Improving Equivariant Networks with Probabilistic Symmetry Breaking

Hannah Lawrence [*1] **Vasco Portilheiro** [*2] **Yan Zhang** [3 4] **Sékou-Oumar Kaba** [5 4]

Abstract

Equivariance builds known symmetries into neural networks, often improving generalization. However, equivariant networks cannot break self-symmetries present in any given input. This poses an important problem: (1) for prediction tasks on symmetric domains, and (2) for generative models, which must break symmetries in order to reconstruct from highly symmetric latent spaces. Thus, equivariant networks are fundamentally limited when applied to these contexts. To remedy this, we present a comprehensive, probabilistic framework for symmetry-breaking, based on a novel decomposition of equivariant *distributions*. Concretely, this decomposition yields a practical method for breaking symmetries in any equivariant network via randomized *canonicalization*, while retaining the inductive bias of symmetry. We experimentally show that our framework improves the performance of group-equivariant methods in modeling lattice spin systems and autoencoding graphs.

1. Introduction

Even in seemingly symmetric domains, there are functions that equivariant networks simply cannot represent. Consider the problem of one-shot prediction of one molecular three-dimensional graph from another, such as a dichlorobenzene molecule from benzene (pictured on the right and left of Fig. 1). Such tasks are relevant in generative modeling of atomic systems (Satorras et al., 2021; Xie et al., 2021) and molecular editing (Liu et al., 2024), for example. Since we are working in 3D space, rotation equivariance is a natural choice—intuitively, rotating the benzene molecule should only affect the rotation of the dichlorobenzene, not the molecule itself. While this approach seems reasonable, a strange problem arises. Since benzene has sixfold rotational symmetry, an equivariant model is unable to output dichlorobenzene, which is not rotationally symmetric. This is a fundamental limitation of equivariant models: when the input is self-symmetric, they cannot break that self-symmetry in the output, as noted by e.g. Smidt et al. (2021).

Self-symmetry arises in a variety of applications, often with more complex groups—e.g. non-trivial graph automorphisms, spin models on symmetric domains, or rotationally self-symmetric point-clouds (Fig. 3). Also noteworthy is the case of generative models and autoencoders that reconstruct from a latent space: by virtue of being embedded in a simple, low-dimensional space, a latent representation often has greater self-symmetry than the input itself. An equivariant decoder must then map the more symmetric latent space to the less symmetric data space, which is just as impossible as predicting dichlorobenzene from benzene. One could discard symmetry structure entirely, but this loses the benefits of equivariance on asymmetric inputs. The question is thus: *how can we retain the inductive bias of symmetry, while resolving the difficulty posed by self-symmetric inputs?*

We examine this question from a probabilistic perspective. Extending the results of Bloem-Reddy & Teh (2020) on probabilistic symmetries, we derive a method that allows us to provably represent *any* equivariant conditional distribution (Section 2). For this, we rely on an external source of randomness, coming from a canonicalization function (Kaba et al., 2023) which has been appropriately randomized. Our theory suggests a simple modification to existing equivariant models allowing them to break symmetries (Section 3), which we validate experimentally (Fig. 2; Section 4). Using our framework, we also show that some recently proposed approaches to symmetry breaking (Kaba & Ravanbakhsh, 2023; Xie & Smidt, 2024) can be modified to represent any equivariant conditional distribution.

2. Representation of Equivariant Distributions

Probabilistic symmetry Bloem-Reddy & Teh (2020) justify the use of equivariant models by showing that, if G

[*]Equal contribution; order determined by coin flip. [1]Department of Electrical Engineering and Computer Science, MIT [2]Gatsby Computational Neuroscience Unit, UCL [3]Samsung, SAIT AI Lab, Montreal [4]Mila, Quebec Artficial Intelligence Institute [5]School of Computer Science, McGill University. Correspondence to: Hannah Lawrence <hanlaw@mit.edu>, Vasco Portilheiro <vascopo@gmail.com>.

Proceedings of the Geometry-grounded Representation Learning and Generative Modeling at 41st International Conference on Machine Learning, Vienna, Austria. PMLR Vol Number, 2024.

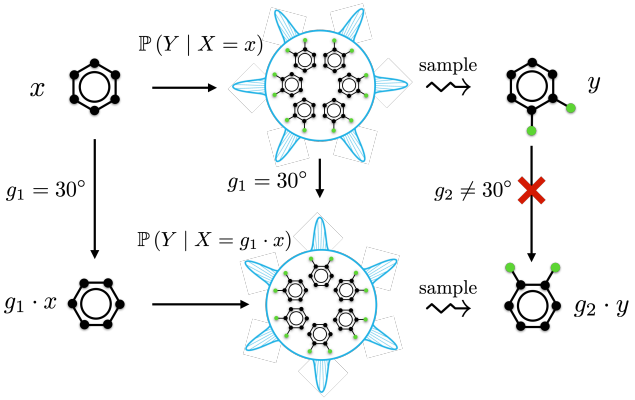


Figure 1. When the input x is rotated, an equivariant distribution $Y|X = x$ rotates. The distribution thus has (at least) the same self-symmetry as x. Individual samples may break this self-symmetry.

acts on $\mathcal{X}$ *freely*[1] and the distribution of a random $X \in \mathcal{X}$ is G-invariant, then $Y|X$ is *conditionally equivariant*[2] if and only if there exists a function $f : \mathcal{X} \times (0,1) \to \mathcal{Y}$ equivariant in x (i.e. $f(gx, \epsilon) = gf(x, \epsilon)$) such that

$$Y \overset{a.s.}{=} f(X, \epsilon),\ \epsilon \sim \mathrm{Unif}(0,1) \perp X. \tag{1}$$

We thus have a *functional representation* of the equivariant distribution, in the sense that it is expressed as a deterministic function of random inputs.

Symmetry breaking The requirement of a free action is a strong one, and amounts to having no input with self-symmetries. For general group actions, we are confronted with *Curie's principle* (Smidt et al., 2021), which states that $G_x \subseteq G_{f(x)}$ when f is equivariant. The equivariance of f therefore means that for any possible values of x and ϵ, the output sample $f(x, \epsilon)$ will have at least the self-symmetry of x. But in a general probabilistic system, Curie's principle only holds *at the level of distributions*. That is, if $Y|X$ is equivariant and $gx = x$, then $Y|X = x$ is sampled from a distribution of which g is a self-symmetry: $\mathbb{P}(Y \in A|X = x) = \mathbb{P}(gY \in A|X = x)$. Nonetheless, a sample of Y need not have any such symmetry (see right panel of Fig. 1). This is what physicists call *spontaneous symmetry breaking* (Beekman et al., 2019).

Randomized canonicalization (the inversion kernel) One way to understand why the representation Eq. (1) is limited to free group actions is that it relies on the existence of an equivariant *canonicalization* function (Kaba et al., 2023). Such functions do not exist when G does not act freely.[3]

Definition 2.1. An *orbit representative map* $\gamma : \mathcal{X} \to \mathcal{X}$ is a G-invariant function such that $\gamma(x) \in [x]$ for all x. A *canonicalization function* $\tau : \mathcal{X} \to G$ is a function such that $x \mapsto \tau(x)^{-1}x$ is an orbit representative map.

The lack of an equivariant canonicalization has been approached from various related angles: a notion of relaxed equivariance (Kaba & Ravanbakhsh, 2023); generalizing frames (Puny et al., 2022) to be weakly equivariant (Dym et al., 2024); generalizing canonicalization to output more than a single orbit representative (Ma et al., 2024). In probability theory, this problem has been addressed via the *inversion kernel* (Kallenberg, 2011). Intuitively, it is the uniform distribution over all ways to canonicalize x.

Definition 2.2. Let $\gamma : \mathcal{X} \to \mathcal{X}$ be an orbit representative map. The *inversion kernel* evaluated at x is $\mathrm{Unif}(gG_{\gamma(x)})$, where $g \in G$ is such that $g\gamma(x) = x$.

Armed with these preliminaries, we state our **main theorem**. It proves the existence of a decomposition of any equivariant $Y|X$ in terms of X, noise ϵ, and additional randomness coming from the inversion kernel, which explicitly captures symmetry breaking. We also derive two functional representations of symmetry generalizing Bloem-Reddy & Teh (2020, Theorem 9).

Theorem 2.3 (Randomized canonicalization)**.** *$Y|X$ is equivariant iff*

$$Y \overset{a.s.}{=} \tilde{g}\phi(\gamma(X), \epsilon) \overset{a.s.}{=} \tilde{g}\phi(\tilde{g}^{-1}X, \epsilon) \tag{2}$$

for some function $\phi : \mathcal{X} \times (0,1) \to \mathcal{Y}$, independent noise $\epsilon \sim \mathrm{Unif}(0,1)$, and $\tilde{g}|X$ distributed according to the inversion kernel for some orbit representative map γ.

The proof (Appendix G.1) follows closely that of Bloem-Reddy and Teh, which bootstraps *deterministic* symmetry from a canonicalizer to that of $Y|X$. We analogously bootstrap the *distributional* symmetry of $\tilde{g}|X$ to that of a model $\tilde{g}\phi(\tilde{g}^{-1}X, \epsilon)$ for Y, where ϕ is an arbitrary function.

Corollary 2.4 (Joint equivariance)**.** *$Y|X$ is equivariant iff $Y \overset{a.s.}{=} f(X, \tilde{g}, \epsilon)$ for a function $f : \mathcal{X} \times G \times (0,1) \to \mathcal{Y}$ jointly equivariant in its first two inputs, noise $\epsilon \sim \mathrm{Unif}(0,1)$, and $\tilde{g}|X$ is distributed according to some inversion kernel.*

Intuitively, our result decomposes the randomness in $Y|X$ into symmetry breaking and independent noise. The next functional representation is in terms of *relaxed equivariance* (Kaba & Ravanbakhsh, 2023).

Definition 2.5. A function $f : \mathcal{X} \to \mathcal{Y}$ is *relaxed equivariant* if for any $x \in \mathcal{X}$ and $g \in G$, we have $hf(x) = f(gx)$ for some $h \in gG_x$.

Corollary 2.6 (Relaxed equivariance)**.** *$Y|X$ is equivariant iff $Y \overset{a.s.}{=} g_X f(X, \epsilon)$ for some $f : \mathcal{X} \times (0,1) \to \mathcal{Y}$ relaxed equivariant in its first input, with $g_X \sim \mathrm{Unif}(G_X)$, and $\epsilon \sim \mathrm{Unif}(0,1)$ independent of X and g_X.*

[1] This means that the self-symmetry group, given by the stabilizer $G_x := \{g \in G : gx = x\}$, of any $x \in \mathcal{X}$ is trivial.

[2] $\mathbb{P}(Y \in A|X = x) = \mathbb{P}(gY \in A|gX = x)$

[3] Suppose $\tau : \mathcal{X} \to G$ is equivariant. For any $h \in G_x$, $h\tau(x) = \tau(hx) = \tau(x)$ and thus $h = \mathrm{id}$.

We justify these symmetry breaking perspectives in Appendices I and J, showing generalization advantages of distributional and relaxed equivariant models.

3. Proposed Method

Following Corollary 2.4 and as shown in Fig. 4, we propose to represent equivariant conditionals using an equivariant neural network f, and pass in $(x, \tilde{g})$ as input (and ϵ, if so desired). This approach can be used whenever the number of input channels to f is flexible. Namely, for any "bias" vector $v \in V$ (where G acts on V), we use the concatenation $x \oplus (\tilde{g}v)$ as input to f; the equivariance of $(x, g) \mapsto x \oplus (gv)$ is clear. The vector v can be fixed, or a learnable parameter; it can also be a matrix, with columns for channels. As shown in the next proposition, if G_v is trivial, our method is fully expressive. (See Appendix H for examples, including when v is a matrix.)

Proposition 3.1. *Let $f : \mathcal{X} \times G \to \mathcal{Y}$ be any function jointly equivariant in its arguments, i.e. $f(hx, hg) = h \cdot f(x, g)$. If $v \in V$ has trivial stabilizer, then there exists an equivariant function $f_0 : \mathcal{X} \times [v] \to \mathcal{Y}$ such that $f(x, g) = f_0(x \oplus gv)$.*

A priori, our method requires sampling $\tilde{g} \sim \mathrm{Unif}(gG_{\gamma(X)})$ for some choice of orbit representatives γ. In some cases, one can define a procedure "by hand": for permutation groups one can use ranking with random tie breaking (e.g. by adding very small noise). Another method for doing so is to generalize a proposal of Kaba et al. (2023) for building relaxed equivariant canonicalizers. For any *energy function* $E : \mathcal{X} \to \mathbb{R}$, consider the set $S(x) = \arg\min_{g \in G} E(g^{-1}x)$. Note that if $g^* \in S(x)$, then $g^* G_x \subseteq S(x)$. If E is "non-degenerate" in the sense that $g^* G_x = S(x)$, i.e. a unique element $\gamma(x) \in [x]$ minimizes the energy, then $S(x) = gG_{\gamma(x)}$ where $g\gamma(x) = x$. That is, we can sample $\tilde{g}$ by sampling uniformly from the minimizer of a non-degenerate energy function, which may be learned as a neural network. For small groups, one may calculate the minimizer directly (or on a fine enough grid). For larger groups, one may optimize the energy by SGD.

Depending on the context, one may also not care about actually getting a *uniform* sample from $gG_{\gamma(x)}$ (i.e. having distributional equivariance of $\tilde{g}$). One case is a prediction task with a group-invariant loss function, where the different ways to break symmetry will result in the same loss value. Then, one may simply use an existing deterministic (relaxed equivariant) canonicalizer to obtain $\tilde{g}$.

3.1. Symmetry breaking with noise injection

Adding noise to inputs is a commonly used heuristic for symmetry breaking (Satorras et al., 2021; Sato et al., 2021; Abboud et al., 2021; Zhao et al., 2024). This involves using a functional model similar to that of Corollary 2.4, but with an arbitrary noise variable Z replacing $\tilde{g}$. It is natural to ask whether this simple heuristic can also represent any equivariant distribution. We show that under some conditions, noise injection can indeed be used to break symmetries and represent any equivariant $Y|X$.

Proposition 3.2 (Noise injection)**.** *Let X, Y, Z be random variables in $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$ respectively, each space acted on by G. The following are equivalent: (1) G acts on $\mathcal{Z}$ freely (up to a set of probability zero) and $Z|X$ is equivariant; (2) $Y|X$ is equivariant iff there exists $f : \mathcal{X} \times \mathcal{Z} \times (0, 1) \to \mathcal{Y}$ jointly equivariant in X and Z such that $Y \overset{a.s.}{=} f(X, Z, \epsilon)$ for noise $\epsilon \sim \mathrm{Unif}(0, 1)$.*

This result implies that one may sample $\tilde{g}$ instead from a general equivariant distribution on G, and still obtain an equivariant $Y|X$.[4] Moreover, for many groups of interest, such as S_n and $O(n)$, simply sampling Z from an isotropic Gaussian satisfies the requirements. However, this method introduces more noise than is *necessary* into the learning process, while the inversion kernel is "minimal" (Appendix C). It is thus likely beneficial to judiciously choose $Z|X$.

The result above also allows us to draw formal connections to the recently proposed *symmetry breaking sets* (SBS) of Xie & Smidt (2024) (Appendix D). In short, our method generalizes SBS while offering certain advantages: theoretically, it enables sampling from the smallest possible "set" (via the inversion kernel), and practically, it is easy to implement via any existing canonicalization method.

4. Experiments

4.1. Predicting ground-states of Ising models

Ising models serve as prototypical examples of physical systems that exhibit symmetry breaking, making them natural candidates to evaluate our framework. Consider a square grid Λ with periodic boundary conditions and a configuration $\sigma \in \{-1, 1\}^\Lambda$ specifying a spin at each site. The Hamiltonian $H(\sigma) = -\sum_{i,j} J^x_{ij} \sigma_i \sigma_j - \sum_{i,j} J^y_{ij} \sigma_i \sigma_j - \sum_i h_j \sigma_j$ assigns an energy to each configuration, where the horizontal interaction $J^x_{ij} \neq 0$ only if i and j are horizontal neighbors and similarly for the vertical J^y_{ij}.

We consider unsupervised training of neural networks to obtain *ground states* of the Ising model given the Hamiltonian (Hu et al., 2017; Carrasquilla & Melko, 2017), i.e. configurations of non-zero probability of the equivariant distribution $P(\sigma|J^x, J^y, h) \propto \exp(-H(\sigma)/T)$ in the $T \to 0$ limit (configurations of minimal energy). This simple problem is of high interest, as brute force optimization scales exponen-

[4] In fact, the stochastic optimization of energy proposed above is akin to Langevin sampling from $p(g|x) \propto \exp(-E(g^{-1}x)/T)$. When $T > 0$ we sample from a general equivariant distribution on G, and we recover the exact case when $T \to 0$.

tially in the number of lattice sites. Monte-Carlo simulation is possible, but does not benefit from the generalization power of neural networks across different Hamiltonians.

Denoting the set of possible Hamiltonian parameters as $\mathcal{J}$, we train a neural network $\phi : \mathcal{J} \to [0,1]^\Lambda$ to output the probability that each spin is up. (See Appendix E for details.) Despite the apparent symmetry of the setup, an equivariant network indeed cannot reliably obtain samples of ground states. This is because while the Hamiltonian is invariant under the automorphism group $p4m$ of the square grid (as a function of both spins and parameters), i.e. $H(g \cdot J^x, g \cdot J^y, g \cdot h, g \cdot \sigma) = H(J^x, J^y, h, \sigma)$, ground-states can in general have less self-symmetry.

We use as baselines: a $p4m$-equivariant G-CNN (Cohen & Welling, 2016; Weiler & Cesa, 2019), a non-equivariant MLP trained with data augmentation sampled from $p4m$, and adding noise $\epsilon \sim \mathcal{N}(0,1)$ to the input of the G-CNN to break symmetry. As we only care about predicting single a ground state, and ground states are symmetrically related, for our method we use a deterministic canonicalizer (see Appendix E) to obtain $\tilde{g}$ (and drop ϵ). We can directly use the average energy over the test set as an evaluation metric. Reaching the ground truth value would guarantee that network outputs the true ground-state for each Hamiltonian. **The proposed method achieves lower energy than all the alternatives** (Table 2, Appendix E). A more in-depth understanding of their shortcomings can be obtained from the predicted phase diagrams (Fig. 2).

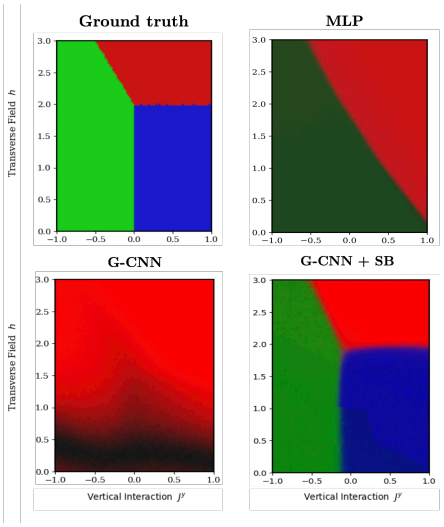


Figure 2. Phase diagrams predicted by the different methods (best viewed in color). For each configuration predicted by the neural network on a test set Hamiltonian, we compute the values of the three order parameters: the ferromagnetic phase (red), the antiferromagnetic phase (green), and the stripes phase (blue). Brighter colors are associated with larger values of the order parameter and black to the disordered phase. As shown, only symmetry-breaking enables recovery of the correct phase diagram.

4.2. Graph Autoencoder

Autoencoders with symmetric latent spaces pose a problem for equivariant models. One example explored in Satorras et al. (2021) is autoencoding graphs using a node-wise latent space $\mathcal{Z} = \mathbb{R}^{n \times f}$, where n is the number of nodes and f is the feature dimension. From an S_n-equivariant embedding in $\mathcal{Z}$, the graph is decoded equivariantly; the presence of an edge between nodes with latents z_i and z_j is a function of $\|z_i - z_j\|$. If A is the adjacency matrix of the graph, its self-symmetries are $G_A = \{g \in S_n : gAg^T = A\}$. However, there may not even *exist* an embedding $z \in \mathcal{Z}$ such that $G_A = G_z$ (see Appendix F for details), which by Curie's principle results in an "overly symmetric" embedding with $G_A \subsetneq G_z$ (Satorras et al., 2021, Figure 3) when *any* equivariant encoder is used.

We consider reconstructing Erdős-Rényi random graphs with edge probability 0.25, using the data from Satorras et al. (2021) and a standard message-passing architecture as the encoder. To break symmetries using our method, we truncate Laplacian positional encodings to the fourth largest singular values, $P \in \mathbb{R}^{n \times 4}$, then apply a learnable dimensionality reduction $w \in \mathbb{R}^4$ to obtain the vector $v = Pw \in \mathbb{R}^n$. The symmetry breaking input $\tilde{g}v$ is obtained by sorting v, letting the sorting algorithm break ties. As baselines, we consider no symmetry breaking (no SB), randomly initialized node features (noise), randomly sampling $\tilde{g}$ from S_n (uniform), and passing in P directly (Laplacian). **Breaking symmetries via our method achieves the lowest error** (Table 1).

Table 1. Cross-entropy loss and reconstruction error

Method	BCE	% Error	Parameters
No SB	28.5	9.7	88,017
Noise	21.9	6.4	88,017
Uniform	17.7	5.0	88,890
Laplacian	18.7	5.3	88,017
L+canon (ours)	**10.8**	**2.8**	88,890

5. Conclusion

By considering symmetry breaking from a probabilistic perspective, we derive a framework which is both theoretically grounded, and flexible and effective in practice. The experiments presented here focus on cases where deterministic symmetry breaking suffices. We leave evaluation of full distributional models of equivariant $Y|X$ to future iterations of this work. In addition to further empirical testing, work remains to sample $\tilde{g}$ in practice for general groups—e.g. testing our proposal of using energy-based modeling—and in exploring the tradeoff between structured symmetry breaking and noise injection. Additionally, it remains to address *partial* symmetry breaking, as treated for example by Xie & Smidt (2024).

References

Abboud, R., Ceylan, İ. İ., Grohe, M., and Lukasiewicz, T. The surprising power of graph neural networks with random node initialization. In *Proceedings of the Thirtieth International Joint Conference on Artifical Intelligence (IJCAI)*, 2021.

Beekman, A. J., Rademaker, L., and van Wezel, J. An introduction to spontaneous symmetry breaking. *SciPost Phys. Lect. Notes*, pp. 11, 2019. doi: 10.21468/SciPostPhysLectNotes.11. URL https://scipost.org/10.21468/SciPostPhysLectNotes.11.

Bloem-Reddy, B. and Teh, Y. Probabilistic symmetries and invariant neural networks. *Journal of Machine Learning Research*, 21(90):1–61, 2020. URL http://jmlr.org/papers/v21/19-322.html.

Carrasquilla, J. and Melko, R. G. Machine learning phases of matter. *Nature Physics*, 13(5):431–434, 2017. doi: 10.1038/nphys4035. URL https://doi.org/10.1038/nphys4035.

Chiu, K. and Bloem-Reddy, B. Non-parametric hypothesis tests for distributional group symmetry, 2023.

Cohen, T. and Welling, M. Group equivariant convolutional networks. In *International conference on machine learning*, pp. 2990–2999, 2016.

Dym, N., Lawrence, H., and Siegel, J. W. Equivariant frames and the impossibility of continuous canonicalization. *arXiv preprint arXiv:2402.16077*, 2024.

Elesedy, B. *Symmetry and Generalisation in Machine Learning*. Ph.D. Thesis, University of Oxford, 2023. URL https://bryn.ai/assets/phd-thesis.pdf.

Elesedy, B. and Zaidi, S. Provably strict generalisation benefit for equivariant models. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 2959–2969. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/elesedy21a.html.

Hu, W., Singh, R. R., and Scalettar, R. T. Discovering phases, phase transitions, and crossovers through unsupervised machine learning: A critical examination. *Physical Review E*, 95(6):062122, 2017.

Kaba, S.-O. and Ravanbakhsh, S. Symmetry breaking and equivariant neural networks. In *Symmetry and Geometry in Neural Representations Workshop, NeurIPS*, 2023.

Kaba, S.-O., Mondal, A. K., Zhang, Y., Bengio, Y., and Ravanbakhsh, S. Equivariance with learned canonicalization functions. In *International Conference on Machine Learning*, pp. 15546–15566. PMLR, 2023.

Kallenberg, O. Invariant palm and related disintegrations via skew factorization. *Probability Theory and Related Fields*, 149:279–301, 2011. URL https://api.semanticscholar.org/CorpusID:121316040.

Kallenberg, O. *Foundations of Modern Probability*. Springer International Publishing, Cham, 2021. ISBN 978-3-030-61871-1. doi: 10.1007/978-3-030-61871-1_9. URL https://doi.org/10.1007/978-3-030-61871-1_9.

Lim, D., Robinson, J., Jegelka, S., Lipman, Y., and Maron, H. Expressive sign equivariant networks for spectral geometric learning. In *ICLR 2023 Workshop on Physics for Machine Learning*, 2023.

Liu, J., Kumar, A., Ba, J., Kiros, J., and Swersky, K. Graph normalizing flows. *Advances in Neural Information Processing Systems*, 32, 2019.

Liu, S., Wang, C., Lu, J., Nie, W., Wang, H., Li, Z., Zhou, B., and Tang, J. Unsupervised discovery of steerable factors when graph deep generative models are entangled. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=wyU3Q4gahM.

Ma, G., Wang, Y., Lim, D., Jegelka, S., and Wang, Y. A canonization perspective on invariant and equivariant learning, 2024.

Mondal, A. K., Panigrahi, S. S., Kaba, O., Mudumba, S. R., and Ravanbakhsh, S. Equivariant adaptation of large pretrained models. In Oh, A., Neumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S. (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 50293–50309. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/9d5856318032ef3630cb580f4e24f823-Paper-Conference.pdf.

Puny, O., Atzmon, M., Smith, E. J., Misra, I., Grover, A., Ben-Hamu, H., and Lipman, Y. Frame averaging for invariant and equivariant network design. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=zIUyj55nXR.

Sato, R., Yamada, M., and Kashima, H. Random features strengthen graph neural networks. In *Proceedings of*

the 2021 SIAM international conference on data mining (SDM), pp. 333–341. SIAM, 2021.

Satorras, V. G., Hoogeboom, E., and Welling, M. E(n) equivariant graph neural networks. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 9323–9332. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/satorras21a.html.

Smidt, T. E., Geiger, M., and Miller, B. K. Finding symmetry breaking order parameters with euclidean neural networks. *Phys. Rev. Research*, 3: L012002, Jan 2021. doi: 10.1103/PhysRevResearch.3.L012002. URL https://link.aps.org/doi/10.1103/PhysRevResearch.3.L012002.

Srinivasan, B. and Ribeiro, B. On the equivalence between positional node embeddings and structural graph representations. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SJxzFySKwH.

Vignac, C. and Frossard, P. Top-n: Equivariant set and graph generation without exchangeability. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=-Gk_IPJWvk.

Wang, R., Hofgard, E., Gao, H., Walters, R., and Smidt, T. E. Discovering symmetry breaking in physical systems with relaxed group convolution, 2024.

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

Xie, T., Fu, X., Ganea, O.-E., Barzilay, R., and Jaakkola, T. Crystal diffusion variational autoencoder for periodic material generation. *arXiv preprint arXiv:2110.06197*, 2021.

Xie, Y. and Smidt, T. Equivariant symmetry breaking sets. *arXiv preprint arXiv:2402.02681*, 2024.

Yan, Q., Liang, Z., Song, Y., Liao, R., and Wang, L. Swingnn: Rethinking permutation invariance in diffusion models for graph generation, 2023.

Zhang, Y., Zhang, D. W., Lacoste-Julien, S., Burghouts, G. J., and Snoek, C. G. M. Multiset-equivariant set prediction with approximate implicit differentiation. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=5K7RRqZEjoS.

Zhao, L., Ding, X., and Akoglu, L. Pard: Permutation-invariant autoregressive diffusion for graph generation, 2024.

Appendices

A. Supplemental Figures

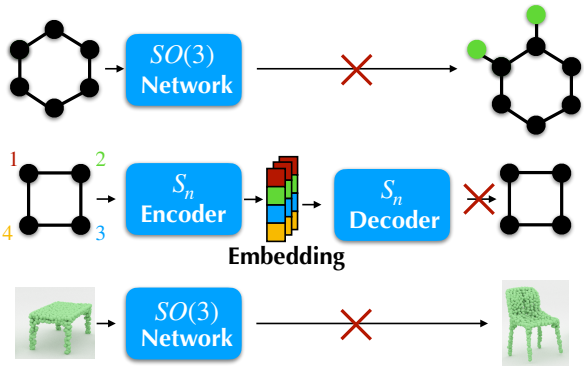


Figure 3. Example applications that require symmetry breaking.

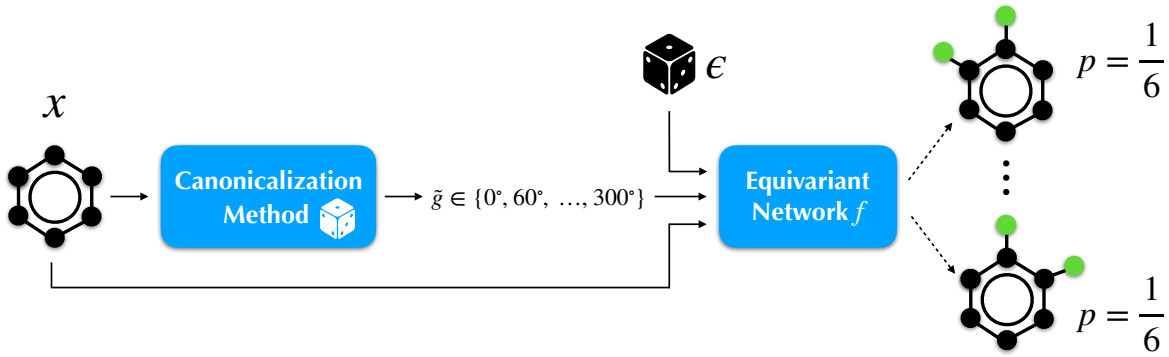


Figure 4. Illustration of our symmetry-breaking method. Here, a die indicates randomness, which is used in the canonicalization method to sample $\tilde{g}$, and optionally as another input to the equivariant network f to capture randomness unrelated to symmetry-breaking.

B. Related work

The problems posed by symmetric inputs in equivariant learning have been noticed in several domains.

In the context of graph representation learning, Srinivasan & Ribeiro (2020) showed that isomorphic structures in graphs must be assigned the same structural representations by equivariant functions. It has been seen in different applications, including generative modelling on graphs (Liu et al., 2019; Satorras et al., 2021; Yan et al., 2023; Zhao et al., 2024) and link prediction (Lim et al., 2023), that this is limiting. Our results formally justify that the symmetry of graph representations can be broken while still preserving distributional symmetry.

Similar issues arise for prediction and generation tasks on sets. Zhang et al. (2022) offered a characterization of this problem. First, equivariant functions are limited in processing multisets - sets with self-symmetries. Second, equivariant functions cannot decode from an invariant latent space to a set. They introduce a notion of *multiset equivariance* of which *relaxed equivariance* is a generalization. Vignac & Frossard (2022) also introduces a generalization of equivariance to address the problem of set generation, which our results encompass.

Other studies have focused on this problem in the context of machine learning for modelling physical systems. Smidt et al. (2021) have first formulated the preservation of symmetry in equivariant neural networks as an analogue to Curie's principle and have proposed to use gradients of equivariant neural networks to identify cases for which symmetry breaking is necessary. Kaba et al. (2023) has identified that for prediction tasks on crystal structures, symmetry breaking can be necessary and has proposed a method based on positional encodings. Wang et al. (2024) propose a flexible method for

implementing and interpreting symmetry breaking based on *relaxed group convolutions*. This notion can be seen as a form of approximate equivariance and is not related to the relaxed equivariance described earlier.

C. The inversion kernel injects minimal noise

We argue that using $\tilde{g}|X$ in place of independent $Z \in \mathcal{Z}$ introduces the least amount of noise possible into the functional representation of Proposition 3.2. We will measure "amount of noise" by conditional entropy, assuming G is finite for ease of exposition. $\mathcal{Z}$ is a disjoint union of orbits. Since entropy is additive, it is minimized if Z is restricted to a single orbit, which assuming G acts freely is isomorphic to G itself. We thus consider Z as a G-valued random variable.

First notice that if Z is independent of X, then it must be uniform on G to be equivariant. It is easy to show the uniform distribution maximizes entropy. To see inversion kernels minimize it, let $g \in G$ be such that $\mathbb{P}\left(Z = g|X = x\right) = p$. If $Z|X$ is equivariant, then for any $h \in G_x$ we have $\mathbb{P}\left(Z = hg|X = x\right) = p$. Thus the support of $Z|X = x$ is at least the size of G_x. Entropy will again be minimized if $Z|X = x$ indeed has a support of this size. This is exactly the case when $Z|X$ is in fact distributed like $\tilde{g}|X$, according to an inversion kernel.

D. Connection to equivariant symmetry breaking sets

Inspired by the idea that symmetry breaking arises from missing information, Xie & Smidt (2024) introduce the notion of *symmetry breaking sets* (SBS). Fixing an input x and some group action, they define a symmetry breaking set $B(x)$ as a G_x-dependent set on which G_x freely. Elements $b \in B(x)$ can then be fed into an equivariant function $f(x, b)$ together with the input x, in order to get outputs unconstrained by Curie's principle. The authors propose sampling the elements $b \in B(x)$ uniformly, in order to be able to represent all possible outputs. Xie and Smidt also emphasize the desideratum that the set $B(x)$ transform equivariantly with x. An *equivariant symmetry breaking set (ESBS)* is thus a G_x-dependent set $B(x)$ such that for all $x \in \mathcal{X}$ and $g \in G$, we have $B(gx) = gB(x)$ and G_x acting freely on $B(x)$. Finally they define an *ideal* ESBS as one in which there is only one element per possible output (i.e. $|B(x)| = |G_x|$), which they argue should make the task of learning f easier. Xie and Smidt show an ideal ESBS may not always exist, deriving a somewhat technical characterization of when it does.

Our work allows us to analyze ESBS through a probabilistic lens. First, our results on noise injection demonstrate that uniformly sampling from an ESBS allows one to represent any equivariant distribution, if one also adds an independent $\epsilon \sim \mathrm{Unif}(0, 1)$ to the input, $f(x, b, \epsilon)$. Additionally, in our methods, $\tilde{g}$ plays the role of a "symmetry breaking element" similar to the elements of an ESBS. In fact, the coset $gG_{\gamma(X)}$ from which $\tilde{g}$ is sampled would satisfy the definition of *ideal* ESBS, except for the fact that it depends on X through more than just its stabilizer. We may thus view the requirement that $B(x)$ depend on x *only* through G_x—which makes it so that an ideal ESBS may not exist—as being overly restrictive.

E. Ising model experimental details

E.1. Ground state of the anisotropic Ising model

We present here an analytical derivation of the Ising model ground states. This is used to obtain the ground-truth values for the average energy in Table 2 and the ground-truth phase diagram in Fig. 2.

The Hamiltonian of the anisotropic Ising model we consider is given by

$$H(\sigma) = -\sum_{i,j} J^x \sigma_i \sigma_j - \sum_{i,j} J^y \sigma_i \sigma_j - \sum_i h \sigma_j \tag{3}$$

We can re-express the Hamiltonian using the variables $b^x_{ij}, b^y_{ij} = [\sigma_i, \sigma_j] \in \{[1, 1], [1, -1], [-1, 1], [-1, -1]\}$ taking values on the horizontal and vertical interaction edges instead of the lattice sites. The variable encodes the value of the spins i and j adjacent to the bond. For convenience, we will write b_{ij} as a one-hot vector, with

$$\mathbf{b}^\alpha_{ij} = \begin{cases} [1, 0, 0, 0] & \text{if } b^\alpha_{ij} = [1, 1] \\ [0, 1, 0, 0] & \text{if } b^\alpha_{ij} = [1, -1] \\ [0, 0, 1, 0] & \text{if } b^\alpha_{ij} = [-1, 1] \\ [0, 0, 0, 1] & \text{if } b^\alpha_{ij} = [-1, -1] \end{cases} \tag{4}$$

In terms of these variables, the Hamiltonian becomes

$$H(\mathbf{b}) = -\sum_{i,j} \begin{bmatrix} J^x \\ -J^x \\ -J^x \\ J^x \end{bmatrix} \mathbf{b}^x_{ij} - \sum_{i,j} \begin{bmatrix} J^y \\ -J^y \\ -J^y \\ J^y \end{bmatrix} \mathbf{b}^y_{ij} - \sum_{i,j} \frac{1}{2} \begin{bmatrix} h \\ 0 \\ 0 \\ -h \end{bmatrix} \left(\mathbf{b}^y_{ij} + \mathbf{b}^x_{ij}\right) \tag{5}$$

For the change of variables to correspond to a valid spin configuration, we additionally need to satisfy the constraints $b_{ij,0} = b_{ij',0}$ and $b_{i'j,1} = b_{ij,1}$ for any i, i', j, j'. Any configuration of variables satisfying such constraint will be an element of the feasible set $\mathcal{B}$.

Finding the ground state therefore corresponds to the optimization problem $\arg\min_{\mathbf{b}\in\mathcal{B}} H(\mathbf{b})$.

Since the form of the Hamiltonian Eq. (6) is a simple sum of non-interacting terms, we first directly minimize to find the ground-states without considering the constraint. We will then verify that the minimum lies in the feasible set $\mathcal{B}$.

We first re-write the Hamiltonian in the following way

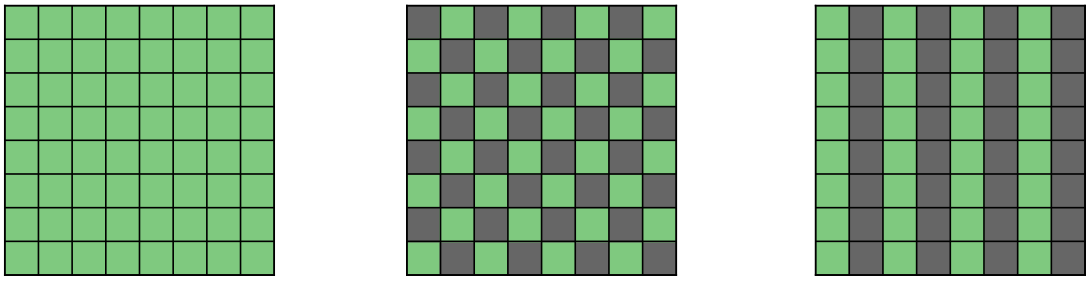
$$H(\mathbf{b}) = -\sum_{i,j} \begin{bmatrix} J^x + h/2 \\ -J^x \\ -J^x \\ J^x - h/2 \end{bmatrix} \mathbf{b}^x_{ij} - \sum_{i,j} \begin{bmatrix} J^y + h/2 \\ -J^y \\ -J^y \\ J^y - h/2 \end{bmatrix} \mathbf{b}^y_{ij} \tag{6}$$

We can then minimize each term independently to obtain the solution

$$\mathbf{b}^\alpha_{ij} = \begin{cases} [1,0,0,0] & \text{if } J^\alpha \geq -\frac{h}{4}, h \geq 0 \\ [0,1,0,0] \text{ or } [0,0,1,0] & \text{if } J^\alpha \leq -\frac{1}{4}|h| \\ [0,0,0,1] & \text{if } J^\alpha \geq \frac{h}{4}, h \leq 0 \end{cases} \tag{7}$$

where $\alpha \in \{x, y\}$. Without loss of generality we consider $h \geq 0$. This leaves us with four possibilities.

1. $J^x \geq -\frac{h}{4}, J^y \geq -\frac{h}{4}$: The ground states is given by $\mathbf{b}^\alpha_{ij} = [1,0,0,0]$, which is in $\mathcal{B}$ and corresponds to the ferromagnetic (FM) phase (see Fig. 5a).
2. $J^x \leq -\frac{h}{4}, J^y \leq -\frac{h}{4}$: Feasible ground states consists of alternating between $\mathbf{b}^\alpha_{ij} = [0,1,0,0]$ and $\mathbf{b}^\alpha_{ij} = [0,0,1,0]$. This is indeed in $\mathcal{B}$ and corresponds to the antiferromagnetic (AFM) phase (see Fig. 5b).
3. $J^x \geq -\frac{h}{4}, J^y \leq -\frac{h}{4}$: Feasible ground states consists of having $\mathbf{b}^x_{ij} = [1,0,0,0]$ and alternating between $\mathbf{b}^y_{ij} = [0,1,0,0]$ and $\mathbf{b}^y_{ij} = [0,0,1,0]$. This is indeed in $\mathcal{B}$ and corresponds to the y stripes phase (S^y) (see Fig. 5c).
4. $J^x \leq -\frac{h}{4}, J^y \geq -\frac{h}{4}$: This is the same as above but with stripes in the x direction.



(a) Ferromagnetic phase (b) Antiferromagnetic phase (c) Stripes phase

Figure 5. Illustration of the different ground-states of the anisotropic Ising model

We can see that the different ground-states are associated with specific types of *orders*. The type of order of an arbitrary spin configuration σ, which is in a sense its closeness to the respective ground-states, can be quantified via *order parameters* (Beekman et al., 2019).

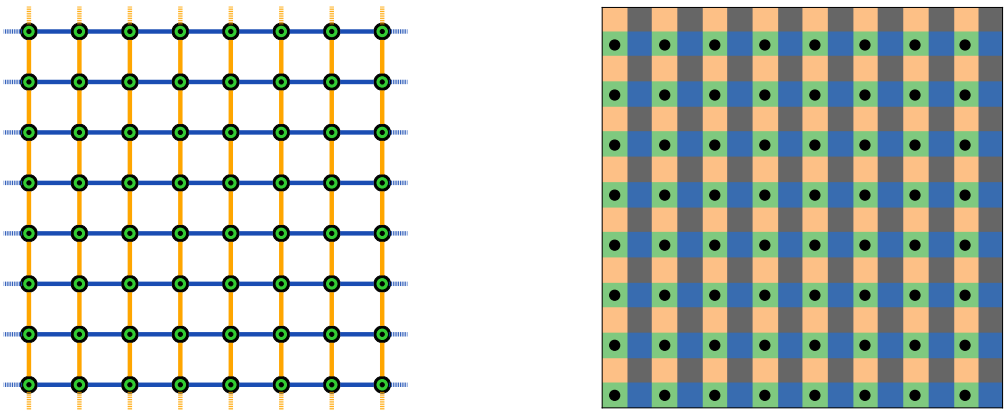
The order parameters are given by

$$O_{\text{FM}} = \frac{1}{N}\sum_i \sigma_i \qquad O_{\text{AFM}} = \frac{1}{N}\sum_i (-1)^{i^x+i^y}\,\sigma_i \qquad O_{S^y} = \frac{1}{N}\sum_i (-1)^{i^y}\,\sigma_i \tag{8}$$

where i^x and i^y are respectively the x and y positions of a spin σ_i. The order parameters take the value 1 if and only if σ is the associated ground-state. In addition, $O_{\text{FM}} + O_{\text{AFM}} + O_{S^y} \leq 1$, which means that the orders are mutually exclusive. This allows us to draw meaningful phase diagrams like the ones in Fig. 2.

E.2. Hamiltonian encoding

For the encoding of the Hamiltonian interaction parameters as input to neural networks, a graph representation would be possible, with interaction parameters J_{ij} encoded as edge attributes and external field values as node attributes (see Fig. 6a). However, this choice would make it much more challenging to use the $p4m$ symmetry of the Hamiltonian. We therefore leverage the structure of the lattice and adopt an image representation, which is complete and unique (see Fig. 6b). We can then conveniently use G-CNNs and MLPs as prediction networks.



(a) Graph encoding (b) Image encoding

Figure 6. Data structures for the encoding of the Hamiltonian interaction parameters. J^x is represented in blue, J^y in yellow and the external field in green. The lattice sites of the spins are represented with black dots. In the image, gray pixels corresponding to the holes between edges, are set to 0.

Through the neural network, we always preserve the size of the image. At the output of the network, we obtain a spin configuration by indexing the image over pixels corresponding to lattice sites (with black dots on figure Fig. 6b). The energy is then computed over the lattice sites.

E.3. Energy results

Table 2. Test set energies of predicted configurations

Method	Average energy	Parameters
G-CNN	-0.73	397K
MLP + aug.	-1.24	3.2M
G-CNN + noise	-1.29	399K
G-CNN + SB (Ours)	-1.47	468K
Ground truth	-1.60	-

E.4. Training setup

Unsupervised training is performed by having the neural network $\phi : \mathcal{J} \to [0,1]^\Lambda$ output the probability that each spin is up by applying a softmax on the last layer; symmetry breaking elements are computed through a deterministic canonicalization given by a G-CNN as in Kaba et al. (2023). We use the EquiAdapt library (Mondal et al., 2023) to implement the canonicalization. While training, we then compute the expectation value of the spin at each site and treat this as the configuration, using the Hamiltonian as a loss function. At evaluation, we sample a spin value for each site using the probability output from the neural network.

We define a training set of Hamiltonian parameters as $J^x = -1$ and sampling $J^y \sim \text{Unif}\,(-3,3)$, $h \sim \text{Unif}\,(0,2)$, with parameters constant over the lattice. The training set is of size 1000. For the validation and test sets, we use $10,000$ regularly sampled Hamiltonian parameter values in the same range.

F. Graph autoencoder details

As discussed in the main body, autoencoders with "very symmetric" latent spaces pose a problem for equivariant models. In the graph auto-encoding setup of Satorras et al. (2021), self-symmetry of a graph with adjacency matrix A arises in the form of its automorphism group, $\{g \in S_n : \ gAg^T = A\}$. For example, the square graph A in Figure 3 has C_4 automorphism group. When using a permutation-equivariant encoder e, $e(A)$ must therefore also have at least C_4 symmetry. Even though the goal of the permutation-equivariant decoder d is to satisfy $d(e(A)) = A$, which would not seem to explicitly require symmetry-breaking, there is a subtle problem: the choice of latent space $\mathcal{Z}$. Precisely, $\mathcal{Z}$ does not contain any node-wise featurization with *precisely* C_4-symmetry: if $\{z_1, \dots, z_4\} \in \mathcal{Z}$ have C_4-symmetry, this implies that $z_1 = \cdots = z_4$.[5] More abstractly, there exist graphs A, with stabilizer (automorphism group) G_A, such that there is no $Z \in \mathcal{Z}$ with $G_Z = G_A$; only $G_A \subsetneq G_Z$. Thus, for these graphs, *any* equivariant encoder will produce an "overly self-symmetric" latent embedding. To reconstruct A, we must therefore break the latent-space symmetry induced by the equivariant encoder.[6]

In our experiments, we follow the setup of Satorras et al. (2021) and break symmetries at the input. Satorras et al. (2021) suggest doing this with random noise as initial node features. However, this breaks *all* permutational symmetry, not just the graph's automorphism group. In contrast, our method is still equivariant to permutations of nodes that are automorphically equivalent. More precisely, we use Laplacian positional encodings to break symmetries of the input graph. Although it is unknown whether detecting graph automorphisms is NP-hard in general, we expect this heuristic to do a reasonable job of interpolating between breaking all permutation symmetry, and falling victim to Curie's principle.

We also include some additional baselines in Table 3, none of which work as well as our proposed method. In both this table and in the main body, % Error is the same metric as used in Satorras et al. (2021).

Table 3. Cross-entropy loss and reconstruction error

Method	**BCE**	**% Error**	**Number of parameters**
Both Laplacian canonicalization and noise injection	12	3.0	88,885
Break symmetry with Laplacian canonicalization, just 1 channel	25.0	6.8	88,239

G. Proofs

The main theorem and its corollaries hold for any G acting *properly* on $\mathcal{X}$, which includes compact groups and actions on Riemannian manifolds by isometries. For more details, including the relationship of proper actions to needed measurability conditions, we refer the reader to Chiu & Bloem-Reddy (2023).

G.1. Proof of Theorem 2.3

Proof. $Y|X$ is equivariant if and only if (Chiu & Bloem-Reddy, 2023, Theorem 3)

$$(\tilde{g}, X) \perp \tilde{g}^{-1}Y \mid \gamma(X) \tag{9}$$

[5] The possible stabilizers of $\mathcal{Z}$, the latent space of node-wise featurizations, is groups of the form $S_{i_1} \times \dots S_{i_k}$, where $i_1 + \cdots + i_k = n$.

[6] Of course, one could also use a different latent space, such as a latent space of matrices. However, this may defeat the purpose of learning an expressive, dimensionality-reduced latent space.

where $\tilde{g}|X$ is distributed according to the inversion kernel associated to γ. By conditional noise outsourcing (Kallenberg, 2021, Proposition 8.20) this is equivalent to there existing a measurable function $\phi : \mathcal{X} \times (0,1) \to \mathcal{Y}$ such that

$$\tilde{g}^{-1}Y \overset{a.s.}{=} \phi(\gamma(X), \epsilon) \tag{10}$$

where $\epsilon \sim \mathrm{Unif}(0,1)$ is independent of $(\tilde{g}, X)$. Rearranging and noting $\tilde{g}\gamma(X) \overset{a.s.}{=} X$ gives the result. □

G.2. Proof of Corollary 2.4

Proof. The forward implication is clear, letting $f(x, g, \epsilon) = g\phi(g^{-1}x, \epsilon)$ with ϕ given in the previous theorem. On the other hand consider a function $f : \mathcal{X} \times G \times (0,1)$ such that $hf(x, g, \epsilon) = f(hx, hg, \epsilon)$, and suppose $Y \overset{a.s.}{=} f(X, \tilde{g}, \epsilon)$. Write $\tilde{g}_x$ for a random variable sampled from the inversion kernel conditioned on $X = x$, independently from $\epsilon \sim \mathrm{Unif}(0,1)$. For any $x \in \mathcal{X}$ and $h \in G$ we have

$$\mathbb{P}\left(hY \in B | hX = x\right) = \mathbb{P}\left(hY \in B | X = h^{-1}x\right) = \mathbb{P}\left(hf(h^{-1}x, \tilde{g}_{h^{-1}x}, \epsilon) \in B\right). \tag{11}$$

By the equivariance of the inversion kernel, $\tilde{g}_{h^{-1}x} \overset{d}{=} h^{-1}\tilde{g}_x$. Applying first this fact and then the equivariance of f we obtain

$$\mathbb{P}\left(hY \in B | hX = x\right) = \mathbb{P}\left(hf(h^{-1}x, h^{-1}\tilde{g}_x, \epsilon) \in B\right) = \mathbb{P}\left(f(x, \tilde{g}_x, \epsilon) \in B\right) = \mathbb{P}\left(Y \in B | X = x\right). \tag{12}$$

Note that for this direction, we only needed the equivariance of the distribution $\tilde{g}|X$, not its restriction to a specific coset. □

G.3. Proof of Corollary 2.6

Proof. Given ϕ from the previous theorem, let f be a relaxed equivariant function such that $f(\gamma(x), \epsilon) = \phi(\gamma(x), \epsilon)$ (which we can always do; see Appendix J). Then we have

$$Y \overset{a.s.}{=} \tilde{g}\phi(\gamma(X), \epsilon) = \tilde{g}f(\gamma(X), \epsilon) = \tilde{g}h^{-1}f(X, \epsilon) \tag{13}$$

for some $h \in G$ (depending on the construction of f) such that $h\gamma(X) = X$. But $\tilde{g} \sim \mathrm{Unif}(hG_{\gamma(X)})$ and $hG_\gamma(X)h^{-1} = G_X$, so $\tilde{g}h^{-1} \sim \mathrm{Unif}(G_X)$.

Suppose on the other hand $Y \overset{a.s.}{=} g_X f(X, \epsilon)$ for a relaxed equivariant f. Write g_x for a uniform random element of G_x independent of $\epsilon \sim \mathrm{Unif}(0,1)$. We then have

$$\mathbb{P}\left(hY \in B | hX = x\right) = \mathbb{P}\left(hY \in B | X = h^{-1}x\right) = \mathbb{P}\left(hg_{h^{-1}x}f(h^{-1}x, \epsilon) \in B\right). \tag{14}$$

Since $g_{h^{-1}x} \overset{d}{=} h^{-1}g_x h$,

$$\mathbb{P}\left(hY \in B | X = h^{-1}x\right) = \mathbb{P}\left(g_x hf(h^{-1}x, \epsilon) \in B\right). \tag{15}$$

By the relaxed equivariance of f, for some $k^{-1} \in h^{-1}G_x$, the above is equal to

$$\mathbb{P}\left(g_x hk^{-1}f(x, \epsilon) \in B\right). \tag{16}$$

Noting that $hk^{-1} \in G_x$, we have $g_x \overset{d}{=} g_x hk^{-1}$ and thus

$$\mathbb{P}\left(hY \in B | hX = x\right) = \mathbb{P}\left(Y \in B | X = x\right). \tag{17}$$

(The invariance of the distribution of $g_X|X$ is used here, though what is strictly needed is equality in distribution of $hg_{h^{-1}x}k^{-1}$ and g_x.) □

G.4. Proof of Proposition 3.1

Proof. Consider any jointly equivariant $f : \mathcal{X} \times G \to \mathcal{Y}$, and suppose $v \in V$ has trivial stabilizer. We may let $f_0(x \oplus u) = f(x, g)$ where g is the unique group element such that $gv = u$. □

G.5. Proof of Proposition 3.2

We decompose the forward and reverse implications in the following four statements:

1. If the action of G on $\mathcal{Z}$ is free except for a measure zero subset, then equivariance of $Y|X$ implies that $Y \overset{a.s.}{=} f(X, Z, \epsilon)$ for an appropriate f and ϵ. Note that equivariance of $Z|X$ is not needed here.

2. If the action of G on $\mathcal{Z}$ is free except for a measure zero subset and $Z|X$ is equivariant, then $Y \overset{a.s.}{=} f(X, Z, \epsilon)$ implies equivariance of $Y|X$.

3. If the action of G on $\mathcal{X} \times \mathcal{Z}$ is *not* free except for a measure zero subset, then $Y \overset{a.s.}{=} f(X, Z, \epsilon)$ is *not* equivalent to equivariance of $Y|X$

4. If $Z|X$ is *not* equivariant, then $Y \overset{a.s.}{=} f(X, Z, \epsilon)$ is *not* equivalent to equivariance of $Y|X$

Proof of 1. Let $\tilde{g}|X$ be a group element distributed according to the inversion kernel. Reusing arguments from our main results we show there exists a map $t : \mathcal{X} \times \mathcal{Z} \times (0, 1) \to G$ equivariant in the first two inputs such that $\tilde{g} \overset{a.s.}{=} t(X, Z, \eta)$ where $\eta \sim \mathrm{Unif}(0, 1)$ is random noise independent of X and Z. By Corollary 2.4 $Y \overset{a.s.}{=} f_0(X, t(X, Z, \eta), \epsilon_0)$ for a jointly equivariant f_0. The unit interval and unit square both being standard probability spaces, there exists a measure preserving bijection $\epsilon \leftrightarrow (\eta, \epsilon_0)$. We thus define $f(X, Z, \epsilon) = f_0(X, t(X, Z, \eta), \epsilon_0)$.

To show the existence of t, one repeats the proof of Theorem 2.3 and Corollary 2.4, but applying Chiu & Bloem-Reddy (2023, Theorem 3) in the special case of an essentially free action. In particular, since (X, Z) has trivial stabilizer almost surely,

$$\tilde{g}^{-1}Y \perp (X, Z) \mid \gamma(X, Z), \tag{18}$$

where $\gamma(X, Z) = (\gamma(X), Z)$ (with a slight abuse of notation). The rest of the proof is identical to that of the forward direction of Corollary 2.4. □

Proof of 2. The proof is identical to that of the reverse direction of Corollary 2.4. □

Proof of 3. Suppose G does not act on $\mathcal{Z}$ essentially freely. Suppose $Y|X$ is equivariant and there is symmetry breaking with non-zero probability, i.e. $\mathbb{P}(G_{X,Z} \not\subseteq G_Y) > 0$. (Such examples are not hard to construct.) Suppose for the sake of contraction that also $Y \overset{a.s.}{=} f(X, Z, \epsilon)$ with some f and ϵ as usual. Curie's principle gives the contradiction: $G_{X,Z} \subseteq G_Y$ at any realization of ϵ. □

Proof of 4. We consider non-equivariant $Z|X$; concretely, suppose Z is constant $z \in \mathcal{Z}$, and that there exists $g \in G$ such that $gz \neq z$. Let $f : \mathcal{X} \times \mathcal{Z} \times (0, 1) \to \mathcal{Y}$ be the jointly equivariant function $f(x, z, \epsilon) = x \oplus z$ (so $\mathcal{Y} = \mathcal{X} \times \mathcal{Z}$). Then if $Y \overset{a.s.}{=} f(X, z, \epsilon)$,

$$\mathbb{P}(gY = y|gX = x) = \mathbb{P}\left(gf(X, z, \epsilon) = y|X = g^{-1}x\right) = \mathbb{1}\{x \oplus gz = y\} \tag{19}$$

$$\neq \mathbb{1}\{x \oplus z = y\} = \mathbb{P}(Y = y|X = x). \tag{20}$$

That is, $Y|X$ is not equivariant. □

H. Breaking symmetry in different groups

In this appendix, we show how to learn symmetry breaking biases $v \in V$ (such that G_v is trivial) for different groups. For any group G, the first basic requirement is for the group to act faithfully on V. We then want to choose V such that we can initialize $v \in V$ and obtain G_v with probability 1 (with the assumption that v is initialized by sampling from an absolutely continuous distribution).

H.1. Permutation groups

For permutation groups such as S_n and $p4m$ (the symmetry group of an image grid), the group admits a faithful representation that maps to permutation matrices acting on $\mathbb{R}^n$. We can therefore choose $V = \mathbb{R}^n$.

In this case, it suffices that all the elements of a symmetry breaking bias v are different. This is captured by the following proposition.

Proposition H.1. *Let G act faithfully by permutation on $\mathbb{R}^n$ and $v \in \mathbb{R}^n$ be such that $v_i \neq v_j$ for any $i \neq j$. Then G_v is trivial. In addition, the set of v not satisfying this condition is of measure zero with respect to the Lebesgue measure.*

Proof. The proof of the first part of the proposition is trivial. For any $g \in G$, $(gv)_i = v_{g^{-1}(i)}$ Since the group action is faithful, for any $g \in G$ except the identity, there is an $i \in [n]$ such that $g^{-1}(i) \neq i$. For this i, $v_{g^{-1}(i)} \neq v_i$, therefore G_v is trivial.

The second part follows the same ideas as Proposition 3 of (Kaba & Ravanbakhsh, 2023). Consider an hyperplane in $\mathbb{R}^n$, defined as $H_{ij} = \{v \in \mathbb{R}^n \mid v_i = v_j\}$. Any v not satisfying the condition is an element of $S = \cup_{i \neq j}^{n} H_{ij}$. An hyperplane H_{ij} defines an $(n-1)$-dimensional space in $\mathbb{R}^n$. Any subspace of $\mathbb{R}^n$ of dimension strictly less than n is of measure zero. The countable union of such subspaces S is also of measure zero. □

H.2. Subgroups of the general linear group

We also consider groups that admit faithful representations as subgroups of $GL(n)$, such as $O(n)$ for point clouds and atomic systems. In this case, we can choose the symmetry breaking bias to be n linearly independent vectors, so that $V = \mathbb{R}^{n \times n}$.

Proposition H.2. *Let $G \subseteq GL(n)$ and $V = \mathbb{R}^{n \times n}$. Assume G acts on V as a product of faithful actions, e.g. $gv \mapsto \left[gv^1, \dots, gv^n\right]$, where v^i is the i-th column vector of v. If the vectors $\left[v^1, \dots, v^n\right]$ are linearly independent, then G_v is trivial. In addition, the set of v such that this condition is not satisfied is of measure zero with respect to the Lebesgue measure.*

Proof. For the first part, we can identify the action of the group and matrix multiplication gv. For any $g \in G_v$, it must be that $gv = v$. Since the columns of v are linearly independent, v is invertible. We therefore have $gvv^{-1} = vv^{-1}$, which implies $g = I$.

For the second part, the idea is similar to the proof of Proposition H.1 above. If two columns of v are not linearly independent, it implies that v is element of a subspace $H_{ij} = \left\{v \in \mathbb{R}^{n \times n} \mid \exists a \in \mathbb{R} \text{ s.t. } v^i = av^j\right\}$ for some i and j. This is a subspace of measure zero since it is of dimension $n^2 - n + 1$. The union of all such subspaces $S = \cup_{i \neq j}^{n} H_{ij}$ is also of measure zero since it is countable. □

I. Generalization benefits of equivariant conditional distributions

Here, we are interested in describing the generalization benefits of using equivariant conditional distributions, when the ground truth distribution $Y|X$ is equivariant. The general outline, which we will fill in below, follows the work of Elesedy & Zaidi (2021) and its straightforward generalization in Elesedy (2023). We similarly assume that X has a G-invariant distribution, that $Y|X$ is equivariant, and the group is compact.[7] Then, considering a Hilbert space of conditional distributions (or rather, their unnormalized counterparts), one can show that the equivariant ones form a subspace. The generalization benefits of assuming equivariance can then be expressed in terms of the projection operator onto that subspace.

In order to follow the program above, we will treat conditional distributions as equivariant functions $f : \mathcal{X} \to \mathbb{P}(\mathcal{Y})$. Furthermore, in order to work in a Hilbert space, we restrict ourselves to distributions with a square integrable density with respect to some measure dy on $\mathcal{Y}$. We assume the measure is invariant under G—the canonical example being $\mathcal{Y}$ a finite-dimensional Euclidean space with Lebesgue measure and G acting orthogonally. Then, we define $\mathbb{P}(\mathcal{Y}) \subset L^2(\mathcal{Y})$ by

$$\mathbb{P}(\mathcal{Y}) = \left\{\psi : \mathcal{Y} \to \mathbb{R} \; s.t. \; \psi(\cdot) \geq 0, \; \int_{y \in \mathcal{Y}} \psi(y) \, dy = 1, \; \int_{y \in \mathcal{Y}} \psi(y)^2 \, dy < \infty\right\} \tag{21}$$

[7] Our previous results held in the more general case of proper group actions.

We will treat densities $p(y|x)$ as members of the Hilbert space $\mathcal{H} = L^2(\mathcal{X}, L^2(\mathcal{Y}), \mathbb{P})$, where the data distribution $\mathbb{P}$ on $\mathcal{X}$ is G-invariant. The inner product between functions $f_1, f_2 \in \mathcal{H}$ is

$$\langle f_1, f_2 \rangle = \int_{\mathcal{X}} \langle f_1(x), f_2(x) \rangle \, \mathbb{P}(dx) = \int_{\mathcal{X}} \int_{\mathcal{Y}} f_1(y|x) f_2(y|x) \, dy \, \mathbb{P}(dx),$$

where for simplicity we use the notation $f(y|x) = (f(x))(y)$ for even those $f \in \mathcal{H}$ which are not probability densities.

G acts on functions $\psi \in L^2(\mathcal{Y})$—and thus on $\mathbb{P}(\mathcal{Y})$—by

$$(g \cdot \psi)(y) = \psi(g^{-1} y).$$

By the assumption of the G-invariance of dy, the inner product on $L^2(\mathcal{Y})$ is invariant: $\langle \psi_1, \psi_2 \rangle = \langle g\psi_1, g\psi_2 \rangle$. Standard arguments from Elesedy (2023, Lemma 3.1) then show that the *Reynolds operator* $\mathcal{R} : \mathcal{H} \to \mathcal{H}$ given by[8]

$$(\mathcal{R} f)(x) = \int_G g^{-1} f(gx) \, \lambda(dg) \tag{22}$$

$$(\mathcal{R} f)(y|x) = \int_G f(gy|gx) \, \lambda(dg) \tag{23}$$

is in fact the orthogonal projection onto the subspace of equivariant functions—i.e. those $f \in \mathcal{H}$ such that $f(g^{-1}y|x) = f(y|gx)$. For the Reynolds operator to be useful to us, it remains to check that it sends normalized conditional densities $p(y|x)$ to normalized conditional densities (which will then be equivariant). But this is clear:

$$\int_{\mathcal{Y}} (\mathcal{R}p)(y|x) \, dy = \int_{\mathcal{Y}} \int_G p(gy|gx) \, \lambda(dg) \, dy = \int_G \int_{\mathcal{Y}} p(gy|gx) \, dy \, \lambda(dg) = 1, \tag{24}$$

where at the end we used the invariance of dy, and the fact that $p(y|x)$ and λ are normalized.

We then consider risk under the $L^2(\mathcal{Y})$ loss. The *generalization gap* between two conditionals $p_1, p_2 \in \mathcal{H}$

$$\Delta(p_1, p_2) = R(p_1) - R(p_2) \tag{25}$$

where R is the risk as measured against the ground truth conditional p^*,

$$R(p) = \int_{\mathcal{X}} ||p(x) - p^*(x)||^2_{L^2(Y)} \, \mathbb{P}(dx) = \int_{\mathcal{X}} \int_{\mathcal{Y}} (p(y|x) - p^*(y|x))^2 \, dy \, \mathbb{P}(dx). \tag{26}$$

We can rewrite that risk as

$$R(p) = ||p - p^*||^2_{\mathcal{H}} = ||p||^2_{\mathcal{H}} - 2\langle p, p^* \rangle_{\mathcal{H}} + ||p^*||^2_{\mathcal{H}}. \tag{27}$$

(We subsequently drop the subscripts for conciseness.) The generalization gap between an arbitrary $p \in \mathcal{H}$ and its equivariant projection $\bar{p} = \mathcal{R}p$ is then given in terms of the orthogonal component $p^\perp = p - \bar{p}$:

$$\Delta(p, \bar{p}) = ||p - p^*||^2 - ||\bar{p} - p^*||^2 \tag{28}$$

$$= ||\bar{p} + p^\perp - p^*||^2 - ||\bar{p} - p^*||^2 \tag{29}$$

$$= ||\bar{p} - p^*||^2 + ||p^\perp||^2 - ||\bar{p} - p^*||^2 = ||p^\perp||^2 \tag{30}$$

where one gets to the the last line by using the orthogonality of $p^\perp$ and $\bar{p}, p^* \in \mathcal{H}_G$.

J. Generalization benefits of relaxed equivariance

We begin by expressing relaxed equivariance more formally, in terms of choices of coset representatives. We then show that relaxed equivariant functions which make the same such choices form a subspace, allowing the orthogonal projection arguments of Elesedy & Zaidi (2021) to be generalized. Furthermore we show that this projection decomposes across different *orbit types* (defined below—loosely, the kinds of stabilizers appearing in the data).

We obtain generalization benefits of relaxed equivariant models much in the same way as for equivariant models, where the improvement now depends on the frequency of the different stabilizers appearing in the data.

[8] Recall that we use λ to denote the (normalized) Haar measure on G.

J.1. Construction of non-trivial relaxed equivariant functions

Recall our that our definition of a relaxed equivariant function $f : \mathcal{X} \to \mathcal{Y}$ is one such that for any $x \in \mathcal{X}$ and $g \in G$, there is a $h \in gG_x$ such that

$$hf(x) = f(gx). \tag{31}$$

This definition is implicit, in the sense that it says nothing about how h is chosen. It is not even clear a priori that (non-equivariant) relaxed equivariant functions exist.

Kaba & Ravanbakhsh (2023, Lemma 9) show that for a fixed choice of orbit representative x_0 (which we prefer here to $\gamma(x)$ for clarity), one can extend a choice of coset representatives $\{\sigma(gG_{x_0}) : g \in G\}$ from cosets of G_{x_0} to the entire orbit: $\sigma(gG_x)$ for any $x \in [x_0]$. The goal is that if some function f is "relaxed equivariant" only on x_0, in the sense that

$$\sigma(gG_{x_0})f(x_0) = f(gx_0)\ \forall g \in G, \tag{32}$$

then in fact for any $x \in Gx_0$ we have the same:

$$\sigma(gG_x)f(x) = f(gx)\ \forall g \in G. \tag{33}$$

An arbitrary choice of coset representatives $\sigma(gG_{x_0})$ for a single x_0 is extended to the entire orbit by defining

$$\sigma(gG_x) := \sigma(ghG_{x_0})\sigma(G_{x_0})\sigma(hG_{x_0})^{-1} \tag{34}$$

for every $x = hx_0$. To see σ is indeed well defined, note that if $x = h'x_0$ (where we think of $h' \neq h$), then $h'G_{x_0} = hG_{x_0}$ and $gh'G_{x_0} = ghG_{x_0}$, and thus the right hand side above remains unchanged; additionally, if $gG_x = kG_x$ then $k = gs$ for some $s \in G_x$ and we have that $\sigma(khG_{x_0}) = \sigma(gshG_{x_0}) = \sigma(ghG_{x_0})$. Furthermore, this definition is consistent in the sense that for $x = x_0$ it recovers $\sigma(gG_x) = \sigma(gG_{x_0})$.

Proposition J.1 (c.f. Lemma 9 of Kaba & Ravanbakhsh (2023))**.** *If f is relaxed equivariant at x_0, for any $x = hx_0$, we have*

$$\sigma(gG_x)f(x) = f(gx)\ \forall g \in G. \tag{33 revisited}$$

Proof. We expand first by the assumed relaxed equivariance at x_0 (32),

$$f(gx) = f(ghx_0) \tag{35}$$
$$= \sigma(ghG_{x_0})f(x_0). \tag{36}$$

We then use the observation that $\sigma(G_{x_0})$ must fix $f(x_0)$ to write

$$f(gx) = \sigma(ghG_{x_0})\sigma(G_{x_0})f(x_0). \tag{37}$$

Using (32) again to write $f(x_0)$ as $\sigma(hG_{x_0})^{-1}f(hx_0)$, we finally have

$$f(gx) = \sigma(ghG_{x_0})\sigma(G_{x_0})\sigma(hG_{x_0})^{-1}f(hx_0) \tag{38}$$
$$:= \sigma(gG_x)f(x), \tag{39}$$

by our definition (34). □

Remark J.2. By Curie's principle, the initial choice of coset representatives for x_0 cannot be fully arbitrary. For all $g \in G_{x_0}$,

$$\sigma(G_{x_0})f(x_0) = \sigma(gG_{x_0})f(x_0) = f(gx_0) = f(x_0), \tag{40}$$

that is, $\sigma(G_{x_0}) \in G_{f(x_0)}$. One can satisfy this constraint trivially by letting the representative of the stabilizer be the identity: $\sigma(G_{x_0}) = \mathrm{id}$. This corresponds to "full symmetry breaking," i.e. no symmetry of the input x_0 is enforced on the output $f(x_0)$, and $\sigma(G_x) = \mathrm{id}$ for all x. Any other choice leads to "partial symmetry breaking." Of course, $\sigma(G_{x_0})$ can only generate a one-generator subgroup of $G_{f(x_0)}$. Thus, we cannot impose that f have more complex partial symmetry breaking only by enforcing (32).

J.2. A Reynolds operator for relaxed equivariance

We restrict ourselves to the setting of a compact group G, with Haar measure λ, and identity stabilizer representatives $\sigma(G_x) = \mathrm{id}$. We define a *generalized Reynolds operator* by

$$(\mathcal{R}f)(x) := \int_G \sigma(gG_x)^{-1} f(gx)\ \lambda(dg) = \int_G \sigma(g^{-1}G_{gx}) f(gx)\ \lambda(dg) \tag{41}$$

Suppose our input data X comes from a G-invariant distribution $\mathbb{P}(dx)$ on $\mathcal{X}$ and that G acts orthogonally on $\mathcal{Y}$, and consider the Hilbert space $\mathcal{H} = L^2((\mathcal{X}, \mathbb{P}), \mathcal{Y})$. We prove below that $\mathcal{R}$ an orthogonal projection onto the subspace of $\mathcal{H}$ consisting of relaxed equivariant functions corresponding to a certain choice of coset representatives $\sigma(gG_x)$. (When referring to relaxed equivariant functions we will mean these in particular.) After showing it is well-defined as a map $\mathcal{H} \to \mathcal{H}$, we prove its image consists of relaxed equivariant functions; then that is it a projection, and indeed and orthogonal one.

To see $\mathcal{R} : \mathcal{H} \to \mathcal{H}$, we mimic the proof of Elesedy & Zaidi (2021, Proposition 23):

$$\begin{aligned}
||\mathcal{R}f||^2 &= \mathbb{E}_X \left[\left|\left| \int_G \sigma(gG_X)^{-1} f(gX) \lambda(dg) \right|\right|^2 \right] \\
&\leq \mathbb{E}_X \left[\int_G \left|\left| \sigma(gG_X)^{-1} f(gX) \right|\right|^2 \lambda(dg) \right] \text{ by Jensen's inequality} \\
&= \mathbb{E}_X \left[\int_G ||f(gX)||^2 \lambda(dg) \right] \text{ by orthogonality} \\
&= \mathbb{E}_X \left[||f(X)||^2 \right] \text{ invariance of } \lambda \\
&= ||f||^2 < \infty.
\end{aligned}$$

We proceed.

Proposition J.3. *The generalized Reynolds operator maps functions to relaxed equivariant functions.*

The proof of the above relies on the following identities.

Lemma J.4. *Suppose $\sigma(G_{x_0}) = \mathrm{id}$ ("full symmetry breaking"). Then $\sigma(G_x) = \mathrm{id}$ for all $x \in [x_0]$, and for any $g, h \in G$ and $x \in [x_0]$ we have*

$$\sigma(ghG_x) = \sigma(gG_{hx})\sigma(hG_x) \tag{42}$$

$$\sigma(gG_x)^{-1} = \sigma(g^{-1}G_{gx}). \tag{43}$$

Proof. Nothing more than definition chasing. Letting $x = kx_0$, on the left hand side we have

$$\sigma(ghG_x) := \sigma(ghkG_{x_0})\sigma(kG_{x_0})^{-1}, \tag{44}$$

and on the right (since $hx = hkx_0$),

$$\sigma(gG_{hx})\sigma(hG_x) := \sigma(ghkG_{x_0})\sigma(hkG_{x_0})^{-1}\sigma(hkG_{x_0})\sigma(kG_{x_0})^{-1}, \tag{45}$$

so the two are equal. The second identity is the first noting $\sigma(g^{-1}G_{gx})\sigma(gG_x) = \sigma(G_x) = \mathrm{id}$. □

Proof of Proposition. For any $h \in G$ and $x \in G_{x_0}$, we have

$$(\mathcal{R}f)(hx) := \int_G \sigma(g^{-1}G_{ghx}) f(ghx)\ \lambda(dg) \tag{46}$$

$$= \int_G \sigma(hg^{-1}G_{gx}) f(gx)\ \lambda(dg), \tag{47}$$

using the invariance of λ when applying a right shift by h^{-1} to g. But (42) gives us

$$\sigma(hg^{-1}G_{gx}) = \sigma(hG_x)\sigma(g^{-1}G_{gx}). \tag{48}$$

Therefore,

$$(\mathcal{R}f)(hx) = \sigma(hG_x)\int_G \sigma(g^{-1}G_{gx})f(gx)\ \lambda(dg) = \sigma(hG_x)\cdot(\mathcal{R}f)(x). \tag{49}$$

□

Remark J.5. It turns out the construction above outputs relaxed equivariant functions without assuming λ is a Haar measure (though this assumption is necessary later in order to show $\mathcal{R}$ is an orthogonal projection). In particular, one can define for any function f and point x the function $f_x : gx \mapsto \sigma(gG_x)f(x)$, which "spreads" the value $f(x)$ along the orbit of x in an equivariant way. It is relaxed equivariant. Noticing that $f_{gx}(x) := \sigma(g^{-1}G_{gx})f(gx)$, the map taking x to $\int_G f_{gx}(x)dg$ defines $\mathcal{R}f$ for any measure on G, which is relaxed equivariant as the integral of relaxed equivariant functions.

To check that $\mathcal{R}$ is a (not necessarily orthogonal) projection, it suffices to show that if f is already relaxed equivariant for some choice of σ (as defined above), then $\mathcal{R}f = f$. In this case, by definition of relaxed equivariance, we have that $f(gx) = \sigma(gG_x)f(x)$. Therefore,

$$\int_G \sigma(gG_x)^{-1}f(gx)\ \lambda(dg) = \int_G \sigma(gG_x)^{-1}\sigma(gG_x)f(x)\ \lambda(dg) = \int_G f(x)\ \lambda(dg) = f(x),$$

and $(\mathcal{R}f)(x) = f(x)$. Note the invariance of λ was not used. Finally, to show $\mathcal{R}$ is an *orthogonal* projection:

$$\begin{aligned}
\langle \mathcal{R}f_1, f_2\rangle &= \int_{\mathcal{X}} \left\langle \int_G \sigma(gG_x)^{-1}f_1(gx)\ \lambda(dg), f_2(x)\right\rangle \mathbb{P}(dx)\\
&= \int_{\mathcal{X}} \left\langle \int_G \sigma(g^{-1}G_{gx})f_1(gx)\ \lambda(dg), f_2(x)\right\rangle \mathbb{P}(dx)\\
&= \int_{\mathcal{X}}\int_G \left\langle \sigma(g^{-1}G_{gx})f_1(gx), f_2(x)\right\rangle\ \lambda(dg)\ \mathbb{P}(dx)\\
&= \int_G\int_{\mathcal{X}} \left\langle \sigma(g^{-1}G_x)f_1(x), f_2(g^{-1}x)\right\rangle\ \mathbb{P}(dx)\ \lambda(dg) \text{ by invariance of } \mathbb{P}(dx)\\
&= \int_G\int_{\mathcal{X}} \left\langle f_1(x), \sigma(gG_{g^{-1}x})f_2(g^{-1}x)\right\rangle\ \mathbb{P}(dx)\ \lambda(dg) \text{by orthogonality and Eq. (42)}\\
&= \int_G\int_{\mathcal{X}} \left\langle f_1(x), \sigma(g^{-1}G_{gx})f_2(gx)\right\rangle\ \mathbb{P}(dx)\ \lambda(dg) \text{ by } \lambda(g) = \lambda(g^{-1}) \text{ for unimodular groups}\\
&= \int_{\mathcal{X}} \left\langle f_1(z), \int_G \sigma(g^{-1}G_{gx})f_2(gx)\ \lambda(dg)\right\rangle \mathbb{P}(dx)\\
&= \int_{\mathcal{X}} \langle f_1(x), \mathcal{R}f_2(x)\rangle\\
&= \langle f_1, \mathcal{R}f_2\rangle.
\end{aligned}$$

J.3. Orbit type decomposition of relaxed equivariance

The moral of this section is that relaxed equivariance looks almost like equivariance in the usual sense, conditioned on the *orbit type* (see below). We will illustrate the sense in which the (generalized) Reynolds operator can be considered independently per orbit type. We then move on to show that—similarly to the orthogonal decomposition of functions into a relaxed equivariant part and its complement—relaxed equivariant functions decompose orthogonally into functions supported on each orbit type.

Definition J.6. The stabilizer G_x transforms under conjugation; that is $G_{gx} = gG_xg^{-1}$. Its orbit under conjugation is thus the set of stabilizers $[G_x] = \{G_{gx} : g \in G\}$, known as the *orbit type* of x.

Under general conditions, for a compact group G acting on $\mathcal{X}$ there are countably many orbit types. Suppose we index these by (non-conjugate) subgroups $H_s \leq G$, $s \in \mathbb{N}$. Then any x has a stabilizer G_x in one of the classes thus indexed: $G_x \in [H_s]$ for some $s \in \mathbb{N}$. We will assume orbit representatives x_0 are chosen such that each $G_{x_0} = H_s$ for some $s \in \mathbb{N}$

The reason we say relaxed equivariance looks almost like equivariance is that symmetrization occurs only with respect to coset representatives:[9]

$$\begin{aligned}(\mathcal{R}f)(x) &= \frac{1}{|G|}\sum_{g\in G}\sigma(gG_x)^{-1}f(gx)\\ &= \frac{1}{|G|}\sum_{g\in G}\sigma(gG_x)^{-1}f(\sigma(gG_x)x)\\ &= \frac{|G_x|}{|G|}\sum_{gG_x\in G/G_x}\sigma(gG_x)^{-1}f(\sigma(gG_x)x)\\ &= \frac{1}{[x]}\sum_{g\in\sigma(G/G_x)}g^{-1}f(gx).\end{aligned}$$

The normalizer is re-expressed by the orbit-stabilizer theorem, showing it does not depend on G_x, only on the orbit type of x. However, the set of representatives $\sigma(G/G_x)$ generally depends not only on the orbit type, but on the stabilizer G_x itself. But for some $k \in G$ and the orbit representative x_0, we have $x = kx_0$ and $G_{x_0} = H_s$, and thus

$$\begin{aligned}(\mathcal{R}f)(x) &= (\mathcal{R}f)(kx_0)\\ &= \sigma(kG_{x_0})\cdot(\mathcal{R}f)(x_0)\\ &= \sigma(kH_s)\cdot\left(\frac{|H_s|}{|G|}\sum_{gH_s\in\sigma(G/H_s)}\sigma(gH_s)^{-1}f\circ\sigma(gH_s)\right)(x_0)\end{aligned}$$

Using the above we see the space S of relaxed equivariant functions decomposes as the direct sum

$$S = \bigoplus_{s\in\mathbb{N}} S_s$$

where

$$\begin{aligned}S_s &:= \{f : G_x \notin [H_s] \Rightarrow f(x) = 0,\ \sigma(gG_x)f(x) = f(gx)\}\\ &= \{f : G_x \notin [H_s] \Rightarrow f(x) = 0,\ gf(x) = f(gx)\ \forall g \in \sigma(G/H_s)\}\end{aligned}$$

To show this, it suffices to note

$$(\mathcal{R}f)(x) = \sum_{s\in\mathbb{N}}(\mathcal{R}_s f)(x)$$

where one can show

$$\begin{aligned}(\mathcal{R}_s f)(x) &:= \mathbb{1}\{G_x \in [H_s]\}(\mathcal{R}f)(x)\\ &= \mathbb{1}\{G_x \in [H_s]\}\cdot\sigma(kH_s)\cdot\left(\frac{|H_s|}{|G|}\sum_{g\in\sigma(G/H_s)}g^{-1}f\circ g\right)(x_0)\end{aligned}$$

are mutually orthogonal projections onto each respective S_s. In particular, this follows from $\mathcal{R}$ being an orthogonal projection and the fact that $\mathbb{1}\{G_x \in [H_s]\}$ is G-invariant as a function of x.

The point is that when x is a random point drawn from a G-invariant distribution, k and x_0 will be independent random variables, by an appropriate "disintegration" of the measure. It turns out we can condition on orbit type—that is, on

[9] For simplicity of exposition, where there would usually appear integrals over Haar measure we will write sums over group elements; the results generalize from finite to compact groups.

$G_x \in [H_s]$. Note that conditioned on orbit type, x still has an invariant distribution. Now consider the individual terms, conditionally on H_s. The term $\sigma(kH_s)$ is chosen uniformly at random from the coset representatives, independent of x_0. The point x_0 is a random orbit representative among those with stabilizer H_s. The term in the middle is just a constant map, obtained by symmetrizing f on the coset representatives of H_s.

J.4. Generalization gap for relaxed equivariant ground truth

We write $\bar{f} = \mathcal{R}f$ and $f^{\perp} = f - \bar{f}$. Below, we appeal to the standard fact that for the generalization gap is $||f^{\perp}||^2$ (Elesedy & Zaidi, 2021); one can follow our proof for the case of densities in Appendix I.

To express the generalization gap, it is helpful to assume that we have a consistent choice of representatives $\sigma(gG_x)$ across all orbits, which only depend G_x—that is, we can write $\sigma(gH)$ for the representative of any gG_x when $G_x = H$. Then the generalization gap can be written by conditioning on the stabilizer being any given subgroup $H \leq G$. For simplicity, consider again a finite G. Then,

$$\begin{aligned}\Delta(f, \bar{f}) &= \mathbb{E}\left[\mathbb{E}\left[||f(X) - \bar{f}(X)||^2 | G_X = H\right]\right] \\ &= \sum_{H \leq G} \mathbb{P}(G_X = H)\mathbb{E}\left[\left|\left|f(X) - \frac{|H|}{|G|}\sum_{g \in \sigma(G/H)} g^{-1}f(gX)\right|\right|^2 \mid G_X = H\right]\end{aligned}$$

Intuitively, this says the generalization gap is measured by, on average over different possible stabilizers, how far away f is from being symmetrized with respect to the coset representatives for a given stabilizer. Another way to see this is to note the interaction term in $||f(X) - \bar{f}(X)||^2 = ||f(X)||^2 - 2\langle f(X), \bar{f}(X)\rangle + ||\bar{f}(X)||^2$ (i.e. the inner product) is

$$\frac{|H|}{|G|}\sum_{g \in \sigma(G/H)} \langle f(X), g^{-1}f(gX)\rangle = \frac{|H|}{|G|}\sum_{g \in \sigma(G/H)} \langle gf(X), f(gX)\rangle \tag{50}$$

by the orthogonality of the G action on $\mathcal{Y}$, and thus measures the alignment between $gf(X) = f(gX)$.