A Summary of the used datasets

All of the public datasets were previously published, either as graph representation learning benchmarks or new datasets for specific graph tasks. The datasets cover a multitude of domains: quantum mechanics, biophysics, bioinformatics, computer vision, social networks, synthetic graphs, and function call graphs. As such, the datasets do not include any personally identifiable information or offensive content. This claim is based both on manual inspection and previous peer-review of the associated publications. Furthermore, no research was performed with human subjects as part of our study. The public benchmarks are listed in Appendix A, Tables 2 and 3. The public and proprietary bio-affinity (high-throughput screening) datasets are listed in Appendix A, Table 1.

Appendix Table 1: Summary of the bio-affinity high-throughput screening datasets.

Availability	Source	Dataset	Size	Splits	Task type	# Tasks
Public	PubChem	AID1949	98,472	No	Regression	1
		AID449762	311,910			
		AID602261	343,811			
Private	Pharmaceutical company		1,013,581	No	Regression	1
			1,482,258			
			1,962,638			

B Experimental design and reporting

Before training each model, a fixed random seed at the beginning of training is set (`pytorch_lightning.seed_everything(0)`). Any convolution-specific hyperparameters (such as the number of attention heads and dropout values for GAT and GATV2) were chosen as reasonable defaults and frozen. As the neural readouts introduce new hyperparameters for the overall GNN architecture, these are also set to reasonable defaults depending on the dataset type and size (Appendix B, Table 4), but are not part of any hyperparameter optimization procedure.

For datasets that did not explicitly provide train, validation, and test sets, we applied an $80\%/10\%/10\%$ split, for five different times on random permutations of the datasets. This procedure was applied on all MoleculeNet datasets, as well as the TUDataset benchmarks. We did not apply any custom splits on MNIST, CIFAR10, and ZINC. The seeds used for the five random permutations are available in **Supplementary File 1** (available on GitHub) and can be directly used with the provided source code. All models are trained with an early stopping mechanism set to a patience value of 30.

Almost all datasets are completely loaded in memory, with the exception of a few complex datasets (such as the REDDIT datasets). If the default batch size of 32 was too large for such datasets, the maximum batch size that allowed the models to be trained on a GPU with 24GB of VRAM was used.

The loss functions used within the deep learning models are set to standard choices, such as mean squared error (MSE) for regression datasets, binary cross-entropy for binary classification datasets and cross-entropy for multi-class datasets. We also generally tried to use the recommended loss functions for the MoleculeNet datasets according to the original publication, in particular using the mean absolute error (MAE) function for certain datasets [29].

The metrics chosen to report the model performance depend on the task type (regression or classification) and the number of tasks (or classes for classification). For binary classification tasks, we report the area under the receiver operating characteristic curve (AUROC) and the Matthews correlation coefficient (MCC), while for classification tasks with more than 2 classes only the MCC is reported. The MCC was recently reported to be a more helpful metric compared to popular choices such as the accuracy or the F_1 score and is considered one of the best summaries of the confusion matrix [14].

For regression tasks, we report the MAE and the coefficient of determination (R^2) for all datasets. R^2 was also recently reported as a regression metric that is more informative compared to the traditional choices of MAE, MSE, RMSE, and others [15].

Appendix Table 2: Summary of all the used benchmarks (datasets), including domain, datasets statistics, random splitting procedures, and source. DeepChem (DC) and PyTorch Geometric (PyG) are released under the MIT license. Several molecular datasets originate from PubChem (public domain). ENZYMES uses CC BY 4.0. reddit_threads and twitch_egos use GPL-3.0. COLORS and TRIANGLES use ECL-2.0. CIFAR10, MNIST, and ZINC use MIT.

Collection	Domain	Dataset	Type	# Tasks	Size	Avg. nodes	Avg. edges	Node attr.	Splits	Source
MoleculeNet	Quantum Mechanics	QM9	Regr.	12	132,480	17.99	37.15	30	5 random	DC
		QM8	Regr.	12	21,747	16.09	32.81	30	5 random	DC
		QM7	Regr.	1	6,834	15.54	30.39	30	5 random	DC
	Physical Chemistry	ESOL	Regr.	1	1,127	13.30	27.38	30	5 random	DC
		FreeSolv	Regr.	1	639	8.76	16.85	30	5 random	DC
		Lipophilicity	Regr.	1	4,200	27.04	59.00	30	5 random	DC
	Biophysics	PCBA	Cls.	128	437,918	25.97	56.22	30	5 random	DC
		HIV	Cls.	1	41,127	25.51	54.94	30	5 random	DC
		BACE	Cls.	1	1,513	34.09	73.72	30	5 random	DC
		BACE	Regr.	1	1,513	34.09	73.72	30	5 random	DC
	Physiology	BBBP	Cls.	1	2,039	24.06	51.91	30	5 random	DC
		SIDER	Cls.	27	1,396	34.36	72.29	30	5 random	DC
TUDataset	Bioinformatics	ENZYMES	Cls.	6	600	32.63	62.14	18	5 random	PyG
		PROTEINS_full	Cls.	2	1,113	39.06	72.82	29	5 random	PyG
	Computer Vision	COIL-DEL	Cls.	100	3,900	21.54	54.24	2	5 random	PyG
		COIL-RAG	Cls.	100	3,900	3.01	3.02	64	5 random	PyG
		Cuneiform	Cls.	30	267	21.27	44.80	3	5 random	PyG
	Social Networks	github_stargazers	Cls.	2	12,725	113.79	234.64	-	5 random	PyG
		IMDB-BINARY	Cls.	2	1,000	19.77	96.53	-	5 random	PyG
		REDDIT-BINARY	Cls.	2	2,000	429.63	497.75	-	5 random	PyG
		REDDIT-MULTI-12K	Cls.	11	11,929	391.41	456.89	-	5 random	PyG
		reddit_threads	Cls.	2	203,088	23.93	24.99	-	5 random	PyG
		twitch_egos	Cls.	2	127,094	29.67	86.59	-	5 random	PyG
		TWITTER-Real-Graph-Partial	Cls.	2	144,033	4.03	4.98	-	5 random	PyG
	Synthetic	COLORS-3	Cls.	11	10,500	61.31	91.03	4	5 random	PyG
		SYNTHETIC	Cls.	2	300	100.00	196.00	1	5 random	PyG
		SYNTHETICnew	Cls.	2	300	100.00	196.25	1	5 random	PyG
		Synthie	Cls.	4	400	95.00	172.93	15	5 random	PyG
		TRIANGLES	Cls.	10	45,000	20.85	32.74	-	5 random	PyG
GNNBenchmarkDataset	Computer Vision	MNIST	Cls.	10	55,000	70.60	564.50	3	Provided	PyG
		CIFAR10	Cls.	10	45,000	117.60	941.20	5	Provided	PyG
ZINC	Drug-like molecules	ZINC	Regr.	1	249,456	23.15	49.80	1	Provided	PyG

Appendix Table 3: Summary of all the used benchmarks (datasets), including domain, datasets statistics, random splitting procedures, and source. PyG, PyTorch Geometric.

Collection	Domain	Dataset	Type	# Tasks	Size	Avg. nodes	Avg. edges	Node attr.	Splits	Source
TUDataset	Small molecules	AIDS	Cls.	2	2,000	15.69	16.20	4	5 random	PyG
		alchemy_full	Regr.	12	202,579	10.10	10.44	3	5 random	PyG
		FRANKENSTEIN	Cls.	2	4,337	16.90	17.88	780	5 random	PyG
		Mutagenicity	Cls.	2	4,337	30.32	30.77	-	5 random	PyG
		MUTAG	Cls.	2	188	17.93	19.79	-	5 random	PyG
		YeastH	Cls.	2	79,601	39.44	40.74	-	5 random	PyG
MalNetTiny	Function call graphs	MalNetTiny	Cls.	5	5,000	1410.31	2859.94	-	5 random	PyG

To simplify reporting the results, for multi-label datasets the appropriate metrics are computed between flattened representations of the predictions and the ground truth values.

C Set Transformer aggregator architectures

Our default Set Transformer architecture, simply referred to as SET TRANSFORMER for the majority of the paper, and ST COMPLEX in Figure 2 and Appendix T, Table 1, uses multiple SAB blocks, where a SAB block is defined as $\text{SAB}(A) = \text{MAB}(A, A)$:

$$\text{ENCODER}(A) \coloneqq \text{SAB}^2(A) \quad \text{and} \quad \text{DECODER}(C) \coloneqq \text{FF}(\text{SAB}^2(\text{PMA}_k(C))) \tag{5}$$

We also evaluate a variation termed ST MINIMAL, with an architecture reduced to:

$$\text{ENCODER}(A) \coloneqq \text{SAB}(A) \quad \text{and} \quad \text{DECODER}(C) \coloneqq \text{FF}(\text{PMA}_k(C)) \tag{6}$$

D Summary of multi-head attention

First of all, we recapitulate the original definition of attention, which works by initializing three learnable matrices, often called Q for *queries*, K for *keys* and V for *values*. For self attention, $Q = K = V$. The attention computation is then defined as

$$\text{ATTENTION}(Q, K, V) \coloneqq \text{softmax}(\frac{QK^\top}{\sqrt{d_k}})V \tag{7}$$

where softmax is defined element-wise as $\text{softmax}(\mathbf{x})_i \coloneqq \frac{\exp(\mathbf{x}_i)}{\sum_j \exp(\mathbf{x}_j)}$

It is often beneficial to perform multiple attention computations concurrently, using different parameters (weights W), called multi-head attention with h heads

$$\text{MULTI-HEAD}(Q, K, V) \coloneqq \text{Concatenate}(\text{head}_1, \ldots, \text{head}_h)W^O \tag{8}$$

$$\text{head}_i \coloneqq \text{ATTENTION}(QW_i^Q, KW_i^K, VW_i^V) \tag{9}$$

E Set Transformer for variable-sized inputs

The SET TRANSFORMER readout supports variable-sized inputs, i.e. it is theoretically possible to start from a non-rectangular (ragged) tensor, such that zero-padding is avoided in the flattened representation. However, as our chosen deep learning library (PyTorch) does not support ragged tensors at the time of writing, our default implementation uses the already defined representations (denoted by H and h in the main text). We did, however, experiment with a computationally inefficient implementation relying on a for-loop instead of batching, without observing notable performance differences (not shown).

F Hyperparameters specific to neural readouts

Appendix Table 4: Summary of the hyperparameters used for neural aggregators for each dataset. The hyperparameter names follow the same notation as introduced in the main text. Dim., dimension.

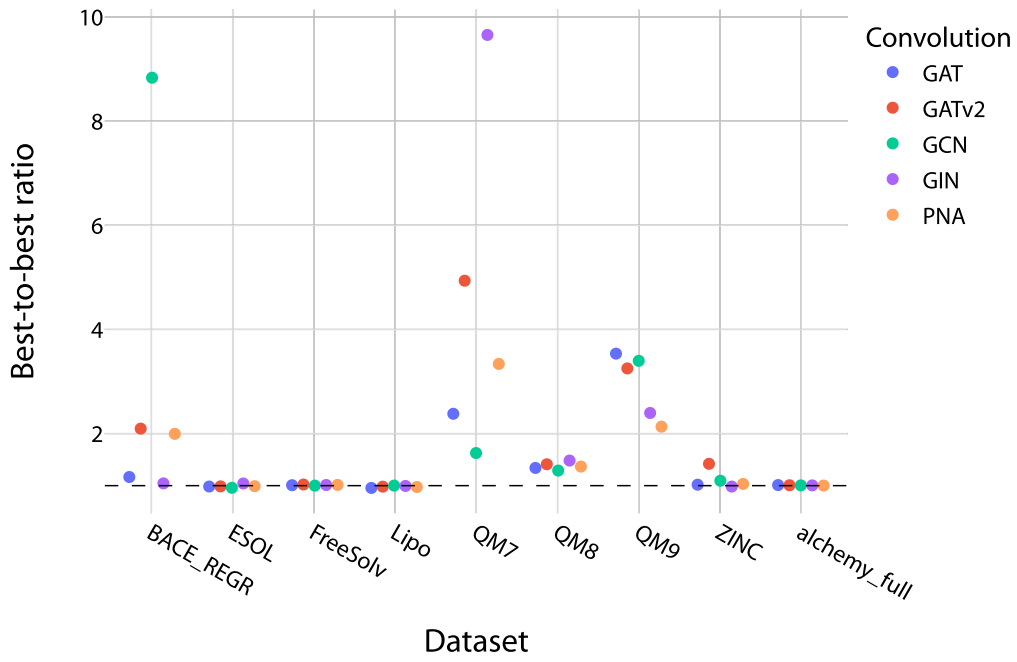
Dataset	MLP		Set Transformer		
	d_1	d_{out}	k	Hidden dim.	n_h
ESOL	64	32	8	32	4
FreeSolv	64	32	8	32	4
Lipo	64	32	8	32	4
BACE_REGR	64	32	8	32	4
BACE_CLS	64	32	8	32	4
BBBP	64	32	8	32	4
SIDER	64	32	8	32	4
QM7	128	64	8	64	8
QM8	128	64	8	64	8
QM9	256	128	8	512	8
PCBA	256	128	8	512	8
HIV	256	128	8	512	4
ENZYMES	256	128	8	64	8
PROTEINS_full	256	128	8	64	8
COIL-DEL	256	128	8	64	8
COIL-RAG	256	128	8	64	8
Cuneiform	256	128	8	64	8
github_stargazers	256	128	8	64	8
IMDB-BINARY	256	128	8	64	8
REDDIT-BINARY	256	128	8	64	8
REDDIT-MULTI-12K	256	128	8	64	8
reddit_threads	256	128	8	64	8
twitch_egos	256	128	8	64	8
TWITTER-Real-Graph-Partial	256	128	8	64	8
COLORS-3	256	128	8	64	8
SYNTHETIC	256	128	8	64	8
SYNTHETICnew	256	128	8	64	8
Synthie	256	128	8	64	8
TRIANGLES	256	128	8	64	8
MNIST	256	128	8	512	4
CIFAR10	256	128	8	512	4
ZINC	256	128	8	512	4

Appendix Table 5: (continued from Appendix F, Table 4) Summary of the hyperparameters used for neural aggregators for each dataset. The hyperparameter names follow the same notation as introduced in the main text. Dim., dimension.

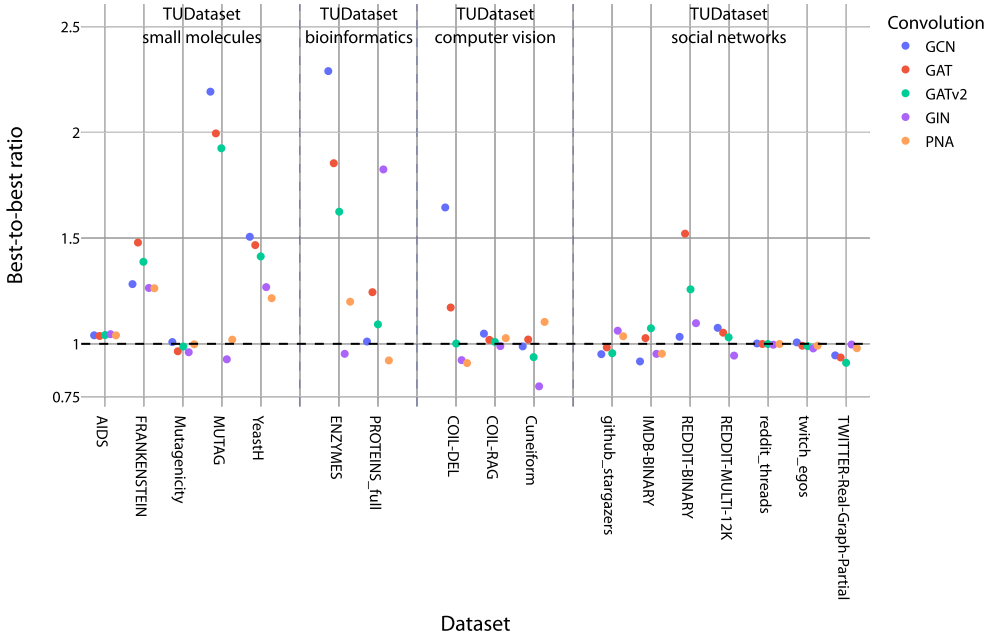
Dataset	MLP		Set Transformer		
	d_1	d_{out}	k	Hidden dim.	n_h
AIDS	256	128	8	256	8
FRANKENSTEIN	256	128	8	256	8
Mutagenicity	256	128	8	256	8
MUTAG	256	128	8	256	8
YeastH	256	128	8	256	8
alchemy_full	256	128	8	256	8
MalNetTiny	256	128	8	192	12

G Best-to-best ratios for all datasets

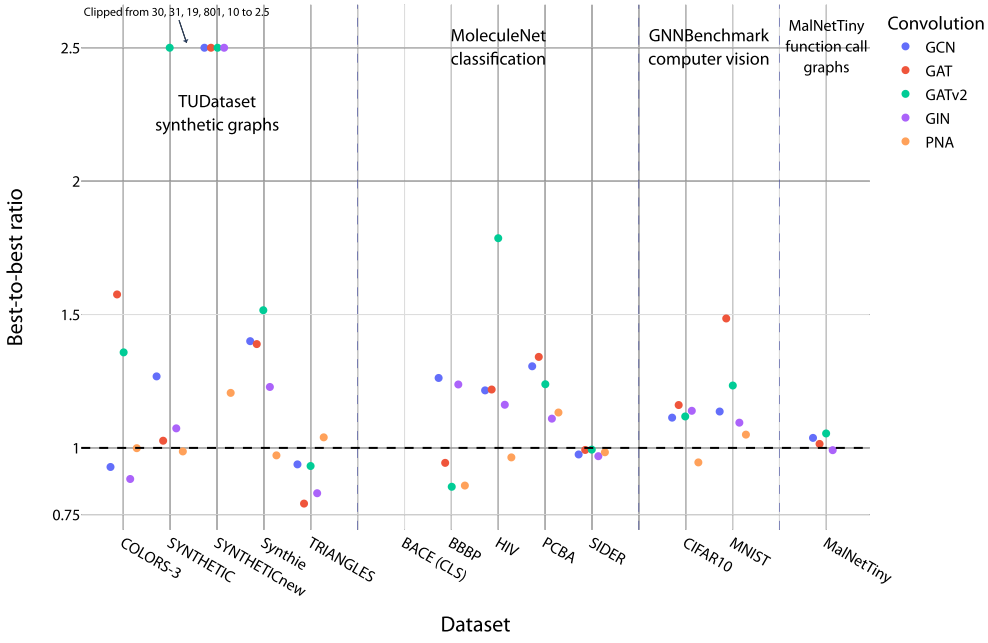
Appendix Figure 1: The performance of the best neural relative to the best standard readout on all regression benchmarks. We use the ratio between the effectiveness scores (R^2), computed by averaging over five random splits of the data. The differences are best appreciated by studying the associated tables (Appendix T, Tables 17, 26 and 31).



Appendix Figure 2: The performance of the best neural relative to the best standard readout on several classification benchmarks. We use the ratio between the effectiveness scores (MCC), computed by averaging over five random splits of the data. The differences are best appreciated by studying the associated tables (Appendix T, Tables 19, 20, 23 to 25 and 30).

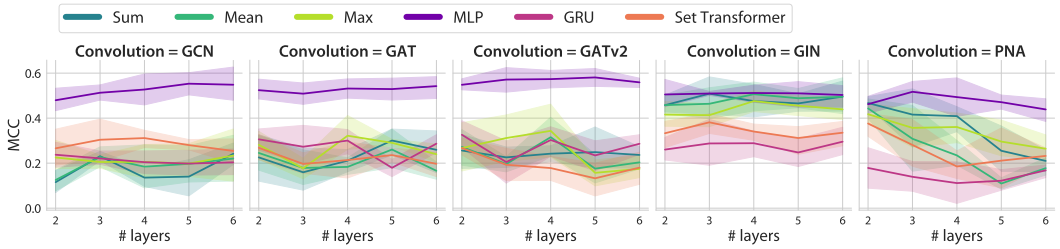


Appendix Figure 3: The performance of the best neural relative to the best standard readout on the rest of the classification benchmarks. We use the ratio between the effectiveness scores (MCC), computed by averaging over five random splits of the data. The differences are best appreciated by studying the associated tables (Appendix T, Tables 18, 21, 22, 28 and 29).



H Deeper GNN models for ENZYMES

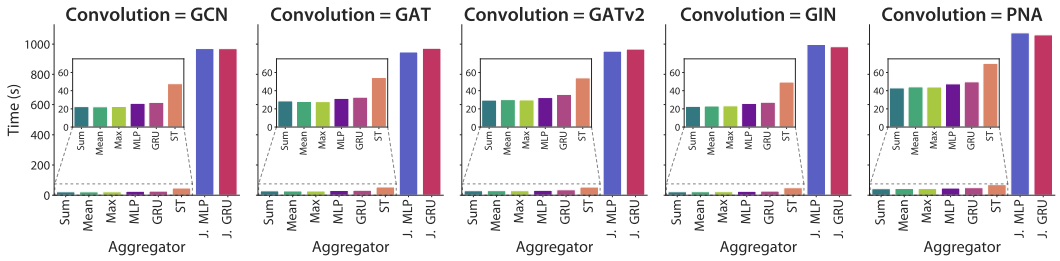
Appendix Figure 4: Increasing the number of neighborhood aggregation iterations or convolutional layers does not produce large differences on ENZYMES. For GCN, the MLP readout improves with deeper networks, whereas most readout are relatively stable regarding the number of layers on GAT, GATV2, and GIN. For PNA, the performance decreases drastically with more than 3 layers for the majority of readouts.



I Time and memory analysis

We benchmarked the elapsed training time and memory utilization on the QM9 dataset (132, 480 data points) for one epoch on a modern high-end GPU (Nvidia RTX 3090, also see Appendix P), averaged from 5 epochs for each model (differences too small to plot error bars). As expected, the JANOSSY variants are not competitive in terms of training time. However, the MLP and GRU aggregators lead to a minimal increase of a few seconds per single epoch compared to the simple classical functions. The full SET TRANSFORMER architecture (also referred to as ST COMPLEX) incurs an increase close to 50%, which is, however, in line with the cost of transitioning from GCN to PNA. The training cost can be minimized by adopting ISAB blocks (trading off performance). The SET TRANSFORMER readout is applicable to large scale datasets, as exemplified by our evaluation which includes SET TRANSFORMER + PNA models with a maximum of 4 GNN layers for datasets of up to 2 million data points (also tested for 5 PNA layers).

Appendix Figure 5: Training times for all aggregators and convolutions on QM9 (seconds). The models are 2-layer GNNs. J., Janossy.

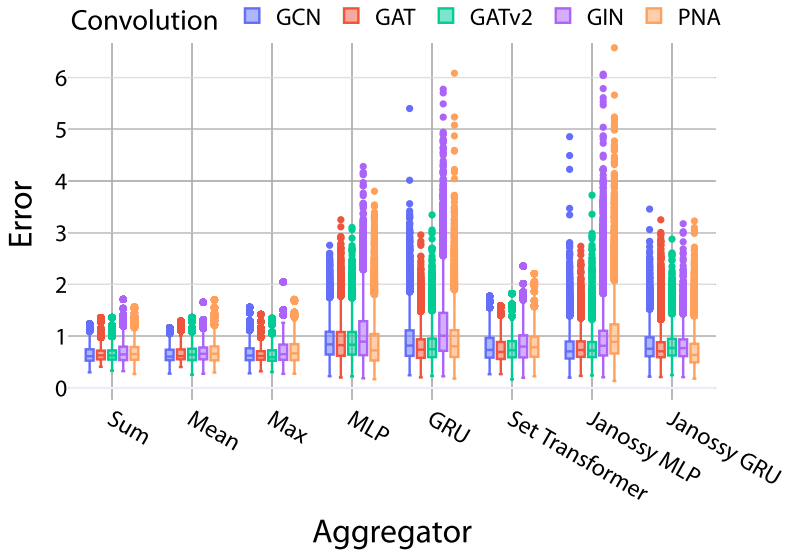


In terms of memory usage, the maximum amount of reserved (higher than allocated) memory as reported by the PyTorch profiler (version 1.10.1) was just under 149MB for the GRU + PNA model, a 27MB increase from the most efficient non-neural aggregator for PNA (mean). It should be noted that it is common for deep learning frameworks such as PyTorch and TensorFlow to automatically reserve or prepare large amounts of memory even if only a portion is allocated during training.

J Robustness to random node permutations of QM9 molecules

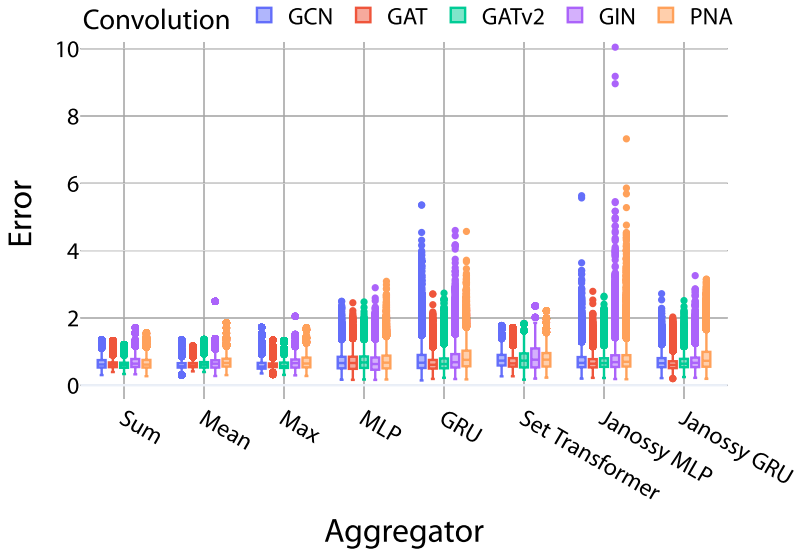
J.1 Random permutations of nodes

Appendix Figure 6: A summary of the error distributions for predictions made on random permutations of 50 randomly selected molecules from the QM9 dataset, presented per-convolution for the arbitrary random permutations strategy.

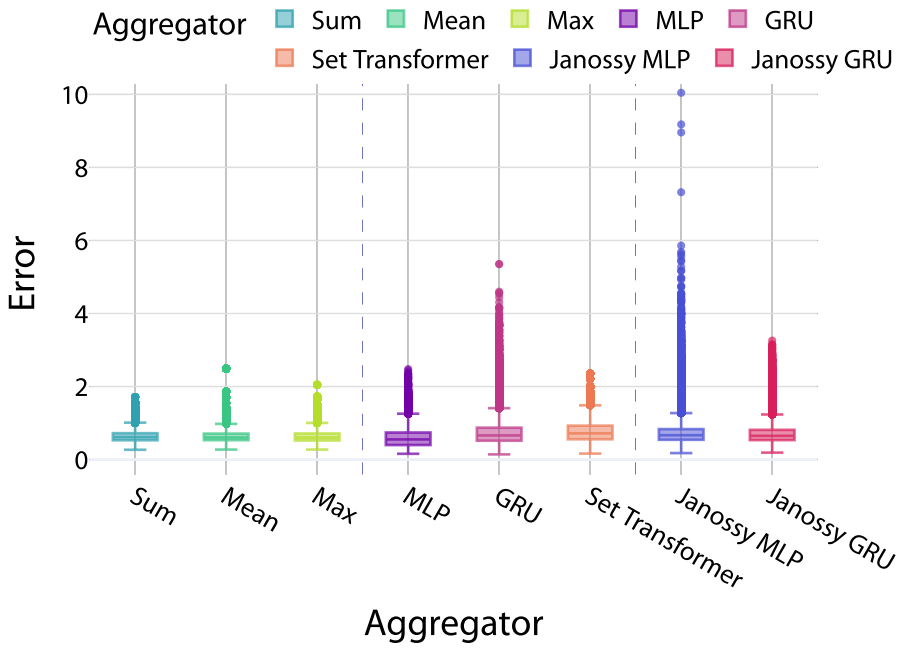


J.2 Random, non-canonical SMILES

Appendix Figure 7: A summary of the error distributions for predictions made on random permutations of 50 randomly selected molecules from the QM9 dataset, presented per-convolution for the random non-canonical SMILES strategy.

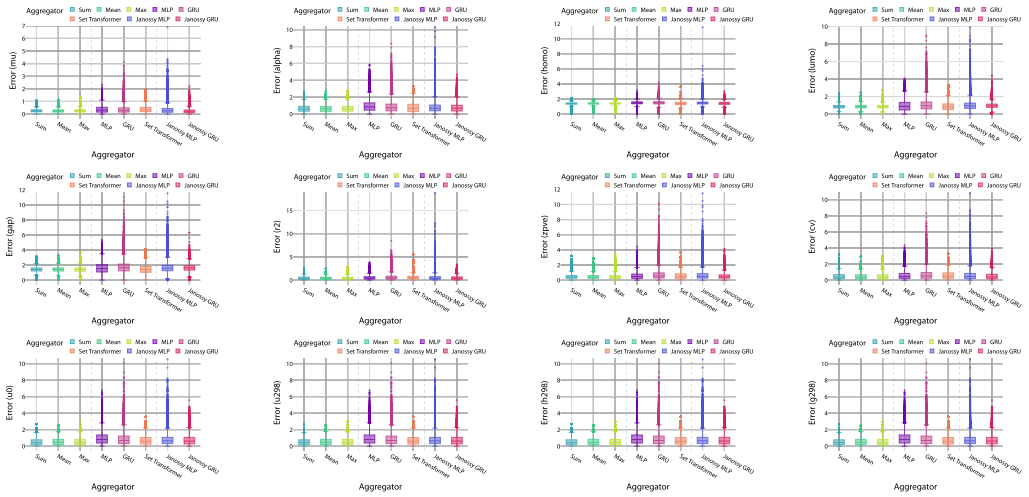


Appendix Figure 8: A summary of the error distributions for predictions made on random permutations of 50 randomly selected molecules from the QM9 dataset for the random non-canonical SMILES strategy, where we selected the top 50 lowest errors for each molecule from the multitude of non-canonical SMILES inputs.



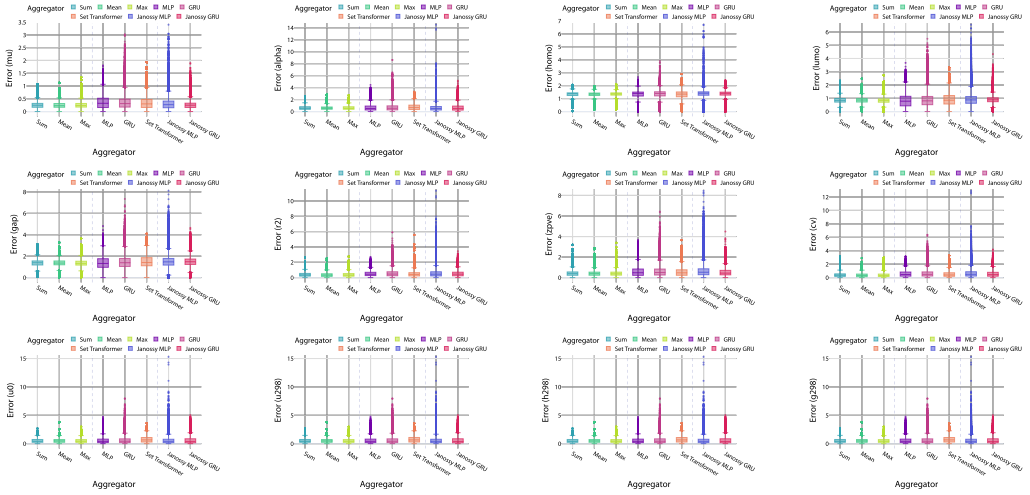
J.3 Error for each QM9 prediction task for random permutations of nodes

Appendix Figure 9: A summary of the error distributions for predictions made on random permutations of 50 randomly selected molecules from the QM9 dataset, presented per QM9 task (12 in total) for the arbitrary random permutations strategy.



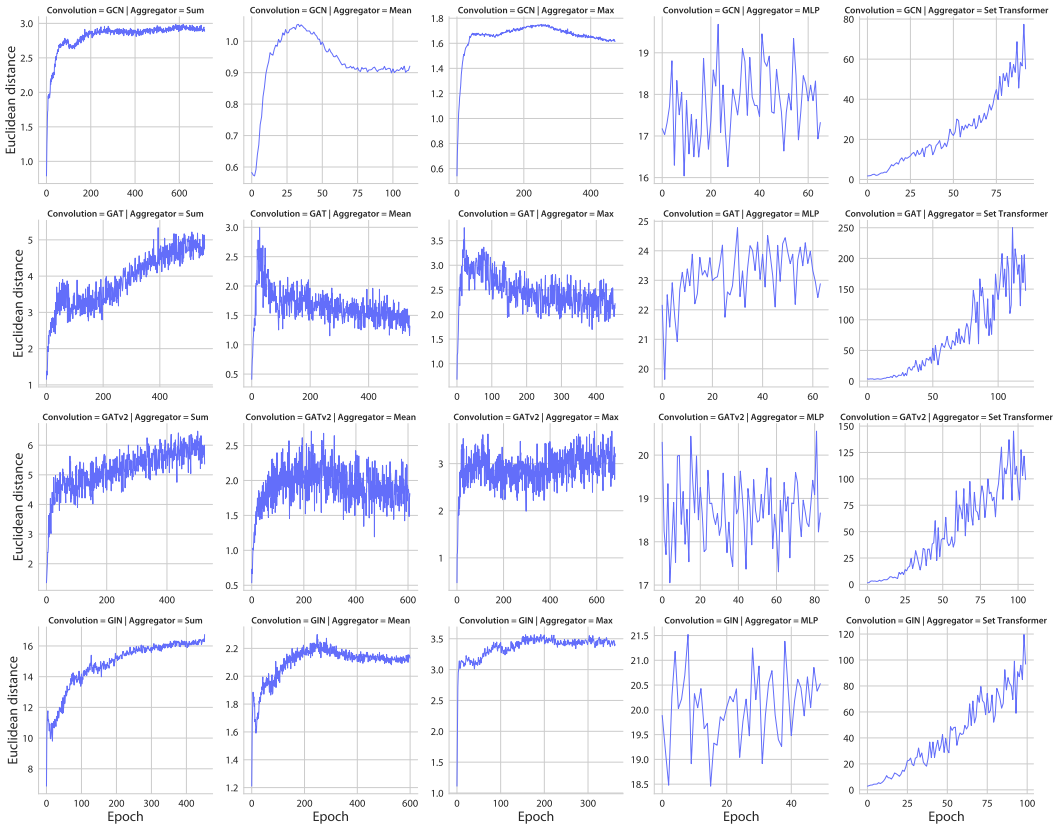
J.4 Error for each QM9 prediction task for random SMILES

Appendix Figure 10: A summary of the error distributions for predictions made on random permutations of 50 randomly selected molecules from the QM9 dataset, presented per QM9 task (12 in total) for the random non-canonical SMILES strategy.

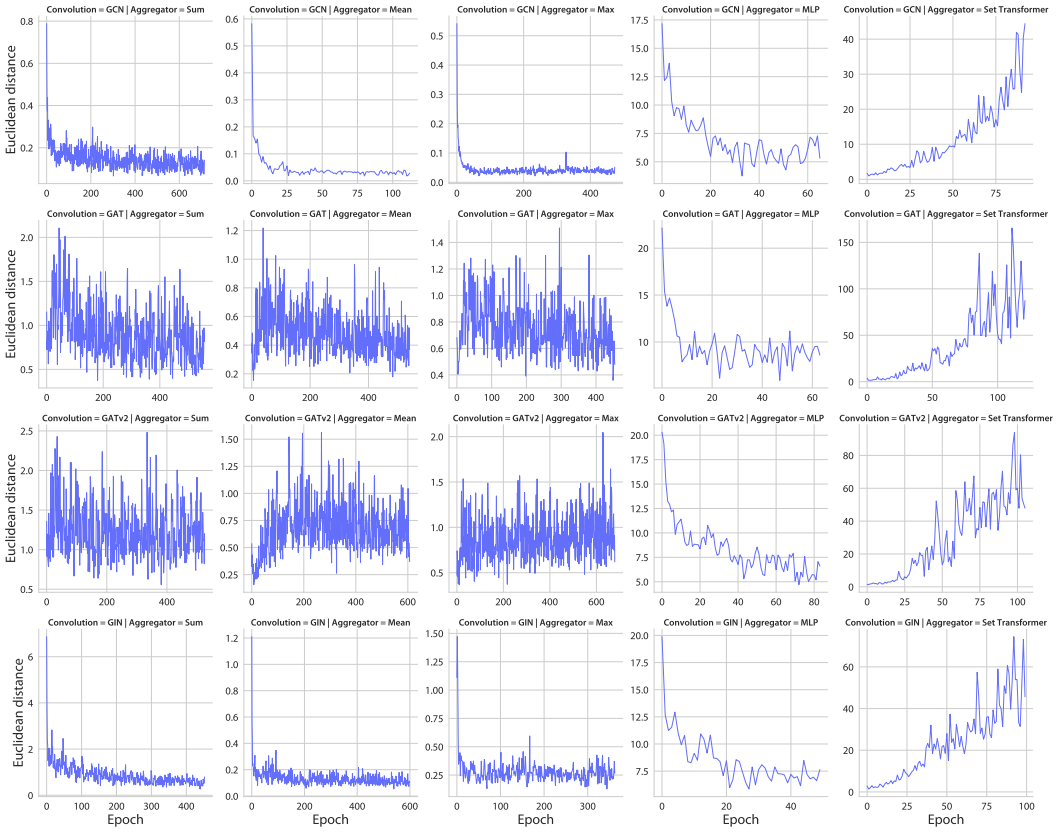


K Distances between graph representations

Appendix Figure 11: The Euclidean distance was computed between the initial graph representation (i.e. after the first epoch) and all subsequent epochs for a random molecule of the QM9 dataset, for multiple graph convolution types and readouts (2-layer GNNs). Generally, models using standard aggregators (sum, mean, max) take a long time to converge (500 to 1,000 epochs), with only minor modifications to the graph representation. The models using neural readouts typically converge in under 100 epochs and are able to explore a much larger hypothesis space, as indicated by the large distances between the initial and final trained representations.



Appendix Figure 12: The Euclidean distance was computed between the graph representations from consecutive epochs for a random molecule of the QM9 dataset (same as Appendix Figure 11), for multiple graph convolution types and readouts (2-layer GNNs). Generally, models using standard aggregators (sum, mean, max) take a long time to converge (500 to $1,000$ epochs), with only minor modifications to the graph representation. The models using neural readouts typically converge in under 100 epochs and are able to explore a much larger hypothesis space, as indicated by the large distances between the initial and final trained representations.

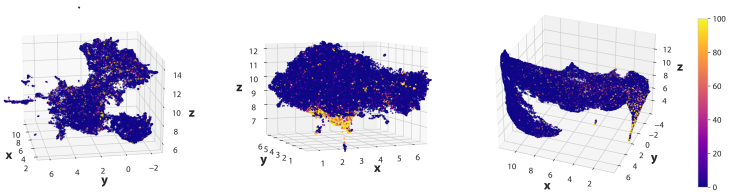
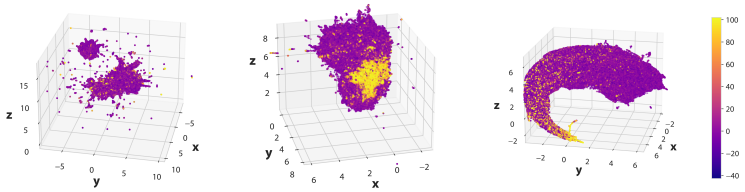
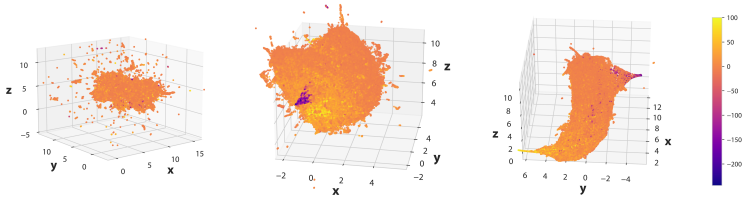


L Visualization of the learned latent space for different aggregators

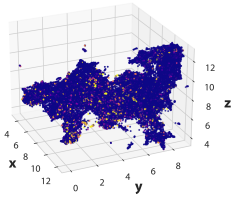
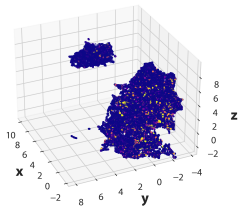
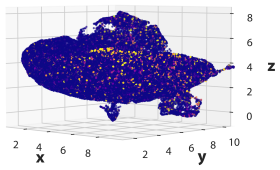
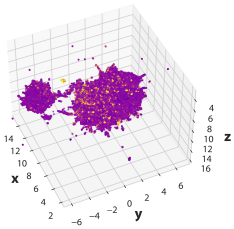
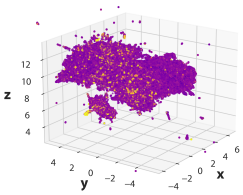
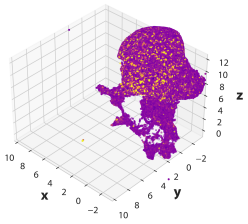
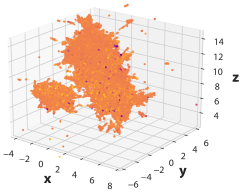
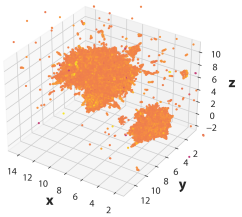
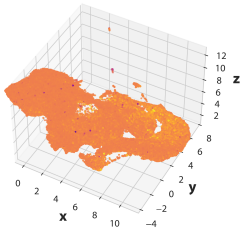
Guided VGAE architecture

We adapted our fixed architecture into a variational graph autoencoder (VGAE, as introduced by Kipf and Welling [19]). The changes include two additional layers for the μ and σ parameters, as well as the reconstruction and regularization losses (as provided by PyTorch Geometric). The graph embeddings learned by the VGAE are fed into a predictor network in an end-to-end fashion, such that the task is supervised.

Appendix Table 6: Visualization of the learned latent space of graphs (molecules) for three recently-introduced bio-affinity datasets, using UMAP projections in 3 dimensions. The figure presents a selection of 3 readouts. Angles are chosen to best highlight the 3D space structure. The activity of molecules, as reported in the bioassay, is illustrated according to the color bars.

Dataset	Sum	MLP	Set Transformer
AID1949			
AID449762			
AID602261			

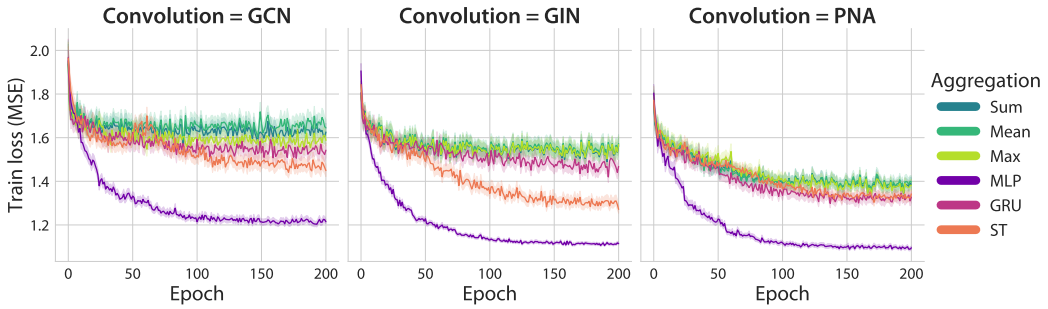
Appendix Table 7: Visualization of the learned latent space of graphs (molecules) for three recently-introduced bio-affinity datasets, using UMAP projections in 3 dimensions. The figure presents the other 3 readouts. Angles are chosen to best highlight the 3D space structure.

Dataset	Mean	Max	GRU
AID1949			
AID449762			
AID602261			

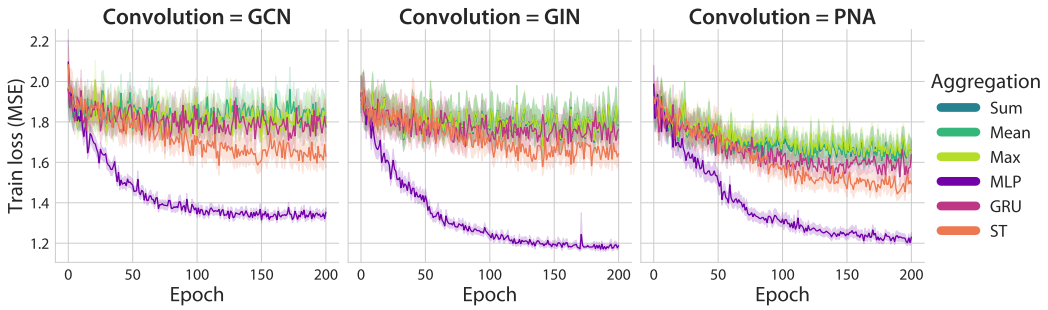
M Train losses plotted for the multi-million scale pharma datasets

M.1 VGAE models

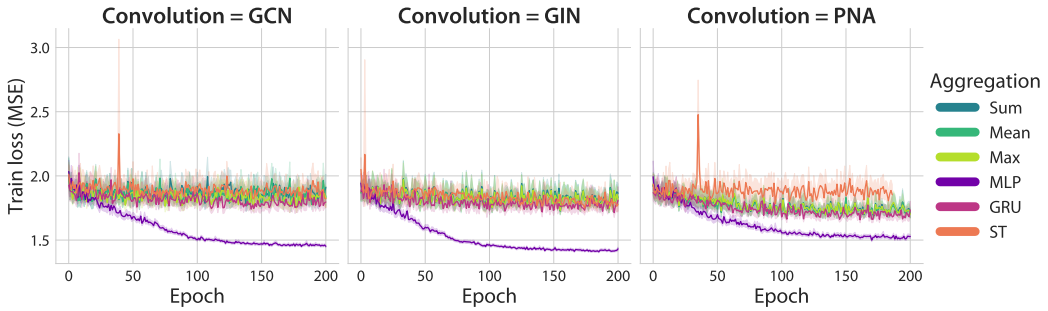
Appendix Figure 13: Train losses for the VGAE models trained on the proprietary dataset with ≈ 1 million molecules.



Appendix Figure 14: Train losses for the VGAE models trained on the proprietary dataset with ≈ 1.5 million molecules.

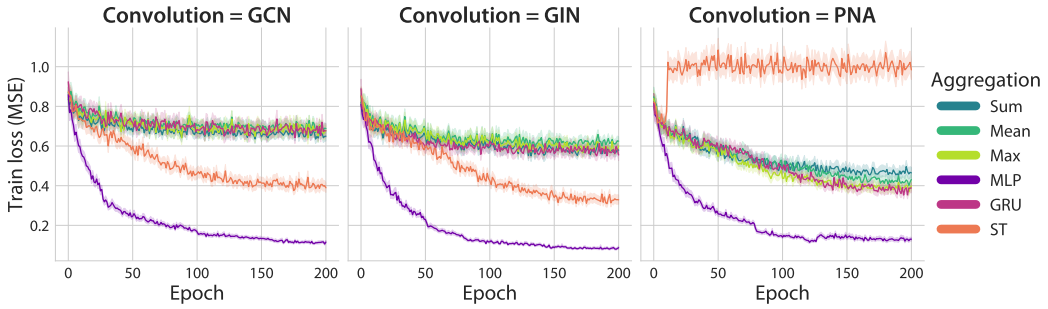


Appendix Figure 15: Train losses for the VGAE models trained on the proprietary dataset with ≈ 2 million molecules.

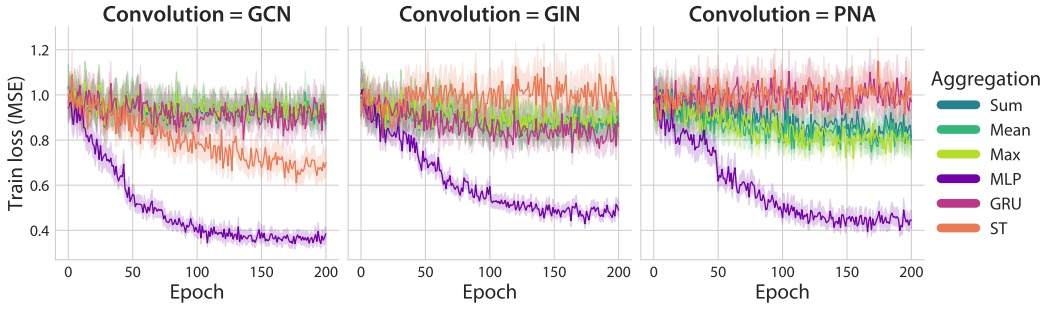


M.2 GNN models

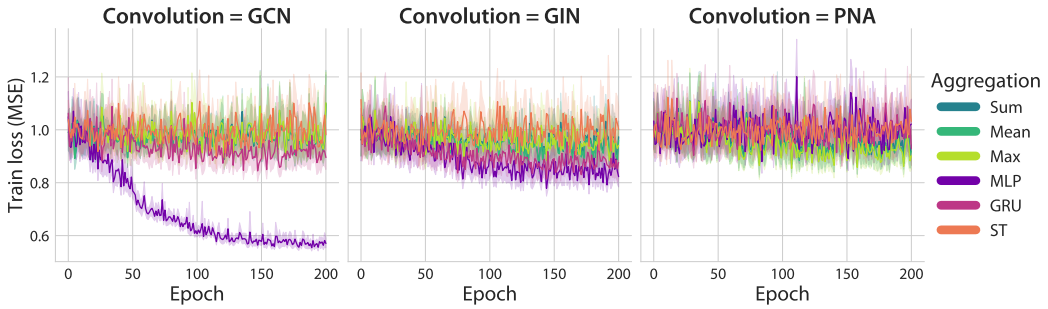
Appendix Figure 16: Train losses for the GNN models trained on the proprietary dataset with ≈ 1 million molecules.



Appendix Figure 17: Train losses for the GNN models trained on the proprietary dataset with ≈ 1.5 million molecules.

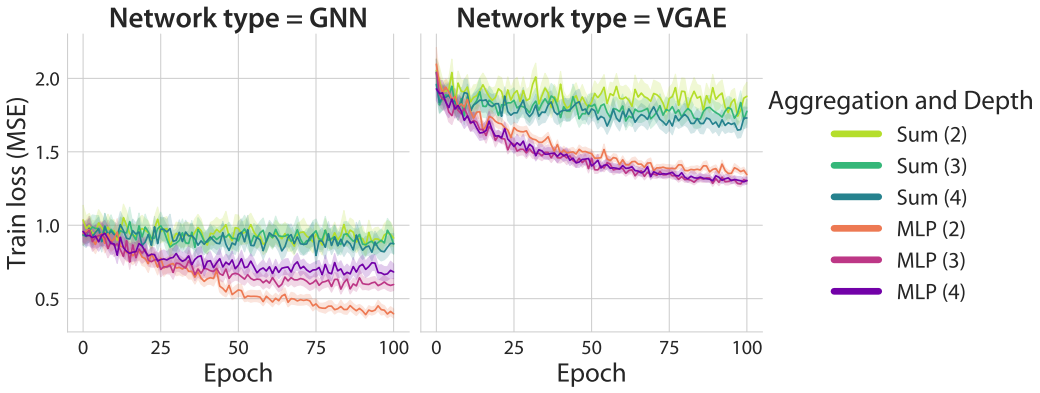


Appendix Figure 18: Train losses for the GNN models trained on the proprietary dataset with ≈ 2 million molecules.



N Train losses for varying GNN depths on the proprietary dataset with 1.5 million molecules

Appendix Figure 19: Train losses (MSE) for the GNN and guided VGAE models trained on the proprietary dataset with ≈ 1.5 million molecules. The illustration includes representative aggregators for each category (sum, respectively MLP) evaluated with GCN layers across 3 different GNN depths (2, 3, and 4 layers). The VGAE is higher as it includes additional loss terms.



O Train metrics for the multi-million scale pharma datasets

Appendix Table 8: Train metrics for the VGAE models trained on the proprietary dataset with $\approx$1 million molecules. A p-value of $< \epsilon$ indicates that the number returned by `scipy` (`stats.pearsonr`) was below the machine precision (thus, reported as 0). A value of 'N/A' indicates that it was not possible to compute the metric.

Dataset	GNN or VGAE	Convolution	Aggregator	MAE	R^2	R^2 p-value
Bio-affinity 1 mil.	VGAE	GCN	Sum	1.13	0.29	$< \epsilon$
			Mean	1.15	0.27	$< \epsilon$
			Max	1.11	0.33	$< \epsilon$
			MLP	**0.68**	**0.78**	$< \epsilon$
			GRU	1.08	0.37	$< \epsilon$
			ST	1.00	0.47	$< \epsilon$
		GIN	Sum	1.07	0.38	$< \epsilon$
			Mean	1.08	0.37	$< \epsilon$
			Max	1.07	0.37	$< \epsilon$
			MLP	**0.55**	**0.86**	$< \epsilon$
			GRU	1.03	0.44	$< \epsilon$
			ST	0.84	0.65	$< \epsilon$
		PNA	Sum	0.93	0.55	$< \epsilon$
			Mean	0.94	0.53	$< \epsilon$
			Max	1.03	0.54	$< \epsilon$
			MLP	**0.56**	**0.86**	$< \epsilon$
			GRU	0.89	0.59	$< \epsilon$
			ST	0.76	0.72	$< \epsilon$

Appendix Table 9: Train metrics for the GNN models trained on the proprietary dataset with ≈1 million molecules. A p-value of $< \epsilon$ indicates that the number returned by `scipy` (`stats.pearsonr`) was below the machine precision (thus, reported as 0). A value of 'N/A' indicates that it was not possible to compute the metric.

Dataset	GNN or VGAE	Convolution	Aggregator	MAE	R^2	R^2 p-value
Bio-affinity 1 mil.	GNN	GCN	Sum	1.10	0.34	$< \epsilon$
			Mean	1.12	0.30	$< \epsilon$
			Max	1.11	0.32	$< \epsilon$
			MLP	**0.47**	**0.89**	$< \epsilon$
			GRU	1.11	0.32	$< \epsilon$
			ST	0.86	0.60	$< \epsilon$
		GIN	Sum	1.03	0.43	$< \epsilon$
			Mean	1.06	0.38	$< \epsilon$
			Max	1.04	0.41	$< \epsilon$
			MLP	**0.40**	**0.92**	$< \epsilon$
			GRU	1.04	0.42	$< \epsilon$
			ST	0.79	0.66	$< \epsilon$
		PNA	Sum	0.94	0.54	$< \epsilon$
			Mean	0.90	0.59	$< \epsilon$
			Max	0.87	0.61	$< \epsilon$
			MLP	**0.46**	**0.87**	$< \epsilon$
			GRU	0.85	0.62	$< \epsilon$
			ST	1.34	N/A	N/A

Appendix Table 10: Train metrics for the VGAE models trained on the proprietary dataset with ≈1.5 million molecules. A p-value of $< \epsilon$ indicates that the number returned by `scipy` (`stats.pearsonr`) was below the machine precision (thus, reported as 0). A value of 'N/A' indicates that it was not possible to compute the metric.

Dataset	GNN or VGAE	Convolution	Aggregator	MAE	R^2	R^2 p-value
Bio-affinity 1.5 mil.	VGAE	GCN	Sum	1.28	0.07	$< \epsilon$
			Mean	1.28	0.05	$< \epsilon$
			Max	1.32	0.07	$< \epsilon$
			MLP	**0.93**	**0.64**	$< \epsilon$
			GRU	1.28	0.10	$< \epsilon$
			ST	1.17	0.29	$< \epsilon$
		GIN	Sum	1.26	0.11	$< \epsilon$
			Mean	1.25	0.10	$< \epsilon$
			Max	1.31	0.10	$< \epsilon$
			MLP	**0.78**	**0.78**	$< \epsilon$
			GRU	1.26	0.13	$< \epsilon$
			ST	1.19	0.27	$< \epsilon$
		PNA	Sum	1.21	0.26	$< \epsilon$
			Mean	1.21	0.26	$< \epsilon$
			Max	1.34	0.24	$< \epsilon$
			MLP	**0.83**	**0.73**	$< \epsilon$
			GRU	1.19	0.30	$< \epsilon$
			ST	1.09	0.43	$< \epsilon$

Appendix Table 11: Train metrics for the GNN models trained on the proprietary dataset with ≈1.5 million molecules. A p-value of $< \epsilon$ indicates that the number returned by `scipy` (`stats.pearsonr`) was below the machine precision (thus, reported as 0). A value of 'N/A' indicates that it was not possible to compute the metric.

Dataset	GNN or VGAE	Convolution	Aggregator	MAE	R^2	R^2 p-value
Bio-affinity 1.5 mil.	GNN	GCN	Sum	1.29	0.07	$< \epsilon$
			Mean	1.29	0.06	$< \epsilon$
			Max	1.29	0.06	$< \epsilon$
			MLP	**0.90**	**0.64**	$< \epsilon$
			GRU	1.28	0.09	$< \epsilon$
			ST	1.15	0.32	$< \epsilon$
		GIN	Sum	1.26	0.13	$< \epsilon$
			Mean	1.27	0.12	$< \epsilon$
			Max	1.27	0.11	$< \epsilon$
			MLP	**1.04**	**0.50**	$< \epsilon$
			GRU	1.25	0.16	$< \epsilon$
			ST	1.30	N/A	N/A
		PNA	Sum	1.25	0.16	$< \epsilon$
			Mean	1.24	0.19	$< \epsilon$
			Max	1.23	0.20	$< \epsilon$
			MLP	**0.96**	**0.56**	$< \epsilon$
			GRU	1.30	0.00	4.22×10^{21}
			ST	1.30	N/A	N/A

Appendix Table 12: Train metrics for the VGAE models trained on the proprietary dataset with ≈2 million molecules. A p-value of $< \epsilon$ indicates that the number returned by `scipy` (`stats.pearsonr`) was below the machine precision (thus, reported as 0). A value of 'N/A' indicates that it was not possible to compute the metric.

Dataset	GNN or VGAE	Convolution	Aggregator	MAE	R^2	R^2 p-value
Bio-affinity 2 mil.	VGAE	GCN	Sum	0.83	0.02	$< \epsilon$
			Mean	0.83	0.03	$< \epsilon$
			Max	0.84	0.06	$< \epsilon$
			MLP	**0.63**	**0.52**	$< \epsilon$
			GRU	0.82	0.09	$< \epsilon$
			ST	0.83	N/A	N/A
		GIN	Sum	0.82	0.06	$< \epsilon$
			Mean	0.82	0.07	$< \epsilon$
			Max	0.83	0.08	$< \epsilon$
			MLP	**0.61**	**0.55**	$< \epsilon$
			GRU	0.82	0.11	$< \epsilon$
			ST	0.82	0.10	$< \epsilon$
		PNA	Sum	0.82	0.15	$< \epsilon$
			Mean	0.82	0.15	$< \epsilon$
			Max	0.84	0.15	$< \epsilon$
			MLP	**0.70**	**0.42**	$< \epsilon$
			GRU	0.81	0.17	$< \epsilon$
			ST	0.83	N/A	N/A

Appendix Table 13: Train metrics for the GNN models trained on the proprietary dataset with ≈2 million molecules. A p-value of $< \epsilon$ indicates that the number returned by `scipy` (`stats.pearsonr`) was below the machine precision (thus, reported as 0). A value of 'N/A' indicates that it was not possible to compute the metric.

Dataset	GNN or VGAE	Convolution	Aggregator	MAE	R^2	R^2 p-value
Bio-affinity 2 mil.	GNN	GCN	Sum	0.83	0.03	$< \epsilon$
			Mean	0.83	0.03	$< \epsilon$
			Max	0.83	0.02	$< \epsilon$
			MLP	**0.68**	**0.43**	$< \epsilon$
			GRU	0.82	0.09	$< \epsilon$
			ST	0.83	N/A	N/A
		GIN	Sum	0.82	0.05	$< \epsilon$
			Mean	0.82	0.06	$< \epsilon$
			Max	0.82	0.05	$< \epsilon$
			MLP	**0.80**	**0.16**	$< \epsilon$
			GRU	0.81	0.12	$< \epsilon$
			ST	0.83	N/A	N/A
		PNA	Sum	0.82	0.05	$< \epsilon$
			Mean	0.82	0.06	$< \epsilon$
			Max	**0.82**	**0.07**	$< \epsilon$
			MLP	0.83	0.00	1.29×10^{-13}
			GRU	0.83	0.00	$< \epsilon$
			ST	0.83	N/A	N/A

P Experimental platform

We used two different platforms for training all the models discussed in the paper. Firstly, a workstation equipped with an AMD Ryzen 5950X processor with 16 cores and 32 threads, an Nvidia RTX 3090 graphics card with 24GB of VRAM, and 64GB of DDR4 RAM. The used operating system is Ubuntu 21.10, with Python 3.9.9, PyTorch 1.10.1 with CUDA 11.3, PyTorch Geometric 2.0.3, and PyTorch Lightning 1.5.7.

Secondly, we used GPU-enabled systems from the AstraZeneca Scientific Computing Platform, generally equipped with Intel processors, Nvidia Tesla V100 GPUs with either 16GB or 32GB, and as much RAM as needed for the experiments. The cloud systems run CentOS Linux 7, with Python 3.9.7, PyTorch 1.8.2 and CUDA 10.2, PyTorch Geometric 2.0.1, and PyTorch Lightning 1.5.5.

Q Statistical significance of dataset attributes for the observed performance

We computed the difference between the best neural aggregator and the best non-neural aggregator for each dataset and for each one of the five random splits (ΔR^2 for regression datasets and ΔMCC for classification datasets). Using these resulting metrics, we fitted multiple linear regression models (`lm` in R) to explain and quantify the relationship between dataset attributes (size, average number of nodes per graph, and others) and the difference in performance.

For the regression datasets, all the dataset attributes from Appendix A, Table 2 except the number of tasks were statistically significant at different levels (Appendix Q, Table 14). Most had a positive relationship to the ΔR^2 except the average number of edges per graph and the number of tasks per dataset. These results suggest that larger regression datasets with more nodes per graph and more node features are likely to see large improvements when using neural aggregators.

For the classification datasets there were no statistically significant dataset attributes (Appendix Q, Table 15), indicating that the proposed techniques are not particularly tied to characteristics like dataset size or graph properties.

Appendix Table 14: Summary of the multiple linear regression model that explains the ΔR^2 in terms of the dataset attributes. p-value significance is indicated by the 'Sign.' column ('***' for < 0.001, '**' for < 0.01, '.' for < 0.1).

Predictors	Estimates	95% Confidence Interval			p-value	Sign.
(Intercept)	-7.06×10^{-01}	-9.88×10^{-01}	–	-4.24×10^{-01}	1.33×10^{-05}	***
Size	1.84×10^{-06}	8.99×10^{-07}	–	2.78×10^{-06}	3.46×10^{-04}	***
Avg. nodes	6.32×10^{-02}	2.71×10^{-02}	–	9.94×10^{-02}	1.14×10^{-03}	**
Avg. edges	-2.36×10^{-02}	-3.92×10^{-02}	–	-8.01×10^{-03}	4.11×10^{-03}	**
Node attr.	1.83×10^{-02}	1.03×10^{-02}	–	2.63×10^{-02}	4.96×10^{-05}	***
Num. tasks	-8.59×10^{-03}	-1.73×10^{-02}	–	9.44×10^{-05}	5.24×10^{-02}	.

Appendix Table 15: Summary of the multiple linear regression model that explains the ΔMCC in terms of the dataset attributes. p-value significance is indicated by the 'Sign.' column ('***' for < 0.001).

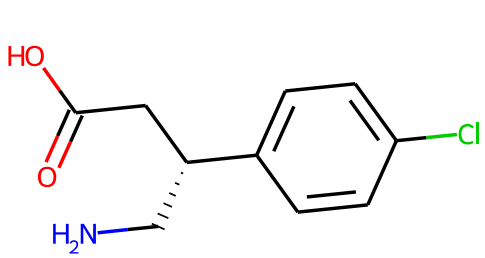
Predictors	Estimates	95% Confidence Interval			p-value	Sign.
(Intercept)	8.83×10^{-02}	5.40×10^{-02}	–	1.23×10^{-01}	1.10×10^{-06}	***
Size	-1.60×10^{-07}	-4.94×10^{-07}	–	1.73×10^{-07}	0.344	
Avg. nodes	2.10×10^{-04}	-2.10×10^{-04}	–	6.29×10^{-04}	0.326	
Avg. edges	-8.56×10^{-05}	-2.65×10^{-04}	–	9.43×10^{-05}	0.349	
Node attr.	-3.36×10^{-05}	-8.60×10^{-05}	–	1.88×10^{-05}	0.207	
Num. tasks	-3.70×10^{-04}	-1.29×10^{-03}	–	5.47×10^{-04}	0.426	

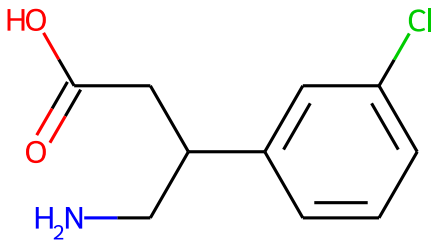
R Similar molecules with different representations

The example illustrates two very similar molecules with greatly different adjacency matrix representations (Appendix R, Figure 20). Despite the similarity, only 6 out of 14 rows are identical in both matrices (rows 1, 2, 3, 4, 6, 8, numbering from 0). There are no rows which occur in both matrices but in different orders.

Appendix Figure 20: Similar molecules with different adjacency representations. The SMILES is provided for both molecules.

(a) `NC[C@H](CC(=O)O)c1ccc(Cl)cc1` **(b)** `C1=CC(=CC(=C1)Cl)C(CC(=O)O)CN`





$$\begin{pmatrix}
0&1&0&0&0&0&0&0&0&0&0&0&0&0\\
1&0&1&0&0&0&0&0&0&0&0&0&0&0\\
0&1&0&1&0&0&0&1&0&0&0&0&0&0\\
0&0&1&0&1&0&0&0&0&0&0&0&0&0\\
0&0&0&1&0&1&1&0&0&0&0&0&0&0\\
0&0&0&0&1&0&0&0&0&0&0&0&0&0\\
0&0&0&0&1&0&0&0&0&0&0&0&0&0\\
0&0&1&0&0&0&0&0&1&0&0&0&0&1\\
0&0&0&0&0&0&0&1&0&1&0&0&0&0\\
0&0&0&0&0&0&0&0&1&0&1&0&0&0\\
0&0&0&0&0&0&0&0&0&1&0&1&1&0\\
0&0&0&0&0&0&0&0&0&0&1&0&0&0\\
0&0&0&0&0&0&0&0&0&0&1&0&0&1\\
0&0&0&0&0&0&0&1&0&0&0&0&1&0
\end{pmatrix}
\begin{pmatrix}
0&1&0&0&0&1&0&0&0&0&0&0&0&0\\
1&0&1&0&0&0&0&0&0&0&0&0&0&0\\
0&1&0&1&0&0&0&1&0&0&0&0&0&0\\
0&0&1&0&1&0&0&0&0&0&0&0&0&0\\
0&0&0&1&0&1&1&0&0&0&0&0&0&0\\
1&0&0&0&1&0&0&0&0&0&0&0&0&0\\
0&0&0&0&1&0&0&0&0&0&0&0&0&0\\
0&0&1&0&0&0&0&0&1&0&0&0&1&0\\
0&0&0&0&0&0&0&1&0&1&0&0&0&0\\
0&0&0&0&0&0&0&0&1&0&1&1&0&0\\
0&0&0&0&0&0&0&0&0&1&0&0&0&0\\
0&0&0&0&0&0&0&0&0&1&0&0&0&0\\
0&0&0&0&0&0&0&1&0&0&0&0&0&1\\
0&0&0&0&0&0&0&0&0&0&0&0&1&0
\end{pmatrix}$$

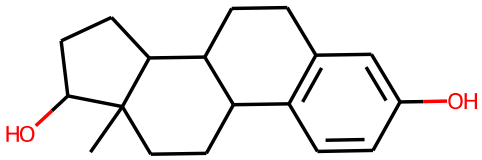
S Similar molecules and the expressiveness of aggregators

As a proof of concept, we selected two well-known molecules with similar structures, estradiol and testosterone (Appendix S, Figure 21). Despite the similarity, their induced biological effects can be very different. We can easily find several examples where the two compounds have different activity levels (active vs inactive) in high-throughput assays from PubChem, for example AID 588544, AID 1347036, AID 624032, or AID 1259394. The compound identifiers (CIDs) for the two compounds are 5757 for estradiol and 6013 for testosterone.

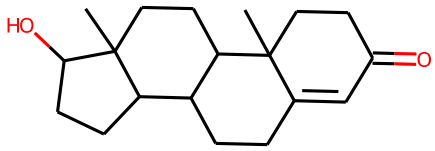
However, GNNs using sum, mean, or max readouts might find it challenging to discern between the two for bio-affinity predictions tasks. On a toy GNN with a single GCN layer, DeepChem featurisation (30 node features), 5 output features from the GCN layer, and random initialization with a seed of 1 (`pytorch_lightning.seed_everything(1)`), the three classical aggregators produced extremely close outputs (Appendix S, Table 16). Although this is only a toy example, it highlights one possible limitation of the simple, existing readout functions.

Appendix Figure 21: Example of similar molecules with different properties.

(a) Estradiol
`CC12CCC3C(C1CCC2O)CCC4=C3C=CC(=C4)O`



(b) Testosterone
`CC12CCC3C(C1CCC2O)CCC4=CC(=O)CCC34C`



Appendix Table 16: Output of the three simple functions for the two similar molecules.

Aggregator	**Molecule**	
	Estradiol	Testosterone
Sum	-39.296	-40.433
Mean	-0.393	-0.385
Max	0.829	0.762

T Detailed metrics for all 39 datasets/benchmarks

The metrics for each layer type, readout, and random split are available in **Supplementary File 2** (available on GitHub).

T.1 MoleculeNet regression models

Appendix Table 17: Detailed metrics (mean ± standard deviation) for the MoleculeNet regression datasets. For QM9, any differences in performance compared to other 2-layer models such as those in Figure 3 might be due to different GNN hyperparameters, such as the output or intermediate node dimension (QM9-specific experiments generally used larger dimensions).

		GCN		GAT		GATv2		GIN		PNA	
Data.	Agg.	MAE	R^2	MAE	R^2	MAE	R^2	MAE	R^2	MAE	R^2
QM9	Sum	0.74 ± 0.00	0.09 ± 0.00	0.76 ± 0.00	0.05 ± 0.01	0.75 ± 0.00	0.08 ± 0.01	0.71 ± 0.00	0.15 ± 0.00	0.70 ± 0.00	0.17 ± 0.01
	Mean	0.73 ± 0.00	0.10 ± 0.00	0.75 ± 0.01	0.07 ± 0.01	0.74 ± 0.00	0.10 ± 0.00	0.72 ± 0.00	0.13 ± 0.00	0.70 ± 0.00	0.16 ± 0.01
	Max	0.73 ± 0.00	0.11 ± 0.00	0.74 ± 0.00	0.10 ± 0.00	0.73 ± 0.00	0.11 ± 0.00	0.70 ± 0.00	0.16 ± 0.00	0.68 ± 0.00	0.20 ± 0.00
	MLP	0.63 ± 0.00	0.31 ± 0.00	0.64 ± 0.00	0.30 ± 0.01	0.64 ± 0.01	0.29 ± 0.01	0.60 ± 0.00	0.38 ± 0.00	0.58 ± 0.00	0.41 ± 0.00
	GRU	0.62 ± 0.01	0.34 ± 0.02	0.62 ± 0.01	0.34 ± 0.01	0.61 ± 0.02	0.35 ± 0.03	0.60 ± 0.00	0.38 ± 0.00	0.60 ± 0.01	0.37 ± 0.02
	ST	0.60 ± 0.01	0.38 ± 0.01	0.63 ± 0.02	0.30 ± 0.04	0.62 ± 0.01	0.32 ± 0.02	0.59 ± 0.00	0.39 ± 0.01	0.57 ± 0.01	0.44 ± 0.01
QM8	Sum	0.58 ± 0.01	0.08 ± 0.01	0.58 ± 0.01	0.08 ± 0.01	0.58 ± 0.01	0.09 ± 0.00	0.59 ± 0.01	0.07 ± 0.01	0.58 ± 0.01	0.09 ± 0.01
	Mean	0.59 ± 0.01	0.07 ± 0.01	0.59 ± 0.01	0.06 ± 0.00	0.59 ± 0.01	0.07 ± 0.01	0.60 ± 0.01	0.07 ± 0.01	0.58 ± 0.01	0.09 ± 0.01
	Max	0.59 ± 0.01	0.07 ± 0.01	0.59 ± 0.01	0.07 ± 0.01	0.59 ± 0.01	0.08 ± 0.01	0.59 ± 0.01	0.07 ± 0.01	0.59 ± 0.01	0.09 ± 0.01
	MLP	0.58 ± 0.01	0.10 ± 0.00	0.57 ± 0.01	0.11 ± 0.01	0.58 ± 0.01	0.10 ± 0.01	0.57 ± 0.01	0.11 ± 0.01	0.56 ± 0.01	0.13 ± 0.02
	GRU	0.57 ± 0.01	0.10 ± 0.01	0.57 ± 0.01	0.11 ± 0.01	0.56 ± 0.01	0.13 ± 0.01	0.58 ± 0.01	0.10 ± 0.01	0.58 ± 0.01	0.10 ± 0.01
	ST	0.59 ± 0.01	0.08 ± 0.01	0.58 ± 0.01	0.10 ± 0.01	0.57 ± 0.01	0.10 ± 0.01	0.59 ± 0.01	0.09 ± 0.01	0.57 ± 0.01	0.12 ± 0.01
QM7	Sum	0.75 ± 0.04	0.15 ± 0.13	0.76 ± 0.03	0.10 ± 0.01	0.78 ± 0.03	0.05 ± 0.05	0.82 ± 0.03	0.00 ± 0.00	0.79 ± 0.03	0.06 ± 0.02
	Mean	0.80 ± 0.03	0.03 ± 0.01	0.79 ± 0.04	0.03 ± 0.02	0.79 ± 0.04	0.03 ± 0.02	0.81 ± 0.03	0.01 ± 0.00	0.79 ± 0.03	0.03 ± 0.02
	Max	0.78 ± 0.04	0.06 ± 0.02	0.78 ± 0.04	0.06 ± 0.03	0.79 ± 0.04	0.03 ± 0.03	0.81 ± 0.04	0.01 ± 0.01	0.78 ± 0.03	0.05 ± 0.03
	MLP	0.79 ± 0.03	0.05 ± 0.02	0.79 ± 0.04	0.06 ± 0.02	0.79 ± 0.04	0.06 ± 0.02	0.80 ± 0.03	0.02 ± 0.01	0.80 ± 0.03	0.02 ± 0.01
	GRU	0.70 ± 0.03	0.24 ± 0.04	0.70 ± 0.03	0.24 ± 0.03	0.69 ± 0.03	0.25 ± 0.05	0.85 ± 0.06	0.09 ± 0.02	0.72 ± 0.02	0.21 ± 0.04
	ST	0.80 ± 0.03	0.03 ± 0.01	0.81 ± 0.04	0.02 ± 0.02	0.81 ± 0.04	0.01 ± 0.01	0.81 ± 0.04	0.01 ± 0.01	0.81 ± 0.04	0.01 ± 0.01
BACE	Sum	1.10 ± 0.00	0.06 ± 0.00	1.05 ± 0.01	0.12 ± 0.05	1.05 ± 0.00	0.14 ± 0.00	1.01 ± 0.00	0.14 ± 0.00	1.06 ± 0.01	0.23 ± 0.15
	Mean	1.05 ± 0.00	0.07 ± 0.00	1.07 ± 0.00	0.02 ± 0.00	1.09 ± 0.00	0.29 ± 0.00	0.97 ± 0.03	0.53 ± 0.02	1.07 ± 0.00	0.15 ± 0.00
	Max	1.05 ± 0.00	0.00 ± 0.00	1.06 ± 0.00	0.50 ± 0.00	1.04 ± 0.00	0.28 ± 0.00	1.04 ± 0.00	0.27 ± 0.00	1.07 ± 0.00	0.29 ± 0.01
	MLP	0.72 ± 0.03	0.63 ± 0.03	0.76 ± 0.02	0.54 ± 0.03	0.75 ± 0.02	0.55 ± 0.03	0.77 ± 0.02	0.56 ± 0.04	0.71 ± 0.03	0.57 ± 0.04
	GRU	0.76 ± 0.00	0.61 ± 0.00	0.77 ± 0.00	0.58 ± 0.00	0.79 ± 0.00	0.60 ± 0.00	0.79 ± 0.00	0.49 ± 0.00	0.79 ± 0.01	0.57 ± 0.01
	ST	1.03 ± 0.00	0.01 ± 0.00	1.04 ± 0.00	0.24 ± 0.01	1.06 ± 0.00	0.40 ± 0.00	1.09 ± 0.00	0.40 ± 0.00	1.09 ± 0.00	0.41 ± 0.00
ESOL	Sum	0.25 ± 0.02	0.89 ± 0.01	0.25 ± 0.02	0.89 ± 0.01	0.24 ± 0.01	0.90 ± 0.01	0.33 ± 0.01	0.81 ± 0.03	0.26 ± 0.03	0.88 ± 0.02
	Mean	0.36 ± 0.03	0.78 ± 0.05	0.33 ± 0.02	0.81 ± 0.02	0.32 ± 0.02	0.82 ± 0.03	0.35 ± 0.04	0.79 ± 0.05	0.34 ± 0.01	0.80 ± 0.02
	Max	0.33 ± 0.02	0.81 ± 0.01	0.38 ± 0.02	0.78 ± 0.03	0.39 ± 0.02	0.78 ± 0.03	0.35 ± 0.02	0.78 ± 0.02	0.33 ± 0.04	0.80 ± 0.04
	MLP	0.30 ± 0.04	0.85 ± 0.02	0.28 ± 0.03	0.85 ± 0.02	0.29 ± 0.03	0.85 ± 0.02	0.33 ± 0.04	0.82 ± 0.03	0.30 ± 0.03	0.84 ± 0.03
	GRU	0.27 ± 0.03	0.85 ± 0.02	0.25 ± 0.02	0.88 ± 0.02	0.25 ± 0.02	0.88 ± 0.02	0.30 ± 0.03	0.85 ± 0.05	0.41 ± 0.16	0.70 ± 0.22
	ST	0.30 ± 0.03	0.86 ± 0.02	0.29 ± 0.03	0.86 ± 0.02	0.28 ± 0.03	0.86 ± 0.02	0.32 ± 0.03	0.84 ± 0.03	0.27 ± 0.02	0.87 ± 0.02
FreeSolv	Sum	0.19 ± 0.04	0.90 ± 0.06	0.20 ± 0.03	0.90 ± 0.05	0.21 ± 0.03	0.89 ± 0.05	0.27 ± 0.05	0.83 ± 0.07	0.22 ± 0.02	0.88 ± 0.05
	Mean	0.26 ± 0.09	0.85 ± 0.12	0.23 ± 0.03	0.89 ± 0.03	0.25 ± 0.02	0.88 ± 0.03	0.27 ± 0.04	0.86 ± 0.03	0.25 ± 0.03	0.88 ± 0.04
	Max	0.20 ± 0.04	0.91 ± 0.04	0.24 ± 0.03	0.89 ± 0.02	0.26 ± 0.02	0.89 ± 0.03	0.25 ± 0.03	0.88 ± 0.02	0.21 ± 0.02	0.91 ± 0.03
	MLP	0.25 ± 0.03	0.89 ± 0.06	0.22 ± 0.01	0.91 ± 0.03	0.21 ± 0.02	0.91 ± 0.04	0.26 ± 0.04	0.87 ± 0.05	0.21 ± 0.03	0.92 ± 0.03
	GRU	0.28 ± 0.21	0.76 ± 0.33	0.20 ± 0.04	0.90 ± 0.06	0.19 ± 0.02	0.90 ± 0.05	0.24 ± 0.06	0.88 ± 0.09	0.25 ± 0.07	0.88 ± 0.05
	ST	0.21 ± 0.03	0.91 ± 0.04	0.21 ± 0.03	0.90 ± 0.04	0.21 ± 0.02	0.91 ± 0.03	0.22 ± 0.03	0.89 ± 0.03	0.19 ± 0.02	0.92 ± 0.02
Lipo	Sum	0.42 ± 0.04	0.68 ± 0.03	0.39 ± 0.04	0.72 ± 0.03	0.39 ± 0.03	0.72 ± 0.02	0.40 ± 0.02	0.70 ± 0.02	0.36 ± 0.03	0.75 ± 0.03
	Mean	0.50 ± 0.07	0.58 ± 0.09	0.48 ± 0.05	0.63 ± 0.04	0.45 ± 0.04	0.66 ± 0.02	0.46 ± 0.05	0.64 ± 0.05	0.44 ± 0.04	0.67 ± 0.03
	Max	0.48 ± 0.03	0.60 ± 0.03	0.55 ± 0.04	0.53 ± 0.02	0.57 ± 0.02	0.49 ± 0.02	0.44 ± 0.03	0.65 ± 0.04	0.42 ± 0.02	0.68 ± 0.04
	MLP	0.54 ± 0.03	0.52 ± 0.03	0.53 ± 0.04	0.54 ± 0.04	0.53 ± 0.03	0.53 ± 0.04	0.53 ± 0.02	0.53 ± 0.03	0.48 ± 0.02	0.61 ± 0.02
	GRU	0.50 ± 0.04	0.58 ± 0.05	0.50 ± 0.08	0.61 ± 0.07	0.52 ± 0.06	0.62 ± 0.06	0.48 ± 0.05	0.61 ± 0.05	0.51 ± 0.14	0.54 ± 0.24
	ST	0.43 ± 0.01	0.68 ± 0.03	0.42 ± 0.02	0.69 ± 0.03	0.41 ± 0.03	0.71 ± 0.03	0.41 ± 0.01	0.70 ± 0.03	0.39 ± 0.03	0.73 ± 0.03

T.2 MoleculeNet classification models

Appendix Table 18: Detailed metrics (mean ± standard deviation) for the MoleculeNet classification datasets.

		GCN		GAT		GATv2		GIN		PNA	
Data.	Agg.	AUROC	MCC	AUROC	MCC	AUROC	MCC	AUROC	MCC	AUROC	MCC
PCBA	Sum	0.53 ± 0.00	0.20 ± 0.00	0.54 ± 0.00	0.22 ± 0.01	0.54 ± 0.00	0.24 ± 0.00	0.55 ± 0.00	0.25 ± 0.01	0.56 ± 0.01	0.27 ± 0.01
	Mean	0.53 ± 0.00	0.19 ± 0.01	0.53 ± 0.00	0.20 ± 0.01	0.54 ± 0.01	0.22 ± 0.01	0.55 ± 0.00	0.24 ± 0.01	0.56 ± 0.01	0.27 ± 0.01
	Max	0.53 ± 0.00	0.20 ± 0.01	0.53 ± 0.00	0.20 ± 0.01	0.53 ± 0.00	0.21 ± 0.01	0.55 ± 0.00	0.24 ± 0.01	0.57 ± 0.01	0.28 ± 0.01
	MLP	0.52 ± 0.01	0.12 ± 0.06	0.52 ± 0.01	0.13 ± 0.04	0.51 ± 0.01	0.12 ± 0.04	0.52 ± 0.02	0.15 ± 0.06	0.52 ± 0.01	0.15 ± 0.05
	GRU	0.53 ± 0.00	0.17 ± 0.01	0.55 ± 0.01	0.24 ± 0.01	0.55 ± 0.01	0.24 ± 0.01	0.55 ± 0.00	0.24 ± 0.01	0.56 ± 0.01	0.25 ± 0.01
	ST	0.56 ± 0.01	0.26 ± 0.01	0.58 ± 0.01	0.30 ± 0.01	0.58 ± 0.00	0.29 ± 0.01	0.57 ± 0.01	0.27 ± 0.01	0.59 ± 0.01	0.31 ± 0.01
BACE	Sum	0.65 ± 0.00	0.32 ± 0.00	0.67 ± 0.00	0.37 ± 0.00	0.70 ± 0.01	0.42 ± 0.02	0.62 ± 0.01	0.28 ± 0.01	0.54 ± 0.01	0.13 ± 0.04
	Mean	0.61 ± 0.01	0.23 ± 0.01	0.61 ± 0.00	0.22 ± 0.01	0.64 ± 0.02	0.28 ± 0.04	0.58 ± 0.00	0.16 ± 0.00	0.66 ± 0.01	0.32 ± 0.02
	Max	0.62 ± 0.01	0.25 ± 0.04	0.70 ± 0.00	0.40 ± 0.00	0.72 ± 0.00	0.43 ± 0.00	0.69 ± 0.02	0.38 ± 0.03	0.47 ± 0.00	-0.09 ± 0.01
	MLP	0.66 ± 0.01	0.32 ± 0.03	0.62 ± 0.00	0.25 ± 0.00	0.65 ± 0.02	0.31 ± 0.03	0.68 ± 0.00	0.36 ± 0.00	0.68 ± 0.01	0.36 ± 0.03
	GRU	0.69 ± 0.00	0.37 ± 0.00	0.70 ± 0.00	0.38 ± 0.00	0.73 ± 0.00	0.45 ± 0.00	0.66 ± 0.01	0.32 ± 0.02	0.63 ± 0.02	0.26 ± 0.03
	ST	0.64 ± 0.00	0.30 ± 0.01	0.69 ± 0.00	0.38 ± 0.00	0.66 ± 0.00	0.36 ± 0.00	0.68 ± 0.01	0.38 ± 0.02	0.65 ± 0.01	0.34 ± 0.01
BBBP	Sum	0.59 ± 0.00	0.21 ± 0.00	0.60 ± 0.00	0.23 ± 0.00	0.62 ± 0.00	0.27 ± 0.01	0.54 ± 0.00	0.11 ± 0.01	0.59 ± 0.01	0.21 ± 0.01
	Mean	0.50 ± 0.00	0.00 ± 0.00	0.52 ± 0.03	0.06 ± 0.08	0.55 ± 0.01	0.15 ± 0.04	0.50 ± 0.00	0.00 ± 0.00	0.58 ± 0.01	0.20 ± 0.02
	Max	0.50 ± 0.00	0.00 ± 0.00	0.63 ± 0.01	0.29 ± 0.02	0.63 ± 0.02	0.30 ± 0.04	0.57 ± 0.01	0.18 ± 0.01	0.62 ± 0.01	0.27 ± 0.03
	MLP	0.63 ± 0.01	0.27 ± 0.03	0.62 ± 0.02	0.27 ± 0.03	0.61 ± 0.02	0.26 ± 0.05	0.57 ± 0.01	0.22 ± 0.02	0.61 ± 0.02	0.24 ± 0.03
	GRU	0.53 ± 0.00	0.12 ± 0.00	0.58 ± 0.00	0.19 ± 0.01	0.58 ± 0.00	0.18 ± 0.01	0.56 ± 0.01	0.14 ± 0.03	0.52 ± 0.02	0.05 ± 0.05
	ST	0.54 ± 0.00	0.11 ± 0.01	0.54 ± 0.00	0.14 ± 0.01	0.53 ± 0.00	0.12 ± 0.00	0.53 ± 0.00	0.11 ± 0.01	0.54 ± 0.01	0.11 ± 0.02
SIDER	Sum	0.74 ± 0.01	0.50 ± 0.02	0.74 ± 0.01	0.50 ± 0.02	0.74 ± 0.01	0.50 ± 0.02	0.74 ± 0.01	0.50 ± 0.02	0.74 ± 0.01	0.50 ± 0.02
	Mean	0.74 ± 0.01	0.50 ± 0.01	0.73 ± 0.01	0.50 ± 0.01	0.73 ± 0.01	0.50 ± 0.02	0.74 ± 0.01	0.51 ± 0.01	0.73 ± 0.01	0.50 ± 0.01
	Max	0.74 ± 0.01	0.50 ± 0.02	0.74 ± 0.01	0.50 ± 0.01	0.74 ± 0.01	0.50 ± 0.02	0.74 ± 0.01	0.50 ± 0.01	0.74 ± 0.01	0.50 ± 0.02
	MLP	0.73 ± 0.01	0.49 ± 0.03	0.73 ± 0.00	0.49 ± 0.01	0.73 ± 0.01	0.48 ± 0.02	0.73 ± 0.00	0.49 ± 0.01	0.73 ± 0.01	0.49 ± 0.01
	GRU	0.73 ± 0.01	0.48 ± 0.02	0.74 ± 0.01	0.50 ± 0.03	0.74 ± 0.01	0.50 ± 0.02	0.73 ± 0.01	0.47 ± 0.02	0.73 ± 0.01	0.49 ± 0.01
	ST	0.73 ± 0.01	0.49 ± 0.02	0.73 ± 0.00	0.50 ± 0.01	0.73 ± 0.01	0.49 ± 0.01	0.73 ± 0.00	0.49 ± 0.01	0.73 ± 0.00	0.49 ± 0.01
HIV	Sum	0.55 ± 0.02	0.15 ± 0.03	0.56 ± 0.02	0.21 ± 0.06	0.55 ± 0.04	0.17 ± 0.08	0.58 ± 0.02	0.23 ± 0.05	0.62 ± 0.01	0.35 ± 0.01
	Mean	0.58 ± 0.03	0.18 ± 0.07	0.51 ± 0.01	0.02 ± 0.02	0.56 ± 0.02	0.12 ± 0.02	0.53 ± 0.03	0.08 ± 0.07	0.57 ± 0.00	0.25 ± 0.01
	Max	0.58 ± 0.03	0.21 ± 0.04	0.56 ± 0.04	0.13 ± 0.09	0.58 ± 0.01	0.19 ± 0.04	0.58 ± 0.02	0.20 ± 0.04	0.58 ± 0.01	0.30 ± 0.04
	MLP	0.55 ± 0.01	0.22 ± 0.01	0.55 ± 0.01	0.23 ± 0.04	0.54 ± 0.01	0.18 ± 0.04	0.56 ± 0.01	0.25 ± 0.03	0.54 ± 0.01	0.23 ± 0.02
	GRU	0.58 ± 0.00	0.26 ± 0.01	0.56 ± 0.01	0.25 ± 0.01	0.56 ± 0.01	0.23 ± 0.02	0.55 ± 0.02	0.16 ± 0.02	0.52 ± 0.02	0.09 ± 0.08
	ST	0.57 ± 0.02	0.23 ± 0.13	0.55 ± 0.03	0.22 ± 0.08	0.59 ± 0.01	0.33 ± 0.00	0.57 ± 0.01	0.27 ± 0.03	0.61 ± 0.02	0.34 ± 0.04

T.3 Social networks classification models

Appendix Table 19: Detailed metrics (mean ± standard deviation) for the social network classification datasets. 'OOM' stands for out-of-memory (RAM).

		GCN		GAT		GATv2		GIN		PNA	
Data.	Agg.	AUROC	MCC	AUROC	MCC	AUROC	MCC	AUROC	MCC	AUROC	MCC
IMDB-BINARY	Sum	0.74 ± 0.03	0.49 ± 0.07	0.73 ± 0.02	0.47 ± 0.05	0.72 ± 0.03	0.44 ± 0.06	0.74 ± 0.03	0.49 ± 0.05	0.72 ± 0.03	0.45 ± 0.05
	Mean	0.73 ± 0.04	0.46 ± 0.08	0.72 ± 0.03	0.45 ± 0.05	0.71 ± 0.03	0.42 ± 0.07	0.72 ± 0.02	0.44 ± 0.03	0.72 ± 0.03	0.46 ± 0.08
	Max	0.72 ± 0.03	0.44 ± 0.06	0.72 ± 0.04	0.44 ± 0.09	0.72 ± 0.05	0.44 ± 0.11	0.74 ± 0.02	0.47 ± 0.03	0.71 ± 0.02	0.43 ± 0.03
	MLP	0.71 ± 0.03	0.42 ± 0.07	0.70 ± 0.02	0.40 ± 0.04	0.71 ± 0.05	0.41 ± 0.10	0.71 ± 0.03	0.42 ± 0.06	0.71 ± 0.05	0.42 ± 0.09
	GRU	0.72 ± 0.04	0.45 ± 0.08	0.74 ± 0.03	0.48 ± 0.06	0.73 ± 0.02	0.46 ± 0.05	0.73 ± 0.04	0.47 ± 0.08	0.68 ± 0.04	0.36 ± 0.09
	ST	0.72 ± 0.04	0.44 ± 0.08	0.72 ± 0.04	0.45 ± 0.07	0.73 ± 0.03	0.47 ± 0.06	0.72 ± 0.06	0.45 ± 0.11	0.72 ± 0.03	0.44 ± 0.06
TWITTER-Real-Graph-Partial	Sum	0.64 ± 0.00	0.28 ± 0.00	0.62 ± 0.01	0.25 ± 0.01	0.62 ± 0.01	0.25 ± 0.01	0.62 ± 0.00	0.24 ± 0.01	0.62 ± 0.00	0.23 ± 0.01
	Mean	0.65 ± 0.01	0.29 ± 0.01	0.64 ± 0.00	0.28 ± 0.01	0.64 ± 0.00	0.29 ± 0.00	0.62 ± 0.01	0.25 ± 0.01	0.62 ± 0.01	0.23 ± 0.01
	Max	0.64 ± 0.01	0.29 ± 0.01	0.63 ± 0.00	0.27 ± 0.01	0.64 ± 0.00	0.28 ± 0.00	0.62 ± 0.01	0.25 ± 0.01	0.61 ± 0.01	0.23 ± 0.02
	MLP	0.62 ± 0.01	0.23 ± 0.01	0.61 ± 0.01	0.23 ± 0.01	0.62 ± 0.01	0.24 ± 0.01	0.62 ± 0.01	0.25 ± 0.01	0.61 ± 0.01	0.22 ± 0.01
	GRU	0.63 ± 0.01	0.28 ± 0.01	0.63 ± 0.00	0.26 ± 0.01	0.63 ± 0.00	0.26 ± 0.01	0.62 ± 0.00	0.25 ± 0.01	0.61 ± 0.01	0.23 ± 0.01
	ST	0.63 ± 0.00	0.25 ± 0.01	0.62 ± 0.00	0.24 ± 0.01	0.62 ± 0.01	0.24 ± 0.01	0.62 ± 0.01	0.24 ± 0.01	0.61 ± 0.00	0.22 ± 0.01
reddit_threads	Sum	0.78 ± 0.00	0.56 ± 0.01	0.78 ± 0.00	0.56 ± 0.01	0.77 ± 0.00	0.56 ± 0.00	0.78 ± 0.00	0.56 ± 0.00	0.77 ± 0.00	0.55 ± 0.01
	Mean	0.77 ± 0.00	0.56 ± 0.01	0.77 ± 0.00	0.55 ± 0.01	0.77 ± 0.00	0.55 ± 0.01	0.77 ± 0.00	0.55 ± 0.01	0.77 ± 0.00	0.56 ± 0.01
	Max	0.77 ± 0.00	0.55 ± 0.01	0.76 ± 0.01	0.55 ± 0.01	0.77 ± 0.00	0.55 ± 0.01	0.77 ± 0.00	0.55 ± 0.00	0.77 ± 0.00	0.55 ± 0.00
	MLP	0.75 ± 0.00	0.50 ± 0.01	0.76 ± 0.00	0.53 ± 0.01	0.76 ± 0.00	0.53 ± 0.01	0.75 ± 0.00	0.51 ± 0.01	0.75 ± 0.00	0.51 ± 0.00
	GRU	0.77 ± 0.00	0.56 ± 0.00	0.77 ± 0.00	0.56 ± 0.01	0.77 ± 0.00	0.56 ± 0.00	0.77 ± 0.00	0.55 ± 0.01	0.77 ± 0.00	0.54 ± 0.00
	ST	0.78 ± 0.00	0.56 ± 0.00	0.77 ± 0.00	0.56 ± 0.01	0.77 ± 0.00	0.55 ± 0.01	0.77 ± 0.00	0.55 ± 0.00	0.77 ± 0.00	0.55 ± 0.01
REDDIT-BINARY	Sum	0.82 ± 0.04	0.65 ± 0.08	0.71 ± 0.03	0.42 ± 0.05	0.74 ± 0.03	0.48 ± 0.06	0.80 ± 0.02	0.61 ± 0.04	OOM	OOM
	Mean	0.76 ± 0.03	0.51 ± 0.06	0.67 ± 0.01	0.36 ± 0.03	0.71 ± 0.02	0.43 ± 0.04	0.78 ± 0.04	0.57 ± 0.08	OOM	OOM
	Max	0.77 ± 0.04	0.53 ± 0.08	0.59 ± 0.04	0.26 ± 0.05	0.50 ± 0.01	0.04 ± 0.05	0.79 ± 0.05	0.58 ± 0.08	OOM	OOM
	MLP	0.84 ± 0.04	0.67 ± 0.08	0.82 ± 0.03	0.64 ± 0.06	0.80 ± 0.03	0.60 ± 0.05	0.83 ± 0.02	0.67 ± 0.05	OOM	OOM
	GRU	0.76 ± 0.02	0.52 ± 0.05	0.77 ± 0.03	0.55 ± 0.06	0.77 ± 0.01	0.53 ± 0.03	0.82 ± 0.05	0.64 ± 0.09	OOM	OOM
	ST	0.78 ± 0.06	0.58 ± 0.10	0.69 ± 0.02	0.40 ± 0.05	0.67 ± 0.03	0.36 ± 0.06	0.80 ± 0.04	0.60 ± 0.08	OOM	OOM
twitch_egos	Sum	0.69 ± 0.01	0.39 ± 0.01	0.69 ± 0.00	0.38 ± 0.01	0.69 ± 0.01	0.39 ± 0.01	0.69 ± 0.00	0.37 ± 0.01	0.69 ± 0.01	0.39 ± 0.01
	Mean	0.69 ± 0.01	0.39 ± 0.01	0.69 ± 0.00	0.39 ± 0.01	0.69 ± 0.00	0.39 ± 0.01	0.69 ± 0.01	0.38 ± 0.01	0.69 ± 0.01	0.39 ± 0.01
	Max	0.69 ± 0.01	0.39 ± 0.01	0.68 ± 0.01	0.38 ± 0.01	0.69 ± 0.00	0.38 ± 0.01	0.69 ± 0.00	0.38 ± 0.01	0.69 ± 0.01	0.39 ± 0.01
	MLP	0.63 ± 0.01	0.26 ± 0.01	0.64 ± 0.01	0.29 ± 0.02	0.65 ± 0.01	0.32 ± 0.02	0.64 ± 0.01	0.29 ± 0.01	0.63 ± 0.01	0.25 ± 0.01
	GRU	0.69 ± 0.01	0.40 ± 0.01	0.69 ± 0.00	0.39 ± 0.01	0.69 ± 0.01	0.39 ± 0.01	0.68 ± 0.01	0.37 ± 0.01	0.69 ± 0.00	0.38 ± 0.01
	ST	0.69 ± 0.01	0.39 ± 0.01	0.69 ± 0.00	0.39 ± 0.01	0.69 ± 0.01	0.39 ± 0.01	0.68 ± 0.00	0.37 ± 0.01	0.69 ± 0.01	0.39 ± 0.01
github_stargazers	Sum	0.64 ± 0.01	0.29 ± 0.02	0.61 ± 0.01	0.24 ± 0.02	0.61 ± 0.01	0.25 ± 0.02	0.61 ± 0.01	0.25 ± 0.01	0.64 ± 0.01	0.29 ± 0.02
	Mean	0.63 ± 0.01	0.28 ± 0.02	0.61 ± 0.01	0.24 ± 0.03	0.61 ± 0.00	0.24 ± 0.01	0.62 ± 0.01	0.25 ± 0.02	0.64 ± 0.01	0.29 ± 0.02
	Max	0.62 ± 0.01	0.26 ± 0.02	0.58 ± 0.02	0.21 ± 0.03	0.59 ± 0.01	0.22 ± 0.02	0.62 ± 0.01	0.24 ± 0.02	0.63 ± 0.01	0.27 ± 0.03
	MLP	0.60 ± 0.01	0.21 ± 0.02	0.60 ± 0.01	0.20 ± 0.02	0.59 ± 0.01	0.19 ± 0.02	0.63 ± 0.01	0.26 ± 0.02	0.63 ± 0.02	0.27 ± 0.04
	GRU	0.62 ± 0.02	0.25 ± 0.04	0.61 ± 0.01	0.24 ± 0.01	0.61 ± 0.01	0.24 ± 0.01	0.63 ± 0.01	0.27 ± 0.02	0.59 ± 0.02	0.18 ± 0.03
	ST	0.63 ± 0.01	0.27 ± 0.03	0.60 ± 0.02	0.23 ± 0.03	0.61 ± 0.01	0.24 ± 0.01	0.62 ± 0.02	0.26 ± 0.03	0.65 ± 0.01	0.30 ± 0.02

Appendix Table 20: Detailed metrics (mean ± sd.) for the REDDIT-MULTI-12K dataset. 'OOM' stands for out-of-memory (RAM). Only the MCC is reported as this is a multi-label task.

Dataset	Aggregator	GCN	GAT	GATv2	GIN	PNA
REDDIT-MULTI-12K	Sum	0.33 ± 0.01	0.29 ± 0.03	0.29 ± 0.03	0.33 ± 0.02	OOM
	Mean	0.28 ± 0.01	0.22 ± 0.01	0.23 ± 0.02	0.29 ± 0.01	OOM
	Max	0.28 ± 0.02	0.11 ± 0.01	0.09 ± 0.02	0.27 ± 0.01	OOM
	MLP	0.24 ± 0.01	0.25 ± 0.02	0.24 ± 0.02	0.27 ± 0.03	OOM
	GRU	0.18 ± 0.03	0.26 ± 0.07	0.25 ± 0.03	0.26 ± 0.02	OOM
	ST	0.36 ± 0.01	0.31 ± 0.02	0.30 ± 0.01	0.31 ± 0.02	OOM

T.4 Synthetic classification models

Appendix Table 21: Detailed metrics (mean ± standard deviation) for the SYNTHETIC and SYNTHETICnew datasets.

		GCN		GAT		GATv2		GIN		PNA	
Dataset	Agg.	AUROC	MCC	AUROC	MCC	AUROC	MCC	AUROC	MCC	AUROC	MCC
SYNTHETIC	Sum	0.87 ± 0.18	0.74 ± 0.36	0.99 ± 0.02	0.97 ± 0.04	0.51 ± 0.01	0.03 ± 0.07	0.94 ± 0.03	0.89 ± 0.06	1.00 ± 0.00	1.00 ± 0.00
	Mean	0.50 ± 0.00	0.00 ± 0.00	0.68 ± 0.23	0.36 ± 0.46	0.50 ± 0.00	0.00 ± 0.00	0.91 ± 0.13	0.86 ± 0.22	0.60 ± 0.22	0.20 ± 0.45
	Max	0.87 ± 0.21	0.75 ± 0.42	0.78 ± 0.25	0.59 ± 0.47	0.50 ± 0.00	0.00 ± 0.00	0.95 ± 0.04	0.91 ± 0.08	0.99 ± 0.01	0.99 ± 0.03
	MLP	0.97 ± 0.03	0.95 ± 0.06	1.00 ± 0.00	1.00 ± 0.00	0.99 ± 0.02	0.97 ± 0.04	0.97 ± 0.03	0.95 ± 0.06	0.98 ± 0.03	0.97 ± 0.06
	GRU	0.87 ± 0.21	0.75 ± 0.42	0.79 ± 0.27	0.59 ± 0.54	0.70 ± 0.27	0.40 ± 0.55	0.99 ± 0.02	0.97 ± 0.04	0.99 ± 0.02	0.99 ± 0.03
	ST	0.69 ± 0.26	0.39 ± 0.53	0.50 ± 0.00	0.00 ± 0.00	0.50 ± 0.00	0.00 ± 0.00	0.92 ± 0.05	0.86 ± 0.08	0.79 ± 0.26	0.57 ± 0.52
SYNTHETICnew	Sum	0.46 ± 0.04	-0.08 ± 0.08	0.51 ± 0.07	0.02 ± 0.15	0.51 ± 0.02	0.04 ± 0.09	0.52 ± 0.08	0.05 ± 0.16	0.76 ± 0.11	0.55 ± 0.19
	Mean	0.50 ± 0.01	-0.00 ± 0.02	0.51 ± 0.02	0.02 ± 0.03	0.50 ± 0.00	0.00 ± 0.00	0.53 ± 0.10	0.06 ± 0.20	0.59 ± 0.12	0.17 ± 0.27
	Max	0.47 ± 0.06	-0.06 ± 0.14	0.50 ± 0.04	-0.01 ± 0.16	0.50 ± 0.00	0.00 ± 0.00	0.53 ± 0.09	0.07 ± 0.19	0.91 ± 0.06	0.83 ± 0.12
	MLP	0.90 ± 0.05	0.82 ± 0.09	0.85 ± 0.06	0.73 ± 0.12	0.89 ± 0.04	0.79 ± 0.08	0.85 ± 0.06	0.71 ± 0.11	1.00 ± 0.00	1.00 ± 0.00
	GRU	0.78 ± 0.16	0.56 ± 0.33	0.77 ± 0.19	0.57 ± 0.32	0.50 ± 0.00	0.00 ± 0.00	0.82 ± 0.06	0.64 ± 0.12	0.99 ± 0.02	0.97 ± 0.04
	ST	0.51 ± 0.02	0.02 ± 0.04	0.51 ± 0.02	0.02 ± 0.03	0.50 ± 0.00	0.00 ± 0.00	0.54 ± 0.03	0.10 ± 0.07	0.78 ± 0.17	0.57 ± 0.34

Appendix Table 22: Detailed metrics (mean ± standard deviation) for the Synthie, TRIANGLES, and COLORS-3 datasets. Only the MCC is reported as this is a multi-label task.

Dataset	Aggregator	GCN	GAT	GATv2	GIN	PNA
Synthie	Sum	0.54 ± 0.06	0.52 ± 0.12	0.46 ± 0.09	0.53 ± 0.11	0.95 ± 0.03
	Mean	0.45 ± 0.11	0.36 ± 0.21	0.52 ± 0.11	0.58 ± 0.07	0.93 ± 0.05
	Max	0.14 ± 0.04	0.24 ± 0.07	0.23 ± 0.07	0.30 ± 0.14	0.59 ± 0.04
	MLP	0.71 ± 0.08	0.72 ± 0.13	0.75 ± 0.05	0.59 ± 0.08	0.93 ± 0.03
	GRU	0.75 ± 0.12	0.48 ± 0.21	0.79 ± 0.11	0.72 ± 0.03	0.79 ± 0.08
	ST	0.37 ± 0.23	0.61 ± 0.08	0.71 ± 0.11	0.60 ± 0.09	0.88 ± 0.12
TRIANGLES	Sum	0.11 ± 0.01	0.18 ± 0.01	0.17 ± 0.01	0.13 ± 0.01	0.10 ± 0.04
	Mean	0.14 ± 0.02	0.17 ± 0.01	0.18 ± 0.02	0.12 ± 0.01	0.11 ± 0.02
	Max	0.12 ± 0.01	0.21 ± 0.01	0.21 ± 0.01	0.14 ± 0.02	0.09 ± 0.01
	MLP	0.10 ± 0.01	0.13 ± 0.01	0.10 ± 0.01	0.10 ± 0.01	0.10 ± 0.05
	GRU	0.10 ± 0.02	0.13 ± 0.04	0.15 ± 0.03	0.12 ± 0.01	0.10 ± 0.02
	ST	0.13 ± 0.01	0.17 ± 0.01	0.20 ± 0.02	0.11 ± 0.02	0.12 ± 0.02
COLORS-3	Sum	0.98 ± 0.00	0.37 ± 0.04	0.38 ± 0.03	0.89 ± 0.01	1.00 ± 0.00
	Mean	0.28 ± 0.04	0.28 ± 0.02	0.26 ± 0.02	0.32 ± 0.01	0.38 ± 0.02
	Max	0.41 ± 0.01	0.20 ± 0.05	0.20 ± 0.05	0.53 ± 0.02	0.59 ± 0.02
	MLP	0.58 ± 0.01	0.48 ± 0.01	0.51 ± 0.02	0.41 ± 0.02	0.55 ± 0.04
	GRU	0.91 ± 0.01	0.59 ± 0.06	0.52 ± 0.11	0.79 ± 0.02	1.00 ± 0.00
	ST	0.83 ± 0.03	0.57 ± 0.06	0.52 ± 0.08	0.76 ± 0.05	0.92 ± 0.04

T.5 Bioinformatics classification models

Appendix Table 23: Detailed metrics (mean ± standard deviation) for the PROTEINS_full dataset.

		GCN		GAT		GATv2		GIN		PNA	
Dataset	Agg.	AUROC	MCC	AUROC	MCC	AUROC	MCC	AUROC	MCC	AUROC	MCC
PROTEINS_full	Sum	0.67 ± 0.06	0.38 ± 0.13	0.52 ± 0.03	0.05 ± 0.07	0.57 ± 0.09	0.14 ± 0.17	0.55 ± 0.04	0.18 ± 0.12	0.64 ± 0.04	0.33 ± 0.07
	Mean	0.69 ± 0.06	0.41 ± 0.11	0.67 ± 0.06	0.35 ± 0.10	0.69 ± 0.07	0.39 ± 0.15	0.52 ± 0.02	0.09 ± 0.11	0.71 ± 0.06	0.47 ± 0.09
	Max	0.69 ± 0.04	0.41 ± 0.08	0.51 ± 0.02	0.06 ± 0.06	0.53 ± 0.03	0.13 ± 0.10	0.57 ± 0.06	0.21 ± 0.16	0.72 ± 0.02	0.44 ± 0.05
	MLP	0.62 ± 0.11	0.26 ± 0.24	0.58 ± 0.12	0.17 ± 0.23	0.59 ± 0.11	0.21 ± 0.22	0.66 ± 0.05	0.33 ± 0.09	0.67 ± 0.06	0.35 ± 0.13
	GRU	0.69 ± 0.02	0.39 ± 0.05	0.71 ± 0.02	0.43 ± 0.04	0.71 ± 0.01	0.43 ± 0.03	0.66 ± 0.03	0.39 ± 0.06	0.65 ± 0.05	0.33 ± 0.10
	ST	0.70 ± 0.04	0.42 ± 0.07	0.67 ± 0.05	0.34 ± 0.09	0.70 ± 0.05	0.39 ± 0.09	0.60 ± 0.06	0.30 ± 0.12	0.62 ± 0.06	0.33 ± 0.15

Appendix Table 24: Detailed metrics (mean ± standard deviation) for the ENZYMES dataset. Only the MCC is reported as this is a multi-label task.

Dataset	Aggregator	GCN	GAT	GATv2	GIN	PNA
ENZYMES	Sum	0.16 ± 0.06	0.21 ± 0.14	0.27 ± 0.05	0.50 ± 0.07	0.42 ± 0.08
	Mean	0.16 ± 0.09	0.24 ± 0.10	0.32 ± 0.15	0.44 ± 0.06	0.39 ± 0.11
	Max	0.20 ± 0.10	0.28 ± 0.11	0.25 ± 0.14	0.41 ± 0.05	0.40 ± 0.11
	MLP	0.47 ± 0.08	0.53 ± 0.04	0.52 ± 0.07	0.48 ± 0.03	0.51 ± 0.07
	GRU	0.24 ± 0.07	0.32 ± 0.10	0.27 ± 0.13	0.38 ± 0.09	0.17 ± 0.05
	ST	0.23 ± 0.12	0.25 ± 0.13	0.28 ± 0.07	0.37 ± 0.04	0.29 ± 0.10

T.6 TUDataset computer vision models

Appendix Table 25: Detailed metrics (mean ± standard deviation) for the TUDataset computer vision datasets. Only the MCC is reported as this is a multi-label task.

Dataset	Aggregator	GCN	GAT	GATv2	GIN	PNA
Cuneiform	Sum	0.51 ± 0.29	0.56 ± 0.25	0.66 ± 0.09	0.73 ± 0.07	0.58 ± 0.34
	Mean	0.00 ± 0.02	0.15 ± 0.25	-0.05 ± 0.04	0.80 ± 0.09	0.03 ± 0.06
	Max	0.26 ± 0.35	0.36 ± 0.33	0.26 ± 0.35	0.77 ± 0.12	0.55 ± 0.30
	MLP	0.50 ± 0.13	0.58 ± 0.17	0.62 ± 0.11	0.64 ± 0.09	0.64 ± 0.11
	GRU	0.18 ± 0.08	0.05 ± 0.09	0.11 ± 0.15	0.51 ± 0.12	0.16 ± 0.14
	ST	0.32 ± 0.19	0.21 ± 0.32	0.40 ± 0.25	0.47 ± 0.15	0.40 ± 0.31
COIL-RAG	Sum	0.91 ± 0.01	0.94 ± 0.02	0.94 ± 0.02	0.95 ± 0.02	0.90 ± 0.02
	Mean	0.89 ± 0.02	0.90 ± 0.02	0.93 ± 0.01	0.94 ± 0.02	0.90 ± 0.02
	Max	0.90 ± 0.02	0.91 ± 0.02	0.93 ± 0.01	0.94 ± 0.02	0.91 ± 0.02
	MLP	0.96 ± 0.01	0.96 ± 0.01	0.95 ± 0.01	0.94 ± 0.02	0.93 ± 0.02
	GRU	0.68 ± 0.11	0.76 ± 0.12	0.72 ± 0.04	0.85 ± 0.04	0.74 ± 0.07
	ST	0.82 ± 0.03	0.82 ± 0.04	0.84 ± 0.04	0.90 ± 0.02	0.82 ± 0.04
COIL-DEL	Sum	0.25 ± 0.04	0.50 ± 0.05	0.61 ± 0.06	0.49 ± 0.04	0.72 ± 0.03
	Mean	0.28 ± 0.02	0.42 ± 0.03	0.57 ± 0.03	0.40 ± 0.05	0.67 ± 0.05
	Max	0.30 ± 0.02	0.44 ± 0.03	0.56 ± 0.03	0.47 ± 0.05	0.70 ± 0.02
	MLP	0.42 ± 0.02	0.58 ± 0.04	0.61 ± 0.04	0.45 ± 0.03	0.55 ± 0.04
	GRU	0.15 ± 0.02	0.33 ± 0.06	0.38 ± 0.04	0.28 ± 0.03	0.23 ± 0.03
	ST	0.49 ± 0.03	0.31 ± 0.04	0.42 ± 0.08	0.40 ± 0.09	0.65 ± 0.03

T.7 ZINC regression models

Appendix Table 26: Detailed metrics for the ZINC dataset. The results are reported only for the provided train/validation/test splits (no custom random splits).

		GCN		GAT		GATv2		GIN		PNA	
Dataset	Agg.	MAE	R^2	MAE	R^2	MAE	R^2	MAE	R^2	MAE	R^2
ZINC	Sum	0.80	0.59	0.86	0.59	1.06	0.47	0.51	0.80	0.36	0.87
	Mean	0.74	0.63	0.90	0.54	1.26	0.43	0.57	0.77	0.39	0.86
	Max	0.88	0.57	1.25	0.50	1.34	0.39	0.64	0.72	0.47	0.86
	MLP	0.68	0.69	0.79	0.60	1.49	0.42	0.50	0.79	0.36	0.88
	GRU	0.92	0.51	1.46	0.06	0.91	0.67	0.56	0.79	0.33	0.89
	ST	0.72	0.64	1.04	0.47	2.39	0.42	0.53	0.79	0.43	0.85

T.8 QM9 regression models with Janossy neural aggregation

Appendix Table 27: Detailed metrics (mean ± standard deviation) for the QM9 dataset, including the two Janossy variants. Any differences in performance compared to other 2-layer models such as those in Figure 3 might be due to different GNN hyperparameters, such as the output or intermediate node dimension (QM9-specific experiments generally used larger dimensions).

		GCN		GAT		GATv2		GIN		PNA	
Data.	Agg.	MAE	R^2	MAE	R^2	MAE	R^2	MAE	R^2	MAE	R^2
QM9	Sum	0.74 ± 0.00	0.09 ± 0.00	0.76 ± 0.00	0.05 ± 0.01	0.75 ± 0.00	0.08 ± 0.01	0.71 ± 0.00	0.15 ± 0.00	0.70 ± 0.00	0.17 ± 0.01
	Mean	0.73 ± 0.00	0.10 ± 0.00	0.75 ± 0.01	0.07 ± 0.01	0.74 ± 0.00	0.10 ± 0.00	0.72 ± 0.00	0.13 ± 0.00	0.70 ± 0.00	0.16 ± 0.01
	Max	0.73 ± 0.00	0.11 ± 0.00	0.74 ± 0.00	0.10 ± 0.00	0.73 ± 0.00	0.11 ± 0.00	0.70 ± 0.00	0.16 ± 0.00	0.68 ± 0.00	0.20 ± 0.00
	MLP	0.63 ± 0.00	0.31 ± 0.00	0.64 ± 0.00	0.30 ± 0.01	0.64 ± 0.01	0.29 ± 0.01	0.60 ± 0.00	0.38 ± 0.00	0.58 ± 0.00	0.41 ± 0.00
	GRU	0.62 ± 0.01	0.34 ± 0.02	0.62 ± 0.01	0.34 ± 0.01	0.61 ± 0.02	0.35 ± 0.03	0.60 ± 0.00	0.38 ± 0.00	0.60 ± 0.01	0.37 ± 0.02
	ST	0.60 ± 0.01	0.38 ± 0.01	0.63 ± 0.02	0.30 ± 0.04	0.62 ± 0.01	0.32 ± 0.02	0.59 ± 0.00	0.39 ± 0.01	0.57 ± 0.01	0.44 ± 0.01
	Janossy MLP	0.67 ± 0.02	0.23 ± 0.04	0.68 ± 0.01	0.22 ± 0.03	0.67 ± 0.01	0.23 ± 0.02	0.61 ± 0.01	0.34 ± 0.02	0.61 ± 0.01	0.34 ± 0.02
	Janossy GRU	0.67 ± 0.02	0.22 ± 0.04	0.67 ± 0.01	0.23 ± 0.02	0.71 ± 0.02	0.15 ± 0.04	0.66 ± 0.03	0.25 ± 0.06	0.63 ± 0.01	0.32 ± 0.02

T.9 MalNetTiny models

Appendix Table 28: Detailed metrics (mean ± standard deviation) for the MalNetTiny dataset. 'OOM' stands for out-of-memory (RAM). Only the MCC is reported as this is a multi-label task.

Dataset	Aggregator	GCN	GAT	GATv2	GIN	PNA
MalNetTiny	Sum	0.81 ± 0.04	0.81 ± 0.05	0.80 ± 0.02	0.87 ± 0.03	OOM
	Mean	0.72 ± 0.02	0.73 ± 0.02	0.75 ± 0.05	0.88 ± 0.01	OOM
	Max	0.81 ± 0.02	0.77 ± 0.04	0.76 ± 0.04	0.89 ± 0.02	OOM
	MLP	0.80 ± 0.02	0.81 ± 0.01	0.80 ± 0.01	0.82 ± 0.02	OOM
	GRU	0.68 ± 0.03	0.69 ± 0.04	0.70 ± 0.02	0.70 ± 0.02	OOM
	ST	0.84 ± 0.03	0.82 ± 0.03	0.84 ± 0.02	0.88 ± 0.03	OOM

T.10 GNNBenchmark computer vision models

Appendix Table 29: Detailed metrics for the MNIST and CIFAR10 datasets. All models use the provided train/validation/test splits (no custom splits), which are used in five different runs and aggregated (mean ± standard deviation). Only the MCC is reported as this is a multi-label task.

Dataset	Aggregator	**GCN**	**GAT**	**GATv2**	**GIN**	**PNA**
MNIST	Sum	0.42 ± 0.01	0.35 ± 0.04	0.52 ± 0.03	0.51 ± 0.01	0.74 ± 0.00
	Mean	0.39 ± 0.01	0.29 ± 0.01	0.33 ± 0.07	0.46 ± 0.01	0.72 ± 0.01
	Max	0.25 ± 0.03	0.40 ± 0.05	0.43 ± 0.17	0.43 ± 0.01	0.71 ± 0.01
	MLP	0.28 ± 0.02	0.28 ± 0.01	0.26 ± 0.04	0.31 ± 0.01	0.47 ± 0.03
	GRU	0.36 ± 0.03	0.25 ± 0.04	0.37 ± 0.09	0.44 ± 0.02	0.66 ± 0.02
	ST	0.48 ± 0.01	0.60 ± 0.03	0.64 ± 0.02	0.56 ± 0.01	0.77 ± 0.00
CIFAR10	Sum	0.28 ± 0.01	0.32 ± 0.01	0.34 ± 0.03	0.31 ± 0.01	0.50 ± 0.00
	Mean	0.27 ± 0.00	0.29 ± 0.01	0.30 ± 0.01	0.30 ± 0.01	0.51 ± 0.00
	Max	0.29 ± 0.00	0.36 ± 0.00	0.35 ± 0.01	0.31 ± 0.00	0.46 ± 0.01
	MLP	0.17 ± 0.00	0.25 ± 0.00	0.19 ± 0.02	0.17 ± 0.00	0.29 ± 0.01
	GRU	0.24 ± 0.01	0.31 ± 0.02	0.29 ± 0.07	0.26 ± 0.01	0.45 ± 0.01
	ST	0.32 ± 0.01	0.42 ± 0.01	0.39 ± 0.01	0.35 ± 0.00	0.48 ± 0.00

T.11 TUDataset small molecules models

Appendix Table 30: Detailed metrics (mean ± standard deviation) for the AIDS, FRANKENSTEIN, MUTAG, and Mutagenicity datasets.

		GCN		GAT		GATv2		GIN		PNA	
Dataset	Agg.	AUROC	MCC	AUROC	MCC	AUROC	MCC	AUROC	MCC	AUROC	MCC
AIDS	Sum	0.97 ± 0.02	0.96 ± 0.03	0.97 ± 0.03	0.95 ± 0.04	0.97 ± 0.02	0.95 ± 0.02	0.96 ± 0.02	0.95 ± 0.02	0.97 ± 0.02	0.96 ± 0.02
	Mean	0.97 ± 0.02	0.95 ± 0.02	0.98 ± 0.02	0.96 ± 0.03	0.97 ± 0.02	0.96 ± 0.03	0.96 ± 0.02	0.95 ± 0.02	0.97 ± 0.02	0.96 ± 0.02
	Max	0.97 ± 0.02	0.96 ± 0.02	0.97 ± 0.02	0.96 ± 0.02	0.97 ± 0.02	0.94 ± 0.03	0.97 ± 0.01	0.95 ± 0.01	0.97 ± 0.02	0.96 ± 0.03
	MLP	1.00 ± 0.00	1.00 ± 0.01	1.00 ± 0.00	1.00 ± 0.01	1.00 ± 0.00	1.00 ± 0.01	1.00 ± 0.00	1.00 ± 0.01	1.00 ± 0.00	1.00 ± 0.01
	GRU	1.00 ± 0.00	1.00 ± 0.01	1.00 ± 0.00	1.00 ± 0.01	1.00 ± 0.00	1.00 ± 0.01	1.00 ± 0.01	0.99 ± 0.01	1.00 ± 0.00	0.99 ± 0.01
	ST	0.97 ± 0.02	0.96 ± 0.03	0.98 ± 0.02	0.96 ± 0.03	0.97 ± 0.02	0.95 ± 0.03	0.96 ± 0.02	0.89 ± 0.09	0.98 ± 0.02	0.97 ± 0.03
FRANKEN STEIN	Sum	0.62 ± 0.01	0.25 ± 0.02	0.60 ± 0.00	0.21 ± 0.02	0.59 ± 0.02	0.19 ± 0.03	0.60 ± 0.03	0.20 ± 0.06	0.65 ± 0.03	0.29 ± 0.06
	Mean	0.60 ± 0.03	0.21 ± 0.06	0.60 ± 0.02	0.20 ± 0.04	0.59 ± 0.03	0.19 ± 0.05	0.61 ± 0.04	0.22 ± 0.08	0.62 ± 0.03	0.25 ± 0.04
	Max	0.60 ± 0.02	0.20 ± 0.03	0.60 ± 0.02	0.20 ± 0.04	0.61 ± 0.03	0.21 ± 0.06	0.62 ± 0.02	0.23 ± 0.05	0.63 ± 0.03	0.25 ± 0.05
	MLP	0.61 ± 0.02	0.23 ± 0.03	0.63 ± 0.02	0.27 ± 0.05	0.65 ± 0.02	0.29 ± 0.05	0.65 ± 0.02	0.30 ± 0.05	0.68 ± 0.01	0.37 ± 0.03
	GRU	0.66 ± 0.03	0.32 ± 0.07	0.66 ± 0.01	0.32 ± 0.02	0.65 ± 0.02	0.29 ± 0.04	0.62 ± 0.03	0.23 ± 0.05	0.58 ± 0.03	0.17 ± 0.07
	ST	0.61 ± 0.03	0.22 ± 0.06	0.59 ± 0.02	0.17 ± 0.04	0.58 ± 0.02	0.16 ± 0.04	0.62 ± 0.02	0.24 ± 0.04	0.63 ± 0.02	0.26 ± 0.04
MUTAG	Sum	0.65 ± 0.18	0.30 ± 0.30	0.66 ± 0.17	0.33 ± 0.28	0.67 ± 0.16	0.36 ± 0.26	0.90 ± 0.08	0.75 ± 0.17	0.84 ± 0.13	0.67 ± 0.23
	Mean	0.68 ± 0.19	0.32 ± 0.30	0.69 ± 0.19	0.35 ± 0.32	0.70 ± 0.20	0.37 ± 0.34	0.90 ± 0.07	0.78 ± 0.14	0.79 ± 0.12	0.53 ± 0.19
	Max	0.71 ± 0.23	0.38 ± 0.42	0.72 ± 0.24	0.40 ± 0.43	0.65 ± 0.22	0.29 ± 0.40	0.91 ± 0.07	0.78 ± 0.14	0.90 ± 0.06	0.75 ± 0.18
	MLP	0.91 ± 0.12	0.84 ± 0.24	0.90 ± 0.15	0.80 ± 0.30	0.84 ± 0.07	0.67 ± 0.14	0.87 ± 0.13	0.71 ± 0.26	0.89 ± 0.11	0.75 ± 0.22
	GRU	0.81 ± 0.19	0.55 ± 0.32	0.78 ± 0.17	0.51 ± 0.29	0.87 ± 0.06	0.71 ± 0.09	0.84 ± 0.14	0.65 ± 0.26	0.88 ± 0.07	0.76 ± 0.10
	ST	0.85 ± 0.08	0.68 ± 0.17	0.85 ± 0.11	0.69 ± 0.18	0.83 ± 0.12	0.69 ± 0.15	0.89 ± 0.09	0.72 ± 0.19	0.83 ± 0.09	0.67 ± 0.13
Mutagen icity	Sum	0.79 ± 0.02	0.59 ± 0.03	0.76 ± 0.01	0.53 ± 0.02	0.79 ± 0.01	0.57 ± 0.02	0.81 ± 0.02	0.63 ± 0.04	0.82 ± 0.02	0.64 ± 0.05
	Mean	0.78 ± 0.02	0.56 ± 0.04	0.77 ± 0.01	0.55 ± 0.03	0.75 ± 0.01	0.52 ± 0.03	0.81 ± 0.01	0.61 ± 0.03	0.80 ± 0.03	0.60 ± 0.06
	Max	0.79 ± 0.02	0.57 ± 0.03	0.74 ± 0.01	0.50 ± 0.02	0.71 ± 0.01	0.44 ± 0.03	0.83 ± 0.01	0.65 ± 0.02	0.81 ± 0.02	0.62 ± 0.04
	MLP	0.73 ± 0.02	0.48 ± 0.03	0.77 ± 0.02	0.53 ± 0.03	0.76 ± 0.01	0.52 ± 0.02	0.79 ± 0.01	0.57 ± 0.02	0.78 ± 0.01	0.56 ± 0.02
	GRU	0.73 ± 0.01	0.47 ± 0.02	0.75 ± 0.01	0.49 ± 0.02	0.74 ± 0.01	0.48 ± 0.03	0.81 ± 0.02	0.61 ± 0.04	0.73 ± 0.02	0.46 ± 0.04
	ST	0.80 ± 0.02	0.59 ± 0.03	0.75 ± 0.04	0.52 ± 0.06	0.78 ± 0.03	0.57 ± 0.05	0.81 ± 0.01	0.63 ± 0.02	0.82 ± 0.01	0.64 ± 0.02
YeastH	Sum	0.54 ± 0.01	0.20 ± 0.02	0.54 ± 0.01	0.20 ± 0.02	0.54 ± 0.00	0.21 ± 0.02	0.57 ± 0.01	0.27 ± 0.03	0.58 ± 0.01	0.29 ± 0.02
	Mean	0.52 ± 0.00	0.16 ± 0.01	0.52 ± 0.00	0.17 ± 0.01	0.53 ± 0.00	0.17 ± 0.01	0.55 ± 0.01	0.24 ± 0.02	0.56 ± 0.01	0.25 ± 0.01
	Max	0.53 ± 0.00	0.17 ± 0.01	0.51 ± 0.00	0.14 ± 0.02	0.52 ± 0.00	0.16 ± 0.01	0.57 ± 0.01	0.26 ± 0.03	0.57 ± 0.01	0.25 ± 0.03
	MLP	0.60 ± 0.00	0.24 ± 0.01	0.60 ± 0.01	0.24 ± 0.01	0.60 ± 0.01	0.25 ± 0.02	0.61 ± 0.01	0.26 ± 0.02	0.60 ± 0.01	0.24 ± 0.02
	GRU	0.54 ± 0.02	0.17 ± 0.02	0.55 ± 0.01	0.19 ± 0.01	0.54 ± 0.01	0.19 ± 0.02	0.58 ± 0.02	0.26 ± 0.02	0.55 ± 0.01	0.17 ± 0.01
	ST	0.62 ± 0.00	0.31 ± 0.01	0.59 ± 0.02	0.30 ± 0.01	0.60 ± 0.01	0.30 ± 0.02	0.64 ± 0.01	0.34 ± 0.02	0.64 ± 0.01	0.35 ± 0.02

Appendix Table 31: Detailed metrics (mean ± standard deviation) for the alchemy_full dataset.

		GCN		GAT		GATv2		GIN		PNA	
Dataset	Agg.	MAE	R^2	MAE	R^2	MAE	R^2	MAE	R^2	MAE	R^2
alchemy_full	Sum	17.78 ± 0.48	0.99 ± 0.00	46.80 ± 2.39	0.98 ± 0.00	46.00 ± 4.02	0.98 ± 0.00	17.49 ± 1.04	0.99 ± 0.00	10.30 ± 0.42	1.00 ± 0.00
	Mean	28.91 ± 0.74	0.97 ± 0.00	56.22 ± 2.65	0.97 ± 0.00	49.94 ± 1.14	0.97 ± 0.00	41.46 ± 4.19	0.97 ± 0.00	16.71 ± 0.50	0.99 ± 0.00
	Max	18.90 ± 0.24	0.99 ± 0.00	78.62 ± 9.31	0.99 ± 0.01	67.37 ± 3.42	0.99 ± 0.00	31.51 ± 9.16	0.99 ± 0.00	13.80 ± 0.99	1.00 ± 0.00
	MLP	12.75 ± 0.74	1.00 ± 0.00	15.87 ± 1.33	1.00 ± 0.00	17.76 ± 1.88	1.00 ± 0.00	20.24 ± 7.66	0.99 ± 0.00	10.14 ± 1.75	1.00 ± 0.00
	GRU	9.25 ± 0.58	1.00 ± 0.00	16.16 ± 1.07	1.00 ± 0.00	15.23 ± 1.54	1.00 ± 0.00	13.35 ± 0.61	0.99 ± 0.00	7.24 ± 0.40	1.00 ± 0.00
	ST	9.58 ± 0.38	1.00 ± 0.00	19.52 ± 8.84	0.99 ± 0.01	15.11 ± 1.27	1.00 ± 0.00	13.83 ± 1.62	1.00 ± 0.00	10.25 ± 1.19	1.00 ± 0.00