Bioinformatics Summer School Report

Name: Mohammad Qadi

Internship Organization: University of Toronto, Faculty of Applied Science, Department of Civil and Mineral Engineering

Supervisor's Name: Seungjae Lee, Jeffrey Siegel

Project title: Indoor Air Quality Data Analysis using Physics Informed Machine Learning

1. Introduction

Indoor air quality (IAQ) affects health, comfort, and productivity for people who spend most of their time indoors. Yet the main drivers of indoor CO_2—human emissions (occupancy/activities) and air exchange with the outdoors—are rarely measured directly, and low-cost sensors add noise and bias. Our goal was to develop a physics informed machine learning (PIML) framework that takes only timestamp and temperature as causal proxies and learns two hidden, time varying drivers: exchange rate ER(t)ER(t) and emissions E(t)E(t). Rather than predicting CO_2 directly, the model predicts ER(t)ER(t) and E(t)E(t) and injects them into the CO_2 mass balance update

$$C_{ik+1} = C_{ik} + ER_k \left(C_{ok} - C_{ik} \right) + E_k,$$

then trains on the physics based next step error. This design yields interpretable drivers and forecasts while staying consistent with the underlying dynamics.

2. Materials and Methodology

Data

- **Signals:** indoor CO_2 (Ci) , outdoor CO_2 (Cok), timestamp, temperature TT.
- **Calendar proxies:** hour-of-day, hour-of-week, and a weekend flag; encoded via small embeddings to capture diurnal/weekly patterns in Ek.
- **Quality control (QC):** drop obvious faults (e.g., Cik <200 ppm), extreme jumps (>300 ppm per step), and very large gaps.

Train/Test Split

- **Forward-chaining split:** first 80% of the timeline for training, last ~20% for testing

Model

- **Two-head LSTM (compact):**
    - **Head ER:** LSTM on temperature sequence → linear head → activation (either bounded sigmoid×cap for physical plausibility, or softplus/linear depending on constraint).

    - **Head E:** LSTM on calendar embeddings (and optionally time-of-day sin/cos) → linear head → activation (softplus for nonnegative emissions, or signed with a tiny mean penalty to absorb bias).

- **Differentiable physics layer:** forms $C_{i_{k+1}}$ using the equation above; loss is computed on this physics-based prediction, not on raw network outputs.

Evaluation

- **One-step (teacher-forced):** next-step error given true Ci.
- **Closed-loop:** seed once and integrate across the test segment (diagnoses drift).
- **Conditional horizon forecasts:** at many start times, seed with true Ci, integrate to **5/10/30/60 min** horizons; because sampling is irregular, we compare at the **exact horizon time** using fractional alignment within the crossing step.
- **Metrics:** RMSE/MAE (ppm) and **R2R^2** (variance explained, scale-free). We report residual summaries and perform QC-consistent scoring.

3. Results and Discussion

Forecast Accuracy

Using the fast conditional-horizon method (subsampled starts, QC mask applied), we obtained on the test split:

- **5 min:** RMSE ≈ **55 ppm**, MAE ≈ 16 ppm, **R2≈0.68R^2 \approx 0.68**
- **10 min:** RMSE ≈ **54 ppm**, MAE ≈ 19 ppm, **R2≈0.69R^2 \approx 0.69**
- **30 min:** RMSE ≈ **54 ppm**, MAE ≈ 27 ppm, **R2≈0.70R^2 \approx 0.70**
- **60 min:** RMSE ≈ **58 ppm**, MAE ≈ 36 ppm, **R2≈0.64R^2 \approx 0.64**

4. Skills and Knowledge Gained

- **Physics-informed ML:** how to embed a differentiable physical law into a neural network and train through it.
- **Modelling under weak identifiability:** leverage-weighted losses, short-horizon free-runs, and gentle regularization to offset weird results.
- **Evaluation design:** differentiating one-step vs. closed-loop vs. conditional forecasts.

5. Challenges and Solutions

- **Sensor noise & bias:** addressed with Huber loss, QC mask, and (optionally) allowing signed EE with a tiny mean penalty.
- **Model "offset" hacks:** the free-run loss and mean-EE regularizer reduce the tendency to fit one-step by inflating EE.

- **cuDNN non-contiguity errors:** ensured **contiguous** tensors, and ran test inference on CPU when necessary for stability.

6. Conclusion

Even with low-cost sensors and only time & temperature as inputs, a compact physics-informed LSTM can recover interpretable, time-varying emissions **E** and a meaningful, albeit less identifiable, exchange rate **ER**. The framework achieves useful short-horizon forecasts (5–60 min, R2∼0.65–0.70R^2 \sim 0.65–0.70) and outperforms a time-invariant baseline, demonstrating that human-driven IAQ patterns can be inferred from proxies when guided by physics. Remaining error and long-horizon drift are consistent with the data's low leverage for ER and sensor limitations. We expect simple additional signals (occupancy/door/fan states), better sensor QC/bias correction, or a light state-space filter around the LSTM to further stabilize ER and extend accurate forecasts to longer horizons.

7. Acknowledgments

Supported by DSI, HAB, KAUST, NTDP, and UofT.
Grateful to Prof. Seungjae Lee and Prof. Jeffery Siegel for their guidance.