EECS 4414: Network Analysis on Water Contamination: Project Progress Report

Saumya Patel
York University
Toronto, ON, Canada
somy2004@my.yorku.ca

Smith Patel
York University
Toronto, ON, Canada
smith04@my.yorku.ca

Krish Patel
York University
Toronto, ON, Canada
krish211@my.yorku.ca

1 Introduction

Lead Contamination is a topic of concern not only across Ontario but also across the world. Although the water in the municipal plants of Ontario is treated according to strict guidelines, the ageing infrastructure can cause metals to leach back into water distribution networks. In 2014, approximately 92% of recorded lead releases totalled 134,000 kg, highlighting the risk posed by legacy materials. Samples from several schools and childcare centres in Toronto were observed to be exceeding 5 ppb, Health Canada's guideline. This calls for the need for proactive and targeted monitoring.[3]

The Toronto Water System is the main focus of this project because of its rich and open datasets: multi-year tap-level sampling, watermain geometries, pipe materials, construction years, and postal-code coordinates. The datasets are not consistent and thus need to be cleaned efficiently. Our team combine municipal watermain attributes (material, diameter, construction year) with tap-level lead tests to study spatial and temporal patterns. Even though overall levels have improved in recent years, but the gains are not uniform across neighbourhoods, likely due to differences in infrastructure age, materials and local corrosion control.

Our project has three main components. First, we assign risk levels (0-3) to every sample in our dataset. This segregation assesses system-wide contamination profiles. Second, we model the city as a postal-are graph and train a Graph Attention network (GAT) to predict year-over-year lead levels. Third, we simulate 6 intervention scenarios—including replacing lead pipes, replacing old pipes, reducing risk at top-priority nodes, and hybrid strategies. These simulations enable analysis for further upgrades and reshape risk patterns. Finally, we test for meaningful differences between high-risk areas from the low-risk ones by generating ROC curves and AUC scores. Together, these components open the doors to deep learning, graph construction, and eventually improvements by intervention.

2 Problem Definition

We model and predict lead contamination in Toronto's drinking-water system using a graph built from watermains and multi-year postal code samples. We extend our objectives beyond just prediction by inculcating contamination categorisation, infrastructure interventions, and analysis against real-life observations. The four interlinked problems that organize our objectives are as follows:

Problem 1 — Contamination Categorisation: Each lead sample is categorised into four risk levels based on ppb thresholds. The categories are as follows:

- **Level 0**: 0–5 ppb (none detectable)
- **Level 1**: 5–10 ppb (minimal contamination; unsafe for children)
- **Level 2**: 10–15 ppb (exceeds regulatory trigger; action required)
- **Level 3**: > 15 ppb (exceeds EPA action level; severe contamination)

Problem 2 — GAT-based prediction: Construct a postal graph by snapping watermain endpoints to nearest postals (nodes = postals, edges = inferred connectivity). Using simple node features (recent 3-year mean lead, node degree), train a lightweight GAT model to predict year-t lead for every postal and compare GAT predictions to measured values on overlapping postal/year pairs. Report MAE, RMSE, and R^2 per year and overall and inspect actual and predicted scatter plots to see where the model tracks (or misses) observed patterns.

Problem 3 — Intervention scenarios: Simulate 6 intervention scenarios to analyse the need for targeted updates.

- **High-Risk Materials Replacement:** Replace all pipes made of CI, CICL, DIP, DICL, and UNK with PVC.
- **Age-Targeted Renewal:** Replace all pipes older than 50 years.
- **Node-Fix Hotspots:** Apply localized remediation at the top 10 highest-risk postal codes.
- **Material-Weighted Priority Replacement:** Replace the top 10% highest-risk pipes using a combined score of material, age, and diameter.

Problem 4: ROC/AUC Model Evaluation: Using the median of real measurements, risk labels for 2024 and 2025 are created by aligning predicted contamination with actual postal-year samples. ROC curves and AUC scores are computed, and the performance is compared. This ensures how efficiently the GAT model distinguishes higher-risk areas from lower-risk areas.

et al.

3 Related Work

Lead is considered to be a primary contaminant in the study of water networks. Both dissolved and particulate lead, even in small amounts, from legacy pipes and solder, create spikes in exposure. Prior work from Schock*et al.* shows that factors such as pipe material, corrosion scales, and historical water chemistry are important for the study of lead levels and also their intervention. We draw inspiration from this source and choose lead as our primary target for analysis and explore intervention strategies[6].

Prior water quality research shows that adding graph context to time series learning improves accuracy over sequence–only models. In particular, Liu*et al.* combine an LSTM with a GAT to model spatio-temporal dependencies across monitoring sites and report consistent gains on river quality prediction tasks[5]. Their principle of creating infrastructure models and graph connectivity is

what we apply to create city-scale nodes (postal areas) and derive edges from the watermain distribution network of Toronto. On the training side, recent ML work emphasises careful optimisation and transparent evaluation for water quality regression. Cao*et al.* discuss practical choices like Adam style optimisers, ELU activations and common error metrics (MAE, RMSE, R^2), highlighting the role of normalisation for stable training and fair comparison [2].

We also draw inspiration from Anaadumba*et al.*, who analysed that if a city is treated as a graph and GAT is used to learn neighbour importance, it is easier to predict the lead risk. We induced the same core idea in our project while shaping it according to Toronto's datasets: postal-area centroids as nodes, watermain lines snapped to the nearest postal areas as edges and predictions up to the pipe level (average of endpoints), so we can relate results to pipe material, diameter, and age.[1]

Saito and Rehmsmeier analyse how useful ROC curves and AUC (Area Under Curve) are in terms of distinguishing high-risk contamination from the low-risk ones. For our project, we induced this knowledge in our postal-level lead-risk classification[7].

4 Methodology

4.1 Overview

To address the problem formalised in Section 2, we follow the six steps given below for our project:

(1) Acquire and clean all relevant data streams, and categorise each measurement into four contamination levels for baseline system understanding.(Section 4.2);

(2) Construct a geospatial network that encodes plausible pathways/adjacency among postal areas using the watermain layer (Section 4.3);

(3) Train predictive models over regions' data within a node over the years(Section 4.4).

(4) Edge-level pipe analysis to relate lead to pipe attributes like age, diameter and material.

(5) Simulate intervention scenarios—including replacing lead pipes, replacing old pipes, reducing risk at top-priority nodes, and hybrid strategies.

(6) Aggregation of node predictions to the edges (pipes) to evaluate how material, diameter, and age relate to the predicted lead risk and also use ROC curves and AUC scores to assess how the model distinguishes high-risk postal areas from the low-risk ones.

4.2 Data Sources and Cleaning

Three primary datasets: multi-year tap-level lead samples, postal-code geospatial coordinates, and Toronto's watermain network. PartialPostalCode are used as the linking key to join the three datasets; all fields are stripped of whitespaces and coerced to consistent types.

Lead samples (2014–2025): For each record we parse `SampleDate`, `PartialPostalCode`, and `LeadAmount (ppm)`. Dates are coerced to a canonical timezone using:

$$t_i = \text{to\_datetime}(\text{SampleDate}_i), \qquad \text{year}_i = \text{year}(t_i)$$

Lead is converted to numeric; non-numeric entries such as "`<0.00005`" are treated as left-censored and set to NaN for interpolation.

Irregular sampling is handled by applying within-postal linear interpolation over time:

$$L_{i,t} = \text{interp\_linear}\big(\{(t, L_{i,t}) : t \in T_i\}\big), \quad i = \text{ postal codes.}$$

This fills gaps only between valid observations:

$$\hat{L}_{i,t} = L_{i,t} \quad \text{for } t \in T_i,$$

$$\hat{L}_{i,t_0-1},\ \hat{L}_{i,t_K+1} \ \text{ remain NaN,}$$

avoiding extrapolation beyond observed years.

A derived ppb value is created for categorization:

$$L_{i,\text{ppb}} = 1000 \cdot \hat{L}_{i,\text{ppm}}.$$

This is followed by categorisation of lead samples into one of four contamination levels:

$$\text{Category}(L_{\text{ppb}}) = \begin{cases} \text{Level 0}, & 0 \le L_{\text{ppb}} < 5, \\ \text{Level 1}, & 5 \le L_{\text{ppb}} < 10, \\ \text{Level 2}, & 10 \le L_{\text{ppb}} < 15, \\ \text{Level 3}, & L_{\text{ppb}} \ge 15. \end{cases}$$

Postal-Code Coordinates: Each postal code in EPSG:4326 is used to create centroids, which after being named to PartialPostalCode, serve as the nodes for graph construction and also target positions for watermain endpoint using cKDTree nearest neighbour queries.

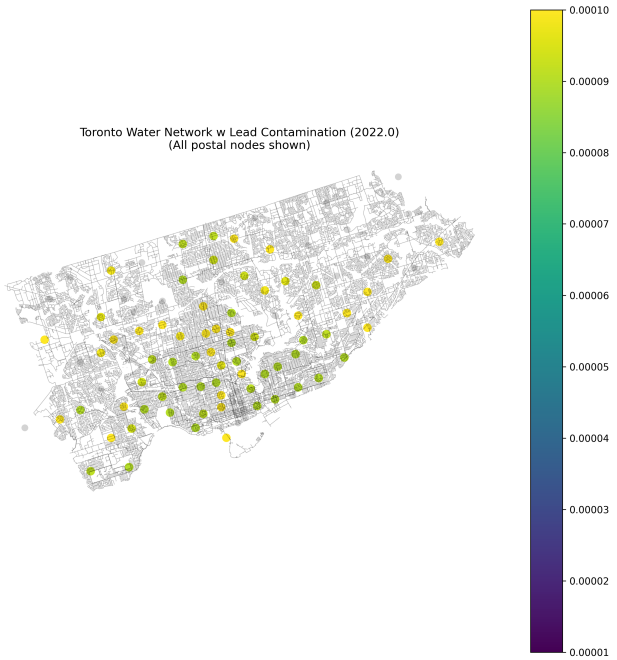


Figure 1: Observed lead concentration by postal centroid in Toronto (2022). Light-gray lines show the watermain network; points mark postal centroids. Color encodes measured/interpolated lead (ppm) on a fixed scale 1×10^{-5}–1×10^{-4} to enable year-over-year comparison; gray points indicate missing data.

Watermain GeoJSON: The water main system includes material, diameter, and construction year. Each MultiLineString is divided into different points that are snapped to the nearest postal centroid using:

$$\text{postal}(p) = \arg\min_j \|p - c_j\|.$$

Edges are then created between the centroids if a pipe connects the nearest endpoints. Each edge e = (u,v) has attributes like:

$$\text{diameter}_e \in \{\text{diameters}\},$$

$$\text{material}_e \in \{\text{materials}\},$$

$$\text{age}_{e,t} = t - \text{construction\_year}_e.$$

All these attributes are used for further analysis.

PanelConstruction : For consistent year-over-year mapping and modeling, we materialize the full Cartesian product of all unique postal codes (from the coordinate table) and all observed years (from the sample data), like:

$$P \times Y = \{\ (p, y) : p \in \text{postal codes},\ y \in \text{observed years}\ \}.$$

We then:

(1) Left-join the interpolated readings onto the grid (`NaN` where no sample exists in that year).
(2) Attach `Latitude`/`Longitude` so every (postal, year) row can be geocoded.

This ensures that map gaps reflect *true* absence of data rather than missing join keys.

4.3 Network Construction

We create a yearly undirected graph G_t whose nodes are postal centroids. Edges approximate infrastructure connectivity by snapping watermain endpoints to their nearest postal centroid using a k-d tree. Each watermain pipeline contributes an edge between its two nearest postal codes, multi segment mains are processed segment wise. If multiple mains connect the same postal pair, we keep a single edge and accumulate their attributes (materials, diameters, construction years) as lists on that edge.

Edge weighting (heuristic): To optionally bias information flow, we assign each edge a scalar weight derived from its accumulated attributes. Materials receive domain-motivated factors (e.g., LEAD > CI/DIP > COPPER > PVC, `UNKNOWN` in the mid range). Diameters are averaged and scaled. The age of the pipes is calculated by: $\text{age}_{e,t} = t - \text{construction\_year}_e$ The intent is to provide a reasonable prior that heavier/older/less inert mains could carry stronger influence between neighboring postals. Our team also run the model on *topology-only* (all weights equal).

Each node p in year t carries two simple, stable features:

(1) **Historic lead signal:** the mean of that postal's measured (or interpolated) lead values over the previous three years:

$$x_{p,t}^{(\text{lead})} = \frac{1}{3} \sum_{\tau=t-3}^{t-1} \hat{L}_{p,\tau}.$$

This anchors predictions in recent local behavior without leaking the target year.
(2) **Graph connectivity:** the node's degree $\deg_{G_t}(p)$ in G_t, capturing how embedded a postal is in the local network induced by snapped mains.

Both features are min–max scaled *within year* to a comparable numeric range for training.

4.4 Learning Model (GAT) and Training Loop

We use a compact 2-layer Graph Attention Network (PyTorch Geometric). The first attention layer reads the two inputs and aggregates information from neighbors (multi-head attention; ELU activation). The second attention layer outputs a single scalar per node: a *normalized* lead estimate for year t.

For each target year t (2018–2026), we train on the immediately preceding three years ($t-3$ to $t-1$). Only nodes with a measured year-t value contribute to the MSE loss. Unlabeled nodes still participate in message passing so the model can use neighborhood context. After 200 epochs (Adam with a small weight decay), we infer values for *all* nodes and de-normalize to ppm using the same window's observed range[5].

[H] InputOutput Lead panel: (PartialPostalCode, date, ppm); postal centroids; watermain pipelines with `material`, `diameter`, `year`. For each target year t: postal-level predictions (ppm), PNG map, appended rows in `gat_predictions_by_year.csv`

Preprocess: Normalize headers, parse dates, set year = year(date); Convert ppm to numeric, within each postal code, linearly interpolate missing ppm, Join centroids to lead panel; build full postal × year panel.

For each target year $t \in \{2018, \ldots, 2026\}$ *Skip* if any of $\{t-3, t-2, t-1\}$ fall before data start.

Graph build G_t Snap each watermain endpoint to the nearest postal centroid via KD-tree. For each line, connect snapped endpoints; merge duplicates. On each edge, accumulate {`materials`, `diameters`, `ages` = t−`construction_year`}; Compute edge weight w_{uv} from material multipliers and mean diameter.

Features & labels For each node (postal), compute 3-year mean lead over $\{t-3, t-2, t-1\}$. Form node features: {scaled 3-year mean, scaled degree}, Mark labeled nodes: those with measured mean in year t.

GAT training Build bidirectional `edge_index`; Two-layer GAT (4 heads → hidden, ELU; 1 head → scalar), dropout 0.6, Train 200 epochs with Adam (lr = 0.005, wd = 5e-4) built in using MSE on labeled nodes[6].

Predict normalized scores for all nodes; invert min–max scaling to ppm, merge with centroids; render dot map over watermains (clip/scale for display). Save `GAT_Lead_Prediction_{t}.png` and append {year, postal, y_{pred}, y_{true}} to `gat_predictions_by_year.csv`.

Why postal centroids? They align with address-level sampling, keep the graph compact and are easy to map. **Why snapping mains?** It turns linear infrastructure into a neighborhood graph that approximates physical adjacency. **Why two features?** A deliberately small feature set highlights the value of relational learning (attention across neighbors) without overfitting.

Quality controls: We validate coordinate ranges (Toronto bounding box), drop exact duplicate records before interpolation, enforce time ordering within postal groups, and render annual maps with a fixed color scale for fair comparison across years.

Limitations. Edge weights are heuristic; results should also be reported with topology-only edges. Per-year min–max scaling changes the numeric range over time; we chiefly use it to stabilize

training and always report predictions in ppm, spatiotemporal GNNs could be used if there is more data.

4.5 Edge Level Pipe Analysis (Built from Our GAT Predictions)

Here we convert postal level GAT predictions into pipe level summaries so we can inspect how predicted lead relates to pipe age, diameter and material.

- Treat each watermain as an **edge** (u, v) between two postal nodes.
- For that edge, take the **average** of the two endpoint predictions as its lead score.
- Aggregate pipe attributes over collapsed segments: *age* = mean of $(t - \text{construction\_year})$, *diameter* = mean of recorded diameters, *material* = a single, stable label (first item of sorted unique set).

CSV files are created to generate relationships between pipe-age vs lead, diameter vs lead, and pipe material vs lead box plots for every year.

4.6 Intervention Scenarios

We created a list of four intervention scenarios that can help upgrade targeted areas where remediation is required. We train GAT models to obtain new predictions for each intervention scenario and rebuild a postal-level graph[4]. The four intervention scenarios are as follows:

- First, we focus on pipe replacement, wherein all high-risk pipe materials (with high variance and means) like CI, CICL, DIP, DICL, and UNK are replaced with PVC pipes (safer). This intervention is significant because of the importance of materials in node features.
- Second, we replace pipes that are older than 50 years. Since older pipes are more susceptible to interior deterioration and lead particulate release, replacing them would result in effective city-wide improvement.
- Third, we focus on pointing out localised hotspots- the highest-risk postal codes. Identifying this is important because if risky postal codes are improved, adjacent areas are indirectly benefited through GAT propagation. This intervention is the easiest to implement and update.
- Fourth, we focus on priority-based replacement. High-risk pipes are identified using combined scores of materials, pipe age, and diameter. This scenario generates the greatest improvement because of all attributes taken into account.

5 Experiment / Evaluation

This section focuses on figure-based findings that are produced from our analysis. The section covers maps, curves and charts that are important to study to visually understand spatial clustering, prediction of models, trends of pipe attributes, and classification of high-risk areas, respectively.

5.1 Contamination Categorization

A bar chart and a pie chart are generated to reveal the number of samples belonging to each category. (Level 0-3). The results show that the majority of the samples belong in Level 0 and Level 1, indicating low lead risk in Toronto. However, samples belonging to Level 2 and Level 3, though in smaller amounts, still pose a threat to the overall network distribution.

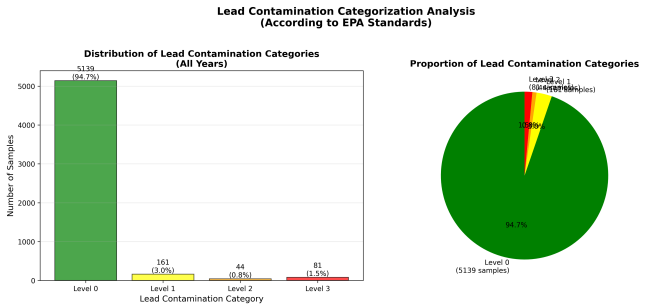
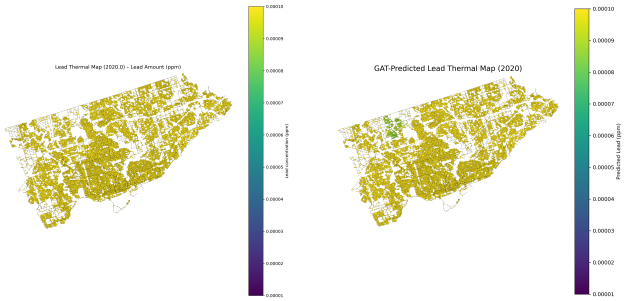


Figure 2: Distribution and proportion of lead contamination categories across all samples (Levels 0–3).

5.2 Actual vs Predicted GAT Maps (2020-2025

A set of predicted GAT maps is generated by training on actual datasets of the past three years for each map. The results are found to be similar to actual maps in terms of structure.

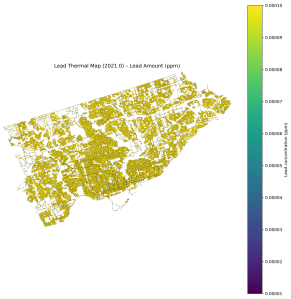
The predicted GAT maps demonstrate a clear downward trend. In the initial set of years (2020-2022), intense colours like yellow dominate, while later years (2023-2025) demonstrate relatively cool colours like blue dominate. Geographical differences are described in the image description of yearly graphs.



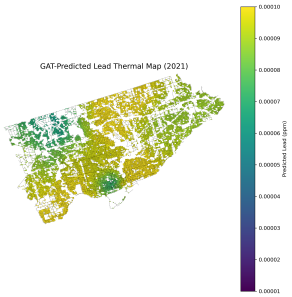
Actual 2020 **Predicted 2020**

2020 Geographical Summary:

Actual: High contamination concentrated in central and midtown areas, with strong yellow clusters across the core. Predicted: Model outputs uniformly high values (yellow everywhere), failing to capture spatial variation. Only a small northwest patch shows slightly lower values.
Overall: Actual map shows localized clusters, while predicted map is flat and overestimates city-wide contamination.



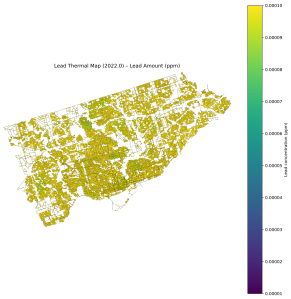
Actual 2021



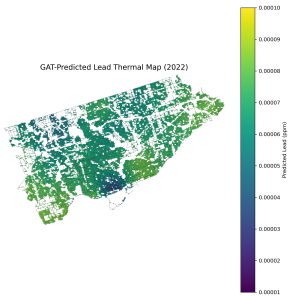
Predicted 2021

2021 Geographical Summary:

Actual: High levels in central Toronto, extending east and west; north and far west remain lower.
Predicted: South-central hotspot emerges; northwest and east show cooler green areas.
Overall: Prediction captures central–south risk but smooths out variation elsewhere.



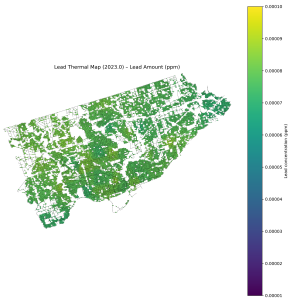
Actual 2022



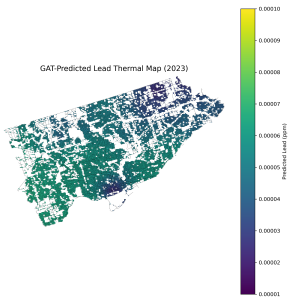
Predicted 2022

2022 Geographical Summary:

Actual: Strong contamination corridor through central and south-central Toronto.
Predicted: Central–south hotspot well captured; west and northwest cool to green more than actual.
Overall: Good spatial alignment with slightly stronger predicted reductions outside the core.



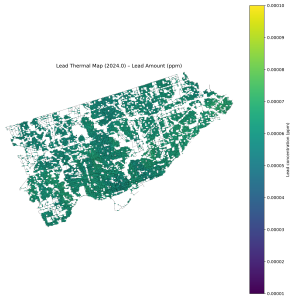
Actual 2023



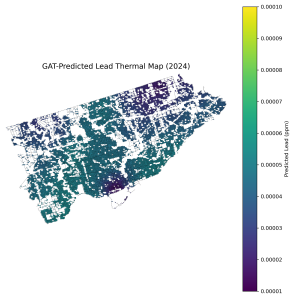
Predicted 2023

2023 Geographical Summary:

Actual: Overall decline with remaining elevated zones in central-east and central-south.
Predicted: City turns mostly green/teal with weak hotspots in central–south and northeast.
Overall: Both show a decline, but predicted reductions are stronger than actual.



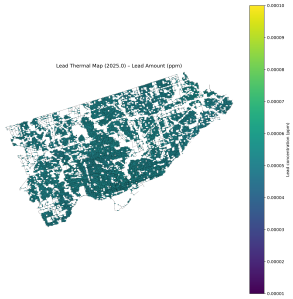
Actual 2024



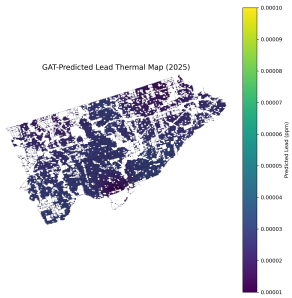
Predicted 2024

2024 Geographical Summary:

Actual: Remaining contamination mainly in the core, with outer regions mostly low.
Predicted: Deep green/blue city-wide, with scattered elevation near south-central areas.
Overall: Prediction captures the shape but shows a much stronger city-wide decline.



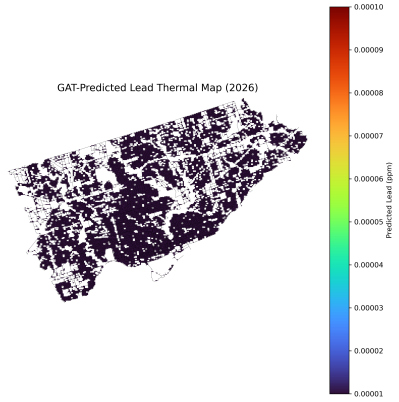
Actual 2025



Predicted 2025

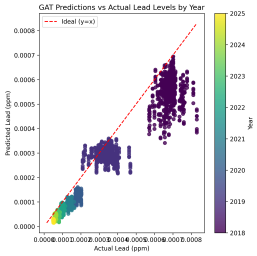
2025 Geographical Summary:

Actual: Low values city-wide with subtle central clustering.
Predicted: Dark blue throughout, indicating very low predicted contamination; tiny elevation near waterfront.
Overall: Both show very low contamination, with predictions suggesting an even sharper final decline.



GAT-Predicted Lead Thermal Map for 2026

This figure shows the model's one-year-ahead prediction of lead concentrations (ppm) across Toronto. The GAT output is uniformly low (dark blue), indicating that the model forecasts minimal contamination city-wide in 2026. Small isolated areas show slightly elevated values, but overall the network's risk is predicted to decline significantly.

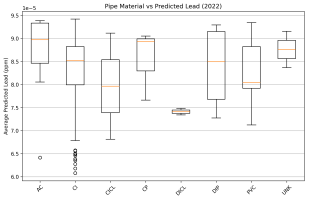


GAT Predictions vs Actual Lead Levels (2018–2025)

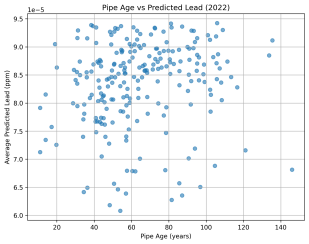
Summary:
The scatter plot compares predicted vs. actual lead levels across all years, with points coloured by year. The 1:1 line (red) represents perfect prediction. Older years (2018–2020) cluster near the lower-left, while later years (2023–2025) shift upward to higher actual and predicted values. The spread indicates systematic underprediction at higher contamination levels and tighter agreement at lower levels.

5.3 Material Analysis

Pipe Age vs Lead Contamination and Pipe Material vs Lead Contamination graphs are generated for each year. These graphs are crucial for analysing the importance of pipe attributes in lead contamination. Due to space constraint, the above two graphs are shown for just one year(2022).



Predicted Lead vs Pipe Material (2022)



Pipe Age vs Predicted Lead (2022)

Summary:

Materials show clear differences in predicted lead, with AC, CP, and PVC exhibiting higher medians, while DICL remains consistently low with minimal spread. Pipe age displays a weak upward trend: older pipes tend to have slightly higher predicted lead, but substantial variation indicates that age alone is not the dominant risk driver.

5.4 Intervention Techniques

The four intervention scenarios (Section 4.6) are simulated and observed in two forms.

First, the pipes are replaced with different strategies:

- High-Risk Material Replacement: The largest red footprint among the four. Densely clustered across midtown, downtown, and suburban areas. These areas are the older sections of Toronto and, hence, are still observed using these pipes.
- Age-Targeted Renewal: Less dense than material replacement. Still mostly concentrated in central/east Toronto.
- Node-Fix Hotspots: 10 high-risk postal nodes (hotspots) are shown here. No pipes replaced here.
- Material-Weighted Priority: More dense than High-Risk Materials but less dense than Age-Targeted. Combined scores of material attributes are calculated for the top 10% highest risk pipes.

Second, a bar graph is generated that shows how the above four intervention scenarios lead to improvements in the mean lead levels in the water distribution network.

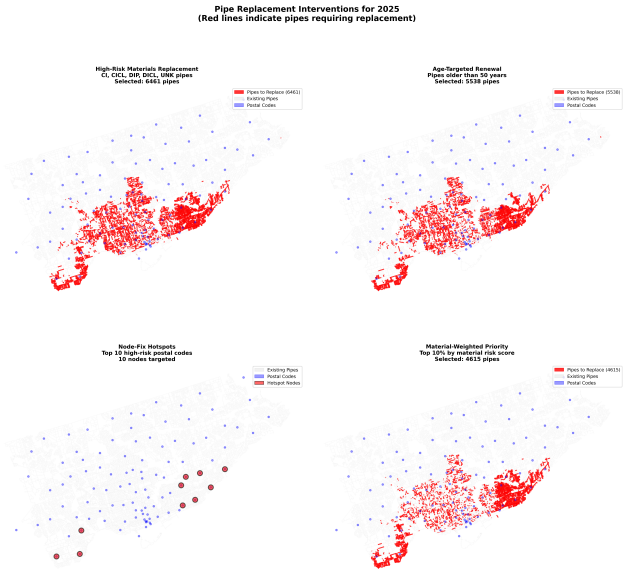


Figure 3: Pipe Replacement Interventions for 2025. Red segments indicate pipes selected for replacement under four strategies: high-risk materials, age-targeted renewal, node-fix hotspots, and material-weighted priority.

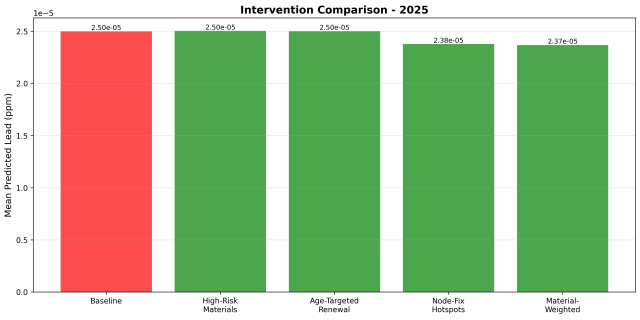
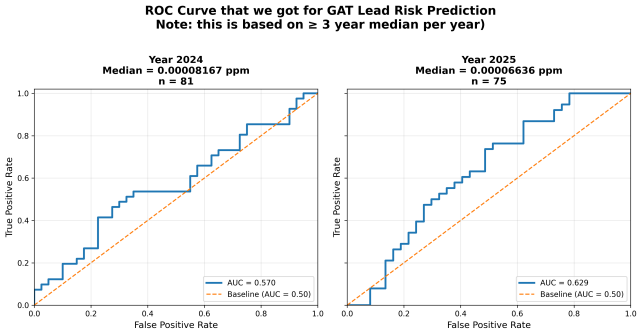


Figure 4: Comparison of mean predicted lead concentrations across four intervention strategies relative to the baseline (2025). Lower bars indicate greater system-wide improvement.

5.5 ROC AUC

The ROC and AUC curves are a method for model evaluation which depicts how our GAT predictions are, with respect to the baseline accuracy. The AUC improves from 0.57 in 2024 to 0.62 in 2025,

showing the improvement of GAT predictions in differentiating high-risk nodes from the low-risk ones[7].



ROC–AUC Performance (2024–2025):

The model shows modest discriminative ability, with AUC improving from 0.57 in 2024 to 0.63 in 2025. Both curves lie close to the diagonal baseline, indicating limited but above-random predictive performance.

Overall Accuracy and Error Magnitudes for this prediction by GAT, our team evaluate GAT predictions against postal/year averages using mean absolute error (MAE), root mean squared error (RMSE), and R^2. In practical terms, MAE corresponds to an average deviation of ≈ 0.052 ppb and RMSE ≈ 0.072 ppb (1 ppm = 1000 ppb), which is small relative to the range of observed values in our dataset.[2]

	MAE (ppm)	RMSE (ppm)	R^2
Overall (2018–2025)	5.2×10^{-5}	7.2×10^{-5}	0.8836

Table 1: Aggregate performance of the GAT model on postal/year pairs with ground truth.

5.6 Future work

We also try to include schools and daycare datasets, the dataset is unorganised and is not geocoded at all; we took the dataset from Ontario's website. After trying for hours we weren't able tot able to find the geocoded locations for the place even after using the many geocoding tools and api's it seems that we need more information than just the place's name in the DWS Name column for the dataset (Note this dataset has approx. 31000 lead samples taken in Ontario including Toronto) we narrowed down the locations to Toronto on based of PHU (Public Health Unit Names) there are approximately 1128 unique locations which are different schools and daycares in Toronto with lead value peaking at 1200 ppb to 0.05 ppb to get this into perspective Canada's Legal limit of Lead is 5 ppb so this dataset contains a lot of values ranging in the min-max value that I mentioned earlier. We tried Google's My Maps in order to get the exact location, but we cannot get the coordinates for those locations in a CSV/XLSX file which mentions coordinates for all the locations. So, we manually found 58 unique locations across Toronto and plotted those on our Watermain pipeline infrastructure for Toronto to get a visualization we have also categorize the lead values above and below the threshold level of 5 ppb into 4 levels (level 0, 1, 2, 3) which in increasing order for lead contamination in water (This are for all locations sample lead values dataset). This dataset would be considered for future work if it were properly organised and structured[8].

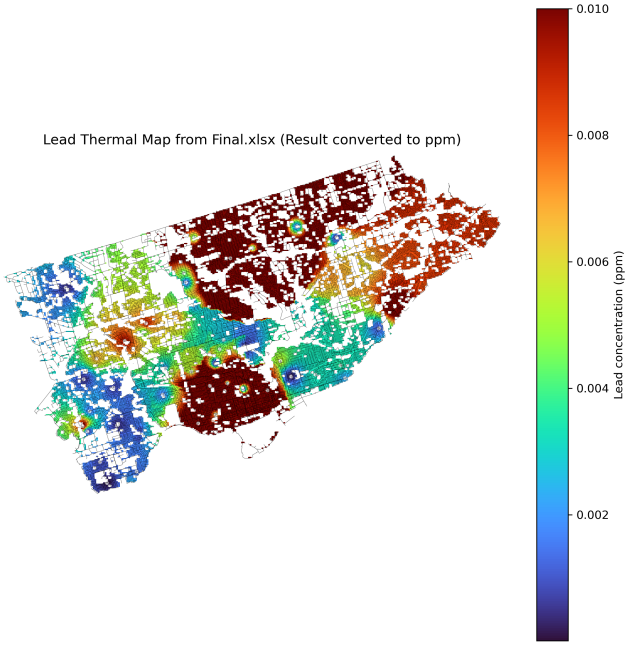


Figure 5: Thermal map of average lead concentration across Toronto. This visualization shows spatial variation in lead levels (converted to ppm) aggregated from our cleaned sample dataset. Higher concentrations appear in red, while lower concentrations appear in blue/green. This map forms a baseline for integrating external datasets, including school and daycare lead-testing records, into future iterations of the model.

6 Conclusion

This project demonstrated a data-driven, graph-based framework for analyzing and forecasting lead contamination across Toronto's drinking water network. By combining multi-year lead sampling data (2014–2025) with watermain infrastructure and postal-code geometry, we constructed a spatial graph representation where postal regions act as nodes and watermains define connectivity. A two-layer Graph Attention Network (GAT) trained on historical data successfully captured both the spatial gradient of contamination higher levels in older, central areas and the gradual temporal decline over recent years. The model achieved strong quantitative performance ($R^2 \approx 0.88$) with low error magnitudes, confirming that relational learning can extract meaningful patterns from sparse, irregular monitoring data.

Beyond this, several other components were also analysed. First, raw samples were categorised into 4 different risk levels. A major part of the samples fall in Level 0-1, which is a positive aspect. However, small amounts still fell in Level 2-3, which should not be

ignored. Second, material analysis was done, which linked the pipe attributes like material and age with lead contamination. Third, intervention scenarios were simulated—covering high-risk material replacement, age-targeted renewal, hotspot remediation, and material-weighted prioritisation. This was followed by analysing how the above simulations could improve the current scenario. Finally, we did ROC and AUC analysis to cover how trained models successfully distinguished high-risk postal areas from the low-risk ones.

In conclusion, categorisation, graph-based prediction, intervention modelling, and classifier-quality evaluation were combined to create a clean and efficient framework for analysing lead contamination patterns across Toronto. This analysis could be utilised as a support system for future targeted lead mitigation projects.

References

[1] Anaadumba R, Bozkurt Y, Sullivan C, Pagare M, Kurup P, Liu B, Alam MAU (2025). *Graph neural network-based water contamination detection from community housing information.* Front. Environ. Eng. 4:1488965. doi: 10.3389/fenve.2025.1488965. https://www.frontiersin.org/journals/environmental-engineering/articles/10.3389/fenve.2025.1488965/full

[2] Cao, Jing; Zhao, Dong; Tian, Chenlei; Jin, Ting; Song, Fei. *Adopting improved Adam optimizer to train dendritic neuron model for water quality prediction.* Mathematical Biosciences and Engineering, 2023, 20(5): 9489–9510. doi: 10.3934/mbe.2023417. ELU, MSE, MAE, RMSE and data normalization. https://www.aimspress.com/article/doi/10.3934/mbe.2023417

[3] Government of Canada, Lead section. https://www.canada.ca/en/environment-climate-change/services/environmental-indicators/releases-harmful-substances-water.html

[4] Homewater Canada. *Lead Pipes in Homes: A Silent Danger That Should Be Addressed.* https://www.homewater.com/blog/lead-pipes-in-homes-a-silent-danger

[5] Liu, Yueyi; Zheng, Hang; Zhao, Jianshi. *Enhanced water quality prediction by LSTM and graph attention network (L-GAT): An analytical study of the Pearl River Basin*, Water Research X, Volume 28, 2025, 100383. ISSN 2589-9147. https://doi.org/10.1016/j.wroa.2025.100383 (https://www.sciencedirect.com/science/article/pii/S2589914725000829)

[6] Schock M. R., Lytle D. A. (2011). *Internal corrosion and deposition control.* In: Water Quality Treatment: A Handbook on Drinking Water, 6th ed. American Water Works Association (AWWA). ISBN: 9780071630115.

[7] Saito T, Rehmsmeier M (2015). *The precision–recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets.* PLOS ONE 10(3): e0118432. doi: 10.1371/journal.pone.0118432 https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0118432

[8] University of Toronto (2024). *Toxic lead showing up in Ontario school and daycare drinking water as evidence of serious health dangers grows.* https://ijb.utoronto.ca/news/toxic-lead-showing-up-in-ontario-school-and-daycare-drinking-water-as-evidence-of-serious-health-dangers-grows/