Supplementary Information (SI) for Environmental Science: Processes & Impacts.
This journal is © The Royal Society of Chemistry 2025

Supplementary material

Hybrid Intelligence for Environmental Pollution: Biodegradability Assessment of Organic Compounds through Multimodal Integration of Graph Attention Networks and QSAR Models

Abbas Salimi[‡], *and Jin Yong Lee*[*,‡]

[‡] Department of Chemistry, Sungkyunkwan University, Suwon 16419, Korea;

Correspondence: Jinylee@skku.edu;

Table S1. Selected models after Cross-validation on 80% of the dataset for training and validation

Table S2. Selected models after Cross-validation on 80% of the dataset for training and validation

Figure S1. Confusion matrices representing the classification performance of four different GAT-based models (A, B, C, and D) on the test set for biodegradability prediction. Each matrix displays the breakdown of True Labels (rows) versus Predicted Labels (columns) for two classes: biodegradable (RB) and non-biodegradable (NRB)., A) GAT1, B) GAT2, C) GAT3, and D) GAT4.

Figure S2. Confusion matrices representing the classification performance of four different GAT-based models (A, B, C, and D) on the test set for biodegradability prediction. Each matrix displays the breakdown of True Labels (rows) versus Predicted Labels (columns) for two classes: biodegradable (RB) and non-biodegradable (NRB)., A) Random Forest, B) Gradient Boosting, and C) eXtreme Gradient Boosting.

Table S3. Weight-averaging ensemble models with grid search

Table S4. 5-fold Stacked models with grid search

Table S1. Selected models after Cross-validation on 80% of the dataset for training and validation

Hidden size	Number layers	Drop out	Batch size	Epoch	lr	Test accuracy	Precision	F1	AUC	Sp	Sn	ER
128	3	0.3	16	400	0.0001	0.84 +/- 0.02	0.80 +/- 0.04	0.79 +/- 0.03	85 +/- 0.02	0.88 +/- 0.03	0.78 +/- 0.02	0.16 +/- 0.02
256	3	0.3	16	400	0.0001	0.84 +/- 0.02	0.81 +/- 0.01	0.79 +/- 0.03	0.85 +/- 0.02	0.89 +/- 0.01	0.78 +/- 0.04	0.16 +/- 0.02
256	4	0.3	16	400	0.0001	0.85 +/- 0.02	0.85 +/- 0.04	0.79 +/- 0.02	0.84 +/- 0.02	0.92 +/- 0.03	0.75 +/- 0.05	0.15 +/- 0.02
512	4	0.3	32	400	0.0001	0.84 +/- 0.02	0.83 +/- 0.04	0.78 +/- 0.02	0.84 +/- 0.02	0.91 +/- 0.02	0.74 +/- 0.02	0.16 +/- 0.02

Table S2. Selected models after Cross-validation on 80% of the dataset for training and validation

Model	Estimators	lr	Max depth	Min_sample leaf	Subsample	Accuracy	Precision	F1	AUC	Sp	Sn	ER
RF	100	-	--	5	--	0.84 +/- 0.02	0.80 +/- 0.04	0.78 +/- 0.02	0.82 +/- 0.02	0.88 +/- 0.03	0.77 +/- 0.03	0.16 +/- 0.02
GB	100	0.1	5		0.7	0.85 +/- 0.02	0.83 +/- 0.03	0.79 +/- 0.02	0.83 +/- 0.02	0.90 +/- 0.02	0.76 +/- 0.03	0.15 +/- 0.02
XGB	100	0.1	8		0.8	0.85 +/- 0.01	0.81 +/- 0.03	0.79 +/- 0.02	0.83 +/- 0.02	0.89 +/- 0.02	0.78 +/- 0.03	0.15 +/- 0.01

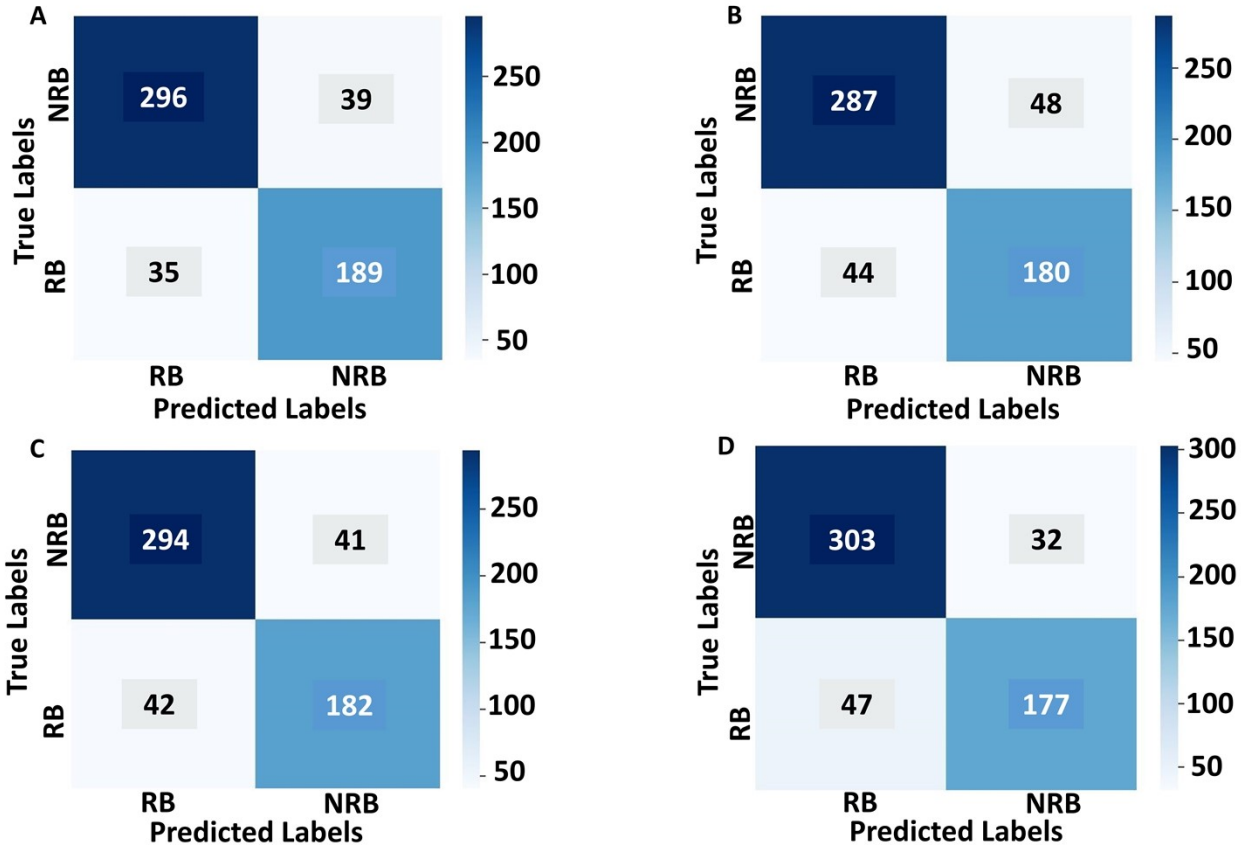

Figure S1. Confusion matrices representing the classification performance of four different GAT-based models (A, B, C, and D) on the test set for biodegradability prediction. Each matrix displays the breakdown of True Labels (rows) versus Predicted Labels (columns) for two classes: biodegradable (RB) and non-biodegradable (NRB)., A) GAT1, B) GAT2, C) GAT3, and D) GAT4.

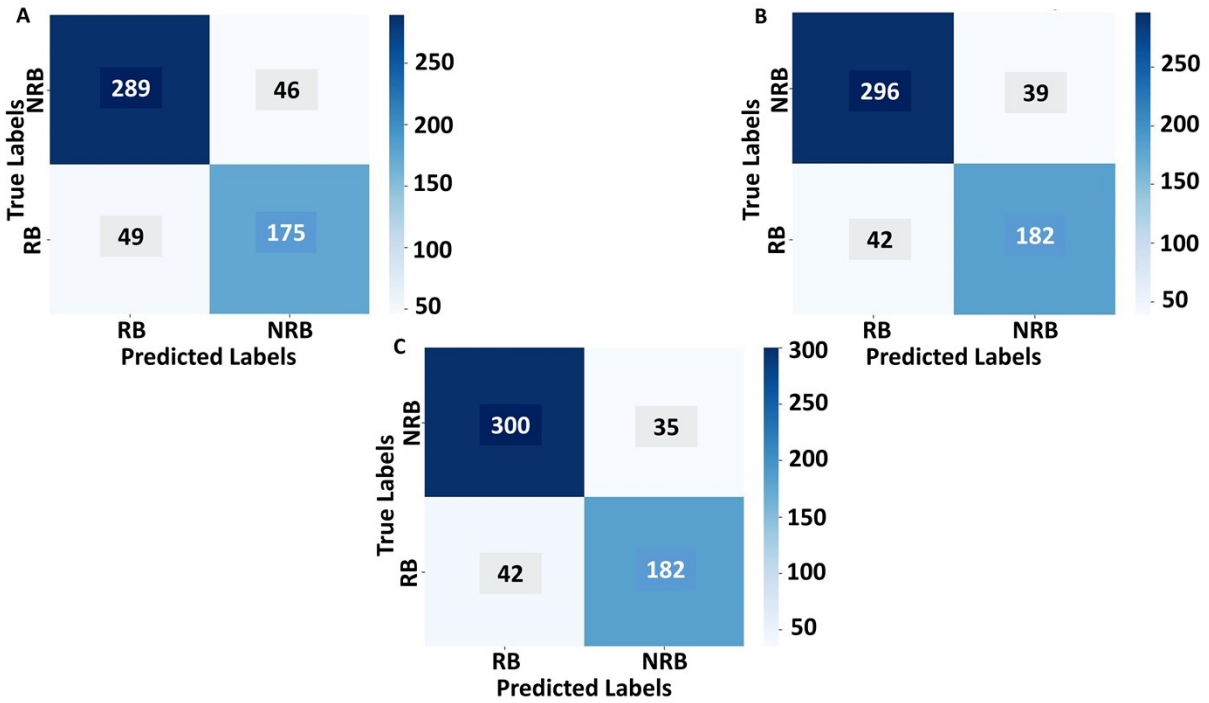

Figure S2. Confusion matrices representing the classification performance of four different GAT-based models (A, B, C, and D) on the test set for biodegradability prediction. Each matrix displays the breakdown of True Labels (rows) versus Predicted Labels (columns) for two classes: biodegradable (RB) and non-biodegradable (NRB)., A) Random Forest, B) Gradient Boosting, and C) eXtreme Gradient Boosting.

Table S3. Weight-averaging ensemble models with grid search

GAT	QSAR	Best Weights	Best Accuracy	Best Specificity (Sp)	Best Sensitivity (Sn)	Best F1 Score	Best Recall	Best AUC	Best precision
1	RF	[0.30, 0.70]	0.84	0.84	0.84	0.81	0.84	0.93	0.77
1	GB	[0.4, 0.6]	0.85	0.86	0.84	0.82	0.84	0.93	0.80
1	XGB	[0.2, 0.8]	0.86	0.89	0.82	0.83	0.82	0.93	0.84
2	RF	[0.2, 0.8]	0.84	0.83	0.84	0.81	0.84	0.93	0.77
2	GB	[0.4, 0.6]	0.85	0.86	0.85	0.82	0.85	0.93	0.80
2	XGB	[0.0, 1.0]	0.86	0.89	0.81	0.82	0.81	0.93	0.84
3	RF	[0.2, 0.8]	0.84	0.84	0.83	0.81	0.83	0.93	0.78
3	GB	[0.4, 0.6]	0.86	0.86	0.85	0.83	0.85	0.93	0.80
3	XGB	[0.2, 0.8]	0.86	0.89	0.82	0.83	0.82	0.93	0.83
4	RF	[0.30, 0.70]	0.85	0.84	0.86	0.82	0.86	0.93	0.78
4	GB	[0.4, 0.6]	0.85	0.86	0.85	0.82	0.85	0.92	0.80
4	XGB	[0.30, 0.70]	0.86	0.88	0.83	0.83	0.83	0.93	0.83

Table S4. 5-fold Stacked models with grid search

GAT	QSAR	Accuracy	Specificity (Sp)	Sensitivity (Sn)	F1 Score	Recall	AUC	precision	Parameters
1	RF	0.86 ± 0.02	0.89 ± 0.03	0.83 ± 0.04	0.83 ± 0.02	0.83 ± 0.04	0.94 ± 0.01	0.83 ± 0.03	{'C': 1, 'penalty': 'l1', 'solver': 'liblinear'}
1	GB	0.87 ± 0.02	0.90 ± 0.04	0.81 ± 0.02	0.83 ± 0.02	0.81 ± 0.02	0.93 ± 0.01	0.85 ± 0.05	{'C': 1, 'penalty': 'l1', 'solver': 'liblinear'}
1	XGB	0.87± 0.02	0.90 ± 0.03	0.81 ± 0.05	0.83 ± 0.03	0.82± 0.05	0.93 ± 0.01	0.85 ± 0.04	'C': 1, 'penalty': 'l1', 'solver': 'liblinear'
2	RF	0.84± 0.02	0.88 ± 0.04	0.83 ± 0.04	0.83 ± 0.02	0.83 ± 0.04	0.94 ± 0.01	0.84 ± 0.04	{'C': 10, 'penalty': 'l1', 'solver': 'liblinear'}
2	GB	0.87 ± 0.02	0.89 ± 0.03	0.80 ± 0.03	0.82 ± 0.02	0.80 ± 0.03	0.93 ± 0.02	0.84 ± 0.03	{'C': 0.1, 'penalty': 'l1', 'solver': 'liblinear'}
2	XGB	0.85 ± 0.03	0.89 ± 0.02	0.80 ± 0.06	0.81 ± 0.04	0.80 ± 0.06	0.93 ± 0.02	0.83 ± 0.03	{'C': 0.1, 'penalty': 'l1', 'solver': 'liblinear'}
3	RF	0.87 ± 0.02	0.91 ± 0.05	0.80 ± 0.04	0.83 ± 0.01	0.80 ± 0.04	0.94 ± 0.01	0.87 ± 0.05	{'C': 1, 'penalty': 'l1', 'solver': 'liblinear'}

3	**GB**	**0.86 ± 0.01**	**0.92 ± 0.04**	**0.79 ± 0.04**	**0.82 ± 0.01**	**0.79 ± 0.04**	**0.93 ± 0.01**	**0.87 ± 0.05**	**{'C': 1, 'penalty': 'l2', 'solver': 'liblinear'}**
3	**XGB**	**0.87 ± 0.02**	**0.91 ± 0.05**	**0.81 ± 0.05**	**0.83 ± 0.03**	**0.81 ± 0.05**	**0.93 ± 0.01**	**0.86 ± 0.06**	**{'C': 10, 'penalty': 'l2', 'solver': 'liblinear'}**
4	**RF**	**0.87 ± 0.02**	**0.91 ± 0.04**	**0.82 ± 0.05**	**0.83 ± 0.02**	**0.82 ± 0.05**	**0.94 ± 0.01**	**0.86 ± 0.04**	**{'C': 10, 'penalty': 'l2', 'solver': 'liblinear'}**
4	**GB**	**0.87 ± 0.02**	**0.91 ± 0.03**	**0.79 ± 0.05**	**0.83 ± 0.03**	**0.79 ± 0.05**	**0.93 ± 0.02**	**0.87 ± 0.04**	**{'C': 10, 'penalty': 'l2', 'solver': 'liblinear'}**
4	**XGB**	**0.87 ± 0.03**	**0.92 ± 0.03**	**0.79 ± 0.06**	**0.83 ± 0.04**	**0.79 ± 0.06**	**0.93± 0.01**	**0.86 ± 0.04**	**{'C': 1, 'penalty': 'l2', 'solver': 'liblinear'}**