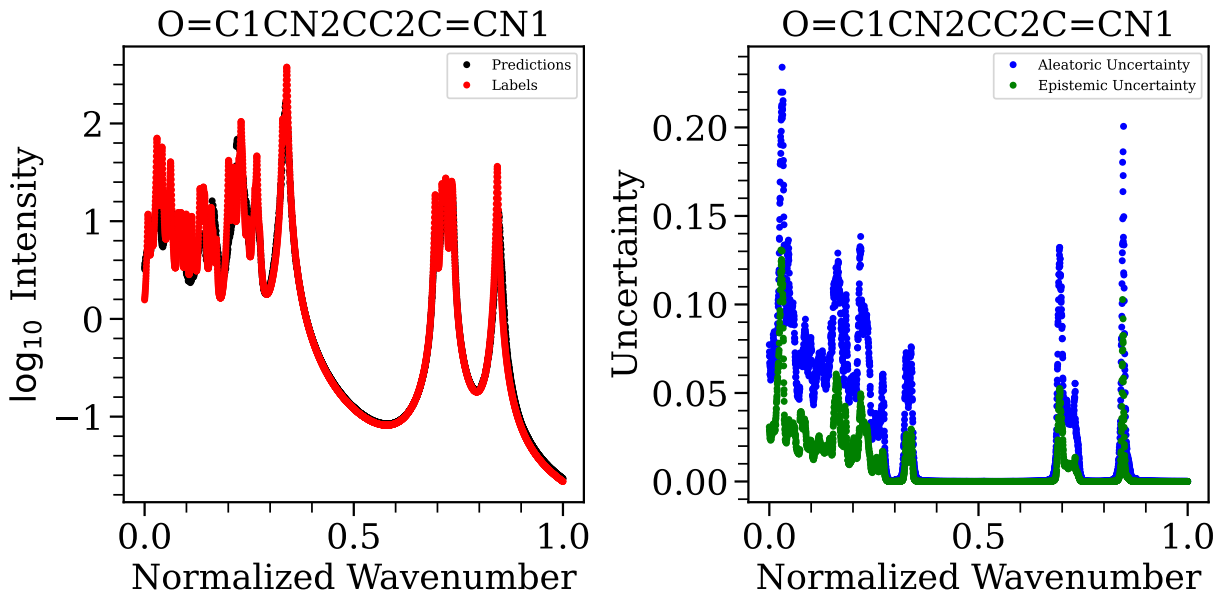

Figure 28: Same as Figure 11 for the molecule given by SMILES string $O = C1CN2CC2C = CN1$.

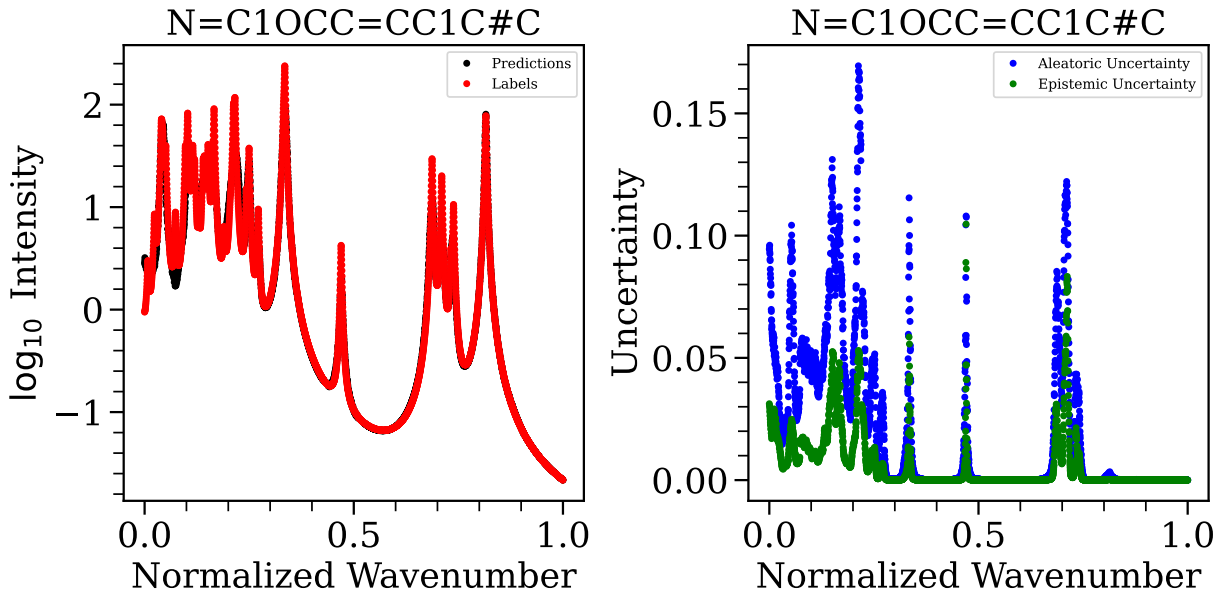

Figure 29: Same as Figure 11 for the molecule given by SMILES string $N = C1OCC = CC1C\#C$.

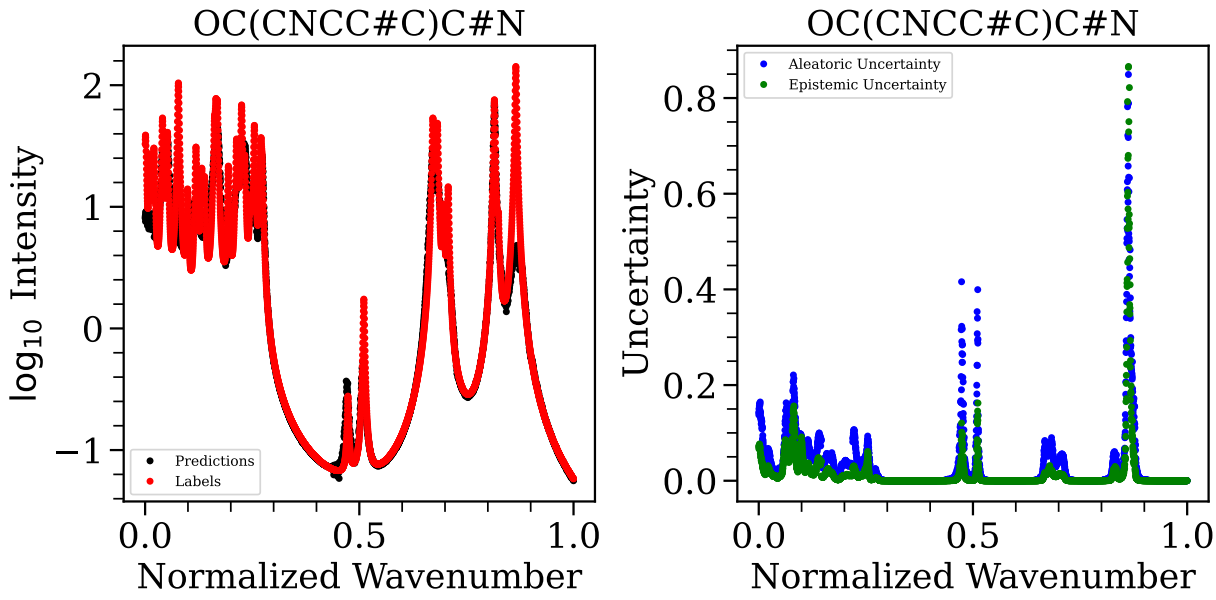

Figure 30: Same as Figure 11 for the molecule given by SMILES string $OC(CNCC\#C)C\#N$.

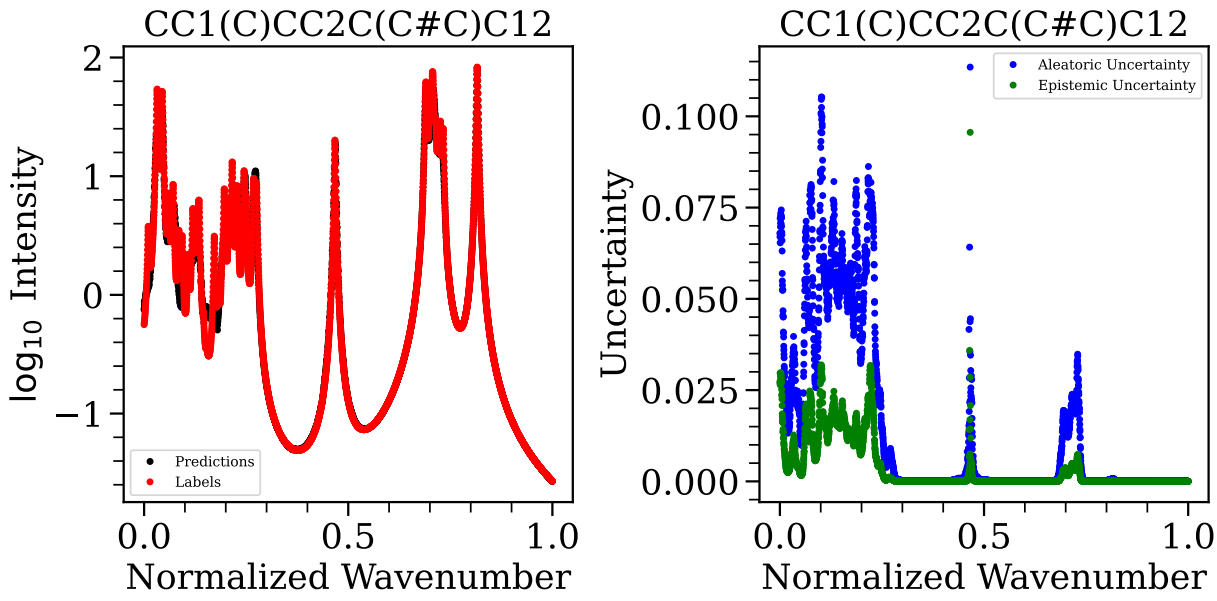

Figure 31: Same as Figure 11 for the molecule given by SMILES string $CC1(C)CC2C(C\#C)C12$.

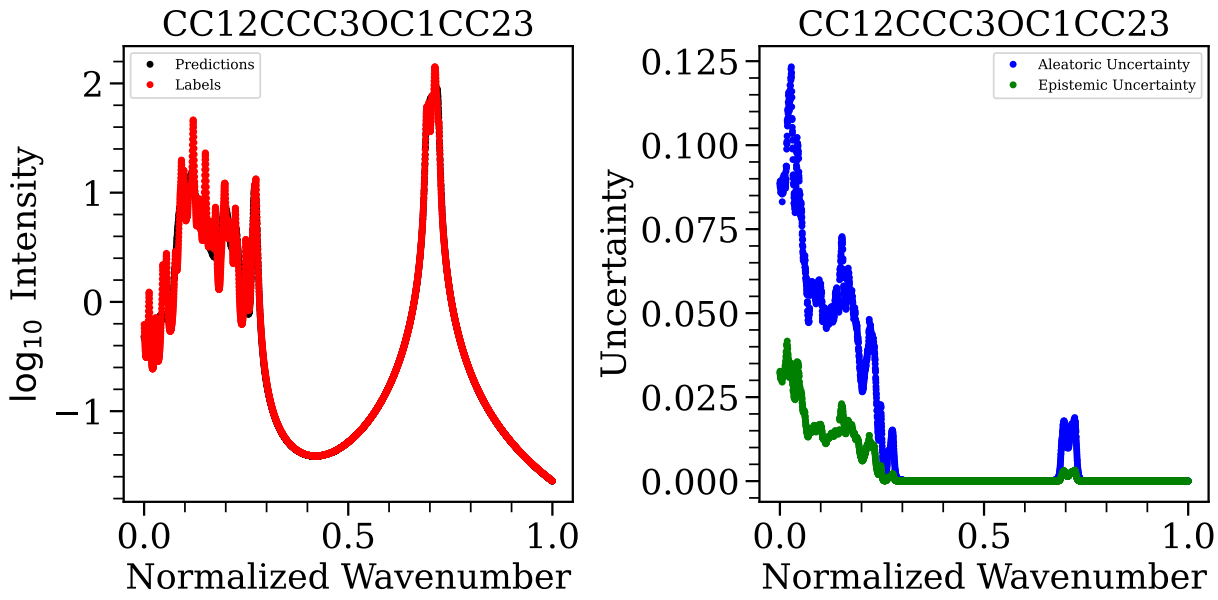

Figure 32: Same as Figure 11 for the molecule given by SMILES string $CC12CCC3OC1CC23$.

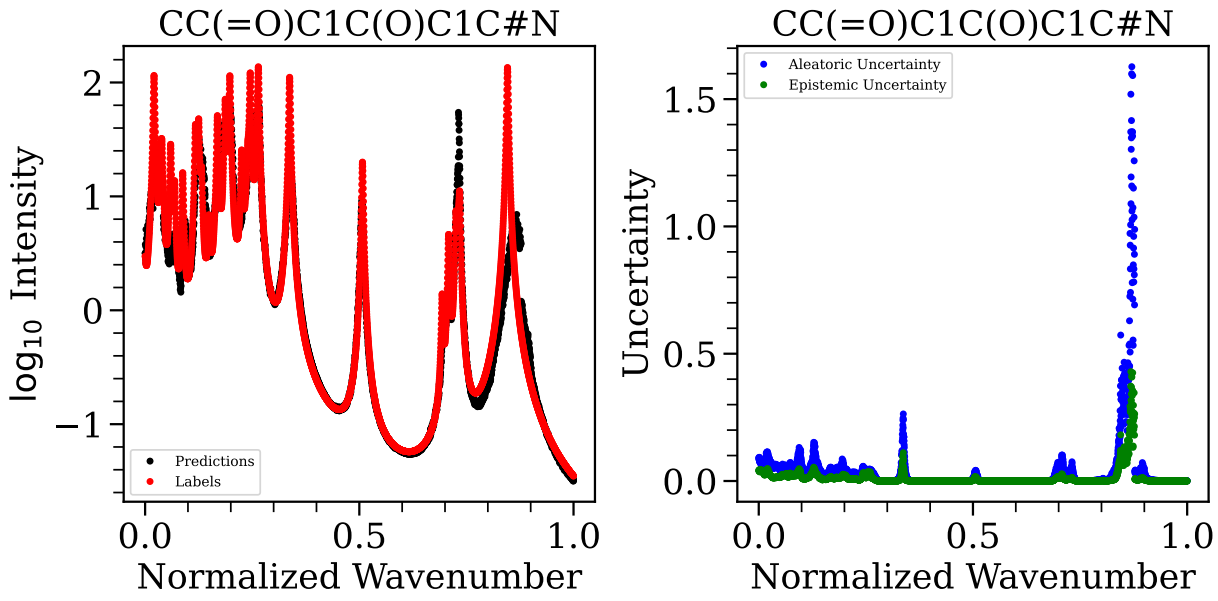

Figure 33: Same as Figure 11 for the molecule given by SMILES string $CC(=O)C1C(O)C1C\#N$.

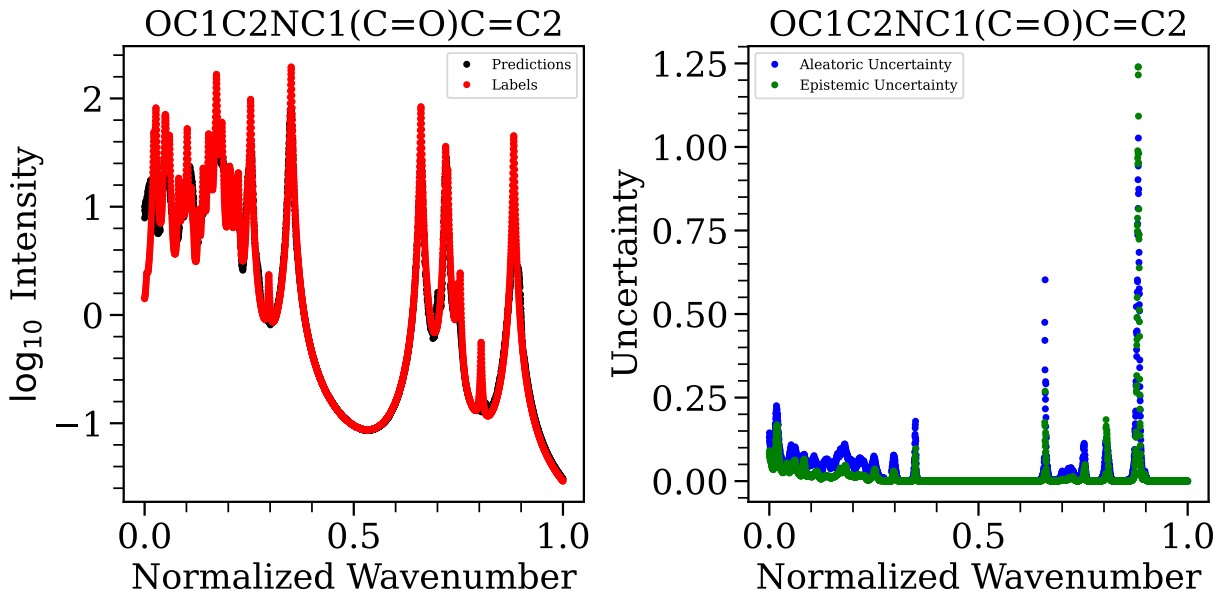

Figure 34: Same as Figure 11 for the molecule given by SMILES string $OC1C2NC1(C=O)C=C2$.

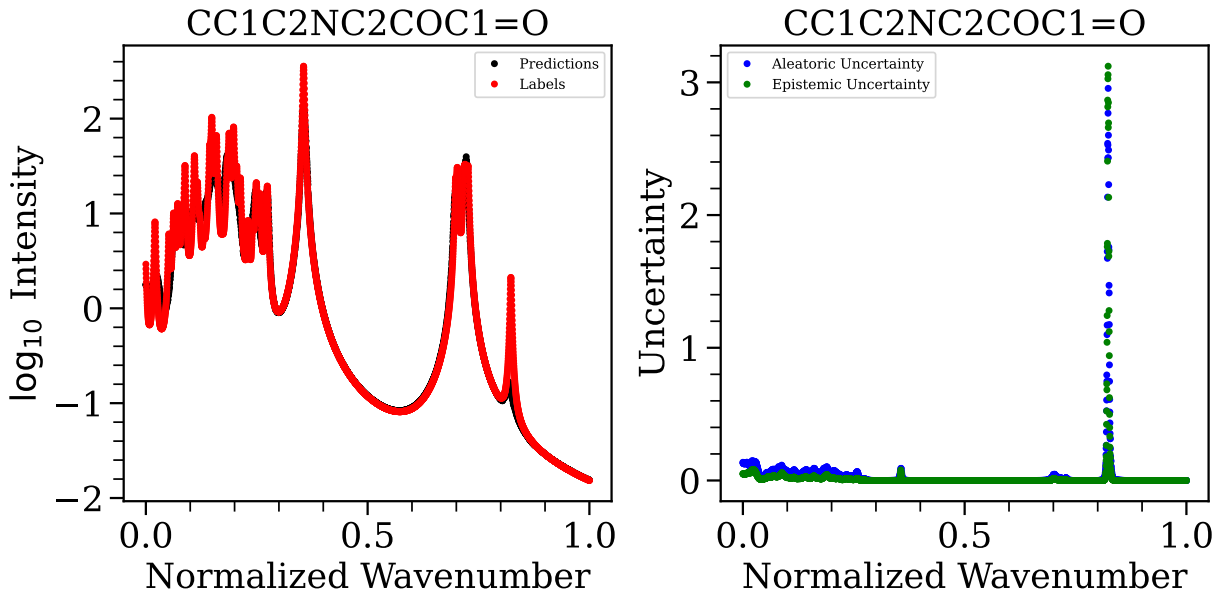

Figure 35: Same as Figure 11 for the molecule given by SMILES string $CC1C2NC2COC1=O$.

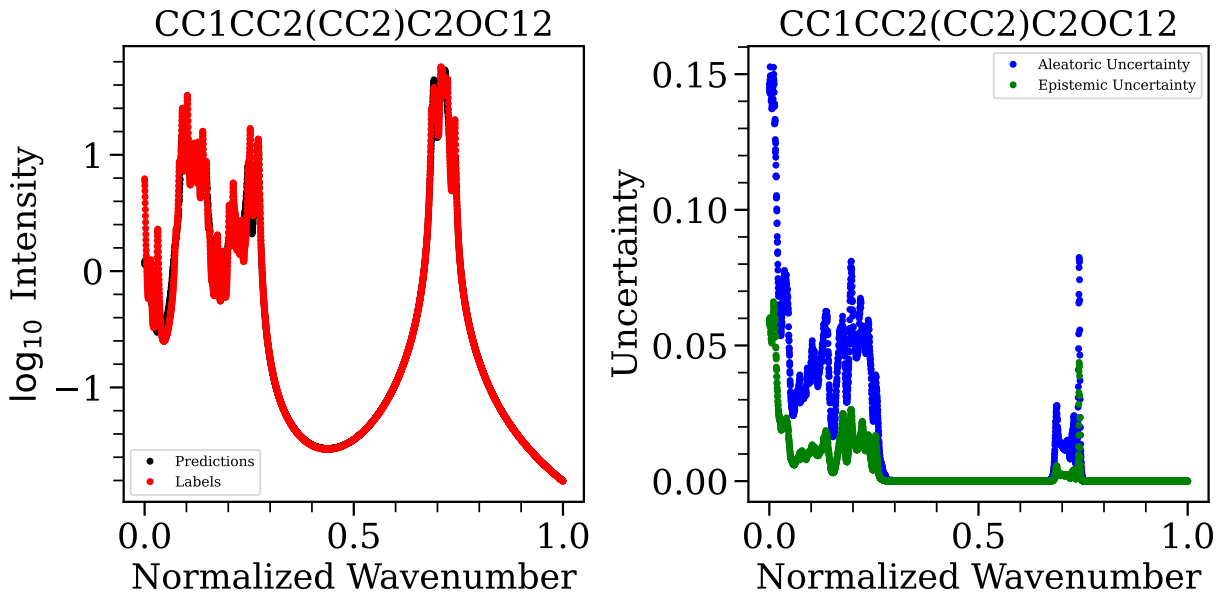

Figure 36: Same as Figure 11 for the molecule given by SMILES string $CC1CC2(CC2)C2OC12$.

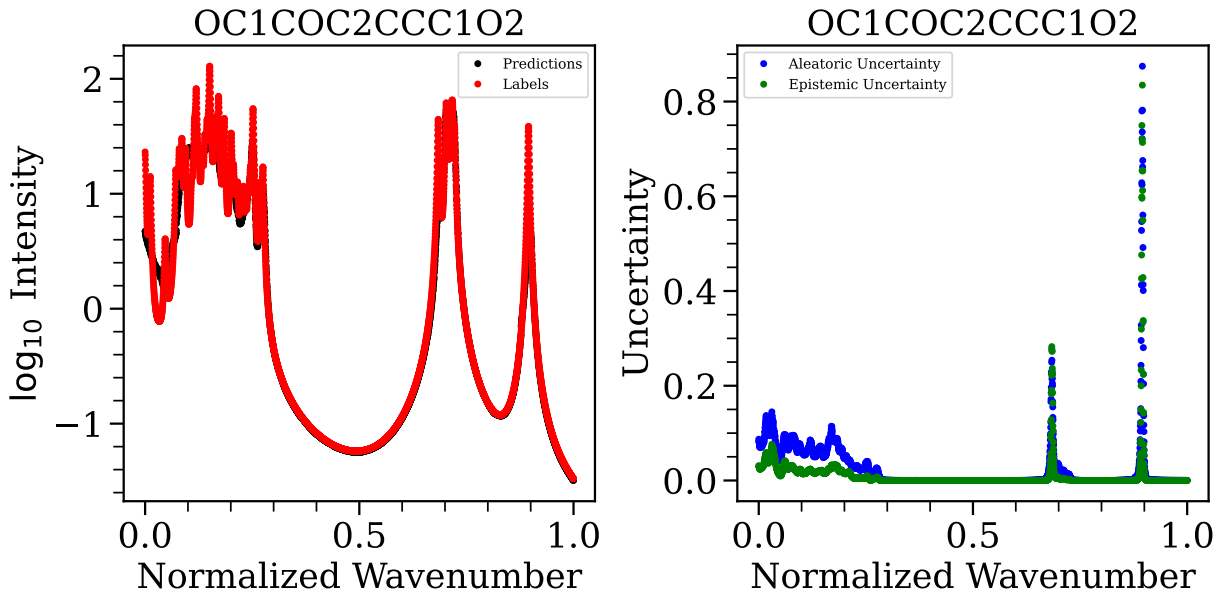

Figure 37: Same as Figure 11 for the molecule given by SMILES string $OC1COC2CCC1O2$.

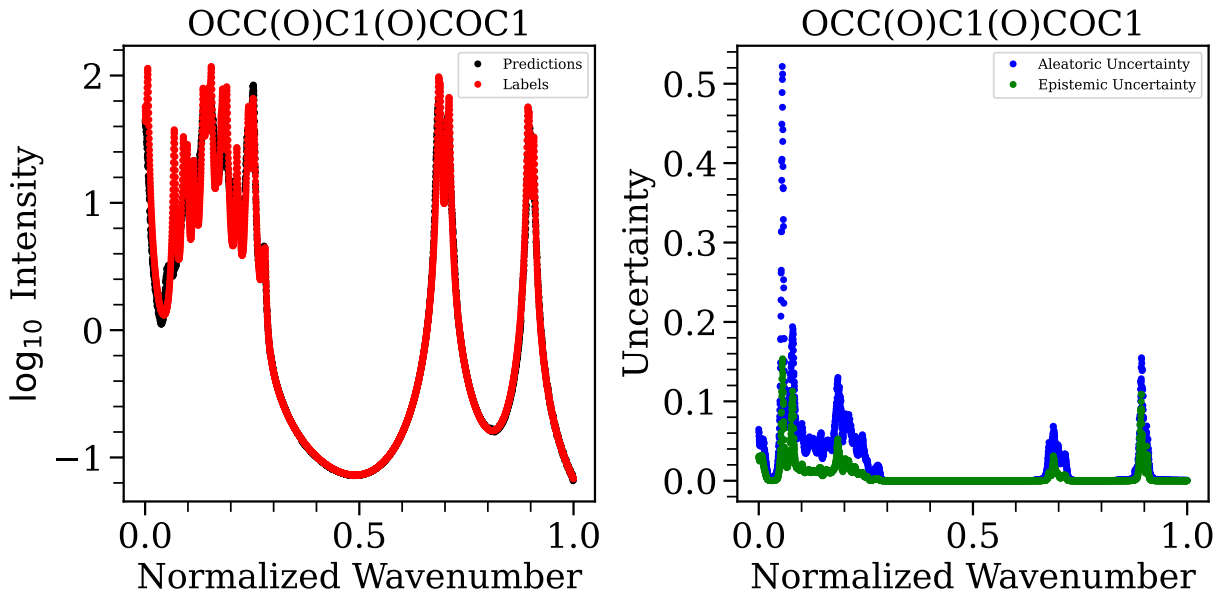

Figure 38: Same as Figure 11 for the molecule given by SMILES string $OCC(O)C1(O)COC1$.

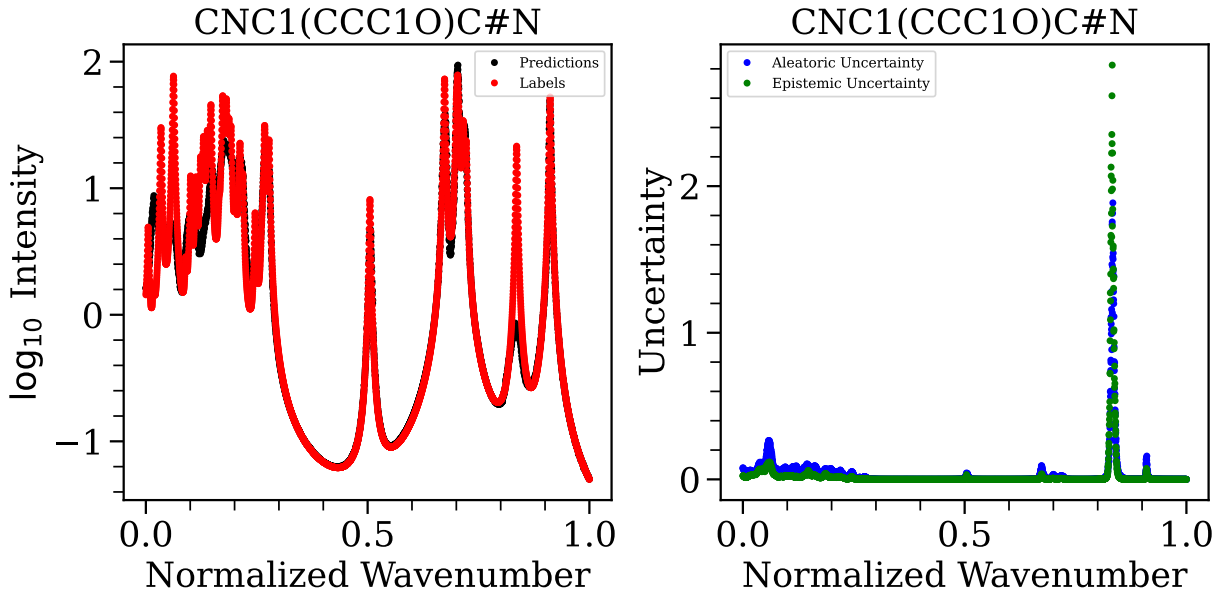

Figure 39: Same as Figure 11 for the molecule given by SMILES string $CNC1(CCC1O)C\#N$.

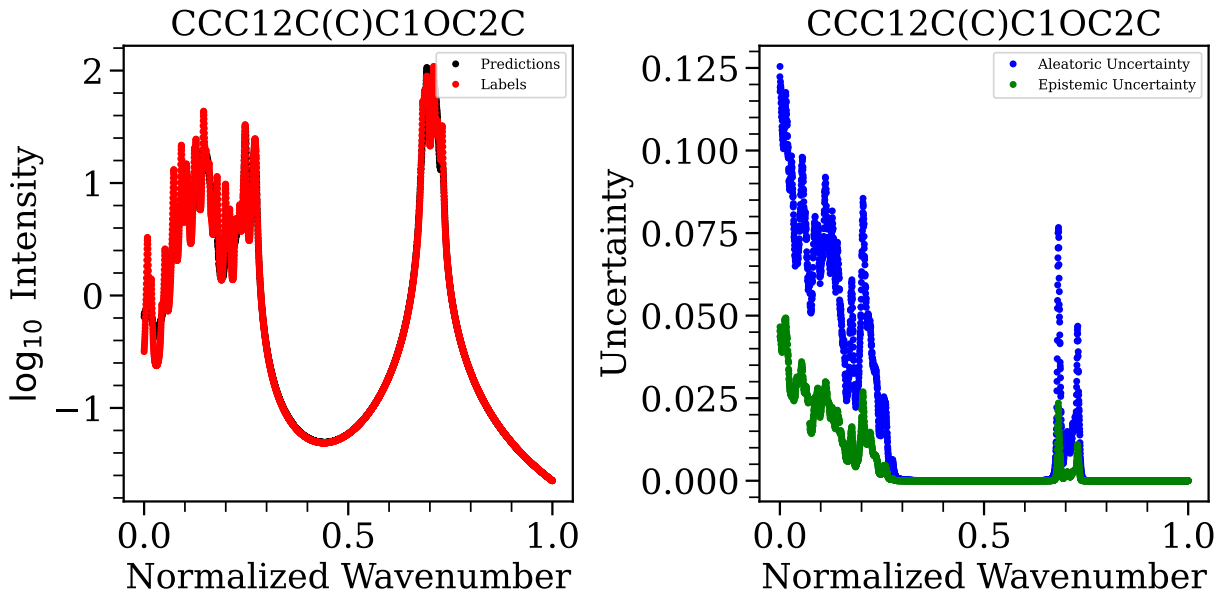

Figure 40: Same as Figure 11 for the molecule given by SMILES string $CCC12C(C)C1OC2C$.

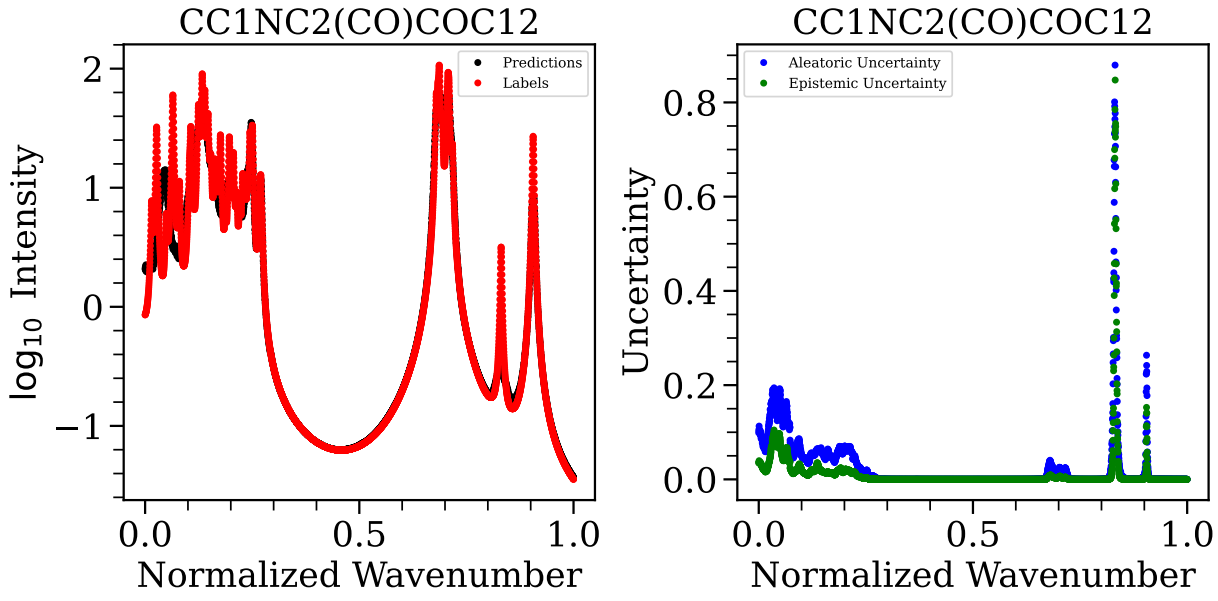

Figure 41: Same as Figure 11 for the molecule given by SMILES string $CC1NC2(CO)COC12$.

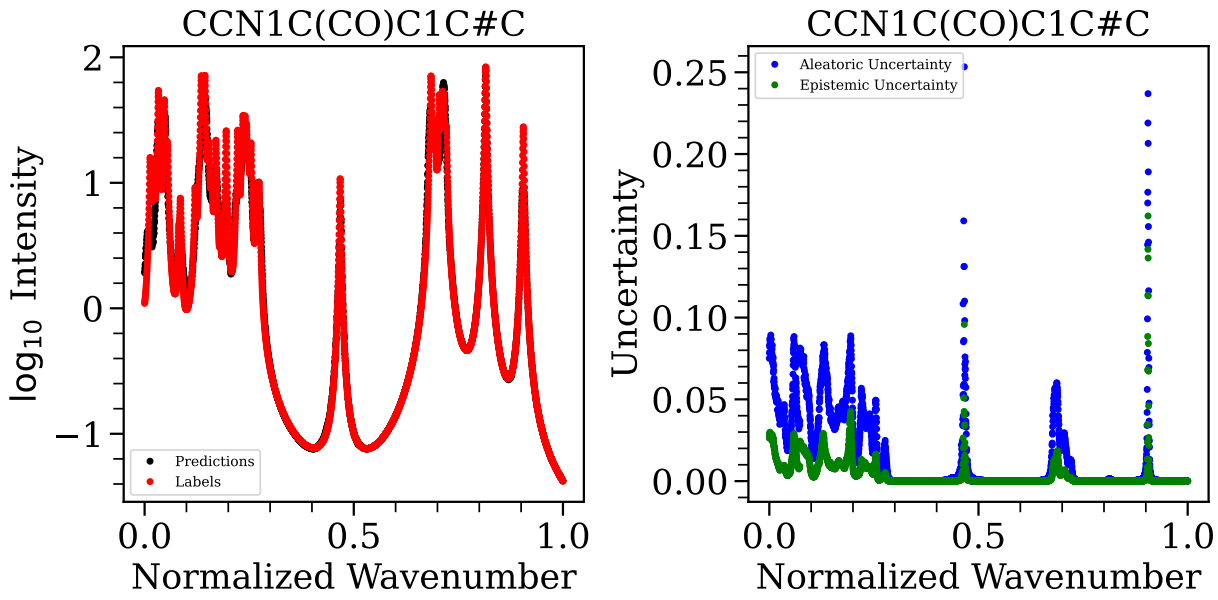

Figure 42: Same as Figure 11 for the molecule given by SMILES string $CCN1C(CO)C1C\#C$.

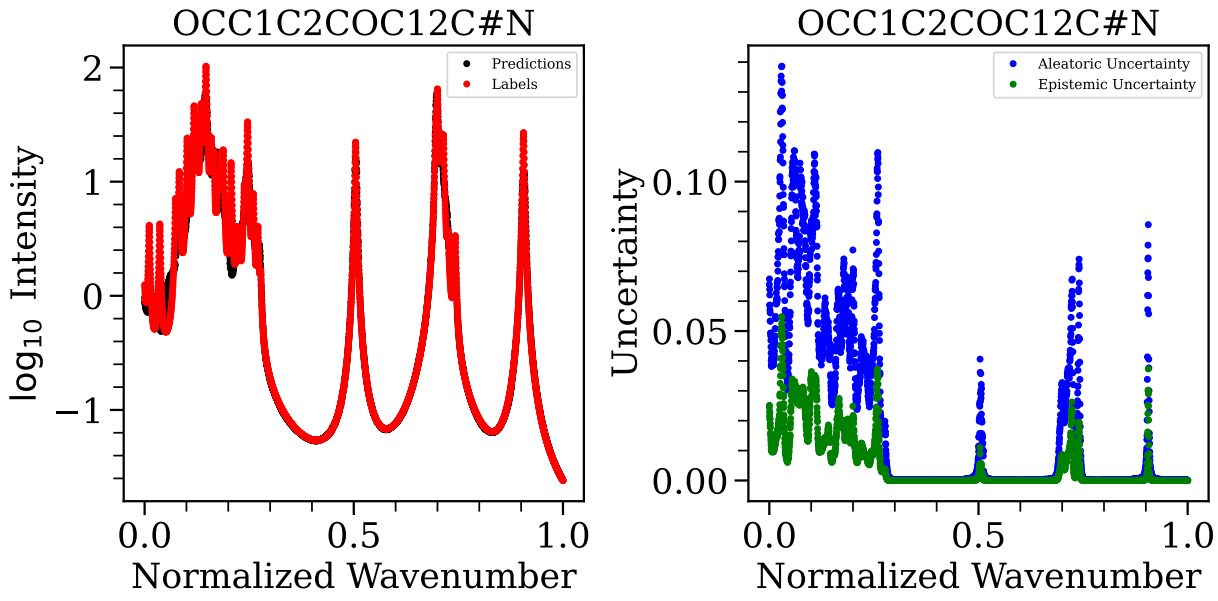

Figure 43: Same as Figure 11 for the molecule given by SMILES string $OCC1C2COC12C\#N$.

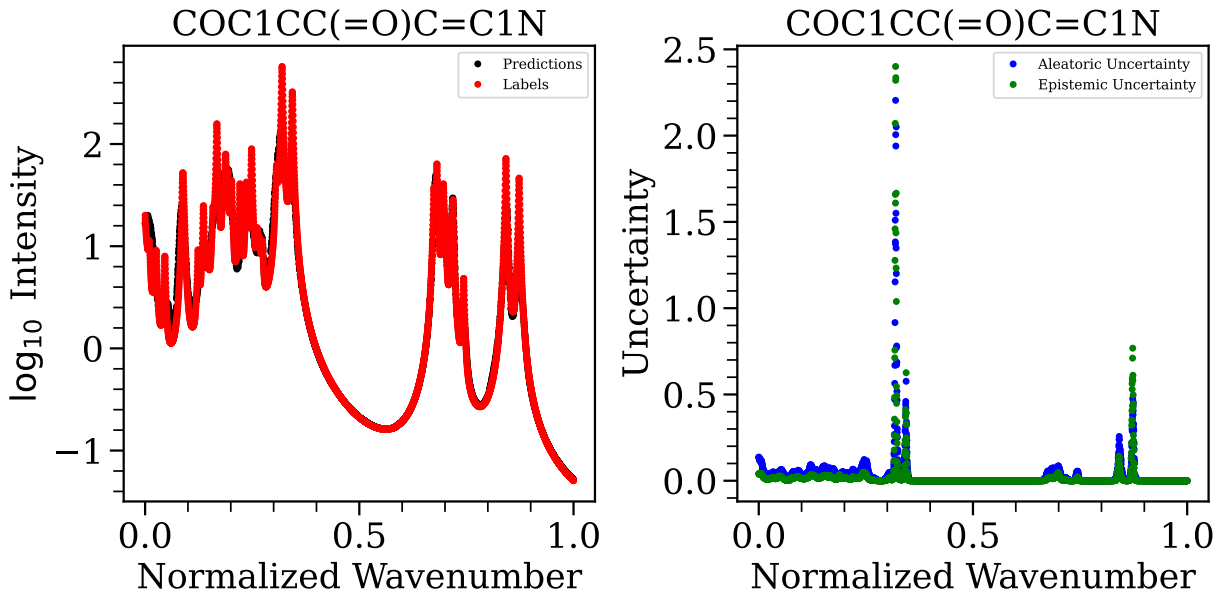

Figure 44: Same as Figure 11 for the molecule given by SMILES string $COC1CC(=O)C=C1N$.

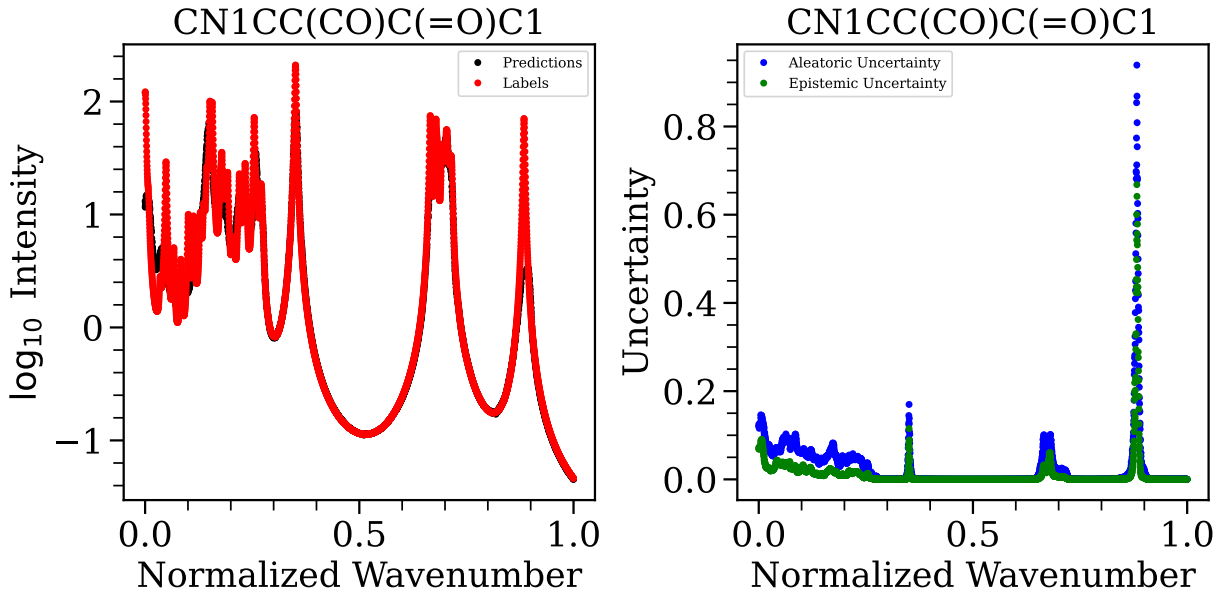

Figure 45: Same as Figure 11 for the molecule given by SMILES string $CN1CC(CO)C(=O)C1$.

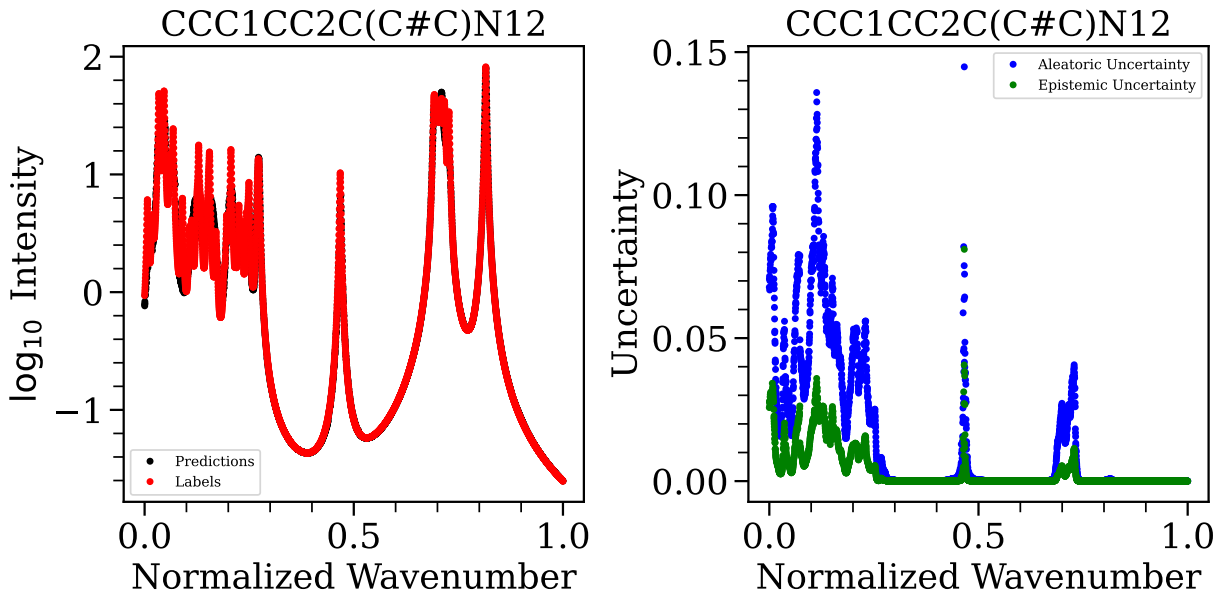

Figure 46: Same as Figure 11 for the molecule given by SMILES string $CCC1CC2C(C\#C)N12$.

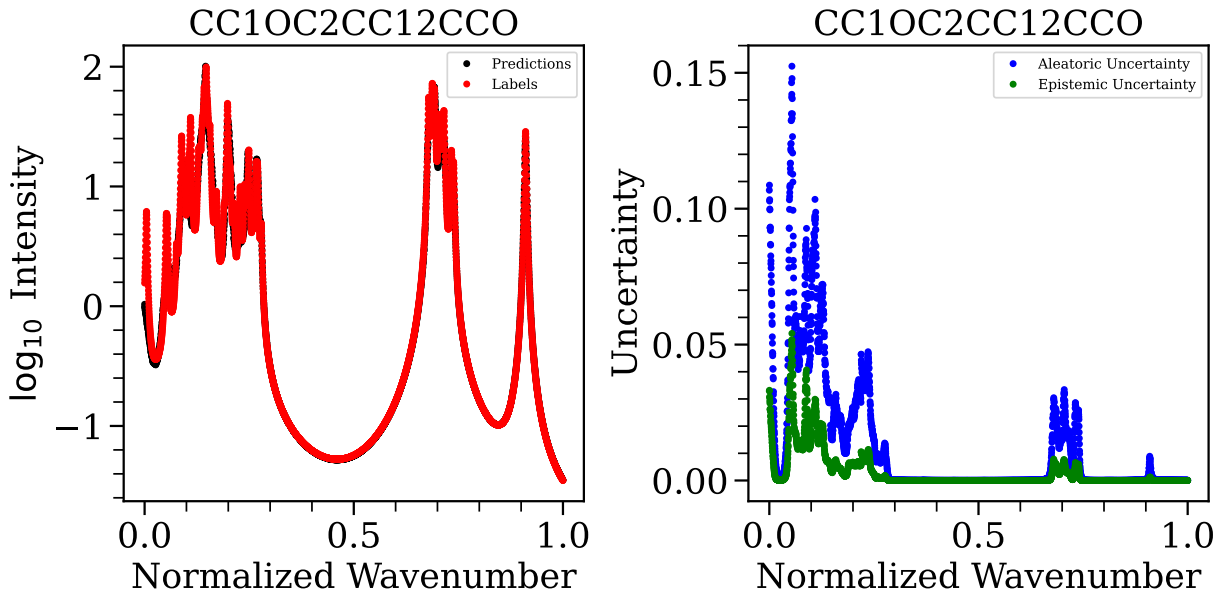

Figure 47: Same as Figure 11 for the molecule given by SMILES string $CC1OC2CC12CCO$.

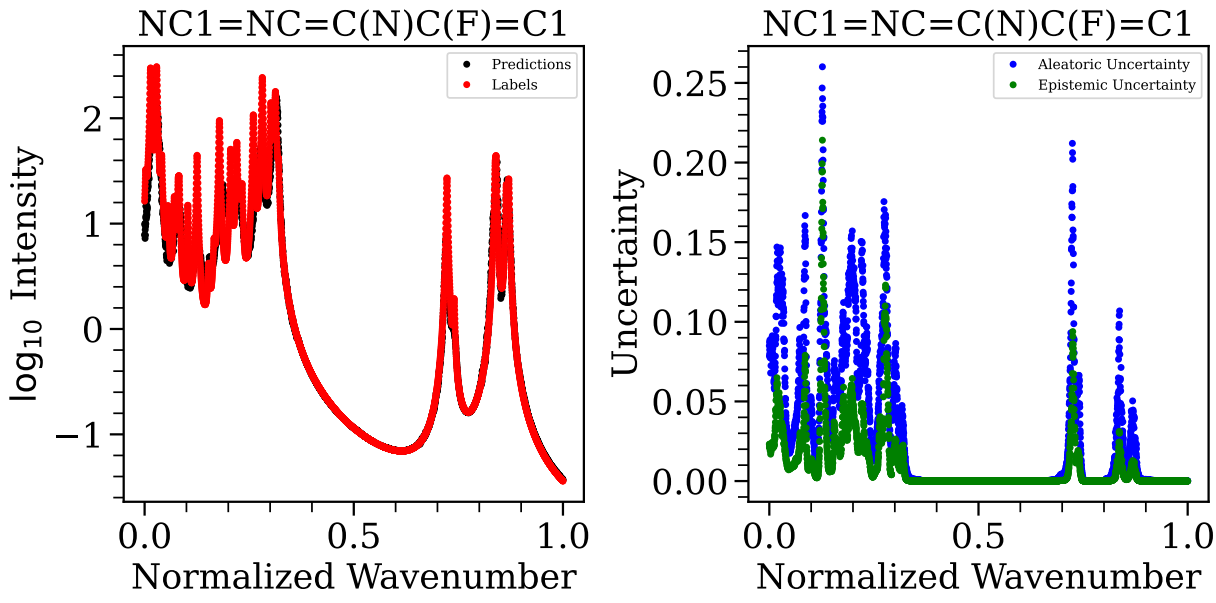

Figure 48: Same as Figure 11 for the molecule given by SMILES string $NC1 = NC = C(N)C(F) = C1$.

L Relation to Classification Error

If we start with

$$c \leq \int_{p=0}^{p=1} r(p)\text{MAX}(p)dp \tag{94}$$

$$= \int_{p=0}^{p=1} r(p)\mathbb{P}\left(f(x) = h_Y(x)|h_P(x) = p\right) dp \tag{95}$$

$$= \mathbb{P}\left(f(x) = h_Y(x)\right) \tag{96}$$

Guo et al. (2017) tells us that accuracy, given by Equation 82, is an unbiased estimator for $\mathbb{P}\left(f(x) = h_Y(x)\right)$ (as long as we let $m = 1$ such that I_m is the entire domain). Notice that Equation 82 is reminiscent of Equation 65. Indeed, the expected value for an indicator function is just the probability. Thus, we may write that

$$\mathbb{P}\left(f(x) = h_Y(x)\right) = \mathbb{E}_{x\sim q}\left[\mathbb{I}(f(x) = h_Y(x))\right] \tag{97}$$

$$= \mathbb{E}_{x\sim q}\left[1 - \mathbb{I}(f(x) \neq h_Y(x))\right] \tag{98}$$

$$= \mathbb{E}_{x\sim q}\left[1\right] - \mathbb{E}_{x\sim q}\left[\mathbb{I}(f(x) \neq h_Y(x))\right] \tag{99}$$

$$= 1 - \text{err}_{\text{cls}}(h). \tag{100}$$

This relationship is useful for comparison with Wang et al. (2024) and provides an alternative way to intuit the accuracy of a correct label prediction.