Predicting Chemical Fate Using Deep Learning: Progress Report

Zach Calhoun

April 3, 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 is PFAS (per-fluorinated alkyl substances), which represents the class of 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 chemicals have values based on observations rather than estimations. For this project, I searched for a dataset that provided the Henry's law constant in a standardized format. That is, the values given for the constant need to have consistent units, be taken at standard temperature and pressure, and clearly indicate the source of the measurement, as it is vital that my model is trained on observed values, rather than calculated values.

3.1 Henry's Law Constants

I found a compilation of Henry's law constants in [6]. This compilation contains over 17,000 values for over 4,600 chemicals. Each chemical had several values given, with sources including actual measurements, and a variety of estimation methods (e.g., a thermodynamical calculation was used to predict Henry's law).

The author of this compilation provided the data through the webpage cited in [7]. Since the data was provided as a large SQL query, I ran the query using PostgreSQL to construct the tables needed to extract the data. I then selected all columns from the `henry` table joined with the `species` table so that I could see a list of Henry's law values for each compound. This query was saved into a CSV file so that analysis could be done outside of SQL.

3.2 Chemical Properties

While the source mentioned above provided the Henry's law constants, this source did not provide the chemical properties needed to build a model. Most notably, many examples of predicting chemical properties from structure in machine learning make use of the Canonical SMILES (Simplified Molecular Input Line Entry System) representation of the chemical. The SMILES representation is a string representation of the chemical from which one can determine both the atoms comprising the molecule as well as the structure of the molecule, with the Canonical flavor of SMILES referencing the fact that each compound's SMILES representation is unique [8].

To get this representation for each molecule, I used the package PubChemPy [9]. This open-source project is a package that accesses PubChem, an online search tool for obtaining chemical information, using PubChem's Power User Gateway [10]. This data source is assumed reliable since it is managed by the National Institute of Health, and the open source package is merely a convenient method for accessing this data.

Upon using PubChemPy, it was discovered that compound queries provided extra information that could be leveraged to predict Henry's law. For each compound, PubChem provides the complexity, molecular weight, TPSA, hydrogen bond acceptor count, hydrogen bond donor count, and a covalent unit count. While other information was provided, I only kept these features from each query as these values pertain to the chemical structure and physiochemical properties of the molecule, which should be useful for building my model.

4 Preprocessing the data

Given that there was one primary dataset, and an auxiliary dataset to fill in gaps, significant preprocessing was required to complete the dataset.

4.1 Joining the Data

First, the compilation of Henry's law constants was reduced to unique compounds with measured values. There were two values for two compounds, so the measurements were averaged to ensure one measured value per compound. Following this reduction, there were 1,025 compounds left over.

For each compound, PubChem was queried to get further information using the chemical's unique identifier (the Inchikey), which the compilation contained as a unique reference. For the 1,025 compounds, only those with 1 unique match in the PubChem database were kept. This resulted in a dataset containing 960 compounds, which was the final dataset used in my analysis.

4.2 Log Transform

Values for Henry's law fall in the range of 0 to infinity. If the value is zero, the compound may vaporize almost immediately out of its aqueous phase into the atmosphere, and an infinite value indicates that the compound strongly prefers the aqueous phase. I use the term infinite here to indicate the theoretical limit, but in practice, a strongly hydrophilic, or water-loving, compound will have a very large constant. In this dataset, the range of observed values is $(1.2 \times 10^{-7}, 2.3 \times 10^{10})$.

To address this large difference, I performed a log-transform on the Henry's law values, to get the distribution shown in Figure 1. This additionally makes sense from a practical point of view; environmental engineers often care about the order of magnitude rather than the actual value when calculating Henry's law.

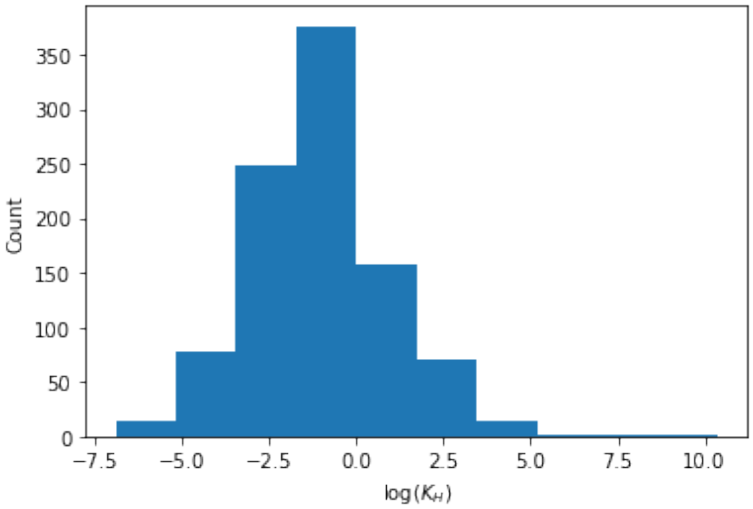

Figure 1: The Distribution of $\log K_H$, where K_H refers to the Henry's Law Constant.

5 Methods and Results

The dataset was compared to the estimated values where they existed in the Henry's law compilation to get a baseline mean squared error (MSE) of 0.2-0.8, depending on the method. However, both measured and estimates only existed for a small subset of the data.

An initial model was built using just the physical properties provided by PubChem. For this initial attempt, the data was split into an 80/20 train/test set. I initially wanted my model to be interpretable, so I tried a simple decision tree regressor model. With a max depth of 6, this model had an MSE of 1.5. I then built a Random Forest Regressor, which decreased the MSE to 1.4. This performance was mostly invariable to changing hyper-parameters, so this set a base line for my models.

Next, I wanted to see how well a model could perform just using the structure. To do this, I converted each chemical to its embedding using the Mol2Vec pre-trained embeddings, as referenced in [11]. The embeddings are a 300 dimensional representation of each substructure within a molecule. For each compound, the substructures were summed to get one 300 dimensional vector for each compound. I then created a KNeighborsRegressor, with the optimal number of neighbors equal to 8 to get a test set MSE of 1.4.

Because 300 dimensions is likely too many dimensions for a dataset of 960 examples, I then ran Principal Component Analysis to attempt to reduce the dimensionality. About 5 principal components explained 80% of the variance, so K Nearest Neighbors was attempted on this reduced dimensionality dataset. This time, an optimal MSE of 1.5 was found with 10 neighbors.

My initial results seems promising, but I believe the information from the molecule substructures might be better aggregated (as opposed to summing the molecule vectors), and by exploring methods for dimensionality reduction.

6 Conclusions

I have created models using physical chemical properties, and models only using the structure, and these models perform very similarly. Given that my initial goal was to learn how to predict environmentally relevant chemical properties using only structure, my initial results suggest that this approach can be applied with success.

My next steps will be to determine whether alternative methods could improve performance. I suspect that my dataset size is preventing performance from increasing, so my focus will be on dimensionality reduction techniques and refined application of the embeddings provided by the Mol2Vec package. Previous research has demonstrated the efficacy of Variational Auto-Encoders applied to chemical footprints to enhance prediction of chemical toxicology [1]. I believe a similar approach may be useful in this domain.

After applying this final approach, I will assimilate my approaches and compare them to the estimation approaches to further explore how my models are performing. I have created several models already, so further analysis is required to evaluate these models to understand the results so that my final report will contain more thorough analysis to better understand this problem space.

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] R. Sander, "Compilation of henry's law constants (version 4.0) for water as solvent," *Atmospheric Chemistry and Physics*, vol. 15, no. 8, pp. 4399–4981, 2015.

[7] R. Sander, "Henry's law constants: Download." `https://henry.mpch-mainz.gwdg.de/henry/download.html`, April 2022.

[8] N. M. O'Boyle, "Towards a universal smiles representation - a standard method to generate canonical smiles based on the inchi," *Journal of Cheminformatics*, vol. 4, no. 1, p. 22, 2012.

[9] Various, "Pubchempy." Online, April 2012.

[10] S. Kim, P. A. Thiessen, E. E. Bolton, and S. H. Bryant, "Pug-soap and pug-rest: web services for programmatic access to chemical information in pubchem," *Nucleic acids research*, vol. 43, no. W1, pp. W605–W611, 2015. 25934803[pmid] PMC4489244[pmcid] gkv396[PII].

[11] S. Jaeger, S. Fulle, and S. Turk, "Mol2vec: Unsupervised machine learning approach with chemical intuition," *Journal of Chemical Information and Modeling*, vol. 58, no. 1, pp. 27–35, 2018. doi: 10.1021/acs.jcim.7b00616.