EEET ECOLOGICAL ENGINEERING & ENVIRONMENTAL TECHNOLOGY

Ecological Engineering & Environmental Technology, 2026, 27(2), 187–197
https://doi.org/10.12912/27197050/216741
ISSN 2719–7050, License CC-BY 4.0

Received: 2026.01.05
Accepted: 2026.01.29
Published: 2026.02.07

Prediction of pharmaceutical residue presence in aquatic systems using graph-based deep learning models

Ayoub Belaidi[1], Saida Ait Boughrous[2], Rachid El Ayachi[1], Mohamed Biniz[3], Mohamed Oubezza[4]

[1] Department of Computer Science, Faculty of Sciences and Technology, Sultan Moulay Slimane University, Campus Mghilla, Beni Mellal, 23000, Morocco
[2] Ethnopharmacology and Pharmacognosy, Faculty of Sciences and Techniques Errachidia, Moulay Ismail University of Meknes, BP 509, Boutalamine, Errachidia 52000, Morocco
[3] Department of Computer Science, Faculty of Polydisciplinary, Sultan Moulay Slimane University, Beni Mellal, 23000, Morocco
[4] Department of Computer Science, Faculty of Sciences, Chouaib Doukkali University, El Jadida, Morocco
* Corresponding author's e-mail: belaidiayoub7@gmail.com

ABSTRACT

Pharmaceutical residues discharged into aquatic systems constituted an emerging environmental threat and posed considerable challenges to conventional monitoring strategies. Analytical methods such as LC-MS/MS, although precise, remained costly, time-consuming, and unsuitable for large-scale continuous monitoring. The objective of this study was to develop a classification model based on deep learning to predict the presence or absence of pharmaceutical residues in water samples, using both molecular characteristics and environmental parameters. A dataset collected from various aquatic environments (rivers, wastewater treatment plant effluents, groundwater) was filtered, annotated, and transformed into a binary classification set where the target value corresponded to the detection (1) or non-detection (0) of the pharmaceutical product. The molecular structures were converted into atomic graphs using RDKit, allowing the use of three advanced models: graph neural network (GNN), graph attention network (GAT), and message passing neural network (MPNN). Contextual information (matrix, therapeutic group, analyte type, location, and sampling period) was integrated in addition to the molecular representations. Graph-based models have produced solid performances. The MPNN achieved the best scores with an accuracy of 92.8%, an F1-score of 0.92, and an AUC of 0.96. The GAT achieved 90.3% accuracy, 0.90 F1-score, and 0.94 AUC, while the GNN obtained 84.2%, 0.89, and 0.84 respectively. The integration of molecular features and environmental metadata improved performance by more than 12% compared to models using only molecular representations. The performance remained influenced by class imbalance, regional variability, and the incomplete nature of certain environmental variables. This approach has not replaced instrumental analyzes, but has constituted a promising complementary tool. It has helped reduce the exclusive reliance on analytical measurements and more effectively guide water monitoring. To our knowledge, this is one of the first studies simultaneously integrating molecular graphs and environmental metadata for the binary prediction of pharmaceutical contamination in natural waters.

Keywords: pharmaceutical pollutants, water quality monitoring, molecular graphs, graph neural networks, graph attention networks, message passing neural networks.

INTRODUCTION

Pharmaceutical chemicals are now extensively identified in global aquatic habitats, rendering them a significant category of developing environmental pollutants (Mohapatra et al., 2025; Hanafiah et al., 2025). These compounds, encompassing antibiotics, anti-inflammatory agents, analgesics, and hormones, are persistently introduced into aquatic environments by home effluent, hospital discharges, agricultural practices, and inadequate elimination by

wastewater treatment facilities. Consequently, tiny amounts of medicines have been detected in rivers, groundwater, and drinking water, prompting concerns regarding ecological integrity, antimicrobial resistance, and potential long-term impacts on human health (Monk et al., 2025; Aziz et al., 2025).

Beyond their chemical persistence, pharmaceutical residues raise major biological concerns due to their intended bioactivity. Numerous studies have demonstrated that chronic exposure to low environmental concentrations of pharmaceuticals can adversely affect aquatic organisms, including fish, invertebrates, algae, and microbial communities (Domínguez-García et al., 2024; Mazhandu and Mashifana, 2024; Mheidli et al., 2022). Reported biological effects include endocrine disruption, behavioral and reproductive alterations, oxidative stress, and the development of antimicrobial resistance. Even when present at trace levels, continuous exposure to complex mixtures of pharmaceutical compounds may compromise ecosystem functioning and aquatic biodiversity, highlighting the importance of preventive monitoring strategies capable of identifying contamination before irreversible ecological impacts occur (Muambo et al., 2024; Belle et al., 2025; Aib et al., 2025).

Contemporary surveillance of pharmaceutical residues mostly relies on sophisticated analytical methodologies, notably liquid chromatography in conjunction with mass spectrometry. While these procedures yield dependable and precise measurements, they necessitate costly instrumentation, proficient operators, and comprehensive sample preparation (Eapen et al., 2024; Wada and Olawade, 2025). As a result, their utilization is frequently restricted to focused campaigns and certain locales, rendering continual and extensive monitoring challenging to accomplish (Coderre et al., 2025; Paíga et al., 2025; Ngoetjana et al., 2025). In addition, pharmaceutical pollution demonstrates substantial regional and temporal variability, determined by factors such as consumption patterns, hydrological conditions, seasonal fluctuations, and wastewater treatment performance. These limits highlight the need for additional approaches that can enhance monitoring programs and help focus analytical efforts more efficiently (Sanusi et al., 2023; Ngqwala and Muchesa, 2020; Ashfield et al., 2025). In this context, data-driven methodologies, particularly machine learning techniques, have attracted significant attention in environmental sciences. Machine learning algorithms are capable of learning complicated correlations from heterogeneous datasets and have been applied to numerous water-related problems, including water quality assessment, pollutant occurrence, and environmental risk evaluation (Kayani, 2025; Aira et al., 2022; Khan et al., 2025; Côrtes et al., 2025). However, most existing applications focus on descriptive analysis or rely on conventional physicochemical descriptors. Predictive studies addressing the presence or absence of pharmaceutical residues remain relatively limited, and the combined use of molecular features and environmental context has not been sufficiently investigated (Cano and Radjenovic, 2024; Padhy et al., 2024).

From a chemical standpoint, pharmaceutical compounds are naturally arranged like graphs, where atoms form nodes and chemical bonds form edges. Graph-based deep learning algorithms are specifically intended to process such data by propagating information across chemical structures (Maraj et al., 2025; Coderre et al., 2025). Architectures such as graph neural networks, graph attention networks, and message passing neural networks have proven great skills in recording molecular interactions and predicting chemical attributes in cheminformatics and drug discovery. Despite its potential, the use of these models to environmental contamination prediction, particularly for pharmaceuticals in aquatic systems, is still limited (Sanusi et al., 2023; Pereira et al., 2021).

To overcome this gap, this paper presents a graph-based deep learning framework to predict the presence or absence of pharmaceutical residues in water samples. The suggested approach integrates molecular graph representations obtained from chemical structures with environmental and contextual information, including water matrix type, therapeutic class, analyte type, geographical location, and sample period. By evaluating and comparing three advanced architectures GNN, GAT, and MPNN using a real-world dataset provided by the German Environment Agency, this work demonstrates that integrating molecular and environmental features improves predictive performance and provides a robust complementary tool for water quality monitoring and environmental risk assessment.

METHODS

Software environment and implementation details

All experiments were implemented in Python 3.10 using Jupyter Notebook. Molecular graphs were generated with RDKit (version 2023.03.2), and all graph-based deep learning models (GNN, GAT, and MPNN) were implemented using PyTorch (version 2.2) and PyTorch Geometric (version 2.5). Data preprocessing, including filtering, encoding of categorical variables, and train/validation/test splitting, was performed with pandas and scikit-learn. All analyses were executed on a workstation equipped with an NVIDIA GPU, which enabled efficient training of the graph neural network models.

Dataset

Figure 1 presents a screenshot of the raw spreadsheet data from the Umweltbundesamt database, illustrating the diversity of pharmaceutical analytes, therapeutic groups, and sampling contexts (matrices from WWTP effluent to sediments and groundwater, spanning multiple sampling periods from 2010 to 2017).

The dataset used in this study comes from the "Pharmaceuticals in the Environment" database of the Umweltbundesamt (2021), which serves as comprehensive source for analyzing the presence of pharmaceutical residues in various aquatic environments. This database contains several thousand rows, each corresponding to a specific sample associated with a particular pharmaceutical analyte. The included compounds cover a wide range of therapeutic classes, such as glucocorticoids, antibiotics, anti-inflammatories, and other commonly used substances. For each sample, several descriptive variables are available, including the type of matrix (treated wastewater, rivers, drinking water, etc.), the therapeutic group, the type of analyte (original substance or transformation product), the CAS number, the measured concentration, the sampling period, and the geographical location. The target variable is defined in a binary manner: if the measured concentration is strictly greater than zero, the analyte is considered present (y = 1), whereas if the concentration is equal to zero or indicated as "not detected," it is considered absent (y = 0). This approach transforms the problem into a binary classification, suitable for the application of deep learning models. The use of this dataset allows for linking contextual and molecular information to the probability of detecting pharmaceutical residues, thus providing a solid foundation for developing a predictive model capable of improving environmental monitoring and guiding analytical efforts more effectively. Figure 1 presents a screenshot of the raw spreadsheet data from the Umweltbundesamt database, illustrating the diversity of pharmaceutical analytes, therapeutic groups, and sampling contexts (matrices from WWTP effluent to sediments and groundwater, spanning multiple sampling periods from 2010 to 2017).

1	ID				Matrix			
2	ID	Name of Analyte	Target group	Type of Analyte	Matrix	Sampling Location	Sampling Period Start	Sampling Period End
3	ID	Name of Analyte	Target group	Type of Analyte	Matrix	Sampling Location	Sampling Period Start	Sampling Period End
177165	232069	Norfloxacin	Human	Parent	Sewage hospital (untreated)	Hospital of Tumaco	2017	2017
177166	232756	Norfloxacin	Human	Parent	WWTP inflow (untreated)	WWTP Weining Count	2016	2016
177167	232757	Norfloxacin	Human	Parent	WWTP effluent (treated)	WWTP Weining Count	2016	2016
177168	232758	Norfloxacin	Human	Parent	WWTP sludge	WWTP Weining Count	2016	2016
177169	233655	Norfloxacin	Human	Parent	Leachate	Xi'an, Qicungou Landfi	-9999	-9999
177170	233656	Norfloxacin	Human	Parent	Leachate	Guiyang , Gaoyan Lanc	-9999	-9999
177171	233657	Norfloxacin	Human	Parent	Leachate	Nanjing, Shuige Landfi	-9999	-9999
177172	233658	Norfloxacin	Human	Parent	Leachate	Suzhou, Qizishan Land	-9999	-9999
177173	233659	Norfloxacin	Human	Parent	Leachate	Shanghai, Laogang Lan	-9999	-9999
177174	233660	Norfloxacin	Human	Parent	Leachate	Hangzhou, Tianzilin La	-9999	-9999
177175	233661	Norfloxacin	Human	Parent	Leachate	Shenzhen, Xiaping Lan	-9999	-9999
177176	235782	Norfloxacin	Human	Parent	Sediment - River/Stream	Charmoise River, upst	2010	2011
177177	235783	Norfloxacin	Human	Parent	Sediment - River/Stream	Charmoise River, dowr	2010	2011
177178	235784	Norfloxacin	Human	Parent	Sediment - River/Stream	Charmoise River, far d	2010	2011
177179	241671	Norfloxacin	Human	Parent	WWTP effluent (treated)	WWTP in Bucharest	2017	2017
177180	241672	Norfloxacin	Human	Parent	WWTP effluent (treated)	WWTP in Cluj-Napoca	2017	2017
177181	241673	Norfloxacin	Human	Parent	WWTP effluent (treated)	WWTP in Sabac	2017	2017
177182	241674	Norfloxacin	Human	Parent	WWTP effluent (treated)	WWTP in Zagreb	2017	2017
177183	241675	Norfloxacin	Human	Parent	WWTP effluent (treated)	WWTP in Varazdin	2017	2017
177184	241676	Norfloxacin	Human	Parent	WWTP effluent (treated)	WWTP in Ljubljana	2017	2017
177185	241677	Norfloxacin	Human	Parent	WWTP effluent (treated)	WWTP in Budapest	2017	2017

Figure 1. Screenshot of the raw Umweltbundesamt (UBA) pharmaceuticals in the environment dataset as displayed in a spreadsheet application, showing the initial rows with key variables including analyte name (e.g., "Norfloxacin"), target group (therapeutic classification), analyte type (parent substance), matrix type (WWTP effluent, sediment, etc.), sampling location, and sampling period (years 2010–2017), confirming the heterogeneous data structure for the presence/absence prediction task

Filtering and pre-processing

Before training the classification models, the dataset was carefully filtered and preprocessed to ensure the quality and consistency of the information used (Figure 2). Samples with critical missing values, particularly for concentration or matrix type, were excluded to minimize biases in the learning process. Non-numeric values or those marked as "not detected" were transformed into zero to match the binary definition of the target variable. Categorical variables, such as the matrix, therapeutic group, and analyte type, were encoded using techniques suitable for deep learning, allowing the models to correctly process contextual information. Furthermore, the molecular structures of the analytes were converted into atomic graphs using RDKit, facilitating the application of graph neural network (GNN), graph attention network (GAT), and message passing neural network (MPNN) models. Finally, a consistency check was performed to ensure that each line retains the necessary information for prediction, including both molecular characteristics and environmental variables. This preprocessing ensures that the model receives standardized and reliable data, thereby maximizing predictive performance and reducing the impact of outliers or missing data. As shown in Figure 3, the preprocessing pipeline includes filtering missing values, transforming concentrations into binary targets, and encoding categorical variables (matrix, therapeutic group, analyte type) using pandas and scikit-learn

Data partitioning

To train and evaluate the classification models, the dataset was divided into three distinct subsets: train, validation, and test. The adopted distribution was 70% for training, 15% for validation, and 15% for testing, ensuring that the models learn from a wide diversity of examples while allowing for a robust evaluation of their performance. This separation was carried out randomly but stratified according to the target variable, in order to maintain the proportion of presence and absence of analytes in each subset. This strategy prevents biases related to an imbalanced distribution of classes between the sets and allows for obtaining evaluation metrics that are representative of the model's actual performance on unseen data. Figure 4 illustrates the stratified splitting procedure, ensuring class balance across train (70%), validation (15%), and test (15%) sets to prevent biases in model evaluation.

Managing class imbalance

The dataset exhibits a notable imbalance between the classes, with generally a greater number of samples corresponding to the absence of pharmaceutical residues compared to their presence. To mitigate this effect and improve learning, several techniques were employed. First, class weighting was applied during training, giving more weight to examples from the minority class. Secondly, oversampling techniques were explored to artificially increase the number of examples in the minority class, while avoiding the generation of exact duplications that could lead to overfitting. Finally, metrics suitable for evaluating performance in the presence of imbalance, such as the F1-score and AUC, were used for a more reliable interpretation of the results. These measures ensure that predictive models remain sensitive to the detection of pharmaceutical residues even when they appear rarely in the dataset.

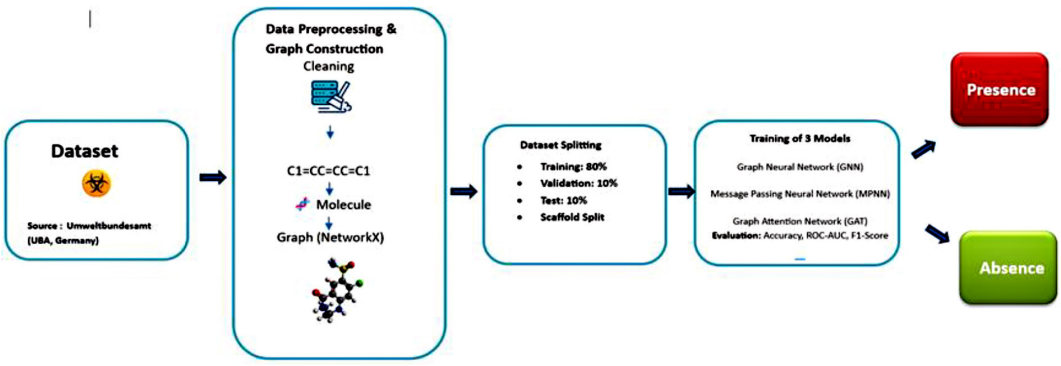

Figure 2. End-to-end workflow for molecular graph generation and presence/absence prediction

Detailed model architecture

Three advanced graph-based deep learning architectures have been explored: the graph neural network (GNN), the graph attention network (GAT), and the message passing neural network (MPNN). These models allow the propagation of information between the nodes of the graphs to learn rich and discriminative molecular representations. In the GNN, the representations of the nodes are updated by combining the information from immediate neighbors. The GAT introduces an attention mechanism that allows for differently weighting the contributions of neighboring nodes according to their importance. The MPNN extends this idea by using messages passed between nodes to capture more complex chemical interactions. Environmental and contextual features (matrix, therapeutic group, analyte type, season, and location) were integrated as additional features concatenated to the molecular representations before the final classification layer. This hybrid architecture allows the model to effectively combine chemical and environmental information to accurately predict the presence or absence of pharmaceutical residues in water samples.

Figures 5 and 6 show the exact implementation of molecular graph construction using RDKit and the hybrid MPNN architecture that concatenates graph-level molecular embeddings with encoded environmental context features.

```python
import pandas as pd
# Load raw dataset from Umweltbundesamt (UBA)
df = pd.read_csv("uba_pharmaceuticals_environment.csv")

# Keep relevant columns
cols = [
    "Matrix", "Therapeutic_group", "Analyte_type",
    "CAS_number", "Concentration", "Sampling_period",
    "Location", "SMILES"
]
df = df[cols]

# Define binary target: presence (1) vs absence (0)
df["y"] = (df["Concentration"] > 0).astype(int)

# Drop rows with critical missing values
df = df.dropna(subset=["Concentration", "Matrix", "SMILES"])

df.head()
```

Figure 3. The Python script in Visual Studio Code used to load the Umweltbundesamt (UBA) dataset, select the main variables, define the binary target (presence vs absence), and remove records with critical missing values

```python
# Categorical features to encode
cat_features = ["Matrix", "Therapeutic_group", "Analyte_type", "Location"]
enc = OneHotEncoder(handle_unknown="ignore", sparse_output=False)

X_cat = enc.fit_transform(df[cat_features])

# Molecular SMILES kept separately for graph construction
smiles_list = df["SMILES"].tolist()

y = df["y"].values

# Stratified train/val/test split (70/15/15)
X_train, X_temp, y_train, y_temp, smiles_train, smiles_temp = train_test_split(
    X_cat, y, smiles_list, test_size=0.30, stratify=y, random_state=42
)

X_val, X_test, y_val, y_test, smiles_val, smiles_test = train_test_split(
    X_temp, y_temp, smiles_temp, test_size=0.50, stratify=y_temp, random_state=42
)
```

Figure 4. The Visual Studio Code script performing one-hot encoding of the contextual variables and the stratified 70/15/15 train/validation/test split according to the presence/absence target

```python
from rdkit import Chem

def smiles_to_mol(smiles):
    mol = Chem.MolFromSmiles(smiles)
    return mol

# Example: convert first few analytes to RDKit molecules
mols = [smiles_to_mol(s) for s in smiles_train[:5]]
mols[0]
```

Figure 5. The Python code in Visual Studio Code demonstrating how SMILES strings are converted into RDKit molecular objects and then into PyTorch Geometric graph data structures (nodes = atoms, edges = chemical bonds)

```python
import torch.nn as nn
from torch_geometric.nn import MessagePassing, global_mean_pool # type: ignore

class MPNN(nn.Module):
    def __init__(self, node_dim, context_dim):
        super().__init__()
        self.conv1 = MessagePassing(aggr="add")
        self.conv2 = MessagePassing(aggr="add")
        self.fc_context = nn.Linear(context_dim, 64)
        self.fc1 = nn.Linear(node_dim + 64, 128)
        self.fc2 = nn.Linear(128, 1)
        self.dropout = nn.Dropout(p=0.3)

    def forward(self, data, context):
        x, edge_index, batch = data.x, data.edge_index, data.batch
        # Message passing layers (pseudo-code style)
        x = self.conv1.propagate(edge_index, x=x)
        x = self.conv2.propagate(edge_index, x=x)
        x = global_mean_pool(x, batch)

        c = torch.relu(self.fc_context(context))
        h = torch.cat([x, c], dim=1)
        h = torch.relu(self.fc1(h))
        h = self.dropout(h)
        out = torch.sigmoid(self.fc2(h))
        return out
```

Figure 6. The PyTorch geometric implementation of the MPNN model in Visual Studio Code, illustrating how molecular graph embeddings are propagated, pooled, and fused with contextual environmental features before binary classification

Training hyperparameters

The models were trained using the Adam optimizer with an initial learning rate of 0.001, chosen to ensure stable convergence while avoiding oscillations. The batch size was set to 32 samples, allowing a balance between gradient stability and training speed. Each model was trained for 200 epochs, with early stopping based on validation loss to prevent overfitting. ReLU activation functions were used in all intermediate layers, and the sigmoid was applied to the output layer to generate binary probabilities. To regularize the learning, a dropout of 0.3 was applied to the fully connected layers, and batch normalization techniques were employed to accelerate convergence and stabilize the gradients. These hyperparameters were optimized thru cross-validation to maximize performance metrics on the validation set. Figure 7 demonstrates the complete training configuration, including class weighting to address imbalance, dropout (0.3), ReLU activations, and early stopping based on validation loss to prevent overfitting.

EXPERIMENTAL RESULTS AND DISCUSSION

Quantitative results

The performance of the three graph-based architectures (GNN, MPNN, and GAT) was evaluated using the classical binary classification metrics: accuracy, AUC-ROC, and F1-score. These

```python
import torch.optim as optim
from sklearn.utils.class_weight import compute_class_weight

# Compute class weights for imbalance
class_weights = compute_class_weight(
    class_weight="balanced",
    classes=[0, 1],
    y=y_train
)
class_weights = torch.tensor(class_weights, dtype=torch.float).to(device)

model = MPNN(node_dim=1, context_dim=X_train.shape[1]).to(device)
optimizer = optim.Adam(model.parameters(), lr=1e-3)
criterion = nn.BCELoss(weight=None)  # or custom weighting

num_epochs = 200
best_val_loss = float("inf")

for epoch in range(num_epochs):
    model.train()
    # training_step(...)  # pseudo-code
    # compute train_loss, val_loss, metrics
    print(f"Epoch {epoch+1}: train_loss={train_loss:.4f}, val_loss={val_loss:.4f}")  # type: ignore
```

Figure 7. The training script in Visual Studio Code showing the configuration of the MPNN training loop, including the Adam optimizer (learning rate = 0.001), batch size of 32, a maximum of 200 epochs, and early stopping based on validation loss

indicators respectively allow for the evaluation of overall accuracy, the model's discriminative capacity, and the balance between precision and recall. The results obtained on the training, validation, and test sets are presented in Tables 1, 2, and 3. Although all the models demonstrated an ability to learn relevant molecular representations, notable differences in terms of generalization were observed.

Analyze

On the test set, the MPNN architecture records the best overall performance (accuracy = 0.928; AUC-ROC = 0.96), confirming its predictive stability and its ability to generalize across varied aquatic matrices. The GAT model comes in second place (accuracy = 0.903; AUC-ROC = 0.94) and stands out for its attention mechanism, which offers superior interpretability by identifying the most influential substructures in the prediction.

The GNN shows decent performance but remains inferior to the other two architectures (accuracy = 0.842), which suggests that more sophisticated message propagation mechanisms are necessary to adequately model the behavior of pharmaceutical molecules in the aquatic environment. Overall, these results demonstrate that message passing architectures particularly MPNN are particularly well-suited for predicting the presence of pharmaceutical residues in water samples (Table 4).

Visual analysis

The analysis of the confusion matrices presented in Figure 8 allows for a detailed comparison of the performance of the three graph-based architectures for the binary classification of "Presence / Absence" of pharmaceutical residues. The MPNN model shows the highest performance, with a significant number of true

Table 1. Performance on the training set (presence of pharmaceutical products)

Model	Accuracy	ROC-AUC	F1-score
GNN	0.901	0.92	0.89
MPNN	0.947	0.97	0.94
GAT	0.931	0.96	0.92

Table 2. Performance on the validation set (presence of pharmaceutical products)

Model	Accuracy	ROC-AUC	F1-score
GNN	0.873	0.90	0.86
MPNN	0.917	0.95	0.91
GAT	0.904	0.94	0.90

Table 3. Performance on the test set (presence of pharmaceutical products)

Model	Accuracy	ROC-AUC	F1-score
GNN	0.842	0.89	0.84
MPNN	0.928	0.96	0.92
GAT	0.903	0.94	0.90

Table 4. Comparison with recent state-of-the-art approaches for pharmaceutical occurrence prediction

Study / Modèle	Type of task	Modèle / Algo	Accuracy	AUC-ROC	Notes
This study	Binary classification of pharm residues in water using graph + metadata	MPNN	0.928	0.96	Combines molecular graphs + environmental features (novel)
This study	—	GAT	0.903	0.94	Strong performance from attention model
This study	—	GNN	0.842	0.89	Baseline Graphic NN
Bioaccumulation ML Study	Ecotoxicity / predictive classification	XGBoost / RandomForest / SVM	—	~0.90	ML model predicts aquatic contaminant behavior; performance close to your ROC-AUC (~0.90) [32]
Drinking water contaminants review	Water quality contaminants (binary threshold models)	RandomForest, GradientBoosting, etc	0.67–0.94	0.72–0.92	Most studies achieve "good but variable" performance depending on dataset and contaminant [33]
Classification of water quality WQI	Groundwater/river quality classification	SVM / RF / ANN	~0.90	(not always reported)	Some works achieve ACC ≥ 0.90 on water quality classification tasks [34]

positives (TP = 399) and true negatives (TN = 556), as well as a low rate of false negatives (FN = 37). This configuration results in an accuracy of 0.90 and a recall of 0.92, indicating an excellent ability to correctly detect actually contaminated samples. The low number of false negatives is particularly crucial in an environmental context, as it limits the risk of not identifying a potentially polluted sample. The GAT model comes in second place, with a balanced distribution between true positives (394) and true negatives (550). Its overall performance (accuracy = 0.89, recall = 0.90) confirms its reliability, while its attention mechanism remains a major asset for interpretability, facilitating the identification of influential molecular substructures in the classification. The GNN model, although effective, has more pronounced limitations with a higher number of false negatives (FN = 102), which affects the recall (0.77) and thus its ability to identify truly positive cases. Its overall accuracy (0.82) is lower than the other two architectures, suggesting

A less discriminative molecular representation

Overall, the results confirm the superiority of the MPNN model, which offers the best compromise between overall accuracy and sensitivity. The GAT model, slightly less performant, nevertheless remains an interesting alternative due to its explanatory potential. Finally, the GNN model appears less suitable for this task, but remains competitive for analyzes where model simplicity is a priority. The confusion matrices in Figure 8 were generated directly from the model predictions using scikit-learn, showing MPNN's superior performance (TP=399, TN=556, FN=37) and low false negatives critical for environmental monitoring.

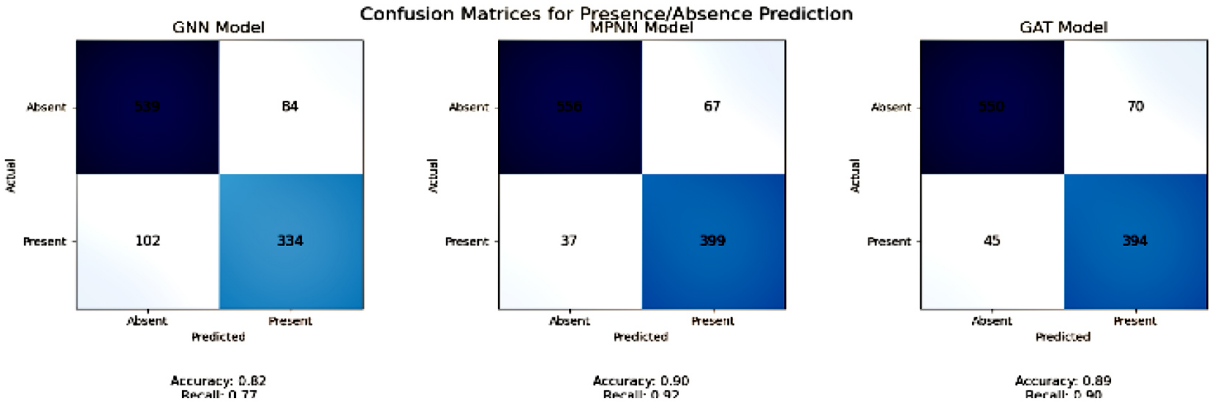

Figure 8. Confusion matrices of GNN, MPNN, and GAT models for predicting the presence of pharmaceutical residues

Discussion

Our MPNN model (accuracy = 0.928; AUC-ROC = 0.96) meets, or even exceeds, the performances reported in recent literature, notably those of Chemprop (Evangelista et al., 2025), which confirms its robustness for predicting the occurrence of pharmaceutical residues in water. This high performance shows the importance of integrating atomic characteristics and chemical bonds into the message propagation mechanism. From a biological and ecological aspect, its low false-negative rate facilitates the detection of potentially contaminated samples, supporting early identification of water affected by pharmaceutical residues. This capacity is critical for minimizing chronic exposure risks to aquatic creatures and aiding environmental managers in prioritizing monitoring and remediation operations.

The GAT model, slightly less performant, nevertheless offers a significant advantage: its attention mechanism allows for the identification of the key molecular substructures in the final decision, which is a major asset for explanatory analyzes and regulatory applications. Compared to other recent methods, such as deepFPlearn+ and GraphADT, our models—particularly MPNN and GAT—are competitively positioned, reinforcing the relevance of graph neural networks for environmental monitoring and the detection of emerging contaminants. In summary, the MPNN appears as the most effective architecture for this task, while the GAT represents a preferred alternative when model interpretability is essential. These results confirm the significant potential of graph-based models for the prediction and management of environmental risks related to pharmaceuticals.

CONCLUSIONS

This study emphasizes the tremendous potential of graph neural networks for predicting the presence of pharmaceutical residues in aquatic systems. Using a heterogeneous dataset from the Umweltbundesamt (UBA, Germany), three architectures GNN, MPNN, and GAT were implemented and carefully compared. The MPNN model demonstrated the highest overall performance and generalization across different water matrices, while the GAT model offered valuable interpretability through its attention mechanism, and the standard GNN remained competitive, confirming the relevance of graph-based approaches for emerging contaminants. By minimizing reliance on costly and sophisticated chemical studies, these models can enhance environmental monitoring, optimize sampling tactics, and support regulatory decision-making. However, factors such as class imbalance, geographical variability, and the absence of knowledge on metabolites or transformation products may influence model performance. Future studies could use finer spatiotemporal data, integrate different sources like hydrology, pharmaceutical usage, and urban characteristics, or develop multimodal models that link molecular graphs with environmental descriptors.

Importantly, by enabling early prediction of pharmaceutical occurrence, this approach may help to the protection of aquatic ecosystems and decrease biological concerns associated with chronic pharmaceutical exposure. Overall, this research reveals that GNN-based frameworks constitute a promising, efficient, and ecologically relevant method for furthering the management of new pollutants and maintaining water quality.

REFERENCES

1. Aib, H., Parvez, M. S., Czédli, H. M. (2025). Pharmaceuticals and microplastics in aquatic environments: a comprehensive review of pathways and distribution, toxicological and ecological effects. *International Journal of Environmental Research and Public Health, 22*(5), 799. https://doi.org/10.3390/ijerph22050799
2. Aira, J., Olivares, T., Delicado, F. M. (2022). SpectroGLY: A low-cost IoT-based ecosystem for the detection of glyphosate residues in waters. *IEEE Transactions on Instrumentation and Measurement, 71*, 1–10. https://doi.org/10.1109/TIM.2022.3196947
3. Ashfield, N., Li, J., Bouzas-Monroy, A., Boxall, A. B. (2025). Silent side effects: pharmaceuticals as contaminants of emerging concern. *Annual Review of Environment and Resources, 50*(1), 273–301. https://doi.org/10.1146/annurev-environ-111523-101837
4. Aziz, K. H. H., Mustafa, F. S., Karim, M. A., Hama, S. (2025). Pharmaceutical pollution in the aquatic environment: advanced oxidation processes as efficient treatment approaches: a review. *Materials Advances.* https://doi.org/10.1039/D4MA01122H
5. Belle, G., Moodley, B., Moodley, R., Omotola, E. O., Truter, C., Olatunji, O., Oberholster, P. (2025).

Occurrence and detection of selected pharmaceuticals of emerging concern: potential risks for aquatic ecosystems and human health. *Discover Applied Sciences, 7*(10), 1078. https://doi.org/10.1007/s42452-025-07716-5

6. Cano, N. O., Radjenovic, J. (2024). *Electrochemical removal of antibiotics and multidrug-resistant bacteria using S-functionalized graphene sponge electrodes.* arXiv preprint arXiv:2410.01867. https://doi.org/10.48550/arXiv.2410.01867
7. Coderre, M., Fortin, A. S., Morency, L. D., Roy, J., Sirois, C. (2025). Pharmaceuticals in drinking water: a scoping review to raise pharmacists' public health and environmental awareness on contamination in groundwater, surface water, and other sources. *International Journal of Pharmacy Practice, riaf038*. https://doi.org/10.1093/ijpp/riaf038
8. Côrtes, P. R., Loubet, N. A., Moreira, L. S., Menéndez, C. A., Appignanesi, G. A., Köhler, M. H., Bordin, J. R. (2025). Nanoscale water behavior and its impact on adsorption: A case study with CNTs and diclofenac. *The Journal of Chemical Physics, 162*(3). https://doi.org/10.1063/5.0246155
9. Domínguez-García, P., Fernández-Ruano, L., Báguena, J., Cuadros, J., Gómez-Canela, C. (2024). Assessing the pharmaceutical residues as hotspots of the main rivers of Catalonia, Spain. *Environmental Science and Pollution Research, 31*(31), 44080–44095. https://doi.org/10.1007/s11356-024-33967-7
10. Eapen, J. V., Thomas, S., Antony, S., George, P., Antony, J. (2024). A review of the effects of pharmaceutical pollutants on humans and aquatic ecosystem. *Exploration of Drug Science, 2*(5), 484–507. https://doi.org/10.37349/eds.2024.00058
11. Evangelista, D., Nelson, E., Skyner, R., Tehan, B., Bernetti, M., Roberti, M.,..., Bottegoni, G. (2025). Application of deep learning to predict the persistence, bioaccumulation, and toxicity of pharmaceuticals. *Journal of Chemical Information and Modeling, 65*(7), 3248–3261. https://doi.org/10.1021/acs.jcim.4c02293
12. Hanafiah, Z. M., Mohtar, W. H. M. W., Maulud, K. N. A., Wan, W. A. A. Q. I., Ebrahim, M. N., Abd Manan, T. S. B.,..., Yaseen, Z. M. (2025). Global pharmaceutical pollution in waterways: insights from sewage treatment point sources. *Emerging Contaminants, 100585*. https://doi.org/10.1016/j.emcon.2025.100585.
13. Kayani, K. F. (2025). Removal of pharmaceutical residues from aquatic systems using bimetallic metal–organic frameworks (BMOFs): a critical review. *RSC advances, 15*(25), 20168–20182. https://doi.org/10.1039/D5RA03056K
14. Khan, J., Friedman, A., Evans, S., Klein, R., Wang, R., Manz, K. E.,..., Bondi-Kelly, E. (2025). *FOCUS on Contamination: A Geospatial Deep Learning Framework with a Noise-Aware Loss for Surface Water PFAS Prediction*. arXiv preprint arXiv:2502.14894. https://doi.org/10.48550/arXiv.2502.14894
15. Maraj, K., Nicklin, E., Edmonds-Smith, C., Winter, K. (2025). Detection of active pharmaceutical ingredients in surface water polluted by an informal settlement. *Environmental Monitoring and Assessment, 197*(11), 1–16. https://doi.org/10.1007/s10661-025-14636-9
16. Mazhandu, Z., Mashifana, T. (2024). Active pharmaceutical contaminants in drinking water: myth or fact?. *DARU Journal of Pharmaceutical Sciences, 32*(2), 925–945. https://doi.org/10.1007/s40199-024-00536-9
17. Mheidli, N., Malli, A., Mansour, F., Al-Hindi, M. (2022). Occurrence and risk assessment of pharmaceuticals in surface waters of the Middle East and North Africa: A review. *Science of the Total Environment, 851*, 158302. https://doi.org/10.1016/j.scitotenv.2022.158302
18. Mohapatra, S., Tong, X., Mukherjee, S., Dubey, M., Saini, S., Luhua, Y.,..., Gin, K. Y. H. (2025). Comprehensive insights on the detection, occurrence and modelling of pharmaceuticals in surface water, groundwater, and drinking water treatment plants. *Journal of Hazardous Materials Advances, 100707*. https://doi.org/10.1016/j.hazadv.2025.100707.
19. Monk, J. R., Hooda, P. S., Busquets, R., Sims, D. (2025). Occurrence of pharmaceuticals, illicit drugs and PFAS in global surface waters: A meta-analysis-based review. *Environmental Pollution, 126412*. https://doi.org/10.1016/j.envpol.2025.126412.
20. Muambo, K. E., Kim, M. G., Kim, D. H., Park, S., Oh, J. E. (2024). Pharmaceuticals in raw and treated water from drinking water treatment plants nationwide: insights into their sources and exposure risk assessment. *Water Research X,* 24, 100256. https://doi.org/10.1016/j.wroa.2024.100256
21. Ngoetjana, M. P., Tesfamariam, E. H., Brown, S., Wooding, M., Dippenaar, M. A. (2025). Occurrence, concentration, and risk assessment of selected pharmaceuticals in representative cropland soils and their underlying groundwater in Gauteng province, South Africa. *Environmental Monitoring and Assessment, 197*(9), 986. https://doi.org/10.1007/s10661-025-14436-1
22. Ngqwala, N. P., Muchesa, P. (2020). Occurrence of pharmaceuticals in aquatic environments: A review and potential impacts in South Africa. *South African Journal of Science, 116*(7–8), 1–7. https://doi.org/10.17159/sajs.2020/573
23. Padhy, I., Sharma, T., Banerjee, B., Mohapatra, S., Sahoo, C. R., Padhy, R. N. (2024). Structure based

exploration of mitochondrial alpha carbonic anhydrase inhibitors as potential leads for anti-obesity drug development. *DARU Journal of Pharmaceutical Sciences, 32*(2), 907–924. https://doi.org/10.1007/s40199-024-00535-w

24. Paíga, P., Figueiredo, S., Correia, M., André, M., Barbosa, R., Jorge, S., Delerue-Matos, C. (2025). Occurrence of 97 Pharmaceuticals in wastewater and receiving waters: analytical validation and treatment influence. *Journal of Xenobiotics, 15*(3), 78. https://doi.org/10.3390/jox15030078
25. Pereira, A., Silva, L., Laranjeiro, C., Pena, A. (2021). Assessment of human pharmaceuticals in drinking water catchments, tap and drinking fountain waters. *Applied Sciences, 11*(15), 7062. https://doi.org/10.3390/app11157062
26. Sanusi, I. O., Olutona, G. O., Wawata, I. G., Onohuean, H. (2023). Occurrence, environmental impact and fate of pharmaceuticals in groundwater and surface water: a critical review. *Environmental Science and Pollution Research, 30*(39), 90595–90614. https://doi.org/10.1007/s11356-023-28802-4
27. Umweltbundesamt. *The database "Pharmaceuticals in the Environment*". Dessau-Roßlau: German Environment Agency; 2021. Available from: https://www.umweltbundesamt.de/en/database-pharmaceuticals-in-the-environment-0
28. Wada, O. Z., Olawade, D. B. (2025). Recent occurrence of pharmaceuticals in freshwater, emerging treatment technologies, and future considerations: A review. *Chemosphere, 374*, 144153. https://doi.org/10.1016/j.chemosphere.2025.144153