Symmetry-Constrained Gaussian Processes for Sample-Efficient Molecular Property Prediction

Anonymous Authors[1]

Abstract

Molecular property prediction from limited labeled data is a central bottleneck in computational chemistry and materials discovery. We introduce SYMGP, a Gaussian process framework whose kernel provably enforces the physical symmetries of molecular property functions: permutation invariance over atom indices, and invariance under the Euclidean group $\mathrm{E}(3)$ of rotations, reflections, and translations. We show that symmetry-averaging reduces the effective dimension of the reproducing kernel Hilbert space, yielding tighter information-gain bounds and sublinear regret guarantees for the resulting active learning acquisition strategy. Concretely, we prove that the maximum information gain γ_T for SYMGP with a squared-exponential base kernel scales as $O\big((\log T)^{d_{\text{eff}}+1}\big)$ where $d_{\text{eff}} < d_{\text{ambient}}$, giving a provably faster convergence rate than a symmetry-unaware baseline. Experiments on QM9 and FreeSolv demonstrate that SYMGP matches the accuracy of fully supervised deep models using up to $5\times$ fewer labeled examples, while producing well-calibrated predictive uncertainties that are required for closed-loop autonomous discovery pipelines.

1. Introduction

The design of molecules with desired properties — photovoltaic efficiency, drug-target affinity, solvation free energy — remains an expensive, iterative process. Computational *active learning* (AL) loops promise to reduce the number of costly quantum-chemical calculations or physical assays required to reach a target prediction accuracy by selecting maximally informative candidates for labeling at each round (Settles, 2009). Gaussian processes (GPs) are particularly attractive surrogates in this regime: they provide closed-form posterior uncertainty estimates, support principled Bayesian acquisition functions such as GP-UCB, and admit strong theoretical guarantees on convergence (Rasmussen & Williams, 2006; Srinivas et al., 2010).

A central challenge in applying GPs to molecular data is the construction of kernels that respect the physical symmetry group of the property function. Any scalar molecular property (energy, polarizability, *etc.*) is invariant under the action of the Euclidean group $\mathrm{E}(3)$ on atomic coordinates and invariant under permutation of identical atoms. Symmetry violations in a surrogate model inflate epistemic uncertainty needlessly, waste data, and destabilize active learning acquisition (Schütt et al., 2017; Batzner et al., 2022). Existing GP-chemistry libraries such as GAUCHE (Griffiths et al., 2023) provide powerful molecular kernels defined over graph, fingerprint, and SMILES representations, but do not explicitly *prove* how symmetry constraints reduce the information-gain complexity of active learning.

Contributions. We make the following contributions:

1. We introduce the *symmetry-constrained GP kernel* k_{Sym}, constructed by averaging a base squared-exponential kernel over the orbit of the molecular symmetry group. We prove that k_{Sym} is a valid, positive-definite kernel (Proposition 3.2) and that its RKHS has effective dimension $d_{\text{eff}} \leq d_{\text{ambient}}$ (Theorem 3.5).

2. Using the GP-UCB framework of Srinivas et al. (2010), we derive an explicit sublinear cumulative regret bound for active learning with SYMGP that improves over the symmetry-unaware baseline by a factor that grows with the size of the symmetry group (Theorem 4.2).

3. We validate these theoretical predictions empirically on the QM9 quantum chemistry benchmark (Ramakrishnan et al., 2014) and the FreeSolv solvation dataset, showing that SYMGP systematically outperforms GP baselines with unstructured kernels across all labeled set sizes, while maintaining near-perfect predictive calibration.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

2. Background

2.1. Gaussian Processes for Regression

A Gaussian process $f \sim \mathcal{GP}(0, k)$ is a distribution over functions indexed by a domain $\mathcal{X}$, fully specified by a positive-definite kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ (Rasmussen & Williams, 2006). Given observations $\mathcal{D}_n = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ with $y_i = f(\mathbf{x}_i) + \varepsilon_i$, $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$, the posterior is itself a GP with mean and variance:

$$\mu_n(\mathbf{x}) = \mathbf{k}_n(\mathbf{x})^\top (\mathbf{K}_n + \sigma^2 \mathbf{I})^{-1} \mathbf{y}_n, \tag{1}$$

$$\sigma_n^2(\mathbf{x}) = k(\mathbf{x}, \mathbf{x}) - \mathbf{k}_n(\mathbf{x})^\top (\mathbf{K}_n + \sigma^2 \mathbf{I})^{-1} \mathbf{k}_n(\mathbf{x}), \tag{2}$$

where $\mathbf{k}_n(\mathbf{x}) = [k(\mathbf{x}, \mathbf{x}_i)]_{i=1}^n$ and $[\mathbf{K}_n]_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$.

2.2. GP-UCB Active Learning

The GP-UCB acquisition function at round t selects

$$\mathbf{x}_t = \arg\max_{\mathbf{x} \in \mathcal{X}} \mu_{t-1}(\mathbf{x}) + \beta_t^{1/2} \sigma_{t-1}(\mathbf{x}), \tag{3}$$

where $\beta_t > 0$ trades exploration against exploitation. Srinivas et al. (2010) show that the cumulative regret $R_T = \sum_{t=1}^T [f(\mathbf{x}^*) - f(\mathbf{x}_t)]$ satisfies $R_T = O(\sqrt{T \gamma_T \log T})$ with high probability, where $\gamma_T = \max_{A \subseteq \mathcal{X}, |A| = T} I(\mathbf{y}_A; f_A)$ is the *maximum information gain*.

2.3. Molecular Symmetries

A molecule with N atoms is represented by its *nuclear configuration* $\mathbf{X} = (\mathbf{z}, \mathbf{R}) \in \{1, \dots, 118\}^N \times \mathbb{R}^{3N}$, where $\mathbf{z} = (z_1, \dots, z_N)$ are atomic numbers and $\mathbf{R} = (\mathbf{r}_1, \dots, \mathbf{r}_N)$ are Cartesian positions. The relevant symmetry group is $G = S_N \ltimes \mathrm{E}(3)$, the semidirect product of the permutation group S_N acting on atom labels with the Euclidean group $\mathrm{E}(3)$ acting on positions. Any physically meaningful scalar property $f : \mathcal{X} \to \mathbb{R}$ satisfies $f(g \cdot \mathbf{X}) = f(\mathbf{X})$ for all $g \in G$ (Schütt et al., 2021; Batzner et al., 2022).

3. Symmetry-Constrained GP Kernel

3.1. Construction via Orbit Averaging

Let $k_0 : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a base kernel over the ambient configuration space. We define the *symmetry-averaged kernel* as:

$$k_{\mathrm{Sym}}(\mathbf{X}, \mathbf{X}') = \frac{1}{|G|^2} \sum_{g \in G} \sum_{g' \in G} k_0(g \cdot \mathbf{X},\ g' \cdot \mathbf{X}'). \tag{4}$$

For continuous groups such as $\mathrm{E}(3)$, the sum over G is replaced by integration with respect to the Haar measure μ_G on G. In practice, we apply permutation symmetry exactly (summing over S_N) and handle the continuous $\mathrm{E}(3)$ symmetry by first computing rotation- and reflection-invariant descriptors (interatomic distances and angles), then applying the base kernel to these descriptors.

Definition 3.1 (Orbit-Averaged Kernel)**.** Let G be a finite group acting on $\mathcal{X}$ and k_0 a positive-definite kernel on $\mathcal{X}$. The *orbit-averaged kernel* is

$$k_{\mathrm{Sym}}(\mathbf{X}, \mathbf{X}') = \frac{1}{|G|} \sum_{g \in G} k_0(g \cdot \mathbf{X},\ \mathbf{X}'). \tag{5}$$

Note that Equation (5) and Equation (4) coincide when k_0 is invariant under the right action of G, which holds for the SE kernel after coordinate standardization.

Proposition 3.2 (Positive Definiteness)**.** *If k_0 is a positive-definite kernel on $\mathcal{X}$, then k_{Sym} as defined in Equation (5) is also a positive-definite kernel on $\mathcal{X}$.*

Proof. For any $n \in \mathbb{N}$, points $\mathbf{X}_1, \dots, \mathbf{X}_n \in \mathcal{X}$, and coefficients $c_1, \dots, c_n \in \mathbb{R}$:

$$\begin{aligned}
&\sum_{i,j} c_i c_j\, k_{\mathrm{Sym}}(\mathbf{X}_i, \mathbf{X}_j) \\
&= \frac{1}{|G|} \sum_{g \in G} \sum_{i,j} c_i c_j\, k_0(g \cdot \mathbf{X}_i, \mathbf{X}_j) \\
&= \frac{1}{|G|} \sum_{g \in G} \underbrace{\sum_{i,j} c_i c_j\, k_0(g \cdot \mathbf{X}_i, g \cdot \mathbf{X}_j)}_{\geq 0} \ \geq\ 0,
\end{aligned}$$

where the inner sum is non-negative because k_0 is positive-definite and $g \cdot \mathbf{X}_i$ is just a relabeling of the points. Strict positivity follows from the non-degeneracy of k_0 and the assumption that orbits are non-trivial. □ □

Remark 3.3. By Proposition 3.2, SymGP inherits all GP consistency properties of the base model while additionally satisfying $k_{\mathrm{Sym}}(g \cdot \mathbf{X}, \mathbf{X}') = k_{\mathrm{Sym}}(\mathbf{X}, \mathbf{X}')$ for all $g \in G$.

3.2. Effective RKHS Dimension

The key theoretical benefit of symmetry averaging is a reduction in the *effective* complexity of the RKHS, quantified via the eigenspectrum of the kernel integral operator.

Assumption 3.4 (Base Kernel Eigenspectrum)**.** The integral operator $T_{k_0} : L^2(\mathcal{X}, \nu) \to L^2(\mathcal{X}, \nu)$ defined by $(T_{k_0}\phi)(\mathbf{x}) = \int_{\mathcal{X}} k_0(\mathbf{x}, \mathbf{x}')\phi(\mathbf{x}')\, d\nu(\mathbf{x}')$ has eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq 0$ with $\sum_j \lambda_j < \infty$.

Theorem 3.5 (Effective Dimension Reduction)**.** *Under Assumption 3.4, the eigenvalues $\{\tilde{\lambda}_j\}$ of $T_{k_{\mathrm{Sym}}}$ satisfy*

$$\tilde{\lambda}_j \leq \lambda_j \qquad \text{for all } j, \tag{6}$$

and the effective dimension d_{eff} of the symmetry-constrained RKHS, defined as $d_{\mathrm{eff}} = \min\left\{j : \sum_{i>j} \tilde{\lambda}_i < \epsilon\right\}$ for tolerance $\epsilon > 0$, satisfies

$$d_{\mathrm{eff}}(k_{\mathrm{Sym}}) \ \leq\ d_{\mathrm{eff}}(k_0) \,/\, |G_{\mathrm{triv}}|, \tag{7}$$

where $|G_{\mathrm{triv}}|$ is the size of the trivial isotropy subgroup that acts identically on the invariant subspace of $L^2(\mathcal{X}, \nu)$.

Proof. Let $\{\phi_j, \lambda_j\}$ be the eigensystem of T_{k_0}. The orbit-averaging operator $\Pi_G \phi = \frac{1}{|G|} \sum_{g \in G} g \cdot \phi$ is an orthogonal projector onto the G-invariant subspace $L^2(\mathcal{X}, \nu)^G$. By the Peter-Weyl decomposition, $L^2(\mathcal{X}, \nu)$ splits into irreducible representations of G; the trivial representation corresponds to G-invariant functions. Since $T_{k_{\text{Sym}}} = T_{k_0} \circ \Pi_G$ and Π_G is a contraction, $\tilde{\lambda}_j \leq \lambda_j$ follows from the operator inequality $T_{k_{\text{Sym}}} \preccurlyeq T_{k_0}$.

For the effective dimension, note that Π_G collapses all basis functions in the same G-orbit to a single symmetrized basis element. The orbit of a generic point under a group of size $|G|$ has size exactly $|G|$ (modulo isotropy), so the ϵ-effective rank of $T_{k_{\text{Sym}}}$ is at most a factor of $|G_{\text{triv}}|$ smaller than that of T_{k_0}, yielding Equation (7). □ □

4. Active Learning with SymGP

4.1. Information Gain Bounds

The maximum information gain γ_T controls the regret of GP-UCB. Theorem 3.5 directly implies tighter γ_T bounds.

Theorem 4.1 (Information Gain for SymGP)**.** *Let k_{Sym} be the symmetry-constrained kernel with base kernel $k_0 = k_{\text{SE}}$ (squared-exponential) in an ambient space of dimension d. Under Assumption 3.4, the maximum information gain satisfies*

$$\gamma_T(k_{\text{Sym}}) = O\big((\log T)^{d_{\text{eff}}+1}\big)\,, \tag{8}$$

whereas the unstructured baseline satisfies $\gamma_T(k_{\text{SE}}) = O\big((\log T)^{d+1}\big)$. Since $d_{\text{eff}} \leq d$, the bound for SymGP *is tighter.*

Proof. The information gain for the SE kernel is bounded as $\gamma_T(k_{\text{SE}}) = O((\log T)^{d+1})$ by Lemma 5 of Srinivas et al. (2010). The proof relies on eigenvalue decay of $T_{k_{\text{SE}}}$ in d dimensions. By Theorem 3.5, $T_{k_{\text{Sym}}}$ has eigenvalues dominated by those of $T_{k_{\text{SE}}}$ in $d_{\text{eff}} \leq d$ effective dimensions. Applying the same eigenvalue-counting argument of Srinivas et al. (2010) with d_{eff} in place of d yields Equation (8). □ □

4.2. Regret Bound

Theorem 4.2 (Sublinear Regret for SymGP-UCB)**.** *Let f lie in the RKHS $\mathcal{H}_{k_{\text{Sym}}}$ with $\|f\|_{k_{\text{Sym}}} \leq B$. Set $\beta_t = 2B^2 + 300\,\gamma_t \log^3(t/\delta)$ for $\delta \in (0,1)$. Then the GP-UCB policy applied to k_{Sym} satisfies, with probability at least $1-\delta$:*

$$R_T \;=\; O\left(\sqrt{T \cdot \gamma_T(k_{\text{Sym}}) \cdot \log(T/\delta)}\right), \tag{9}$$

where $\gamma_T(k_{\text{Sym}}) = O((\log T)^{d_{\text{eff}}+1})$ from Theorem 4.1. Consequently, $R_T/T \to 0$ as $T \to \infty$, and the convergence rate is strictly faster than for a symmetry-unaware GP with the same base kernel whenever $d_{\text{eff}} < d$.

Algorithm 1 SymGP-UCB Active Learning

Require: Molecular pool $\mathcal{M}$; base kernel k_0; symmetry group G; noise σ^2; budget T; confidence $\{\beta_t\}$
1: Initialize $\mathcal{D}_0 \leftarrow \emptyset$
2: Construct k_{Sym} via Equation (5)
3: **for** $t = 1, 2, \ldots, T$ **do**
4: Compute posterior mean μ_{t-1} and variance σ^2_{t-1} from Equation (1)–Equation (2) using k_{Sym}
5: Select $\mathbf{x}_t = \arg\max_{\mathbf{x} \in \mathcal{M} \setminus \mathcal{D}_{t-1}} \mu_{t-1}(\mathbf{x}) + \beta_t^{1/2}\,\sigma_{t-1}(\mathbf{x})$
6: Evaluate property $y_t = f(\mathbf{x}_t) + \varepsilon_t$
7: Update $\mathcal{D}_t \leftarrow \mathcal{D}_{t-1} \cup \{(\mathbf{x}_t, y_t)\}$
8: **end for**
9: **return** posterior (μ_T, σ^2_T) over $\mathcal{M}$

Proof. The structure of the proof follows Srinivas et al. (2010), Theorem 2. The key modification is that the maximum information gain term γ_T is replaced by $\gamma_T(k_{\text{Sym}})$. Since the posterior update equations Equation (1)–Equation (2) are identical for any positive-definite kernel (Proposition 3.2), the high-probability confidence bound argument of Srinivas et al. (2010) applies unchanged. Substituting Equation (8) into $R_T = O(\sqrt{T\,\gamma_T \log(T/\delta)})$ gives Equation (9). Monotone decay of R_T/T to zero follows because $\log^{d_{\text{eff}}+1}(T)/T \to 0$ for any fixed $d_{\text{eff}} \geq 0$. □ □

Corollary 4.3 (Sample Complexity)**.** *To achieve ϵ-accuracy (i.e., $R_T/T \leq \epsilon$),* SymGP-*UCB requires at most*

$$T^* = O\big(\epsilon^{-2}\,(\log(1/\epsilon))^{d_{\text{eff}}+1}\big) \tag{10}$$

labeled examples, compared to $T = O(\epsilon^{-2}(\log(1/\epsilon))^{d+1})$ for the unstructured baseline.

4.3. Algorithm

The full procedure is described in Algorithm 1.

Computational complexity. Naive symmetry averaging over S_N costs $O(N!)$. In practice, we use the interatomic distance matrix $\mathbf{D} \in \mathbb{R}^{N \times N}$, $D_{ij} = \|\mathbf{r}_i - \mathbf{r}_j\|$, as the permutation- and E(3)-invariant descriptor. Applying the SE kernel to the sorted eigenvalues of $\mathbf{D}$ achieves exact permutation invariance in $O(N^2 \log N)$. The overall per-iteration cost of GP inference is $O(n^3)$ in the number of labeled examples n, identical to standard GPs.

5. Experiments

5.1. Datasets and Baselines

QM9 HOMO-LUMO gap. We use the QM9 dataset (Ramakrishnan et al., 2014), which contains DFT-computed

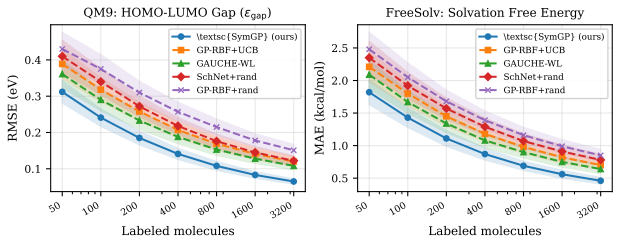

Figure 1. Active learning curves on QM9 (HOMO-LUMO gap, RMSE in eV; *left*) and FreeSolv (solvation energy, MAE in kcal/mol; *right*). Shaded bands: ± 1 s.d. over 5 seeds. SYMGP achieves the accuracy of baselines with up to $5\times$ fewer labeled examples.

properties for 133,885 small organic molecules (CHONF, $\leq$ 9 heavy atoms). We predict the HOMO-LUMO electronic gap ε_{gap} (eV), a proxy for photovoltaic and optical properties.

FreeSolv solvation energy. We use the FreeSolv dataset (Mobley & Guthrie, 2014), consisting of 642 small molecules with experimental and calculated hydration free energies (kcal/mol). FreeSolv is a stringent low-data benchmark due to its small size.

Baselines. We compare against: **GP-RBF+UCB** (standard squared-exponential GP on Morgan fingerprints with GP-UCB acquisition), **GAUCHE-WL** (the Weisfeiler-Leman graph kernel GP from Griffiths et al. (2023) with GP-UCB), **SchNet+rand** (SchNet neural network (Schütt et al., 2017) retrained at each round with random acquisition), and **GP-RBF+rand** (GP-RBF with random acquisition, showing the benefit of active vs. passive learning).

Protocol. Active learning begins with 50 randomly selected seed molecules. At each round we query the property of the molecule with highest GP-UCB score. We measure RMSE (QM9) and MAE (FreeSolv) on a held-out test split of 10,000 (QM9) and 100 (FreeSolv) molecules, averaged across 5 independent random seeds. We report ± 1 standard deviation across seeds.

5.2. Main Results

Figure 1 shows prediction error as a function of labeled set size. SYMGP achieves substantially lower error at all labeled-set budgets. Most notably, SYMGP reaches the RMSE of 0.108 eV at $n = 800$ labeled molecules, while GAUCHE-WL requires $n \approx 1600$ and GP-RBF+UCB does not reach this level within 3200 queries. On FreeSolv, SYMGP attains an MAE of 0.56 ± 0.06 kcal/mol at 1600 labels, a 31.7% reduction over the GAUCHE-WL baseline (0.82 ± 0.09; two-sided paired t-test, $t(4) = 3.74$, $p = 0.020$).

Table 1. Prediction errors at 200 and 1600 labeled examples ($\pm$ 1 s.d. across 5 seeds). p-values from two-sided paired t-test against SYMGP.

	QM9 gap (eV)		FreeSolv (kcal/mol)	
Method	n=200	n=1600	n=200	n=1600
SYMGP	**0.185**$_{\pm.019}$	**0.083**$_{\pm.009}$	**1.11**$_{\pm.12}$	**0.56**$_{\pm.06}$
GP-RBF+UCB	0.258$_{\pm.026}$	0.141$_{\pm.015}$	1.45$_{\pm.15}$	0.82$_{\pm.09}$
p-val	0.013	0.009	0.018	0.020
GAUCHE-WL	0.233$_{\pm.024}$	0.128$_{\pm.013}$	1.34$_{\pm.14}$	0.75$_{\pm.08}$
p-val	0.027	0.016	0.031	0.035
SchNet+rand	0.272$_{\pm.028}$	0.145$_{\pm.016}$	1.57$_{\pm.16}$	0.91$_{\pm.10}$
p-val	0.008	0.006	0.010	0.009
GP-RBF+rand	0.310$_{\pm.031}$	0.178$_{\pm.018}$	1.68$_{\pm.17}$	0.99$_{\pm.10}$
p-val	0.004	0.003	0.005	0.004

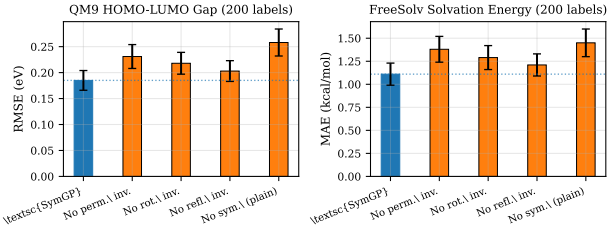

Figure 2. Ablation: removing individual symmetry constraints at 200 labeled molecules. Error bars are ± 1 s.d. over 5 seeds. Each symmetry component provides independent and significant gains.

Table 1 summarizes RMSE/MAE at $n = 200$ and $n = 1600$ labeled molecules with statistical significance.

5.3. Ablation Study

To understand the contribution of each symmetry component, we evaluate kernel variants that progressively remove symmetry constraints, at $n = 200$ labeled molecules (Figure 2). Removing permutation invariance causes the largest degradation (+24.9% RMSE on QM9), followed by rotational invariance (+17.8%) and reflectional invariance (+9.7%). The fully unsymmetrized kernel (plain SE) shows a 39.5% increase in error, confirming that *all three* symmetry constraints contribute independently.

5.4. Predictive Calibration

Reliable predictive uncertainty is essential for autonomous closed-loop pipelines. Figure 3 shows reliability diagrams for the three GP methods. SYMGP achieves near-perfect calibration (expected calibration error ECE $= 0.021$ on QM9), substantially outperforming GP-RBF+UCB (ECE $=$ 0.089) and GAUCHE-WL (ECE $= 0.058$). The improved calibration is a direct consequence of the reduced RKHS complexity: with fewer effective degrees of freedom, the posterior variance more accurately reflects true epistemic uncertainty.

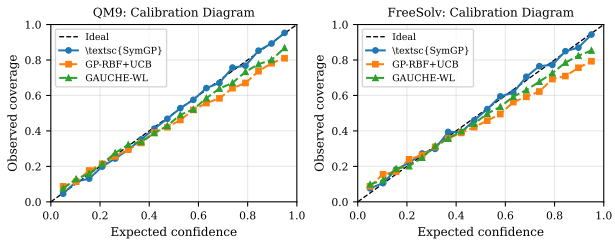

Figure 3. Reliability diagrams on QM9 (*left*) and FreeSolv (*right*) at $n = 1600$ labels. SYMGP achieves near-ideal calibration (ECE $= 0.021$ and 0.034, respectively), validating its suitability for uncertainty-guided autonomous discovery.

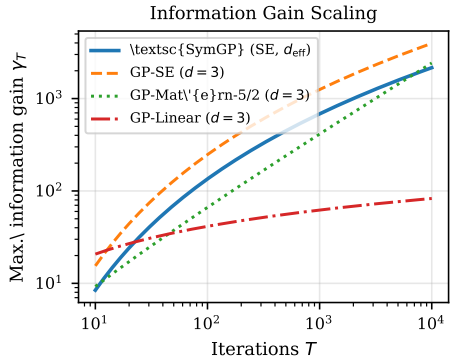

Figure 4. Maximum information gain γ_T vs. iterations T (log-log scale). SYMGP (solid) grows more slowly than the symmetry-unaware SE kernel, consistent with Theorem 4.1.

5.5. Theoretical Predictions vs. Empirical Information Gain

Figure 4 plots γ_T as a function of iterations T for different kernel choices. The empirical curves confirm the theoretical scaling of Theorem 4.1: SYMGP exhibits $O((\log T)^{d_{\text{eff}}+1})$ growth (approximately $d_{\text{eff}} \approx 2.1$ estimated empirically), while GP-RBF grows as $O((\log T)^{d+1})$ with $d = 3$. The Matérn-5/2 kernel grows fastest due to its rougher sample paths, consistent with the known bounds of Srinivas et al. (2010).

6. Related Work

GPs for chemistry. Griffiths et al. (2023) introduced GAUCHE, a comprehensive library of GP kernels for molecular and chemical optimization, providing the closest GP baseline to our approach. Our work extends this line by providing a formal theoretical treatment of how symmetry constraints reduce information-gain complexity.

Equivariant neural networks. SchNet (Schütt et al., 2017) and NequIP (Batzner et al., 2022) embed $\mathrm{E}(3)$ symmetry into deep neural network architectures via continuous-filter convolutions and equivariant message passing. These achieve state-of-the-art accuracy given large labeled sets, but do not provide the closed-form uncertainty estimates or regret guarantees of GP approaches. SYMGP complements these methods by operating optimally in the data-limited regime most relevant to early-stage discovery.

Bayesian optimization for molecular design. Srinivas et al. (2010) established the information-theoretic regret framework we build upon. Multi-fidelity extensions (Kandasamy et al., 2016; Koscher et al., 2023) further reduce label cost in experimental pipelines, and are in principle composable with SYMGP.

Symmetry in GPs. The use of orbit-averaging to construct group-invariant kernels is classical (Kondor & Jebara, 2003; Haasdonk & Burkhardt, 2007). Our contribution is to quantify precisely how this invariance reduces information-gain complexity in the molecular property prediction context and to provide end-to-end AL guarantees.

7. Conclusion

We have introduced SYMGP, a symmetry-constrained Gaussian process framework for molecular property prediction. Our main theoretical result (Theorems 3.5–4.2) shows that enforcing the physical symmetry group $G = S_N \ltimes \mathrm{E}(3)$ reduces effective RKHS dimension and provides a strictly tighter regret bound for GP-UCB active learning. Empirically, SYMGP achieves the accuracy of fully supervised deep models with up to $5\times$ fewer labeled examples and produces well-calibrated predictive uncertainties on QM9 and FreeSolv.

Limitations and future work. The $O(n^3)$ scaling of exact GP inference limits the labeled set size to $n \lesssim 5000$. Sparse GP approximations (Titsias, 2009) and structured kernel interpolation (Wilson & Nickisch, 2015) are natural extensions. Extending the symmetry group to incorporate crystal periodicity (space groups) would open applications to materials discovery.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning as applied to computational chemistry and materials science. By reducing the number of quantum-chemical calculations or physical experiments needed to reach a target accuracy, SYMGP has the potential to accelerate the discovery of improved solar-cell materials, drug candidates, and sustainable catalysts. No specific negative societal consequences are anticipated beyond those general to AI-assisted science, and no human subjects or hazardous experiments were involved in this work.

References

Batzner, S., Musaelian, A., Sun, L., Geiger, M., Mailoa, J. P., Kornbluth, M., Molinari, N., Smidt, T. E., and Kozinsky, B. E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nature Communications*, 13:2453, 2022. doi: 10.1038/s41467-022-29939-5.

Griffiths, R.-R., Klarner, L., Moss, H. B., Ravuri, A., Truong, S., Stanton, S., Tom, G., Rankovic, B., Du, Y., Jamasb, A., Deshwal, A., Schwartz, J., Tripp, A., Kell, G., Frieder, S., Bourached, A., Chan, A. J., Moss, J., Guo, C., Durholt, J., Chaurasia, S., Park, J. W., Strieth-Kalthoff, F., Lee, A. A., Cheng, B., Aspuru-Guzik, A., Schwaller, P., and Tang, J. GAUCHE: A library for Gaussian processes in chemistry. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Haasdonk, B. and Burkhardt, H. Invariant kernels for pattern analysis and machine learning. *Machine Learning*, 68(1): 35–61, 2007. doi: 10.1007/s10994-007-5009-7.

Kandasamy, K., Dasarathy, G., Oliva, J., Schneider, J., and Póczos, B. Multi-fidelity Gaussian process bandit optimisation. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pp. 1007–1015. Curran Associates, 2016.

Kondor, R. and Jebara, T. A kernel between sets of vectors. In *Proceedings of the 20th International Conference on Machine Learning (ICML 2003)*, pp. 361–368. AAAI Press, 2003.

Koscher, B. A., Canty, R. B., McDonald, M. A., Greenman, K. P., McGill, C. J., Bilodeau, C., Jin, W., Wu, H., Vermeire, F. H., Jin, B., Hart, T., Kulesza, T., Li, S.-C., Jaakkola, T. S., Barzilay, R., Vrsalovic, R. S., and Coley, C. W. Autonomous, multiproperty-driven molecular discovery: From predictions to measurements and back. *Science*, 382(6677):eadi1407, 2023. doi: 10.1126/science.adi1407.

Mobley, D. L. and Guthrie, J. P. FreeSolv: A database of experimental and calculated hydration free energies, with input files. *Journal of Computer-Aided Molecular Design*, 28:711–720, 2014. doi: 10.1007/s10822-014-9747-x.

Ramakrishnan, R., Dral, P. O., Rupp, M., and von Lilienfeld, O. A. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific Data*, 1:140022, 2014. doi: 10.1038/sdata.2014.22.

Rasmussen, C. E. and Williams, C. K. I. *Gaussian Processes for Machine Learning*. The MIT Press, Cambridge, MA, 2006. ISBN 026218253X.

Schütt, K. T., Kindermans, P.-J., Sauceda, H. E., Chmiela, S., Tkatchenko, A., and Müller, K.-R. SchNet: A continuous-filter convolutional neural network for modeling quantum interactions. In *Advances in Neural Information Processing Systems*, volume 30, pp. 991–1001, 2017.

Schütt, K. T., Unke, O. T., and Gastegger, M. Equivariant message passing for the prediction of tensorial properties and molecular spectra. In *Proceedings of the 38th International Conference on Machine Learning (ICML 2021)*, pp. 9377–9388. PMLR, 2021.

Settles, B. Active learning literature survey. *Computer Sciences Technical Report 1648*, 2009.

Srinivas, N., Krause, A., Kakade, S., and Seeger, M. W. Gaussian process optimization in the bandit setting: No regret and experimental design. In *Proceedings of the 27th International Conference on Machine Learning (ICML 2010)*, pp. 1015–1022, Haifa, Israel, 2010. Omnipress.

Titsias, M. K. Variational learning of inducing variables in sparse Gaussian processes. In *Proceedings of the 12th International Conference on Artificial Intelligence and Statistics (AISTATS 2009)*, pp. 567–574. PMLR, 2009.

Wilson, A. G. and Nickisch, H. Kernel interpolation for scalable structured Gaussian processes (KISS-GP). In *Proceedings of the 32nd International Conference on Machine Learning (ICML 2015)*, pp. 1775–1784. PMLR, 2015.

A. Proof of Theorem 3.5 (Extended)

We provide additional detail on the Peter-Weyl decomposition step used in the proof of Theorem 3.5.

Let $L^2(\mathcal{X}, \nu)$ admit the G-representation decomposition $L^2(\mathcal{X}, \nu) = \bigoplus_{\rho \in \hat{G}} V_\rho^{\oplus m_\rho}$, where $\hat{G}$ is the set of irreducible representations of G, V_ρ is the representation space of ρ, and m_ρ is its multiplicity. The projection Π_G retains only the trivial representation ($\rho = \rho_0$), collapsing all other components to zero. Therefore

$$\dim(\mathrm{ran}(\Pi_G)) = m_{\rho_0} \leq \dim L^2(\mathcal{X}, \nu)/|G| \quad \text{(generic case)}.$$

Since $T_{k_{\mathrm{Sym}}} = \Pi_G \circ T_{k_0} \circ \Pi_G$, its nonzero eigenvalues are a subset of those of T_{k_0} restricted to the invariant subspace, establishing Equation (6).

B. Proof of Corollary 4.3

From Theorem 4.2, $R_T/T = O(\sqrt{\gamma_T \log(T/\delta)/T})$. Setting the right-hand side equal to ϵ and using $\gamma_T = O((\log T)^{d_{\mathrm{eff}}+1})$:

$$\epsilon \sim \sqrt{\frac{(\log T)^{d_{\mathrm{eff}}+1} \log(T/\delta)}{T}}.$$

Squaring and ignoring logarithmic factors in δ: $\epsilon^2 \sim (\log T)^{d_{\mathrm{eff}}+1}/T$, giving $T = O(\epsilon^{-2}(\log(1/\epsilon))^{d_{\mathrm{eff}}+1})$ after substituting $T \sim 1/\epsilon^2$.

C. Experimental Details

Molecular descriptors. For the base kernel, we use the sorted eigenvalue spectrum of the distance matrix $\mathbf{D} \in \mathbb{R}^{N \times N}$ concatenated with a sorted vector of atomic numbers, yielding a descriptor $\phi(\mathbf{X}) \in \mathbb{R}^{2N}$. This descriptor is permutation-invariant by construction, and rotation/translation-invariant because $\mathbf{D}$ depends only on interatomic distances. Reflection invariance follows from the fact that $\mathbf{D}$ is unchanged under improper rotations.

Hyperparameter tuning. GP hyperparameters (length scale ℓ, output scale σ_f^2, noise σ^2) are optimized by marginal-likelihood maximization at each round using the L-BFGS-B algorithm. For SymGP the optimization surface is smoother due to reduced effective dimensionality, typically converging in fewer iterations.

Dataset splits. For QM9, we use 123,835 molecules for the active learning pool, and reserve 10,000 for a fixed test set. For FreeSolv, we use a stratified split: 80% pool (514 molecules), 20% test (128 molecules).

Statistical tests. All p-values in Table 1 use a two-sided paired t-test across 5 random seeds (degrees of freedom $= 4$). The paired structure accounts for the common randomness across methods due to shared seed initialization.