Interpretable Deep Learning for Toxicity Prediction

Aranya Banerjee, Kevin Boby, Samuel Lam, Jeffrey Li, David Polefrone,
Robert San, Erika Schlunk, Sean Wynn, Colin Yancey
Mentor: Professor Soheil Feizi **Librarian:** Ms. Lindsay Inge Carpenter

GEMSTONE Honors College University of Maryland

PROBLEM

- Late-stage failures in animal and human drug testing make drug development expensive and time-consuming [1].
- Pharmaceutical regulators and developers increasingly leverage deep learning models to detect adverse health effects earlier in the process.
- However, the crucial step of model interpretation has not yet been thoroughly explored.

RESEARCH QUESTIONS

- How can false positive and false negative rates benchmark model interpretability?
- How can model interpretation across data representations identify toxic features?

METHODOLOGY

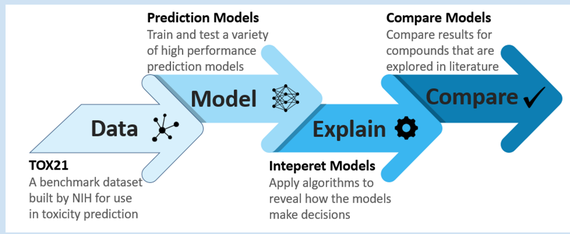


Figure 1

DATA REPRESENTATIONS

- **Graphs:** Atoms are treated as nodes and bonds as edges.
- **Bits:** 0's and 1's indicate the absence and presence of atoms and functional groups.

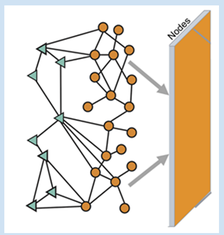


Figure 2

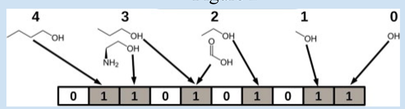
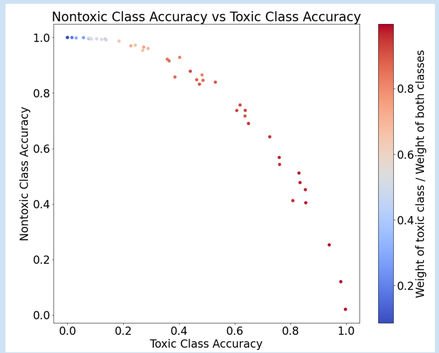


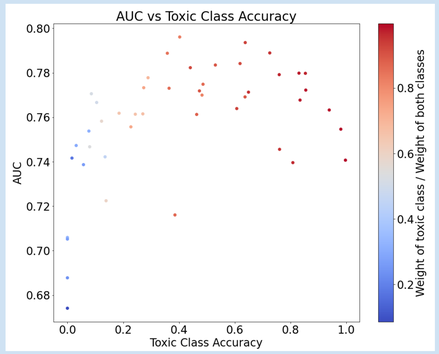
Figure 3

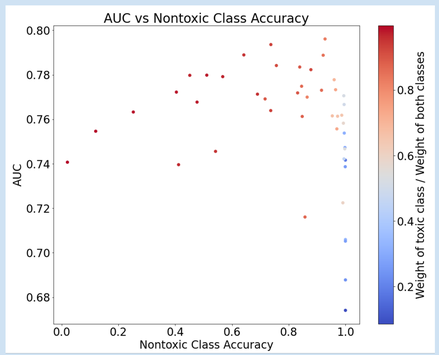
DATA ANALYSIS: INTERPETABLE METRICS

- Previous works benchmark model performance only by Area Under the Curve (AUC), a metric agnostic to class imbalances [2].
- However, we observe a significant difference between data representations in AUC's relationship with class-specific accuracies as class weighting changes, which previous works omit from their methodology.

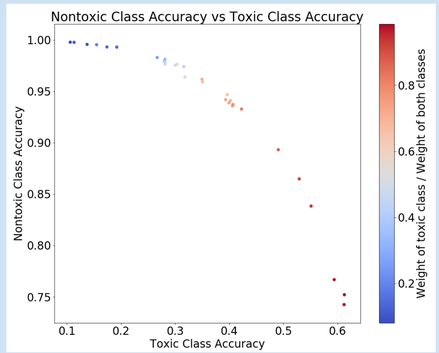
Graph Model

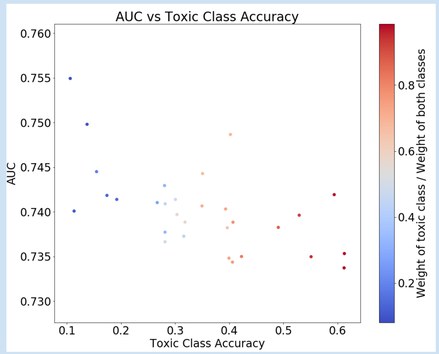


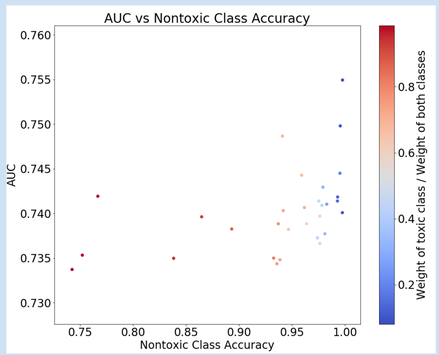




Bit Model







INTERPRETATION COMPARISON

- Using gradient-based and sampling algorithms, we are able to isolate features in compounds that contribute to positive predictions for both representations [3].
- We plan to qualitatively and quantitatively compare the interpretations for similarity across data representations.

Interpretation: Bit Model

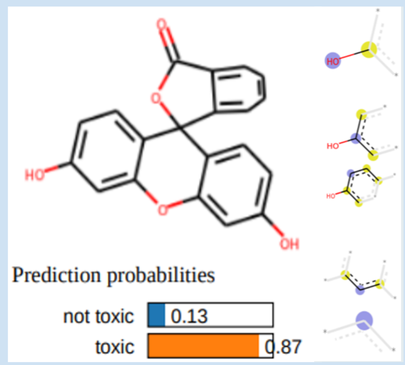


Figure 4

Interpretation: Graph Model

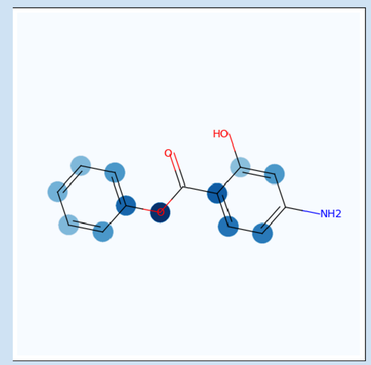


Figure 5

EXPLORING OTHER REPRESENTATIONS

- Electron density maps provide the model with a more physically descriptive input space than bit or graph representations [4].
- We plan to develop explanations for this new input and compare their interpretability against previous inputs

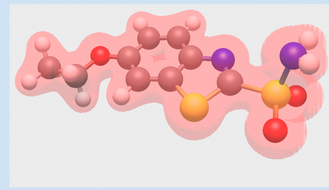


Figure 6: Electron density field of a sampled molecule.

ACKNOWLEDGEMENTS

We would like to thank Professor Feizi, Professor Srinivasan, the Gemstone faculty and staff, and all of our friends and family who helped us in our efforts.

CITATIONS