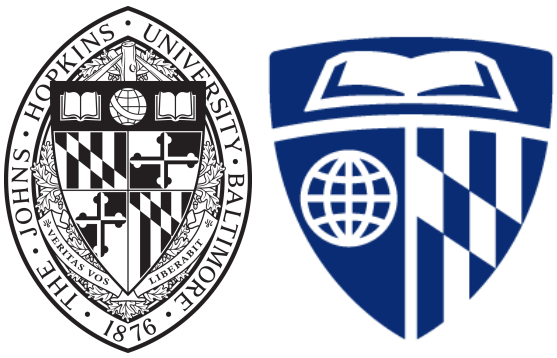
# Relationships across Chemical Exposure Concentrations: Modeling the PFAS Chemical Group Through Classification and Regression

by

Chloe Warren

Master of Science in Engineering in Data Science

2022




Johns Hopkins University

Department of Applied Mathematics and Statistics

**Abstract**

Per- and polyfluoroalkyl substances (PFAS) are a group of synthetic chemicals that have been heavily used for manufacturing purposes since the 1950's[1]. The extent of human exposure to this chemical group is overwhelmingly large. Notable contamination concentration levels have been found in indoor and ambient air, house dust, drinking water, and food[2]. One of the greatest concerns surrounding the PFAS group is that the majority of health implications associated with exposure to such chemicals are widely inconclusive as of today[3]. This is due to many factors, but especially as a result of how little is still known about the PFAS chemical group. One issue that has been brought up is the lack of historical data on concentration levels, amongst other missing properties, such as undocumented chemicals and missing lab data[3]. Our goal was to see if we could find a promising method to fill in missing concentration levels from the 2013-2014 NHANES laboratory data on a chemical within the PFAS chemical group. We were able to achieve a model with 87.86% accuracy on classification of Perfluorodecanoic acid detection using a Stepwise Logistic Regression model with 3 forward steps. This led us to believe that the results showed promising options for filling in missing data gaps for other PFAS-related research opportunities.

**Introduction**

NHANES is a survey on health and nutrition information collected from a randomly selected population within the U.S. household population[4]. The NHANES [National Health and Nutrition Examination Survey] program is run by the NCHS [National Center for Health Statistics], a part of the CDC [Center for Disease Control and Prevention]. Amongst their available laboratory data, 8 PFAS chemicals are included. We set out to model both the binary output of detect versus non-detect, as well as the concentration levels recorded.

More than 3000 PFAS chemicals have been introduced to the global market since the 1950's[3]. With significant exposure found in communities that are especially at risk due to location, and little known about the definite health effects related to exposure to PFAS chemicals, there is a pressing need for more research on this chemical group. One study found significant exposure with some outgoing drinking water having sums of PFAS above 10,000 ng/L originating from firefighting foams in Ronneby, Sweden[5]. Our paper is less concerned with the health effects of the PFAS group given our lack of chemical expertise. PFAS chemicals are found in water not only from runoff and disposal, but also due to their water solubility and persistence towards a stable end product, it is often transported through water currents and aerosols that led to distribution in wildlife and humans[3].

Our paper is less concerned with the health effects of the PFAS group given our lack of chemical expertise, and more concerned on what kinds of research could be conducted on said health effects using imputed or modeled data to fill in historical gaps and levels of unknown chemicals within the PFAS group. One large issue is that since these chemicals are synthesized at a large

rate for industry purposes, many of the chemical identities remain unknown so there are not adequate regulations or accurate testing ability in place for this group[3]. This exploration is in response to the call to action for researchers to develop methods to fill in data gaps on the PFAS group as well as modeling levels for unknown chemicals that do not yet have testing methods.

Originally, we had planned to only use multiple imputation to achieve this goal. However, during our data exploration phase, we found promising results from predicting binary detection values of Perfluorodecanoic acid (Figure A) with logistic regression on other PFAS chemical levels. We decided it was worthwhile to weigh this option of modeling as just as important to explore as multiple imputation.

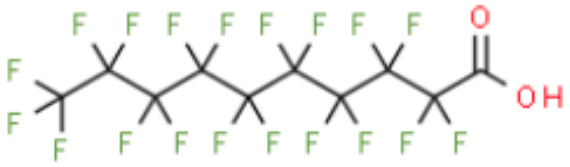



**Figure A: Molecular Structure of Perfluorodecanoic acid**

## Methods

### Data

We used data from the NHANES 2013-2014 laboratory data available on the CDC's website. NHANES is a survey on health and nutrition information collected from a randomly selected population within the U.S. household population. The NHANES [National Health and Nutrition Examination Survey] program is run by the NCHS [National Center for Health Statistics], a part of the CDC [Center for Disease Control and Prevention].

The selection process consists of 4 stages. The first stage PSUs [Primary Selection Units] are selected from a frame of all counties, the second stage PSUs are selected from area segments defined by 2000 census data, the third stage goes into DUs [Dwelling Units] on which the randomly selected subsample is screened for potential sampled participants, the fourth stage is subsample selection based on the results of the screening[4]. The screening rate is designed to produce the desired diversity within the sample, which is further outlined in the sample design documentation[4].

For our analysis, we sized down the original data selected from the site by only including chemical groups we found to be useful for investigation. This was done with the help of Dr. Judy LaKind, an exposure scientist introduced to us through our advisor, Dr. Daniel Naiman.
We further edited down the dataset eventually used to perform this project, which will be later discussed. After this further development, the largest dataset used for this project consisted of

2124 subjects with 66 chemical measures. This dataset included the PFAS group, as well as other outside chemical groups, such as Phthalates and Plasticizers Metabolites.

**Preprocessing**

Since we were dealing with both binary and continuous measurement versions of each chemical, We decided to make a threshold to decide to keep only the binary version for chemicals where over 70% of the samples were considered non-detect. After choosing these chemicals, we removed the continuous concentration level of the sample record for said chemicals. Non-detect versus detect was decided based on if the concentration level was below the lower limit detection level recorded for the chemical in question (Perfluorodecanoic acid) on the NHANES documentation.

For all of our models, both classification and regression based, the natural log of the concentration levels were taken before any training was done. For specific models (Nearest Neighbors, 5 Nearest Neighbors, Lasso, Naive Bayes, Neural Network) we used standard scaled versions of the logged data.

**Chemical Selection**

The NHANES Laboratory database is extensive and choosing the right chemical groups for this project needed a professional's opinion. Dr. Daniel Q Naiman and Dr. Judy LaKind are credited with designing and overseeing the selection of all chemicals used in our study. Naiman and Lakind introduced us to this project and stayed heavily involved throughout the process to give their expert advice and leadership.

From the 82 available data files on the database, 3 were used in the main analysis. Earlier versions of the main code that drove this analysis included significantly more files, but led to the decision that the large majority of chemicals used for modeling were poor features. This selection was based on both LaKind's knowledge of chemical groups related to the PFAS groups, as well as supplemental code written by Naiman to find the chemicals that shared the same test subjects as well as optimizing shared missingness with the PFAS group. The missing values were then dropped from the dataset for the modeling purposes, but imputation of missing values is discussed in the partner paper to this study[refer to anna's paper].

**Datasets**

During the chemical selection process, the question came up of how modeling would differ depending on what kind of dataset we used for our modeling. Thus 3 datasets were developed. The first, named combination dataset, was the dataset created by Naiman; this dataset consisted of 2124 subjects with 66 chemical measures. The second, named PFAS dataset, was created by removing from the combination dataset any chemical measures that were not from the PFAS group; this dataset consisted of 2124 subjects with 16 chemical measures. The third, named

outside dataset, was created by removing the PFAS dataset from the combination dataset; this dataset consisted of 2124 subjects with 50 chemical measures.

## Models

For all of our classification models we randomly selected 80% of the data to be training data and 20% of the data to be testing data. For all of our Regression models we randomly selected 20% of the data to be our testing data, and split the remaining 80% into 20% development data and 80% training data.

In total, 13 classification models and 5 regression models were used in this project. The classification models were not highly developed beyond their default packages in Sklearn, the regression models were all developed for improvement beyond their default packages in Sklearn. Therefore, for the sake of efficient summary, we will only be going through the best performing classification model and the best performing regression model.

The best performing classifier was stepwise logistic regression with 3 forward steps performed on the combination dataset (Figure 1.2). The selection of these features were based on adding the feature (chemical) that resulted with the lowest BIC score and highest log likelihood. This was done until 3 were added to the model.

The best performing regressor was lasso regression performed on the combination dataset (Figures 2.1-2.2). This model was improved from its original accuracy by optimizing the alpha parameter. The goal of the Sklearn lasso algorithm is minimize the following objective function:

$$min_w \frac{1}{2n}||Xw - y||_2^2 + \alpha||w||_1$$

Where n is number of samples, X is the x-value, y is the y-value, α is a constant, and $||w||_1$ is the $l_1$-norm of the coefficient vector. We found the optimal alpha by utilizing Sklearn's lassoCV model which selects the best possible model parameters to use via a cross-validation estimator. The optimal alpha was different for each run given that the training set was selected at random each time, so there is no "one size fits all" selection for alpha. Figure 3.1 shows the breakdown of lasso coefficients as a function of alpha for one scenario in which an alpha of 0.010311061330098667 was selected.

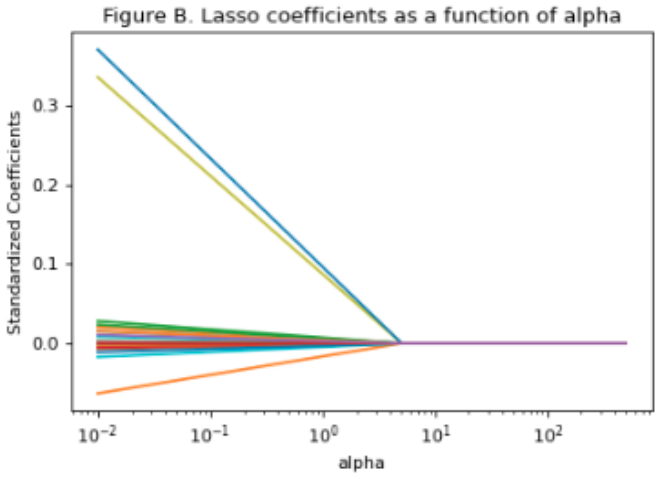



**Figure B. Lasso Coefficients as a Function of Alpha**

## Results

| Classifier | Accuracy on Combination dataset | Accuracy on PFAS dataset | Accuracy on Outside dataset |
|---|---|---|---|
| Logistic Regression | 85.929412 | 81.105991 | 70.164706 |
| Stepwise Logistic Regression (3 steps) | 87.858824 | 83.548387 | 70.164706 |
| Stepwise Logistic Regression | 86.211765 | 79.124424 | 70.164706 |
| XGBOOST | 86.235294 | 86.382488 | 70.047059 |
| Random Forest | 86.305882 | 85.691244 | 68.517647 |
| Nearest Neighbors | 76.800000 | 81.221198 | 59.905882 |
| 5 Nearest Neighbors | 80.376471 | 84.147465 | 67.247059 |
| Lasso | 45.552941 | 45.092166 | 70.164706 |
| LDA | 86.117647 | 86.866359 | 70.141176 |
| SVM (Linear Kernel) | 13.929412 | 20.875576 | 19.835294 |
| SVM (Polynomial Kernel) | 13.388235 | 20.875576 | 19.835294 |
| SVM (Radial Basis Function Kernel) | 13.600000 | 20.875576 | 19.835294 |
| Naive Bayes | 61.247059 | 73.870968 | 20.964706 |

**Table 1.1. Accuracy for Classification Methods Per dataset**

| Regressor | Average difference from averaged true level | R-Squared | MSE | RMSE |
|---|---|---|---|---|
| Lasso | 0.118063 | 0.683540 | 0.022488 | 0.149959 |
| Decision Tree | 0.122739 | 0.660326 | 0.024137 | 0.155362 |
| Random Forest | 0.120611 | 0.686415 | 0.022283 | 0.149276 |
| OLS | 0.119139 | 0.685822 | 0.022326 | 0.149417 |
| Neural Network | 0.119318 | 0.676679 | 0.022975 | 0.151576 |

**Table 2.1. Accuracy for Regression Methods on Combination dataset**

| Regressor | Average difference from averaged true level | R-Squared | MSE | RMSE |
|---|---|---|---|---|
| Lasso | 0.118063 | 0.689071 | 0.022888 | 0.151286 |
| Decision Tree | 0.122739 | 0.625908 | 0.027537 | 0.165943 |
| Random Forest | 0.120611 | 0.675041 | 0.02392 | 0.154662 |
| OLS | 0.119139 | 0.685026 | 0.023185 | 0.152267 |
| Neural Network | 0.119318 | 0.689483 | 0.022857 | 0.151186 |

**Table 2.2. Accuracy for Regression Methods on PFAS dataset**

| Regressor | Average difference from averaged true level | R-Squared | MSE | RMSE |
|---|---|---|---|---|
| Lasso | 0.224792 | 0.010114 | 0.075284 | 0.274379 |
| Decision Tree | 0.229072 | -0.067588 | 0.081193 | 0.284945 |
| Random Forest | 0.234025 | -0.136238 | 0.086415 | 0.293963 |
| OLS | 0.225684 | 0.001328 | 0.075952 | 0.275594 |
| Neural Network | 0.223582 | 0.016850 | 0.074772 | 0.273444 |

**Table 2.3. Accuracy for Regression Methods on Outside dataset**

Tables 1.1-2.3 summarize the numerical accuracy results that were obtained with each classifier and regressor per dataset. The outside dataset performed significantly worse than the other datasets for training on all counts. Interestingly the PFAS and combination datasets performed strikingly comparable, with regression models performing almost identically. This is most likely due to the feature selection process involved in the development of many of the regression models. Most feature selection algorithms run on the combination dataset resulted in a selection of chemicals in the PFAS group. With the exception of the outside dataset, lasso regression with an optimized alpha parameter performed the best consistently. As for classifiers, there was more of a distinction between the results trained on the combination versus PFAS datasets. With the

exception of SVM with a linear kernel, linear model-based classifiers performed very well consistently.

## Discussion

Using classification and regression to predict Perfluorodecanoic acid showed promising results. We were able to achieve high accuracy on both counts, but with specifically chosen training sets performed on one specific chemical. Due to the niche nature of our methods, this is not proof of the generalization of this technique. However, the results show promise for further development of these methods to be used in filling in data gaps on PFAS chemicals for other research opportunities. The most noteworthy takeaway from this study is to respond to the call for researchers to research methods to fill critical data gaps for the big picture of decision making, rather than filling with high certainty[3]. This is useful for other researchers that want to study health effects of the PFAS group, as well as for missing historical exposure data at individual sites with high exposure that would be interesting for epidemiological studies[3].

One of the biggest limitations of this study was that we lacked expert knowledge on the PFAS chemical group beyond research done for the purposes of contextualization. This was supplemented by guidance LaKind provided, but there are still many results we are not able to explain with certainty. One notable discovery during the analysis was finding that the chemical concentration levels of Perfluorodecanoic acid given by the NHANES laboratory data was finite. Finite in that there were a finite number of possible recorded concentration levels (ng/mL).

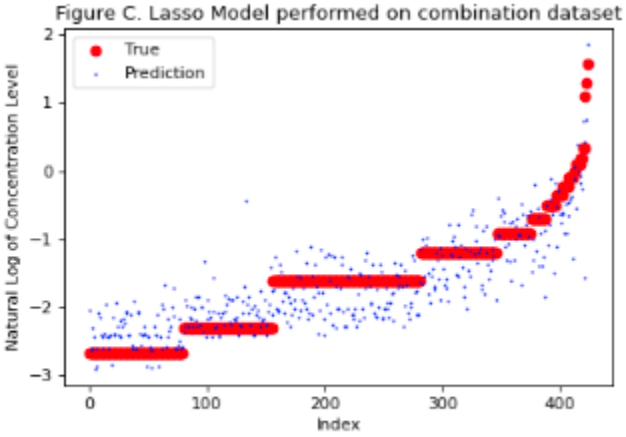
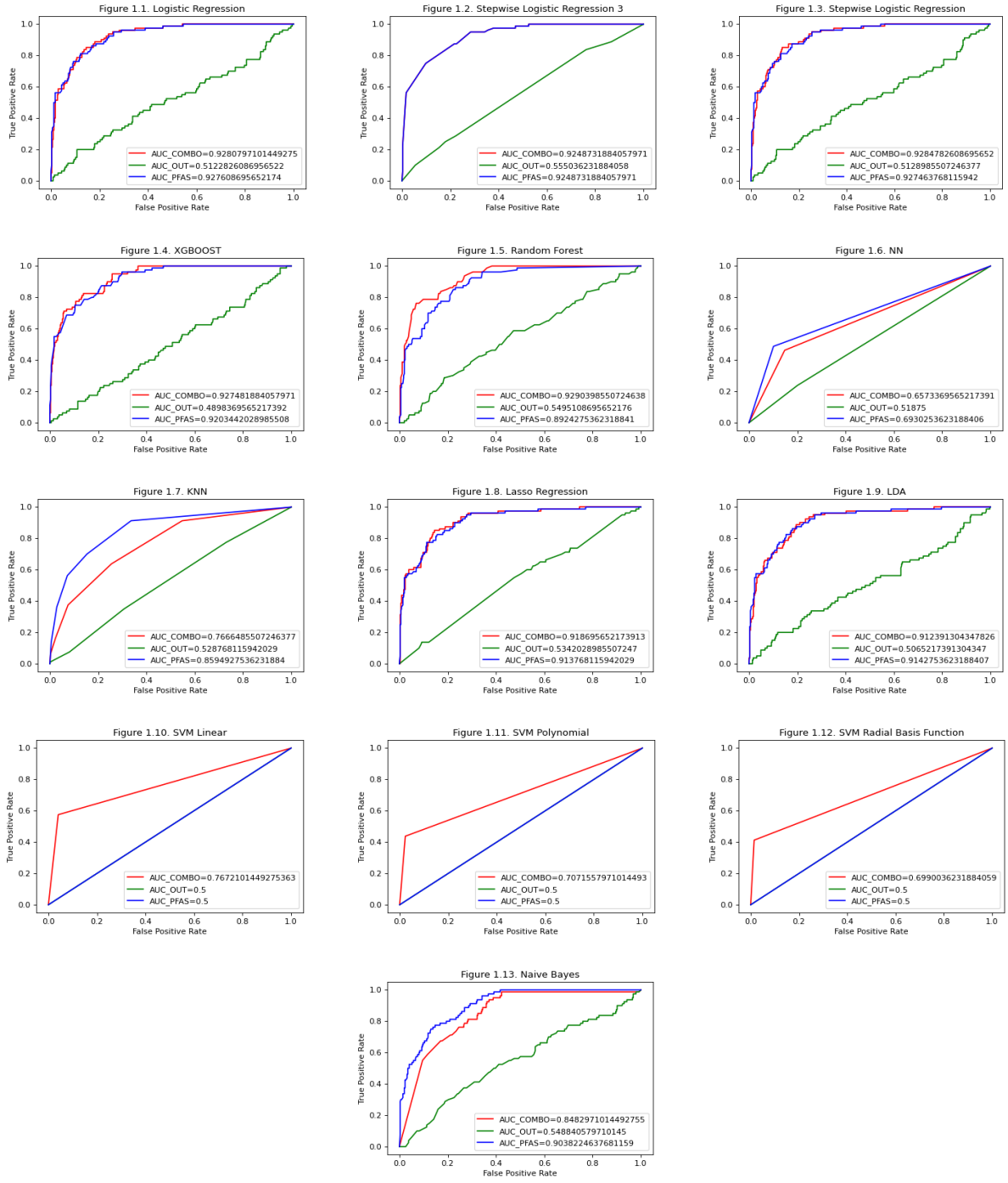


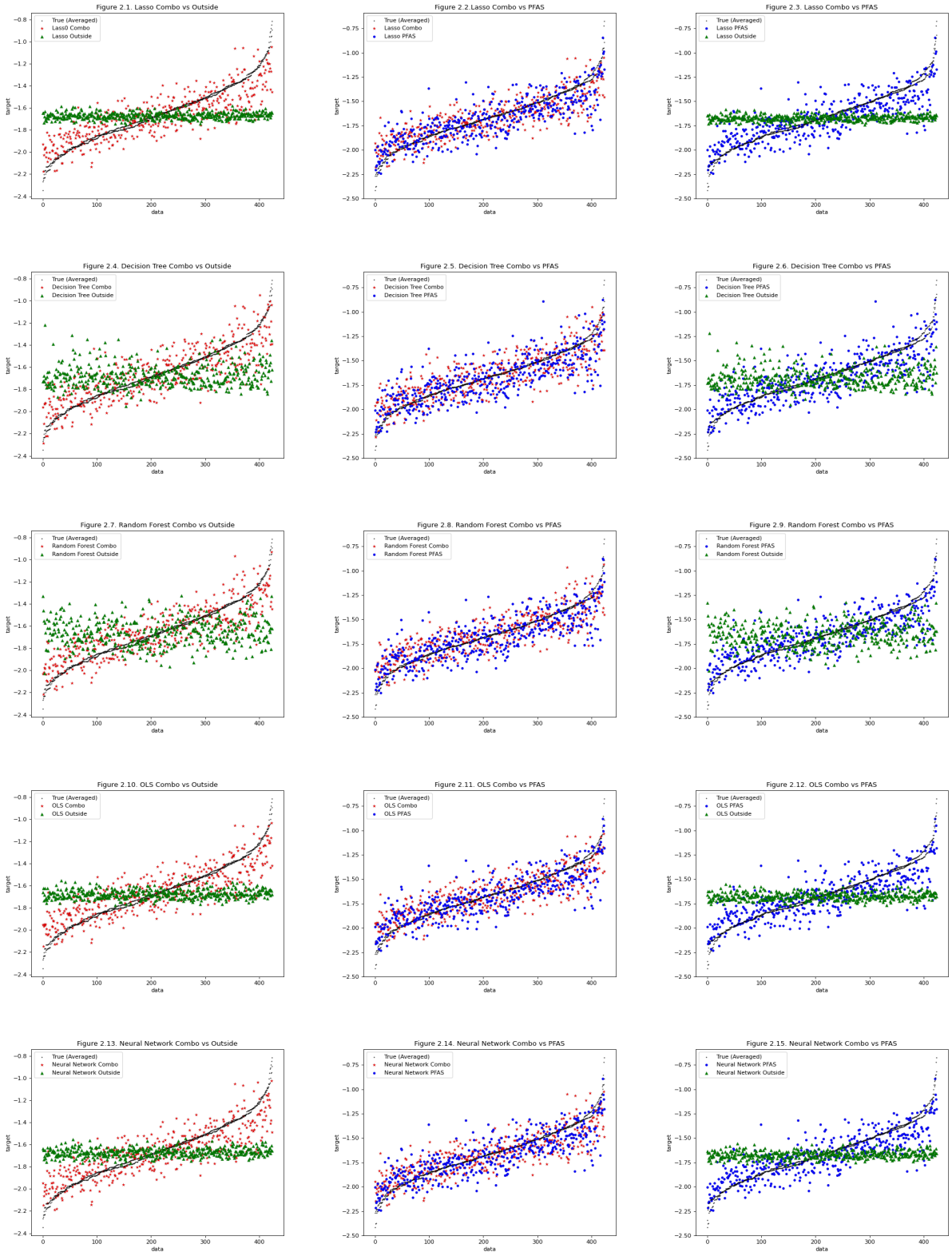
**Figure C. Lasso Model Performed on Combination Dataset**

To give more context on this phenomenon, plots were initially created to visualize regression accuracy on a single run, where two scatterplots were compared. To achieve this we joined the predictions with the actual recorded levels in a dataframe based on subject matching, then ordered by the actual level from lowest to largest, and the dataframe was reindexed. The scatterplots are the index of the record as the x-value, and the associated level as the y-value. As you can see, the actuals have finite values (Figure C). However, we could not find an explanation for this in the NHANES documentation, and do not have an explanation for this. We did run further analysis to group our predictions into the closest value in the finite levels from the training set, but found that this resulted in less accurate results. For reference, Figures 2.1-2.15 were created using averaged levels across 10 runs, which is why it doesn't reflect this phenomenon.

# Supplemental Materials




**Figures 1.1-1.13. ROC curves for different groups per Classification**
(Datasets by color: Red: Combination, Blue: PFAS, Green: Outside)




**Figures 2.1-2.15. Plots for different groups per Regression Method**
(Datasets by color: Red: Combination, Blue: PFAS, Green: Outside)

## Bibliography


1. Buck, R.C., Franklin, J., Berger, U., Conder, J.M., Cousins, I.T., de Voogt, P., Jensen, A.A., Kannan, K., Mabury, S.A. and van Leeuwen, S.P. Perfluoroalkyl and polyfluoroalkyl substances in the environment: Terminology, classification, and origins. *Integr Environ Assess Manag,* 2011, 7: 513-541. https://doi.org/10.1002/ieam.258.
2. Hermann Fromme, Sheryl A. Tittlemier, Wolfgang Völkel, Michael Wilhelm, Dorothee Twardella. Perfluorinated compounds – Exposure assessmen for the general population in western countries, *International Journal of Hygiene and Environmental Health*, Volume 212, Issue 3, 2009, Pages 239-270, ISSN 1438-4639, https://doi.org/10.1016/j.ijheh.2008.04.007.
3. Zhanyun Wang, Jamie C. DeWitt, Christopher P. Higgins, and Ian T. Cousins. A Never-Ending Story of Per- and Polyfluoroalkyl Substances (PFASs)? *Environmental Science & Technology,* 2017 *51* (5), 2508-2518.https://doi.org/10.1021/acs.est.6b04806
4. Johnson CL, Dohrmann SM, Burt VL, Mohadjer LK. National Health and Nutrition Examination Survey: Sample design, 2011–2014. National Center for Health Statistics. Vital Health Stat 2(162). 2014. (Link to file)
5. Yiyi Xu, Christel Nielsen, Ying Li, Sofia Hammarstrand, Eva M. Andersson, Huiqi Li, Daniel S. Olsson, Karin Engström, Daniela Pineda, Christian H. Lindh, Tony Fletcher, Kristina Jakobsson. Serum perfluoroalkyl substances in residents following long-term drinking water contamination from firefighting foam in Ronneby, Sweden. *Environment International,* Volume 147, 2021, 106333, ISSN 0160-4120. https://doi.org/10.1016/j.envint.2020.106333.