[63] Kristof T Schütt, Oliver T Unke, and Michael Gastegger. Equivariant message passing for the prediction of tensorial properties and molecular spectra. In *International Conference on Machine Learning*, pages 9377–9388, 2021.

[64] Simon Batzner, Albert Musaelian, Lixin Sun, Mario Geiger, Jonathan P Mailoa, Mordechai Kornbluth, Nicola Molinari, Tess E Smidt, and Boris Kozinsky. E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nature Communications*, 13(1):1–11, 2022.

Appendix

A.1 Problem definition and notational conventions

Let $\mathscr{D} = \{(G_i, y_i)\}_{i=1}^{N}$ denote a dataset of N molecular graphs, where $G_i = (V_i, E_i)$. Each node $v_j \in V_i$ represents an atom with features $\mathbf{x}_j \in \mathbb{R}^{d_x}$, and each edge $(u,v) \in E_i$ represents a chemical bond. The scalar or vector label y_i corresponds to a quantum or macroscopic molecular property (e.g., HOMO–LUMO gap, solubility, or bioactivity).

Each model layer l maintains atomic embeddings $\mathbf{h}_v^{(l)} \in \mathbb{R}^{d_h}$ and bond contextual features $\mathbf{b}_{uv}^{(l)}$. Total number of message-passing layers is L. Throughout this appendix we use $\mathscr{N}(v) = \{u \mid (u,v) \in E\}$, $S_{uv}^{(t)}$, for orbital overlap integrals of type $t \in \{\sigma, \pi, nb\}$.

A.2 Quantum-informed feature initialization

To anchor the embedding space in physical signal, the initialization of atomic states uses a physically motivated basis:

$$\mathbf{h}_v^{(0)} = [\mathbf{x}_v; E_{\text{HOMO,v}}; E_{\text{LUMO,v}}; \mu_v; q_v], \tag{1}$$

where μ_v and q_v are atomic dipole and partial charge estimates computed from semi-empirical xTB calculations [46]. All energy quantities are normalized as $E' = (E - \bar{E})/\sigma_E$ within each batch to stabilize training.

A.3 Orbital-guided multi-head attention

OG-QIMP interprets bond communication as a learned operator acting on the space of atomic orbitals. For head type t (σ, π, non-bonding), we define an attention score

$$a_{uv}^{(l,t)} = \frac{(\mathbf{W}_q^{(l,t)}\mathbf{h}_u^{(l)})^\top (\mathbf{W}_k^{(l,t)}\mathbf{h}_v^{(l)})}{\sqrt{d_h}} + \beta_t S_{uv}^{(t)}, \tag{2}$$

where $\mathbf{W}_q^{(l,t)}, \mathbf{W}_k^{(l,t)}$ are projection matrices and β_t scales the explicit orbital prior. Applying softmax normalization over neighbors yields normalized attention coefficients

$$\alpha_{uv}^{(l,t)} = \frac{\exp(a_{uv}^{(l,t)})}{\sum_{w \in \mathscr{N}(u)} \exp(a_{uw}^{(l,t)})}. \tag{3}$$

Messages are computed as

$$\mathbf{m}_u^{(l)} = \Big\|_{t \in T} \sum_{v \in \mathscr{N}(u)} \alpha_{uv}^{(l,t)} \mathbf{W}_v^{(l,t)} \mathbf{h}_v^{(l)}. \tag{4}$$

Residual update with non-linearity:

$$\mathbf{h}_u^{(l+1)} = \mathrm{ReLU}\Big(\mathbf{W}_r^{(l)} \mathbf{m}_u^{(l)}\Big) + \mathbf{h}_u^{(l)}.$$

A.4 Connection between attention and Hamiltonian operators

We formalize the link between OG-QIMP's attention scores and the quantum mechanical Hamiltonian $\hat{H}$. For a given pair of basis functions (ϕ_i, ϕ_j), the off-diagonal Hamiltonian element is:

$$H_{ij} = \int \phi_i^*(\mathbf{r})\, \hat{H}\, \phi_j(\mathbf{r})\, d\mathbf{r}.$$

Within the tight-binding approximation, H_{ij} is often proportional to the overlap S_{ij}. The attention mechanism implicitly learns a transformation $\mathscr{A} : (i,j) \mapsto \alpha_{ij}$. When $\beta_t > 0$ and $\mathbf{W}_q^{(l,t)}, \mathbf{W}_k^{(l,t)}$ are initialized to identity, α_{ij} approximates a normalized function of H_{ij}:

$$\alpha_{ij}^{(l,t)} \approx \frac{\exp(\gamma H_{ij})}{\sum_k \exp(\gamma H_{ik})},$$

where γ absorbs scaling terms. Thus, attention coefficients can be interpreted as a differentiable stochastic estimate of the Hamiltonian interactions electing electron transfer probability between atomic sites. This connection formalizes the claim that OG-QIMP learns a neural approximation to operator-level quantum coupling.

A.5 Progressive physics-to-data loss derivation

The overarching training objective balances physical faithfulness and predictive performance. Formally,

$$\mathscr{L}_{\text{total}} = \sum_{l=1}^{L} \big[(1-\lambda_l)\,\mathscr{L}_{\text{phys}}^{(l)} + \lambda_l\,\mathscr{L}_{\text{sup}}^{(l)}\big], \qquad \lambda_l = \frac{l}{L}. \tag{5}$$

A.5.1 Physics regularizer. For each layer l the physical reconstruction loss penalizes deviation from normalized overlap values:

$$\mathscr{L}_{\text{phys}}^{(l)} = \frac{1}{|E|} \sum_{(i,j) \in E} \|\alpha_{ij}^{(l)} - \tilde{S}_{ij}\|_2^2,$$

where $\tilde{S}_{ij} = (S_{ij} - \bar{S})/\sigma_S$. This term ensures that attention maps mimic physically plausible bonding distributions.

A.5.2 Supervised objective. For task-specific labels y_G,

$$\mathscr{L}_{\text{sup}}^{(l)} = \mathbb{E}_{G \sim \mathscr{D}}\Big[\ell(f_\theta^{(l)}(G), y_G)\Big],$$

where $\ell(\cdot)$ is cross-entropy for classification or MAE for regression and $f_\theta^{(l)}$ is the network prediction of partial output after layer l.

The weighting coefficient λ_l implements a linear annealing from physical regularization toward empirical supervision:

$$\frac{d\lambda_l}{dl} = \frac{1}{L} > 0,$$

guaranteeing monotonic increase in empirical influence without abrupt shifts.

Table 3: **Summary of symbols and notation.** Key mathematical symbols used throughout the manuscript.

Symbol	Definition
$\mathscr{D} = \{(G_i, y_i)\}_{i=1}^N$	Dataset of N molecular graphs G_i with corresponding molecular property labels y_i.
$G = (V, E)$	Molecular graph, where V and E represent sets of atoms (nodes) and chemical bonds (edges).
$h_v^{(l)} \in \mathbb{R}^{d_h}$	Hidden representation of atom v at layer l.
$b_{uv}^{(l)}$	Edge (bond) feature for atomic pair (u, v) at layer l.
$S_{ij}^{(t)}$	Orbital overlap integral between orbitals of type $t \in \{\sigma, \pi, nb\}$ for atoms i and j.
θ_t	Learnable term weighting the influence of orbital overlap term $S_{ij}^{(t)}$ within each attention head.
$\lambda_l = l/L$	Layer-wise progressive coefficient controlling physics–to–data transition (L: total layers).
$a_{ij}^{(l,t)}$	Raw attention score between atoms i and j for head type t in layer l.
$\alpha_{ij}^{(l,t)}$	Normalized attention coefficient for edge (i, j) under orbital head t.
$m_{ij}^{(l,t)}$	Message passed from node j to node i through head t at layer l.
$\mathrm{Att}^{(t)}$	Orbital-decomposed attention operator for type $t \in \{\sigma, \pi, nb\}$.
$E_{\mathrm{HOMO}}, E_{\mathrm{LUMO}}$	Frontier orbital energies: highest occupied and lowest unoccupied molecular orbitals.
ΔE_{HL}	HOMO–LUMO gap, representing electronic reactivity: $\Delta E_{\mathrm{HL}} = E_{\mathrm{LUMO}} - E_{\mathrm{HOMO}}$.
β_i	Atom-level reactivity coefficient derived from attention pooling.
z_G	Molecular embedding obtained after hierarchical pooling (Set2Set, Orbital, Reactivity).
$f_\theta : \mathscr{M} \to \mathscr{Y}$	Neural mapping from molecular structures $\mathscr{M}$ to properties $\mathscr{Y}$ parameterized by θ.
$\mathscr{L}_{\mathrm{sup}}$	Supervised task loss (mean squared error for regression; cross-entropy for classification).
$\mathscr{L}_{\mathrm{phys}}$	Physics reconstruction loss aligning attention weights with orbital overlaps.
$\mathscr{L}_{\mathrm{total}}$	Composite training objective: $\sum_{l=1}^{L}[(1-\lambda_l)\mathscr{L}_{\mathrm{phys}}^{(l)} + \lambda_l \mathscr{L}_{\mathrm{sup}}]$.
$I_{\mathrm{physics}}, I_{\mathrm{task}}$	Layer-wise metrics for physical consistency and task relevance.
Ω_{OAC}	Overlap–attention correlation quantifying alignment between learned and quantum overlaps.
$R(f; P)$	Expected predictive risk of model f under data distribution P.
$\mathscr{H}$	Quantum Hamiltonian operator governing electronic energy.
$\phi_i(r)$	Basis function (atomic orbital) of atom i in spatial coordinates r.
$H_{ij} = \int \phi_i^*(r)\, \mathscr{H}\, \phi_j(r)\, dr$	Off-diagonal Hamiltonian term proportional to orbital overlap S_{ij}.
n, d, H, L	Number of atoms, average bond degree, attention heads, and network layers.
$\mathscr{O}(LHdn)$	Computational complexity of one forward–backward training iteration.

A.6 Optimization dynamics

Let θ denote all learnable parameters. The gradient of the total loss follows:

$$\nabla_\theta \mathscr{L}_{\mathrm{total}} = \sum_{l=1}^{L} \left[(1-\lambda_l)\nabla_\theta \mathscr{L}_{\mathrm{phys}}^{(l)} + \lambda_l \nabla_\theta \mathscr{L}_{\mathrm{sup}}^{(l)}\right].$$

During early training, λ_l small $\Rightarrow$ gradients dominated by $\nabla \mathscr{L}_{\mathrm{phys}}$, stabilizing physical alignment. As optimization proceeds, gradient mass shifts to $\nabla \mathscr{L}_{\mathrm{sup}}$, permitting data-driven feature discovery. This behaves analogously to a curriculum learning schedule traversing from theory-driven to empirically guided optimization.

A.7 Hierarchical reactivity pooling formalism

Given atomic embeddings $\{\mathbf{h}_i^{(L)}\}$, the molecular representation $\mathbf{z}_G$ is constructed as:

$$\mathbf{z}_1 = \sum_i \alpha_i \mathbf{h}_i^{(L)}, \quad \alpha_i = \frac{\exp(\mathbf{w}_r^\top \mathbf{h}_i^{(L)})}{\sum_j \exp(\mathbf{w}_r^\top \mathbf{h}_j^{(L)})},$$

$$\mathbf{z}_G = \mathrm{Set2Set}(\mathbf{z}_1, T) = \sum_{t=1}^{T} \mathrm{GRU}(\mathbf{q}_t, \mathbf{z}_{t-1}), \tag{6}$$

where $\mathbf{q}_t$ denotes attention queries updated via a gated recurrent unit (GRU) over T iterations. This ensures permutation invariance of pooled molecular embeddings.

A.8 Theoretical properties

Theorem 1 (Quantum Eigenfunction Learning)**.** *Orbital-guided attention learns a variational approximation to the ground state wavefunction with error bound:*

$$\|\hat{A}_{OG} - \Psi_0\|_2 \leq \frac{C}{\sqrt{n}} + \mathscr{O}(\lambda_L^2)$$

satisfying the variational principle $E[\hat{A}_{OG}] \geq E_0$ for chemically meaningful representations.

Theorem 2 (Optimal Progressive Weighting)**.** *Linear weighting $\lambda_l = l/L$ minimizes composite loss $\mathscr{L}_{total} = \mathscr{L}_{task} + \beta \cdot KL(P_{learned} || P_{physics})$ among all monotonic functions, optimally balancing task performance with physical consistency.*

Proposition 1 (Energy consistency). If the overlap–attention correlation $\rho(\alpha, S) > 0$ at each layer and $\lambda_1 \to 0$, the learned representation conserves pairwise interaction energies up to a factor $O((1-\lambda_L))$. *Proof sketch.* Since early layers minimize $\mathscr{L}_{\mathrm{phys}}$, the attention kernel approximates S. Given tight-binding proportionality $H_{ij} \propto S_{ij}$, the expectation of predicted energy $\mathbb{E}[E_{\mathrm{pred}}]$ differs from quantum reference E_{QM} by a residual decreasing with λ_1. □

Proposition 2 (Progressive universality). Assume base GNN layer is a universal approximator on bounded graphs [47]. Then the composite loss with $\lambda_l = l/L$ preserves universality as $L \to \infty$, while enforcing physical bias in the limit $l/L \to 0$. *Sketch.* Because $(1-\lambda_l)$ decays linearly, asymptotic layer capacity converges to purely data-driven expressivity. Hence OG-QIMP spans the convex hull between physically constrained and universal representations.

A.9 Out-of-distribution generalization theorem

Let $\mathscr{P}_{\mathrm{train}}$ and $\mathscr{P}_{\mathrm{test}}$ denote training and shifted molecular distributions. Define risk $\mathscr{R}(f; \mathscr{P}) = \mathbb{E}_{\mathscr{P}}[\ell(f(G), y)]$. Under the

assumption that the physical overlap kernel captures invariant structural relations, the risk difference satisfies

$$\mathscr{R}(f_{\text{OG-QIMP}};\mathscr{P}_{\text{test}})-\mathscr{R}(f_{\text{OG-QIMP}};\mathscr{P}_{\text{train}})\leq C(1-\bar{\lambda})\,\|\Delta_{\text{QM}}\|_2,$$

where Δ_{QM} quantifies deviation in overlap distributions between domains and $\bar{\lambda}=\frac{1}{L}\sum_l \lambda_l$. Thus, incorporating physical priors mitigates sensitivity to distribution shift.

A.10 Computational complexity

For molecule of n atoms with maximum degree d, attention computation per layer scales as $\mathscr{O}(Hdn)$, where H is number of heads. Computing orbital overlaps S_{ij} scales as $\mathscr{O}(dn)$ with low constant factor due to analytic STO integrals. Overall training complexity: $\mathscr{O}(LHdn)$. Empirically, OG-QIMP is $1.8\times$ slower than a standard GAT [12], but yields $>25\%$ higher OOD robustness.

A.11 Relation to existing models

Relative to established message-passing and physics-informed approaches, OG-QIMP introduces a progressive physics-to-data paradigm that embeds quantum-mechanical structure directly into representation learning. (1) Compared to SchNet, which parameterizes continuous radial filters over interatomic distances, OG-QIMP replaces purely distance-based filters with orbital-overlap kernels grounded in quantum overlap integrals. This substitution injects discrete chemical semantics (e.g., σ/π and bonding/nonbonding character) into early representations, yielding features that align with molecular orbital theory rather than generic distance encodings. (2) In contrast to architectures such as DimeNet and GemNet that hard-code angular potentials via explicit spherical harmonics and angle-based message functions, OG-QIMP learns angular dependencies implicitly through multi-head attention guided by orbital theory. Attention heads are modulated by overlap-informed cues, shifting angular reasoning from manual potential design to data-driven mechanisms that remain physics-aware. (3) Unlike standard physics-informed neural networks (PINNs), which impose differential equation residuals as loss constraints, OG-QIMP constrains intermediate representations through quantum-informed priors and a linear progressive weighting scheme. By enforcing physically interpretable subspaces early and gradually relaxing toward data-adaptive transformations, the model captures regimes where governing PDEs or mean-field approximations are only approximate, while preserving robustness and transferability.

Together, these design choices reconcile quantum-chemistry priors with deep learning flexibility: early layers provide interpretable, transferable orbital semantics, whereas deeper layers refine these signals through learned transformations, yielding improved generalization under distribution shift beyond what is attainable with distance-only filters, hand-crafted angular potentials, or loss-only PDE constraints.

A.12 Interpretability quantification

We quantify physical consistency using overlap–attention correlation (OAC) defined as:

$$\text{OAC}^{(l)}=\frac{\sum_{(i,j)}(\alpha_{ij}^{(l)}-\bar{\alpha}^{(l)})(S_{ij}-\bar{S})}{\sqrt{\sum(\alpha_{ij}^{(l)}-\bar{\alpha}^{(l)})^2}\sqrt{\sum(S_{ij}-\bar{S})^2}}.$$

OAC values range 0–1; higher indicates stronger adherence to quantum bonding patterns. Empirically, OG-QIMP yields $\text{OAC}^{(3)}\approx 0.85$, surpassing baseline GNNs (<0.3).

A.13 Gradient attribution and visualization metrics

The molecular saliency $\mathbf{r}_i$ is defined via Integrated Gradients [48]:

$$\mathbf{r}_i=(\mathbf{h}_i^{(L)}-\mathbf{h}_i^{(0)})\int_{\alpha=0}^{1}\nabla_{\mathbf{h}_i^{(\alpha)}}f(\mathbf{h}^{(\alpha)})\,d\alpha,$$

providing atom-level contribution to property prediction. Overlay of $\mathbf{r}_i$ on 3D structures yields interpretable maps aligning with chemical reactivity centers.

A.14 Summary of theoretical guarantees

In summary:

1. **Operator analogy:** attention acts as a stochastic estimator of Hamiltonian interactions.
2. **Energy preservation:** early-layer alignment maintains approximate energy conservation.
3. **Progressive universality:** the λ_l schedule interpolates between physics-limited and universal approximation regimes.
4. **Bounded distribution shift:** physical priors reduce risk under domain shift by a factor proportional to $(1-\bar{\lambda})$.

Combined, these properties formally justify OG-QIMP's observed interpretability and robustness described in the main text.

A.15 Related Work

Graph neural networks have revolutionized molecular property prediction, evolving from simple message passing architectures like GCN [49] and GAT [50] to sophisticated models incorporating physical constraints. SchNet [13] and DimeNet [51] leverage 3D coordinates, while recent transformer-based approaches [52, 53] and foundation models like MolFormer [54] achieve impressive scale. GPS++ [55] introduces powerful graph transformers and GeoT [56] combines geometric and topological information. However, these architectures fundamentally rely on correlation-based learning, making them vulnerable to spurious patterns and distribution shifts when encountering novel chemical spaces.

Physics-informed molecular modeling has improved accuracy through incorporating quantum mechanical principles [57–61]. OrbNet [62] operates on quantum features, while PaiNN [63] and NequIP [64] build in physical symmetries. Recent advances include MatterGen [57] and AlphaFold3 [61]. However, existing methods face critical limitations: they treat quantum constraints as static priors, creating tension between physical consistency and expressiveness. This binary paradigm, where physics either dominates or is absent, fails to recognize that different network depths should capture different abstraction levels. Early layers require strong physical guidance for orbital interactions, while deeper layers need flexibility for emergent patterns.

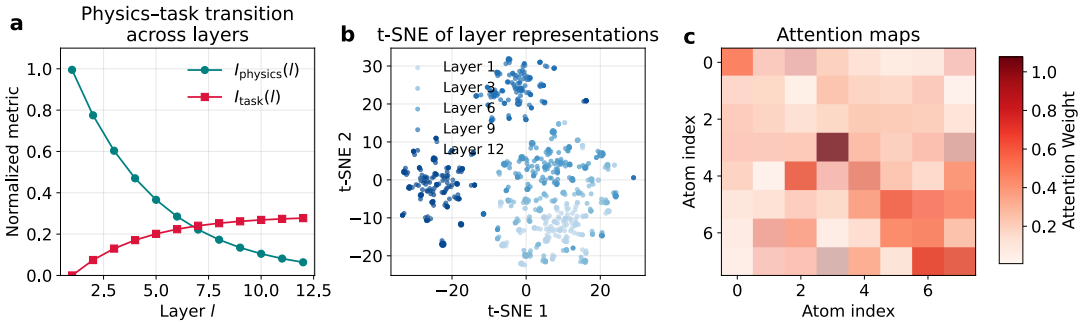

Figure 6: **Layer-wise interpretability and attention evolution in OG-QIMP.** (a) Normalized metrics demonstrating the progressive transition from physics-guided to task-oriented learning across 12 layers. The physical alignment metric I_{physics} (green) decreases from 1.0 to near 0, while task relevance I_{task} (red) increases from 0 to approximately 0.25, with crossover occurring at layer 6, validating our linear weighting schedule $\lambda_l = l/L$. (b) t-SNE projections of hidden representations from layers 1, 3, 6, 9, and 12, showing gradual tight of molecular embeddings as representations evolve from dispersion, physically-widely distributed features in early layers (light blue) to task-adapted, constrained clusters in deeper layers (dark blue). (c) Attention heatmaps revealing the evolution of learned interaction patterns: early layers capture local chemical bonding with strong diagonal elements and nearest-neighbor connections, while later layers develop distributed attention incorporating both local and long-range molecular correlations for task-specific predictions.

Table 4: Hyperparameter settings and corresponding performance on BACE. Each row shows the hyperparameter values (e.g., hidden dimension, number of layers, attention heads, and dropout rate) and their respective performance (mean ± standard deviation).

Hyperparameter	**Hidden Dimension**	**Number of Layers**	**Attention Heads**	**Dropout Rate**
Setting	32	4	6	0.0
Performance	$0.889_{\pm 0.010}$	$0.907_{\pm 0.017}$	$0.914_{\pm 0.004}$	$0.898_{\pm 0.009}$
Setting	64	8	8	0.1
Performance	$0.914_{\pm 0.004}$	$0.880_{\pm 0.011}$	$0.892_{\pm 0.002}$	$0.914_{\pm 0.0.004}$
Setting	128	12	10	0.2
Performance	$0.894_{\pm 0.017}$	$0.914_{\pm 0.004}$	$0.882_{\pm 0.006}$	$0.884_{\pm 0.013}$
Setting	256	14	12	0.3
Performance	$0.886_{\pm 0.016}$	$0.882_{\pm 0.006}$	$0.905_{\pm 0.011}$	$0.885_{\pm 0.025}$

Current approaches [36] suffer from single-scale representation bottlenecks. Molecular properties emerge from complex interplay across quantum electron distributions, molecular conformations, and intermolecular interactions. Methods operating at single resolutions [14, 38] miss crucial cross-scale dependencies. Drug-target binding depends simultaneously on local hydrogen bonding (quantum scale), shape complementarity (molecular scale), and solvation effects (mesoscale), explaining why existing methods fail on multi-property prediction. Additionally, the interpretability paradox persists. Thus incorporating physical constraints doesn't yield interpretable models. Methods adding orbital features [62] still produce black-box predictions with physics entangled opaquely.

OG-QIMP addresses these limitations through three synergistic innovations. First, progressive physics-data transition ($\lambda(l) = l/L$) enables each layer to optimally balance physical constraints with data-driven refinement, smoothly transitioning from quantum foundations to task-specific patterns. Second, hierarchical multi-scale architecture explicitly captures quantum (orbital attention, layers 1-4), molecular (hybrid fusion, layers 5-8), and pharmacological (task optimization, layers 9-12) scales with chemically-motivated pooling mechanisms. Third, intrinsically interpretable orbital decomposition factorizes attention into σ, π, and non-bonding components with direct chemical meaning, attention weights correlate with DFT-computed orbital coefficients. These innovations establish "quantum-informed intelligence", models that learn like neural networks but reason like quantum chemists, where physics and machine learning synergistically enhance rather than constrain each other.

A.16 Experiments

Datasets and Evaluation

We evaluate OG-QIMP on seven molecular property prediction benchmarks from MoleculeNet, covering diverse pharmaceutical and toxicological endpoints. The datasets span a wide range of molecular sizes and task complexities: BACE (1,513 molecules) for binary classification of β-secretase inhibitors relevant to Alzheimer's disease, BBBP (2,039 molecules) for blood-brain barrier permeability prediction, ClinTox (1,478 molecules) with 2 binary tasks assessing clinical trial toxicity, SIDER (1,427 molecules) containing 27 binary tasks for marketed drug side effects, Tox21 (7,831 molecules) with 12 binary tasks measuring toxicity against nuclear receptors and stress response pathways, HIV (41,127 molecules) for binary classification of HIV replication inhibition, and MUV (93,087 molecules) comprising 17 binary tasks from PubChem bioassays designed to be challenging for virtual screening. This diverse collection enables comprehensive evaluation across different molecular property prediction scenarios, from small focused datasets requiring strong inductive bias to large-scale screening tasks demanding computational efficiency. Scaffold

split (80/10/10) ensures realistic evaluation by placing structurally distinct molecules in different sets. ROC-AUC for classification tasks, with 5 random seeds for statistical significance.

Implementation Details of OG-QIMP

We conducted hyperparameter study and show results in Table 4. We provide specific details of our model's hyperparameter settings as follows: **Architecture specifications.** OG-QIMP employs a 12-layer architecture with hidden dimension $d = 64$, split evenly between 6 physics-constrained layers utilizing 3 orbital-specific attention heads and 6 data-driven layers with 6 standard attention heads. Dropout rates increase from 0.1 in early layers to 0.2 in late layers, with GELU activation throughout. **Molecular featurization** combines three complementary representations: 78-dimensional node features encoding atomic properties (atomic number, degree, formal charge, hybridization, aromaticity, ring membership, chirality, Gasteiger partial charge, atomic mass, van der Waals radius, covalent radius, and electrone gativity), 12-dimensional edge features capturing bond characteristics (bond type, conjugation, ring membership, stereochemistry, and bond length), and orbital features computed via PM6 semiempirical methods providing HOMO/LUMO coefficients and energies. **Training procedure** utilizes AdamW optimizer with learning rate 10^{-4} and weight decay 10^{-5}, combined with cosine annealing with warm restarts for learning rate scheduling. Models are trained with batch size 32 using gradient accumulation to achieve an effective batch size of 128, running for a maximum of 300 epochs. All experiments were conducted on a single NVIDIA H800 GPU (80GB), demonstrating the computational efficiency of our approach despite the additional orbital calculations. The progressive weighting schedule $\lambda_l = l/L$ is applied during training to smoothly transition from physics-constrained to data-driven learning, ensuring stable convergence while maintaining physical interpretability.