APPROXIMATE EQUIVARIANCE VIA PROJECTION-BASED REGULARISATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Equivariance is a powerful inductive bias in neural networks, improving generalisation and physical consistency. Recently, however, non-equivariant models have regained attention, due to their better runtime performance and imperfect symmetries that might arise in real-world applications. This has motivated the development of approximately equivariant models that strike a middle ground between respecting symmetries and fitting the data distribution. Existing approaches in this field usually apply sample-based regularisers which depend on data augmentation at training time, incurring a high sample complexity, in particular for continuous groups such as $SO(3)$. This work instead approaches approximate equivariance via a projection-based regulariser which leverages the orthogonal decomposition of linear layers into equivariant and non-equivariant components. In contrast to existing methods, this penalises non-equivariance at an operator level across the full group orbit, rather than point-wise. We present a mathematical framework for computing the non-equivariance penalty exactly and efficiently in both the spatial and spectral domain. In our experiments, our method consistently outperforms prior approximate equivariance approaches in both model performance and efficiency, achieving substantial runtime gains over sample-based regularisers.

1 INTRODUCTION

Over the past few years, equivariance has been proven to be a powerful design principle for machine learning models across chemistry (Thomas et al., 2018; Satorras et al., 2021; Brandstetter et al., 2022; Hoogeboom et al., 2022; Xu et al., 2024), physics (Bogatskiy et al., 2020; Spinner et al., 2024; Brehmer et al., 2025), robotics (Hoang et al., 2025), and engineering (Toshev et al., 2023).

Recently, however, there has been a shift back towards non-equivariant models, most prominently AlphaFold-3 (Abramson et al., 2024). Non-equivariant architectures often allow more flexible feature parameterisations and can be easier to optimise because the search is not restricted to an equivariant hypothesis class. This broader parameter space may enable the optimiser to find better minima than if it was confined to strictly equivariant models (Pertigkiozoglou et al., 2024). Moreover, many existing equivariant architectures rely on specialised tensor products to preserve symmetry (Weiler & Cesa, 2019; Brandstetter et al., 2022), which can be less efficient to compute on modern GPUs than dense matrix–vector operations.

At the same time, recent work demonstrates that equivariance remains a valuable inductive bias even at scale (Brehmer et al., 2024), and, for example, state-of-the-art molecular property prediction models continue to leverage it (Liao & Smidt, 2023; Liao et al., 2024; Fu et al., 2025). This motivates approaches that retain the benefits of equivariance without incurring its full constraints or computational costs.

A common approach here is to promote equivariance in otherwise non-equivariant architectures at the level of samples - for example via data augmentation (as in AlphaFold-3 (Abramson et al., 2024)) or pointwise equivariance penalties (Bai et al., 2025). In this work, we take a different perspective and introduce *projection-based equivariance regularisation*, a novel framework which allows tuning equivariance into any neural architecture on the operator level, thereby directly affecting the model weights.[1] Our primary contributions are:

- We propose a theoretically-grounded approach to regularise general machine learning models towards exact equivariance.
- Making use of the orthogonal decompostion of functions into equivariant and non-equivariant components, we are able to penalise non-equivariance on an operator level over the whole group orbit.

[1] Source code will be released with the camera-ready version.

- We show how to efficiently calculate the closed-form projection by working in the Fourier domain, allowing efficient regularisation for continuous groups such as $SO(n)$.
- We empirically demonstrate improvements over existing approaches for approximate equivariance, consistently achieving better task performance and have especially large gains in run-time over sample-based regularisers.

1.1 Related Work

A growing body of work relaxes strictly equivariant architectures to better capture approximate or imperfect symmetries in data. Finzi et al. (2021) model departures from symmetry by adding a small non-equivariant "residual" pathway to an otherwise equivariant network. Romero & Lohit (2022) introduce partial group convolutions that activate only on a subset of group elements. For discrete groups, Wang et al. (2022c) propose relaxed group convolutions, later extended by Wang et al. (2024) to expose symmetry-breaking mechanisms; Hofgard et al. (2024) further generalise this framework to continuous groups. Veefkind & Cesa (2024) introduce a learnable non-uniform measure over the group within steerable CNNs, yielding partially equivariant SCNNs whose degree of symmetry breaking is explicitly encoded in the learned measure. Samudre et al. (2025) instead enforce approximate equivariance through group-matrix–structured convolutional layers with low displacement rank, so that symmetry and its controlled violation are encoded as proximity to the group-matrix manifold, leading to highly parameter-efficient CNNs for discrete groups. McNeela (2024) introduce Lie-algebra convolutions with a non-strict equivariance bias, and van der Ouderaa et al. (2022) relax translation equivariance using spatially non-stationary convolution kernels. On graphs, Huang et al. (2023) develop approximately automorphism-equivariant GNNs. A complementary line of work studies how to measure equivariance (or its violation) and use it in training objectives Finzi et al. (2021); van der Ouderaa et al. (2022); Gruver et al. (2023); Otto et al. (2023); Petrache & Trivedi (2023). Another common approach focusses on regularisation towards equivariance. Bai et al. (2025) penalise pointwise deviations from equivariance constraints, Kouzelis et al. (2025) incorporate approximate symmetry in VAEs for generative modelling, and Zhong et al. (2023) apply related ideas to depth and normal prediction. Finally, Pertigkiozoglou et al. (2024) improve the training behavior of equivariant models by learning a non-equivariant model and projecting it into the equivariant subspace at test time.

2 Background

Notation. For vector spaces V and V', we denote the *identity* on V by I_V and write $\mathrm{Hom}(V, V')$ for the algebra of *linear homomorphisms* $V \to V'$. We write $\mathrm{Hom}(V, V) = \mathrm{End}(V)$. For $T \in \mathrm{Hom}(V, V')$, its *conjugate transpose* is denoted by $T^* : V' \to V$ and the group of *unitary operators* is $U(V) = \{T : TT^\star = T^\star T = I_V\}$. We can define a *norm* on the space of operators between normed spaces $(V, \|\cdot\|)$ and $(V', \|\cdot\|_{V'})$ by $\|T\| = \sup_{\|v\|_V=1} \|T(v)\|_{V'}$. The *Kronecker delta* $\delta_{x,y}$ is equal to 1 if $x = y$ and 0 otherwise.

Unitary representations. Given a group G, a *unitary representation* is a homomorphism $\pi : G \to U(V_\pi)$ into the unitary operators on a Hilbert space V_π; we call the pair (V_π, π) a G-module. Two representations $\pi : G \to U(V_\pi)$ and $\pi' : G \to U(V_{\pi'})$ are said to be *isomorphic* if there exists a unitary $U : V_\pi \to V_{\pi'}$ with $\pi(g) = U\,\pi'(g)\,U^{-1}$ for all $g \in G$. A representation is *irreducible* if it is not isomorphic to a direct sum of non-zero representations $\pi \oplus \pi'$ where $\pi \oplus \pi' : G \to U(V \oplus V')$ is defined by $(\pi \oplus \pi')(g)(v, v') = (\pi(g)v, \pi'(g)v')$.

Haar measure. Let G be a compact group. The *Haar measure* λ is the unique *bi-invariant* and *normalised* measure, i.e. for all Borel sets $E \subset G$ and every $g \in G$ we have $\lambda(gE) = \lambda(Eg) = \lambda(E)$, and $\lambda(G) = 1$. We can view the Haar measure as a uniform distribution over the group G. Indeed, if G is discrete, the Haar measure becomes the discrete uniform measure with $\lambda(\{g\}) = \frac{1}{|G|}$ for all $g \in G$.

Equivariance and G-smoothing. Let $T : (V, \pi) \to (V', \pi')$ be a (bounded) linear map between G-modules. We say T is *G-equivariant* if $T\big(\pi(g)v\big) = \pi'(g)\,T(v)$ for all $g \in G$, $v \in V$. If the action on V' is trivial ($\pi'(g) = I_{V'}$), we call T *invariant*. Averaging over G yields the *G-smoothing (Reynolds) operator*

$$P(T) \;=\; \int_G \pi'(g)^*\, T\, \pi(g)\, d\lambda(g). \tag{1}$$

Projection onto the equivariant subspace. When π, π' are unitary, P is the orthogonal projector (with respect to the Hilbert–Schmidt inner product) from $\mathrm{Hom}(V, V')$ onto the closed subspace of G-equivariant linear maps (Elesedy & Zaidi, 2021). The following structural decomposition will be useful.

Algorithm 1: Pseudo-code for the equivariant projection for finite (left) and continuous groups (right).

Projection for finite groups

```
def project_finite(W, group, rho_in,
    rho_out):
    W_proj = zeros_like(W)
    for g in group:
        W_proj += rho_out[g].conj().T @ W @
     rho_in[g]
    return W_proj / len(group)
```

Projection for continuous groups

```
def project_continuous(K, irreps,
    spatial_axes):
    K_hat = fftn(K, axes=spatial_axes)
    for pi in irreps:
        K_hat[pi] = mask_and_average(K_hat[
    pi])
    return ifftn(K_hat, axes=spatial_axes)
```

Lemma 2.1 (Elesedy & Zaidi (2021), Lemma 1)**.** *Let $\mathcal{H} \subset \{(V,\pi) \to (V',\pi')\}$ be a function space that is closed under P (i.e. $P(T) \in \mathcal{H}$ whenever $T \in \mathcal{H}$). Define*

$$S \;=\; \{T \in \mathcal{H} :\; T \text{ is } G\text{-equivariant}\}, \qquad A \;=\; \ker P \;=\; \{T \in \mathcal{H} :\; P(T) = 0\}. \tag{2}$$

Then P is an orthogonal projection with range S and kernel A, and hence $\mathcal{H} = S \oplus A$.

In particular, every $T \in \mathcal{H}$ orthogonally decomposes uniquely as $T = P(T) + \big(T - P(T)\big)$, where $P(T)$ is the G-equivariant component S and $T - P(T) \in A$ is its G-anti-symmetric component. Moreover, we have the following:

Corollary 2.2. *A function $T : (V,\pi) \to (V',\pi')$ is G-equivariant if and only if $P(T) = T$.*

3 Equivariant Projection Regularisation

Motivated by these observations, we propose a simple framework for learning (approximately) equivariant models: Let $\mathcal{H}$ be a hypothesis class and $L_{\text{task}}(T)$ a task-specific loss function for $T \in \mathcal{H}$. We learn T by solving

$$T^* \in \arg\inf_{T \in \mathcal{H}} \; L_{\text{task}}(T) \;+\; \lambda_G \,\|P(T)\| \;+\; \lambda_\perp \,\|T - P(T)\|, \tag{3}$$

where $\lambda_G, \lambda_\perp \geq 0$ are hyperparameters. Intuitively, increasing $\lambda_\perp$ (or decreasing λ_G) penalises $\|T - P(T)\|$ more strongly, which encourages $P(T) = T$, steering the solution toward stronger equivariance according to Lemma 2.1.

In what follows, we provide a theoretical justification for using $\|T - P(T)\|$ as a regulariser. Recalling that $P(T)$ denotes the closest equivariant operator to T, we show that the distance $\|T - P(T)\|$ is quantitatively equivalent to a natural measure of non-equivariance, the *equivariance defect*.

3.1 Bounding the Equivariance Error

Definition 3.1 (Equivariance defect)**.** *Let T be a function between G-modules with actions π_{in} and π_{out}. The equivariance defect at $g \in G$ is*

$$\Delta_g(T) \;\coloneqq\; \pi_{\text{out}}(g) \circ T \;-\; T \circ \pi_{\text{in}}(g), \tag{4}$$

and the worst-case defect is

$$\mathcal{E}(T) \;\coloneqq\; \sup_{g \in G} \|\Delta_g(T)\|. \tag{5}$$

By Lemma 2.1 (Elesedy & Zaidi, 2021), the quantity $\mathcal{E}(T)$ vanishes if and only if T is G-equivariant. The next lemma shows that this defect is effectively controlled, up to constants, by the distance to the equivariant subspace measured by the projection P.

Lemma 3.2. *For every (Lipschitz) function T between G-modules with unitary actions,*

$$\|T - P(T)\| \;\leq\; \mathcal{E}(T) \;\leq\; 2\,\|T - P(T)\|. \tag{6}$$

Proof. See Appendix A.1 □

Lemma 3.2 shows that regularising by $\mathcal{E}(T)$ or by $\|T - P(T)\|$ is equivalent up to a factor of 2. Thus, minimising $\|T - P(T)\|$ minimises the worst-case defect.

In practice, T will be some type of neural network architecture and is hence a composition of functions. The following bound decomposes the global defect of a network into per-layer defects, weighted by downstream Lipschitz constants.

Lemma 3.3. *Let $T = f_k \circ f_{k-1} \circ \cdots \circ f_1$ be a composition of Lipschitz maps between G-modules with unitary actions, and set $L_m := \mathrm{Lip}(f_m)$. Then*

$$\mathcal{E}(T) \;\le\; \sum_{i=1}^{k} \Big(\prod_{m \neq i}^{k} L_m \Big)\, \mathcal{E}(f_i). \tag{7}$$

Proof. See Appendix A.2 □

The bound above immediately yields the following corollary for standard feed-forward networks first shown by Kim et al. (2023).

Corollary 3.4 (Kim et al. (2023))**.** *Let*

$$T \;=\; W^{(S)} \circ \sigma_{S-1} \circ W^{(S-1)} \circ \cdots \circ \sigma_1 \circ W^{(1)} \tag{8}$$

be an S-layer network where each linear map $W^{(l)}$ acts between G-modules with unitary actions and each activation σ_l is G-equivariant and Lipschitz. Then

$$\mathcal{E}(T) \;\le\; C \sum_{l=1}^{S} \big\| W^{(l)} - P\big(W^{(l)}\big) \big\|, \tag{9}$$

for a constant $C > 0$ depending only on the operator norms of the $W^{(l)}$, the Lipschitz constants of the σ_l, and (when working on a bounded input domain) its radius.

Proof. See Appendix A.3. □

3.2 Projection in Fourier Space

The previous section motivates the use of the norm of the projection operator as a regulariser. When the projection operator in Equation 1 is efficiently computable in the spatial domain, e.g., for small finite groups (see Section 4.3), this is straightforward; Algorithm 1 provides pseudo-code for this case. However, in many applications, the group is large (for instance, uncountably infinite, as in $SO(n)$, the group of rotations about the origin in $\mathbb{R}^n$; see Section 4.1). In such cases, the integral in Equation 1 rarely admits a closed-form solution.

We therefore switch to the spectral domain. We assume the following setup, which is in line with the geometric deep learning blueprint (Bronstein et al., 2021) that constructs equivariant networks as a composition of equivariant linear layers with equivariance-preserving non-linearities. Let G be a compact group with normalised Haar measure λ, and consider linear maps $T : L^2(G) \to L^2(G)$ on the Hilbert space of square-integrable complex functions,

$$L^2(G) = \{f : G \to \mathbb{C}\}, \qquad \langle f, h \rangle = \int_G f(g)\, \overline{h(g)}\, d\lambda(g). \tag{10}$$

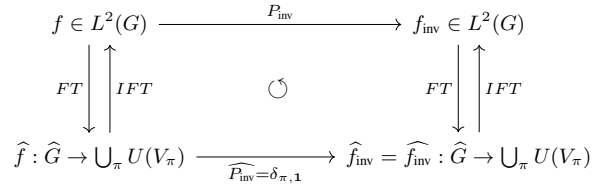
We study equivariance with respect to the (left) regular representation $\tau : G \to U(L^2(G))$ defined by

$$(\tau(g)f)(x) = f(g^{-1}x), \qquad x, g \in G. \tag{11}$$

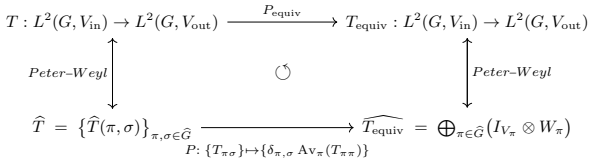
We denote by $\widehat{G}$ the set of equivalence classes of finite-dimensional irreducible representations of G and call it the *unitary dual* of G. Each $[\pi] \in \widehat{G}$ has a representative $\pi : G \to U(V_\pi)$ with $d_\pi = \dim V_\pi$. For $f \in L^2(G)$, we define the *(non-abelian) Fourier transform* as

$$\widehat{f}(\pi) \;:=\; \int_G f(g)\, \pi(g)^*\, d\lambda(g) \;\in \mathrm{End}(V_\pi). \tag{12}$$

In the following, using tools from Fourier analysis on compact groups, we will derive that the projection operator of Equation 1 can be computed efficiently in Fourier space.



(a) Invariance: keep only the trivial representation and discards all other frequencies.



(b) Equivariance: zero out off-diagonals ($\pi \neq \sigma$) and average within each frequency block to obtain $I_{V_\pi} \otimes B_\pi$.

Figure 2: Commutative diagrams showing how to apply the projection operator in Fourier space.

Theorem 3.5 (Informal)**.** *Equivariant linear maps are block-diagonal in the frequency domain (one block per irreducible representation). Hence, the projection onto equivariant subspaces acts by zeroing out all off-diagonal terms.*

Hence, whenever an efficient Fourier transform is available (e.g., on regular grids) or the model is already parameterised spectrally (e.g., eSEN (Fu et al., 2025)), imposing equivariance reduces to diagonalising the relevant linear operators in the spectral domain.

3.3 Invariant functions in Fourier space

In this subsection, we show that an invariant function $f \in L^2(G)$ only has trivial non-zero Fourier coefficients.

Lemma 3.6. *Let $f \in L^2(G)$ be left invariant with respect to the regular representation τ, i.e. $f(hg) = f(g)$ for all $h, g \in G$. Then $\widehat{f}(\pi)$ is non-zero if and only if π is the trivial representation $\mathbf{1} : g \mapsto I_{\mathbb{C}}$.*

Proof. See Appendix A.4. □

Corollary 3.7. *Let $f \in L^2(G)$ be any function on G and set P_{inv} to be the invariant projection. Then $\widehat{P_{inv}(f)}(\pi) = \widehat{f}(\pi)\delta_{\pi,\mathbf{1}}$.*

Proof. See Appendix A.5 □

In Figure 2a, we schematically depict how we can exploit the simple structure of the projection in the spectral domain $\widehat{P_{\text{inv}}}$ to efficiently calculate the smoothing operator P_{inv}.

3.4 Equivariant maps are diagonal across frequencies

Having shown the projection in the spectral domain for the invariant case, we now turn towards the case of equivariance. It turns out that in this case the projection acts by zeroing out all off-diagonal terms and averages over the rest. We can formalise this as follows:

Theorem 3.8. *Let $T : L^2(G) \to L^2(G)$ be a linear function which is equivariant with respect to the (left) regular representation, i.e. $\tau(g) \circ T = T \circ \tau(g)$ for all $g \in G$. Then T decomposes as follows:*

$$\widehat{T} \cong \bigoplus_{\pi \in \widehat{G}} I_{V_\pi} \otimes B_\pi \tag{13}$$

for some $B \in End(V_\pi^*)$ (one for each π). Equivalently, on Fourier coefficients:

$$\widehat{T(f)}(\pi) = \widehat{f}(\pi)\widehat{k}(\pi) \tag{14}$$

with $B_\pi \cong \widehat{k}(\pi)^*$.

Proof. See Appendix A.6. □

This means that an equivariant linear map T does not mix between irreps; it is block-diagonal. We now show what this means for the projection of a general linear operator T.

Corollary 3.9. *Let $T : L^2(G) \to L^2(G)$ be linear and set $P_{equiv}(T)$ to be its equivariant projection. Then for each $[\pi] \in \widehat{G}$, there exists $B_\pi \in End(V_\pi^*)$ such that for all $f \in L^2(G)$,*

$$\widehat{P(T)(f)}(\pi) = \widehat{f}(\pi)\, B_\pi. \tag{15}$$

3.5 Vector-valued signals and fiber-wise projection

Thus far we treated scalar signals $f \in L^2(G)$. In many applications (e.g. steerable CNNs Cohen & Welling (2017), tensor fields) one works with vector-valued signals taking values in a finite-dimensional unitary G-module (V, ρ). Define

$$L^2(G, V) \;\cong\; L^2(G) \otimes V \qquad \text{with action} \qquad \big((\tau \otimes \rho)(g) f\big)(x) = \rho(g)\, f(g^{-1}x). \tag{16}$$

More generally, for an operator $T : L^2(G, V_{\text{in}}) \to L^2(G, V_{\text{out}})$ we measure equivariance with respect to the pair of actions $\tau \otimes \rho_{\text{in}}$ (on the domain) and $\tau \otimes \rho_{\text{out}}$ (on the codomain), i.e.

$$(\tau \otimes \rho_{\text{out}})(g) \circ T \;=\; T \circ (\tau \otimes \rho_{\text{in}})(g) \qquad \forall g \in G. \tag{17}$$

As in the scalar case, P_{equiv} is an idempotent, self-adjoint projection onto the equivariant subspace and we can analogously show that a projected map T will have block-diagonal structure. Indeed, a Peter–Weyl–type decomposition yields the following (details in Appendix B):

Theorem 3.10. *Let $T : L^2(G, V_{\text{in}}) \to L^2(G, V_{\text{out}})$ be linear. Then the equivariant projection decomposes as*

$$\widehat{P_{\text{equiv}}(T)} \;\cong\; \bigoplus_{\pi \in \widehat{G}} \Big(I_{V_\pi} \otimes W_\pi \Big) \tag{18}$$

with

$$W_\pi \;=\; \int_G \big(\pi(g)^* \otimes \rho_{\text{out}}(g)\big)\, \widehat{T}(\pi, \pi)\, \big(\pi(g) \otimes \rho_{\text{in}}(g)^{-1}\big)\, d\lambda(g). \tag{19}$$

In particular, every equivariant T is block-diagonal across frequencies and acts as the identity on V_π and as an intertwiner on the fiber–multiplicity space $V_\pi^* \otimes V$.

Hence, the equivariant projection can be computed efficiently in Fourier space. Given a linear map T, we (i) compute the Fourier transform of the matrix representation of T to obtain the frequency blocks $\widehat{T}(\pi, \sigma)$; (ii) zero all off-diagonal blocks, setting $\widehat{T}(\pi, \sigma) \leftarrow 0$ for $\pi \neq \sigma$; (iii) for each π, project $\widehat{T}(\pi, \pi)$ onto $\mathrm{Hom}_G(\pi^* \otimes \rho_{\text{in}}, \pi^* \otimes \rho_{\text{out}})$ using the averaging formula for B_π above; and (iv) apply the inverse Fourier transform to obtain $P_{\text{equiv}}(T)$ in the spatial domain.

This procedure is illustrated by the commutative diagram in Figure 2b, and a corresponding pseudo-code implementation is given in Algorithm 1 on the right.

3.6 Asymptotic cost

We now want to briefly comment on the computational complexity of calculating the projection for both finite and continuous groups.

Finite groups. For finite groups we use Equation 1 directly. For a linear layer with weights $W \in \mathbb{C}^{d_{\text{out}} \times d_{\text{in}}}$ and $N_\ell = d_{\text{out}} d_{\text{in}}$ parameters, the projection evaluates $\pi_{\text{out}}(g)^* W \pi_{\text{in}}(g)$ for each $g \in G$, where $\pi_{\text{out}}(g)$, $\pi_{\text{in}}(g)$ are the representation matrices. Each step costs $O(d_{\text{out}}^2 d_{\text{in}}) + O(d_{\text{out}} d_{\text{in}}^2)$, which is $O(d_{\text{out}}^3)$ under $d_{\text{in}} \sim d_{\text{out}}$. Since $N_\ell \sim d_{\text{out}}^2$, this is $O(N_\ell^{3/2})$ per group element, and $O\big(|G|\, N_\ell^{3/2}\big)$.

Continuous groups. For continuous groups, we use the Fourier-domain projection. For a kernel on a d-dimensional grid of size $S = k^d$, FFTs cost $O(S \log S)$ per input-output channel pair, so per layer $O(d_{\text{out}} d_{\text{in}} S \log S)$. Masking and averaging in spectral space cost $O(P_\ell)$ with $P_\ell = d_{\text{out}} d_{\text{in}} S = N_\ell$, so the overall cost is $O\big(d_{\text{out}} d_{\text{in}} S \log S\big)$. If weights are already stored in irreducible spectral blocks, the projection reduces to masking and averaging only, giving $O(N_\ell)$ per layer; this is precisely the regime of steerable CNNs, where kernels are parameterised directly in such blocks.

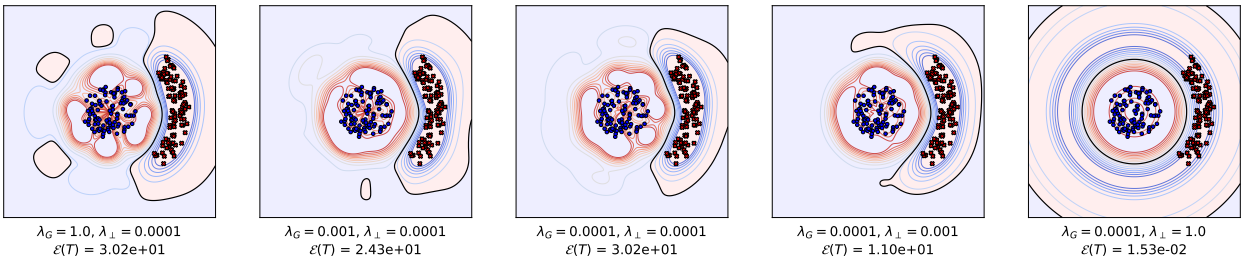


Figure 3: Controlling the degree of learned $SO(2)$ invariance by tuning the parameters λ_G and $\lambda_\perp$, which penalise the projections of the equivariant and non-equivariant components, respectively.

4 Experiments

In this section, we conduct three sets of experiments to demonstrate the feasibility and efficiency of our approach to learn (approximate) equivariance from data. For implementation details and information on hyperparameters, see Appendix C.

4.1 Example: Learned $SO(2)$ invariance

We first want to illustrate the approach in Section 3 on a simple toy problem (Figure 3). The task is binary classification on two point clouds in $\mathbb{R}^2$. Using polar coordinates (r, θ), we sample an inner disk-shaped cloud (blue, label $+1$), and the outer angular section of an annulus (red, label -1). We then train an approximately $\mathrm{SO}(2)$-invariant MLP with the following structure on this dataset: We first project inputs $(x, y) \in \mathbb{R}^2$ onto circular harmonics up to degree M, adding C radial channels via radial embedding functions, to obtain equivariant irreps features $H \in \mathbb{C}^{(2M+1)\times C}$. We then apply two fully connected complex linear layers

$$L_1 : \mathbb{C}^{(2M+1)\times C} \to \mathbb{C}^{(2M+1)\times C_{\text{hid}}}, \qquad L_2 : \mathbb{C}^{(2M+1)\times C_{\text{hid}}} \to \mathbb{C}^{(2M+1)\times C_{\text{hid}}},$$

followed by an $\mathrm{SO}(2)$-equivariant tensor product. Lastly, we extract the invariant component and pass its real part through a final real-valued linear head $L_{\text{final}} : \mathbb{R}^{C_{\text{hid}}} \to \mathbb{R}$ to produce the scalar logit. For a more in-depth description of this architecture, see Appendix C.1.

In this setting, the projection onto the equivariant subspace reduces to masking. Let $W_i \in \mathbb{C}^{((2M+1)C)\times((2M+1)C)}$ denote the flattened weight matrix of an intermediate linear layer. Define the mask $M \in \mathbb{R}^{((2M+1)C)\times((2M+1)C)}$ by

$$M_{(m_1,c_1),(m_2,c_2)} = \delta_{m_1,m_2},$$

i.e., only blocks with matching harmonic order m are kept. The projected weights are $P(W_i) = M \odot W_i$, where $\odot$ denotes elementwise multiplication. The overall objective is

$$L = L_{\text{task}} + \lambda_G \sum_i \|W_i\| + \lambda_\perp \sum_i \|W_i - M \odot W_i\|,$$

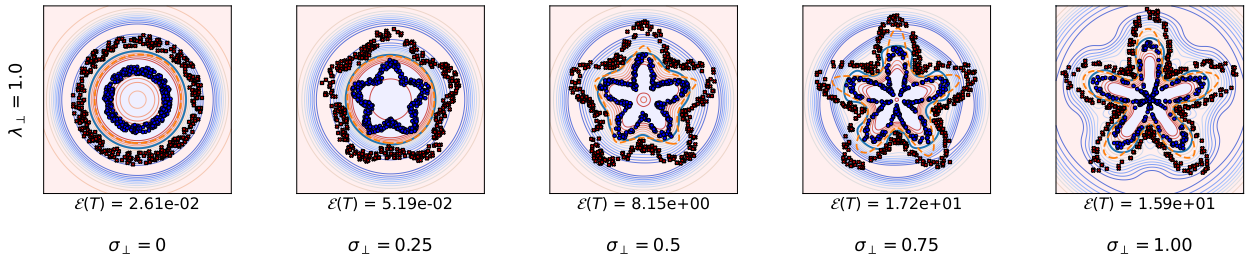


Figure 4: Effect of increasing angular perturbation at fixed projection strength. Each panel shows the decision boundary and level sets of the approximately $\mathrm{SO}(2)$-invariant network (blue) and an MLP (orange) trained with a fixed non-equivariant penalty $\lambda_\perp = 1.0$ on datasets with growing angular "wave" amplitude $\sigma_\perp$ (left to right). As $\sigma_\perp$ increases, the decision boundary becomes more angle-dependent and the learned classifier departs from perfect radial symmetry only where required to fit the data, while remaining nearly circular elsewhere. The empirical invariance defect $\mathcal{E}(T)$ for each setting is reported beneath the corresponding panel.

with $\lambda_G, \lambda_\perp \geq 0$ and L_{task} the standard classification loss.

In Figure 3, we compare trained models across different values of $(\lambda_G, \lambda_\perp)$. From left to right, we first reduce λ_G and then increase $\lambda_\perp$, enforcing progressively stronger invariance. For a full 2D grid for different combinations of $(\lambda_G, \lambda_\perp)$ see Figure 7 in Appendix C.1. As the regularisation intensifies, the decision boundary becomes increasingly $\mathrm{SO}(2)$-invariant, confirming that the proposed projection-based regulariser effectively pushes the model toward invariance. Consistently, the empirical equivariance defect

$$\mathcal{E}_{\text{emp}}(T) \;=\; \sum_{k,l} \Big\| \rho_{\text{out}}(g_l)\, T(x_k) \;-\; T\big(\rho_{\text{in}}(g_l)\, x_k\big) \Big\| \tag{20}$$

with k ranging over data samples and g_l drawn as random rotations in $\mathrm{SO}(2)$, decreases from left to right.

In a second experiment, we probe the behaviour of the regulariser when the target function departs from exact $\mathrm{SO}(2)$-invariance by making the labels increasingly dependent on the polar angle. Starting from two concentric rings, we introduce an angular "wave" perturbation of amplitude $\sigma_\perp$ in the radial direction, such that for $\sigma_\perp = 0$ the data distribution is rotationally symmetric, whereas larger $\sigma_\perp$ produce interlocking rings (Figure 4). We train the approximately $\mathrm{SO}(2)$-invariant network with projection-based regularisation alongside a plain MLP baseline on these datasets and compare both the learned decision boundaries and the empirical defect $\mathcal{E}_{\text{emp}}(T)$. As $\sigma_\perp$ increases, the regularised model departs from strict invariance only insofar as needed to fit the angularly perturbed rings. This illustrates how the projection penalty (even for constant values of $\lambda_\perp$) furnishes a tunable bias toward invariance that can be gradually traded off against fitting angle-dependent structure in the data. For a full grid, where we also vary the value of $\lambda_\perp$, see Figure 5 in Appendix C.1.

4.2 Imperfectly Symmetric Dynamical Systems

In this section, we follow the experimental design of Wang et al. (2022c) and evaluate our regulariser when applied to their relaxed group and steerable convolutional layers. Using `PhiFlow` (Holl & Thuerey, 2024), we generate 64×64 two-dimensional smoke advection–diffusion simulations with varied initial conditions under relaxed symmetries. Each network is trained to predict the velocity field one step ahead.

To test generalisation, we consider two out-of-distribution settings. In the *Future* setting, models predict velocity fields at time steps that are absent from the training distribution, while remaining within spatial regions that were seen during training. In the *Domain* setting, we evaluate at the same time indices as training but at spatial locations that were not seen. The data are produced to break specific symmetries in a controlled way: for *translation*, we generate series for 35 distinct inflow positions and split the domain horizontally into two subdomains with different buoyancy forces so that plumes diffuse at different rates across the interface; for *discrete rotation*, we simulate 40 combinations of inflow position and buoyancy, where the inflow pattern alone is symmetric under 90° rotations about the domain centre but a position-dependent buoyancy factor breaks rotational equivariance; and for *scaling*, we run 40 simulations with different time steps Δt and spatial resolutions Δx to disrupt scale equivariance.

We compare the relaxed group convolutional networks (`RGroup`) and relaxed steerable CNNs (`RSteer`) introduced by Wang et al. (2022c) with several baselines: a standard CNN (`Conv`), an equivariant convolutional network (`Equiv`) (Weiler & Cesa, 2019; Sosnovik et al., 2020), Residual Pathway Priors (`RPP`) (Finzi et al., 2021), a locally connected network with an explicit equivariance penalty in the loss (`CLNN`) and `Lift` (Wang et al., 2022a). We indicate the addition of our regulariser with the suffix `+Reg`.

Across these settings, incorporating our regulariser preserves performance when approximate translation equivariance holds and delivers substantial improvements in the rotation and scaling regimes. In short, the penalty promotes the desired approximate equivariance where symmetry is only partially present, without degrading accuracy where the symmetry is already well aligned with the data.

4.3 CT-Scan Metal Artifact Reduction

We compare our approach with a sample-based equivariance penalty on metal artefact reduction (MAR) for CT scans. Metal implants introduce characteristic streaking artefacts that obscure clinically relevant structures. The task is to map a corrupted slice to its artefact-reduced counterpart.

We use the AAPM CT-MAR Grand Challenge datasets (AAPM, 2022a;b), comprising 14,000 head and body CT slices with synthetic metal artefacts (Table 2 and Appendix C.3, Figure 6 for a visual comparison). The datasets were generated with the open-source CT simulation environment XCIST (Wu et al., 2022), using a hybrid data-simulation framework that combines publicly available clinical images (Yan et al., 2018; Goren et al., 2017) and virtual metal objects.

Table 1: Results on three synthetic smoke-plume datasets exhibiting approximate symmetries. We report means and standard deviations of pixel-wise MSE over 5 random seeds. *Future* indicates that the test set occurs after the training period; *Domain* indicates that training and test sets come from different spatial regions. Adding our proposed *equivariance regulariser* (+Reg) consistently improves performance.

Model		Conv	Equiv	Rpp	CLCNN	Lift	RGroup	+Reg	RSteer	+Reg
Translation	Future	—	0.94±0.02	0.92±0.01	0.92±0.01	0.87±0.03	**0.71±0.01**	**0.72±0.01**	—	—
	Domain	—	0.68±0.05	0.93±0.01	0.89±0.01	0.70±0.00	**0.62±0.02**	**0.62±0.01**	—	—
Rotation	Future	1.21±0.01	1.05±0.06	0.96±0.10	0.96±0.05	0.82±0.08	0.82±0.01	0.80±0.01	0.80±0.00	**0.79±0.00**
	Domain	1.10±0.05	0.76±0.02	0.83±0.01	0.84±0.10	0.68±0.09	0.73±0.02	0.67±0.01	0.67±0.01	**0.58±0.00**
Scaling	Future	0.83±0.01	0.75±0.03	0.81±0.09	1.03±0.01	0.85±0.01	0.80±0.01	0.81±0.00	0.70±0.01	**0.62±0.01**
	Domain	0.95±0.02	0.87±0.02	0.86±0.05	0.83±0.05	0.77±0.02	0.88±0.01	0.88±0.02	0.73±0.01	**0.69±0.01**

Following Bai et al. (2025), we adapt three convolution-based architectures `ACDNet` (Wang et al., 2022b), `DICDNet` (Wang et al., 2021) and `OSCNet` (Wang et al., 2023) by encouraging rotation equivariance with respect to the discrete group C_4 (rotations by multiples of $90°$). We compare the unregularised baselines, the sample-based regulariser of Bai et al. (2025), and the same networks equipped with our projection-based regulariser. Additionally, we compare with Residual Pathway Priors (RPPs) (Finzi et al., 2021) and a train-then-project variant, in which we first train a non-equivariant model and then project its linear layers onto the equivariant subspace at test time using our projection operator.

For steerable CNN layers whose channels are organised into orientation groups of four, the layer-wise projection acting on a kernel $K \in \mathbb{R}^{C'_{\text{out}} \times C'_{\text{in}} \times 4 \times 4 \times s \times s}$ is

$$P_{\text{equiv}}(K) \;=\; \tfrac{1}{4} \sum_{r=0}^{3} S^r \left(\text{rot}_r K\right) S^{-r}, \tag{21}$$

where S is the 4×4 cyclic-shift matrix on orientation channels and rot_r rotates the spatial kernel by $90°r$. For a derivation of this expression, see Appendix C.3.2.

In contrast, Bai et al. (2025) penalise a term that samples both a data point and a group element. For each sample x they draw a random $r \in C_4$ and add

$$L_{\text{equiv}}(x, r) \;=\; \left\| S^r \,\text{rot}_r\, K(x) \;-\; K\big(S^r \,\text{rot}_r x\big) \right\|^2 \tag{22}$$

to the task loss. This requires an extra forward pass for each sampled rotation and each data sample, with asymptotic cost $O(N_{\text{samples}} \cdot \text{cost}_{\text{forward}})$ where N_{samples} is the number of sampled group elements and $\text{cost}_{\text{forward}}$ is the cost of a single forward pass. By contrast, as derived in Section 3.6, our projection-based regulariser $\|P_{\text{equiv}}(\cdot)\|$ incurs a cost that is linear in the number of parameters, does not sample rotations or data, introduces no extra forward passes, and has zero estimator variance.

We report peak signal-to-noise ratio (PSNR), structural similarity (SSIM), and training throughput on a single A100 GPU under two regimes. In the fixed-batch setting, we use batch size 4 for all methods. In the max-feasible setting, the sample-based regulariser remains at batch size 4 (limited by the extra forward/activation memory), whereas the baselines and our projection-based regulariser scale to batch size 12 due to unchanged per-sample compute and memory.

Our projection-based regulariser delivers competitive or superior reconstruction quality, surpassing the sample-based penalty in all metrics across all settings but one, and improving over the unregularised baselines in most cases. Owing to the extra forward pass in Equation 22, the sample-based approach is constrained to smaller batch sizes and lower throughput. Even under the fixed-batch protocol, its throughput is 42–47% lower than ours; under the max-feasible protocol, the gap widens to 54–61%. These results indicate that projection-based regularisation achieves stronger C_4-equivariance with better hardware efficiency by avoiding per-sample group sampling. Similarly, due to the overparameterisation of the equivariant subspace, RPPs incur slower runtime during both training and inference and require more learnable parameters, and still underperform our approach in reconstruction quality.

5 Conclusion

In this work, we introduced projection-based regularisation - a theoretically grounded approach to learned equivariance which directly penalises model weights and regularises over the entire group instead of only point-wise, per-sample

Table 2: CT-scan metal artefact reduction on the AAPM challenge dataset. We compare three baseline models in their vanilla form, the sample-based regulariser of Bai et al. (2025), a train-then-project approach, Residual Pathway Priors of Finzi et al. (2021), and our projection-based regulariser. We report PSNR/SSIM, training throughput (for batch sizes 4 and 12, where stable) and inference throughput, epoch wall-clock time, and peak memory usage. Sample-based regularisation is limited to batch sizes ≤ 4, whereas the baselines and our method scale to batch size 12.

Model	#params	Throughput (no./GPU-s)		Epoch	Memory	AAPM	
		Train ↑	Inference ↑	time (s) ↓	(GB)	PSNR ↑	SSIM ↑
`ACDNet` (Wang et al., 2022b)	4.2M	4.90/5.16	8.40	1108	11.08	42.08	0.9559
+ sample-based (Bai et al., 2025)	4.2M	2.54/2.54	8.38	2011	21.99	40.02	**0.9623**
+ test-time projection	4.2M	–	–	–	–	23.63	0.8384
+ RPP (Finzi et al., 2021)	6.9M	3.49/4.14	5.37	1455	11.15	37.12	0.9413
+ projection-based (ours)	4.2M	4.25/4.99	7.44	1202	11.11	**42.68**	0.9620
`DICDNet` (Wang et al., 2021)	4.3M	8.38/9.72	11.86	632	10.90	41.44	0.9468
+ sample-based (Bai et al., 2025)	4.3M	4.05/4.05	10.15	1303	23.93	41.47	0.9464
+ test-time projection	4.3M	–	–	–	–	41.59	0.9602
+ RPP (Finzi et al., 2021)	6.6M	3.10/6.10	6.93	1028	12.08	39.42	0.9481
+ projection-based (ours)	4.3M	5.77/7.82	10.11	782	12.05	**41.52**	**0.9605**
`OSCNet` (Wang et al., 2023)	4.3M	8.59/9.86	12.00	624	10.37	**42.36**	0.9596
+ sample-based (Bai et al., 2025)	4.3M	4.05/4.05	10.13	1304	23.93	41.50	0.9593
+ test-time projection	4.3M	–	–	–	–	41.37	0.9609
+ RPP (Finzi et al., 2021)	6.6M	4.51/6.14	6.92	1016	12.08	39.45	0.9507
+ projection-based (ours)	4.3M	5.66/7.87	10.14	769	12.05	41.88	**0.9612**

regularisation. For operators for which no closed-form solution of the projection can be computed efficiently in the spatial domain, we provide a general framework for computing the projection efficiently in Fourier space by masking. The experiments demonstrate that across synthetic and real-world experiments, covering both finite and continuous symmetry groups, the proposed approach improves both task performance and runtime.

Limitations and future work. A limitation of the proposed approach is that the penalty term needs to be derived anew for each model architecture and group operation. Also, current experiments only evaluate the proposed method for relatively simple groups. In future work, we plan to extend this approach to more complex group structures consisting of several subgroups with applications in e.g. material sciences.

5.1 Reproducibility Statement

Reproducibility Statement: We performed our experiments on public datasets and included all necessary hyperparameters in Appendix C. Throughout Section 3, we clearly state all our theoretical assumptions, in particular in the statements of Theorems 3.8 and 3.10. We will publish the source code with evaluation scripts to reproduce the experiments with the camera-ready version.

References

American Association of Physicists in Medicine AAPM. Aapm ct metal artifact reduction (ct-mar) grand challenge, 2022a. URL `https://www.aapm.org/GrandChallenge/CT-MAR/`.

American Association of Physicists in Medicine AAPM. Aapm ct metal artifact reduction (ct-mar) grand challenge benchmark tool, 2022b. URL `https://github.com/xcist/example/tree/main/AAPM_datachallenge/`.

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

Yulu Bai, Jiahong Fu, Qi Xie, and Deyu Meng. A regularization-guided equivariant approach for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2300–2310, June 2025.

Alexander Bogatskiy, Brandon Anderson, Jan Offermann, Marwah Roussi, David Miller, and Risi Kondor. Lorentz group equivariant neural network for particle physics. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 992–1002. PMLR, 13–18 Jul 2020. URL `https://proceedings.mlr.press/v119/bogatskiy20a.html`.

Johannes Brandstetter, Rob Hesselink, Elise van der Pol, Erik J Bekkers, and Max Welling. Geometric and physical quantities improve e(3) equivariant message passing. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=_xwr8gOBeV1.

Johann Brehmer, Sönke Behrends, Pim De Haan, and Taco Cohen. Does equivariance matter at scale? In *NeurIPS 2024 Workshop on Symmetry and Geometry in Neural Representations*, 2024. URL https://openreview.net/forum?id=L4gb2wvVhM.

Johann Brehmer, Víctor Bresó, Pim de Haan, Tilman Plehn, Huilin Qu, Jonas Spinner, and Jesse Thaler. A lorentz-equivariant transformer for all of the lhc, 2025. URL https://arxiv.org/abs/2411.00446.

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.

Taco S Cohen and Max Welling. Steerable cnns. In *International Conference on Learning Representations*, 2017.

Li Deng. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.

Bryn Elesedy and Sheheryar Zaidi. Provably strict generalisation benefit for equivariant models. In *International conference on machine learning*, pp. 2959–2969. PMLR, 2021.

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.

Xiang Fu, Brandon M Wood, Luis Barroso-Luque, Daniel S. Levine, Meng Gao, Misko Dzamba, and C. Lawrence Zitnick. Learning smooth and expressive interatomic potentials for physical property prediction. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=R0PBjxIbgm.

Nir Goren, Tony Dowrick, James Avery, and David Holder. UCLH Stroke EIT Dataset – Radiology Data (CT), 2017. URL https://doi.org/10.5281/zenodo.8383704.

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.

Tai Hoang, Huy Le, Philipp Becker, Vien Anh Ngo, and Gerhard Neumann. Geometry-aware RL for manipulation of varying shapes and deformable objects. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=7BLXhmWvwF.

Elyssa Hofgard, Rui Wang, Robin Walters, and Tess Smidt. Relaxed equivariant graph neural networks. *ELLIS Workshop on Geometry-grounded Representation Learning and Generative Modeling, ICML*, 2024.

Philipp Holl and Nils Thuerey. Φ_{flow} (PhiFlow): Differentiable simulations for pytorch, tensorflow and jax. In *International Conference on Machine Learning*. PMLR, 2024.

Emiel Hoogeboom, Víctor Garcia Satorras, Clément Vignac, and Max Welling. Equivariant diffusion for molecule generation in 3D. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 8867–8887. PMLR, 17–23 Jul 2022.

Ningyuan (Teresa) Huang, Ron Levie, and Soledad Villar. Approximately equivariant graph networks. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

Hyunsu Kim, Hyungi Lee, Hongseok Yang, and Juho Lee. Regularizing Towards Soft Equivariance Under Mixed Symmetries, 2023. URL https://arxiv.org/abs/2306.00356.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Theodoros Kouzelis, Ioannis Kakogeorgiou, Spyros Gidaris, and Nikos Komodakis. EQ-VAE: Equivariance regularized latent space for improved generative image modeling. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=UWhW5YYLo6.

Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. Cifar-10 (canadian institute for advanced research). URL http://www.cs.toronto.edu/~kriz/cifar.html.

Yi-Lun Liao and Tess Smidt. Equiformer: Equivariant graph attention transformer for 3d atomistic graphs. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=KwmPfARgOTD.

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

Daniel McNeela. Almost equivariance via lie algebra convolutions, 2024. URL https://arxiv.org/abs/2310.13164.

Samuel E Otto, Nicholas Zolman, J Nathan Kutz, and Steven L Brunton. A unified framework to enforce, discover, and promote symmetry in machine learning. *arXiv preprint arXiv:2311.00212*, 2023.

Stefanos Pertigkiozoglou, Evangelos Chatzipantazis, Shubhendu Trivedi, and Kostas Daniilidis. Improving equivariant model training via constraint relaxation. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA, 2024. Curran Associates Inc. ISBN 9798331314385.

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

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

Ashwin Samudre, Mircea Petrache, Brian Nord, and Shubhendu Trivedi. Symmetry-based structured matrices for efficient approximately equivariant networks. In Yingzhen Li, Stephan Mandt, Shipra Agrawal, and Emtiyaz Khan (eds.), *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 1171–1179. PMLR, 03–05 May 2025. URL https://proceedings.mlr.press/v258/samudre25a.html.

Víctor Garcia Satorras, Emiel Hoogeboom, and Max Welling. E(n) equivariant graph neural networks. In Marina Meila and Tong Zhang (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.

Ivan Sosnovik, Michał Szmaja, and Arnold Smeulders. Scale-equivariant steerable networks. In *International Conference on Learning Representations*, 2020.

Jonas Spinner, Victor Bresó, Pim de Haan, Tilman Plehn, Jesse Thaler, and Johann Brehmer. Lorentz-equivariant geometric algebra transformers for high-energy physics. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 22178–22205. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/277628cff838927d869cd1f671328ce0-Paper-Conference.pdf.

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

Artur Toshev, Gianluca Galletti, Johannes Brandstetter, Stefan Adami, and Nikolaus A Adams. E (3) equivariant graph neural networks for particle-based fluid mechanics. In *ICLR 2023 Workshop on Physics for Machine Learning*, 2023.

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. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 49249–49309. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/veefkind24a.html.

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, 2022a.

Hong Wang, Yuexiang Li, Nanjun He, Kai Ma, Deyu Meng, and Yefeng Zheng. Dicdnet: Deep interpretable convolutional dictionary network for metal artifact reduction in ct images. *IEEE Transactions on Medical Imaging*, 41(4): 869–880, 2021.

Hong Wang, Yuexiang Li, Deyu Meng, and Yefeng Zheng. Adaptive convolutional dictionary network for ct metal artifact reduction. In *The 31st International Joint Conference on Artificial Intelligence*. IEEE, 2022b.

Hong Wang, Qi Xie, Dong Zeng, Jianhua Ma, Deyu Meng, and Yefeng Zheng. Oscnet: Orientation-shared convolutional network for ct metal artifact learning. *IEEE Transactions on Medical Imaging*, 2023.

Rui Wang, Robin Walters, and Rose Yu. Approximately Equivariant Networks for Imperfectly Symmetric Dynamics. *Proceedings of the 39th International Conference on Machine Learning*, 2022c. doi: 10.48550/arXiv.2201.11969. URL http://arxiv.org/abs/2201.11969.

Rui Wang, Elyssa Hofgard, Han Gao, Robin Walters, and Tess E. Smidt. Discovering symmetry breaking in physical systems with relaxed group convolution. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

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

M. Wu, P. FitzGerald, J. Zhang, W. P. Segars, J. Yu, Y. Xu, and B. De Man. XCIST – an open access x-ray/ct simulation toolkit. *Physics in Medicine and Biology*, 2022.

Minkai Xu, Jiaqi Han, Aaron Lou, Jean Kossaifi, Arvind Ramanathan, Kamyar Azizzadenesheli, Jure Leskovec, Stefano Ermon, and Anima Anandkumar. Equivariant graph neural operator for modeling 3d dynamics. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 55015–55032, 2024.

Ke Yan, Xiaosong Wang, Le Lu, and Ronald M. Summers. DeepLesion: Automated mining of large-scale lesion annotations and universal lesion detection with deep learning. *Journal of Medical Imaging*, 2018.

Yuanyi Zhong, Anand Bhattad, Yu-Xiong Wang, and David Forsyth. Improving equivariance in state-of-the-art supervised depth and normal predictors. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 21775–21785, 2023.

A Proofs in Section 3

A.1 Proof of Lemma 3.2

Proof of Lemma 3.2. By definition of P,

$$T - P(T) = \int_G \pi_{\text{out}}(g)^* \big(\pi_{\text{out}}(g) \circ T - T \circ \pi_{\text{in}}(g)\big)\, d\lambda(g) = \int_G \pi_{\text{out}}(g)^*\, \Delta_g(T)\, d\lambda(g). \tag{23}$$

Pre-/post-composition with the unitaries $\pi_{\text{out}}(g)^*$ preserves the Lipschitz seminorm, and the seminorm of an average is at most the average of the seminorms. Hence

$$\|T - P(T)\| \;\le\; \int_G \|\Delta_g(T)\|\, d\lambda(g) \;\le\; \sup_{g\in G} \|\Delta_g(T)\| \;=\; \mathbb{E}(T), \tag{24}$$

giving the lower bound. For the upper bound, note that $P(T)$ is G-equivariant, and therefore,

$$\Delta_g(T) = \pi_{\text{out}}(g)\big(T - P(T)\big) - \big(T - P(T)\big)\pi_{\text{in}}(g). \tag{25}$$

Taking Lipschitz seminorms and using that $\pi_{\text{in/out}}(g)$ are unitaries,

$$\|\Delta_g(T)\| \;\le\; \|T - P(T)\| + \|T - P(T)\| \;=\; 2\,\|T - P(T)\|. \tag{26}$$

Finally, take the supremum over $g \in G$ to obtain $\mathbb{E}(T) \le 2\|T - P(T)\|$. □

A.2 Proof of Lemma 3.3

Proof of Lemma 3.3. For any composable maps A, B, the equivariance defect satisfies the chain rule

$$\Delta_g(A \circ B) \;=\; (\Delta_g A) \circ B \;+\; A \circ (\Delta_g B). \tag{27}$$

Applying this repeatedly to $f_k \circ \cdots \circ f_1$ yields the telescoping identity

$$\Delta_g(T) = \sum_{i=1}^{k} \big(f_k \circ \cdots \circ f_{i+1}\big) \circ \Delta_g(f_i) \circ \big(f_{i-1} \circ \cdots \circ f_1\big). \tag{28}$$

Taking norms and using $\|X \circ Y\| \le \text{Lip}(X)\,\|Y\|$ together with $\text{Lip}(f_j) = L_j$ to obtain

$$\|\Delta_g(T)\| \;\le\; \sum_{i=1}^{k} \left(\prod_{m=i+1}^{k} L_m\right) \|\Delta_g(f_i)\| \left(\prod_{m=1}^{i-1} L_m\right). \tag{29}$$

Finally, take $\sup_{g\in G}$ on both sides and note that $\mathbb{E}(T) = \sup_g \|\Delta_g(T)\|$ and $\mathbb{E}(f_i) = \sup_g \|\Delta_g(f_i)\|$ to obtain the stated bound. □

A.3 Proof of Corollary 3.4

Proof of Corollary 3.4. To fit into the framework of Lemma 3.3, we choose

$$f_{2k-1} \coloneqq W^{(k)}, \qquad f_{2k} \coloneqq \sigma_k,$$

so that

$$L_{2k-1} \coloneqq \|W^{(k)}\|, \quad L_{2k} \coloneqq \text{Lip}(\sigma_k),$$

for $k = 1, \ldots, 2S-1$. By construction $E(\sigma_k) = 0$ for all k, hence $E(f_{2k}) = 0$. Plugging this into Equation 7 and noting that the even indices do not contribute, we obtain

$$E(T) \;\le\; \sum_{k=1}^{S} \left(\prod_{m\neq 2k-1} L_m\right) E\big(W^{(k)}\big). \tag{30}$$

Now note that the product over $m \neq 2k-1$ contains

- all activation Lipschitz constants $L_{2j} = \text{Lip}(\sigma_j)$, $j = 1, \ldots, 2S-1$,

- all weight norms $L_{2r-1} = \|W^{(r)}\|$ with $r \neq k$.

Thus

$$\prod_{m \neq 2k-1} L_m = \left(\prod_{j=1}^{S-1} \mathrm{Lip}(\sigma_j) \right) \left(\prod_{\substack{r=1 \\ r \neq k}}^{S} \|W^{(r)}\| \right),$$

and Equation 30 becomes

$$E(T) \;\leq\; \left(\prod_{j=1}^{S-1} \mathrm{Lip}(\sigma_j) \right) \sum_{k=1}^{S} \left(\prod_{\substack{r=1 \\ r \neq k}}^{S} \|W^{(r)}\| \right) E\big(W^{(k)}\big). \tag{31}$$

Next use Lemma 3.2, which states that for each linear layer

$$E\big(W^{(k)}\big) \;\leq\; 2 \big\| W^{(k)} - P\big(W^{(k)}\big) \big\|.$$

Substituting this into Equation 31 yields

$$E(T) \;\leq\; 2 \left(\prod_{j=1}^{S-1} \mathrm{Lip}(\sigma_j) \right) \sum_{k=1}^{S} \left(\prod_{\substack{r=1 \\ r \neq k}}^{S} \|W^{(r)}\| \right) \big\| W^{(k)} - P\big(W^{(k)}\big) \big\|. \tag{32}$$

Define

$$C := 2 \left(\prod_{j=1}^{S-1} \mathrm{Lip}(\sigma_j) \right) \max_{1 \leq k \leq S} \prod_{\substack{r=1 \\ r \neq k}}^{S} \|W^{(r)}\|. \tag{33}$$

Then, for every k,

$$2 \left(\prod_{j=1}^{S-1} \mathrm{Lip}(\sigma_j) \right) \prod_{\substack{r=1 \\ r \neq k}}^{S} \|W^{(r)}\| \;\leq\; C,$$

and Equation 32 implies

$$E(T) \;\leq\; C \sum_{k=1}^{S} \big\| W^{(k)} - P\big(W^{(k)}\big) \big\|.$$

This is exactly Eq. (9), with the dependence of C on the norms $\|W^{(k)}\|$ and Lipschitz constants $\mathrm{Lip}(\sigma_j)$ made explicit in Equation 33. □

A.4 Proof of Lemma 3.6

Proof of Lemma 3.6. We define the invariance operator of a function $f \in L^2(G)$ as

$$f_{\text{inv}}(g) = \int_G f(hg)\, d\lambda(h) \tag{34}$$

The Fourier coefficients of this are

$$\widehat{f_{\text{inv}}}(\pi) = \int_G f_{\text{inv}}(g)\, \pi(g)^*\, d\lambda(g) \tag{35}$$

$$= \int_G \left(\int_G f(hg)\, d\lambda(h) \right) \pi(g)^*\, d\lambda(g) \tag{36}$$

$$= \int_G f(x) \left(\int_G \pi(h^{-1}x)^*\, d\lambda(h) \right) d\lambda(x) \quad \text{substituting } x = hg \implies g = h^{-1}x \tag{37}$$

$$= \int_G f(x) \left(\int_G \pi(h^{-1})^*\, d\lambda(h) \right) \pi(x)^*\, d\lambda(x) \tag{38}$$

$$= \int_G f(x) \left(\int_G \pi(h)^*\, d\lambda(h) \right) \pi(x)^*\, d\lambda(x). \quad \text{invariance of Haar measure} \tag{39}$$

Define $A_\pi := \int_G \pi(h)^*\, d\lambda(h) \in \mathrm{End}(V_\pi)$. Note that A_π is π-equivariant; indeed, for all $g \in G$,

$$\pi(g)\, A_\pi = \int_G \pi(g)\, \pi(h)^*\, d\lambda(h) \tag{40}$$

$$= \int_G \pi(gh^{-1})\, d\lambda(h) \tag{41}$$

$$= \int_G \pi(k)^* \pi(g)\, d\lambda(k) \qquad \text{substituting } k = ghg^{-1} \implies gh^{-1} = k^{-1}g \tag{42}$$

$$= A_\pi\, \pi(g), \tag{43}$$

Hence by Schur's lemma (since π is irreducible), we have

$$A_\pi \in \mathrm{End}G(V_\pi) \;\cong\; \{\, \lambda I : \lambda \in \mathbb{C}\}.$$

So $A_\pi = \lambda I$ for some $\lambda \in \mathbb{C}$.

Now,

$$\mathrm{tr}\, A_\pi = \int_G \mathrm{tr}\big(\pi(h)^*\big)\, d\lambda(h) = \int_G \overline{\chi_\pi(h)}\, d\lambda(h) \;=\; \overline{\int_G \chi_\pi(h)\, d\lambda(h)}. \tag{44}$$

But the characters χ_π are orthonormal, so denoting the trivial representation $g \mapsto 1$ by $\mathbf{1}$, i.e. have $\chi_{\mathbf{1}}(g) = 1$ for all g, we have

$$\int_G \chi_\pi(g)\, d\lambda(g) = \int_G \chi_\pi(g)\, \overline{\chi_{\mathbf{1}}(g)}\, d\lambda(g) = \langle \chi_\pi(g), \chi_{\mathbf{1}}(g)\rangle_{L^2(G)} = \delta_{\pi,\mathbf{1}}. \tag{45}$$

Finally, this gives

$$d_\pi \lambda \;=\; \mathrm{tr}\, A_\pi \;=\; \Longrightarrow\; \lambda \;=\; \frac{\delta_{\pi,\mathbf{1}}}{d_\pi} = \begin{cases} 0, & \pi \neq \mathbf{1}, \\ \frac{1}{d_\pi}, & \pi = \mathbf{1}. \end{cases} \tag{46}$$

Substituting this into the above yields

$$\widehat{f_{\text{inv}}}(\pi) = \frac{1}{d_\pi} \widehat{f}(\pi) \delta_{\pi,\mathbf{1}}. \tag{47}$$

□

A.5 Proof of Corollary 3.7

Proof of Corollary 3.7. Since P_{inv} is a projection onto the G-invariant subspace, $P_{\text{inv}}(f)$ is always invariant. Hence, by Lemma 3.6, $\widehat{P_{\text{inv}}(f)}(\pi)$ is zero for all $\pi \neq \mathbf{1}$. Now note that by invariance, $P_{\text{inv}}(f)(g) = c$ for all $g \in G$ for some $c \in \mathbb{C}$. We then calculate

$$\widehat{P_{\text{inv}}(f)}(\mathbf{1}) = \int_G P_{\text{inv}}(f)(g)\; \mathbf{1}(g)^*\, d\lambda(g) = \int_G P_{\text{inv}}(f)(g)\; d\lambda(g) = \int_G c\, d\lambda(g) = c. \tag{48}$$

At the same time

$$\widehat{f}(\mathbf{1}) = \int_G\; f(g)\; \mathbf{1}(g)^* d\lambda(g) = \int_G\; f(g) d\lambda(g) = c. \tag{49}$$

which concludes the proof. □

A.6 Proof of Theorem 3.8

Proof of Theorem 3.8. By the Peter–Weyl theorem there is a unitary isomorphism

$$L^2(G) \;\cong\; \bigoplus_{\pi \in \widehat{G}} V_\pi \otimes V_\pi^*,$$

under which the left regular action is $\tau(g) \cong \bigoplus_\pi \big(\pi(g) \otimes I_{V_\pi^*}\big)$. Any linear map T on $L^2(G)$ becomes a block matrix $T = (T_{\pi\to\pi'})_{\pi,\pi'}$ with $T_{\pi\to\pi'} \in \mathrm{Hom}(V_\pi \otimes V_\pi^*,\, V_{\pi'} \otimes V_{\pi'}^*)$. The equivariance condition $\tau(g)T = T\tau(g)$ for all g reads, blockwise,

$$\big(\pi'(g) \otimes I_{V_{\pi'}^*}\big)\, T_{\pi\to\pi'} \;=\; T_{\pi\to\pi'}\, \big(\pi(g) \otimes I_{V_\pi}\big) \qquad \forall g \in G.$$

Thus each $T_{\pi\to\pi'}$ is an intertwiner from $\pi\otimes\mathbf{1}$ to $\pi'\otimes\mathbf{1}$. By Schur's lemma, if $\pi\not\cong\pi'$ then $T_{\pi\to\pi'}=0$. Hence T is block-diagonal across distinct irreps:

$$T \;\cong\; \bigoplus_{\pi} T_\pi, \qquad T_\pi\in \mathrm{End}(V_\pi\otimes V_\pi^*).$$

Now choose $A_\pi\in\mathrm{End}(V_\pi)$, $B_\pi\in\mathrm{End}(V_\pi^*)$ such that

$$T \;\cong\; \bigoplus_{\pi} A_\pi\otimes B_\pi.$$

Again by Schur's lemma, A_π must be a scalar multiple of I_{V_π}; this scalar can be absorbed into B_π, which gives the desired decomposition in Equation 13. □

B Details on Vector-valued Signals

Fourier description. Peter–Weyl yields the unitary decomposition

$$L^2(G) \;\cong\; \bigoplus_{\pi\in\widehat{G}} V_\pi\otimes V_\pi^*, \qquad L^2(G,V) \;\cong\; \bigoplus_{\pi\in\widehat{G}} V_\pi\otimes\big(V_\pi^*\otimes V\big),$$

where G acts by π on the first tensor factor and trivially on V_π^*, while the fiber transforms by ρ. Accordingly, any bounded linear map $T: L^2(G,V_{\mathrm{in}})\to L^2(G,V_{\mathrm{out}})$ admits a block form

$$\widehat{T} \;\cong\; \Big(\widehat{T}(\pi,\sigma)\Big)_{\pi,\sigma\in\widehat{G}}, \qquad \widehat{T}(\pi,\sigma): V_\sigma\otimes\big(V_\sigma^*\otimes V_{\mathrm{in}}\big)\longrightarrow V_\pi\otimes\big(V_\pi^*\otimes V_{\mathrm{out}}\big).$$

Averaging annihilates all off-diagonal ($\pi\neq\sigma$) blocks and, on each frequency π, orthogonally projects $\widehat{T}(\pi,\pi)$ onto the intertwiner space $\mathrm{Hom}_G\big(\pi^*\otimes\rho_{\mathrm{in}}\,,\,\pi^*\otimes\rho_{\mathrm{out}}\big)$.

Theorem B.1 (Theorem 3.10 restated)**.** *Let $T: L^2(G,V_{\mathrm{in}})\to L^2(G,V_{\mathrm{out}})$ be linear. Then*

$$\widehat{P_{\mathrm{equiv}}(T)} \;\cong\; \bigoplus_{\pi\in\widehat{G}} \Big(I_{V_\pi}\otimes B_\pi\Big), \tag{50}$$

$$B_\pi \;=\; \int_G \big(\pi(g)^*\otimes\rho_{\mathrm{out}}(g)\big)\,\widehat{T}(\pi,\pi)\,\big(\pi(g)\otimes\rho_{\mathrm{in}}(g)^{-1}\big)\,d\lambda(g), \tag{51}$$

with $B_\pi\in\mathrm{Hom}_G\big(\pi^*\otimes\rho_{\mathrm{in}},\,\pi^*\otimes\rho_{\mathrm{out}}\big)$. In particular, every equivariant T is block-diagonal across frequencies and acts as the identity on V_π and as an intertwiner on the fiber–multiplicity space $V_\pi^*\otimes V$.

Sketch. Decompose both domain and codomain via Peter–Weyl and write $\widehat{T}$ in blocks $\widehat{T}(\pi,\sigma)$. Conjugation by $(\tau\otimes\rho)$ restricts, on the (π,π) block, to the representation $\pi^*\otimes\rho_{\mathrm{out}}$ on the codomain multiplicity and $\pi^*\otimes\rho_{\mathrm{in}}$ on the domain multiplicity. Averaging is the orthogonal projection onto the commutant, hence onto $\mathrm{Hom}_G(\pi^*\otimes\rho_{\mathrm{in}},\pi^*\otimes\rho_{\mathrm{out}})$, and kills $\pi\neq\sigma$ by Schur orthogonality. The displayed formula is the explicit Bochner average of that projection. □

C Implementation Details

In this section, we provide additional information on the implementation details of all of our experiments.

C.1 Example: Learned $SO(2)$ invariance

Data generation. Using polar coordinates (r,θ), we sample the inner cloud (blue, label $+1$) by drawing $r\sim\mathrm{Unif}[0,1]$ and $\theta\sim\mathrm{Unif}[0,2\pi)$, and the outer cloud (red, label -1) by drawing $r\sim\mathrm{Unif}[2.3,3]$ and $\theta\sim\mathrm{Unif}\big[-\frac{\pi}{4},\frac{\pi}{4}\big)$.

Feature map and network. We project inputs $(x,y)\in\mathbb{R}^2$ onto circular harmonics up to degree $M=4$ with $C=4$ radial channels as follows: viewing (x,y) as a complex number $z\in\mathbb{C}$ with $r=|z|$ and $\hat{z}=z/r$, define radial basis functions

$$b_n(r) \;=\; \exp\Big(-\frac{(r-c_n)^2}{2\sigma^2}\Big), \quad \sigma=0.5, \quad c_n \text{ uniform in } [0,4],\; n=1,\dots,C.$$

Form the order-m harmonic features by $h^{(m)}(r, \hat{z}) = \big(b_n(r)\,\hat{z}^{\,m}\big)_{n=1}^{C}$ for $m = -M, \dots, M$, and concatenate across m to obtain the embedding

$$H \in \mathbb{C}^{(2M+1)\times C}.$$

We then apply two fully connected complex linear layers

$$L_1 : \mathbb{C}^{(2M+1)\times C} \to \mathbb{C}^{(2M+1)\times C_{\text{hid}}}, \qquad L_2 : \mathbb{C}^{(2M+1)\times C_{\text{hid}}} \to \mathbb{C}^{(2M+1)\times C_{\text{hid}}},$$

followed by an $\mathrm{SO}(2)$-equivariant tensor product:

$$h'_{m_{\text{out}}} = \sum_{m_1+m_2=m_{\text{out}}} h_{m_1}\, h_{m_2},$$

with complex multiplication applied channel-wise. We then extract the invariant component h'_0 and pass its real part through a final real-valued linear head $L_{\text{final}} : \mathbb{R}^{C_{\text{hid}}} \to \mathbb{R}$ to produce the scalar logit.

We then train a new model for each combination of $\lambda_G, \lambda_\perp$ (see Figure 3) using the Adam optimiser Kingma & Ba (2014) for 200 epochs with a learning rate of 0.003. We use a binary cross-entropy loss as task-specific loss.

Angular perturbation experiment. To study the interaction between the projection regulariser and violations of exact $\mathrm{SO}(2)$ symmetry, we construct a family of "wavey" ring datasets parameterised by an amplitude $\sigma_\perp \geq 0$. For each $\sigma_\perp$ we independently sample angles $\theta_+, \theta_- \sim \mathrm{Unif}[0, 2\pi)$ and define class-conditional radii

$$r_+(\theta_+) = r_{\text{in}} + \sigma_\perp \sin(f\theta_+) + \epsilon_{\text{in}}, \qquad r_-(\theta_-) = r_{\text{out}} + \sigma_\perp \sin(f\theta_-) + \epsilon_{\text{out}},$$

with $(r_{\text{in}}, r_{\text{out}}) = (1.1, 2.2)$, frequency $f = 5$ and independent jitters $\epsilon_{\text{in}} \sim \mathrm{Unif}[-b_{\text{in}}, b_{\text{in}}]$, $\epsilon_{\text{out}} \sim \mathrm{Unif}[-b_{\text{out}}, b_{\text{out}}]$ for $(b_{\text{in}}, b_{\text{out}}) = (0.15, 0.22)$. Mapping $(r_\pm, \theta_\pm)$ to Cartesian coordinates yields two noisy rings labelled $+1$ (inner) and -1 (outer). In Figure 5, we consider $\sigma_\perp \in \{0, 0.5, 0.75, 1.0\}$, sample 350 points per class, and split the data into 80% training and 20% test. For each $(\sigma_\perp, \lambda_\perp)$ we then train (i) the approximately $SO(2)$-invariant architecture described above (blue lines), and (ii) a plain real-valued MLP on the raw coordinates (orange).

We see that even for a fixed value of $\lambda_\perp$, the regulariser allows us to capture different effective levels of invariance as the data depart from rotational symmetry; see, for instance, the row with $\lambda_\perp = 1.0$, where the learned classifier remains nearly invariant for small $\sigma_\perp$ and gradually departs from invariance as the angular modulation strengthens. For strongly broken $\mathrm{SO}(2)$ symmetry (e.g. $\sigma_\perp = 1.0$), the decision boundary remains "as radially symmetric as possible": away from the perturbed regions the contours revert to circular rings, and in the region between the two classes, around each arm of the star-shaped pattern, the classifier exhibits consistent behaviour across angles.

C.2 Imperfectly Symmetric Dynamical Systems

For each baseline, relaxed group convolution (`RGroup`) and relaxed steerable CNN (`RSteer`), and for each symmetry setting, we conduct a hyperparameter sweep over learning rate, batch size, hidden width, number of layers, and the number of rollout steps used to compute prediction errors during training, using the same search ranges as Wang et al. (2022c) (see Table 3). We also tune the number of filter banks for group-convolution models and the coefficient for the non-equivariance penalty $\lambda_\perp$ for relaxed weight-sharing models. The input sequence length is fixed to 10. To ensure a fair comparison, we cap the total number of trainable parameters for every model at no more than 10^7.

Table 3: Hyperparameter tuning range for the asymetric smoke simulation data.

LR	Batch size	Hid-dim	Num-layers	Num-banks	#Steps for Backprop	$\lambda_\perp$
$10^{-2} \sim 10^{-5}$	$8 \sim 64$	$64 \sim 512$	$3 \sim 6$	$1 \sim 4$	$3 \sim 6$	$0, 10^{-2}, 10^{-4}, 10^{-6}$

C.3 CT Scan Metal Artifact Reduction

C.3.1 Hyperparameters

For the most part, we use the same hyperparameters as Bai et al. (2025). We train for 80 epochs with a batch size of 12 for the baselines and our projection-based regulariser, and a batch size of 4 for the sample-based regulariser. We set the patch size at 256×256. Optimization uses Adam Kingma & Ba (2014) (β_1=0.5, β_2=0.999) with initial learning

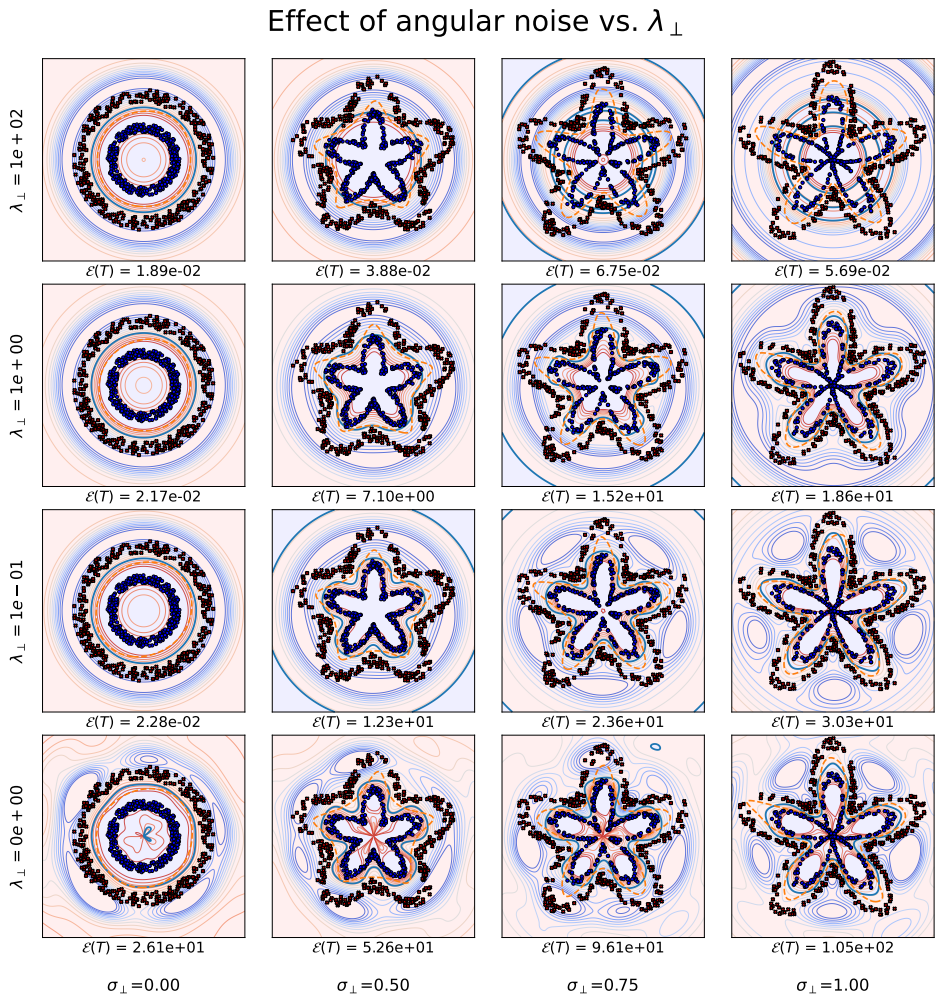


Figure 5: Effect of angular perturbations and projection strength. Columns vary the angular wave amplitude $\sigma_\perp$, rows vary the non-equivariant penalty weight $\lambda_\perp$. Blue contours show level sets of the approximately $\mathrm{SO}(2)$-invariant network and points denote training samples. Orange dashed lines are the decision boundary of a non-equivariant MLP. The value $\mathcal{E}(T)$ underneath each panel is the empirical invariance defect, demonstrating that larger $\lambda_\perp$ keeps the classifier close to invariant even as the Bayes decision boundary becomes increasingly angle-dependent.

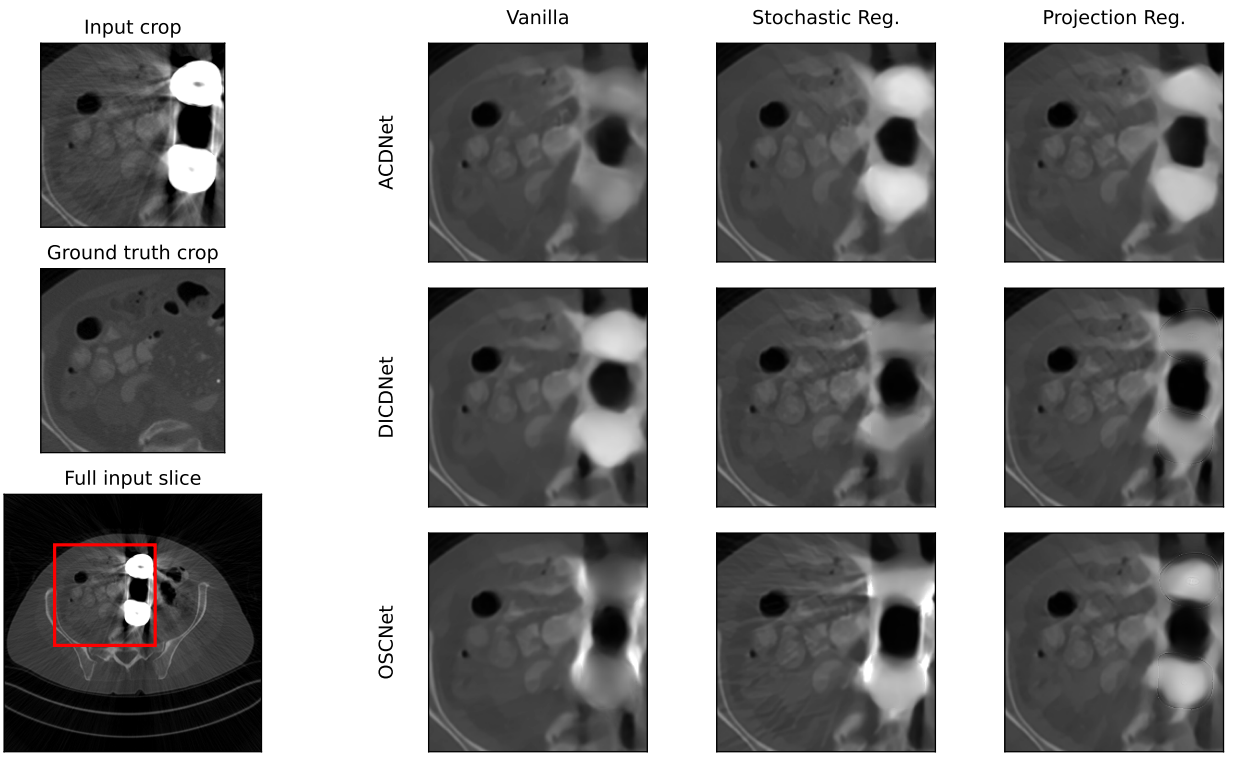


Figure 6: Qualitative comparison of the baseline methods (left column) with each of the sample-based (middle column) and our projection-based regulariser (right column) on the metal artefact reduction task. We show a cropped pelvic slice containing two metallic implants that generate artefacts.

Table 4: Hyperparameters for the CT-MAR experiments.

Parameter	**Value**
N (feature maps)	8
N_p (concat channels)	35
d (dict. filters)	32
Residual blocks / ResNet	3
Stages T	10

rate η_0=2×10^{-4} and a MultiStepLR scheduler (milestones at epochs $\{50, 100, 150, 200\}$, decay factor γ=0.5). The model hyperparameters are summarised in Table 4.

The scalar weight for sample-based regulariser is set at 10^6. To set ours, we performed a hyperparameter sweep over the set $\{1.0, 10^{-1}, \ldots, 10^{-6}\}$ and chose $\lambda_G = 1.0$.

C.3.2 Projection onto the C_4-equivariant kernel subspace

We consider steerable CNN layers whose input and output feature spaces are arranged in orientation groups of four (regular-representation channels) for the discrete rotation group $C_4 = \{0, 1, 2, 3\}$ (multiples of 90°). Let

$$K \in \mathbb{R}^{C'_{\text{out}} \times C'_{\text{in}} \times 4 \times 4 \times s \times s}$$

denote an $s \times s$ convolution kernel with output block index $p \in \{1, \ldots, C'_{\text{out}}\}$, input block index $q \in \{1, \ldots, C'_{\text{in}}\}$, orientation indices $\alpha, \beta \in \{0, 1, 2, 3\}$, and spatial indices $(i, j) \in \{0, \ldots, s-1\}^2$. Let S be the 4×4 cyclic-shift matrix so that the channel representations of C_4 act by $\rho_{\text{out}}(r) = S^r$ and $\rho_{\text{in}}(r) = S^r$ for $r \in \{0, 1, 2, 3\}$. Write rot_r for rotation of the spatial kernel by $90°r$ (with exact index permutation on the discrete grid).

The natural action of C_4 on kernels combines spatial rotation with orientation-channel permutations:

$$\big(\mathcal{A}(r)\,K\big) \;=\; \rho_{\text{out}}(r)\,\big(\text{rot}_r K\big)\,\rho_{\text{in}}(r)^{-1} \;=\; S^r\,\big(\text{rot}_r K\big)\,S^{-r}. \tag{52}$$

The orthogonal projector onto this subspace is the (finite) Haar average of the action:

$$P(K) \;=\; \tfrac{1}{4}\sum_{r=0}^{3} S^r\,\big(\text{rot}_r K\big)\,S^{-r}. \tag{53}$$

Index-wise, for any $(p, q, \alpha, \beta, i, j)$, this reads

$$\big[P(K)\big]_{p,\alpha;\,q,\beta}[i,j] \;=\; \tfrac{1}{4}\sum_{r=0}^{3} \big[\text{rot}_r K\big]_{p,\alpha-r;\,q,\beta-r}[i,j]. \tag{54}$$

Since equation 53 is the average of unitary (permutation + rotation) actions, P is an orthogonal projector: $P^2 = P$ and $P^\top = P$. In practice, equation 53 yields an efficient, exact implementation requiring only four 90° rotations and two inexpensive orientation-channel permutations per term.

D Sensitivity with respect to λ_G and $\lambda_\perp$

We study the sensitivity of our method to the scalar weights λ_G and $\lambda_\perp$ through two ablation experiments. First, we repeat the experiment from Section 4.1 on approximate $SO(2)$ invariance in 2D for $\lambda_G, \lambda_\perp \in \{0, 0.001, 0.01, 0.1\}$; the resulting decision boundaries are shown in Figure 7. When the penalty on the orthogonal component dominates (e.g. $\lambda_\perp = 0.1$ and $\lambda_G \in \{0, 0.001, 0.01\}$), the decision boundary becomes essentially rotationally invariant. In the regime $\lambda_\perp \approx \lambda_G$, the regulariser effectively reduces to standard Tikhonov (ℓ_2) regularisation and no longer induces a geometric inductive bias. For $\lambda_\perp < \lambda_G$, the learned level sets increasingly depend on angular information.

Training setup. Additionally, we study learned translation equivariance on a perfectly translation-equivariant task: image classification on MNIST Deng (2012) and CIFAR Krizhevsky et al.. Figure 8 reports the classification accuracy for different values of λ_G and $\lambda_\perp$. As expected in this setting, models with a stronger equivariance bias perform better: the best results are generally obtained for $\lambda_G = 0$, and accuracy increases as $\lambda_\perp$ grows. In Figure 9, we show the corresponding equivariance defect for each $(\lambda_G, \lambda_\perp)$ pair. This defect remains largely unchanged when varying λ_G at fixed $\lambda_\perp$, and decreases sharply as $\lambda_\perp$ increases, consistent with the role of $\lambda_\perp$ as the primary control on the non-equivariant component.

E Sensitivity with respect to norm

In this ablation, we study the impact of the choice of matrix norm in the projection regulariser. We consider the following norms. First, the spectral norm

$$\|A\|_2 = \max_{\|x\|_2=1} \|Ax\|_2, \tag{55}$$

which is equal to the largest singular value of A. Second, the Frobenius norm

$$\|A\|_F = \sqrt{\sum_{i,j} a_{i,j}^2}. \tag{56}$$

Third, the (entrywise) infinity norm

$$\|A\|_\infty = \max_{i,j} |a_{i,j}|. \tag{57}$$

Finally, we consider mixed (p, q)-norms, defined row-wise as

$$\|A\|_{p,q} = \left(\sum_i \left(\sum_j |a_{i,j}|^p\right)^{\frac{q}{p}}\right)^{\frac{1}{q}}, \tag{58}$$

for $p, q \in \{1, 2, 3\}$. The corresponding results are reported in Table 5. We can see that the choice of norm has only a modest effect on both computational cost and reconstruction quality. Training and inference throughput, as well as epoch time, are nearly identical across all norms, except for the spectral norm, which is about 10–15% slower per

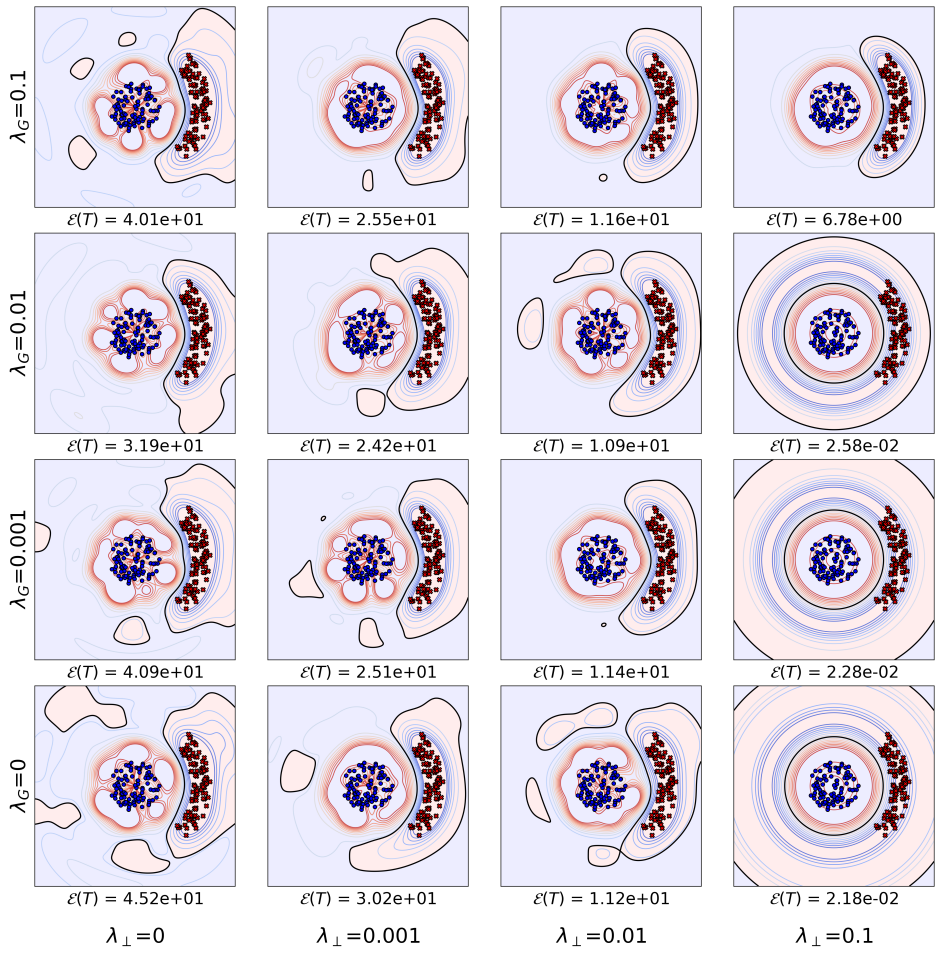


Figure 7: Controlling the degree of learned $SO(2)$ invariance by varying the values of λ_G and $\lambda_\perp$ over the grid $\{0, 0.001, 0.01, 0.1\}$.

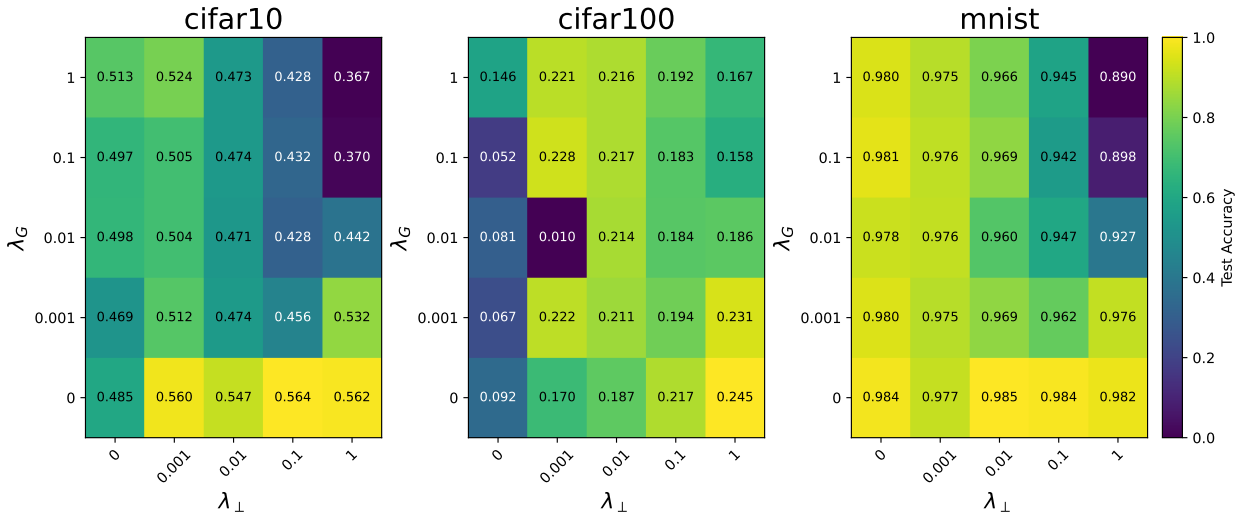


Figure 8: Classification accuracy on the CIFAR and MNIST datasets for models trained with varying values of λ_G and $\lambda_\perp$.

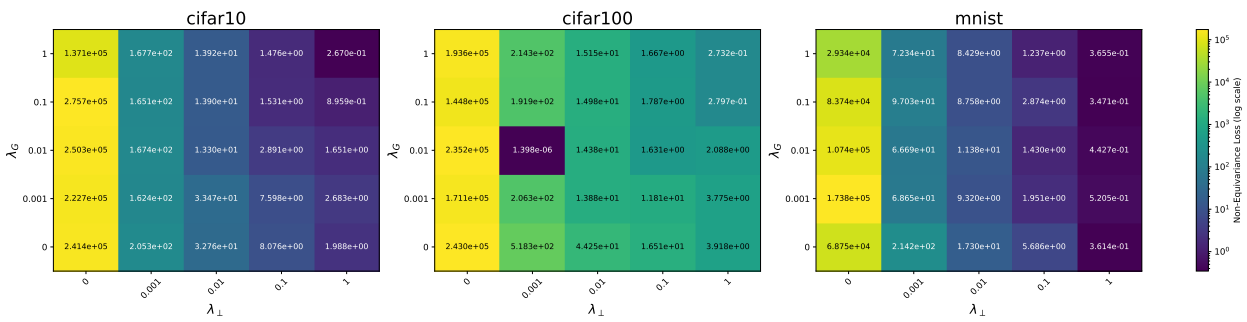


Figure 9: Equivariance defect for models trained on the CIFAR and MNIST classification tasks for varying values of λ_G and $\lambda_\perp$.

Table 5: Results for the CT scan metal artifact reduction task from Section 4.3 for different matrix norms. We report throughput during training and inference as well as total epoch time; the performance metrics are PSNR/SSIM. We consider the spectral, Frobenius (which we use by default in Section 4.3) and infinity norms, as well as the (p, q)-norms for $p, q \in \{1, 2, 3\}$.

Norm	Throughput (no./GPU-s)		Epoch	AAPM	
	Train ↑	Inference ↑	time (s) ↓	PSNR ↑	SSIM ↑
Spectral	6.59	10.16	877	39.25	0.9318
Frobenius	7.22	10.11	778	38.48	**0.9457**
Infinity	7.73	10.12	777	35.61	0.9153
$(1, 1)$	7.63	10.13	785	35.57	0.8864
$(1, 2)$	7.12	10.14	785	38.05	0.9365
$(1, 3)$	7.12	10.13	785	38.67	0.9391
$(2, 1)$	7.65	10.14	783	39.33	0.9299
$(2, 2)$	7.65	10.13	783	38.24	0.9430
$(2, 3)$	7.61	10.14	786	38.18	0.9299
$(3, 1)$	7.59	10.14	787	39.10	0.9304
$(3, 2)$	7.32	10.10	810	39.54	0.9322
$(3, 3)$	7.18	10.16	780	37.86	0.9346

epoch, as expected given the need to estimate the largest singular value. In terms of image quality, several choices yield very similar PSNR/SSIM, with the Frobenius and (p, q)-norms for $(p, q) \in (2, 2), (1, 3), (3, 3)$ all lying within roughly 1 dB PSNR and 0.01 SSIM of each other. Norms that emphasise elementwise extremal behaviour, such as the infinity norm and the $(1, 1)$-norm, lead to clear degradation in both PSNR and SSIM, indicating that these penalties are too stiff and effectively underfit the reconstruction task. Since the spectral norm brings no systematic performance gains while incurring a noticeable runtime overhead, and more aggressive entrywise norms harm reconstruction quality, we adopt the Frobenius norm as our default in Section 4.3.

F Declaration of LLM use

We used LLMs to aid in the writing process for proof-reading, spell checking, and polishing writing.