Electronic Supplementary Material (ESI) for Physical Chemistry Chemical Physics.

SpectraFP: A new spectra-based descriptor to aid in cheminformatics, molecular characterization and search algorithm applications.

Jefferson R. Dias-Silva,* Vitor M. Oliveira, Flávio O. Sanches-Neto, Renan Z. Wilhelms, and Luiz H. K. Q. Junior

Institute of Chemistry, Federal University of Goiás, Goiânia

E-mail: jrichardquimica@gmail.com

Text S1. Brief introduction of Random Forest, XGBoost, Extra Trees, Gradient Boosting and Neural Network

Random Forest (RF) is one of the most used algorithms in regression and classification models.[1–3] RF is an ensemble method based on a set of decision trees, with each tree having a collection of random variables. The RF algorithm employs randomness when developing the tree architecture, which results in a great diversity producing a better model when compared to other decision tree models.[4,5]

The Gradient Boosting (GB) algorithm is a machine learning technique for regression and classification problems that generates a prediction model as a collection of weak prediction models, typically decision trees. As with previous reinforcement methods, it develops the model in stages and generalizes them to enable the optimization of an arbitrary loss function. The goal of the algorithm is to generate a series of weak models, each of which aims to minimize the error of the previous model, through the use of a loss function.[6,7]

Another algorithm used to predict the functional groups was eXtreme Gradient Boosting (XGB) – a method based on a Gradient Boosting decision tree.[8]This algorithm was developed under the same GB framework and to be highly efficient, flexible, and portable. XGB provides a parallel tree reinforcement that solves many data science problems quickly and accurately and has therefore been widely used in recent literature.[9–12]

The ExtraTrees (Extremely Randomized Trees) machine learning method generates a large number of decision trees at random and then combines the outputs of each tree to determine the final answer. Its main difference is that this process is extremely random, thereby contributing to models that are more generalizable. This entire procedure is quite similar to what occurs in the Random Forest method, which likewise generates decision trees at random and uses each tree to determine the outcome. The distinction between the two algorithms resides in the amount of random processes employed by each.[13–15]

Finally, the last algorithm used in this work was the neural network (NN). This algorithm

is based on a series of units organized and connected in sequential layers.[16,17] Neural network architecture involves an input layer, several layers and an output layer. The units are the neurons, with neurons within the same layer acting in parallel and transforming the input values received from the previous layer into a scalar value.

The hyperparameters of each machine learning model used in this work are listed in JSON file called "hyperparameters.json".

Text S2. Brief introduction of similarity and Tanimoto coefficient

One of the fundamental issues in the development of QSAR/QSPR models is the model's applicability domain.[18–20] The developed models are limited to a set of data they are trained on – thus limiting their domain. The larger the data set used, the greater the applicability domain of that model. However, the data set available for a given protocol is still small. In this sense, it is useful to use methods that search for similarity between the training set used in the model and for a data set not trained by the model – either the test set or a new available data. Among these methodologies, the most used approach to calculate the similarity between two compounds is through the Tanimoto coefficient.[21] An easily understood example to show how to calculate similarity based on the Tanimoto coefficient can be found here: https://docs.eyesopen.com/toolkits/python/graphsimtk/measure.html.

Text S3. Brief introduction of SHAP

Another fundamental requirement in developing a QSAR model is understanding how the model performed a prediction. In this sense, a recent unified approach to interpreting model predictions, SHAP (SHapley Additive exPlanations),[22] which sheds light on the black box of ML algorithms, was developed to elucidate the most important features learned by the

model.[23] The SHAP approach was derived from cooperative game theory – primarily developed to estimate the importance of each player on a team.[24] For this, a reward for each player is carried out depending on their importance and their contributions to the result of a game. For our case – using SpectraFP – the Shapley values provide a solution for assigning a fair or reasonable reward to each of the spectraFP variables. The following equation is used to calculate the Shapley value Φ_i:

$$\Phi_i(p) = \sum_{S \subseteq n/i} \frac{|S|!(n - |S| - 1)!}{n!}(p(S \cup i) - p(S)) \quad (1)$$

where $p(S)$ corresponds to the ouput of the ML model to be explained using a set S of features, and n is the complete set of all features. The final contribution or Shapley value of feature $i(\Phi_i)$ is determined as the average of its contributions across all possible permutations of a feature set. The predictions for all possible subsets $S \subseteq N$ are calculated because the effect of withholding a feature depends on all other features in the model.

Text S4. SHAP analysis

The Spectral Database for Organic Compounds (SDBS)[25] was consulted for all experimental ^{13}C chemical shifts referenced in the SHAP discussions below. Furthermore, the spectra of the compounds discussed below can be accessed at https://sdbs.db.aist.go.jp.

FG2 group (Alcohol)

In the 50-80 ppm range of an NMR spectra, C-O bonds are regularly observed. Clearly, the fluctuation of a given signal is caused by the electronegativity of nearby atoms deshielding the carbon atom. For example, the signal for the C-O carbon of the 2-propanol compound is 64.04 ppm, while the similar signal for the 1-amino-2-propanol compound is 68.01. Following this logic and observing the most significant variables brought by the SHAP analysis for the FG2 group (Figure S1), it is reasonable to conclude that these variables fall within

the range of chemical shifts observed in the laboratory, thereby validating that the models make decisions based on physically coherent variables for this group. In addition, the most important variables of the two models, GB and XGB, presented very similar values, with variations of approximately 62-77 ppm in both cases. However, it is observable that the first most important variable (65.8-66.4) for the GB model had a significantly bigger influence on the model decision than the other variables, whereas the XGB model exhibited a broader distribution of importance across the other variables.

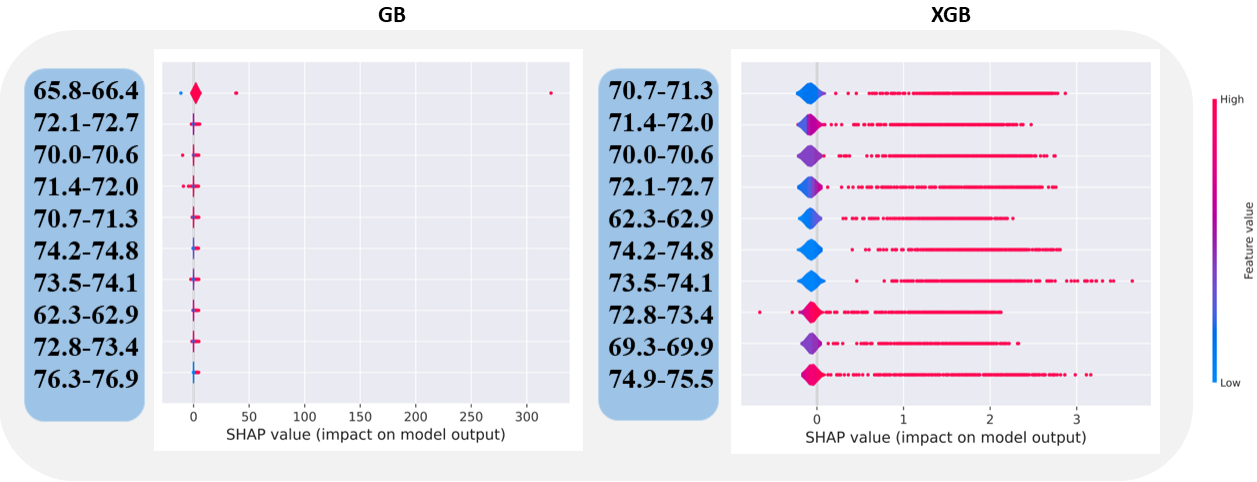

Figure S1: SHAP analysis of GB and XGB models for group FG2.

FG4 group (Carboxylic Acid Derivatives)

The FG4 group has a high level of "model understanding" complexity due to its high coverage, i.e., there are numerous forms of carboxylic acid derivatives, such as amides and the exchange of the C-OH group for a halogen (e.g., C-Br), therefore the chemical shifts to characterise this entire range of groups can vary significantly. Despite this, the typical range of ^{13}C signals from this group is between 150 and 200 ppm. Figure S2 depicts the important variables used by the GB and XGB models to determine the existence or absence of this group inside a spectra. The GB model presented variables in a wider range of spectra areas, with the smallest variable represented by 95.9 ppm and the largest variable represented by

205.7 ppm. The ppm ranges 176.4-177.0, 166.6-167.2, and 129-130.1 have a higher impact on the final decision, as the first three factors carry a far greater weight than the remaining features. The first two variables explain the vast majority of carboxylic acid derivatives, whereas the third variable refers to the alpha aromatic carbon attached to the carboxylic acid group. On the other hand, the first nine most important variables for the XGB model are concentrated between 165.9 and 172.1 ppm, resulting in less noise in the model's decision and a more uniform coverage of derivatives such as esters, amides, anhydrides, and acid halides. Moreover, the variable weights (SHAP value) are more uniformly distributed throughout the first 10 variables compared to the GB, indicating that more variables would be considered in the model outcome.

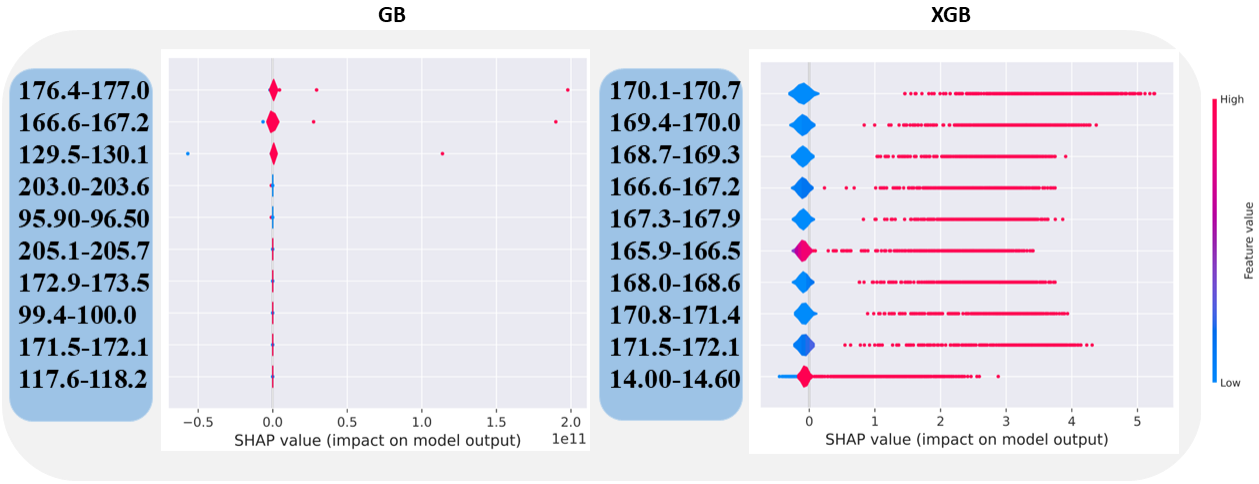

Figure S2: SHAP analysis of GB and XGB models for group FG4.

FG5 group (Carboxylic Acid or Ester)

The FG5 group includes carboxylic acids and esters since the most significant carbon (C=O) in these two groups has a similar chemical shift region. In addition, it is usual to find the alpha carbonyl carbon between 20 and 30 ppm; for instance, the alpha carbonyl carbon of propionic acid (Figure S3 A) and ethyl propionate (Figure S3 B) is 27.6 and 27.7 ppm respectively. The chemical shift of the carbons of the ester part decreases as they get further

away from the oxygen atom; carbon 4 of ethyl propionate has a chemical shift of 14.3 ppm, while carbon 2 is 60.26 ppm. It is common to find carbons similar to carbon 2 of ethyl propionate in the range between 50 to 65 ppm.

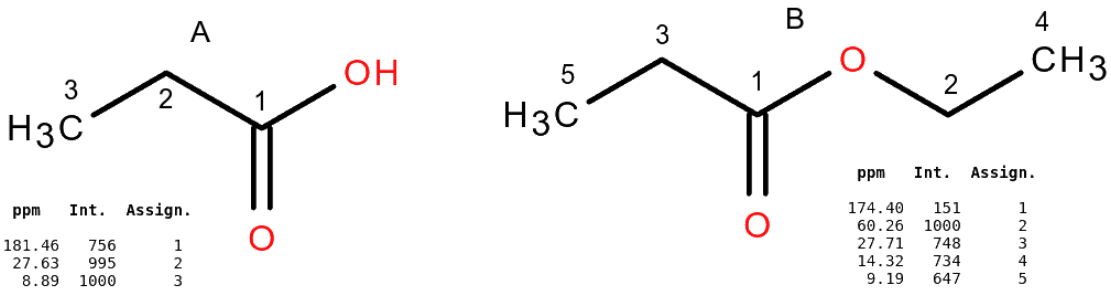

Figure S3: Structural representation of propionic acid (A) and ethyl propionate (B) and their chemical shift signals according to the SDBS database.

The SHAP analysis for the GB and XGB models pertaining to the FG5 group is depicted in Figure S4. The most significant variables in both models were nearly identical. In addition, it is reasonable to conclude that all chemical shift ranges deemed crucial for model selection are consistent with the experimental chemical shifts. Notably, the ten most relevant variables have similar weights in regard to model decisions, therefore the two models tend to make decisions based on carbonyl carbons as well as shielded carbons close to the group C=O or C-O.

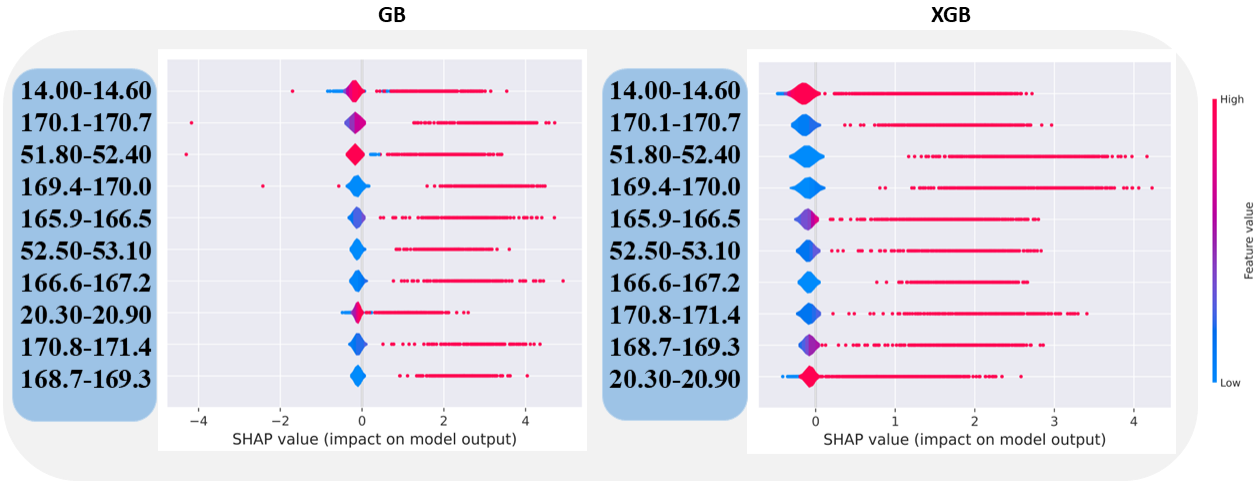

Figure S4: SHAP analysis of GB and XGB models for group FG5.

FG6 group (Aromatic compounds)

The FG6 group refers to carboxylic and heterocyclic aromatic groups. Aromatic carbons are well known and classic in ^{13}C NMR spectra, having a chemical shift range between 125 to 170 ppm. Figure S5 shows the ten most important features for the GB and XGB models; the chemical displacement ranges for the two models are nearly identical, ranging from 126.0 to 134.3 ppm for the GB model and 126.0 to 139.2 ppm for the XGB model. Thus, all of the most important variable of both models fall within the expected range of chemical shift, and it can be argued that the models forecast based on factors that are consistent.

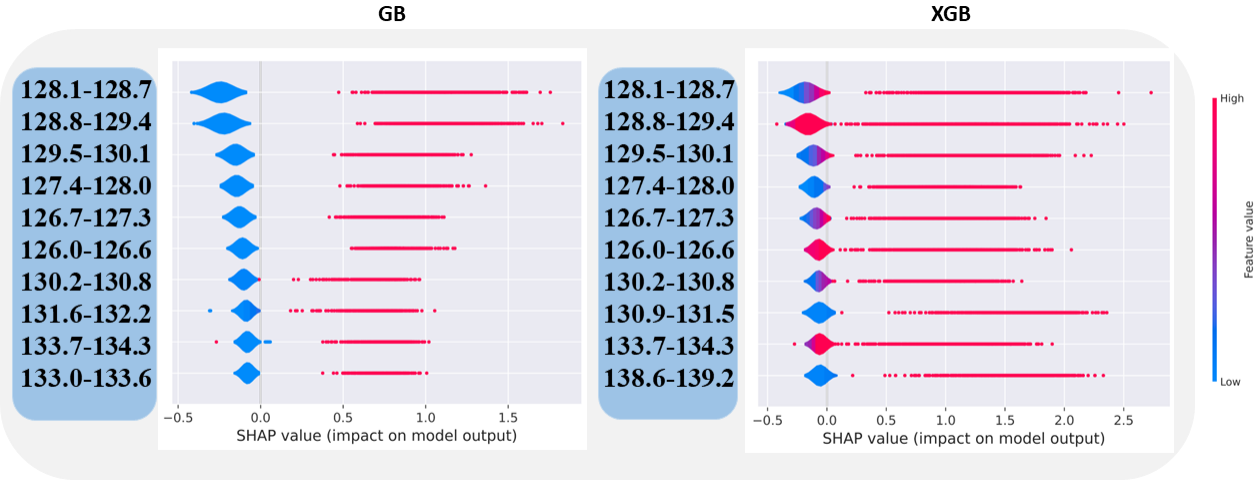

Figure S5: SHAP analysis of GB and XGB models for group FG6.

Table S1: Results of statistical parameters $F1$ and κ, from external validation, for all five models developed to predict the six functional groups.

External Validation										
	RF		**XGB**		**ET**		**GB**		**NN**	
	$F1$	κ	$F1$	κ	$F1$	κ	$F1$	κ	$F1$	κ
FG1	0.960	0.917	0.963	0.915	0.960	0.916	0.961	0.913	0.959	0.908
FG2	0.844	0.661	0.839	0.665	0.847	0.661	0.837	0.667	0.848	0.670
FG3	0.836	0.654	0.839	0.651	0.830	0.648	0.850	0.670	0.845	0.670
FG4	0.876	0.734	0.875	0.724	0.872	0.727	0.871	0.728	0.870	0.718
FG5	0.855	0.680	0.868	0.705	0.854	0.681	0.868	0.700	0.855	0.679
FG6	0.931	0.724	0.934	0.745	0.932	0.718	0.936	0.755	0.938	0.764

Table S2: External MCC values for the trained models of all FG with random sets of the dependent variable (y) drawn from three random seeds.

Y-randomization															
	RF			XGB			ET			GB			NN		
	seed 1	seed 2	seed 3	seed 1	seed 2	seed 3	seed 1	seed 2	seed 3	seed 1	seed 2	seed 3	seed 1	seed 2	seed 3
FG1	0.010	0.063	0.022	0.073	0.033	0.043	0.008	0.039	0.023	0.053	0.041	0.104	0.036	0.033	0.001
FG2	0.014	0.034	0.041	0.005	0.033	0.015	0.019	0.018	0.027	0.011	0.037	0.040	0.008	0.024	0.040
FG3	0.034	0.010	0.066	0.015	0.003	0.046	0.033	0.005	0.040	0.001	0.041	0.023	0.021	0.017	0.067
FG4	0.032	0.024	0.048	0.056	0.062	0.049	0.015	0.024	0.080	0.030	0.110	0.048	0.053	0.017	0.002
FG5	0.047	0.012	0.028	0.019	0.018	0.030	0.031	0.017	0.009	0.038	0.017	0.009	0.015	0.040	0.007
FG6	0.005	0.031	0.025	0.001	0.036	0.042	0.001	0.048	0.032	0.004	0.007	0.027	0.029	0.002	0.005

Table S3: Thresholds of similarity, number of compounds outside the AD (NCOAD), and corresponding MCC for each model for all FGs.

Applicability Domain															
	RF			XGB			ET			GB			NN		
	Threshold	MCC	NCOAD	Threshold	MCC	NCOAD	Threshold	MCC	NCOAD	Threshold	MCC	NCOAD	Threshold	MCC	NCOAD
FG1	0.15	0.917	0	0.15	0.915	0	0.17	0.927	1	0.17	0.925	1	0.15	0.908	0
FG2	0.15	0.662	0	0.15	0.668	0	0.15	0.661	0	0.15	0.668	0	0.15	0.671	0
FG3	0.17	0.662	0	0.17	0.654	0	0.17	0.653	0	0.17	0.671	0	0.17	0.672	0
FG4	0.17	0.735	0	0.17	0.724	0	0.17	0.727	0	0.19	0.742	3	0.17	0.718	0
FG5	0.18	0.680	0	0.18	0.706	0	0.18	0.682	0	0.2	0.751	8	0.18	0.679	0
FG6	0.2	0.732	0	0.17	0.749	0	0.17	0.730	0	0.17	0.760	0	0.17	0.765	0

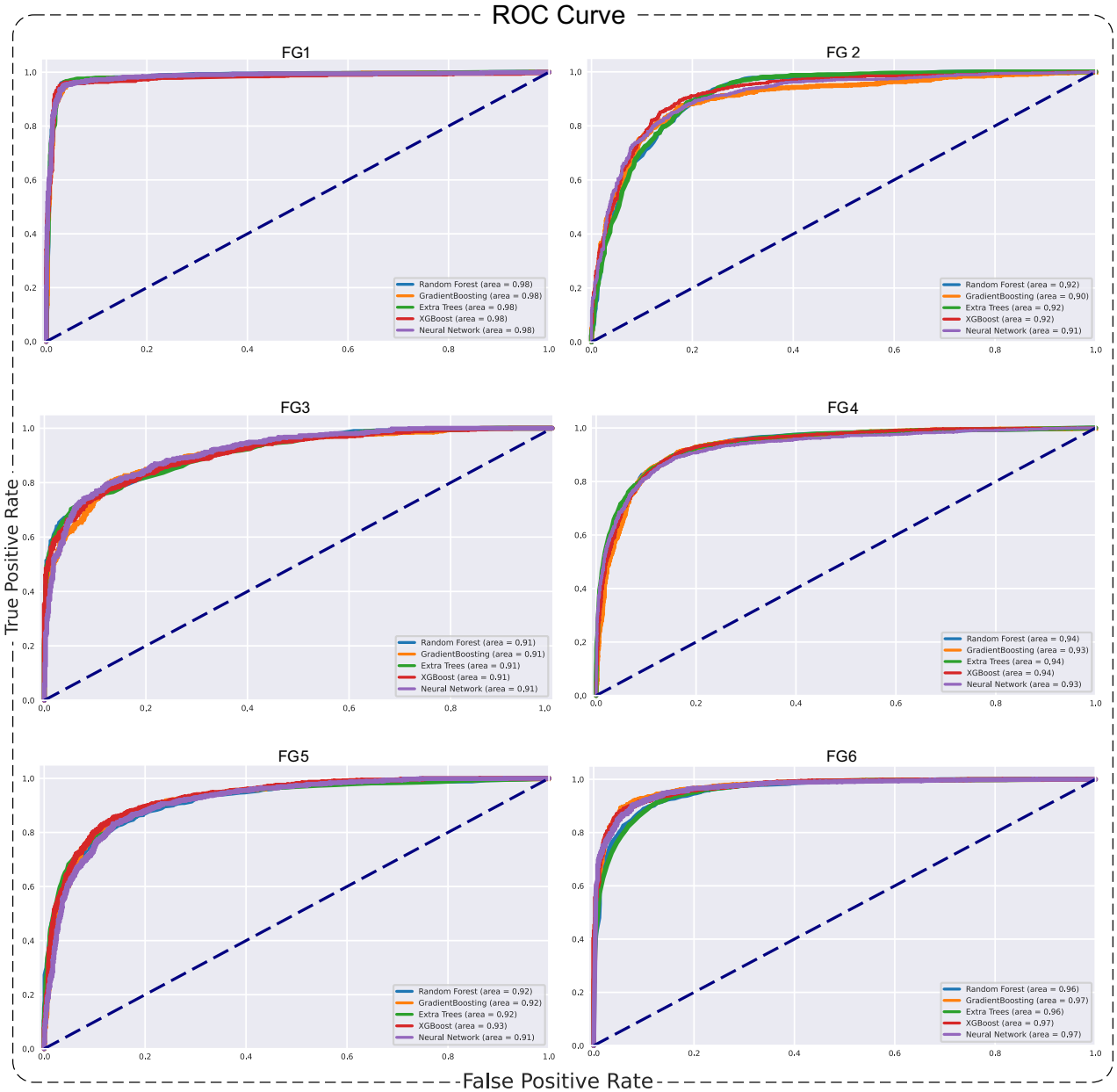

Figure S6: ROC Curve of all five ML models, as well as the respective AUC values covering all FGs, in which the value 1 of AUC represents a perfect classification and 0.5 indicates random classification. The AUC value is 0.5 when the curve is superimposed on the dashed line.

References

(1) You, H.; Ma, Z.; Tang, Y.; Wang, Y.; Yan, J.; Ni, M.; Cen, K.; Huang, Q. Comparison of ANN (MLP), ANFIS, SVM, and RF models for the online classification of heating value of burning municipal solid waste in circulating fluidized bed incinerators. *Waste management* **2017**, *68*, 186–197.

(2) Hu, S.; Liu, H.; Zhao, W.; Shi, T.; Hu, Z.; Li, Q.; Wu, G. Comparison of machine learning techniques in inferring phytoplankton size classes. *Remote Sensing* **2018**, *10*, 191.

(3) Nitze, I.; Schulthess, U.; Asche, H. Comparison of machine learning algorithms random forest, artificial neural network and support vector machine to maximum likelihood for supervised crop type classification. *Proceedings of the 4th GEOBIA, Rio de Janeiro, Brazil* **2012**, *79*, 3540.

(4) Xing, J.; Wang, H.; Luo, K.; Wang, S.; Bai, Y.; Fan, J. Predictive single-step kinetic model of biomass devolatilization for CFD applications: A comparison study of empirical correlations (EC), artificial neural networks (ANN) and random forest (RF). *Renewable energy* **2019**, *136*, 104–114.

(5) Welbl, J. Casting random forests as artificial neural networks (and profiting from it). German Conference on Pattern Recognition. 2014; pp 765–771.

(6) Friedman, J. H. Greedy function approximation: a gradient boosting machine. *Annals of statistics* **2001**, 1189–1232.

(7) Friedman, J. Stochastic gradient boosting. Department of Statistics. 1999.

(8) Chen, T.; Guestrin, C. Xgboost: A scalable tree boosting system. Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. 2016; pp 785–794.

(9) Meng, Y.; Yang, N.; Qian, Z.; Zhang, G. What makes an online review more helpful: an interpretation framework using XGBoost and SHAP values. *Journal of Theoretical and Applied Electronic Commerce Research* **2020**, *16*, 466–490.

(10) Li, P.; Zhang, J.-S. A new hybrid method for China's energy supply security forecasting based on ARIMA and XGBoost. *Energies* **2018**, *11*, 1687.

(11) Sun, B.; Lam, D.; Yang, D.; Grantham, K.; Zhang, T.; Mutic, S.; Zhao, T. A machine learning approach to the accurate prediction of monitor units for a compact proton machine. *Medical physics* **2018**, *45*, 2243–2251.

(12) Torlay, L.; Perrone-Bertolotti, M.; Thomas, E.; Baciu, M. Machine learning–XGBoost analysis of language networks to classify patients with epilepsy. *Brain informatics* **2017**, *4*, 159–169.

(13) Geurts, P.; Ernst, D.; Wehenkel, L. Extremely randomized trees. *Machine learning* **2006**, *63*, 3–42.

(14) Jiang, C.; Zhao, P.; Li, W.; Tang, Y.; Liu, G. In silico prediction of chemical neurotoxicity using machine learning. *Toxicology research* **2020**, *9*, 164–172.

(15) Barta, G. Identifying biological pathway interrupting toxins using multi-tree ensembles. *Frontiers in Environmental Science* **2016**, *4*, 52.

(16) Yang, Z.-Y.; Dong, J.; Yang, Z.-J.; Lu, A.-P.; Hou, T.-J.; Cao, D.-S. Structural analysis and identification of false positive hits in luciferase-based assays. *Journal of Chemical Information and Modeling* **2020**, *60*, 2031–2043.

(17) Wenzel, J.; Matter, H.; Schmidt, F. Predictive multitask deep neural network models for ADME-Tox properties: learning from large data sets. *Journal of chemical information and modeling* **2019**, *59*, 1253–1268.

(18) Roy, P. P.; Kovarich, S.; Gramatica, P. QSAR model reproducibility and applicability: A case study of rate constants of hydroxyl radical reaction models applied to polybrominated diphenyl ethers and (benzo-) triazoles. *Journal of computational chemistry* **2011**, *32*, 2386–2396.

(19) Gramatica, P. External evaluation of QSAR models, in addition to cross-validation: verification of predictive capability on totally new chemicals. *Molecular informatics* **2014**, *33*, 311–314.

(20) Gadaleta, D.; Mangiatordi, G. F.; Catto, M.; Carotti, A.; Nicolotti, O. Applicability domain for QSAR models: where theory meets reality. *International Journal of Quantitative Structure-Property Relationships (IJQSPR)* **2016**, *1*, 45–63.

(21) Tanimoto, T. T. Elementary mathematical theory of classification and prediction. **1958**,

(22) Lundberg, S. M.; Erion, G.; Chen, H.; DeGrave, A.; Prutkin, J. M.; Nair, B.; Katz, R.; Himmelfarb, J.; Bansal, N.; Lee, S.-I. From local explanations to global understanding with explainable AI for trees. *Nature machine intelligence* **2020**, *2*, 56–67.

(23) Zhong, S.; Zhang, K.; Wang, D.; Zhang, H. Shedding light on "Black Box" machine learning models for predicting the reactivity of HO radicals toward organic compounds. *Chemical Engineering Journal* **2021**, *405*, 126627.

(24) Lundberg, S. M.; Lee, S.-I. A unified approach to interpreting model predictions. *Advances in neural information processing systems* **2017**, *30*.

(25) Saito, T.; Hayamizu, K.; Yanagisawa, M.; Yamamoto, O.; Wasada, N.; Someno, K.; Kinugasa, S.; Tanabe, K.; Tamura, T.; Hiraishi, J. Spectral database for organic compounds (sdbs). *National Institute of Advanced Industrial Science and Technology (AIST)* **2006**,