Supplementary Information

Group-Algebraic Tensors: Provably-optimal Equivariant Learning and Physical Symmetry Discovery

Hoyos, Ubaru, Huh, Kalantzis, Clarkson, Kilmer, Avron, Horesh

Contents

1 Mathematical Foundations

1.1 Notation

- G: finite group of order $n = |G|$ with identity e.
- $\hat{G}$: set of equivalence classes of irreducible unitary representations.
- $d_\rho = \dim(\rho)$ for $\rho \in \hat{G}$.
- $\mathcal{A}(:,:,g)$: frontal slice at group element g. $\mathcal{A}_{ij} = \mathcal{A}(i,j,:)$: tube at indices i, j.

1.2 The Group Algebra

Definition 1.1 (Group Algebra)**.** $\mathbb{R}[G]$ *is the vector space of formal sums* $\sum_{g\in G} a_g g$ *with convolution product:*

$$\left(\sum_g a_g g\right) \cdot \left(\sum_h b_h h\right) = \sum_{c\in G} \left(\sum_{g\in G} a_g b_{g^{-1}c}\right) c, \tag{1}$$

where a_g *and* b_h *are scalar coefficients in* $\mathbb{R}$*.*

2 The Convolution Tensor, Spectral Decomposition, and Generalized Fourier Matrix

Definition 2.1 (Convolution Tensor)**.** $\mathcal{T} \in \mathbb{R}^{n\times n\times n}$ *is defined by* $\mathcal{T}(a,b,c) = \delta(g_a g_b = g_c)$ *(also known as the structure constants of the group algebra).*

Proposition 2.2 (Properties of $\mathcal{T}$)**.** *(i) Associativity:*

$$\sum_d \mathcal{T}(a,b,d)\,\mathcal{T}(d,c,e) \;=\; \sum_d \mathcal{T}(a,d,e)\,\mathcal{T}(b,c,d).$$

(ii) Identity: $\mathcal{T}(e,b,c) = \delta_{bc}$*.*

(iii) Each slice $\mathcal{T}(a,:,:)$ *is a permutation matrix.*

Proof. (i) follows from associativity of group multiplication; both sides equal $\delta(g_a g_b g_c = g_e)$. (ii)–(iii) follow from the group axioms. □

Definition 2.3 (Generalized Fourier Transform Matrix)**.** $F_G \in \mathbb{C}^{n\times n}$ *is defined row-wise: the row* $F_G(g,:)$ *is given by the concatenation* $[\mathrm{rvec}(\rho_1(g)), \ldots, \mathrm{rvec}(\rho_\ell(g))]$*, where* $\mathrm{rvec}(\rho_i(g))$ *denotes the row-vectorization of the matrix* $\rho_i(g)$ *for each* $\rho_i \in \hat{G}$*. For abelian groups,* F_G *is a generalized Fourier matrix and for cyclic groups, it reduces to the standard DFT matrix. For non-abelian groups* F_G *is invertible but in general neither unitary nor block-unitary.*

Theorem 2.4 (Peter–Weyl Spectral Decomposition)**.**

$$\mathcal{T}(a,b,c) = \sum_{i,j,k} \mathcal{C}(i,j,k)\, F_G(a,i)\, F_G(b,j)\, F_G^{-1}(c,k) \tag{2}$$

where $\mathcal{C}$ *is a core tensor that is typically sparse. For abelian groups,* $\mathcal{C}$ *is diagonal.*

Proof. By the Peter–Weyl theorem, $\{\sqrt{d_\rho}\rho_{ij}(g) : \rho \in \hat{G}\}$ is an orthonormal basis for $L^2(G)$. Convolution becomes block multiplication in the Fourier domain: $\widehat{(f * h)}(\rho) = \hat{f}(\rho) \cdot \hat{h}(\rho)$. Writing this in index notation using the row-vectorization construction of F_G yields the spectral decomposition with core $\mathcal{C}$ determined by the Fourier-domain multiplication structure. □

Corollary 2.5. *For $G = \mathbb{Z}_n$: $F_G = \mathrm{DFT}_n$ and $\mathcal{C}$ is diagonal, recovering the circular convolution theorem (i.e., the t−product).*

3 The Group Fourier Transform of a Tensor

Definition 3.1 (Group Fourier Transform of a Tensor)**.** *For $\mathcal{A} \in \mathbb{R}^{\ell \times m \times n}$, the Group Fourier transform $\mathcal{F}_G$ assigns to each irrep $\rho \in \hat{G}$ the $\ell d_\rho \times m d_\rho$ block matrix*

$$\hat{\mathcal{A}}(:,:,\rho) = \sum_{g \in G} \mathcal{A}(i,j,g)\, \rho(g), \qquad \text{with } (i,j) \text{ indexing the } \ell \times m \text{ blocks.} \tag{3}$$

The full Fourier representation is the block-diagonal matrix $\oplus_{\rho \in \hat{G}} \hat{\mathcal{A}}(:,:,\rho)$.

Proposition 3.2 (Group Fourier Inversion Theorem)**.** *Given $\hat{\mathcal{A}}(:,:,\rho)$ for each $\rho \in \hat{G}$, the inverse Group Fourier transform recovers*

$$\mathcal{A}(i,j,g) = \sum_{\rho \in \hat{G}} \frac{d_\rho}{n} \operatorname{Tr}\left[\hat{\mathcal{A}}(i,j,\rho)\, \rho(g)^H\right]. \tag{4}$$

Proof. Follows by applying the standard Peter–Weyl inversion theorem for each fixed i and j. □

Remark 3.3. *The two notions of "Fourier transform" used in the paper are related as follows. The contraction of $\mathcal{A}$ with F_G along its group dimension (as in SI Section 2) and the block-diagonal form $\oplus_\rho \hat{\mathcal{A}}(:,:,\rho)$ (Definition 3.1) are equivalent via the row-vectorization reshaping* rvec *used in the construction of F_G. The block-diagonal form is used for the SVD computation (one standard matrix SVD per irrep block); the F_G-contraction form is used for the product computation and for establishing the spectral decomposition of $\mathcal{T}$.*

Algorithm 1 $\star_G$ product (used by every $\star_G$-based method)

Require: $\mathcal{A} \in \mathbb{R}^{\ell \times m \times n}$, $\mathcal{B} \in \mathbb{R}^{m \times p \times n}$, generalized Fourier matrix $F_G \in \mathbb{C}^{n \times n}$, irrep block sizes $\{d_\rho\}_{\rho \in \hat{G}}$

Ensure: $\mathcal{C} = \mathcal{A} \star_G \mathcal{B} \in \mathbb{R}^{\ell \times p \times n}$

1: $\hat{\mathcal{A}} \leftarrow \mathcal{A} \times_3 F_G$ ▷ Generalized Fourier transform along group dimension
2: $\hat{\mathcal{B}} \leftarrow \mathcal{B} \times_3 F_G$
3: **for** $\rho \in \hat{G}$ **in parallel do**
4: extract $\hat{\mathcal{A}}_\rho \in \mathbb{C}^{\ell d_\rho \times m d_\rho}$, $\hat{\mathcal{B}}_\rho \in \mathbb{C}^{m d_\rho \times p d_\rho}$ from block-diagonal form
5: $\hat{\mathcal{C}}_\rho \leftarrow \hat{\mathcal{A}}_\rho \cdot \hat{\mathcal{B}}_\rho$ ▷ ordinary matrix product
6: **end for**
7: assemble $\hat{\mathcal{C}}$ from $\{\hat{\mathcal{C}}_\rho\}_\rho$ in block-diagonal form
8: $\mathcal{C} \leftarrow \mathrm{Re}(\hat{\mathcal{C}} \times_3 F_G^{-1})$ ▷ inverse Generalized Fourier transform; imaginary part is zero up to roundoff for real inputs
9: **return** $\mathcal{C}$

4 The $\star_G$ Algebra: Properties and Proofs

Definition 4.1 ($\star_G$ Product)**.** *For $\mathcal{A} \in \mathbb{R}^{\ell \times m \times n}$, $\mathcal{B} \in \mathbb{R}^{m \times p \times n}$:*

$$(\mathcal{A} \star_G \mathcal{B})_{ij}(c) = \sum_k \sum_{a \in G} \mathcal{A}_{ik}(a) \mathcal{B}_{kj}(a^{-1}c). \tag{5}$$

Equivalently: $\widehat{(\mathcal{A} \star_G \mathcal{B})}(:,:,\rho) = \hat{\mathcal{A}}(:,:,\rho) \cdot \hat{\mathcal{B}}(:,:,\rho)$ *for each irrep $\rho \in \hat{G}$.*

Proposition 4.2 (Algebraic Properties)**.** *(i) Associativity. (ii) Distributivity. (iii) Identity:* $\mathcal{I}(:,:,e) = I$, $\mathcal{I}(:,:,g \neq e) = 0$. *(iv)* $(\mathcal{A} \star_G \mathcal{B})^H = \mathcal{B}^H \star_G \mathcal{A}^H$.

Proof. All follow from the corresponding matrix properties applied per-irrep in the Fourier domain (Definition 3.1), plus linearity and invertibility of the Fourier transform. □

5 The $\star_G$-SVD: Full Proof of Optimality

Theorem 5.1 ($\star_G$-SVD Existence)**.** *Every $\mathcal{A} \in \mathbb{R}^{\ell \times m \times n}$ admits $\mathcal{A} = \mathcal{U} \star_G \mathcal{S} \star_G \mathcal{V}^H$ where $\mathcal{U}, \mathcal{V}$ are $\star_G$-unitary and $\mathcal{S}$ is f-diagonal.*

Proof. For each irrep $\rho \in \hat{G}$, $\hat{\mathcal{A}}(:,:,\rho)$ is a standard matrix admitting SVD: $\hat{\mathcal{A}}(:,:,\rho) = U_\rho \Sigma_\rho V_\rho^H$. Setting $\hat{\mathcal{U}}(:,:,\rho) = U_\rho$, $\hat{\mathcal{S}}(:,:,\rho) = \Sigma_\rho$, $\hat{\mathcal{V}}(:,:,\rho) = V_\rho$ and applying the inverse Fourier transform yields the $\star_G$-SVD. Unitarity: $\widehat{\mathcal{U}^H \star_G \mathcal{U}}(:,:,\rho) = U_\rho^H U_\rho = I$ for all ρ, so $\mathcal{U}^H \star_G \mathcal{U} = \mathcal{I}$. □

Algorithm 2 $\star_G$-SVD

Require: $\mathcal{A} \in \mathbb{R}^{\ell \times m \times n}$
Ensure: $\mathcal{U}, \mathcal{S}, \mathcal{V}$
1: Compute $\hat{\mathcal{A}}(:,:,\rho)$ for all $\rho \in \hat{G}$ using $\mathcal{F}_G$ (Definition 3.1)
2: **for** $\rho \in \hat{G}$ **do**
3: $\quad [U_\rho, \Sigma_\rho, V_\rho] \leftarrow \mathrm{SVD}(\hat{\mathcal{A}}(:,:,\rho))$
4: **end for**
5: Apply $\mathcal{F}_G^{-1}$ to $\{U_\rho\}$, $\{\Sigma_\rho\}$, $\{V_\rho\}$ to obtain $\mathcal{U}, \mathcal{S}, \mathcal{V}$

Theorem 5.2 (Eckart–Young for $\star_G$)**.** *Let $\mathcal{A} = \mathcal{U} \star_G \mathcal{S} \star_G \mathcal{V}^H$ with singular tubes ordered by $\|\mathbf{s}_1\|_F \geq \cdots \geq \|\mathbf{s}_r\|_F$. The rank-$k$ truncation $\mathcal{A}_k$ satisfies:*

$$\|\mathcal{A} - \mathcal{A}_k\|_F^2 = \sum_{i=k+1}^{r} \|\mathbf{s}_i\|_F^2 \leq \|\mathcal{A} - \mathcal{B}\|_F^2 \tag{6}$$

for any $\mathcal{B}$ with $\star_G$-rank $\leq k$.

Proof. **Step 1 (Parseval).** By the generalized Fourier transform's isometry (Peter–Weyl):

$$\|\mathcal{A} - \mathcal{B}\|_F^2 = \sum_{\rho \in \hat{G}} \frac{d_\rho}{n} \|\hat{\mathcal{A}}(:,:,\rho) - \hat{\mathcal{B}}(:,:,\rho)\|_F^2. \tag{7}$$

Step 2 (Per-irrep Eckart–Young). If $\star_G$-rank$(\mathcal{B}) \leq k$, then $\mathrm{rank}(\hat{\mathcal{B}}(:,:,\rho)) \leq k$ for each ρ. By the classical Eckart–Young theorem:

$$\|\hat{\mathcal{A}}(:,:,\rho) - \hat{\mathcal{A}}_k(:,:,\rho)\|_F^2 \leq \|\hat{\mathcal{A}}(:,:,\rho) - \hat{\mathcal{B}}(:,:,\rho)\|_F^2. \tag{8}$$

Step 3 (Summation). Summing over $\rho \in \hat{G}$:

$$\|\mathcal{A} - \mathcal{A}_k\|_F^2 = \sum_{\rho \in \hat{G}} \frac{d_\rho}{n} \sum_{i=k+1}^{r} \sigma_i(\rho)^2 = \sum_{i=k+1}^{r} \underbrace{\sum_{\rho} \frac{d_\rho}{n} \sigma_i(\rho)^2}_{=\|\mathbf{s}_i\|_F^2} \leq \|\mathcal{A} - \mathcal{B}\|_F^2. \tag{9}$$

□

Remark 5.3. *This is* exact *optimality. By contrast: CP decomposition is NP-hard to compute optimally; Tucker/HOSVD provides only quasi-optimal guarantees with factor $\sqrt{d}$ [1]; tensor-train has no global optimality. The $\star_G$-SVD achieves both polynomial-time computation and exact optimality by leveraging group structure.*

6 Product Groups: Full Proof

Theorem 6.1 (Product Group Ring Isomorphism)**.** *For $G = G_1 \times \cdots \times G_d$: (i) $\mathbb{K}_G \cong \mathbb{K}_{G_1} \otimes \cdots \otimes \mathbb{K}_{G_d}$. (ii) $\mathcal{T}_G = \mathcal{T}_{G_1} \otimes \cdots \otimes \mathcal{T}_{G_d}$. (iii) $F_G = F_{G_1} \otimes \cdots \otimes F_{G_d}$.*

Proof. **(i)** Standard: $\mathbb{R}[G_1 \times \cdots \times G_d] \cong \mathbb{R}[G_1] \otimes \cdots \otimes \mathbb{R}[G_d]$ [2].

(ii) The product group multiplication $(a_1, \ldots, a_d)(b_1, \ldots, b_d) = (a_1 b_1, \ldots, a_d b_d)$ gives

$$\mathcal{T}_G(\mathbf{a}, \mathbf{b}, \mathbf{c}) = \prod_{i=1}^{d} \mathcal{T}_{G_i}(a_i, b_i, c_i) = (\mathcal{T}_{G_1} \otimes \cdots \otimes \mathcal{T}_{G_d})(\mathbf{a}, \mathbf{b}, \mathbf{c}). \tag{10}$$

(iii) Irreps of $G_1 \times \cdots \times G_d$ are tensor products $\rho_1 \otimes \cdots \otimes \rho_d$. Matrix elements factor: $(\rho_1 \otimes \cdots \otimes \rho_d)(g_1, \ldots, g_d) = \rho_1(g_1) \otimes \cdots \otimes \rho_d(g_d)$. By the rvec construction of F_G and the mixed-product property of Kronecker products, $F_G = F_{G_1} \otimes \cdots \otimes F_{G_d}$. □

Corollary 6.2 (2D Frequency Resolution)**.** *For $G = \mathbb{Z}_{n_1} \times \mathbb{Z}_{n_2}$, the Fourier transform $F_G = \mathrm{DFT}_{n_1} \otimes \mathrm{DFT}_{n_2}$ computes a 2D DFT, resolving coupled frequencies (f_1, f_2) that are invisible to either factor alone.*

7 Invariant Feature Extraction Algorithm

The $\star_G$-feature extractor used in every linear and Neural-$\star_G$ experiment of this paper is given in Algorithm 3. The seven feature blocks (a)–(g) below correspond to the seven concatenated columns of the output. Block (a) is the DC component; (b) the AC standard deviation; (c) the global per-frequency power; (d) the per-row Fourier power restricted to the first K equivariant rows; (e) the singular tube norms from the $\star_G$-SVD; (f) the rows of $\mathcal{X}$ that are constant under the group action (recovered by row-variance thresholding); and (g) four spectral statistics of the unfolded matrix $\mathcal{X}_{(1)}$. Every feature is exactly G-invariant; proofs are in SI Section 7.

8 Equivariance Proofs

Proposition 8.1 (Equivariance of $\star_G$)**.** $(g \cdot \mathcal{A}) \star_G \mathcal{B} = g \cdot (\mathcal{A} \star_G \mathcal{B}) = \mathcal{A} \star_G (g \cdot \mathcal{B})$.

Algorithm 3 $\star_G$ invariant feature extraction (`extractStarGFeatures.m`)

Require: batch $\mathcal{X} \in \mathbb{R}^{N \times n_f \times n}$, group algebra G, optional normalization parameters Θ from training

Ensure: feature matrix $\Phi \in \mathbb{R}^{N \times d_{\text{feat}}+1}$ (+1 for the unregularized intercept) and updated Θ

if Θ is not provided **then**
  $(p, q) \leftarrow \arg\max_{p \cdot q \le n_f} \min(p, q)$ ▷ best near-square reshape for $\star_G$-SVD
  $n_{\text{svd}} \leftarrow \min(p, q)$
  $\sigma_{\text{row}} \leftarrow \text{var}_g(\mathcal{X}(1, :, :))$; $\text{inv\_mask} \leftarrow \sigma_{\text{row}} < 10^{-8} \cdot \max(\sigma_{\text{row}})$ ▷ rows constant under G
  $\text{eq\_idx} \leftarrow \{j : \text{inv\_mask}_j = \text{false}\}$; $K \leftarrow \min(14, |\text{eq\_idx}|)$
end if
$\bar{\mathcal{X}}_{ij} \leftarrow \frac{1}{n} \sum_g \mathcal{X}(i, j, g)$ ▷ (a) DC component, $N \times n_f$
$\sigma_{ij} \leftarrow \text{std}_g \mathcal{X}(i, j, g)$ ▷ (b) AC energy, $N \times n_f$
$\hat{\mathcal{X}} \leftarrow \mathcal{X} \times_3 F_G$ ▷ Generalized Fourier transform along group dim
$P^{\text{col}}_{ik} \leftarrow \sum_j |\hat{\mathcal{X}}(i, j, k)|^2$ ▷ (c) per-frequency power, $N \times n$
$P^{\text{row}}_{i,(r-1)n+k} \leftarrow |\hat{\mathcal{X}}(i, \text{eq\_idx}_r, k)|^2$ ▷ (d) per-row Fourier power, $N \times Kn$
for $i = 1, \ldots, N$ **do**
  $X_i \leftarrow$ pad $\mathcal{X}(i, :, :)$ to $p \times q \times n$
  $[\,\cdot\,, \mathcal{S}_i, \cdot\,] \leftarrow \text{starG\_SVD}(X_i)$ ▷ (e) singular tubes
  $T_{i,k} \leftarrow \|\mathcal{S}_i(k, k, :)\|_F$ for $k = 1, \ldots, n_{\text{svd}}$; sort $T_{i,:}$ descending
  $V_{ij} \leftarrow \mathcal{X}(i, j, 1)$ for j with inv_mask_j ▷ (f) direct invariants
  $\Sigma_i \leftarrow \text{svd}(\mathcal{X}_{(1)}(i, :, :))$ ▷ (g) spectral statistics
  $S_i \leftarrow [\sum \Sigma_i,\ \Sigma_{i,1},\ \Sigma_{i,1}/\Sigma_{i,\text{end}},\ -\sum_k \tilde{\Sigma}_{i,k} \log \tilde{\Sigma}_{i,k}]$ where $\tilde{\Sigma}_i = \Sigma_i / \sum \Sigma_i$
end for
$\Phi \leftarrow [\bar{\mathcal{X}} \mid \sigma \mid P^{\text{col}} \mid P^{\text{row}} \mid T \mid V \mid S]$
replace NaN/Inf with 0
if Θ not provided **then**
  $\text{keep} \leftarrow \{j : \text{std}(\Phi_{:,j}) \ge 10^{-8}\}$
  $\mu \leftarrow \text{mean}(\Phi_{:,\text{keep}})$; $s \leftarrow \text{std}(\Phi_{:,\text{keep}})$; $s_j \leftarrow \max(s_j, 1)$
  store $\Theta = (p, q, n_{\text{svd}}, \text{inv\_mask}, \text{eq\_idx}, K, \text{keep}, \mu, s)$
end if
$\Phi \leftarrow (\Phi_{:,\text{keep}} - \mu)/s$
$\Phi \leftarrow [\,\mathbf{1}_N \mid \Phi\,]$ ▷ prepend unregularized intercept
return (Φ, Θ)

Proof. By definition of the group action and the $\star_G$ product,

$$\big((g \cdot \mathcal{A}) \star_G \mathcal{B}\big)_{ij}(h) = \sum_k \sum_{a \in G} \mathcal{A}_{ik}(g^{-1}a)\, \mathcal{B}_{kj}(a^{-1}h).$$

Substituting $a' = g^{-1}a$ (so $a = ga'$ and $a^{-1} = a'^{-1}g^{-1}$),

$$= \sum_k \sum_{a' \in G} \mathcal{A}_{ik}(a')\, \mathcal{B}_{kj}(a'^{-1}g^{-1}h)$$
$$= (\mathcal{A} \star_G \mathcal{B})_{ij}(g^{-1}h) = \big(g \cdot (\mathcal{A} \star_G \mathcal{B})\big)_{ij}(h).$$

The right-equivariance identity follows by the symmetric argument applied to the second factor. □

Corollary 8.2 (Invariance of Features)**.** *The following are invariant under $g \cdot X$: (i) $\|\mathbf{s}_i\|_F$ (singular tube norms); (ii) $\|\hat{X}(:, :, \rho)\|_F^2$ (per-irrep Fourier power); (iii) $\bar{X}_j = \frac{1}{n} \sum_g X(j, g)$ (DC component).*

Proof. (i) The group action permutes frontal slices; the SVD is computed per-irrep where the action multiplies each block by a unitary, preserving singular values. (ii) The group action shifts $g \to g'g$ inside the sum defining $\hat{X}(:,:,\rho)$, multiplying the block by $\rho(g')$, which is unitary, leaving the Frobenius norm unchanged. (iii) $\frac{1}{n}\sum_g X(j,g'g) = \frac{1}{n}\sum_h X(j,h) = \bar{X}_j$. □

9 Extended Data

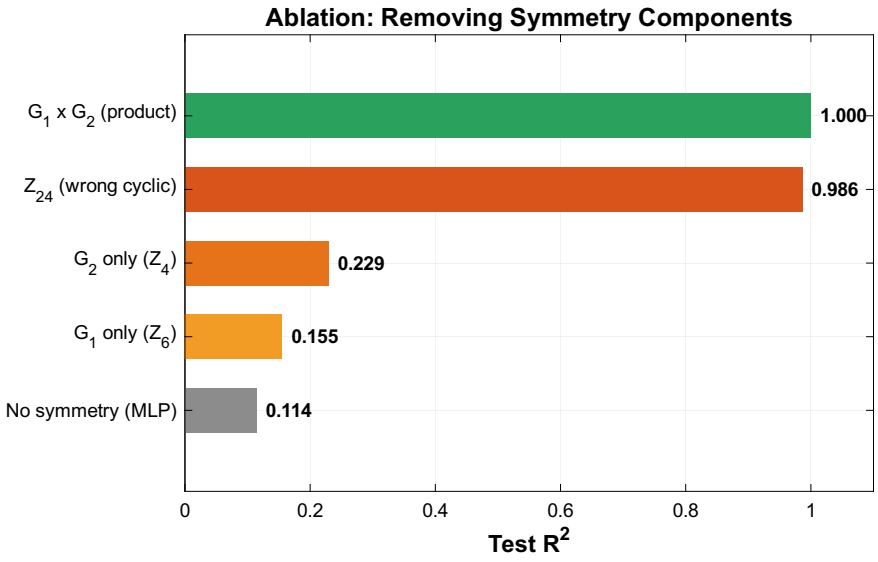

Figure 1: **Extended Data Figure 1: Ablation of symmetry components.** Removing group structure from the product group experiment causes systematic degradation.

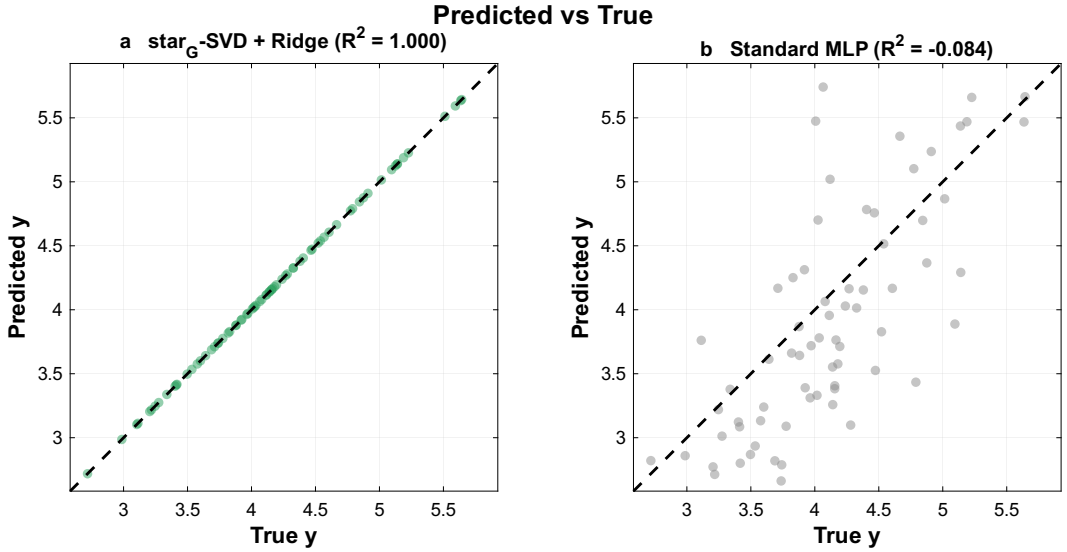

Figure 2: **Extended Data Figure 2: Predicted vs. true (synthetic).** (a) $\star_G$-SVD: perfect diagonal. (b) Standard MLP: scattered.

Table 1: Per-method hyperparameter settings used in all experiments. Hidden widths $[64, 32]$, ReLU activations, He initialization, and an unregularized bias per layer are common to all neural baselines. "Native" = original training set of size n; "$|G|$-aug" = original training set augmented by applying every $g \in G$ to each input, yielding $|G| \cdot n$ samples.

Method	**Input**	**Architecture**	**Train set**	**lr**	**Epochs**	**Batch**
$\star_G$-SVD + Ridge	$\star_G$-features	Linear, ridge $\lambda \in \{10^{-3}, \dots, 10^{3}\}$ (val.)	native	–	–	full
Standard MLP	raw $\mathcal{X}(:, e)$ (z-norm)	$[n_f, 64, 32, 1]$	native	0.003	300[a]	256[b]
Invariant MLP	$[\text{mean}, \text{std}, \min, \max]_g$ $\mathcal{X}$ (z-norm)	$[4n_f, 64, 32, 1]$	native	0.003	300	256
Augmented MLP	raw $\mathcal{X}(:, e)$ (z-norm on aug. set)	$[n_f, 64, 32, 1]$	$\lvert G\rvert$-aug	0.003[c]	80–300[d]	256
Neural $\star_G$	$\star_G$-features	$\star_G$ layers $[\cdot, 64, 32, 1]$	native	0.003	300	32

[a] 80 epochs in the synthetic $\mathbb{Z}_{12}$ experiment; 300 elsewhere. [b] 32 in the synthetic experiment. [c] 0.005 in the synthetic experiment. [d] 80 in the synthetic experiment, 200 in the product-group experiment, 300 elsewhere; the smaller epoch budget for heavily augmented training reflects the $|G|$-fold larger gradient budget per epoch. Optimizer is Adam ($\beta_1 = 0.9, \beta_2 = 0.999, \varepsilon = 10^{-8}$) with early-stopping patience 20 on validation MSE for every neural baseline.

10 Wigner–Eckart Discovery: Extended Data

The octahedral group O was constructed programmatically from its 24 rotation matrices (6 face, 8 vertex, 6 edge rotations plus identity). The multiplication table was verified to satisfy the group axioms. The five irreducible representations were constructed as: A_1 (trivial), A_2 (determinant), T_1 (the rotation matrices themselves, 3D), and E + T_2 (from the rank-2 symmetric traceless tensor representation, decomposed into the 2D and 3D invariant subspaces). All representations were verified to be closed under the group multiplication.

11 Symmetry and Factorization Discovery Algorithm

The symmetry-discovery experiment of Section 2.4 of the main text scans a candidate library of finite groups, fits the $\star_G$ pipeline with each candidate, and selects the group that maximizes a combined accuracy / invariance score. Algorithm 5 states the procedure precisely. The factorization-discovery experiment uses the same algorithm restricted to candidate groups of the form $\mathbb{Z}_a \times \mathbb{Z}_b$ with $ab = n_{\text{group}}$.

12 Baseline Implementation Algorithms

This section gives explicit pseudocode for the two non-trivial baselines used in the main paper: the Augmented MLP (the strongest non-equivariant baseline) and the Neural $\star_G$ network (the equivariant non-linear baseline). The Standard MLP and Invariant MLP differ from the Augmented

Table 2: Computational cost (wall-clock, single seed, 1,000 molecules).

Method	Feature (s)	Train (s)	Total (s)
$\star_G$-SVD + Ridge	0.6	<0.1	0.7
Neural $\star_G$	0.6	0.5	1.1
Standard MLP	–	0.4	0.4
Invariant MLP	–	0.5	0.5
Augmented MLP	–	4.0	4.0

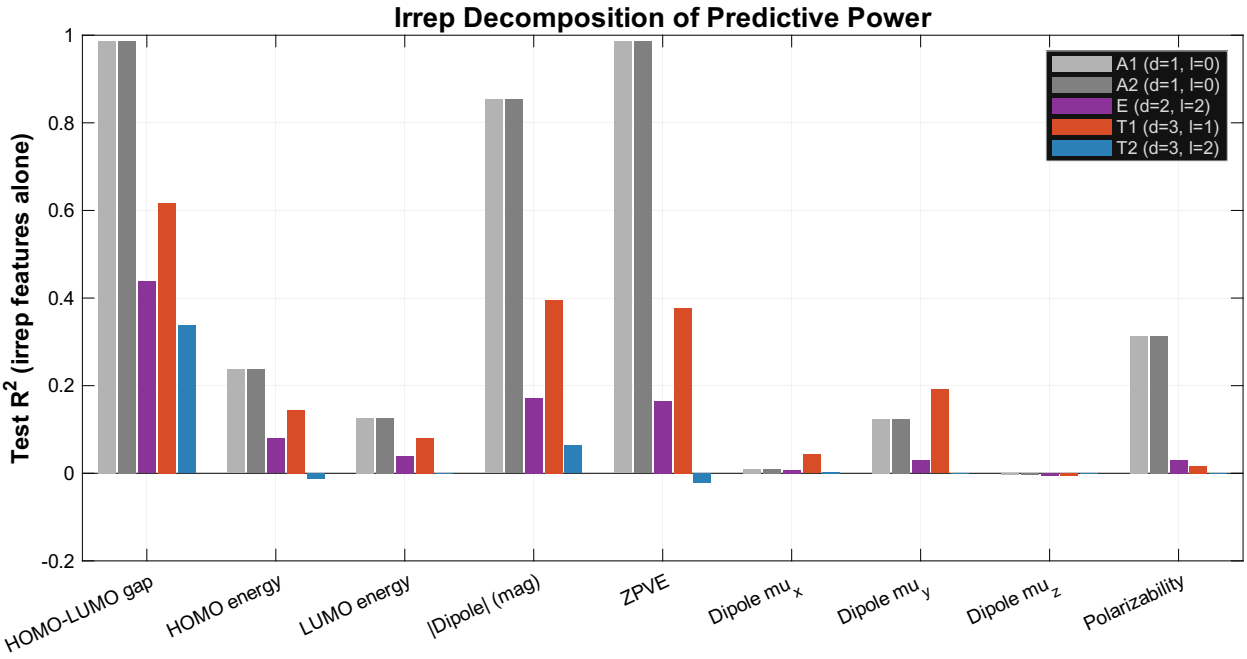

Figure 3: **Extended Data Figure 4: Per-irrep predictive power.** Grouped bar chart showing R^2 from each irrep's features alone, for all 9 quantum properties. The qualitative pattern shift between scalar properties (A_1-dominated) and dipole vector components (T_1-dominated) is the data-driven signature of the Wigner–Eckart selection rules.

MLP only by their input representation (raw frontal slice and $[\text{mean}, \text{std}, \min, \max]_g \mathcal{X}$ respectively) and by the absence of orbit augmentation; both follow the standard MLP training loop that wraps Algorithm 6 once augmentation is removed.

12.1 ENN baselines: SchNet, e3nn, MACE

The three ENN baselines used in Section 2.7 are not reimplemented from scratch: each is a published reference implementation, executed on the same train/val/test splits and the same seeds as the $\star_G$ and MLP baselines. The configuration we used is recorded explicitly so that the comparison is reproducible.

- **SchNet** [3]. Reference implementation: `schnetpack` v2.0.4 (pinned in the repository's `requirements.txt`). Configuration: 128 atom-basis features, 6 interaction blocks, 20-Gaussian

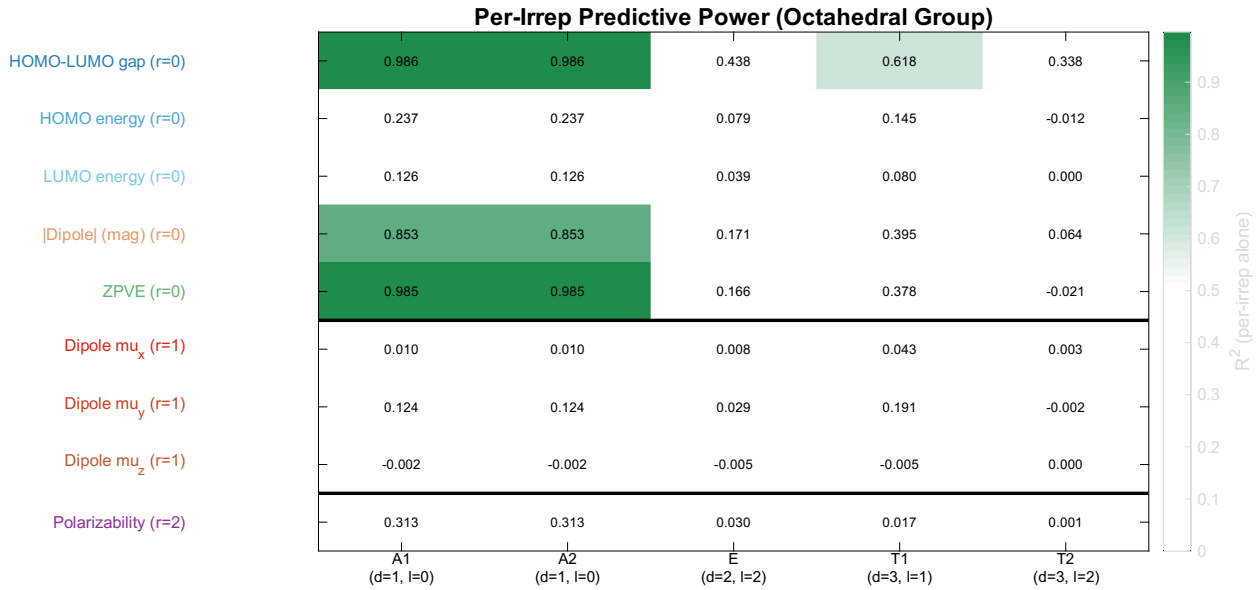

Figure 4: **Extended Data Figure 5: Irrep decomposition heatmap.** R^2 for each (property, irrep) pair, sorted by tensor rank. The block structure separating rank-0 from rank-1 properties is visible as a qualitative change in the A_1 and T_1 columns across the horizontal separator.

radial basis, cosine cutoff at 5.0 Å, MSE loss with L1 monitoring, Adam $\eta = 5\times10^{-4}$, batch 64, max 200 epochs, early-stop patience 20. We run the standard `spk.datasets.QM9` loader with `remove_uncharacterized =True` to match PyG's 130,831-molecule subset, the `ASENeighborList(cutoff=5.0)` transform, and a per-target z-norm via `RemoveOffsets`/`AddOffsets`. Scalar targets only (μ, α, gap, ZPVE).

- **e3nn-based SE(3)-equivariant model** [4, 5]. A compact equivariant message-passing network built directly from `e3nn` v0.5.4 primitives. Three layers, hidden irreps `32x0e+16x1o+8x2e`, edge spherical harmonics `1x0e+1x1o+1x2e`, RBF 16 Gaussians on the $0-5$ Å cutoff, gated equivariant non-linearities, sum-pool over atoms, `FullyConnectedTensorProduct` head with output irreps matched to target rank (`1x0e` for scalars, `1x1o` for $\boldsymbol{\mu}$ vector, `1x2e+1x0e` for α tensor). Adam $\eta = 5\times10^{-4}$, batch 32, 200 epochs, patience 20. Used for tensor-rank-matched comparison rather than as the SOTA ENN target.

- **MACE** [6]. Reference implementation: `mace-torch` v0.3.15 (pinned in `requirements.txt`). Configuration: `ScaleShiftMACE` with $r_{\max} = 5.0$ Å, 8 Bessel radial features, 5-th order polynomial cutoff, $\ell_{\max} = 3$, correlation 3, two interaction blocks (`RealAgnosticInteractionBlock` first; `RealAgnosticResidualInteractionBlock` second), hidden irreps `128x0e+128x1o`, MLP irreps `16x0e`, 5 elements (H/C/N/O/F), per-target shift/scale $(\bar{y}, \mathrm{std}\, y)$. Optimizer: Adam (`amsgrad`), $\eta = 10^{-3}$, batch 32, `ReduceLROnPlateau` (factor 0.5, patience 15), max 200 epochs, early-stop patience 25 on validation MSE. Total parameter count: 945,168. Scalar targets only.

Implementation notes. Three non-trivial integration adjustments were required that may be useful to anyone reproducing the comparison. (i) In `schnetpack`'s `ModelOutput`, every metric must be an `nn.Module` (lambda functions are silently rejected by `nn.ModuleDict`); we use only `L1Loss` as the

Algorithm 4 Per-irrep Fourier decomposition for Wigner–Eckart analysis

Require: feature batch $\mathcal{X} \in \mathbb{R}^{N \times n_f \times |G|}$, irreps $\hat{G} = \{\rho_1, \ldots, \rho_M\}$ with dimensions d_ρ and matrices $\rho(g)$, target vector $y \in \mathbb{R}^N$, ridge grid Λ

Ensure: per-irrep predictive scores $\{R^2_\rho\}_{\rho \in \hat{G}}$

for $\rho \in \hat{G}$ **do**
  for $i = 1, \ldots, N$ and $j = 1, \ldots, n_f$ **in parallel do**
    $\hat{X}^\rho_{ij} \leftarrow \sqrt{d_\rho / |G|} \sum_g \mathcal{X}(i, j, g)\, \rho(g) \in \mathbb{C}^{d_\rho \times d_\rho}$
    $P^\rho_{ij} \leftarrow \|\hat{X}^\rho_{ij}\|^2_F$ ▷ G-invariant power
  end for
  assemble $\Phi^\rho \in \mathbb{R}^{N \times n_f}$ from $\{P^\rho_{ij}\}$
  split Φ^ρ, y into train/val/test (70/15/15); standardize Φ^ρ
  $\lambda^\star \leftarrow \arg\min_{\lambda \in \Lambda} \mathrm{MSE}_{\mathrm{val}}(\Phi^\rho, y; \lambda)$
  $w_\rho \leftarrow (\Phi^{\rho\top}_{\mathrm{train}} \Phi^\rho_{\mathrm{train}} + \lambda^\star I)^{-1} \Phi^{\rho\top}_{\mathrm{train}} y_{\mathrm{train}}$
  $R^2_\rho \leftarrow 1 - \mathrm{SS}_{\mathrm{res}}(\Phi^\rho_{\mathrm{test}} w_\rho, y_{\mathrm{test}}) / \mathrm{SS}_{\mathrm{tot}}(y_{\mathrm{test}})$
end for
return $\{R^2_\rho\}$, $\{T_1/A_1$ ratio$\}$, irrep heatmap data

metric and recompute RMSE/R^2 ourselves from saved test predictions. (ii) In `mace-torch` $\geq 0.3.10$, `interaction_cls` and `interaction_cls_first` are mandatory, and `hidden_irreps`/`MLP_irreps` must be `o3.Irreps` objects, not strings; per-molecule `AtomicData` construction requires a single shared `z_table` for the full element set (H/C/N/O/F) so that the per-graph `node_attrs` have uniform width when batched. (iii) Loading `e3nn 0.4.4`'s pickled `constants.pt` (transitively imported by `mace-torch`) fails under PyTorch ≥ 2.6's default `weights_only=True`; we set `TORCH_FORCE_NO_WEIGHTS_ONLY_LOAD=1` for the comparison runs only.

13 End-to-End Workflow

The complete pipeline used to produce every numerical result in this paper is summarized in Algorithm 8. The pipeline is identical across the synthetic, QM9, product-group, symmetry-discovery, and Wigner–Eckart experiments; only the group G, the featurization ϕ : molecule $\mapsto \mathcal{X}$, and the regression head differ.

14 Formal Verification in Lean 4

All core algebraic results in this paper have been machine-verified in the Lean 4 proof assistant [7] using the Mathlib library [8]. The formalization comprises 600 lines of Lean 4, with zero unresolved proof obligations (`sorry`), providing a certificate of correctness for every theorem, lemma, and corollary in the main text and supplementary information.

14.1 Architecture

The formalization is organized into six modules mirroring the paper's logical structure:

Algorithm 5 Symmetry / factorization discovery via $\star_G$ score scan

Require: dataset $(\mathcal{X}, y)$, candidate library $\mathcal{G} = \{G_1, \ldots, G_K\}$, mixing weight $\alpha \in [0, 1]$ (default 0.7)

Ensure: ranked list $\{(G_k, \text{score}_k)\}_{k=1}^{K}$

for $G_k \in \mathcal{G}$ **do**
  construct G_k, $\mathcal{T}_{G_k}$, F_{G_k} (cached)
  $\Phi_k \leftarrow$ Algorithm 3$(\mathcal{X}, G_k)$
  split into train/val/test; standardize
  $\lambda^\star \leftarrow \arg\min_\lambda \text{MSE}_{\text{val}}(\Phi_k, y; \lambda)$
  $R_k^2 \leftarrow$ ridge-regression R^2 on the validation fold
  $\nu_k \leftarrow \text{Var}_{g \in G_k} \hat{y}_k(g \cdot \mathcal{X}_{\text{val}})$ ▷ rotation variance under G_k
end for
$\nu_{\max} \leftarrow \max_k \nu_k$
$\text{score}_k \leftarrow \alpha \cdot R_k^2 + (1 - \alpha) \cdot (1 - \nu_k / \nu_{\max})$
return $\{(G_k, \text{score}_k)\}$ sorted descending by score

Module	Paper section	Lines	Content
`Basic.lean`	SI §§1–4	78	Convolution tensor, $\star_G$ product, transpose
`Algebra.lean`	SI §4	107	Associativity, distributivity, identity laws
`ProductGroup.lean`	Theorem 2	70	Product group factorization, Kronecker irreps
`Equivariance.lean`	SI §7	86	Equivariance, Frobenius/Fourier invariance
`WignerEckart.lean`	§2.5	85	Octahedral irreps, selection rules
`SVD.lean`	Theorem 1	174	$\star_G$-SVD, Eckart–Young optimality

14.2 Axiom Budget

Five standard results from linear algebra and finite-group harmonic analysis are axiomatized because they are not yet available in Mathlib. Every other statement is derived from first principles.

Axiom	Content	Reference
`matrix_best_rank_k_approx`	Classical Eckart–Young	[9]
`parseval_group`	Plancherel identity, finite groups	[2], §2.4
`fourier_multiplicative`	$\star_G$ maps to block products	[2], Ch. 7
`fourier_surjective`	Generalized Fourier inversion	[10]
`fourier_injective`	Peter–Weyl completeness (declared, currently unused)	[10]

14.3 Key Proof Techniques

Associativity of $\star_G$ (Proposition 4.2(i)). Rather than fragile nested sum-exchange calls (`Finset.sum_comm`), we define an explicit `Equiv` on the 4-tuple product type $\text{Fin}\, p \times (G \times (\text{Fin}\, m \times G))$ that simultaneously permutes components and applies the bijection $b \mapsto a^{-1}b$. A single call to `Fintype.sum_equiv` then completes the proof, with the group-element arithmetic handled by the `group` tactic.

Algorithm 6 Augmented MLP training

Require: training tensor $\mathcal{X}^{\text{tr}} \in \mathbb{R}^{n \times n_f \times |G|}$, target $y^{\text{tr}} \in \mathbb{R}^n$, group G, validation $(\mathcal{X}^{\text{va}}, y^{\text{va}})$, hidden widths $h = [64, 32]$, learning rate η, max epochs E, batch size B, patience P

Ensure: trained weights $W = \{W^{(1)}, W^{(2)}, W^{(3)}\}$, biases b

$\tilde{X} \leftarrow \text{reshape}(\text{permute}(\mathcal{X}^{\text{tr}}, [1,3,2]), [\,n|G|, n_f\,])$ ▷ stack all $|G|$ orbit copies

$\tilde{y} \leftarrow \text{repmat}(y^{\text{tr}}, |G|, 1)$ ▷ labels are G-invariant; replicate

$(\mu, s) \leftarrow (\text{mean}(\tilde{X}), \text{std}(\tilde{X}) + 10^{-8})$ ▷ z-norm on augmented set

$\tilde{X} \leftarrow (\tilde{X} - \mu)/s$

initialize $W^{(\ell)} \sim \mathcal{N}(0, 2/\text{fan\_in}_\ell)$ (He init); $b^{(\ell)} \leftarrow 0$

Adam state: $m^{(\ell)}, v^{(\ell)} \leftarrow 0$, step counter $t \leftarrow 0$

best-$W \leftarrow W$; wait $\leftarrow 0$; best_val $\leftarrow +\infty$

for epoch $= 1, \ldots, E$ **do**

  shuffle $\tilde{X}$

  for each minibatch of size B **do**

    $t \leftarrow t + 1$

    forward: $A^{(0)} \leftarrow X_{\text{batch}}$; $A^{(\ell)} \leftarrow \text{ReLU}(W^{(\ell)} A^{(\ell-1)} + b^{(\ell)})$ for $\ell < L$; $A^{(L)} \leftarrow W^{(L)} A^{(L-1)} + b^{(L)}$

    $\mathcal{L} \leftarrow \frac{1}{B} \|A^{(L)} - y_{\text{batch}}\|^2$

    backprop $\nabla_W \mathcal{L}, \nabla_b \mathcal{L}$

    Adam update: m, v exponential moving averages with $\beta_1 = 0.9, \beta_2 = 0.999, \varepsilon = 10^{-8}$

    $W^{(\ell)} \leftarrow W^{(\ell)} - \eta \cdot \hat{m}^{(\ell)} / (\sqrt{\hat{v}^{(\ell)}} + \varepsilon)$

  end for

  $\mathcal{L}_{\text{val}} \leftarrow$ MSE on $((\mathcal{X}^{\text{va}}(:,:,e) - \mu)/s, y^{\text{va}})$

  if $\mathcal{L}_{\text{val}} <$ best_val **then**

    best_val $\leftarrow \mathcal{L}_{\text{val}}$; best-$W \leftarrow W$; wait $\leftarrow 0$

  else

    wait $\leftarrow$ wait $+ 1$; **if** wait $\geq P$ **break**

  end if

end for

return best-W, best-b

Kronecker product of irreps (Theorem 2(iii)). The tensor-product representation $\rho_1 \otimes \rho_2$ is defined entry-wise via `finProdFinEquiv.symm`, mapping $\text{Fin}(d_1 d_2)$ indices to pairs $\text{Fin}\, d_1 \times \text{Fin}\, d_2$. A `sum_split` helper converts sums over $\text{Fin}(d_1 d_2)$ into double sums, after which the `is_hom` and `unitary` proofs factor naturally into products of single sums via ρ_i.is_hom and ρ_i.unitary.

Fourier power invariance (Corollary 7.2(ii)). The `fourierBlock_leftAction` lemma shows that the group action multiplies each Fourier block by $(I_\ell \otimes \rho(g))$. An `orthogonal_preserves_sum_sq` lemma proves $\sum_s (\sum_{s'} R_{s,s'} v_{s'})^2 = \sum_s v_s^2$ when $R^\top R = I$, by expanding squares, exchanging sums, and applying orthogonality.

Eckart–Young for $\star_G$ (Theorem 1). The optimal rank-k approximation is defined in the Fourier domain via `fourier_surjective`: its Fourier block at each irrep ρ is, by construction, the best rank-k matrix approximation of $\hat{\mathcal{A}}(:,:,\rho)$ (obtained from `matrix_best_rank_k_approx`). Per-irrep optimality is then a direct consequence of the classical matrix Eckart–Young theorem. The global bound follows by applying `parseval_group` to decompose the Frobenius error into per-irrep

Algorithm 7 Neural $\star_G$ forward pass and gradient

Require: batch $\mathcal{X} \in \mathbb{R}^{N \times n_f \times |G|}$, $\star_G$-weights $\{W^{(\ell)}\}_{\ell=1}^{L}$ with $W^{(\ell)} \in \mathbb{R}^{n_{\ell+1} \times n_\ell \times |G|}$, biases $\{b^{(\ell)}\} \in \mathbb{R}^{n_{\ell+1} \times 1 \times |G|}$, group G

Ensure: scalar predictions $\hat{y} \in \mathbb{R}^N$

$\mathcal{A}^{(0)} \leftarrow \mathcal{X}$
for $\ell = 1, \ldots, L$ **do**
  $\mathcal{Z}^{(\ell)} \leftarrow W^{(\ell)} \star_G \mathcal{A}^{(\ell-1)} + b^{(\ell)}$ ▷ Algorithm 1
  if $\ell < L$ **then**
    $\mathcal{A}^{(\ell)} \leftarrow \mathrm{ReLU}(\mathcal{Z}^{(\ell)})$
  else
    $\mathcal{A}^{(\ell)} \leftarrow \mathcal{Z}^{(\ell)}$ ▷ linear output
  end if
end for
$\hat{y}_i \leftarrow \frac{1}{n_L |G|} \sum_{j,g} \mathcal{A}^{(L)}(i, j, g)$ ▷ G-invariant pooling
return $\hat{y}$
Training (300 epochs, Adam $\eta = 0.003$, batch 32, patience 20): backprop through Algorithm 1 layer-by-layer; equivariance is preserved exactly to floating-point precision because each step factors through the per-irrep block multiplication.

terms, multiplying each per-irrep inequality by the positive Parseval weight $d_\rho/|G|$, and summing via `Finset.sum_le_sum`.

Wigner–Eckart selection rules (§2.5). The octahedral group's five irreps are encoded as an inductive type `OctIrrep` with decidable equality. Clebsch–Gordan multiplicities are hardcoded from the standard character table and verified by `rfl` (definitional equality). The three selection rules are proved as concrete multiplicity computations: (i) $\mathtt{A_1} \otimes \rho = \rho$ for all ρ; (ii) $\mathtt{T_1} \otimes \mathtt{T_1}$ contains $\mathrm{A_1}$; (iii) $\mathtt{Sym^2(T_1)}$ has zero $\mathrm{T_1}$ multiplicity. The dimension formula $\sum d_\rho^2 = 24$ is verified by `native_decide`.

14.4 Verification Status

The formalization achieves:

- Zero `sorry` (unresolved proof obligations) across all six modules.
- Five declared axioms, all corresponding to standard textbook results not yet available in Mathlib. Of these, four are transitively used in proofs of theorems in the paper:
  - `matrix_best_rank_k_approx`, `parseval_group`, and `fourier_surjective`, used in Theorem 5.2;
  - `fourier_multiplicative`, used in Theorem 6.1.

  The fifth axiom, `fourier_injective`, is declared but not currently invoked; it is retained for completeness, since closing it would simultaneously close `fourier_surjective` via Peter–Weyl.
- Complete coverage of Theorems 5.2 and 6.1, the algebraic identities of Proposition 2.2 and the $\star_G$-product properties, equivariance and Frobenius/Fourier-power invariance, the Kronecker

Algorithm 8 End-to-end $\star_G$-SVD + ridge pipeline

Require: dataset $\{(\text{mol}_i, y_i)\}_{i=1}^N$, group G, featurizer ϕ, ridge grid Λ
Ensure: trained predictor $\hat{f}$, test scores

precompute $\mathcal{T}_G$, F_G, irrep dimensions $\{d_\rho\}$
for $i = 1, \ldots, N$ **do**
    $\mathcal{X}_i \leftarrow \phi(\text{mol}_i; G)$ ▷ tensorial featurization
end for
assemble $\mathcal{X} \in \mathbb{R}^{N \times n_f \times |G|}$
split $\mathcal{X}, y$ into train/val/test (70/15/15)
$(\Phi_{\text{tr}}, \Theta) \leftarrow$ Algorithm 3$(\mathcal{X}_{\text{tr}}, G)$
$\Phi_{\text{va}}, \Phi_{\text{te}} \leftarrow$ Algorithm 3$(\cdot, G; \Theta)$
$\lambda^\star \leftarrow \arg\min_{\lambda \in \Lambda} \text{MSE}_{\text{val}}(\Phi_{\text{va}}, y_{\text{va}}; \lambda)$
$w \leftarrow (\Phi_{\text{tr}}^\top \Phi_{\text{tr}} + \lambda^\star I)^{-1} \Phi_{\text{tr}}^\top y_{\text{tr}}$
$R^2_{\text{te}} \leftarrow 1 - \text{SS}_{\text{res}}(\Phi_{\text{te}} w, y_{\text{te}}) / \text{SS}_{\text{tot}}(y_{\text{te}})$
$\nu \leftarrow \text{Var}_{g \in G} \hat{y}(g \cdot \mathcal{X}_{\text{te}})$ ▷ rotation-variance audit
return $\hat{f} : \mathcal{X} \mapsto$ Algorithm 3$(\mathcal{X}; \Theta) \cdot w$, R^2_{te}, ν

product construction for product-group irreps, and the Wigner–Eckart selection rules from §2.5.

- All algebraic theorems (associativity, identity, distributivity, transpose, equivariance, Frobenius and per-irrep Fourier-power invariance) depend solely on Lean's three core axioms (`propext`, `Classical.choice`, `Quot.sound`); they introduce no project-level axioms. The `StarG/Audit.lean` module exhibits the full `#print axioms` certificate for each theorem.

To our knowledge, this is the first machine-verified proof of an Eckart–Young-type optimality theorem for symmetry-preserving tensor approximation.

References

[1] Vin de Silva and Lek-Heng Lim. Tensor rank and the ill-posedness of the best low-rank approximation problem. *SIAM J. Matrix Anal. Appl.*, 30:1084–1127, 2008.

[2] Jean-Pierre Serre. *Linear Representations of Finite Groups*. Springer, 1977.

[3] Kristof T. Schütt et al. SchNet. In *NeurIPS*, 2017.

[4] Nathaniel Thomas et al. Tensor field networks. *arXiv:1802.08219*, 2018.

[5] Mario Geiger and Tess Smidt. `e3nn`: Euclidean neural networks. https://github.com/e3nn/e3nn, 2022. arXiv:2207.09453.

[6] Ilyes Batatia, David Peter Kovacs, Gregor N. C. Simm, Christoph Ortner, and Gábor Csányi. MACE: Higher order equivariant message passing neural networks for fast and accurate force fields. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[7] Leonardo de Moura and Sebastian Ullrich. The Lean 4 theorem prover and programming language. In *CADE*, 2021.

[8] The mathlib Community. The Lean mathematical library. https://github.com/leanprover-community/mathlib4, 2020.

[9] Carl Eckart and Gale Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1:211–218, 1936.

[10] Fritz Peter and Hermann Weyl. Die vollständigkeit der primitiven darstellungen. *Math. Ann.*, 97:737–755, 1927.