VARIATIONAL QUANTUM ATTENTION FOR MOLECULAR GRAPH LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Molecular property prediction is central to computational drug discovery, where graph neural networks learn to weight neighboring atomic environments during message passing. Yet it remains unclear how variational quantum circuits alter learned attention behavior in molecular graphs. We introduce an edge-aware variational quantum attention mechanism for molecular graph learning, in which the receiving atom, neighboring atom, and connecting bond jointly determine the quantum attention state. Across five molecular property and bioactivity prediction tasks, QGAT achieves competitive performance relative to GATv2, with a consistent improvement on BBBP across all six evaluated circuit ansatzes. We further compare how the quantum and classical attention scores weight molecular structure beyond accuracy. In the Verubecestat BACE1 inhibitor series, QGAT achieves a higher Spearman correlation than GATv2 and assigns positive attributions to several structural changes consistent with reported structure-activity relationships (SARs). This case study shows that the two attention mechanisms can exhibit different prediction and attribution behavior across structurally related BACE1 analogues, while broader validation is required to determine how consistently these differences generalize across chemical series and targets. Circuit ablations further show that performance depends on the circuit design. Together, these results show that variational quantum attention can serve as a viable alternative molecular attention parameterization while inducing circuit- and chemistry-dependent behavior distinct from a matched classical scorer.

1 INTRODUCTION

Molecular property prediction plays a central role in computational drug discovery, where predictive models are used to prioritize compounds before costly experimental evaluation (Berdigaliyev & Aljofan, 2020). Molecular graphs provide a natural representation for this task: atoms are represented as nodes, bonds are represented as edges, and molecular properties arise from interactions among local chemical environments. Graph neural networks (GNNs) exploit this graph structure by propagating information along molecular bonds and have been widely applied to physicochemical and bioactivity prediction tasks (Gilmer et al., 2017; Faria et al., 2025). Graph attention networks (GATs) (Brody et al., 2022; Veličković et al., 2017) further introduce learnable attention coefficients that adaptively weight neighboring atoms during message aggregation. Beyond their role in prediction, these learned attention weights characterize how local structural information is differentially weighted during message passing.

We observe that the attention-scoring step in GAT provides a natural entry point for hybrid quantum-classical modeling, since it operates on local atom-pair representations rather than on the entire molecular graph. This local and edge-wise formulation avoids replacing the full node encoder and eliminates the need to encode high-dimensional molecular features directly into a limited quantum state space. By contrast, fully quantum graph models can require quantum resources that grow with graph size; representative formulations encode an n-node graph using n qubits (Sein et al., 2026; Raj et al., 2026; Sauvage et al., 2026). Such constructions provide a direct quantum representation of graph structure, but substantially increase the cost of simulation on molecular tasks. For these reasons, we use variational quantum circuits (VQCs) (Abbas et al., 2021; Du et al., 2021; Schuld et al., 2021; Hsu et al., 2025b) specifically as the attention scorer while retaining the classical pipeline for

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

feature transformation, message aggregation, readout and prediction. This design isolates the effect of the quantum scoring function and allows us to examine how replacing the classical scorer changes molecular prediction and weighting behavior.

Recent quantum graph attention network (QGAT) models have demonstrated that VQCs can be used to generate attention coefficients from quantum encoded graph representations (Ning et al., 2025). Existing quantum graph attention formulations are primarily developed for generic graph settings and do not explicitly incorporate molecular bond attributes into the quantum attention computation. This distinction is important for molecular graphs, where the relation between neighboring atoms is chemically defined by the connecting bond rather than by graph adjacency alone (Ma et al., 2022). Performance alone also does not reveal how replacing a classical attention scorer with VQC changes the weighting of molecular structure during message passing. This leads to two questions: can an edge-aware VQC provide a competitive molecular attention scorer, and does its circuit induced parameterization lead to systematically different molecular weighting behavior?

To address these questions, we formulate an edge-aware molecular variational quantum attention mechanism in which the receiving atom, neighboring atom, and connecting bond jointly determine the attention score. The remaining message passing pipeline is kept unchanged and classical, making the attention scorer the principal architectural intervention between QGAT and its GATv2 counterpart.

We evaluate the QGAT across five molecular property and bioactivity prediction tasks using scaffold splits, and then examine how replacing the classical attention scorer changes learned molecular weighting. We analyze the resulting behavior using substructure attribution, attention weights, and integrated gradients (IG), and relate these differences to experimentally reported structure-activity relationships (SARs) in the Verubecestat (MK-8931) BACE1 inhibitor series (Scott et al., 2016).

In summary, our contributions are as follows:

- We introduce an edge-aware variational quantum attention mechanism for molecular graphs that jointly incorporates atom and bond information into the quantum attention computation while retaining the surrounding classical message passing pipeline.
- We evaluate QGAT across five molecular property and bioactivity tasks under scaffold splits and compare it with a matched GATv2 baseline. Across the evaluated circuit family, QGAT achieves competitive predictive performance relative to matched GATv2, including a consistent improvement on BBBP across all evaluated ansatzes; circuit ablations further reveal strong task dependence on ansatz and variational depth.
- We characterize the resulting prediction behavior using substructure masking, attention analysis, and IG. In the Verubecestat series, QGAT achieves a higher activity ranking correlation than GATv2 and produces several attribution patterns that are directionally consistent with reported SAR.

2 Related Work

Hybrid quantum–classical machine learning has increasingly been explored for drug-related prediction, typically by using classical neural networks to extract or compress molecular and biological representations before applying parameterized quantum circuits as encoding or prediction modules (Santagati et al., 2024; Li et al., 2026). Representative studies have applied VQC to protein–ligand binding-affinity, drug-response, and molecular property prediction (Domingo et al., 2023; Jeong et al., 2025; Sagingalieva et al., 2023; Lu et al., 2025; Dong et al., 2023). These studies demonstrate diverse forms of quantum–classical integration, but the quantum component is typically used for representation encoding or downstream prediction rather than for determining graph attention coefficients.

Recent studies in quantum graph learning have incorporated variational quantum circuits (VQCs) into attention-based message passing. Ning et al. (2025) showed that VQCs can generate multi-head attention coefficients from quantum encoded node representations, while Faria et al. (2025) used quantum graph encoders to modulate neighboring-node contributions. Our work builds on this line by studying variational quantum attention in molecular graphs, where both atom and bond features are incorporated directly into the attention computation. This raises a question beyond whether

quantum attention can match classical performance: does replacing the classical scorer with a VQC change which molecular information the model relies on when making predictions? Answering this question is important because two models with similar accuracy can nevertheless rely on different structural signals, which may affect their behavior on specific chemical series or local structural modifications.

3 PRELIMINARIES

We briefly introduce the graph attention network and quantum computing operations used throughout this work.

3.1 GRAPH ATTENTION NETWORKS

GATs (Veličković et al., 2017; Brody et al., 2022) aggregate information from neighboring nodes using learned attention coefficients. For a receiving node u and a neighbor v, an attention scorer first produces an unnormalized logit e_{uv} from their node representations. The coefficient assigned to v is then obtained by normalizing the logits over the neighborhood $\mathcal{N}_u$:

$$\alpha_{uv} = \frac{\exp(e_{uv})}{\sum_{w \in \mathcal{N}_u} \exp(e_{uw})}, \tag{1}$$

The coefficient α_{uv} determines the contribution of node v to the updating of node u. Multi-head GATs compute this operation in parallel across multiple heads.

3.2 QUANTUM MACHINE LEARNING

Quantum Machine Learning (QML) combines parameterized quantum circuits with classical learning pipelines to construct trainable hybrid models (Cerezo et al., 2022; Biamonte et al., 2017; Benedetti et al., 2019). A typical workflow consists of three stages: a classical transformation that maps the input into a feature representation, a quantum data encoding procedure that prepares the corresponding quantum state, and a VQC whose measurement outcomes are returned to the classical computation.

Let an input be denoted by $\mathbf{x} \in \mathbb{R}^d$. A classical transformation $f_{\boldsymbol{\theta}_c}(\cdot)$, parameterized by $\boldsymbol{\theta}_c$, maps the input to a lower dimensional feature vector

$$\mathbf{z} = f_{\boldsymbol{\theta}_c}(\mathbf{x}) \in \mathbb{R}^{n_f} \tag{2}$$

To encode $\mathbf{z}$ into a quantum state, we use amplitude encoding. The feature vector is written as $\tilde{\mathbf{z}} \in \mathbb{R}^{2^{n_q}}$ with $\sum_i |\tilde{z}_i|^2 = 1$. An n_q-qubit state is then prepared as

$$|\psi(\tilde{\mathbf{z}})\rangle = V(\tilde{\mathbf{z}})|0\rangle^{\otimes n_q} = \sum_{i=0}^{2^{n_q}-1} \tilde{z}_i |i\rangle, \tag{3}$$

where $V(\tilde{\mathbf{z}})$ denotes the state preparation operation. Amplitude encoding represents a 2^{n_q}-dimensional vector using n_q qubits, reducing the qubit requirement for high dimensional inputs. However, preparing an arbitrary amplitude encoded state may require substantial circuit depth.

The encoded state is subsequently processed by a variational circuit $U(\boldsymbol{\theta})$ composed of parameterized single-qubit rotations and entangling gates (e.g. CNOTs). For a circuit with N_b variational blocks,

$$U(\boldsymbol{\theta}) = \prod_{\ell=1}^{N_b} U_\ell\big(\boldsymbol{\theta}^{(\ell)}\big) \tag{4}$$

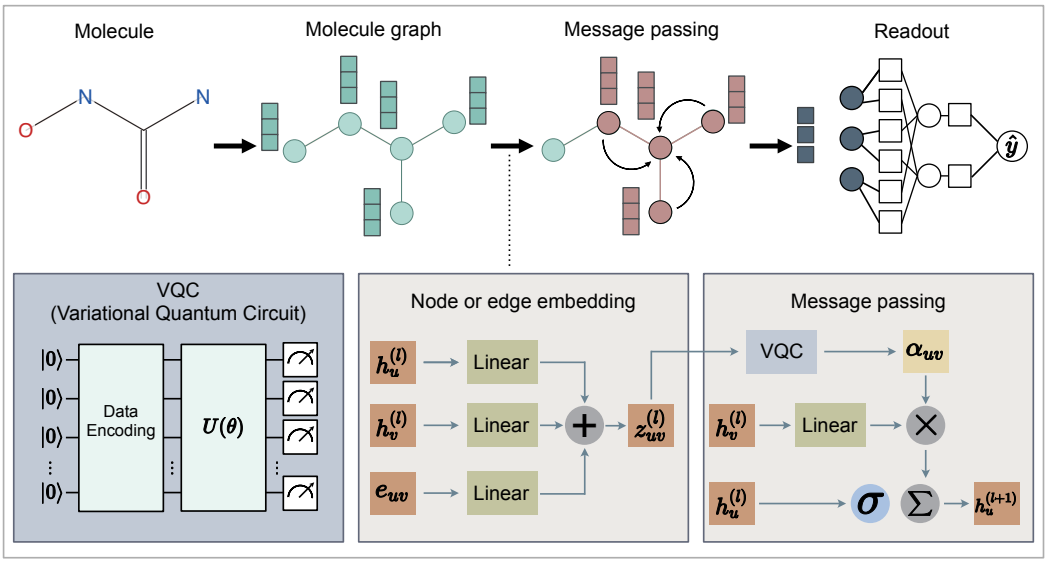


Figure 1: Overview of the edge-aware QGAT architecture for molecular property prediction.

where the structure of each block U_ℓ is determined by the chosen circuit ansatz. The resulting quantum state is

$$|\psi(\tilde{\mathbf{z}}, \boldsymbol{\theta})\rangle = U(\boldsymbol{\theta})V(\tilde{\mathbf{z}})|0\rangle^{\otimes n_q}. \tag{5}$$

Classical information is extracted from the quantum state through expectation values of observables. For an observable operator $\hat{O}_k$, the corresponding readout is:

$$q_k = \langle\psi(\tilde{\mathbf{z}}, \boldsymbol{\theta})| \hat{O}_k |\psi(\tilde{\mathbf{z}}, \boldsymbol{\theta})\rangle \tag{6}$$

These expectation values provide continuous classical outputs from the quantum circuit and are used in our model to construct the attention logits described in Section 4.

4 Method

We adapt QGAT (Ning et al., 2025) to molecular graphs by incorporating atom and bond representations directly into the quantum attention computation. The quantum module replaces the attention scorer while the message passing and readout prediction remain classical. Figure 1 summarizes the resulting architecture.

4.1 Molecular graph representation

Each molecule is represented as a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ denotes the atoms and $\mathcal{E}$ denotes the covalent bonds. Each atom u is associated with a node feature vector $\mathbf{h}_u \in \mathbb{R}^d$, and each bond (u, v) is associated with an edge feature vector $\mathbf{e}_{uv} \in \mathbb{R}^{d_e}$. Full details of the molecular graph construction and node and edge descriptors are provided in Appendix A.3.

4.2 Edge-aware variational quantum attention

For each edge (u, v), the receiving node, neighboring node, and edge features are independently projected into a common multi-head representation. Rather than concatenating the three projected representations, we combine them through element-wise addition:

$$\mathbf{z}_{uv} = \mathbf{W}_u \mathbf{h}_u + \mathbf{W}_v \mathbf{h}_v + \mathbf{W}_e \mathbf{e}_{uv} \in \mathbb{R}^{H d_{\text{out}}} \tag{7}$$

where H is the number of attention heads and d_{out} is the latent dimension per head. Direct concatenation would increase the latent dimension to $3Hd_{\text{out}}$, requiring either additional qubits or an extra compression projection before quantum encoding. We therefore combine the three projected representations by addition and set the resulting dimension to $Hd_{\text{out}} = 2^{n_q}$. The fused representation can then be normalized and directly amplitude encoded without an additional dimensionality reduction layer. In our configurations, Hd_{out} is 64 or 128, corresponding to 6 or 7 qubits. A direct angle encoding would require 64 or 128 qubits. Therefore, amplitude encoding is particularly relevant because the attention circuit is evaluated for molecular edges throughout message passing.

The encoded state is processed by a variational quantum circuit $U(\boldsymbol{\theta})$. Each variational block consists of trainable single-qubit rotations and entangling operations. We evaluate six circuit ansatzes with different rotation and entangling structures; their complete circuit definitions are provided in Appendix B.1.

For each attention head, one Pauli-Z expectation value is used as the corresponding quantum attention logit:

$$e_{uv}^{(k)} = \langle \psi(z_{uv}) | U^\dagger(\theta) \hat{O}_k U(\theta) | \psi(z_{uv}) \rangle \tag{8}$$

where $\hat{O}_k$ denotes the observable associated with the k-th measured qubit. In this work, we choose $\hat{O}_k = Z_k$, i.e., the Pauli-Z operator acting on the k-th qubit. We use configurations satisfying $H \leq n_q$, assigning each attention head to a distinct Pauli-Z observable. Because these observables commute and are diagonal in the computational basis, their expectation values can be estimated from the same set of measurement outcomes.

Since Pauli-Z expectation values are bounded within $[-1, +1]$, the range of pairwise logit differences is also bounded, which limits the sharpness of the subsequent softmax distribution. We therefore introduce a head-specific learnable scale s_k:

$$\tilde{e}_{uv}^{(k)} = s_k e_{uv}^{(k)} \tag{9}$$

The scaled logits are normalized over the incoming neighborhood u using softmax:

$$\alpha_{uv}^{(k)} = \frac{\exp \tilde{e}_{uv}^{(k)}}{\sum_{w \in N_u} \exp \tilde{e}_{uw}^{(k)}} \tag{10}$$

where $\mathcal{N}_u$ denotes the neighbors of receiving node u. The resulting attentions are used to weight classical neighbor messages. The output of the H attention heads is concatenated as:

$$\mathbf{h}'_u = \sigma\left(\Big\Vert_{k=1}^{H} \sum_{v \in \mathcal{N}_u} \alpha_{uv}^{(k)} \mathbf{W}^{(k)} \mathbf{h}_v \right) \tag{11}$$

where $\|$ denotes concatenation across attention heads and $\mathbf{W}^{(k)}$ denotes the value projection matrix for the k-th head, and σ is a non-linear activation function.

5 Experiments and Results

5.1 Experimental Settings

We evaluate QGAT on five molecular property and bioactivity prediction benchmarks covering both classification and regression tasks. BACE and BBBP are binary classification datasets, while FreeSolv and ESOL are regression benchmarks. We additionally construct a BACE1 bioactivity regression task from ChEMBL (Gaulton et al., 2012). We focus on molecular property and bioactivity tasks. Large-scale or highly sparse multitask benchmarks, such as HIV, Tox21, SIDER, and MUV

Table 1: Results on molecular property prediction tasks. Performance is reported as the mean ± standard deviation over five scaffold folds. The best and second best results for each metric are highlighted in red and blue, respectively.

Model	Classification		Regression			
	BACE	BBBP	FreeSolv	ESOL	ChEMBL-BACE	
	AUC ↑	AUC ↑	RMSE ↓	RMSE ↓	Spearman ↑	RMSE ↓
No. of molecules	1,513	2,039	642	1,128	4,580	
Avg. no. of atoms	65	46	18	26	59	
D-MPNN	0.7591 ± 0.0305	0.6713 ± 0.0230	2.5079 ± 0.1136	1.0740 ± 0.0813	0.6520 ± 0.0266	0.9804 ± 0.0261
AttentiveFP	0.7872 ± 0.0337	0.6469 ± 0.0273	1.9674 ± 0.3156	0.9672 ± 0.0666	0.6455 ± 0.0398	0.9970 ± 0.0456
GIN	0.7977 ± 0.0195	0.6670 ± 0.0207	2.2618 ± 0.7023	0.8997 ± 0.0391	0.7115 ± 0.0109	0.9257 ± 0.0122
GPS	0.7864 ± 0.0227	0.6580 ± 0.0414	2.0111 ± 0.2729	$\mathbf{0.8637 \pm 0.0372}$	0.6457 ± 0.0225	0.9831 ± 0.0250
KAGAT	0.8005 ± 0.0231	0.6568 ± 0.0253	2.3897 ± 0.2029	1.1366 ± 0.0451	0.6679 ± 0.0216	0.9738 ± 0.0207
m-GAT	0.8244 ± 0.0072	0.6942 ± 0.0303	1.8374 ± 0.1506	0.9497 ± 0.0680	0.7166 ± 0.0294	0.8985 ± 0.0385
GATv2	0.8210 ± 0.0202	0.6808 ± 0.0111	$\mathbf{1.7268 \pm 0.2214}$	0.9389 ± 0.0660	0.7313 ± 0.0201	0.8895 ± 0.0330
QGAT	$\mathbf{0.8283 \pm 0.0133}$	$\mathbf{0.7157 \pm 0.0095}$	1.7400 ± 0.2033	0.9103 ± 0.0317	$\mathbf{0.7357 \pm 0.0147}$	$\mathbf{0.8800 \pm 0.0188}$

are therefore not included in the present study. We use ROC-AUC for classification and RMSE for regression. ChEMBL-BACE additionally reports Spearman rank correlation to measure whether compounds are ordered consistently with their experimental activities, which is particularly relevant for compound prioritization in drug discovery. Dataset statistics and preprocessing details are provided in Appendix A.1.

To evaluate generalizing beyond closely related molecular structures, we group molecules according to their Bemis–Murcko scaffold splitting (Bemis & Murcko, 1996). Scaffold splitting reduces train–test structural overlap and evaluates generalization to unseen chemical scaffolds than random splitting. We first reserve 10% of each dataset as a fixed scaffold based test set. The remaining 90% is partitioned over five folds using scaffold groups, resulting in 72% training, 18% validation, and 10% test data in each fold. Results are reported as mean ± standard deviation across the five scaffold folds. Further details of the splitting procedure are provided in Appendix A.2.

To ensure fair comparisons, we adopt consistent experimental settings across all models, including D-MPNN (Yang et al., 2019), AttentiveFP (Xiong et al., 2020), GIN (Xu et al., 2019), GPS (Rampášek et al., 2022), KA-GAT (Li et al., 2025), m-GAT (Mathew et al., 2025), and GATv2 (Brody et al., 2022). GATv2 serves as the primary classical attention baseline because it retains the same graph representation and message passing framework while using a classical attention scorer. All models use the same molecular descriptors, hidden dimension, number of attention heads where applicable, and training settings. Details of the hyperparameter configurations and input feature constructions are provided in Appendix A.4.

5.2 Main Results on Molecular Property Prediction

We compare QGAT with seven graph learning baselines across five molecular property prediction tasks. Table 1 summarizes the strongest observed QGAT configuration for each task together with classical graph-learning baselines. Across tasks, QGAT remains competitive with the matched GATv2 baseline, although the magnitude and direction of the difference depend on the dataset and circuit configuration.

The largest difference between QGAT and GATv2 is observed on BBBP, where the mean ROC-AUC increases from 0.6808 to 0.7157. QGAT also achieves higher mean performance on BACE, BBBP, ESOL, and both ChEMBL-BACE metrics, whereas GATv2 obtains a slightly lower RMSE on FreeSolv. These results indicate that the effect of variational quantum attention is task dependent rather than uniformly beneficial across all molecular prediction settings.

Across datasets, QGAT uses approximately 0.5%–1.3% fewer trainable parameters than GATv2. These results show that VQC scorer can replace the classical attention scorer without a systematic loss in performance or an increase in model size.

5.3 Ablation Study: The Effect of Variational Quantum Circuit Design

We examine two aspects of the variational circuit design: the circuit ansatz and the number of variational layers N_b. Rather than searching over unrestricted circuit structures, we consider six shallow circuit ansatzes constructed from parameterized single-qubit gates and CNOT entangling gates. The candidates differ in rotation ordering and CNOT connectivity while keeping the overall circuit scale comparable. A representative circuit block and the complete definitions of all six ansatzes are provided in Appendix B.1.

Table 2 compares the six circuit designs across the molecular prediction tasks. No single ansatz consistently dominates across datasets, indicating that circuit architecture is not a neutral implementation detail but a task-dependent modeling choice.

Table 2: Performance across six variational circuit ansatzes. Results are reported as the mean ± standard deviation across five scaffold folds. The best result for each evaluation metric is highlighted in red. Ansatz 6 results on ChEMBL-BACE are omitted due to the substantially higher computational cost of this configuration under our simulation budget.

Ansatz	BACE	BBBP	FreeSolv	ESOL	ChEMBL-BACE	
	AUC ↑	AUC ↑	RMSE ↓	RMSE ↓	Spearman ↑	RMSE ↓
1	**0.8283 ± 0.0133**	0.6920 ± 0.0069	**1.7400 ± 0.2033**	0.9278 ± 0.0564	0.7263 ± 0.0083	0.8955 ± 0.0140
2	0.8136 ± 0.0132	0.7023 ± 0.0092	1.7435 ± 0.0805	0.9962 ± 0.1096	0.7228 ± 0.0152	0.9041 ± 0.0185
3	0.8143 ± 0.0131	**0.7157 ± 0.0095**	1.7942 ± 0.1943	0.9806 ± 0.0941	0.7227 ± 0.0184	0.9041 ± 0.0261
4	0.8138 ± 0.0154	0.7000 ± 0.0148	1.8142 ± 0.1723	0.9342 ± 0.0377	**0.7357 ± 0.0147**	**0.8800 ± 0.0188**
5	0.8063 ± 0.0105	0.7058 ± 0.0182	1.8783 ± 0.2774	0.9288 ± 0.0559	0.7185 ± 0.0121	0.9070 ± 0.0239
6	0.8044 ± 0.0219	0.6866 ± 0.0150	1.8838 ± 0.2699	**0.9103 ± 0.0317**	–	–

We further vary the number of variational layers over $N_b \in \{1, 2, 3, 4\}$. The complete results are reported in Appendix B.2. Increasing circuit depth does not lead to monotonic improvements. BACE and FreeSolv obtain their best performance with a single variational layer, whereas BBBP performs best with three layers and ESOL and ChEMBL-BACE perform best with two layers. Across the evaluated tasks, the strongest configurations are obtained with one to three variational layers, and additional depth does not provide a consistent benefit.

5.4 Analysis of Molecular Attention and Attribution

We next compare QGAT and GATv2 at two distinct levels: attention behavior during message passing and attribution with respect to the final prediction. Substructure mask explanation (SME) (Wu et al., 2023) measures changes in model output after masking predefined molecular substructures, while integrated gradients (IG) provides signed feature level attributions with respect to the final prediction. We additionally inspect learned attention weights to characterize how neighboring information is weighted during message passing.

Functional group attribution on ESOL. We first apply SME to the ESOL task, where functional group effects admit a comparatively direct physicochemical interpretation. Table 3 summarizes representative attribution values, while the complete distributions are reported in Appendix C.2. Both models recover broadly similar attribution directions for several functional groups. Hydrophobic substituents such as CF_3 and tBu receive negative contributions, whereas several polar groups exhibit positive contributions to predicted solubility. The two models nevertheless differ in attribution magnitude. In particular, QGAT assigns weaker positive contributions than GATv2 to several carbonyl-containing groups, including $C(=O)N$, $C(=O)CH_3$, and $C(=O)OMe$. These results indicate that QGAT preserves several broad structure–solubility trends while exhibiting distinct sensitivities to individual functional groups.

Case study on the Verubecestat (MK-8931) BACE1 inhibitor series. We next examine the Verubecestat series reported during BACE1 inhibitor modification (Scott et al., 2016). No exact molecular matches are present in the ChEMBL-BACE dataset; one analogue has a nearest Tanimoto similarity of 0.92, whereas all remaining compounds have maximum similarity at or below 0.67. We therefore treat this series as an external medicinal chemistry case study rather than a strict chemical out-of-distribution benchmark. This series is particularly informative because it combines a documented medicinal chemistry optimization trajectory with experimentally characterized bind-

Table 3: Representative key functional group attributions produced by GATv2 and QGAT on ESOL.

Functional groups	GATv2	QGAT
$-\mathrm{OH}$	0.411	0.475
$-\mathrm{NH_2}$	0.283	0.120
$-\mathrm{CF_3}$	-1.023	-1.092
$-\mathrm{tBu}$	-0.611	-0.523
$-\mathrm{C(=O)N}$	0.267	0.052
$-\mathrm{C(=O)CH_3}$	0.275	0.084
$-\mathrm{C(=O)OMe}$	0.496	0.185

ing information. Additional results on the Umibecestat (Machauer et al., 2021) and LY3202626 (McKinzie et al., 2021) series are reported in Appendix D.3.

Across the series, QGAT achieves a higher Spearman correlation than GATv2 (0.8264 vs. 0.7954), whereas GATv2 obtains a slightly lower RMSE (0.9920 vs. 1.0306). Because the ChEMBL-BACE model is trained primarily on standardized $\mathrm{IC_{50}}$ values whereas the Verubecestat series is evaluated using K_i, we use Spearman correlation as the primary measure of whether the models recover the experimental activity ordering and report RMSE only as a reference.

The two models also exhibit different attention patterns for Verubecestat (Figure 2). Both assign relatively high attention to the fluoropyridine substituent extending toward the S3 region, while QGAT additionally emphasizes the central amide region, the iminothiadiazinane dioxide scaffold, and the S1-facing fluorophenyl moiety. These regions correspond to structural features involved in the experimentally determined BACE1 binding mode. In the BACE1-Verubecestat cocrystal structure (PDB: 5HU1), the two hydrogen bond donors of the iminothiadiazinane dioxide core interact with the catalytic residues $\mathrm{Asp^{32}}$ and $\mathrm{Asp^{228}}$, while the amide NH lies within hydrogen bonding distance of the $\mathrm{Gly^{230}}$ carbonyl. The fluoropyridine group occupies the S3 region, with its fluoro substituent extending into the S3 subpocket, whereas the fluorophenyl group projects toward the S1 pocket (Scott et al., 2016). The QGAT attention pattern overlaps with several ligand regions involved in the experimentally determined binding mode. Because the model uses ligand-derived features only, this correspondence should be interpreted as a qualitative alignment of model weighting with known binding-relevant regions rather than as a reconstruction of protein–ligand interactions.

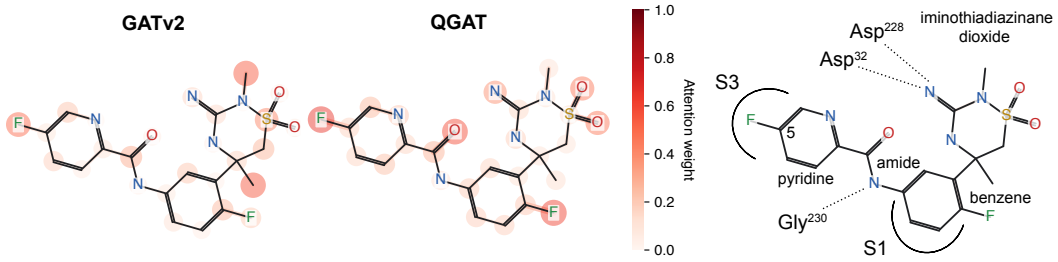


Figure 2: Attention visualization for Verubecestat. The right panel highlights the structural groups and binding interactions discussed in the text, with dashed lines indicating hydrogen bonds. QGAT attention scores are transformed for visualization to enhance contrast.

Attention weights describe how neighboring information is weighted during message aggregation, but they do not indicate whether a feature contributes positively or negatively to the final prediction. We therefore complement the attention analysis with IG. As shown in Figure 3, QGAT assigns a positive contribution to the 5-fluoropyridine substituent, whereas GATv2 assigns a negative contribution. This direction is consistent with the reported SAR showing that substitution at the pyridyl 5-position improves BACE1 affinity relative to the corresponding unsubstituted analogue **28** (Scott et al., 2016). Both models assign positive contributions to the pyridine nitrogen, consistent with its role in forming an intramolecular hydrogen bond with the amide NH that helps lock the preferred picolinamide conformation.

The attribution analysis also reveals limitations. Both models assign a non-positive contribution to the fluorophenyl fluoride, despite matched analogues demonstrating an approximately fivefold potency improvement from this substitution. Thus, the Verubecestat analysis identifies several structural modifications for which QGAT attribution is directionally consistent with reported SAR and differs from GATv2, while also revealing features that neither model captures correctly.

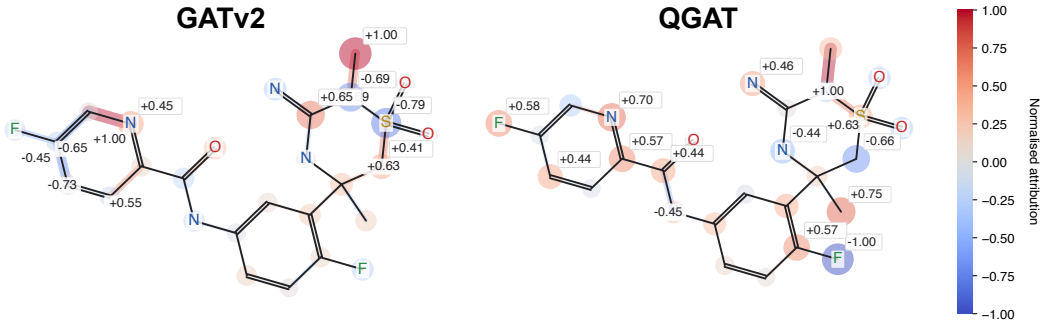


Figure 3: Integrated gradients attribution for Verubecestat.

Activity cliff analysis. Finally, we examine local structural changes associated with activity cliffs (ACs). We define an AC pair as two structurally similar compounds (Stumpfe et al., 2014; 2019) with a potency difference of $|\Delta pK_i| > 1.0$ (Chen et al., 2025). Two pairs in the Verubecestat series satisfy this criterion: compounds **21**–**24** and **30**–**32**. As shown in Appendix Figures 13 and 14, QGAT exhibits a broader redistribution of IG attribution around the modified pyridyl-amide region for compounds **21**–**24** than GATv2. For compounds **30**-**32**, both models respond to the benzene-to-pyridine replacement, while QGAT again shows a broader redistribution across the modified region. These observations are qualitative and are used to characterize differences in attribution response rather than to establish a general advantage in AC sensitivity.

6 Discussion and Conclusion

Our results establish edge-aware variational quantum attention as a viable and experimentally competitive mechanism for molecular graph learning. QGAT achieves performance comparable to the matched GATv2 baseline, with the most consistent gain observed on BBBP. Beyond performance, QGAT and GATv2 exhibit distinct molecular weighting and attribution behavior in the Verubecestat series. QGAT achieves a higher activity ranking correlation and assigns attribution directions to several structural modifications that are consistent with reported SAR, while neither model captures all experimentally observed effects. Together, these findings show that replacing the classical scorer with a VQC can alter prediction behavior beyond aggregate accuracy, although they do not establish universally superior chemical interpretability.

The performance of QGAT also depends on the variational circuit design. No single ansatz performs best across all tasks, and increasing the number of variational layers does not consistently improve predictive performance. These results indicate that circuit architecture is an important modeling choice for molecular quantum attention, rather than a design factor that consistently yields performance improvements.

Several limitations remain. First, the present framework is ligand-based and does not explicitly incorporate protein or binding site information, limiting its ability to model interaction specific affinity prediction. Second, the quantum circuits are evaluated using classical simulation rather than real quantum hardware. Third, the current chemical interpretation is based primarily on qualitative attribution analyses and a single case study. The present evidence therefore does not establish generally superior chemical interpretability, and broader quantitative evaluation is needed. Future work should explore protein-aware extensions and quantitative protocols for evaluating chemically relevant attention and attribution.

AI USE STATEMENT

In this work, we used generative AI tools to provide feedback on research methodology and experimental design, assist with the interpretation of experimental results, and support software code development and revision. We also used generative AI tools to assist with literature summarization, manuscript drafting and editing, paper organization, reference formatting, and brainstorming of titles and presentation strategies. All AI-assisted outputs were reviewed by the authors for correctness, consistency with the experimental results, and alignment with the cited literature. The authors take full responsibility for the final content of this work, including all text, claims, analyses, and artifacts produced with the assistance of generative AI.

REPRODUCIBILITY STATEMENT

We provide the dataset construction and preprocessing procedures, scaffold-splitting protocol, molecular graph features, training settings, and model hyperparameters in Appendix A. The variational circuit architectures, circuit depth experiments, and simulated noise evaluation protocols are described in Appendix B. Additional substructure attribution and Verubecestat analyses are provided in Appendices C and D. The implementation will be released publicly upon publication.

REFERENCES

Amira Abbas, David Sutter, Christa Zoufal, Aurelien Lucchi, Alessio Figalli, and Stefan Woerner. The power of quantum neural networks. *Nat. Comput. Sci.*, 1(6):403–409, Jun 2021. ISSN 2662-8457. doi: 10.1038/s43588-021-00084-1.

Guy W. Bemis and Mark A. Murcko. The properties of known drugs. 1. molecular frameworks. *J. Med. Chem.*, 39(15):2887–2893, July 1996. doi: 10.1021/jm9602928.

Marcello Benedetti, Erika Lloyd, Stefan Sack, and Mattia Fiorentini. Parameterized quantum circuits as machine learning models. *Quantum science and technology*, 4(4):043001, 2019. doi: 10.1088/2058-9565/ab4eb5.

N. Berdigaliyev and M. Aljofan. An overview of drug discovery and development. *Future Medicinal Chemistry*, 12:939–947, 2020.

Jacob Biamonte, Peter Wittek, Nicola Pancotti, Patrick Rebentrost, Nathan Wiebe, and Seth Lloyd. Quantum machine learning. *Nature*, 549(7671):195–202, Sep 2017. doi: 10.1038/nature23474.

Shaked Brody, Uri Alon, and Eran Yahav. How attentive are graph attention networks? In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=F72ximsx7C1.

Marco Cerezo, Guillaume Verdon, Hsin-Yuan Huang, Lukasz Cincio, and Patrick J Coles. Challenges and opportunities in quantum machine learning. *Nature Computational Science*, 2(9): 567–576, Mar 2022. doi: 10.1038/s42254-022-00425-7.

Xu Chen, Dazhou Yu, Liang Zhao, and Fang Liu. Aces-gnn: can graph neural network learn to explain activity cliffs? *Digital Discovery*, 4(8):2062–2074, 2025. doi: 10.1039/d5dd00012b.

John S. Delaney. Esol: Estimating aqueous solubility directly from molecular structure. *J. Chem. Inf. Comput. Sci.*, 44(3):1000–1005, 2004. doi: 10.1021/ci034243x.

L. Domingo, M. Djukic, C. Johnson, and F. Borondo. Binding affinity predictions with hybrid quantum-classical convolutional neural networks. *Scientific Reports*, 13:17951, Oct 2023. doi: 10.1038/s41598-023-45269-y.

Lina Dong, Yulin Li, Dandan Liu, Ye Ji, Bo Hu, Shuai Shi, Fangyan Zhang, Junjie Hu, Kun Qian, Xianmin Jin, and Binju Wang. Prediction of protein-ligand binding affinity by a hybrid quantum-classical deep learning algorithm. *Advanced Quantum Technologies*, 6(9):2300107, 2023. doi: https://doi.org/10.1002/qute.202300107.

Yuxuan Du, Min-Hsiu Hsieh, Tongliang Liu, Shan You, and Dacheng Tao. Learnability of quantum neural networks. *PRX Quantum*, 2(4):040337, Nov 2021. doi: 10.1103/PRXQuantum.2.040337.

Arthur M. Faria, Mehdi Djellabi, Igor O. Sokolov, and Savvas Varsamopoulos. Quantum graph attention networks: Trainable quantum encoders for inductive graph learning. *arXiv preprint arXiv:2509.11390*, 2025.

Anna Gaulton, Louisa J. Bellis, A. Patricia Bento, Jon Chambers, Mark Davies, Anne Hersey, Yvonne Light, Shaun McGlinchey, David Michalovich, Bissan Al-Lazikani, and John P. Overington. Chembl: a large-scale bioactivity database for drug discovery. *Nucleic Acids Research*, 40:D1100–D1107, 2012. doi: 10.1093/nar/gkr777.

Justin Gilmer, Samuel S. Schoenholz, Patrick F. Riley, Oriol Vinyals, and George E. Dahl. Neural message passing for quantum chemistry. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1263–1272. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/gilmer17a.html.

Kuang-Cheng Hsu, Pei-Hua Wang, Bo-Han Su, and Yufeng Jane Tseng. A robust and interpretable graph neural network-based protocol for predicting p-glycoprotein substrates. *Briefings in Bioinformatics*, 26(4):bbaf392, Jul 2025a. doi: 10.1093/bib/bbaf392.

Yu-Chao Hsu, Kuan-Cheng Chen, Tai-Yue Li, and Nan-Yow Chen. Quantum adaptive excitation network with variational quantum circuits for channel attention. In *2025 IEEE International Conference on Quantum Computing and Engineering (QCE)*, volume 02, pp. 344–349, 2025b. doi: 10.1109/QCE65121.2025.10347.

Ye Hu, Dagmar Stumpfe, and Jürgen Bajorath. Computational exploration of molecular scaffolds in medicinal chemistry. *Journal of Medicinal Chemistry*, 59:4062–4076, May 2016. doi: 10.1021/acs.jmedchem.5b01746.

S.G. Jeong, K.H. Moon, and W.J Hwang. Hybrid quantum neural networks for efficient protein-ligand binding affinity prediction. *EPJ Quantum Technology*, 12:120, 2025. doi: 10.1140/epjqt/s40507-025-00419-1.

Greg Landrum. RDKit: Open-source cheminformatics software, 2016. Available at https://www.rdkit.org/.

Jing Li, Leyi Wei, Henry H Y Tong, and Quan Zou. Quantum computing applications in drug discovery. *Briefings in Bioinformatics*, 27:bbag274, May 2026. doi: 10.1093/bib/bbag274.

Longlong Li, Yipeng Zhang, Guanghui Wang, and Kelin Xia. Kolmogorov–arnold graph neural networks for molecular property prediction. *Nature Machine Intelligence*, 7:1346–1354, 2025. doi: 10.1038/s42256-025-01087-7.

Min Lu, Lei Du, Ziwei Cui, Yiming Zhao, Qipeng Yan, Jianyu Zhao, Ye Li, Menghan Dou, Qingchun Wang, Yu-Chun Wu, and Guo-Ping Guo. Quantum-embedded graph neural network architecture for molecular property prediction. *J. Chem. Inf. Model.*, 65(15):8057–8065, Aug 2025. doi: 10.1021/acs.jcim.5c01019.

Hehuan Ma, Yatao Bian, Yu Rong, Wenbing Huang, Tingyang Xu, Weiyang Xie, Geyan Ye, and Junzhou Huang. Cross-dependent graph neural networks for molecular property prediction. *Bioinformatics*, 38(7):2003–2009, March 2022. doi: 10.1093/bioinformatics/btac039.

Rainer Machauer, Rainer Lueoend, Konstanze Hurth, Siem J. Veenstra, Heinrich Rueeger, Markus Voegtle, Marina Tintelnot-Blomley, Jean-Michel Rondeau, Laura H. Jacobson, Grit Laue, Karen Beltz, and Ulf Neumann. Discovery of umibecestat (cnp520): A potent, selective, and efficacious β-secretase (bace1) inhibitor for the prevention of alzheimer's disease. *J. Med. Chem.*, 64(20):15262–15279, Oct 2021. doi: 10.1021/acs.jmedchem.1c01300.

Ines Filipa Martins, Ana L. Teixeira, Luis Pinheiro, and Andre O. Falcao. A bayesian approach to in silico blood-brain barrier penetration modeling. *J. Chem. Inf. Model.*, 52(6):1686–1697, 2012. doi: 10.1021/ci300124c.

Ezek Mathew, Kyle A. Emmitte, and Jin Liu. I-gat: Interpretable graph attention networks for ligand optimization. *ACS Omega*, 10(30):32968–32986, 2025.

David L. McKinzie, Leonard L. Winneroski, Steven J. Green, Erik J. Hembre, Jon A. Erickson, Brian A. Willis, Scott A. Monk, Christopher D. Aluise, Thomas K. Baker, Jose E. Lopez, Jörg Hendle, James P. Beck, Richard A. Brier, Leonard N. Boggs, Anthony R. Borders, Patrick J. Cocke, Pablo Garcia-Losada, Stephen L. Lowe, Brian M. Mathes, Patrick C. May, Warren J. Porter, Stephanie L. Stout, David E. Timm, Brian M. Watson, Zhixiang Yang, and Dustin J. Mergott. Discovery and early clinical development of ly3202626, a low-dose, cns-penetrant bace inhibitor. *J. Med. Chem.*, 64(12):8076–8100, Jun 2021. doi: 10.1021/acs.jmedchem.1c00489.

D. L. Mobley and J. P. Guthrie. Freesolv: a database of experimental and calculated hydration free energies, with input files. *J Comput Aided Mol Des*, 28(7):711–720, 2014. doi: 10.1007/s10822-014-9747-x.

Michael A. Nielsen and Isaac L. Chuang. *Quantum Computation and Quantum Information: 10th Anniversary Edition*. Cambridge University Press, Cambridge, 2010.

An Ning, Tai Yue Li, and Nan Yow Chen. Quantum graph attention network: A novel quantum multi-head attention mechanism for graph learning. *arXiv preprint arXiv:2508.17630*, 2025. URL https://doi.org/10.48550/arXiv.2508.17630.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. In *NIPS-W*, 2017.

Snehal Raj, Brian Coyle, Léo Monbroussou, André J. Ferreira-Martins, Renato M. S. Farias, and Elham Kashefi. Scalable message-passing quantum graph neural networks in the weisfeiler–leman hierarchy. *arXiv preprint*, 2026.

Ladislav Rampášek, Mikhail Galkin, Vijay Prakash Dwivedi, Anh Tuan Luu, Guy Wolf, and Dominique Beaini. Recipe for a general, powerful, scalable graph transformer. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Asel Sagingalieva, Mohammad Kordzanganeh, Nurbolat Kenbayev, Daria Kosichkina, Tatiana Tomashuk, and Alexey Melnikov. Hybrid quantum neural network for drug response prediction. *Cancers*, 15:2705, 2023. doi: 10.3390/cancers15102705.

R. Santagati, Aspuru-Guzik, R. A., Babbush, et al. Drug design on quantum computers. *Nat. Phys.*, 20:549–557, Mar 2024. doi: 10.1038/s41567-024-02411-5.

Frédéric Sauvage, Pranav Kalidindi, Frederic Rapp, and Martín Larocca. Quantum machine learning models for graphs. *arXiv preprint arXiv:2607.00698*, 2026.

Maria Schuld, Ryan Sweke, and Johannes Jakob Meyer. Effect of data encoding on the expressive power of variational quantum-machine-learning models. *Phys. Rev. A*, 103(3):032430, Mar 2021. doi: 10.1103/PhysRevA.103.032430.

Jack D. Scott et al. Discovery of the 3-imino-1,2,4-thiadiazinane 1,1-dioxide derivative verubecestat (mk-8931)–a β-site amyloid precursor protein cleaving enzyme 1 inhibitor for the treatment of alzheimer's disease. *J. Med. Chem.*, 59:10435–10450, Dec 2016. doi: 10.1021/acs.jmedchem.6b00307.

Paul San Sebastian Sein, Theodor Iosif, Tilen G. Limbäck-Stokin, Kin Ian Lo, and Yidong Liao. Quantum graph convolutional networks: Implementation and trainability analysis. *arXiv*, Sep 2026. URL https://doi.org/10.48550/arXiv.2609.19983.

Sukin Sim, Peter D. Johnson, and Alán Aspuru-Guzik. Expressibility and entangling capability of parameterized quantum circuits for hybrid quantum-classical algorithms. *Advanced Quantum Technologies*, 2(12):1900070, 2019. doi: 10.1002/qute.201900070.

Dagmar Stumpfe, Ye Hu, Dilyana Dimova, and Jürgen Bajorath. Recent progress in understanding activity cliffs and their utility in medicinal chemistry. *J. Med. Chem.*, 57(1):18–28, Jan 2014. doi: 10.1021/jm401120g.

Dagmar Stumpfe, Huabin Hu, and Jürgen Bajorath. Evolving concept of activity cliffs. *ACS Omega*, 4(11):14360–14368, Sep 2019. doi: 10.1021/acsomega.9b02221.

Govindan Subramanian, Bharath Ramsundar, Vijay Pande, and Rajiah Aldrin Denny. Computational modeling of β-secretase 1 (bace-1) inhibitors using ligand based approaches. *J. Chem. Inf. Model.*, 56:1936–1949, 2016. doi: 10.1021/acs.jcim.6b00290.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 3319–3328. PMLR, 2017.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2017.

Conghao Wang, Gaurav Asok Kumar, and Jagath C. Rajapakse. Drug discovery and mechanism prediction with explainable graph neural networks. *Scientific Reports*, 15(179), Jan 2025. doi: 10.1038/s41598-024-83090-3.

Kai Wang and Zhen-Yang Peng. Quantum purification for amplitude damping noise. *arXiv*, Sep 2025. URL https://doi.org/10.48550/arXiv.2509.05709.

Zhenqin Wu, Bharath Ramsundar, Evan N. Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S. Pappu, Karl Leswing, and Vijay Pande. Moleculenet: a benchmark for molecular machine learning. *Chem. Sci.*, 9(2):513–530, 2018. doi: 10.1039/C7SC02664A.

Zhenxing Wu, Jike Wang, Hongyan Du, et al. Chemistry-intuitive explanation of graph neural networks for molecular property prediction with substructure masking. *Nat Commun*, 14(2585), May 2023. doi: 10.1038/s41467-023-38192-3.

Tian Xie and Jeffrey C. Grossman. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Phys. Rev. Lett.*, 120:145301, Apr 2018. doi: 10.1103/PhysRevLett.120.145301.

Zhaoping Xiong, Dingyan Wang, Xiaohong Liu, Feisheng Zhong, Xiaozhe Wan, Xutong Li, Zhaojun Li, Xiaomin Luo, Kaixian Chen, Hualiang Jiang, and Mingyue Zheng. Pushing the boundaries of molecular representation for drug discovery with the graph attention mechanism. *J. Med. Chem.*, 63:8749–8760, 2020.

Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? In *International Conference on Learning Representations*, 2019. URL `https://openreview.net/forum?id=ryGs6iA5Km`.

Kevin Yang, Kyle Swanson, Wengong Jin, Connor Coley, Philipp Eiden, Hua Gao, Angel Guzman-Perez, Timothy Hopper, Brian Kelley, Miriam Mathea, Volker Settels Andrew Palmer, Tommi Jaakkola, Klavs Jensen, and Regina Barzilay. Analyzing learned molecular representations for property prediction. *Journal of Chemical Information and Modeling*, 59(8):3370–3388, Aug 2019. doi: 10.1021/acs.jcim.9b00237.

Appendix

TABLE OF CONTENTS

A Experimental Details **16**

A.1 Datasets and Preprocessing 16

A.2 Scaffold Splitting 16

A.3 Graph Representation Method 17

A.4 Hyperparameter and Training Details 17

B Variational Quantum Circuit Details **18**

B.1 Circuit Architectures 18

B.2 Effect of Variational Depth 19

B.3 Post-training Robustness to Simulated Quantum Noise 20

C Substructure Attribution Analysis **21**

C.1 Substructure Mask Explanation 21

C.2 Complete Attribution Distributions 23

D Additional Analysis of External Medicinal Chemistry Series **23**

D.1 Integrated Gradients for the Verubecestat Series 23

D.2 Activity Cliff Analysis for the Verubecestat Series 24

D.3 Additional External Series: Umibecestat and LY3202626 24

A Experimental Details

A.1 Datasets and Preprocessing

We evaluate the models on five molecular property and bioactivity prediction datasets spanning classification and regression tasks. Table 4 summarizes the dataset size, prediction tasks, and target properties used in this study.

BACE BACE (Subramanian et al., 2016) dataset measures the binding results for a set of inhibitors of β-secretase 1 (BACE-1) and is gathered in MoleculeNet (Wu et al., 2018).

BBBP The Blood-Brain Barrier Penetration (BBBP) (Martins et al., 2012) is a binary classification benchmark containing measurements of whether a small molecule can penetrate the central nervous system.

ESOL The ESOL (Delaney, 2004) dataset contains aqueous solubility values (solubility in mols per liter) for small organic molecules in log scale.

FreeSolv FreeSolv (Mobley & Guthrie, 2014) contains experimental and calculated hydration free energy values for small molecules in water.

ChEMBL-BACE We construct a BACE1 bioactivity regression dataset from ChEMBL (Gaulton et al., 2012). We retain measurements associated with a single human BACE1 target and use reported IC_{50} values. IC_{50} values are transformed to pIC_{50}. The resulting activity values are subsequently normalized for model training.

Table 4: Summary for the molecular property and bioactivity datasets used in this study

Dataset	Task	No. Tasks	Molecules	Target
BACE	Classification	1	1,513	Binding results for a set of inhibitors of human BACE1
BBBP	Classification	1	2,039	Permeability properties of the BBB penetration
FreeSolv	Regression	1	642	Hydration free energy
ESOL	Regression	1	1,128	Water solubility
ChEMBL-BACE	Regression	1	4,580	Binding results for a set of inhibitors of human BACE1

A.2 Scaffold Splitting

A molecular scaffold describes the core structural framework of a molecule. In the Bemis–Murcko scaffolds (Bemis & Murcko, 1996), the scaffold is composed primarily of ring systems and the linkers connecting them, while terminal side chains are removed. Consequently, molecules that differ in peripheral substituents can still share the same scaffold if they are built around the same ring-and-linker framework.

Scaffolds are particularly relevant in medicinal chemistry because compounds within the same chemical series are often generated by modifying substituents around a conserved core (Hu et al., 2016). As a result, structurally related analogues can share both a scaffold and similar physicochemical or biological properties. Under a random split, close analogues may therefore appear in both the training and test set, allowing a model to be evaluated on molecules that are structurally similar to those observed during training (Yang et al., 2019).

To provide a more stringent evaluation of structural generalization, we use scaffold splitting. Each molecule is first assigned a Bemis–Murcko scaffold using RDKit (Landrum, 2016), and molecules sharing the same scaffold are grouped together. Scaffold groups are then assigned to the training, validation, and test partitions without splitting molecules from the same scaffold across partitions (Wu et al., 2018; Yang et al., 2019). Thus, the evaluation emphasizes generalization to molecular cores that are not represented in the corresponding training set (Wu et al., 2018).

In our experiments, we first reserve 10% of the molecules as a fixed scaffold based test set. The remaining molecules are partitioned into five folds, producing an 72% training and 18 % validation set for each fold.

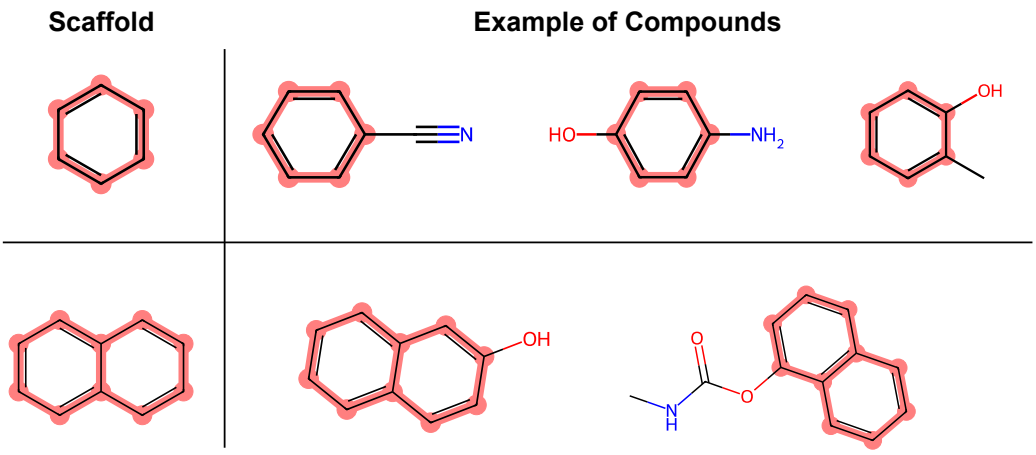


Figure 4: Examples of molecular scaffolds and compounds sharing the corresponding core structures. Highlighted atoms indicate the Bemis–Murcko scaffold, while peripheral substituents vary among compounds within the same scaffold group.

A.3 Graph Representation Method

We represent each molecule as a bidirectional graph $G = (V, E)$, where atoms, including explicit hydrogen atoms, correspond to nodes (V) and each covalent bond is represented by two parallel edge types (E), namely a chemical bond edge and a charge–distance edge. Each atom is encoded by a 92-dimensional vector following the atom representation used in crystal graph convolutional neural network (Xie & Grossman, 2018), incorporating atomic number, group, period, electronegativity and covalent radius. Both edge types are represented by 21-dimensional feature vectors. For a chemical bond edge, the first 15 dimensions encode the direction, bond type, bond distance, and ring membership, while the remaining six dimensions are set to zero. For a charge–distance edge, the first 15 dimensions are set to zero, whereas the final six dimensions contain the partial charges q_i and q_j, their pairwise product $q_i q_j$, and the inverse-distance features r_{ij}^{-1}, r_{ij}^{-6}, and r_{ij}^{-12}. The charge and inverse distance terms provide information about the local electronic environment and three-dimensional geometry of the bonded atom pair. The inverse-distance terms are used as nonlinear geometric descriptors rather than as explicit estimates of interaction energies. Detailed descriptions of the feature vectors are summarized in Table 5.

Table 5: Feature Types used for molecular graph construction

Entity	Feature Type	Description	Type	Size
Atom	CGCNN	Atomic number, radius, and electronegativity	One-Hot	92
Chemical bond edge	Bond direction	None, beginwedge, begindash, etc.	One-Hot	7
	Bond type	Single, Double, Triple, or Aromatic	One-Hot	4
	Bond length	r_{ij} and r_{ij}^2	Float	2
	Ring membership	Whether the bond belongs to a ring	One-Hot	2
Charge-distance edge	Partial charges	Atomic charges in molecule	Float	3
	Distance between atoms	Distance terms $(r_{ij}^{-1}, r_{ij}^{-6}, r_{ij}^{-12})$	Float	3

A.4 Hyperparameter and Training Details

All models were implemented in PyTorch (Paszke et al., 2017) and trained on NVIDIA H200 GPUs. Unless otherwise specified, we use the Adam optimizer with a learning rate of 1×10^{-3} with weight decay 1×10^{-4}, and dropout rate of 0.05. All models are trained for a maximum of 500 epochs. For classification tasks, we use binary cross-entropy loss, whereas regression tasks are optimized using mean squared error loss. Validation performance is used only for checkpoint selection within each trained configuration. The implementation will be released publicly upon publication.

Table 6 summarizes the model configurations used for each dataset. The hidden dimension, number of attention heads, number of layers, learning rate, weight decay, dropout rate, maximum training epochs, and batch size are kept identical between QGAT and the corresponding GATv2 baseline for each task. The quantum module introduces computational and memory overhead during classical circuit simulation, particularly because attention scores must be evaluated for molecular edges and multiple attention heads. We therefore restrict the search space to relatively compact hidden dimensions and four attention heads rather than performing an exhaustive architecture search over substantially larger models.

Table 6: GAT and QGAT hyperparameters for different datasets.

Dataset	Model	Ansatz	Hidden	Heads	Layers	N_b	Qubits	Batch	Params.
BACE	QGAT	1	16	4	2	1	6	64	13,153
BACE	GATv2	–	16	4	2	–	–	64	13,249
BBBP	QGAT	3	32	4	4	3	7	64	73,385
BBBP	GATv2	–	32	4	4	–	–	64	73,729
FreeSolv	QGAT	1	32	4	4	1	7	64	73,273
FreeSolv	GATv2	–	32	4	4	–	–	64	73,729
ESOL	QGAT	6	16	4	2	2	6	64	13,201
ESOL	GATv2	–	16	4	2	–	–	64	13,377
ChEMBL BACE	QGAT	4	32	4	4	2	7	256	73,345
ChEMBL BACE	GATv2	–	32	4	4	–	–	256	73,729

Unless otherwise stated, quantum expectation values during training are computed using exact state-vector simulation. Noisy quantum simulations are treated separately in the corresponding robustness experiments rather than during the standard training procedure.

B Variational Quantum Circuit Details

B.1 Circuit Architectures

We consider a constrained family of shallow parameterized quantum circuits. Following the hardware-efficient circuit family studied by Sim et al. (2019), we construct each ansatz from parameterized single-qubit rotation gates and CNOT entangling gates.

For a rotation around axis $a \in \{x, y, z\}$, the corresponding single-qubit gate is

$$R_a(\theta) = \exp(-i\frac{\theta}{2}\sigma_a) \tag{12}$$

where σ_a denotes the Pauli matrix associated with the rotation axis.

Because rotations about different axes generally do not commute, their ordering changes the resulting variational unitary. For example,

$$R_Y(\theta_1)R_Z(\theta_2) \neq R_Z(\theta_2)R_Y(\theta_1) \tag{13}$$

Accordingly, changing the rotation sequence modifies the local parameterization even when the circuit uses the same gate set and a similar number of trainable parameters. Figure 5 summarizes the six circuit templates considered in our experiments. The candidates share the same general design principle of shallow single-qubit rotations followed by CNOT gates, while varying the rotation sequence and entangling connectivity. Previous studies have shown that the choice of parameterized gates, entangling structure, and circuit depth can substantially affect properties such as expressibility and entangling capability (Sim et al., 2019); however, these circuit structures do not by themselves determine predictive performance for a given learning task. We therefore evaluate the six candidates directly on the molecular prediction tasks rather than assuming a universally preferred circuit structure. As reported in Table 2, no single ansatz performs best across all datasets, indicating that performance is sensitive to the circuit parameterization and task.

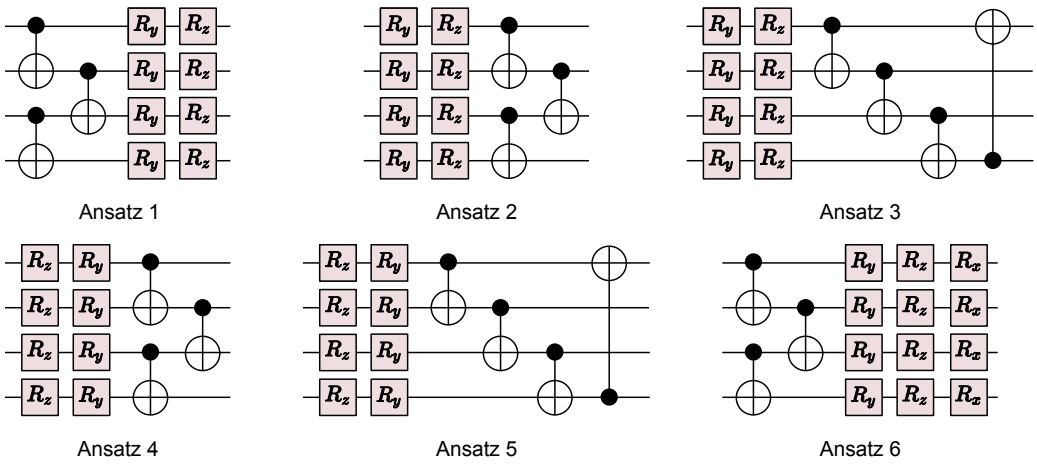


Figure 5: Six ansatzes evaluated in this study. For illustration, the circuits are shown with $n_q = 4$ qubits and $N_b = 1$ block. All ansatzes use shallow hardware-efficient circuits composed of parameterized single-qubit rotations and CNOT entangling gates, while differing in rotation patterns and entangling connectivity.

B.2 Effect of Variational Depth

We examine whether increasing the depth of the VQC consistently improves performance. Table 7 reports the results obtained by varying the number of variational layers while keeping the remaining model hyperparameters fixed for each task.

The optimal circuit depth is task-dependent, and increasing the number of blocks does not lead to performance improvements. BACE and FreeSolv achieve their best performance with a single block. In contrast, ESOL and ChEMBL-BACE perform best with two blocks, while BBBP reaches its highest ROC-AUC with three blocks.

Deeper circuits generally provide limited additional benefits beyond these task-specific optima. For example, BACE performance decreases from 0.8283 with one block to approximately 0.80 with two and three blocks, while FreeSolv RMSE gradually increases as the circuit depth grows. A similar degradation is observed for BBBP and ESOL at four blocks after reaching their respective optima at shallower depths.

These results suggest that the benefit of additional circuit depth is not universal. Relatively shallow circuits are sufficient for the evaluated molecular property prediction tasks, whereas increasing circuit depth may introduce additional optimization complexity without providing a corresponding improvement in performance.

Table 7: Effect of variational circuit depth on QGAT performance. Results are reported as mean $\pm$ standard deviation across five folds. The best result for each task is shown in red. Unreported results are denoted by "–" due to higher computational cost.

N_b	BACE	BBBP	FreeSolv	ESOL	ChEMBL-BACE	
	AUC ↑	AUC ↑	RMSE ↓	RMSE ↓	Spearman ↑	RMSE ↓
1	**0.8283 ± 0.0133**	0.7019 ± 0.0164	**1.7400 ± 0.2033**	0.9263 ± 0.0275	0.7265 ± 0.0198	0.9018 ± 0.0130
2	0.8023 ± 0.0235	0.7125 ± 0.0095	1.7697 ± 0.1798	**0.9103 ± 0.0317**	**0.7357 ± 0.0147**	**0.8800 ± 0.0188**
3	0.8018 ± 0.0126	**0.7157 ± 0.0095**	1.8498 ± 0.2731	0.9218 ± 0.0538	–	–
4	0.8119 ± 0.0118	0.6919 ± 0.0153	1.8816 ± 0.2340	0.9377 ± 0.0645	–	–

B.3 Post-training Robustness to Simulated Quantum Noise

We evaluate the post-training sensitivity of QGAT to two simulated noise channels: depolarizing noise and amplitude damping. All models are trained under noiseless simulation and subsequently evaluated with noise injected into the VQC during inference.

Depolarizing noise. For a single-qubit gate acting on qubit q, we apply the depolarizing channel (Nielsen & Chuang, 2010)

$$\mathcal{E}_{\mathrm{dep},1Q}(\rho) = (1 - p_{1Q})\rho + \frac{p_{1Q}}{3} \sum_{P \in \{X,Y,Z\}} P_q \rho P_q \tag{14}$$

where p_{1Q} denotes the total probability of a non-identity single-qubit Pauli probability. For a CNOT acting on control qubit c and target qubit t, we use a joint two-qubit depolarizing channel:

$$\mathcal{E}^{(c,t)}_{\mathrm{dep},2Q}(\rho) = (1 - p_{2Q})\rho + \frac{p_{2Q}}{15} \sum_{P \in \mathcal{P}_2 \setminus \{I \otimes I\}} P \rho P^\dagger, \tag{15}$$

where $\mathcal{P}_2 = \{I, X, Y, Z\}^{\otimes 2}$ and p_{2Q} denotes the total probability assigned to the 15 non-identity two-qubit Pauli operators.

Amplitude damping.

Amplitude damping describes irreversible relaxation from the excited state $|1\rangle$ toward the ground state $|0\rangle$. For a single qubit, the channel is defined as (Nielsen & Chuang, 2010; Wang & Peng, 2025)

$$\mathcal{E}_{\mathrm{AD}}(\rho) = K_0 \rho K_0^\dagger + K_1 \rho K_1^\dagger \tag{16}$$

with Kraus operators

$$K_0 = \begin{bmatrix} 1 & 0 \\ 0 & \sqrt{1 - p_{\mathrm{AD}}} \end{bmatrix}, \qquad K_1 = \begin{bmatrix} 0 & \sqrt{p_{\mathrm{AD}}} \\ 0 & 0 \end{bmatrix} \tag{17}$$

where $p_{\mathrm{AD}} \in [0, 1]$ denotes the damping probability. After each single-qubit rotation, the amplitude damping channel is applied to the corresponding qubit.

For a CNOT acting on qubits c and t, amplitude damping is applied independently to both participating qubits using the same damping strength p_{AD}:

$$\mathcal{E}^{(c,t)}_{p_{\mathrm{AD}}}(\rho) = \sum_{i,j \in \{0,1\}} (K_i^{(c)} \otimes K_j^{(t)}) \rho (K_i^{(c)} \otimes K_j^{(t)})^\dagger. \tag{18}$$

Equivalently, this local two-qubit channel can be represented by the four product Kraus operators $M_{ij} = K_i \otimes K_j$, with $i, j \in \{0, 1\}$. This model assumes independent relaxation of the two qubits and does not include correlated decay.

Noisy circuit simulation. Amplitude encoding is performed without noise and initially prepares a state vector. For noisy inference, the state is converted to the density matrix. Noise is applied after every gate. Specifically, each gate rotation is followed by the corresponding single-qubit noise channel, whereas each CNOT is followed by either the joint two-qubit depolarizing channel in Eq. 15 or amplitude-damping channels in Eq. 18. The resulting evolution can be summarized as

$$\rho \xrightarrow{U} U\rho U^\dagger \xrightarrow{\mathcal{E}} \mathcal{E}(U\rho U^\dagger). \tag{19}$$

The noise channels are evaluated directly through density matrix evolution by summing the corresponding Pauli or Kraus branches.

Evaluation protocol. We evaluate the channel strengths $p \in \{0.001, 0.01, 0.1, 0.2\}$ where p denotes p_{dep} for depolarizing noise and p_{AD} for amplitude damping. For the reported experiments, the same strength is assigned to single- and two-qubit operations within each noise model.

We focus on post-training inference sensitivity rather than training under noisy conditions. Noisy training would require differentiating through density matrix evolution after each quantum operation, substantially increasing the computational and memory cost relative to the noiseless state vector training used in the main experiments.

As shown in Figures 6 and 7, QGAT remains largely stable under weak noise. At channel strengths up to 0.01, the performance remains close to the noiseless setting across all tasks. In contrast, at larger noise strengths, the degradation becomes task dependent. BACE and FreeSolv exhibit gradual performance deterioration as the noise strength increases, whereas ESOL and ChEMBL-BACE show a substantially sharper degradation at $p \geq 0.1$. Amplitude damping generally produces slightly smaller performance losses than depolarizing noise at the larger error rates.

BBBP is comparatively insensitive to both noise models, its ROC-AUC changes only marginally even at the largest noise strength. This observation suggests that the prediction on BBBP is relatively stable under perturbations to the circuit, although the mechanism underlying this behavior remains unclear.

Under matched numerical channel parameters, amplitude damping generally produces smaller performance changes than depolarizing noise at the larger tested values. Overall, QGAT shows limited performance degradation under weak simulated noise, whereas stronger noise produces larger and task-dependent degradation.

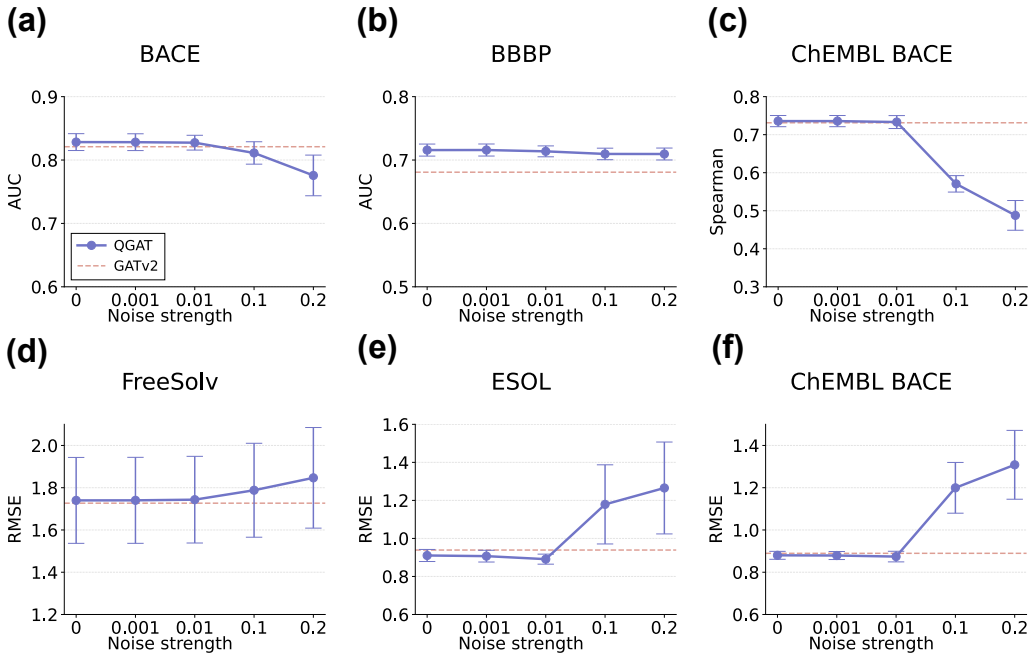


Figure 6: Performance of QGAT at inference stage under depolarizing noise. Models are trained under noiseless simulation and evaluated with depolarizing noise injecting into the VQC at different channel strengths. Error bars indicate variation across five scaffold folds.

C Substructure Attribution Analysis

C.1 Substructure Mask Explanation

Substructure Mask Explanation (SME) (Wu et al., 2023) is a perturbation-based interpretation method that estimates the contribution of a molecular substructure by comparing the model predic-

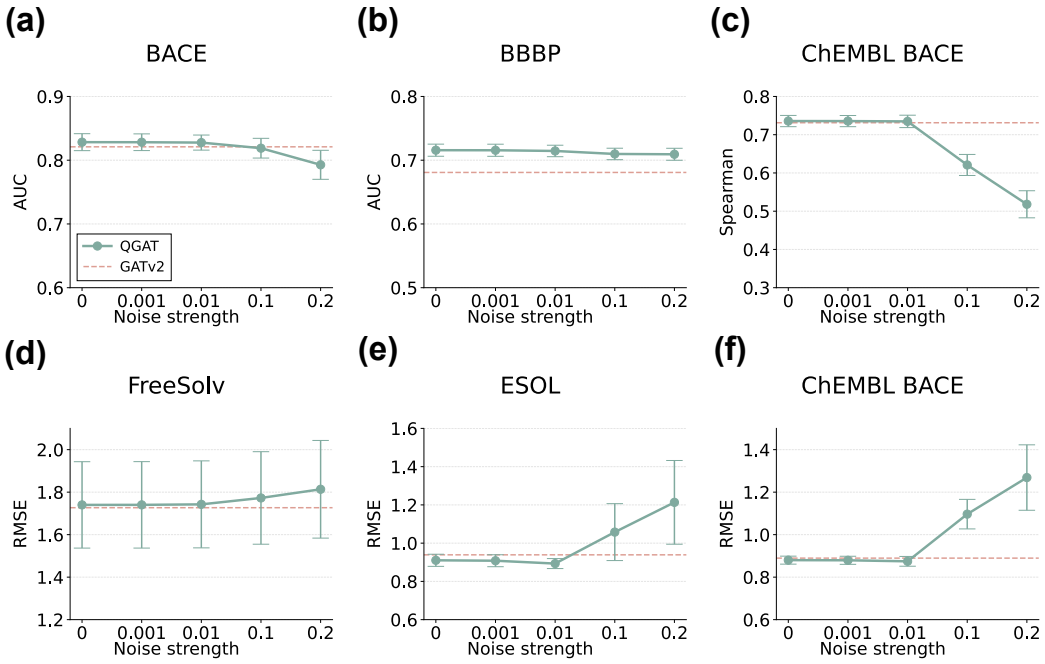


Figure 7: Performance of QGAT at inference stage under amplitude damping noise. Models are trained under noiseless simulation and evaluated with amplitude damping injecting into the VQC at different channel strengths. Error bars indicate variation across five scaffold folds.

tion before and after masking that substructure. By aggregating the attribution scores over functional groups occurrences across the entire dataset, we examine how specific chemical groups influence the predictions of each model.

Following the original SME formulation, masking is performed after message passing rather than by deleting atoms or bonds from the input graph. The complete molecular graph is first processed by the GNN to obtain the final node representations, after which the contributions of nodes belonging to the selected substructure are suppressed during graph readout. We adapt this procedure to the pooling operators used by our models.

$$\{h_v\}_{v=1}^{N} = f_\theta^{\mathrm{GNN}}(G) \tag{20}$$

where h_v is the learned representation of atom v after the final message passing layers. The graph structure and message passing process are therefore kept unchanged, and masking is applied only at the readout layer.

For a functional-group occurrence $S \subseteq V$, we define the binary node mask

$$m_v^{(S)} = \begin{cases} 0, & v \in S, \\ 1, & v \notin S \end{cases} \tag{21}$$

Let $\mathcal{R}$ denote the pooling used by the corresponding model configuration. The unmasked molecular representation is:

$$g = \mathcal{R}\left(\{h_v\}_{v \in V}\right) \tag{22}$$

The corresponding masked molecular representation is

$$g^{(S)} = \mathcal{R}\left(\{m_v^{(S)} h_v\}_{v \in V};\ S\right) \tag{23}$$

where $\mathcal{R}_S$ applies the same readout rule while suppressing the contribution of nodes in S. For mean pooling, masked node embeddings are set to zero while the denominator remains the original number of nodes N. For max pooling, masked nodes are excluded from the element-wise maximum. Depending on the model configuration, the mean and max pooled representations are either concatenated or combined by element-wise addition.

The original and masked predictions are then computed using the same prediction head:

$$\hat{y} = f_\theta^{\text{head}}(g), \qquad \hat{y}^{(S)} = f_\theta^{\text{head}}\left(g^{(S)}\right). \tag{24}$$

The attribution of substructure occurrence S is defined as

$$A_S = \hat{y} - \hat{y}^{(S)}. \tag{25}$$

A positive attribution indicates that masking the substructure decreases the model output, whereas a negative attribution indicates that masking the substructure increases it. The sign therefore reflects the direction of the model output and should not be interpreted as an intrinsically favorable or unfavorable chemical effect.

C.2 Complete Attribution Distributions

To complement the representative functional group analysis reported in the main text, we provide the complete SME attribution distributions for all evaluated datasets. For each functional group, the distribution contains attribution scores from all matched occurrences in the corresponding dataset, thereby retaining variation across different molecular contexts rather than reporting only the dataset-level mean.

Figures 8–12 summarize the full attribution distributions for GATv2 and QGAT. Across datasets, many functional groups show similar attribution directions between the two models, while differences are more apparent in attribution magnitude and distributional spread for selected groups. Several functional groups also exhibit broad distributions spanning both positive and negative values, indicating that their contributions are context dependent rather than determined solely by functional-group identity. These distributions are therefore used to characterize model specific attribution patterns across molecular contexts, rather than to assign universal favorable or unfavorable effects to individual chemical groups.

D Additional Analysis of External Medicinal Chemistry Series

D.1 Integrated Gradients for the Verubecestat Series

Integrated Gradients (IG) (Sundararajan et al., 2017; Hsu et al., 2025a; Wang et al., 2025) is a gradient-based attribution method that quantifies the contribution of input features to a model prediction relative to a reference input. Rather than evaluating the gradient at the observed input, IG accumulates gradients along a straight line path from a baseline x' to the input x.

Let $F(x)$ denote the model output of interest. The IG of the i-th input feature is defined as

$$\text{IG}_i(x) = (x_i - x_i') \int_{\alpha=0}^{1} \frac{\partial F(x' + \alpha(x - x'))}{\partial x_i} d\alpha \tag{26}$$

where x' denotes the baseline input and α parameterizes the interpolation path between x' and x. The resulting attribution reflects the accumulated sensitivity of the model output to feature i along this path.

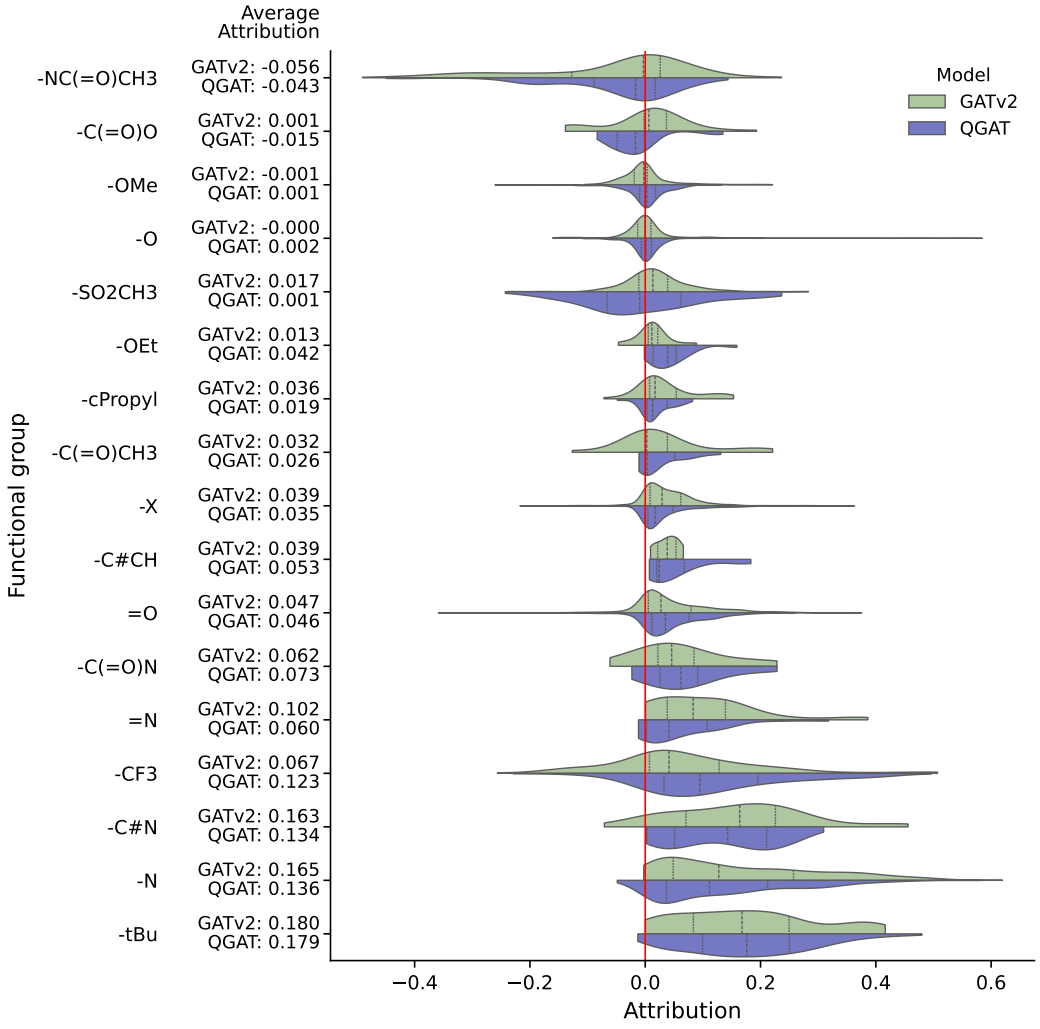


Figure 8: The attributions of different functional groups in the whole BACE dataset. Green colors represent GATv2 model, while blue colors represent QGAT model.

D.2 Activity Cliff Analysis for the Verubecestat Series

We provide the complete integrated gradient visualizations for the two AC pairs identified in the Verubecestat series, compounds **21**–**24** and **30**–**32**. As described in the main text, these pairs satisfy the AC criterion of a potency difference $|\Delta pK_i| > 1.0$ between structurally similar analogues.

Figures 13 and 14 show the IG attribution maps for both GATv2 and QGAT. The visualization is intended to complement the main text analysis by showing the full spatial redistribution of attribution induced by the corresponding structural modifications. Because only two AC pairs are available in this series, we treat these examples as qualitative case studies rather than as a statistical evaluation of AC prediction.

D.3 Additional External Series: Umibecestat and LY3202626

We further evaluate two independent medicinal chemistry case studies, Umibecestat (CNP520) (Machauer et al., 2021) and LY3202626 (McKinzie et al., 2021). Because these sets are small and contain several compounds with high similarity to the training data, we do not treat them as

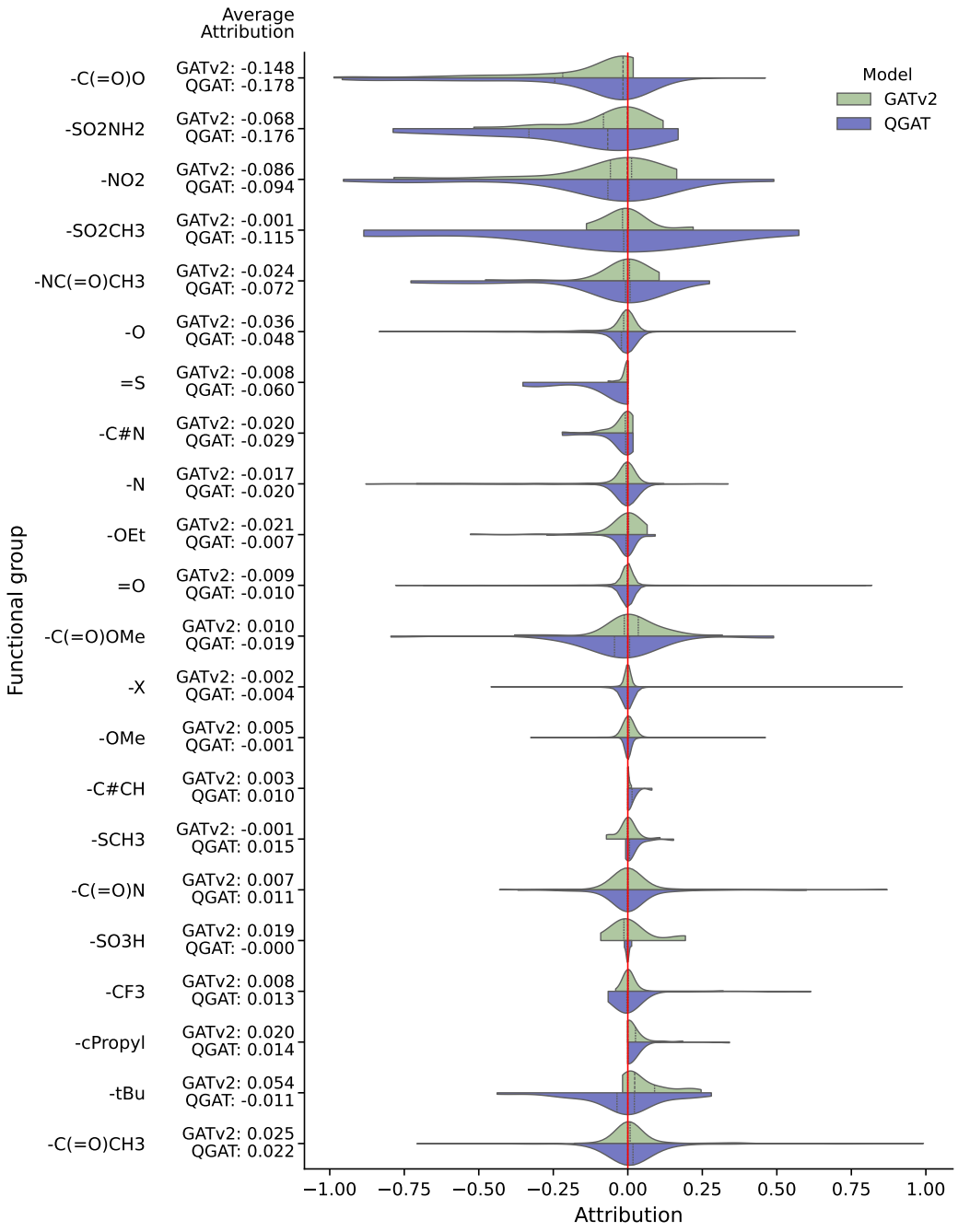


Figure 9: The attributions of different functional groups in the whole BBBP dataset. Green colors represent GATv2 model, while blue colors represent QGAT model.

chemical OOD benchmarks. Instead, they are used to assess whether the ranking behavior observed for Verubecestat generalizes qualitatively across additional analogue series.

For the Umibecestat series, QGAT yields a modestly higher mean Spearman correlation than GATv2 (0.692 vs. 0.655), while RMSE remains comparable. For the LY3202626 series, the two models achieve nearly identical mean Spearman correlation (0.788 vs. 0.784), whereas GATv2 obtains lower RMSE.

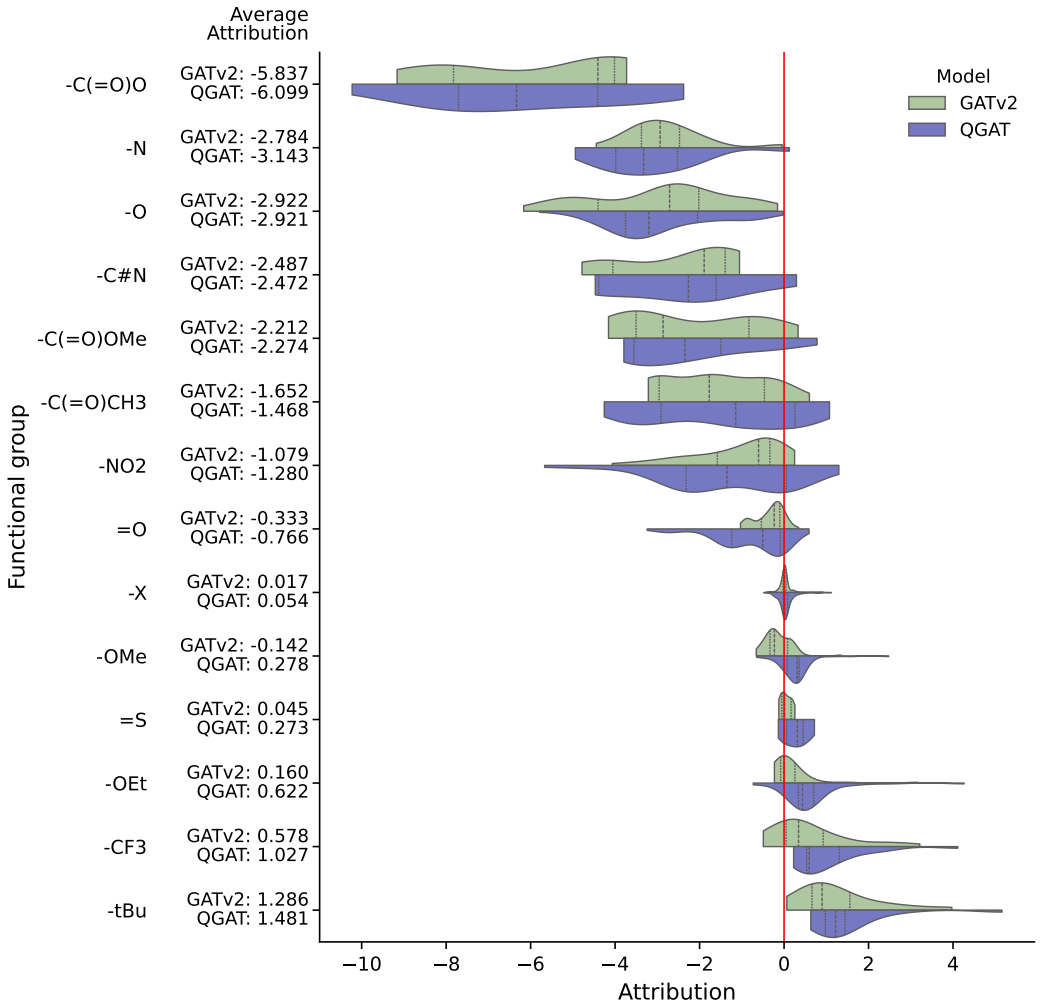


Figure 10: The attributions of different functional groups in the whole FreeSolv dataset. Green colors represent GATv2 model, while blue colors represent QGAT model.

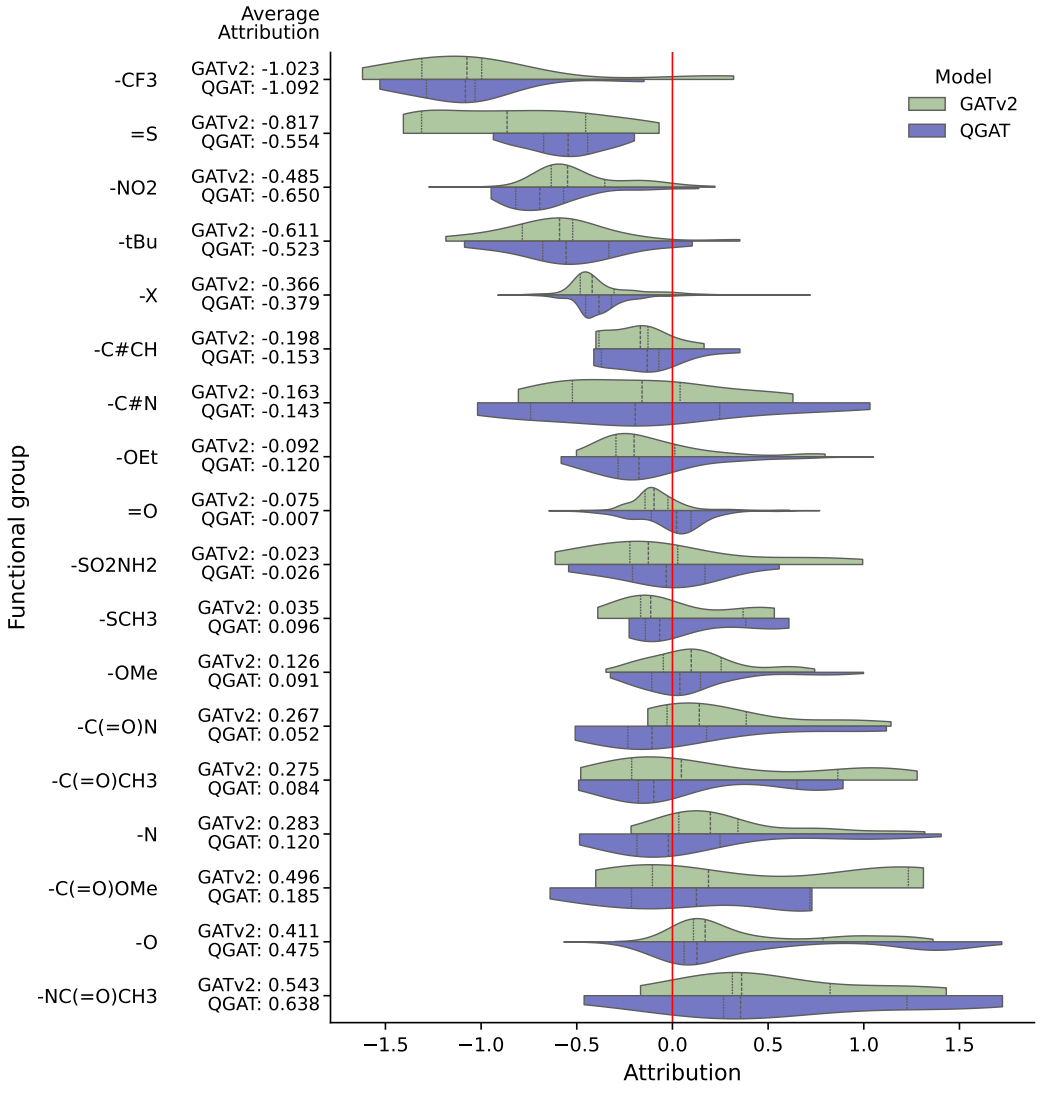


Figure 11: The attributions of different functional groups in the whole ESOL dataset. Green colors represent GATv2 model, while blue colors represent QGAT model.

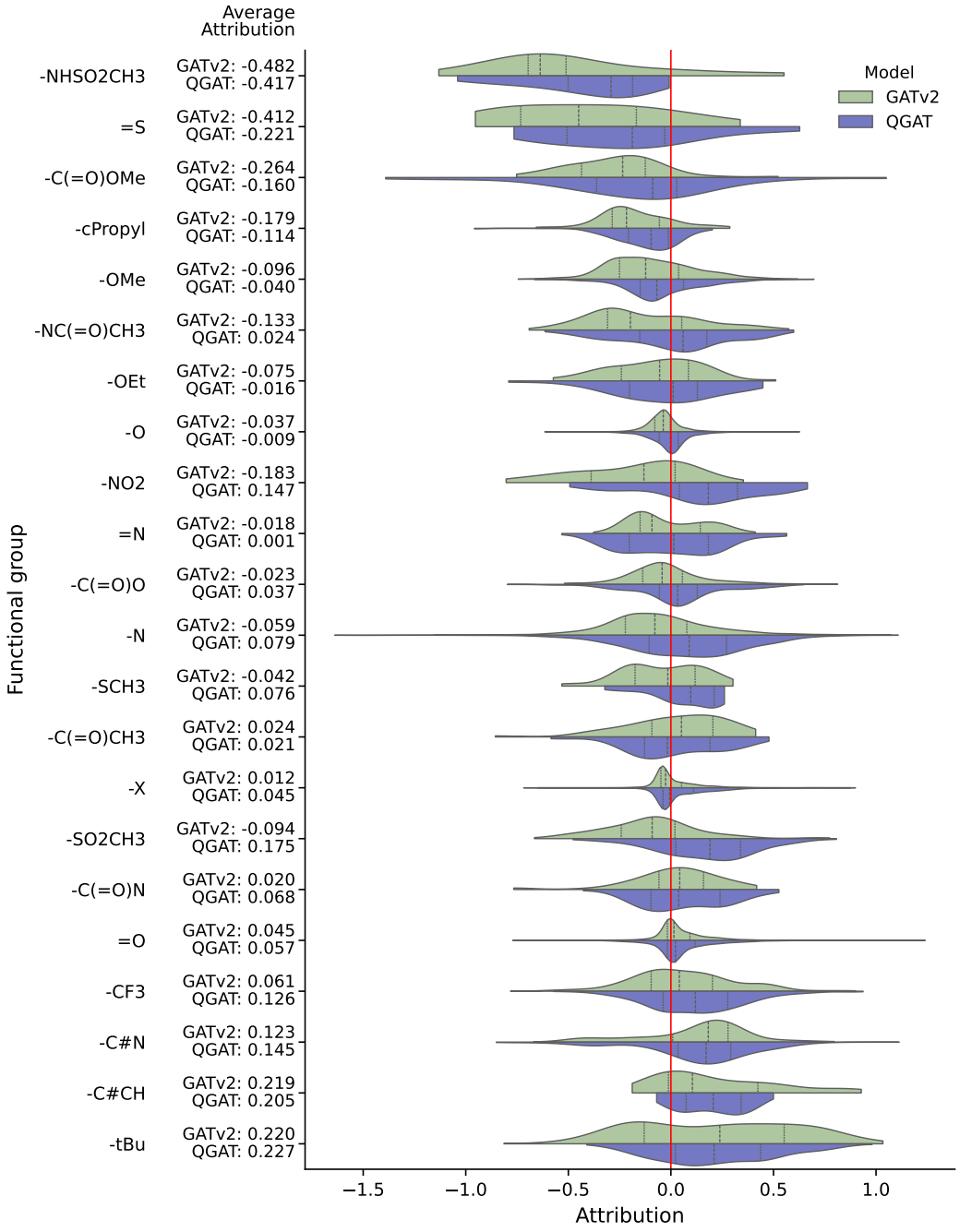


Figure 12: The attributions of different functional groups in the whole ChEMBL-BACE dataset. Green colors represent GATv2 model, while blue colors represent QGAT model.

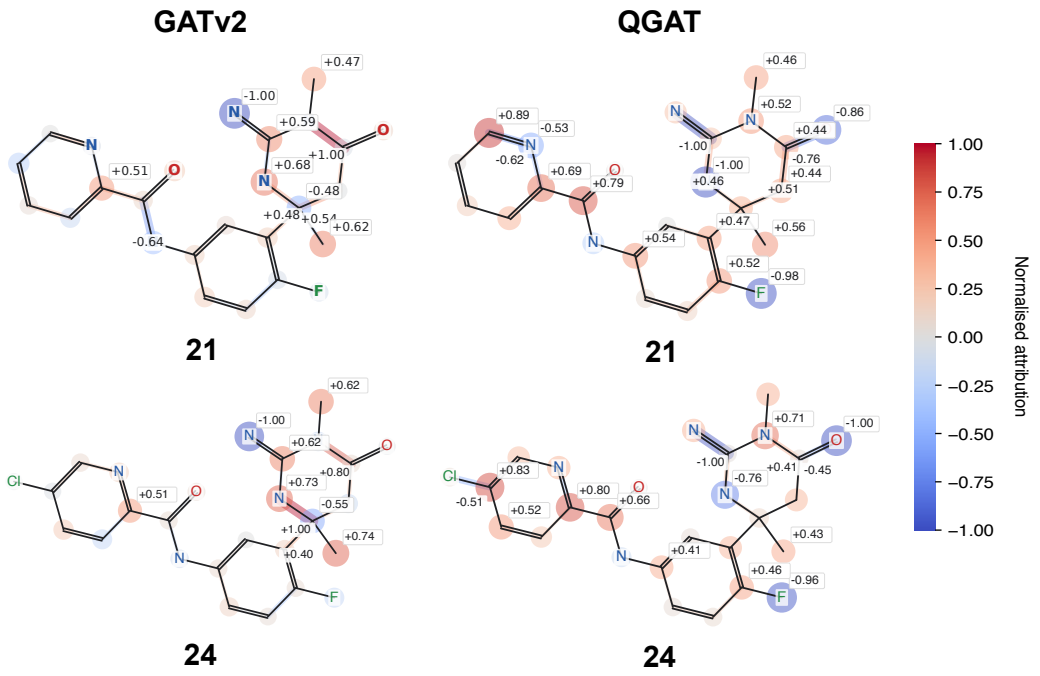


Figure 13: IG attribution maps for the AC pairs identified in the Verubecestat series: compounds **21**–**24**. Attributions are normalized within each molecule for visualization. The figure provides the complete atom level attribution patterns corresponding to the qualitative AC analysis discussed in the main text.

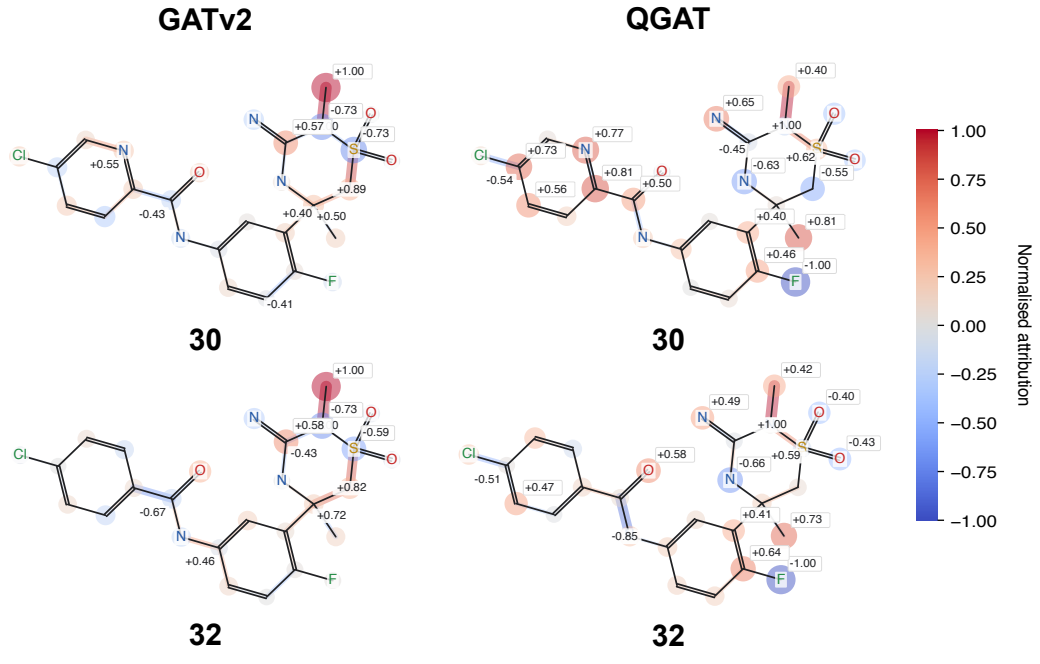


Figure 14: IG attribution maps for the AC pairs identified in the Verubecestat series: compounds **30**–**32**. Attributions are normalized within each molecule for visualization. The figure provides the complete atom level attribution patterns corresponding to the qualitative AC analysis discussed in the main text.