Providing the 'best' lipophilicity assessment in a drug discovery environment

George Chang, Nathaniel Woody and Christopher Keefer
Computational ADMET Group, Medicine Design, Pfizer Inc., Groton, CT 06340, USA

Abstract:

Lipophilicity is a fundamental structural property that influences almost every aspect of drug discovery. Within Pfizer, we have two complementary high-throughput screens for measuring lipophilicity as a distribution coefficient (LogD) – a miniaturized shake-flask method (SFLogD) and a chromatographic method (ELogD). The results from these two assays are not the same (see Figure 1), with each assay being applicable or more reliable in particular chemical spaces. In addition to LogD assays, the ability to predict the LogD value for virtual compounds is equally vital. Here we present an in-silico LogD model, applicable to all chemical spaces, based on the integration of the LogD data from both assays. We developed two approaches towards a single LogD model – a Rule-based and a Machine Learning approach. Ultimately, the Machine Learning LogD model was found to be superior to both internally developed and commercial LogD models.

Introduction

Assessment of lipophilicity is fundamental to understanding the properties of a molecule. Within the drug discovery environment, additional consideration of the ionization of the compound in the aqueous phase, especially at pH 7.4, is necessary. Hence, distribution coefficient (LogD) is the lipophilicity descriptor of choice. While there are many methods for determining LogD (Kempinska, Chmiel et al. 2019), partitioning between n-octanol and an aqueous buffer (i.e. shake flask) remains the most direct method, though chromatographic methods have become popular indirect experimental approaches. To meet the demands of thousands of LogD determinations, rapid, high capacity and compound-sparing assays needed to be developed.

Within Pfizer, we have two assays for assessing LogD – a miniaturized shake flask method (SFLogD) (Stopher and McClean 1990) and a chromatographic method (ELogD) (Lombardo, Shalaeva et al. 2000, Lombardo, Shalaeva et al. 2001). Each assay has limitations that affect which assay is the most appropriate for a particular compound.

For our SFLogD assay, measured LogD values > 3 appear to be an underestimation, which has also been noted by others(Low, Blasco et al. 2016). The SFLogD underestimation is evident when comparing experimental values of SFLogD and ELogD for compounds within the Pfizer sample collection where SFLogD < ELogD when SFLogD > 3 (Figure 1). There may be multiple causes of this lower than anticipated value including cross contamination of the extracted aqueous sample by n-octanol with high compound concentration due to increased lipophilicity. Inadequate shaking/equilibration and the presence of DMSO for compound solubilization are other proposed causes of this underestimation (Low, Blasco et al. 2016).

For our ELogD assay, while it is an indirect estimation of LogD, it does not appear to suffer from high LogD limitations. However, it is limited in its inability to estimate LogD for acids or zwitterions due to specific interaction of anions with the stationary phase (Lombardo, Shalaeva et al. 2001). Internally, acids are excluded from this assay.

Given the complementary nature and limitations of these two high-throughput (HTS) assays, Pfizer has implemented a screening paradigm where compounds are first submitted to the SFLogD assay. Those with SFLogD > 3 are automatically submitted to the ELogD assay for assessment of underestimation by SFLogD. It is acknowledged that highly lipophilic acids, where only SFLogD values are obtained, may be underestimated in this paradigm.

In addition to the availability of HTS lipophilicity assays, a predictive lipophilicity model for use during the design phase is equally critical. We initially built Machine Learning models for each assay to complement the two LogD assays. However, both models suffer from the same limitations as their respective assays, namely underprediction for the SFLogD model and inability to predict acidic compounds for the ELogD model. Therefore, users of the models are stuck with the same wet data quandary when looking at the models, i.e. which model should I trust for this compound? Introduction of a single model for both assays became paramount. Analysis of commercial programs (i.e. ACD or MoKa) indicated that they were inadequate for serving as a reliable single source of predictions that matched our internal assay output. Ultimately, we launched two hybrid models, where each can provide users with a single best estimate of LogD. Development of these two hybrid models will be the focus of this paper. For differentiation and for the remainder of this paper, reference to measured will contain a ‘w’ (i.e. wet) as prefix, while predicted lipophilicity will contain a ‘c’ (i.e. computed) as prefix.

Methods

The ELogD and SFLogD models are non-linear regression models built on an internal database of compounds. The internal database contains approximately 200k unique structures with wSFLogD data and approximately 80k unique structures with wELogD data. Models on both datasets are constructed using descriptors generated by MoKa (Molecular Discovery Ltd 2020) , alvaDesc (Mauri 2020), and counts of a set of internally developed SMARTS fragments (calculated using OpenEye OEChemTK (OpenEye Scientific Software 2020). For a MoKa pKa prediction where no value is given (e.g., neutrals), the value of ApKa is set to 14 and BpKa is set to 0. Compounds that failed descriptor generation due to problematic substructures are removed from the dataset prior to model construction. Datasets are averaged by unique structure and are modeled as LogD values (i.e. untransformed). All error metrics are also calculated in the LogD space.

Models were constructed using XGBoost (Chen, T.; Guestrin, C. In *XGBoost: A Scalable Tree Boosting System*). Hyperparameters and descriptors are optimized using a 5-fold cross-validation scheme, where 80% of the data was selected for training and the remaining 20% is predicted. Model statistics are generated from these cross-validated prediction values. A final model, using all training data and the selected parameters and descriptors, is constructed and published internally for LogD predictions. Model optimization is performed only once and subsequent newly measured data is added to the training set of the published model.

To create a single LogD calculator, two variants that integrate data from the ELogD and SFLogD assays were developed - Rule-based PFLogD and Machine Learning PFLogD.

The Rule-based PFLogD model generates a prediction based on a set of heuristic rules for integrating predictions from the ELogD and SFLogD models. The basic rules are depicted in Figure 2 and were created with cooperation from assay scientists and medicinal chemists. Supplementing these rules is set of exceptions to handle corner cases that have been identified during application of this heuristic model.

The Machine Learning (ML) PFLogD model is built on a collection of LogD values that is derived from integrating predictions from the ELogD and SFLogD models. The integration follows the rules shown in Figure 2. The cSFLogD and cELogD are from 5-fold CV predictions. Only structures with measured LogD in one of the two assays are included in the training set. The supplementary exceptions used in the Rule-based PFLogD are not considered, hence measured LogD values are not directly used in the integration algorithm. A model, based on this integrated set of predictions as input, is optimized using the same methodology as described above for the SFLogD and ELogD models.

Lipophilicity models

The lipophilicity model (cSFLogD or cELogD) developed for each assay (wSFLogD or wELogD) is able to reasonably predict its respective assay as shown in Figures 3 and 5. However, the models are limited in their ability to predict the complementary assay (i.e. cELogD to predict wSFLogD or cSFlogD to predict wELogD) as shown in Figures 4 and 6. The inability of cELogD to predict acids is evident in Figure 4 where the acidic compounds (colored in red) have significantly higher cELogD compared to wSFLogD. The underestimation of SFLogD, whether experimental or predicted, is evidenced by cELogD > wSFLogD or wELogD > cSFLogD for highly lipophilic compounds, as shown in Figures 4 and 6, respectively.

The performance of these models is shown in Table 1. Statistics of the models on their own experimental data (i.e., cSFLogD on wSFLogD) are based on 5-fold cross-validation. Statistics against the complementary assay (i.e., cSFLogD on wELogD) are external predictions. There is sufficient disagreement between the experimental datasets that neither model is able to successfully predict the complementary dataset.

Analysis of commercial lipophilicity models' ability to predict our assays are shown Figures 7 – 10. For prediction of wSFLogD by ACD or MoKa, while there are no apparent non-linear deviations, both appear to overpredict at very high and underpredict at very low wSFLogD values (Figures 7 and 8). For wELogD, while the overall prediction pattern for these commercial models is closer to unity, their prediction accuracies for this dataset is poor as indicated by the low percentage within +/- 0.5 (see Table 2).

From a modeling perspective, it is more important to model our internal assays instead of an external data source - i.e. higher prediction accuracy of our internally generated LogD values even if there is a discrepancy between our internal and published values. Ultimately, the role of the model is to provide an equally accurate prediction of the experiment for all compounds and their effectiveness is measured by how well they predict the assay values. As an example, even though the underprediction of very low wSFLogD by ACD and MoKa may be justified by correlation to an external data source, for internal structure activity relationship (SAR) development, it is more important for our model to predict the eventually acquired experimental value.

With the goal of introducing a single LogD model that is able to predict both wSFLogD and wElogD, we developed two variants that integrate data from these two assays - Rule-based PFLogD and Machine Learning PFLogD.

The Rule-based PFLogD model generates a prediction based on a set of heuristic rules for integrating predictions from the ELogD and SFLogD models. The rules are described in Figure 2. Supplementing these rules was an evolving and ever-increasing set of exceptions to handle the integration of measured LogD values at different LogD ranges. The initial advantage of this model is in its simplicity and the integration of available experimental data. However, there is a high performance and management cost with this approach. Firstly, when experimental data was integrated became a critical question for the model. Batch differences, structure normalization, and data update frequency all cause significant maintenance problems and confusion. Second, constant re-analysis and refinement of these "expert" rules were required to determine if they are still valid or need to be updated.

As an alternative, a machine learning model – ML-PFLogD, was developed. This model is more consistent with current modeling approaches where structure-based descriptors of the input structure are used for the prediction. The training set for this model is derived from the cSFLogD and cELogD predictions that are integrated using the rules shown in Figure 2. The cSFLogD and cELogD are from 5-fold CV predictions. No measured values are directly used in the integration and hence, none of the exceptions used in the Rule-based PFLogD are considered. All the compounds in the training set must contain a measured LogD value in one of the two assays, and the size of the training set for this model increases as compounds are tested in these assays.

Predictions of our wSFLogD and wELogD assays by ML-PFLogD are shown in Figures 11 and 12. The desired non-linear deviation for high lipophilicity compounds in wSFLogD is present, addressing the underestimation of the wSFLogD assay. However, the underprediction of low wSFLogD compounds, present with ACD and MoKa, is absent with ML-PFLogD (Figure 11). The deviation of PFLogD prediction

of wELogD is linear across the range of lipophilicity and appears to be smaller than those of ACD and MoKa (Figure 12). The prediction statistics of these models against both datasets are summarized in Table 2. While ACD and MoKa are inadequate at predicting either internal assay, the ML-PFLogD model performs well against both datasets.

Summary

Pfizer has implemented two complementary high-throughput lipophilicity assays – wSFLogD and wELogD, and a screening paradigm to address the limitations of each assay – i.e. underestimation of high wSFLogD, and the overestimation of acidic compounds in wELogD. We have also built LogD models for each of these two assays. While each model performs well against its own assay, it performs poorly against the complementary assay. We developed two hybrid models – Rule-based PFLogD and Machine Learning (ML) PFLogD, in order to provide users with a single, best LogD prediction that was applicable for all compounds. The input for these hybrid models is a blend of experimental and/or predicted SFLogD and ELogD values, where the integration follows a set of heuristic rules. The Rule-based PFLogD, while simplistic in its prediction algorithm, has a high cost in performance and management. The machine learning based ML-PFLogD was found to be superior to commercial models (e.g. ACD or MoKa) and our internal models for predicting the LogD values generated by our assays, as well as simpler to maintain.

Reference:

Kempinska, D., et al. (2019). "State of the art and prospects of methods for determination of lipophilicity of chemical compounds." Trends in Analytical Chemistry **113**: 54-73.

Lombardo, F., et al. (2001). "ElogDoct: A Tool for Lipophilicity Determination in Drug Discovery. 2. Basic and Neutral Compounds." Journal of Medicinal Chemistry **44**: 2490-2497.

Lombardo, F., et al. (2000). "ElogPoct: a tool for lipophilicity determination in drug discovery." Journal of Medicinal Chemistry **43**(15): 2922-2928.

Low, Y., et al. (2016). "Optimised method to estimate octanol water distribution coefficient (logD) in a high throughput format." European Journal of Pharmaceutical Sciences : Official Journal of the European Federation for Pharmaceutical Sciences **92**: 110-116.

Mauri, A. (2020). alvaDesc: A tool to calculate and analyze molecular descriptors and fingerprints, Humana Press Inc.

Molecular Discovery Ltd (2020). "MoKa pKa " 3.2.6.

OpenEye Scientific Software (2020). OpenEye Toolkits 2020.1.0.

Stopher, D. and S. McClean (1990). "An improved method for the determination of distribution coefficients." Journal of Pharmacy and Pharmacology **42**(2): 144-144.

Table 1 - Performance of SFLogD and ELogD models against experimental data

	Predicted SFLogD		*Predicted ELogD*	
	R^2	% within 0.5	R^2	% within 0.5
Experimental SFLogD	0.889	85.4	0.413	49.1
Experimental ELogD	0.699	47.3	0.905	83

Table 2 - Comparison of ACD, MoKa and Machine Learning PFLogD models against experimental data

	ACD		*MoKa*		*ML PFLogD*	
	R^2	% within 0.5	R^2	% within 0.5	R2	% within 0.5
Experimental SFLogD	0.517	33.4	0.594	41.3	0.816	71.0
Experimental ELogD	0.489	31.3	0.549	38.6	0.859	75.9

Figure 1 – Experimental SFLogD vs Experimental ELogD. Magenta lines are +/- 1.

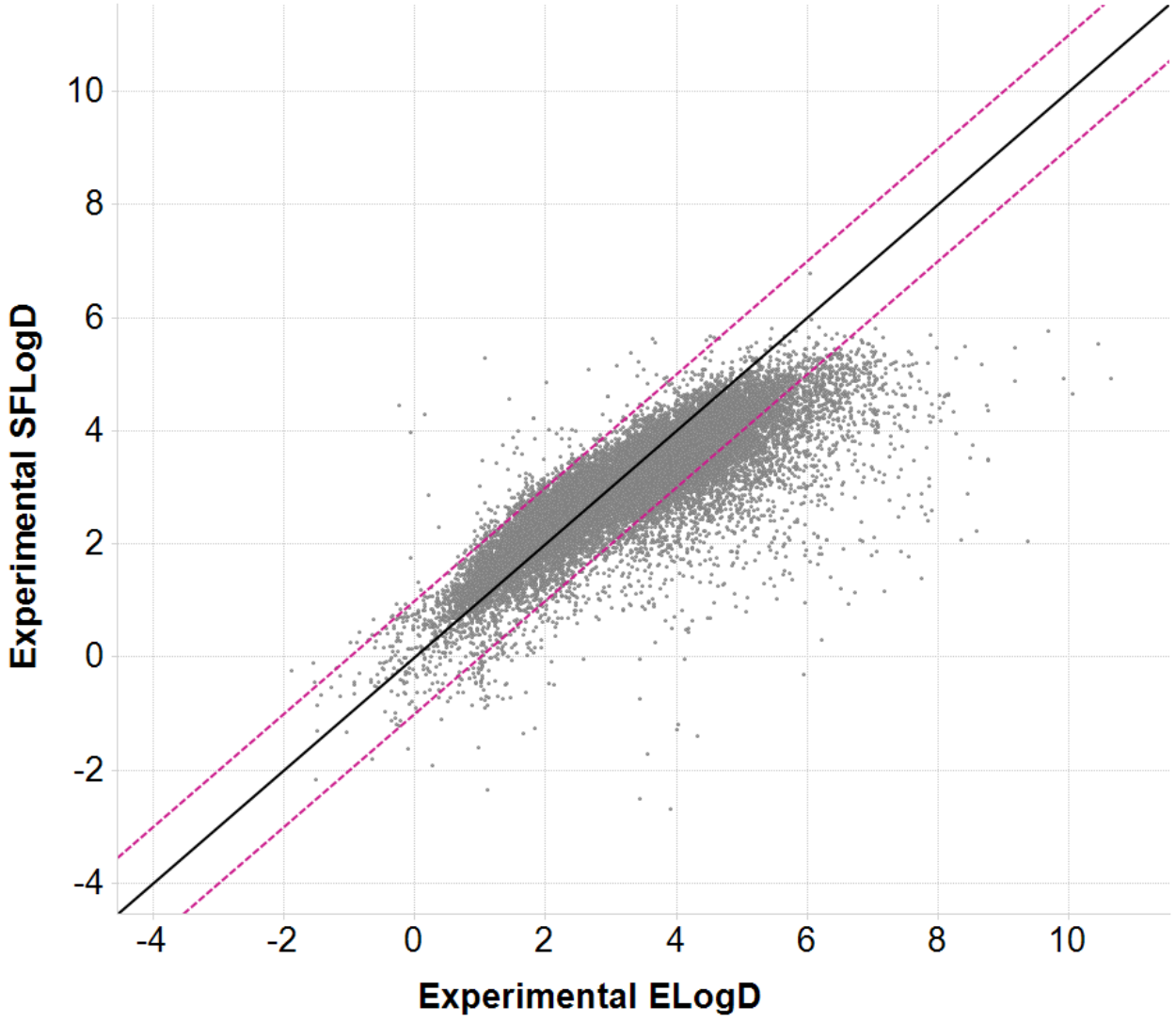

Figure 2 – Heuristic rules for integrating experimental or predicted SFLogD and ELogD for PFLogD models

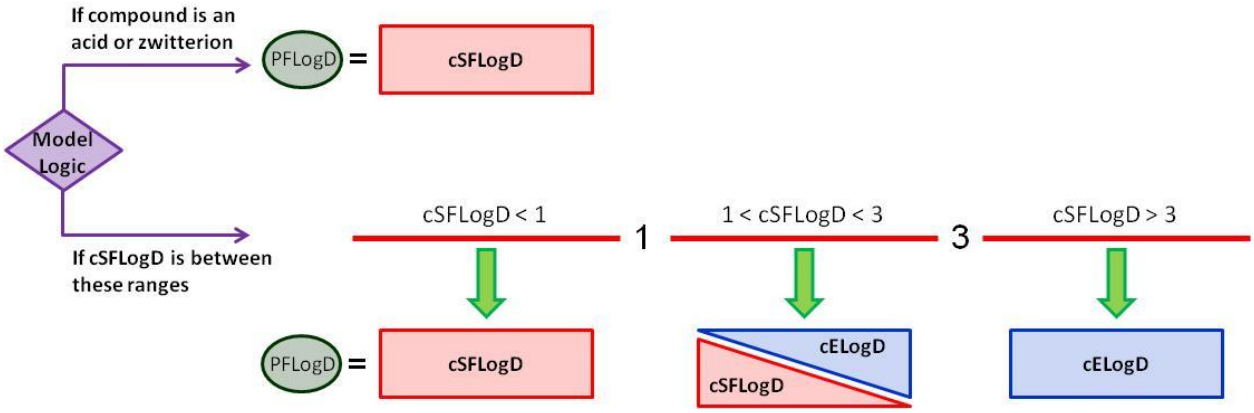

Figure 3 – Experimental SFLogD vs SFLogD Model. Magenta lines are +/- 1.

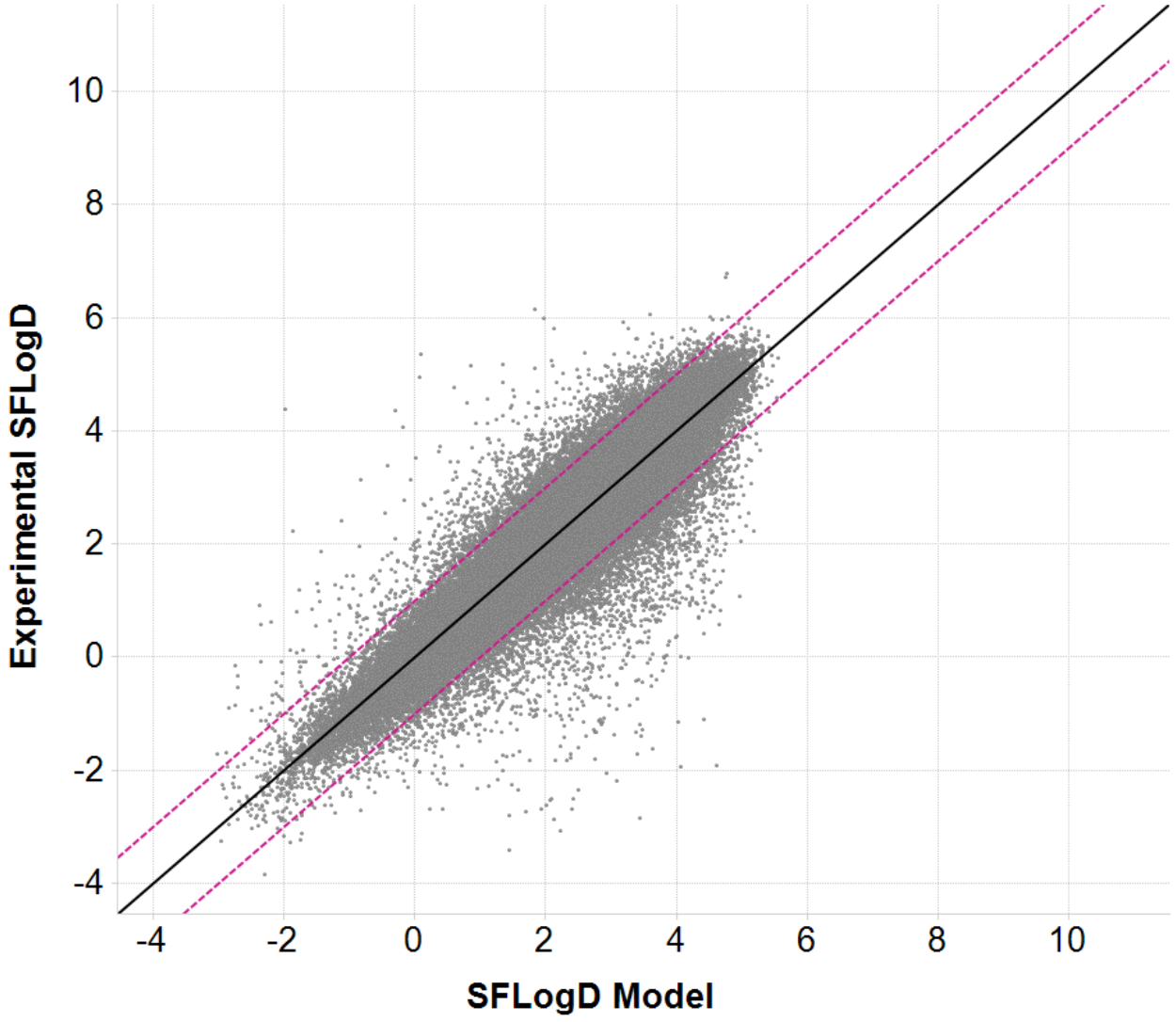

Figure 4 – Experimental SFLogD vs ELogD Model. Magenta lines are +/- 1. Points are colored by ionization where blue is basic, green is neutral and red is acidic.

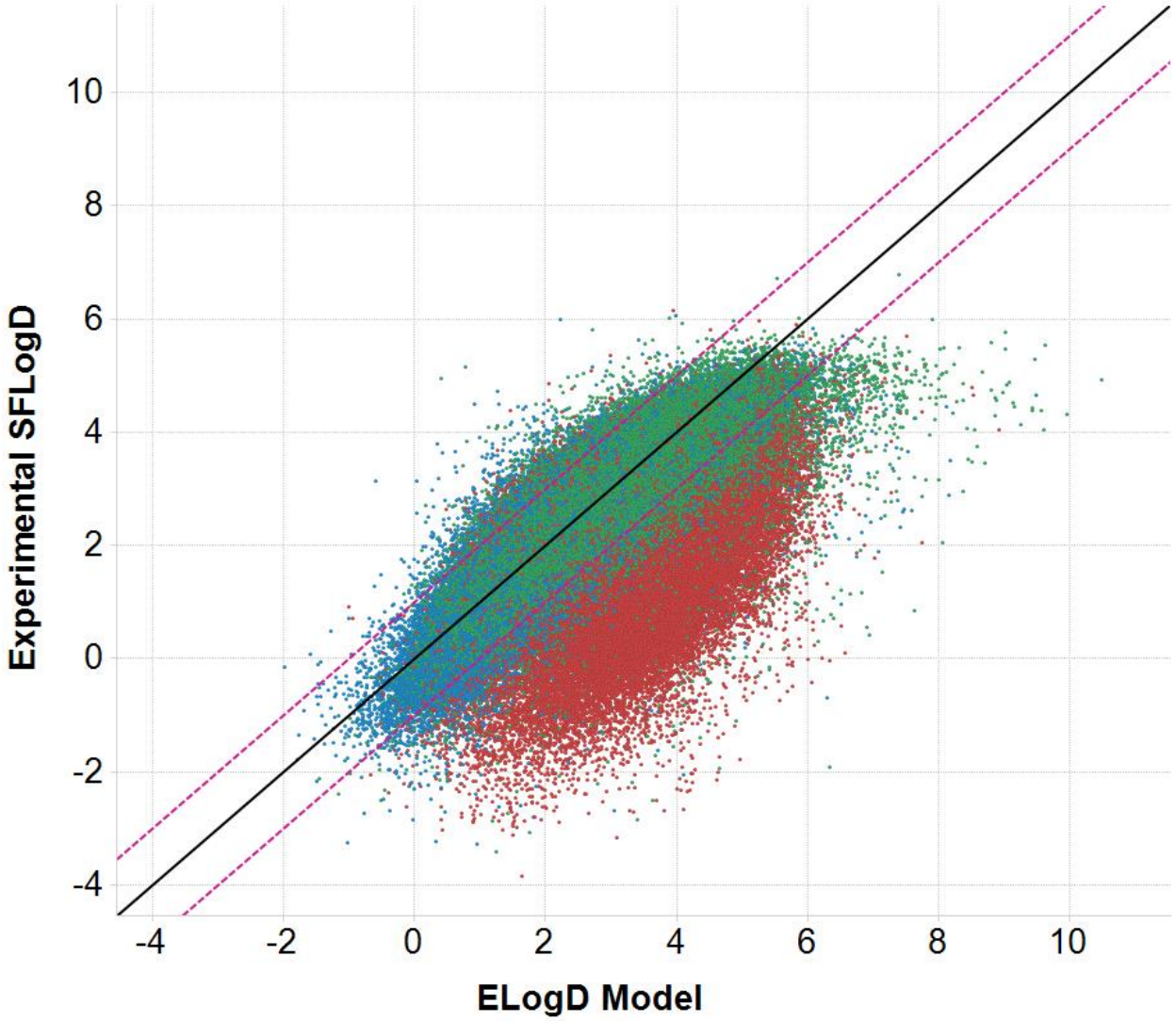

Figure 5 – Experimental ELogD vs ELogD Model. Magenta lines are +/- 1.

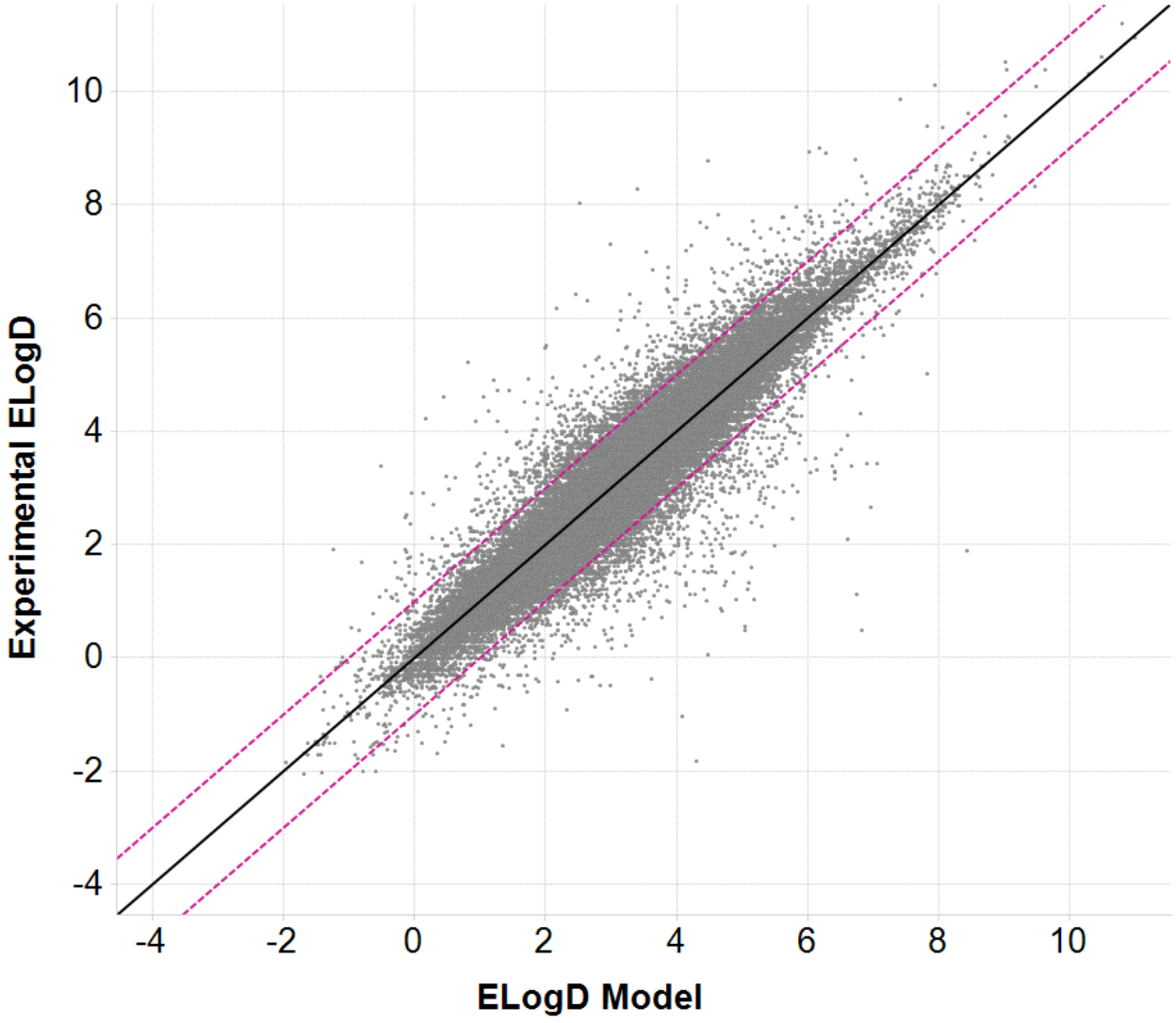

Figure 6 – Experimental ELogD vs SFLogD Model. Magenta lines are +/- 1. Points are colored by ionization where blue is basic and green is neutral.

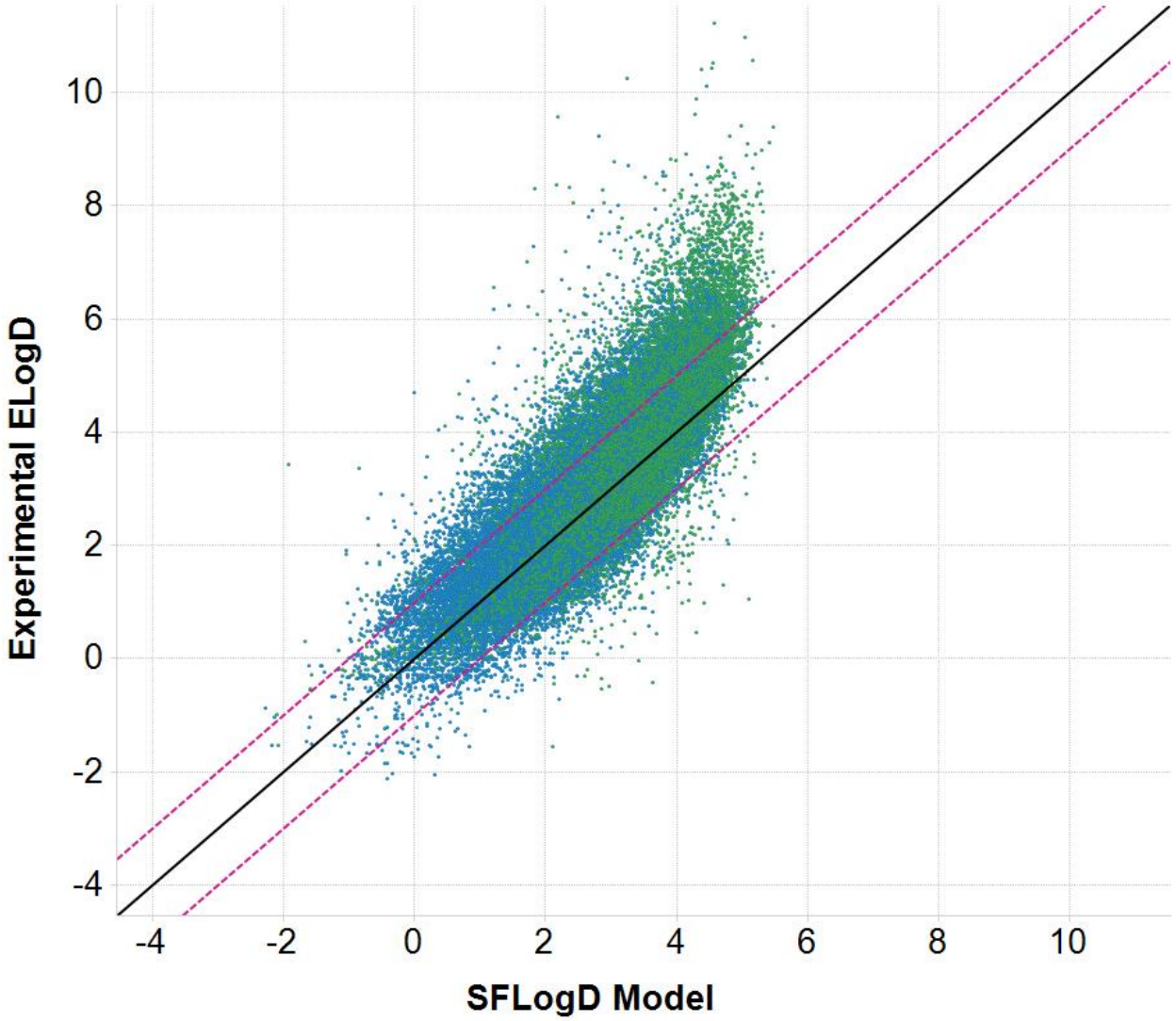

Figure 7 – Experimental SFLogD vs ACD Model. Magenta lines are +/- 1. Points are colored by ionization where blue is basic, green is neutral and red is acidic.

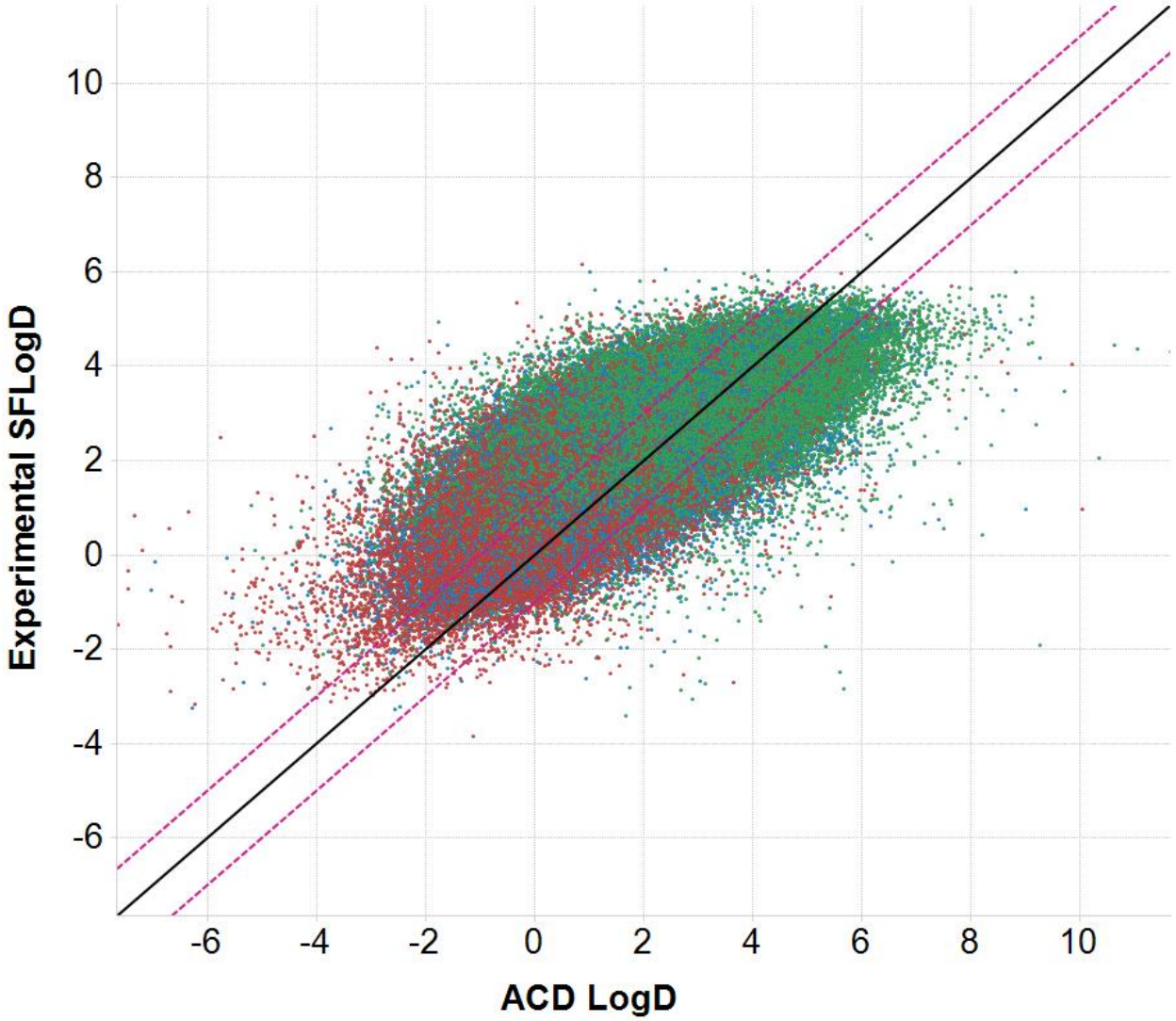

Figure 8 – Experimental SFLogD vs MoKa Model. Magenta lines are +/- 1. Points are colored by ionization where blue is basic, green is neutral and red is acidic.

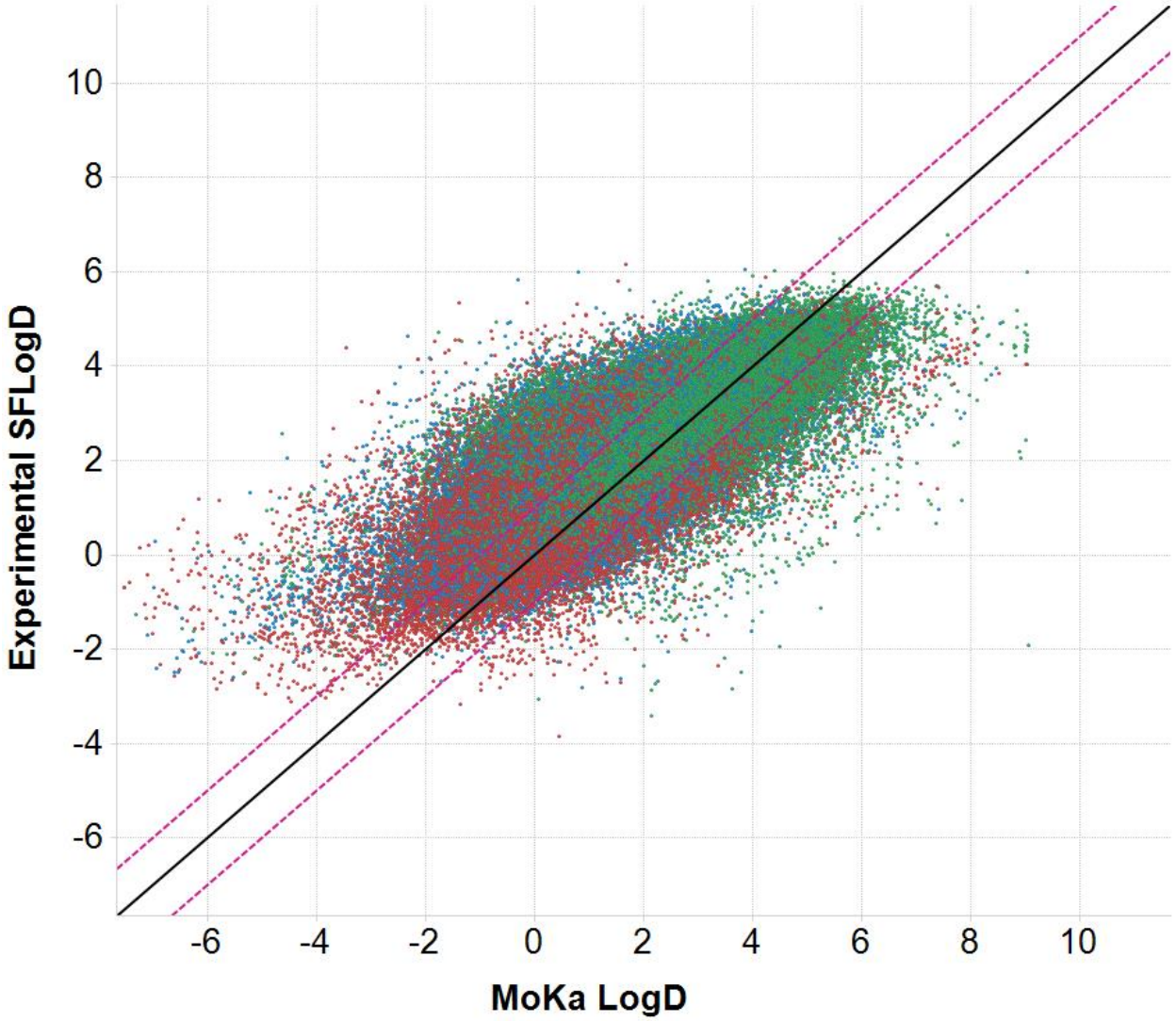

Figure 9 – Experimental ELogD vs ACD Model. Magenta lines are +/- 1. Points are colored by ionization where blue is basic and green is neutral.

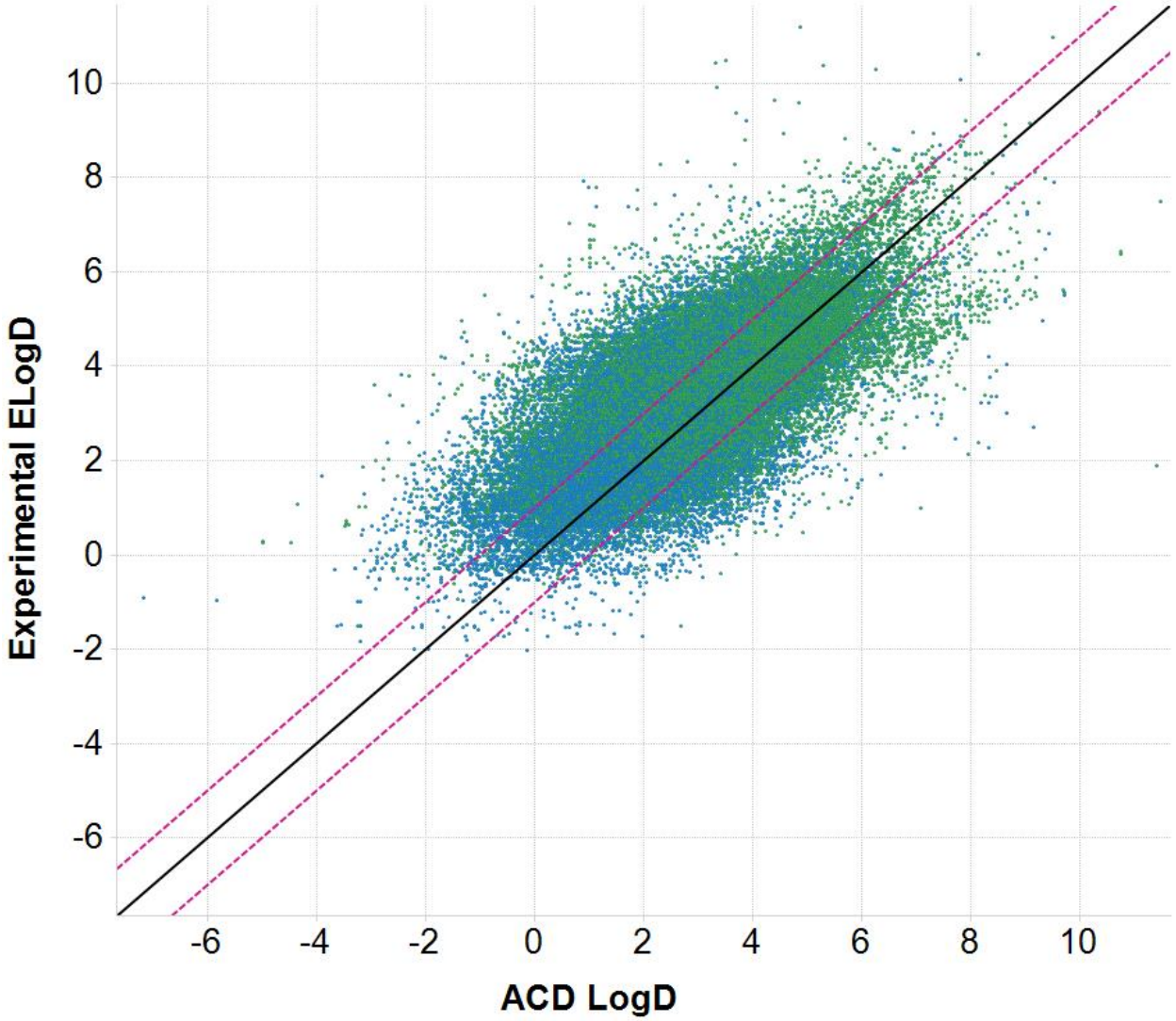

Figure 10 – Experimental ELogD vs MoKa Model. Magenta lines are +/- 1. Points are colored by ionization where blue is basic and green is neutral.

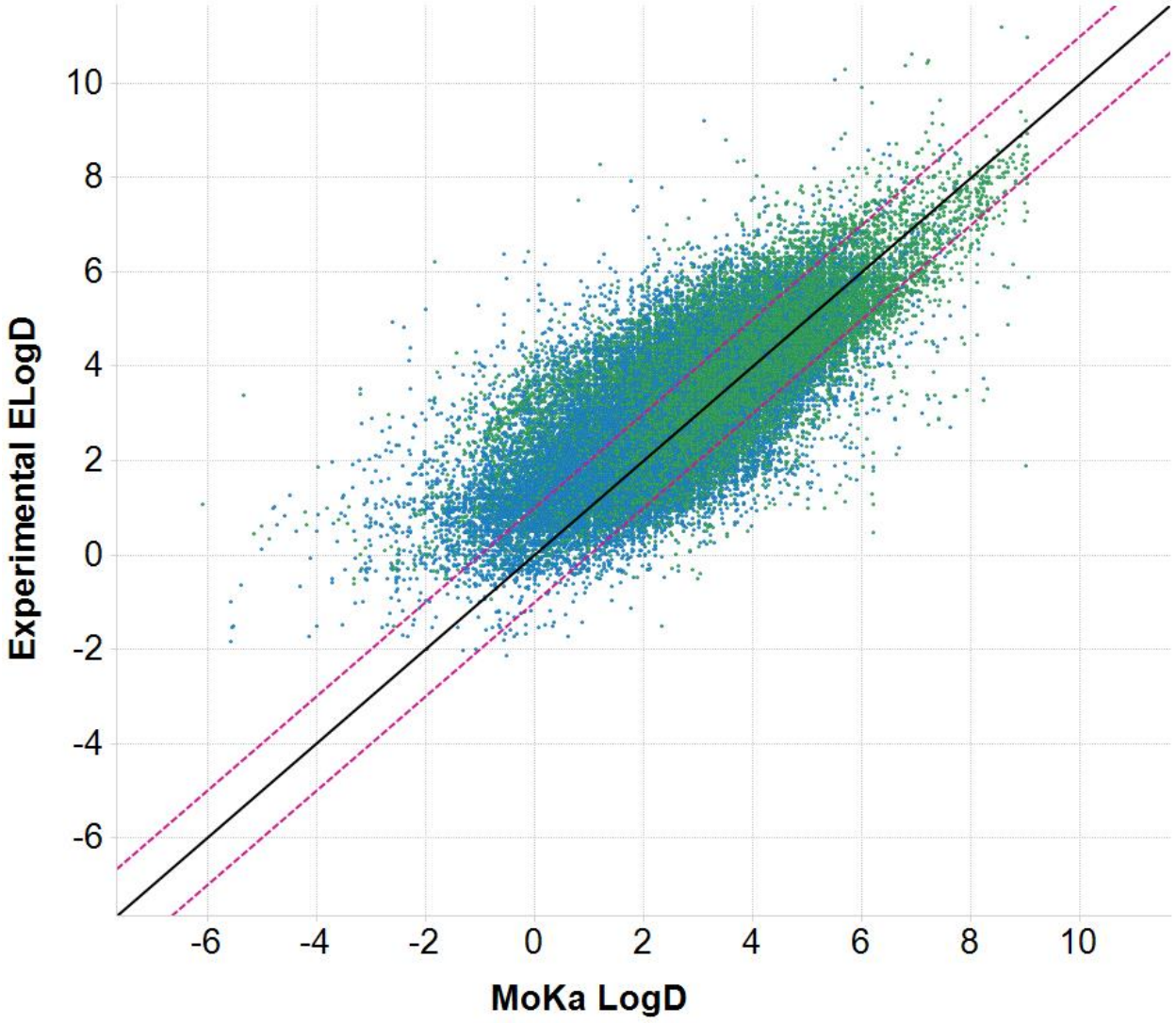

Figure 11 – Experimental SFLogD vs Machine Learning PFLogD Model. Magenta lines are +/- 1. Points are colored by ionization where blue is basic, green is neutral and red is acidic.

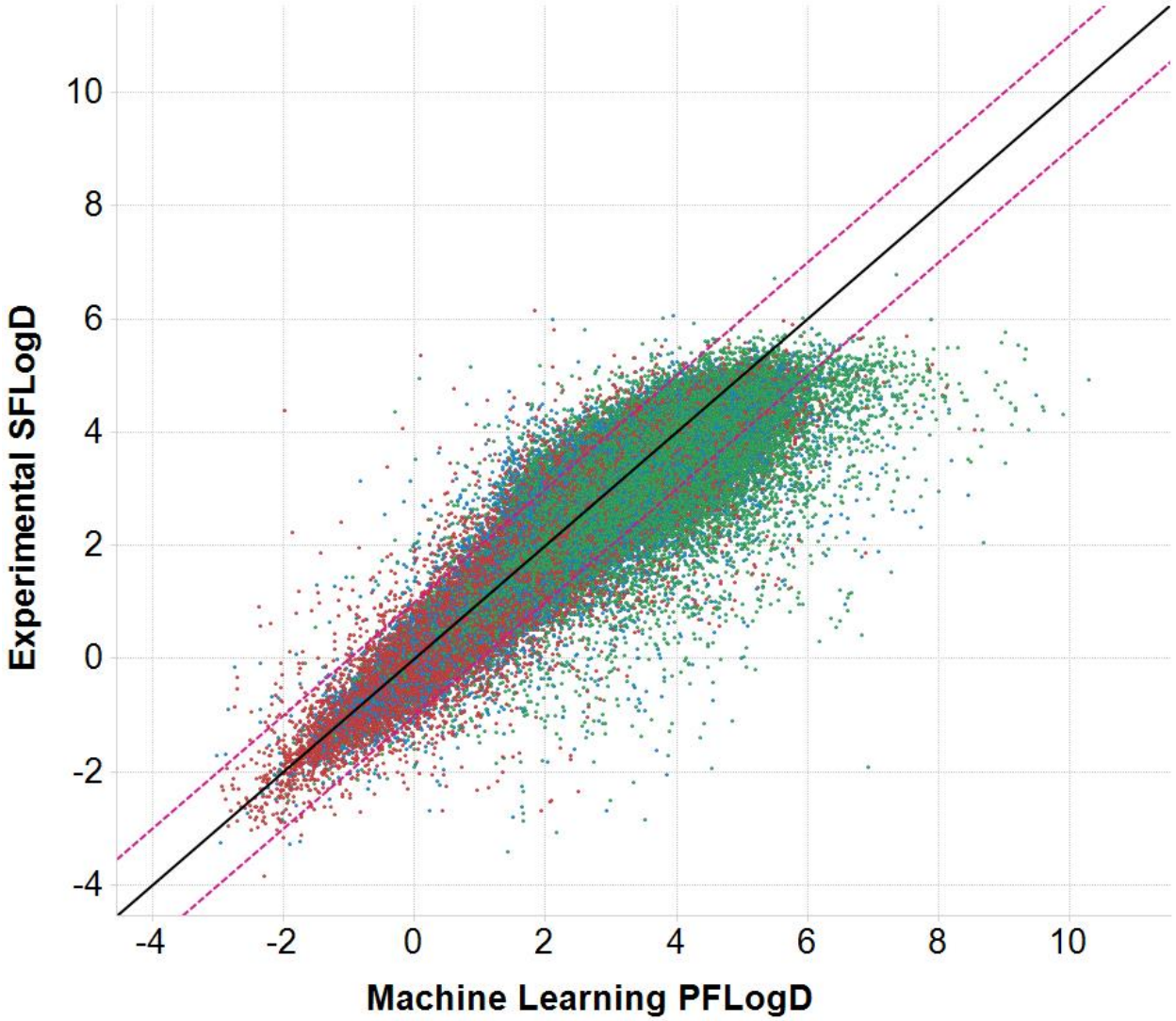

Figure 12 – Experimental ELogD vs Machine Learning PFLogD Model. Magenta lines are +/- 1. Points are colored by ionization where blue is basic and green is neutral.