Predicting Chemical Fate Using Deep Learning

Zach Calhoun

March 4, 2022

1 Project Objective

For my final project, I want to explore how deep learning can be used to predict Henry's Law, a physical chemical property often used by environmental engineers to predict the fate of pollutants in the environment. Existing research has explored the ability of graph neural nets and variational auto-encoders to model chemical behavior well, but these have yet to be applied in the field of environmental chemistry, with most research seemingly in drug design and toxicology [1]. Thus, my goal through this project is to learn about modeling chemicals using deep learning within the context of environmental chemistry, by building a model to predict Henry's law based on chemical structure.

2 Problem Statement

Roughly 10 million new chemicals are synthesized each year, and the EPA is only equipped to screen a fraction of these chemicals for environmental risk [2]. Of particular concern are so-called PBT (persistent, bioaccumulative, and toxic) chemicals. These chemicals are unique because they do not break down, they are particularly toxic, and because they bioaccumulate, they risk making their way into the food chain [3].

Perhaps the most infamous example of a PBT is DDT, the fertilizer used during the 20th century that was found to be highly toxic to bird populations, inspiring Rachel Carson's *Silent Spring*. The more recent example making headlines today are PFAS (perfluorinated alkyl substances), which are chemicals commonly found in non-stick pans or in water-repellant clothing. In both cases, the chemicals were synthesized for a particular purpose, and the risk to society was discovered decades later. Decades later, however, is far too late. Both PFAS and DDT are now ubiquitous, found in remote locations such as the Arctic, as well as in a large percentage of the population's bloodstream [4].

To prevent this pattern from repeating itself, the EPA conducts environmental risk assessments using known and estimated chemical properties to predict chemical fate and toxicity in the environment. However, the EPA can only fully study a fraction of the compounds they need to, and estimations can be unreliable [2]. For example, the Bond Contribution Method is frequently used to predict Henry's law [5]. This method is based on a small set of chemicals, and its output may be unreliable, since it only focuses on the bonds in the molecule rather than the structure as a whole. Deep learning, and specifically, Graph Neural

Networks and Variational Auto-encoders both represent models that may be able to more effectively represent the impact of molecular structure on Henry's law.

If we could better predict this property, then environmental chemists could more accurately assess the true risk of new compounds on the market, allowing society to better prevent the widespread adoption of chemicals that pose significant environmental threats.

3 Data Summary

Henry's law (also called the air/water partitioning coefficient) is chosen because it is a well studied partitioning coefficient, meaning a lot of chemical's have values based on observations rather than estimations. I will reference two data sources to get started.

1. EPA CompTox Dashboard. The EPA maintains a database of over 350,000 chemicals and their chemical properties. From this dashboard, we would like to restrict our analysis to only the chemicals that are organic and that contain an observed value for Henry's law (rather than an estimated value). The organic compound restriction is placed to reduce the complexity of the graphs used in our analysis [6].

2. Meylan et al's 1991 paper on the Bond Contribution Method for estimating Henry's law [5]. This paper includes data for 345 organic compounds, on which the coefficients for the Bond Contribution Method are based. The Bond Contribution method is still referenced in the literature today as an estimation method for Henry's law, so performance will be benchmarked against this method.

   .

4 Approach

The approach is outlined in Table 1. The first step is perhaps the most important, since there appears to be inconsistent results for graph neural networks and variational auto encoders, so the optimal architecture should be thoroughly considered prior to development [**?**].

Table 1: Proposed timeline for the project

Date	Goal
Early March	Research Graph Neural Networks and Variational Auto-Encoders
Early/Mid March	Decide on approach
Mid March	Develop prototype of model
Late March	Test prototype and compare results to bond contribution method
Early April	Write up on results

Lastly, for simplicity, measurements will be restricted to standard temperature and pressure. Henry's law varies as a function of ambient conditions, but for the scope of this project, we will restrict our model to 25°C and 1 atm pressure. This will keep our model simple.

References

[1] M. Lovric, T. Duricic, H. T. N. Tran, H. Hussain, E. Lacic, M. A. Rasmussen, and R. Kern, "Should we embed in chemistry? a comparison of unsupervised transfer learning with pca, umap, and vae on molecular fingerprints," *Pharmaceuticals*, vol. 14, no. 8, p. 758, 2021.

[2] G. A. Burton, R. Di Giulio, D. Costello, and J. R. Rohr, "Slipping through the cracks: Why is the u.s. environmental protection agency not funding extramural research on chemicals in our environment?," *Environmental Science and Technology*, vol. 51, no. 2, pp. 755–756, 2017. doi: 10.1021/acs.est.6b05877.

[3] M. Matthies, K. Solomon, M. Vighi, A. Gilman, and J. V. Tarazona, "The origin and evolution of assessment criteria for persistent, bioaccumulative and toxic (pbt) chemicals and persistent organic pollutants (pops)electronic supplementary information (esi) available. see doi: 10.1039/c6em00311g," *Environmental Science: Processes and Impacts*, vol. 18, no. 9, pp. 1114–1128, 2016.

[4] G. Czub, F. Wania, and M. S. McLachlan, "Combining long-range transport and bioaccumulation considerations to identify potential arctic contaminants," *Environmental Science & Technology*, vol. 42, no. 10, pp. 3704–3709, 2008. doi: 10.1021/es7028679.

[5] W. M. Meylan and P. H. Howard, "Bond contribution method for estimating henry's law constants," *Environmental Toxicology and Chemistry*, vol. 10, no. 10, pp. 1283–1293, 1991.

[6] E. P. Agency, "Comptox dashboard." `https://comptox.epa.gov/dashboard/`.

[7] D. Jiang, Z. Wu, C.-Y. Hsieh, G. Chen, B. Liao, Z. Wang, C. Shen, D. Cao, J. Wu, and T. Hou, "Could graph neural networks learn better molecular representation for drug discovery? a comparison study of descriptor-based and graph-based models," *Journal of Cheminformatics*, vol. 13, no. 1, p. 12, 2021.