

# Learning relationships between chemical and physical stability for drug development

Jonathan Fine<sup>1,3</sup>, Prageeth R. Wijewardhane<sup>1</sup>, Sheik Dawood Beer Mohideen<sup>1</sup>, Katelyn Smith<sup>3</sup>, Jameson Bothe<sup>3</sup>, Yogita Krishnamachari<sup>4</sup>, Alexandra Andrews<sup>3</sup>, Peter Wuelfing<sup>5</sup>, Yong Liu<sup>3</sup>, Gaurav Chopra<sup>\*,1,2</sup>

<sup>1</sup>Department of Chemistry, Purdue University, West Lafayette, IN, USA

<sup>2</sup>Department of Computer Science (by courtesy), Purdue University, West Lafayette, IN, USA

<sup>3</sup>Analytical Research & Development, MRL, Merck & Co., Inc., Kenilworth, NJ, USA

<sup>4</sup>Sterile and Specialty Products, Pharmaceutical Sciences, MRL, Merck & Co., Inc., Kenilworth, NJ, USA

<sup>5</sup>Discovery Pharmaceutical Sciences, Pharmaceutical Sciences, MRL, Merck & Co., Inc., Kenilworth, NJ, USA

## Abstract

Chemical and physical stabilities are two key features considered in pharmaceutical development. Chemical stability is typically reported as a combination of potency and degradation product. For peptide products, it is common to measure physical stability via aggregation or fibrillation using the fluorescent reporter Thioflavin T. Executing stability studies is a lengthy process and requires extensive resources. To reduce the resources and shorten the process for stability studies during the development of a product, we introduce a machine learning based model for predicting the chemical stability over time using both the formulation conditions as well as the aggregation curve. In this work, we explore the relationships between the formulation, stability time point, and the measurements of chemical stability and achieve a coefficient of determination on a random test set of 0.945 and a mean absolute error (MAE) of 0.421 when using a multilayer perceptron (MLP) neural network for total degradation. Similarly, we achieve a coefficient of determination of 0.908 and an MAE of 1.435 when predicting the potency using a random forest model. When measurements of physical stability are included into the model, the MAE in the prediction of TD decreases to 0.148 for the MLP model. Using a similar strategy for the prediction of potency, the MAE decreases to 0.705 for the random forest model. Therefore, we can conclude two important points: first, chemical stability can be modeled using machine learning techniques and second there is a relationship between the physical stability of a peptide and its chemical stability.

## Introduction

Peptides are an important class of biomolecules that have been rapidly developed as therapeutical agents that cover areas such as metabolic disorders, skin infections, and oncology.<sup>1,2</sup> Among many challenges in developing acceptable commercial peptide drug product is to demonstrate acceptable chemical and physical stabilities. This is accomplished through systematic stability studies, in which peptides as active pharmaceutical ingredient (API) or in the formulation matrix are stored at certain conditions, their chemical and physical stabilities are monitored over time to understand the effect of environmental parameters such as temperature, moisture and light.<sup>3,4</sup> There are reported computer programs such as *ASAPprime*<sup>5</sup> utilizing humidity corrected Arrhenius equation, which predicts the stability of APIs using storage conditions such as temperature and humidity.<sup>3,5</sup> Potency, individual degradation product and total degradation product are key quality attributes of peptide drug products and studied in chemical stability.<sup>6</sup> High performance liquid chromatography (HPLC) is one of the most common analytical techniques for measurement of potency and degradation product. Peptides, especially in solution during storage or manufacture, can undergo conformational change from  $\alpha$  helical structure to  $\beta$  sheets, then aggregate or form fibrils. These pathways of physical instability are undesirable due to their potential risk of immunogenic responses and impacting bioavailability.<sup>7,8</sup> Thus, aggregation is a key physical stability parameter of formulated peptide product and need to be evaluated. One method for measuring the aggregation of peptides is through the use of Thioflavin T (ThioT). The knowledge gained from stability studies will be used to guide the design of formulation, establishment of shelf life and determination of proper storage and packaging conditions. Depending on the stage of drug development, the duration of stability study ranges

from a couple of months to several years, which requires tremendous number of resources, and the process is lengthy.

Recently, chemists have started to apply machine learning techniques to model various chemical processes for catalyst design<sup>9,10</sup>, prediction of organic reactions<sup>11,12</sup>, reactions in a mass spectrometer<sup>13</sup>, and for the analysis of chemical spectra. Moreover, Lai and coworkers have developed machine learning based methodologies to determine molecular features which are responsible for the viscosity behavior and the aggregation of therapeutic antibodies.<sup>14,15</sup> In short, machine learning is a process where an algorithm is given input variables (x) and an output variable (y). This algorithm then adjusts a set of weights, biases, and other parameters to predict the output variable from the input variables. Given enough weights and an optimized algorithm, the result will be a model that can always predict the output from the inputs. Therefore, one needs to ensure that a given model is robust with high predictive performance on data that was not used to directly train the model. A popular technique for ensuring that this scenario is the case is referred to as k-fold cross validation where the available data is divided into various portions called folds.<sup>16</sup> For example, the available data can be divided into 5-folds each containing 20% of the entire data set. One of these folds is then removed and referred to as the validation set and the remaining training set is used to train the model. The model is then evaluated on the validation set and the performance of the model is calculated. This process is repeated until all the folds has been used as the validation set and statistics can be calculated on how well the model did for the various folds.

One interesting application of machine learning in chemistry is the ability to understand how underlying variables (x) relate to the output (y). The use of machine learning to understand these

variables has been applied to understanding how IR and MS spectra relate to functional groups<sup>17</sup> and how functional groups in organic molecules relate to chemical reactivity<sup>13</sup>. When performing this type of analysis there are two important points to consider. First, machine learning models can only learn from the variables they are given, and second machine learning models can “overfit” to a particular data point set of variables<sup>18,19</sup>. This second point can be addressed by using cross-validation. Recent works have shown that machine learning models built on small training sets can be successful for both understanding chemistry and for useful predictions.<sup>13,20–</sup>

22

To our knowledge, these techniques have not been used to understand the relationships between storage conditions, chemical stability, and physical stability in drug formulation. In this work, we will outline how an initial design of experiment (DoE) was used to generate training data for both the chemical and physical stability of a peptide drug molecule in formulation matrix. Then, we will use the training data from this DoE to create machine learning models to predict the chemical stability from both the storage conditions and the physical stability to develop a relationship between these variables.

## Methods

### Stability study for Peptide A

Pharmaceutical development was triggered for Peptide A, sequence H<sub>2</sub>SQGTFTSDK(yE<sub>16</sub>)SKYLDARAAQDFVQWLLDT-NH<sub>2</sub>,<sup>23</sup> necessitating the design of a robust and safe formulation with optimized chemical and physical stability. Early in development it was determined that Peptide A had risks in both its chemical and physical stabilities. The primary chemical degradation products were aspartic acid isomerization at multiple sites and reactivity

at the N-terminal histidine. Fibrillation was identified as the primary mechanism of physical degradation. Optimizing the formulation for Peptide A was a balancing act due to the contrasting pH dependent trends necessary for chemical and physical stability. The chemical stability and key degradation products were observed to be accelerated under more alkaline conditions for the pH range evaluated (7.0-8.7). In contrast, the physical stability was inverted where Peptide A was stabilized by more alkaline conditions tending to form fibrils at neutral or acidic pH's.

We sought out to design an optimized formulation by setting up a stability study composed of Peptide A formulations spanning a wide range of pH's, buffers, and excipient concentrations. The stability study utilized both long term (5°C) and accelerated (25°C, 40°C) conditions with a target storage condition of being refrigerated. Given that pH was closely linked to chemical and physical stability formulations were designed to explore a wide pH range of 7.5-8.7 to define a formulation with an optimized pH. In addition to pH, compositions of multiple excipients were varied including histidine (0 to 10 mM) and a proprietary excipient B (EB) (0-0.5% w/v) which were targeted to improve chemical and physical stability, respectively. The addition of histidine was to prevent the growth of a specific degradant and is not expected to have a dramatic effect neither on the potency nor the total degradation of the peptide. Therefore, the addition of this excipient can be used as a case study for changes to the formulation that are not designed to make a dramatic impact on the output variables. The goal of the stability study was to optimize the pH and excipient compositions and determine if it was achievable to design a formulation with acceptable physical and chemical stability over the course of Peptide A's shelf life.

The values measured in the above DoE experiment are easily translated into features usable for a machine learning model. The pH, %EB, and output of the model are all numeric features of the

model. While it would be possible to treat some of these values as one-hot encoded vectors, the true utility of the model would be to use for the optimization of these parameters. Since this ability is not possible using a one-hot encoded vector, we opted to treat these values numerically. Additionally, the natural choice to treat the prediction of potency (measured using the percent label claim of the product, %LC) and Total Degradation (TD) as numeric values forces the creation of regression models. However, the buffer type used as well as the presence of histidine were treated as factors. The first choice is required as the use of a phosphate buffer versus a tris buffer cannot be modeled numerically. The second choice is also required as only concentration of histidine was used in the DoE experiment.

#### [Interpretable machine learning for predicting potency and total degradation.](#)

A shotgun approach was used to evaluate how ML models perform for the calculation of potency (%LC) and Total Degradation (TD) from the variables in the DoE. The models were evaluated using k-fold cross validations (k=2, 5, 10 and 20) as well as a bootstrap method provided by Caret package in R<sup>24,25</sup>. The ML models evaluated are a Generalized Linear Model<sup>26</sup> (GLM), Random Forest<sup>27,28</sup> (RF), Partial Least Squares<sup>29</sup> (PLS), K Nearest Neighbor (KNN), Multi-Layer Perceptron<sup>17</sup> (MLP), and Support Vector Machine<sup>30,31</sup> with a linear kernel (SVM<sup>L</sup>) and a radial kernel (SVM<sup>R</sup>). Since both TD and %LC are represented to these models as a number, these models were treated as regression problems. We decided to use the coefficient of determination ( $R^2$ ) to compare the performance of the models and select the best model from the data as this measure is easily understood by analytical chemists. More conventional measures of error such as Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) were calculated as well. The equations for these measures of error are given below where  $\hat{y}$  is the experimentally measured observation

and  $y$  is the value predicted by a model. Additionally,  $\bar{y}$  is the mean of the experimentally measured values.

$$R^2 = \frac{\sum(\hat{y}_i - y_i)^2}{\sum(\hat{y}_i - \bar{y})^2} \quad 1$$

$$MAE = \frac{\sum|\hat{y}_i - y_i|}{n} \quad 2$$

$$RMSE = \sqrt{\frac{\sum(\hat{y}_i - y_i)^2}{n}} \quad 3$$

### Model training, validation, and testing

To evaluate the performance of the model on an external test (blind) set, the data available for modeling was split into a training and test set. Assignment of an observation to either set was random and 80% of the data was set aside for the training set while 20% was used for the test set. Then, the training set was then split into a training and validation set as per the cross-validation procedure and the above three measurements of model performance ( $R^2$ , MAE, RMSE) are calculated for the validation set for a given set of hyperparameters (i.e. parameters used to control the training performance of model itself). The model and associated hyperparameters which performs the best (as measured by these metrics) is then selected and a new model is trained using the same method and hyperparameters using the entire training set. The test set identified at the beginning of the procedure was then used to evaluate this final model.

### Results and Discussion

Total degradation is linear with respect to time point while label claim is not linear.

Before attempting to model the data, we examined how the total degradation (TD) and percent label claim (%LC) changed with time-point. To do so, we plotted these variables, separated by



the storage conditions, as a function of time-point (**Figure 2a** for TD and **Figure 2b** for %LC). The plot for TD shows a linear trend, which is confirmed by the coefficient of determination ( $R^2$ ) values for each sample (**Table 1**) as 16 out of the 18 samples have an  $R^2$  greater than 0.95 and 13 samples have an  $R^2$  greater than or equal to 0.98. In contrast, there are only 5 samples where the %LC has an  $R^2$  value greater than 0.95 and only a single sample has an  $R^2$  greater 0.98. Interestingly, all five samples with a high  $R^2$  were stored at the same pH and for these storage conditions, the %LC was less than or equal to 80% at the 24-month time-point. All other pH storage conditions had a %LC greater than 90%, therefore these 5 high  $R^2$  values may be false indicator of linearity resulting from a strong outlier affect. While  $R^2$  values alone cannot be used to determine whether a relationship is best modeled by a linear function, these results indicate that the TD measurement is more linear with respect to time than %LC. It should also be noted that some samples with low potency were noted to have a high degree of fibrillation, leading to samples with a high viscosity. This made the samples difficult to analyze and is not a good measurement of the true chemical degradation and problematic samples were removed from the available training data. In the next section, we explore different machine learning models for predicting these two measurements of chemical degradation.

#### Regression models predict total degradation.

The first machine learning models we investigate are interpretive models known to perform well in most general cases. Here we use Decision Trees (DT), Generalized Linear Models (GLM), k-Nearest Neighbor regression (KNN), Partial Least Squares (PLS), linear Support Vector Machine ( $SVM^L$ ), radial Support Vector Machine ( $SVM^R$ ), Random Forest (RF), and a Multi-Layer Perceptron (MLP) to predict TD and %LC from the storage conditions (pH, buffer type, %EB, histamine

concentration) and the time-point. We evaluated each machine learning model using both bootstrapping cross-validation and a k-fold cross-validation using k=2, 5, 10, and 20. To compare the models to each other and across validation techniques, we calculated the  $R^2$  of the validation set. These results are summarized in **Figure 3a** and **Table S1** with additional metrics shown in **Figures S1**. It can be seen that SVM<sup>R</sup>, RF, and MLP models out-perform the other models and within error of each other. For these models, the performance of models trained on 50% of the available data (k=2) is similar to those trained on 95% of the available data (k=20). This indicates that these models can be trained with a small amount of data and still be predictive. Since these models all perform within error of each other, we compared the true TD to the predicted TD for each fold for the three models. The results for k=2 cross-validation is shown in **Figures 3b-d** and for all cross-validation schemes in **Figures S2-S4**. These results show that the MLP model is best at producing a one-to-one prediction with the true TD. The RF and SVM<sup>R</sup> models fail to predict large TD values (above 7.5%) while the MLP model is able to predict the TD for these values. Therefore, we conclude that the MLP model performs best for this form of chemical degradation.

Given the success of the MLP model, we decided to investigate the internals of the model to attempt to gain insight. The entire model is displayed in the top of **Figure 4**. It can be seen that the 5 input variables go through two linear transformations: first a 5->5 vector transformation and second from the second vector to a scalar. The second transformation weights and sums the previous layer in a completely positive manner. The first transformation, however, does include negative transformations into account. The weights for the “Histidine” input variable are all negative. Additionally, we calculated the importance of all the importance variables using Olden’s method (bottom of **Figure 4**).<sup>32</sup> It is not surprising that the most important variable is the Time-

point, followed by the pH (see the above discussion of **Figure 2**). While these results do not yield any insights into how these variables affect chemical degradation, they do show that this type of analysis can be used to reveal known relationships between the input variables and the output variables.

Results of the Train-Validate-Test paradigm are given in **Table 2**. The MAE of the MLP model is 0.421 for the test set and 0.267 for the full training set, indicating that a tuned version of this model performs well for data not available during training. This represents a mean percent error of 15.6% for the prediction of TD for time-points past the first month of the study given the average total degradation past this is 3.0%. Therefore, we would like to further optimize the prediction of TD using additional information, and idea which will be discussed in a later section of this work.

#### Regression models predict percent label claim.

The successful modeling of TD using storage conditions was repeated for the prediction of potency (quantitated as %LC). In **Figure 5a** it can be seen that the RF model achieved an  $R^2$  greater than 0.90 and **Figure S5** indicates that the RF model has the lowest RMSE and MAE. Additionally, the relationship between the true observations and predictions does not show any significant patterns (**Figure 5b**, **Figure S6**). For other interpretable machine learning models, points with true low %LC values are predicted to be too high (**Figures S7-S8**). These results indicate that the RF model is able to capture the relationship between the input variables and %LC. An analysis of the variable importance is given in **Figure 5c**. As with the prediction of TD, the most important variable for the prediction of %LC is the Timepoint of the data. The second most important variable is the type of buffer used to store the peptide. This result is interesting as the only pH

condition associated with a change in buffer type is the pH 8.7 condition, indicating that the RF model is using this variable as a surrogate for the highest pH condition. Since the highest pH condition shows the greatest decrease in %LC, this selection is justified. The third most important variable is the pH, which likely explains the remaining variance in the %LC curve not already covered by the selection of buffer type. Finally, the EB is least important variable that is used by the RF model and likely adjusts for small changes in %LC. It is interesting to note that the amount of histidine is not used in the model to predict %LC and is assigned an importance of zero, similar to the result from the TD model.

As with the TD predictions, the Train-Validate-Test paradigm<sup>16</sup> was used to evaluate the performance of the tuned model, and the results of these tests is given in **Table 3**. The MAE for the RF model is 1.435 for the test set and 0.779 for the training set. Since the mean %LC value for timepoints after 1 month is 94.7%, the percent error for this model is 1.52%. It is interesting to note that, relative to the magnitude of the mean value, the final %LC model does better than the final TD model, even though the trend is less linear with respect to time-point. This indicates that the ML models are able to account for non-linearity in the data.

#### [Incorporating physical degradation as an input feature for chemical degradation](#)

To improve the prediction results for the %LC and TD models, we decided to include the physical degradation of the peptide as measured by the ThioT curve measured at each time-point for the storage conditions (data shown in **Figure 6** for the initial **(6a)**, 1-month **(6b)**, 6-month **(6c)** and 12-month **(6d)** time-points). Here, the ThioT curve was given to the model as a normalized vector where each point of the curve measured over the 48-hour experiment became an independent input to the model. Since this value was measured every 15 minutes, there are 192 points

available to the model. We used a rolling average with a window of 5 to smooth out the curve and trained the models using only the resulting ThioT curve as well as a model containing both the ThioT as well as the storage conditions used in the previous sections. The results of this modeling are shown in **Table 4** for MLP, RF, and SVM models.

All three investigated models are able to predict the TD using solely the ThioT data as input. The RF, MLP, and SVM models achieve an MAE of 0.387, 0.310, and 0.662, respectively. However, The  $R^2$  values for these models trained on only ThioT data are all below 0.5. This result can be explained by the models predicting low TD values with high fidelity and high TD values poorly. Inclusion of storage conditions improves the predictions significantly, decreasing the MAE of the models to 0.121, 0.132, and 0.148, respectively and increases all the  $R^2$  values to be greater than 0.9. These results show that the inclusion of ThioT data into the model is beneficial. Furthermore, the MAE values for the TD prediction are approximately half that for the model trained with ThioT versus the model trained on only storage conditions, indicating that there is value in adding physical degradation measures to the model.

For the prediction of %LC, the RF, SVM, and MLP models are able to predict %LC from the ThioT information alone (albeit poorly) with MAEs of 3.04, 2.78, and 4.131, respectively. However, the  $R^2$  value for both models are quite low at values of 0.254, 0.304, and 0.106. A potential explanation of this performance is that the models predict high %LC values with great accuracy but are not able to capture the overall trend of the data as the %LC values decrease. Inclusion of the storage conditions decreases the MAEs to 0.705, 0.976, and 1.367 for the RF, SVM, and MLP models, and increases the  $R^2$  to 0.963 and 0.931, 0.880 for the three models, respectively. The

improvement in model performance indicates that these models are able to properly weigh the storage conditions and physical degradation to predict a non-linear response variable.

To better understand how the addition of the ThioT curve data improves the model, we performed feature importance calculations using the RF model. These results are shown for the time-points of the ThioT curve in **Figure 7** and show that the initial points of the ThioT curve are most important to the model performance. The same RF model was used to visualize the learned relationship between the Physical stability (ThioT curve) and chemical stability data (%LC or TD) for each time point. These models are visualized as 2D contour plots in **Figure 8** for both %LC and TD. Predicted values are plotted as a function of two variables: i) Samples – different sample conditions at each time point ii) pH – pH value of the considered sample. The grouping of the samples in the figure (X1, X2, etc.) was done such that the pH of four group members were different as to sample the four different pH conditions in the original experimental setup. There is no other relationship between the samples in these groups, therefore movement across the X-axis only means different samples having the same pH. These grouped samples are tabulated in the **Table S2** with storage conditions. The ML learned “function” shown in **Figure 8** has captured special conditions such as sample R3 solely based on physical stability data. Sample R3 (pH = 8.7, Buffer = Tris, EB=0.50, and Histidine = 0) at time-point 12 showed the maximum true TD (5.2131) as well as a low true %LC values (87.5178). Both functional surfaces in **Figure 8** show higher and lower values for TD and %LC, respectively, around sample R3. Further, both increasing TD and decreasing %LC with time have been captured by the function developed by the model. This is demonstrated as the %LC functional surface gets more dark blue areas and TD functional surface gets more dark red areas over time for some sample conditions such as “R3”, indicating a

decrease of the peptide's stability over time. These functional surfaces demonstrate a relationship between the physical degradation and chemical degradation. Both ML learned functions are compared with surfaces generated by true data in **Figure S9** and **Figure S10** for TD and %LC, respectively. Therefore, we show that the learned functional surfaces have a very good correlation with true data. Moreover, the ML learned function by the RF model trained on both ThioT curve data and storage conditions in **Figure S11** also shows similar relationship to the surfaces shown in **Figure 8**. A limitation of this work is that it does not determine how physical and chemical degradation are related on a molecular level. In future works, we will use molecular mechanics to investigate this relationship to develop detailed mechanistic models.

## Conclusion

We have investigated the relationship between the storage conditions, physical and chemical stability of Peptide A using interpretive machine learning techniques. These machine learning models trained only on the storage conditions of the peptide are able to reliably predict the total degradation (TD) and percent label claim (%LC, a measure of potency) with  $R^2$  values greater than 0.90 and a mean absolute error less than 0.5 for the prediction of TD and less than 1.5 for the prediction of %LC. These models were then analyzed to understand how the underlying variables relate to the chemical degradation of Peptide A to show that the pH and %EB combined with the timepoint of the experiment are important features of the model while the usage of histidine does not appear to be important for modeling the chemical degradation of the peptide. We believe that these modeling efforts are useful for understanding the degradation of Peptide A as a function of its storage conditions. In conclusion, we provide a framework to develop the

relationships between chemical and physical degradation for future studies in drug development and in chemical sciences.

### Conflict of Interest

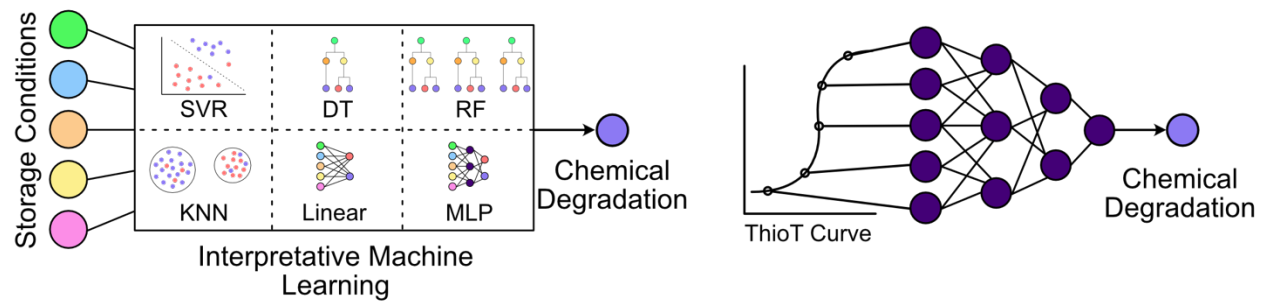
Gaurav Chopra is the Director of Merck-Purdue Center for Measurement Science funded by Merck Sharp & Dohme Corp., a subsidiary of Merck & Co., Inc., Kenilworth, NJ, U.S.A.

### Acknowledgements

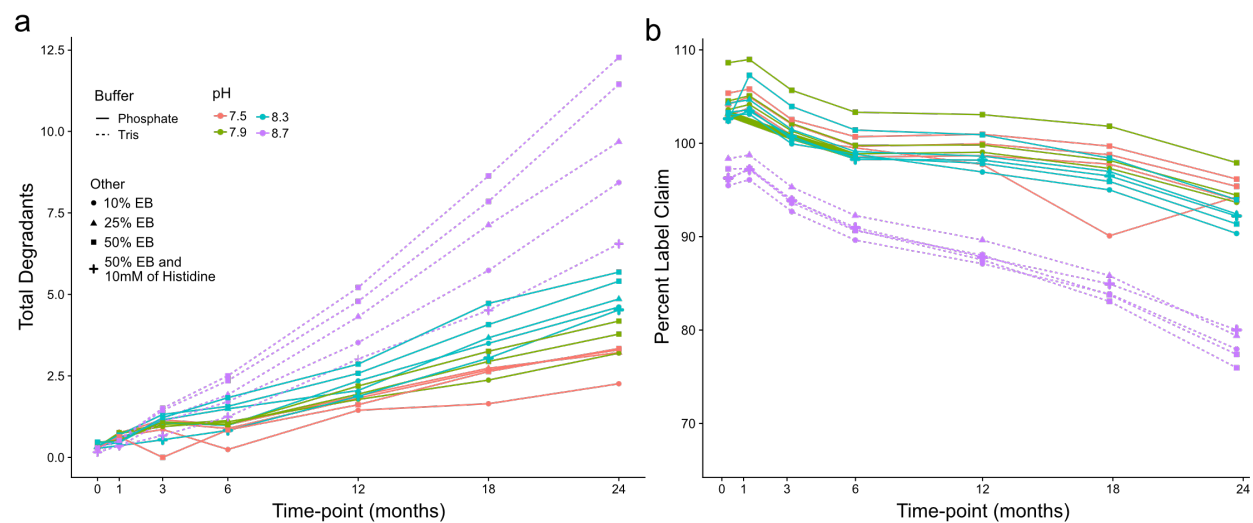
This work is funded, in part by, the Merck-Purdue Center award # 40002619 on 'Machine Learning Methods to Elucidate Peptide Aggregation' to Gaurav Chopra.



## Figures and Tables



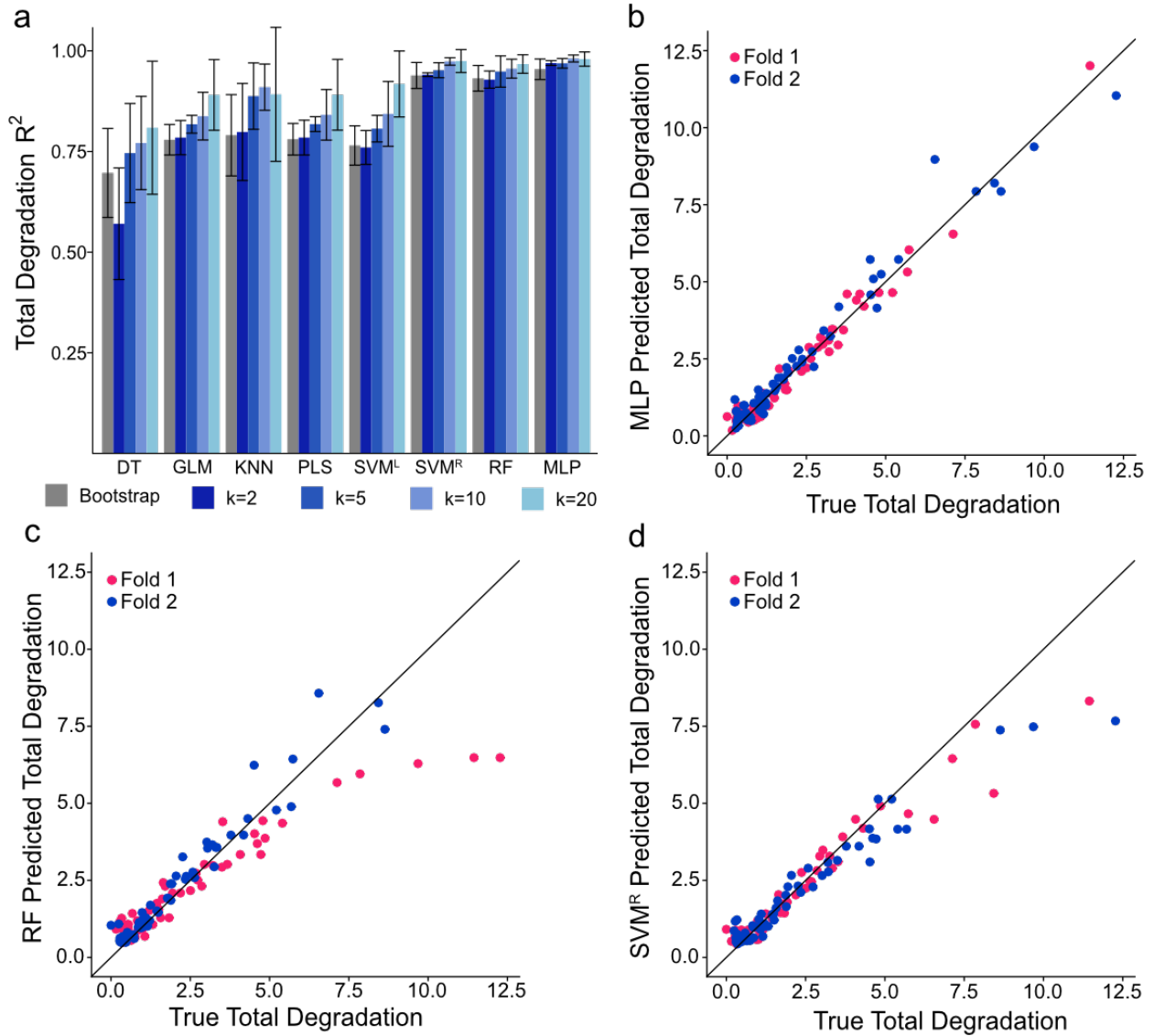
**Figure 1:** Relationship between physical and chemical degradation. (left) Methodology for predicting chemical degradation as a function of storage conditions and time. Interpretative machine learning predicts total degradation of Peptide A as this appears to be a linear function with respect to time. Second, we will present a model to predict the change in percent label claim using changes in physical degradation as measured by a ThioT curve. Using this methodology, we will show that there is a link between the physical degradation of the peptide and the chemical degradation as represented by total degradation and percent label claim.



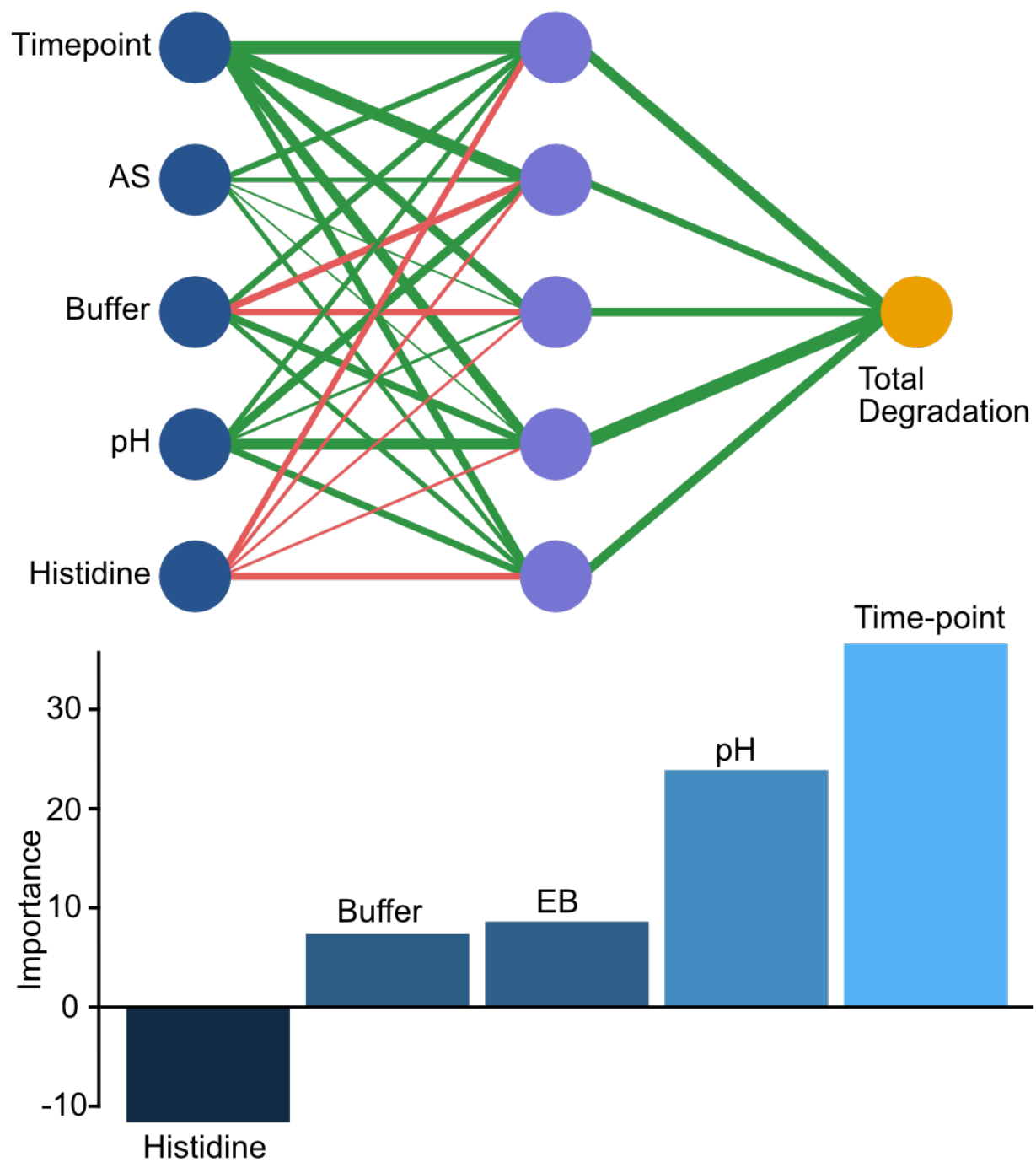
**Figure 2:** Training data for (a) total degradants and (b) percent label claim as measured by HPLC. The relationship between the total amount of degradants and time-point follows a linear trend (see **Table 1**). The slope of this trend is a function of pH, the type of buffer used, concentration of histidine, and the percentage of EB used. The relationship between time-point and the percent label claim is less obvious.

**Table 1:** Coefficient of determination ( $R^2$ ) for Total Degradation and Percent Label claim (%LC) as a linear function of time-point. Values greater than or equal to 0.95 are bolded.

Sample ID	pH	EB	Histidine	Buffer	TD $R^2$	%LC $R^2$
O1	7.5	0.1	0	Phosphate	0.856	0.010
O2	7.5	0.25	0	Phosphate	<b>0.970</b>	0.838
O3	7.5	0.5	0	Phosphate	<b>0.973</b>	0.703
T1	7.5	0.5	0	Phosphate	<b>0.950</b>	0.571
P1	7.9	0.1	0	Phosphate	<b>0.986</b>	0.852
P2	7.9	0.25	0	Phosphate	0.854	0.750
P3	7.9	0.5	0	Phosphate	<b>0.983</b>	0.750
T2	7.9	0.5	0	Phosphate	<b>0.986</b>	0.510
Q1	8.3	0.1	0	Phosphate	<b>0.986</b>	0.944
Q2	8.3	0.25	0	Phosphate	<b>0.981</b>	0.858
Q3	8.3	0.5	0	Phosphate	<b>0.991</b>	0.911
S1	8.3	0.5	10	Phosphate	<b>0.980</b>	0.889
T3	8.3	0.5	0	Phosphate	<b>0.992</b>	0.695
R1	8.7	0.1	0	Tris	<b>0.989</b>	<b>0.976</b>
R2	8.7	0.25	0	Tris	<b>0.993</b>	<b>0.979</b>
R3	8.7	0.5	0	Tris	<b>0.992</b>	<b>0.985</b>
S2	8.7	0.5	10	Tris	<b>0.992</b>	<b>0.979</b>
T4	8.7	0.5	0	Tris	<b>0.990</b>	<b>0.978</b>



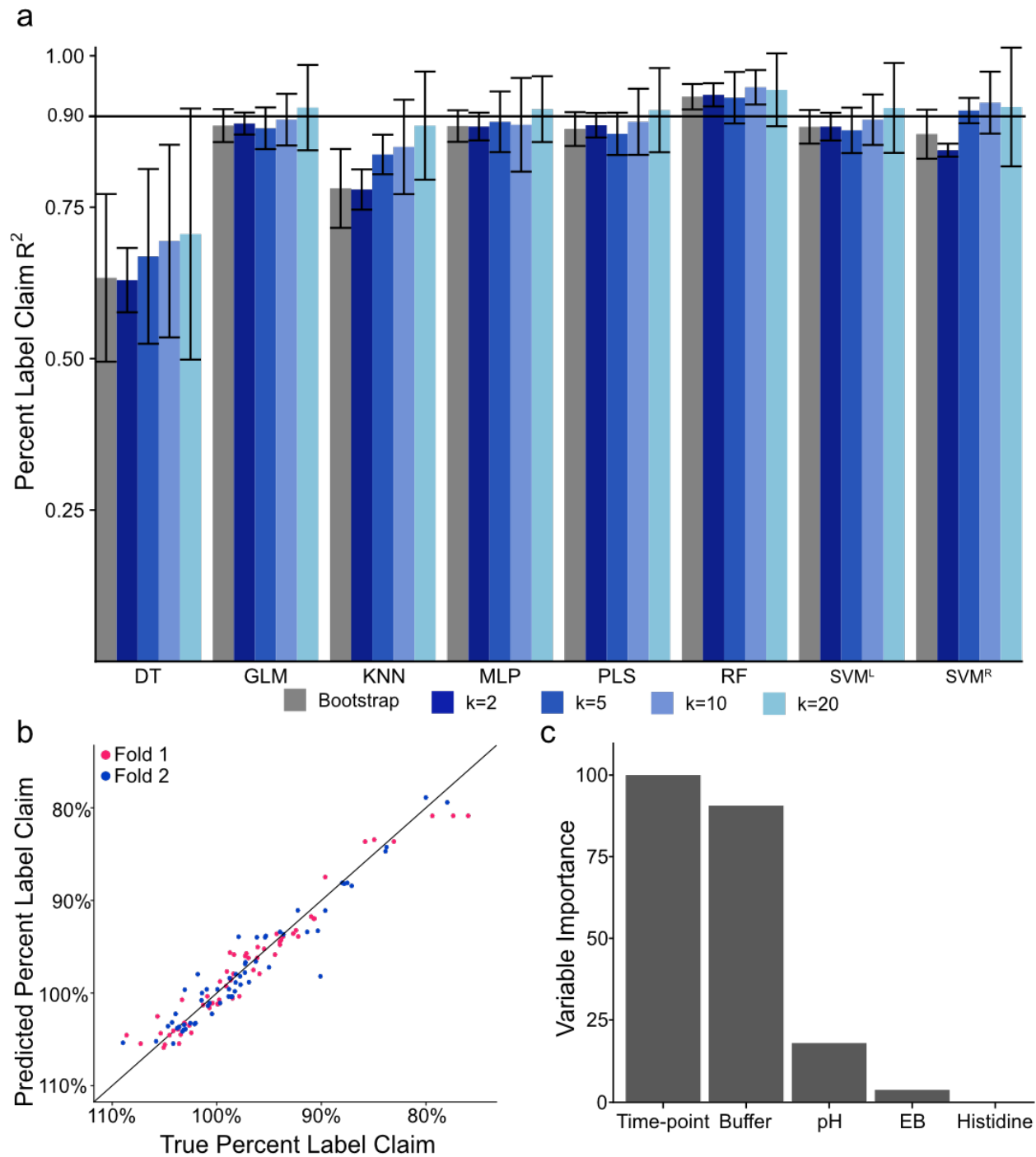
**Figure 3:** (a) The ability of each ML model to predict total degradation as a function of the five input variables is shown above. A look at how the prediction of each model correlates with true total degradation for the (b) MLP model, (c) random forest model, and (d) radial SVM model. These plots show the MLP model is able to predict the total degradation at higher values in the plot. These values correspond to the values at higher pH (see **Figure 2**).



**Figure 4:** The final MLP model when trained with hyperparameters derived from the 2-fold cross validation model. Here, the thickness of the lines indicates the magnitude of the given weight. Positive values are green while negative values are red. The bias values are set to zero in this model and therefore these values are not shown. The importance of each feature is shown in the Olden plot at the bottom on the figure. This shows that the Time-Point and pH are two most important features in the model.

**Table 2:** Test set results for the prediction of TD using only the storage conditions. Hyperparameters for each method were determined using k-fold cross validation with k= 2.

Set	Method	R <sup>2</sup>	RMSE	MAE
Test	MLP	0.945	0.531	0.421
Test	RF	0.937	0.584	0.412
Test	SVM <sup>R</sup>	0.974	0.414	0.311
Train	MLP	0.980	0.374	0.267
Train	RF	0.974	0.450	0.257
Train	SVM <sup>R</sup>	0.972	0.530	0.281

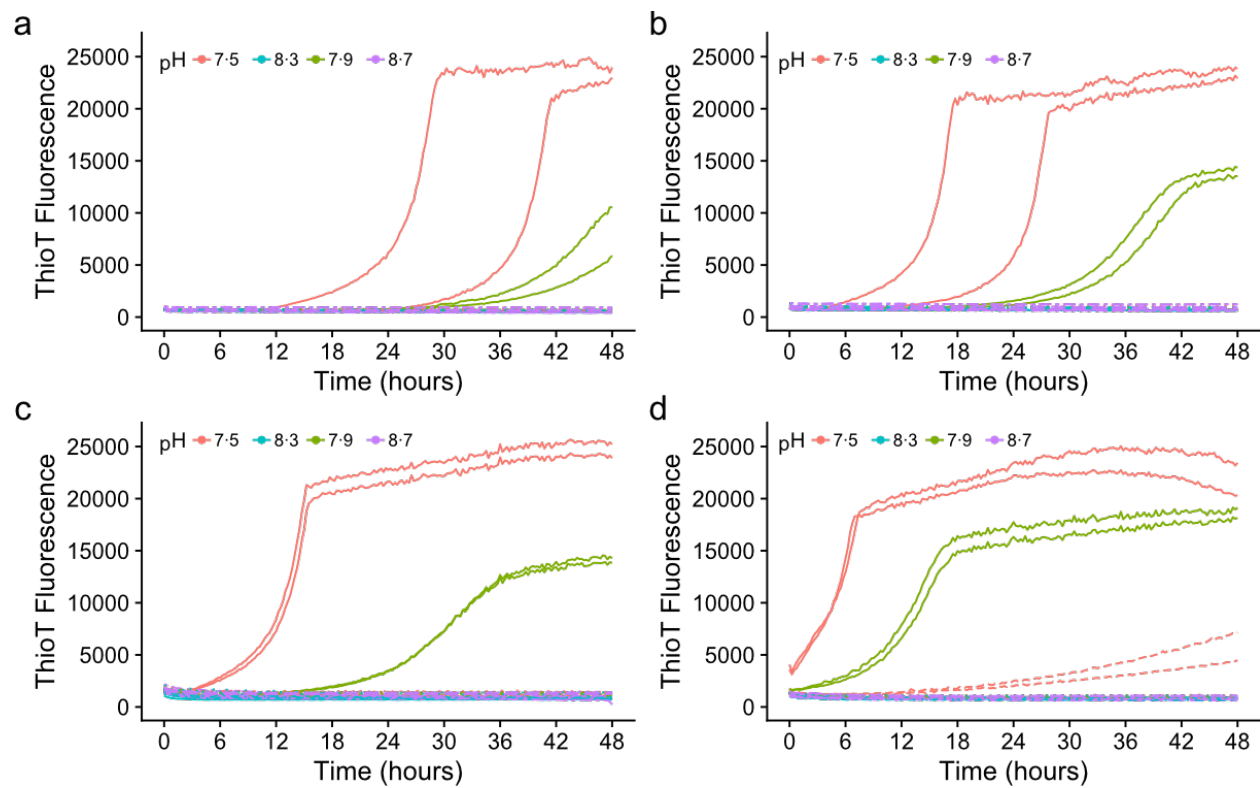


**Figure 5:** The ability of each model to predict %LC from the storage conditions is shown in (a). Of these models, only RF consistently exceeds an  $R^2$  value greater than 0.90 for all cross-validation schemes. A plot of the predicted %LC value versus the true %LC value is shown in (b). These results show that the model is also able to predict %LC from storage conditions.

**Table 3:** Test set results for the prediction of %LC using only the storage conditions. Hyperparameters for each method were determined using k-fold cross validation with k= 2.

Set	Method	R <sup>2</sup>	RMSE	MAE
Test	MLP	0.918	3.285	2.864
Test	RF	0.908	2.174	1.435
Test	SVM <sup>R</sup>	0.905	2.200	1.715
Train	MLP	0.893	3.434	2.836
Train	RF	0.978	1.010	0.779
Train	SVM <sup>R</sup>	0.959	1.516	1.053

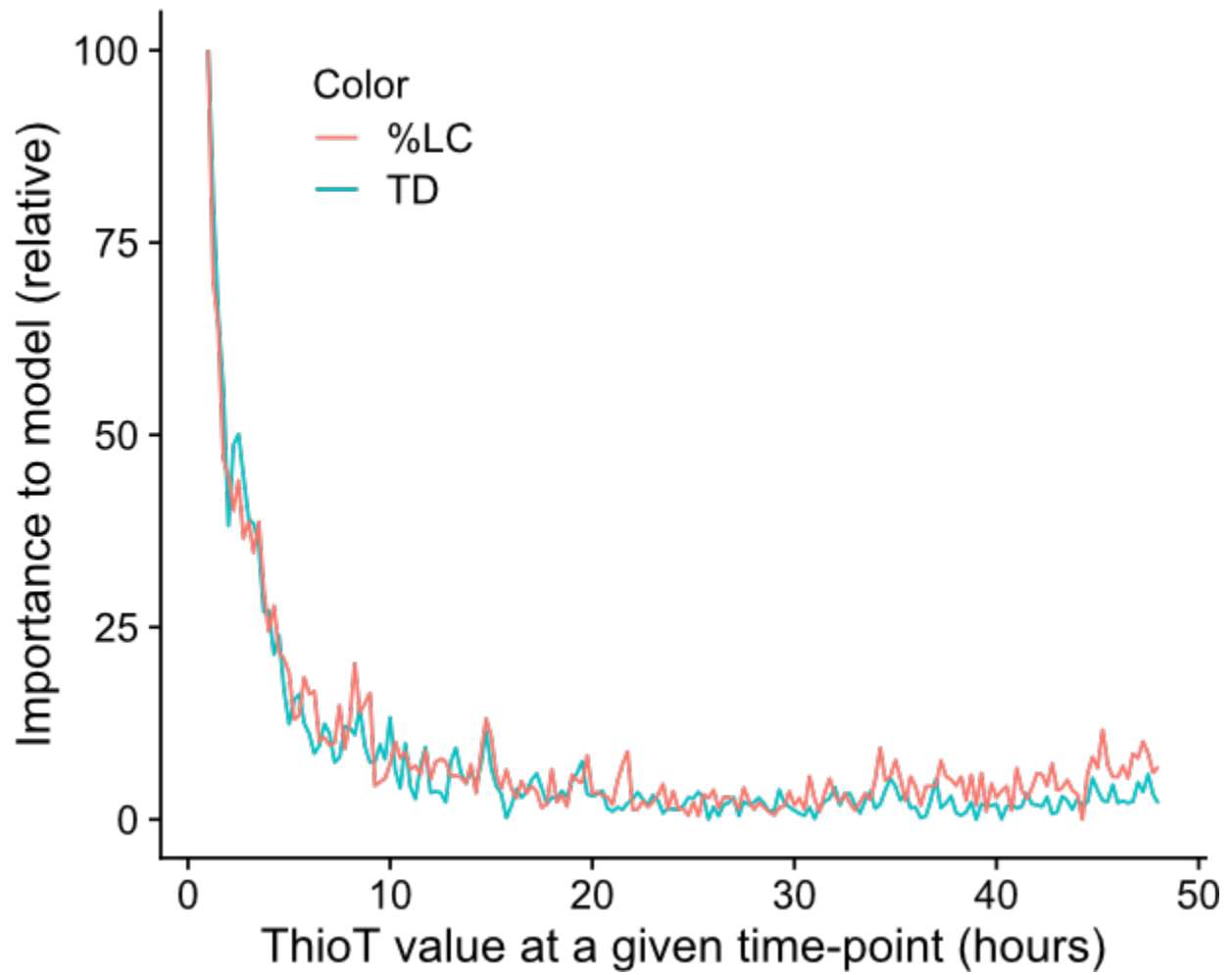




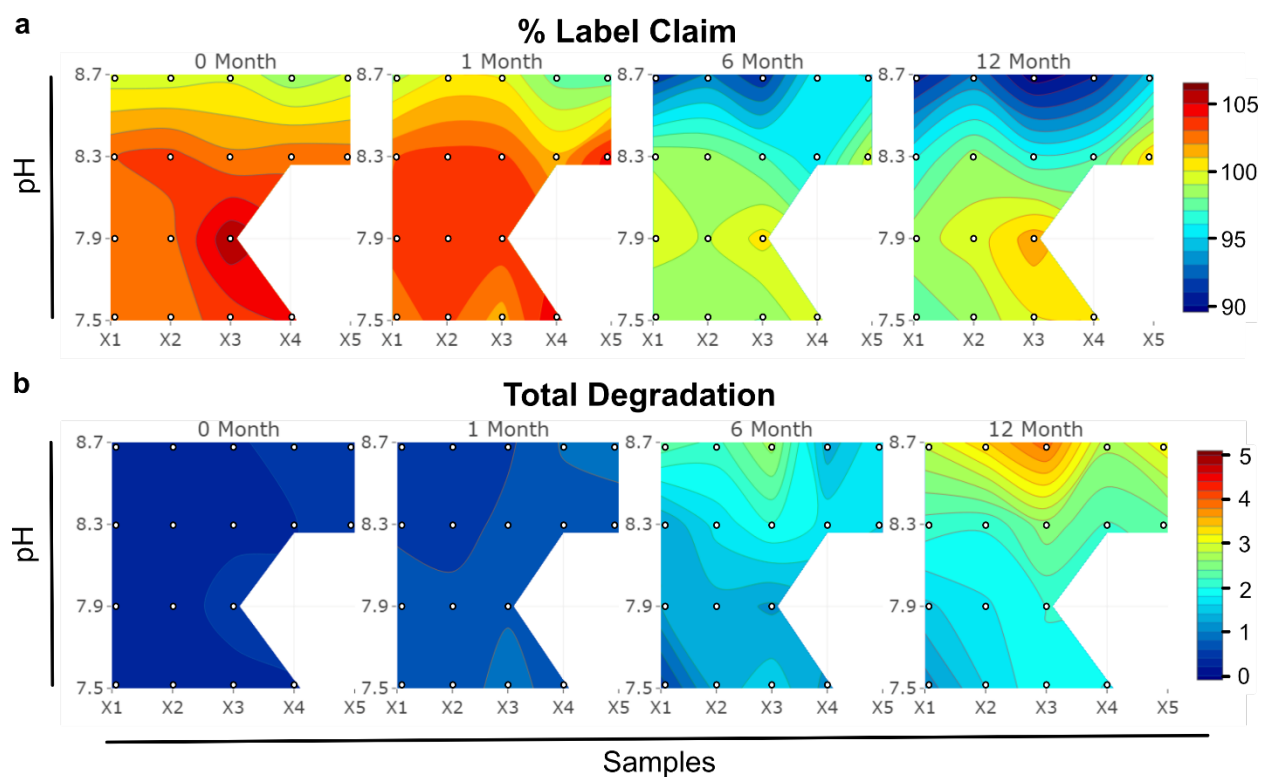
**Figure 6:** ThioT curves for the (a) initial, (b) 1-month, (c) 6-month, and (d) 12-month time-points.

**Table 4:** Test and training set results for the prediction of %LC and TD from the only the ThioT curve and from the ThioT curve with the storage conditions explicitly given to the model.

%LC					
Set	Method	Input	MAE	R <sup>2</sup>	RMSE
Test	RF	ThioT Only	3.037	0.254	3.761
		With Conditions	0.705	0.963	0.848
	SVM	ThioT Only	2.779	0.304	3.769
		With Conditions	0.976	0.931	1.174
	MLP	ThioT Only	4.131	0.106	4.667
		With Conditions	1.367	0.880	1.738
Train	RF	ThioT Only	1.620	0.894	2.147
		With Conditions	0.465	0.987	0.633
	SVM	ThioT Only	2.878	0.468	4.270
		With Conditions	0.852	0.963	1.128
	MLP	ThioT Only	4.226	0.155	5.102
		With Conditions	1.491	0.902	1.775
TD					
Set	Method	Input	MAE	R <sup>2</sup>	RMSE
Test	RF	ThioT Only	0.387	0.457	0.596
		With Conditions	0.121	0.937	0.188
	SVM	ThioT Only	0.310	0.494	0.493
		With Conditions	0.132	0.954	0.172
	MLP	ThioT Only	0.662	0.315	0.850
		With Conditions	0.148	0.957	0.180
Train	RF	ThioT Only	0.254	0.916	0.400
		With Conditions	0.103	0.983	0.156
	SVM	ThioT Only	0.405	0.611	0.774
		With Conditions	0.124	0.985	0.173
	MLP	ThioT Only	0.709	0.344	0.976
		With Conditions	0.188	0.957	0.264



**Figure 7:** Importance of a ThioT measurement at a given time for the prediction of %LC and TD using a Random Forest model.



○ True Samples:

X1 = O1, P1, Q1, R1    X3 = O3, T2, Q3, R3    X5 = NA, NA, S1, S2

X2 = O2, P3, Q2, R2    X4 = T1, NA, T3, T4

**Figure 8:** Relationship of chemical and physical stability using the Random Forest model trained only on ThioT assay data to predict a) % LC and b) TD. The Y-axis shows the pH and the movement across the X-axis only means Different samples having the same pH. Positions of true samples are labeled on the contour surface using white circles.

## References

- (1) D'Addio, S. M.; Bothe, J. R.; Neri, C.; Walsh, P. L.; Zhang, J.; Pierson, E.; Mao, Y.; Gindy, M.; Leone, A.; Templeton, A. C. New and Evolving Techniques for the Characterization of Peptide Therapeutics. *J. Pharm. Sci.* **2016**, *105* (10), 2989–3006.
- (2) Market, P. T. Peptide Therapeutics Market (by Applications, by Route of Administration, and by Marketing Status)—Global Industry Analysis, Size, Share, Growth, Trends and Forecast 2014–2020. 2015.
- (3) Waterman, K. C. The Application of the Accelerated Stability Assessment Program (ASAP) to Quality by Design (QbD) for Drug Product Stability. *AAPS PharmSciTech* **2011**, *12* (3), 932–937.
- (4) Sengupta, P.; Chatterjee, B.; Tekade, R. K. Current Regulatory Requirements and Practical Approaches for Stability Analysis of Pharmaceutical Products: A Comprehensive Review. *Int. J. Pharm.* **2018**, *543* (1–2), 328–344.
- (5) Li, H.; Nadig, D.; Kuzmission, A.; Riley, C. M. Prediction of the Changes in Drug Dissolution from an Immediate-Release Tablet Containing Two Active Pharmaceutical Ingredients Using an Accelerated Stability Assessment Program (ASAP Prime®). *AAPS Open* **2016**, *2* (1), 1–9.
- (6) Guideline, I. C. H. H. T. Stability Testing of New Drug Substances and Products. *Q1A (R2)*, *Curr. step* **2003**, *4*, 1–24.
- (7) Rosenberg, A. S. Effects of Protein Aggregates: An Immunologic Perspective. *AAPS J.* **2006**, *8* (3), E501–E507.
- (8) Jiskoot, W.; Randolph, T. W.; Volkin, D. B.; Middaugh, C. R.; Schöneich, C.; Winter, G.; Friess, W.; Crommelin, D. J. A.; Carpenter, J. F. Protein Instability and Immunogenicity: Roadblocks to Clinical Application of Injectable Protein Delivery Systems for Sustained Release. *J. Pharm. Sci.* **2012**, *101* (3), 946–954.
- (9) Nielsen, M. K.; Ahneman, D. T.; Riera, O.; Doyle, A. G. Deoxyfluorination with Sulfonyl Fluorides: Navigating Reaction Space with Machine Learning. *J. Am. Chem. Soc.* **2018**, *140* (15), 5004–5008. <https://doi.org/10.1021/jacs.8b01523>.
- (10) Ahneman, D. T.; Estrada, J. G.; Lin, S.; Dreher, S. D.; Doyle, A. G. Predicting Reaction Performance in C–N Cross-Coupling Using Machine Learning. *Science* (80-. ). **2018**, *360* (6385), 186–190. <https://doi.org/10.1126/science.aar5169>.
- (11) Gao, H.; Struble, T. J.; Coley, C. W.; Wang, Y.; Green, W. H.; Jensen, K. F. Using Machine Learning to Predict Suitable Conditions for Organic Reactions. *ACS Cent. Sci.* **2018**, *4* (11), 1465–1476. <https://doi.org/10.1021/acscentsci.8b00357>.
- (12) Coley, C. W.; Green, W. H.; Jensen, K. F. Machine Learning in Computer-Aided Synthesis Planning. *Acc. Chem. Res.* **2018**, *51* (5). <https://doi.org/10.1021/acs.accounts.8b00087>.
- (13) Fine, J.; Kuan-Yu Liu, J.; Beck, A.; Alzarini, K. Z.; Ma, X.; Boulos, V. M.; Kenttämä, H. I.; Chopra, G. Graph-Based Machine Learning Interprets and Predicts Diagnostic Isomer-Selective Ion-Molecule Reactions in Tandem Mass Spectrometry. *Chem. Sci.* **2020**, *11* (43), 11849–11858. <https://doi.org/10.1039/d0sc02530e>.
- (14) Lai, P.-K.; Fernando, A.; Cloutier, T. K.; Kingsbury, J. S.; Gokarn, Y.; Halloran, K. T.; Calero-Rubio,

- C.; Trout, B. L. Machine Learning Feature Selection for Predicting High Concentration Therapeutic Antibody Aggregation. *J. Pharm. Sci.* **2021**, *110* (4), 1583–1591.
- (15) Lai, P.-K.; Fernando, A.; Cloutier, T. K.; Gokarn, Y.; Zhang, J.; Schwenger, W.; Chari, R.; Calero-Rubio, C.; Trout, B. L. Machine Learning Applied to Determine the Molecular Descriptors Responsible for the Viscosity Behavior of Concentrated Therapeutic Antibodies. *Mol. Pharm.* **2021**, *18* (3), 1167–1175.
  - (16) Xu, Y.; Goodacre, R. On Splitting Training and Validation Set: A Comparative Study of Cross-Validation, Bootstrap and Systematic Sampling for Estimating the Generalization Performance of Supervised Learning. *J. Anal. Test.* **2018**, *2* (3), 249–262. <https://doi.org/10.1007/s41664-018-0068-2>.
  - (17) Fine, J. A.; Rajasekar, A. A.; Jethava, K. P.; Chopra, G. Spectral Deep Learning for Prediction and Prospective Validation of Functional Groups. *Chem. Sci.* **2020**, *11* (18), 4618–4630. <https://doi.org/10.1039/c9sc06240h>.
  - (18) Ma, J.; Sheridan, R. P.; Liaw, A.; Dahl, G. E.; Svetnik, V. Deep Neural Nets as a Method for Quantitative Structure-Activity Relationships. *J. Chem. Inf. Model.* **2015**, *55* (2), 263–274. <https://doi.org/10.1021/ci500747n>.
  - (19) Wallach, I.; Heifets, A. Most Ligand-Based Classification Benchmarks Reward Memorization Rather than Generalization. *J. Chem. Inf. Model.* **2018**, *58* (5). <https://doi.org/10.1021/acs.jcim.7b00403>.
  - (20) Szucs, R.; Brown, R.; Brunelli, C.; Heaton, J. C.; Hradski, J. Structure Driven Prediction of Chromatographic Retention Times: Applications to Pharmaceutical Analysis. *Int. J. Mol. Sci.* **2021**, *22* (8). <https://doi.org/10.3390/ijms22083848>.
  - (21) P. Jethava, K.; Fine, J.; Chen, Y.; Hossain, A.; Chopra, G. Accelerated Reactivity Mechanism and Interpretable Machine Learning Model of N-Sulfonylimines toward Fast Multicomponent Reactions. *Org. Lett.* **2020**, *22* (21), 8480–8486. <https://doi.org/10.1021/acs.orglett.0c03083>.
  - (22) Wen, Y.; Amos, R. I. J.; Talebi, M.; Szucs, R.; Dolan, J. W.; Pohl, C. A.; Haddad, P. R. Retention Index Prediction Using Quantitative Structure-Retention Relationships for Improving Structure Identification in Nontargeted Metabolomics. *Anal. Chem.* **2018**, *90* (15), 9434–9440. <https://doi.org/10.1021/acs.analchem.8b02084>.
  - (23) Kapoor, Y.; Milewski, M.; Dick, L.; Zhang, J.; Bothe, J. R.; Gehrt, M.; Manser, K.; Nissley, B.; Petrescu, I.; Johnson, P.; Burton, S.; Moseman, J.; Hua, V.; Grunewald, T.; Tomai, M.; Smith, R. Coated Microneedles for Transdermal Delivery of a Potent Pharmaceutical Peptide. *Biomed. Microdevices* **2020**, *22* (1), 1–10. <https://doi.org/10.1007/s10544-019-0462-1>.
  - (24) Max, K.; Weston, S.; Keefer, C.; Engelhardt, A.; Cooper, T.; Mayer, Z.; Kenkel, B.; Team, R. C.; Benesty, M.; Lescarbeau, R.; Ziem, A.; Scrucca, L.; Tang, Y.; Candan, C. Classification and Regression Training. 2016, p 198.
  - (25) Kuhn, M. Building Predictive Models in R Using the Caret Package. *J. Stat. Software, Artic.* **2008**, *28* (5), 1–26. <https://doi.org/10.18637/jss.v028.i05>.
  - (26) Friedman, J.; Hastie, T.; Tibshirani, R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *J. Stat. Softw.* **2010**, *33* (1), 1–22. <https://doi.org/10.18637/jss.v033.i01>.

- (27) Breiman, L. Random Forests. *Mach. Learn.* **2001**, 9 (1), 5–32. <https://doi.org/10.1186/1478-7954-9-29>.
- (28) Wright, M. N.; Ziegler, A. Ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *J. Stat. Softw.* **2017**, 77 (1). <https://doi.org/10.18637/jss.v077.i01>.
- (29) Sanchez, G. PLS Path Modeling with R. *R Packag. Notes* **2013**, 235. <https://doi.org/citeulike-article-id:13341888>.
- (30) Li, H.; Liang, Y.; Xu, Q. Support Vector Machines and Its Applications in Chemistry. *Chemom. Intell. Lab. Syst.* **2009**, 95 (2), 188–198. <https://doi.org/10.1016/j.chemolab.2008.10.007>.
- (31) Fan, R. E.; Chang, K. W.; Hsieh, C. J.; Wang, X. R.; Lin, C. J. LIBLINEAR: A Library for Large Linear Classification. *J. Mach. Learn. Res.* **2008**, 9 (2008), 1871–1874.
- (32) Olden, J. D.; Joy, M. K.; Death, R. G. An Accurate Comparison of Methods for Quantifying Variable Importance in Artificial Neural Networks Using Simulated Data. *Ecol. Modell.* **2004**, 178 (3–4), 389–397. <https://doi.org/10.1016/j.ecolmodel.2004.03.013>.