Rosetta Machine Learning Models Accurately Classify Positional Effects of Thioamides on Proteolysis

Sam Giannakoulias,[a] Sumant Shringari,[a] Chunxiao Liu,[a,b] Hoang Anh T. Phan,[a] Taylor M. Barrett,[a] John J. Ferrie,*[,c] and E. James Petersson*[,a]

[a]Department of Chemistry, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA

ABSTRACT

Thioamide substitutions of the peptide backbone have been shown to stabilize therapeutic and imaging peptides toward proteolysis. In order to rationally design thioamide modifications, we have developed a novel Rosetta custom score function to classify thioamide positional effects on proteolysis in substrates of serine and cysteine proteases. Peptides of interest were docked into proteases using the FlexPepDock application in Rosetta. Docked complexes were modified to contain thioamides parameterized through the creation of custom atom types in Rosetta based on *ab intio* simulations. Thioamide complexes were simulated and the resultant structural complexes provided features for machine learning classification as the decomposed values of the Rosetta score function. An ensemble, majority voting model was developed to be a robust predictor of previously unpublished thioamide proteolysis holdout data. Theoretical control simulations with pseudo-atoms that modulate only one physical characteristic of the thioamide show differential effects on prediction accuracy by the optimized voting classification model. These pseudo-atom model simulations, as well as statistical analyses of the full thioamide simulations, implicate steric effects on peptide binding as being primarily responsible for thioamide positional effects on proteolytic resistance.

INTRODUCTION

Hundreds of peptide therapeutics are under clinical development globally, sixty of which have already been approved for clinical use in the United States.[1] Peptide therapeutics comprise an interesting and diverse portion of FDA approved drugs. Peptides have been used in both cancer diagnosis and treatment, as well as directly as hormones such as GLP-1, or to disrupt protein–protein interactions.[2-3] Due to their diverse biochemical properties and applications, peptides have become a standard therapeutic strategy both academically and industrially resulting in tens of new peptide-based pharmaceuticals entering clinical trials annually.[1, 4]

While peptides can potentially act as external modulators of many biochemical systems of interest, they have intrinsic limitations on their ability to act as effective therapeutics. Cellular instability due to degradation by proteolytic enzymes and membrane impermeability due to the need for desolvation of the backbone as well as polar and charged residues prove to be major challenges for peptides as therapeutics.[5- 6-8] The popularity of cyclic peptides has grown in order to combat both issues.[9-12] Cyclic peptides typically exhibit increased stability in serum as opposed to their linear counterparts due to an inability of the cyclic peptide to adopt the extended conformation necessary to fit into a protease active site. Alternative methods for improving peptide serum stability have been achieved through peptidomimetic substitutions such as β-amino acids, *N*-alkyl glycine residues (peptoids), and thioamides.[13-19] Thioamides are a single atom substitution of oxygen-to-sulfur substitution of an amide bond, and have been shown to impart peptides with resistance to degradation by proteases if placed at or near the scissile bond.[17-20]

One of the principal challenges associated with designing proteolytically resistant thioamide containing peptides is identifying the position of thioamide incorporation. To date, there is no reliable way of rationally designing a thioamide containing peptide that will be resistant to

proteolysis. Our laboratory has previously shown that for a subset of the cysteine protease cathepsins (Cts), the positions where incorporation of a thioamide will impart resistance to proteolysis differ between each protease even for a singular peptide sequence.[19] This is particularly interesting as it suggests that even homologous proteases (35-59% sequence identity with conserved active site residues) with the same mechanism process the same peptide substrate differently. Additionally, we have shown similarly complex results for a subset of the trypsin-like serine proteases.[18] The positional differences found within the serine proteases were also distinctly different compared to the cysteine proteases.[18-19] While such findings would not be surprising in the context of sidechain modifications, one might expect the interactions of the backbone amides to be similar across proteases as this is integral to the serine/cysteine protease mechanism. Indeed, some previous studies of thioamides indicated that this is the case.[21-22]

Due to the wide array of potential applications for stabilized peptides, the need for an accurate computational method of predicting where a thioamide can imbue proteolytic stability for a general peptide sequence and protease combination is of great importance. The two previously noted works have attempted to use Rosetta modeling in order to explain the positional effects of thioamides by modeling the bound all-amide substrates, however, the identification of peptide-protein interactions was insufficient for establishing a clear mechanism for the thioamide activity or intuition for predicting thioamide impacts on proteolysis for new peptide/protease combinations.[18-19] In this work, we demonstrate that explicitly modeling the thioamide peptides in Rosetta provides such predictability, but that simply using the structures or energies from the thioamide simulations is insufficient. Rather, the simulations are used as inputs for machine learning to create a custom score function (Thio_Class) which accurately predicts thioamide proteolysis effects for holdout sets including novel data for cathepsin L. Lastly, through feature

analysis as well as the use of theoretical amide analogs with specific mixtures of amide and thioamide parameters, we identify which physical characteristics of the thioamide most influence the predictability of protease resistance.

COMPUTATIONAL METHODS

The work herein utilizes the Rosetta modeling suite, PyRosetta, and sci-kit learn.[23-25] Instruction on how to access our data and utilize our models can be found on our github at https://github.com/Sam-Giannakoulias/RML_ThioClass/tree/master/Anaconda. Development of a classification model which can accurately predict thioamide effects on proteolysis was accomplished using an approach similar to the one used by Shringari *et al*. in the prediction of $\Delta\Delta G$ of mutations at protein-protein interfaces.[26] The general approach in this work is to simulate the experimental complexes of interest and extract structural information from Rosetta Simulations as inputs for machine learning.

Experimental Dataset

The data used for PyRosetta simulation and machine learning are derived from previously published thioamide scanning data across different proteases and peptides as well as one previously unpublished data set.[18-19] The data set spans ten protease/peptide combinations where six include cysteine proteases (Cts B, Cts K, Cts L, Cts S, Cts V, and papain) and the remaining four include serine proteases (chymotrypsin, kallikrein, trypsin with Lys-containing substrates, and trypsin with Arg-containing substrates). Thioamide positions correspond to the amides on either side of the scissile bond (Fig 1). Those which are N-terminal to the scissile bond are considered PX positions (non-primed positions) where the value of X increases with each amide towards the N-terminus. Positions which are C-terminal to the scissile bond are considered PX'

positions (primed positions) where the value of X increases with each amide towards the C-terminus. The μ in the peptide sequences denotes a 7-methoxycoumarin-4-yl-alanine amino acid. For each position in the inhibitor peptides, a value of 1 or -1 was assigned to denote imbued resistance on proteolysis, which was defined as a (>1.7 fold) decrease in the cleavage rate and can be viewed for all members of the dataset in Table S2a/b compared to the control peptides.

Flexible Peptide Docking

Experimental complexes were simulated using the flexible peptide docking application in Rosetta.[27] In order to run FlexPepDock, both the peptides and proteases had to be prepared as inputs for docking. For proteases where a structure of the protease in complex with a peptide inhibitor existed, the native peptide inhibitor was trimmed and mutated to the sequence of interest using PyMOL and PyRosetta respectively. If an experimental structure of the protease in complex with a peptide inhibitor representing both primed and non-primed positions was not available, the corresponding peptides of interest were prepared externally using PyRosetta. The peptides were then manually docked into the protease active site maintaining proximity of the active site residue to scissile bond.

All of the protease-peptide complexes of interest were then simulated using the FlexPepDock application with the prepack and refine flags in Rosetta in order to optimize the binding interactions of the peptides in the context of the protease. The FlexPepDock refine protocol functions primarily by iteratively optimizing the backbone geometry of the peptide in addition to the rigid-body orientation of the peptide relative to the protease. Finally, "on the fly" side-chain optimizations are performed. Each of the initial complexes were simulated 100 times in the flexible

docking protocol and the output structures were sorted based on their full atom total score. The lowest energy complex from each FlexPepDock was then carried forward.

Thioamide Patches

Backbone thioamides were introduced into PyRosetta simulations using appropriate patch files. In order to accurately populate the thioamide patch file, *N*-acetyl-thioalanine methyl amide was optimized in Gaussian at the Hartree-Fock level of theory with a 6-31G(d) basis set according to Renfrew *et al*.[28-29] Atomic charges were calculated using CHELPG. Using these data, patch files for converting the carbonyl oxygen to a sulfur in the nth residue and adjusting the charge parameters of an amide to that of a thiomide for the nitrogen in the n + 1 residue were created. The patch file for the sulfur was written such that the mainchain oxygen atom was set to a virtual atom and replaced by a custom atom type TS. The atom type TS has the van der Waals radius of sulfur and the bond length of the carbon sulfur double bond from the HF/6-31G(d) calculation. We assigned the appropriate CHELPG charge to the sulfur and allowed hydrogen bonding interactions with the thioamide sulfur to be counted towards the hydrogen bonding score terms, as Lee *et al*. demonstrated that thioamide sulfurs have the ability to act as hydrogen bond acceptors albeit to a lesser extent compared to their oxoamide counterparts. The patch file for the nitrogen atom simply altered the charge to match that from CHELPG. Alternative patch files which retained subsets of these properties were created for the control simulations and can be found in the supplemental section titled Pseudo-atom Modeling.

"Local Relax" and Feature Generation

Relaxed complexes were modified to insert the thioamide at the appropriate locations P3-P3'. Following mutation, all complexes were put into five independent unconstrained relax trajectories. Relaxes were performed such that only the residues of the peptide and residues which contained a

C_α within 8 Å of the C_α of the residue containing the sulfur of the thioamide. We denote this as a "local relax", which has previously been used to show good correlations for protein design.[30] For each position in the peptide, a control simulation corresponding to a local relax without insertion of the thioamide was also performed. The locally relaxed complexes were scored with the beta_nov16 score function. The score terms of the Rosetta full atom score function were recorded for every residue and averaged over the five local relax simulations for both the all-amide and thioamide complexes. Following averaging, score deltas were computed, where thioamide scores were subtracted from the all-amide scores from both the protease and peptide. Features for machine learning were extracted as scores corresponding to the residue containing the sulfur of the thioamide, the residue containing the nitrogen of the thioamide, and the average of the residues which were locally relaxed.

Feature Engineering

Hypothesis testing was performed to identify which features if any were significantly different between the all-amide simulation and the thioamide simulations. The Wilcoxon signed rank test was used for hypothesis testing as there is no expectation of a normal distribution of features. Hypothesis testing demonstrated that almost half of the features were significantly different from the all-amide simulation.

In order to reduce the dimensionality of our features given the size of the experimental data set, features were selected by ranking the most important features through the SelectKBest module in sklearn with the χ^2 distribution as the score function.[25] The following nine machine learning algorithms were used with their default parameters to coarsely assess the effect of prediction accuracy as a function of number of features: Logistic (LOG), Support Vector (SVM), K Nearest Neighbors (KNN), Gaussian Naïve Bayes (GNB), Gradient Boosted Random Forest (GBT),

Gaussian Process (GPC), Ridge (KR), Stochastic Gradient Descent (SGD), and Linear Discriminant Analysis (LDA) classifiers were used to average the prediction accuracy from five-fold cross validation trials for 2 through 75 features.

Logistic Classification (LOG)

Logistic classification is the process of making a decision informed by logistic regression.[31] Logistic regression utilizes a quantitative response variable given by the logarithm of the odds of being classified in the ith group of a binary or higher order class response. The transformation of the response variable yields a continuous probability distribution which is bounded between 0 and 1. The transformation is that of a sigmoid function which takes the following form.

$$\sigma(Z) = \frac{1}{1+ e^{-z}} \tag{1}$$

If the response is binary as is in the case in this study, a logistic regression model is given as a weighted linear combination of input features plus a bias term in each instance.

$$p\left(y^{(i)} = 1 \mid x^{(i)}, w\right) = 1 - \left(\frac{1}{1+ e^{\left(w^T x^{(i)} + b\right)}}\right) \tag{2}$$

$$p\left(y^{(i)} = 0 \mid x^{(i)}, w\right) = 1 - \left(\frac{1}{1+ e^{\left(w^T x^{(i)} + b\right)}}\right) \tag{3}$$

Finally, in the logistic regression, the classification is made after probabilities are computed when the weights in the linear combination of terms has minimized the negative log likelihood or in other words, optimizing the average cross entropy which is given by Eqn. 4.[31]

$$J(\Omega) = -\frac{1}{m}\sum p_i log(y_i) + (1 - p_i)\log\left(1 - y_i\right) \tag{4}$$

Support Vector Classification

Support vector machines (SVMs) are used for classification tasks as they are highly effective at linearly separating data.[32] SVMs function through optimization of the margin of the data-separating hyperplane which takes the form:

$$w^T x + b = 0 \tag{5}$$

Optimization of the margin guaranties the lowest rate of misclassification as it provides the maximally wide boundary between the response classes. Due to noise or insufficiently descriptive features, a hard margin SVM which strictly splits the data may not be found. Therefore, soft margin, or error tolerant margins are found instead. Slack parameters ξ_i are introduced to create such tolerance. The form of the hyperplane can therefore be generalized to the following:

$$y_i(w^T x_i + b) \geq 1 - \xi_i \, , i = 1,2,3 \ldots . , n \tag{6}$$

The primal problem is now set to minimize $\frac{1}{2} w^T w + C \sum_{i=1}^{n} \xi_i$ with respect to w where the C parameter controls the extent with which the support vector machine is error tolerant.

Finally, the optimal weights and biases are found through Lagrange multipliers $\propto_i$.[32]

$$w = \sum_{i=1}^{n} \propto_i y_i x_i \, , b = \frac{\left(\sum_{i=1}^{n} (C - \propto_i) \propto_i y_i - \sum_{-=1}^{n} (C - \propto_i) \propto_i w^T x_i\right)}{\sum_{i=1}^{n} (C - \propto_i) \propto_i y_i} \tag{7}$$

K Nearest Neighbors Classification

In K nearest neighbors (KNN) classification, training data is clustered through an unsupervised method.[33] Following clustering, the dissimilarity measure is computed with the Minkowski distance which is given by:

$$distance(x_1, x_2) = \left(\sum_{i=1}^{n}(x_{1i} - x_{2i})^q\right)^{\frac{1}{q}} \quad (8)$$

Where q is a small positive tunable constant. Following computation of all distances, classification is determined by the value of K, or the number of nearest neighbors which are used to assign a cluster to the classified testing data.[33]

Gradient Boosted Random Forest Classification

Random forests are ensemble-based learning algorithms. Forests are comprised of *n* collections of de-correlated decision trees. Random Forest models utilize several decision trees as votes and ultimately use majority voting to make predictions. The architecture of a decision tree consists of a top node which is recursively split at nodal points until a terminal node is reached at which point a decision is made. Nodal points are split unambiguously based off a value called "entropy." Entropy is a measure of homogeneity within a subset data and is given by Eqn. 9, where p and q are the frequency of a feature in a class

$$entropy = -plog_2(p) - qlog_2(q) \quad (9)$$

The bounds of entropy are 0 and 1. When entropy is calculated using a feature, if classification is maximally split, which is to say half the points have been classified as 1 and the other half -1, entropy has been maximized at a value of 1. If there is any amount of unequal classification,

entropy will decrease all the way down to a value of 0 where all points are classified homogenously as either 1 or -1. Random forest methods split nodal points on features that maximize entropy. When feature importance is analyzed following classification in a random forest algorithm, the values for entropies at the nodal points are used to quantify the magnitude of importance a feature brings to the overall model. Finally, gradient boosting is the technique of further optimizing the ensemble behavior of the random forest through optimization of a loss function.[34]

Gaussian Process Classification

Similarly to Bayesian based classification algorithms, Gaussian Process Classification (GPC) assumes that the distribution of feature space takes the functional form of a Gaussian distribution.[35] GPC's function through application of the Gaussian distribution to a latent function which produces a continuous probability distribution that is then transformed by a logistic function. The predictions of GPC can be made by computing the following:

$$\bar{\pi}_* \simeq \mathrm{E}_q[\pi_*|X, y, x_*] = \int \sigma(f_*) q(f_*|X, y, x_*) df_* \quad (10)$$

The implementation used specifically in sklearn applies the Laplace approximation for the binary classification.[35]

Gaussian Naïve Bayes Classification

The Naïve Bayes classification algorithm is, in principal, based on applying Bayes' theorem with naïve independence assumptions between model features.[36] In Gaussian Naïve Bayes classification, there is an implicit assumption that the continuous features are normally distributed. For each feature, the mean and Bessel corrected variance are computed. Then, for any feature, the probability distribution can be computed as the following:

$$p(x = v\,|C_k) = \frac{1}{\sqrt{2\pi\sigma_k^2}} e^{-\left(\frac{(v-\mu_k)^2}{2\sigma_k^2}\right)} \quad (11)$$

For each feature and the respective distribution, test features can serve as inputs into the distributions to calculate likelihoods which are then multiplied together under independence assumptions. Finally, this is repeated for distributions coming from known -1 and 1 labels and whichever score is more likely is classified as such.[36]

Linear Discriminant Analysis

Linear discriminant analysis (LDA) is typically a dimensionality reduction technique, but can also be used in classification problems.[37] For a binary response class, the first assumption made in LDA is that the two conditional probability distributions are normally distributed with the following mean (μ) and covariance (∑) parameters.

$$p(x|y = -1), (\mu_{-1}, \sum_{-1}) \quad (12)$$

$$p(x|y = 1), \ \ (\mu_1, \sum_1) \quad (13)$$

Like the Gaussian Naïve Bayes classifier, classification can be predicted by computing likelihoods and selecting for the more likely distribution. LDA, however, has an additional intrinsic assumption that each of the feature random variables have the same finite variance. This is called the homoscedastic assumption and the implications are that the response class covariances are equal and have full rank.

$$\sum_{-1} = \sum_{1} = \sum \tag{14}$$

Finally, this leaves us with the criterion that for an arbitrary input vector to be in a given response class, is given purely by a linear combination of the known observations for some threshold T.[37]

$$c = \frac{1}{2}(T - \mu_{-1}{}^{T}\sum^{-1}\mu_{-1} + \mu_{1}{}^{T}\sum^{-1}\mu_{1}) \tag{15}$$

Ridge Classification

Like Logistic classification, Kernel Ridge (KR) classification utilizes Ridge Regression in order to classify the output.[38] Ridge regression with a linear kernel utilizes only the α parameter to introduce bias. The bias used in Ridge regression is L2 regularization which is a penalization of the sum of squared weights scaled by the α parameter. As α approaches zero, Ridge regression approaches ordinary linear regression. Ridge regression can be described in the following way:

$$\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \alpha \sum_{j=1}^{p} \beta_j^2 \tag{16}$$

Loss Function L2 Regularization

Here y_i are experimental values, x_i and β_j are the features and their corresponding weighting factors respectively. The α parameter acts as a tunable scaling factor for the L2 bias. The sum is performed over all experimental values, i, and features, j, to the respective total number of each, n and p. Ridge classification can be thought of simply as an SVM which utilizes a linear kernel and least squares loss.

Stochastic Gradient Descent Classification

Stochastic gradient descent (SGD) is not a classification method in and of itself, but rather a method for optimizing a loss function. In sklearn the SGD classifier allows for tuning of different loss functions alongside different regularization techniques. SGD optimizes a loss function through computation of the gradient at a random single point instead of the whole set of data like in traditional gradient descent.[39]

Parameter Tuning

For each of the ten different protease-peptide combinations, the corresponding six-point scanning data was used as a holdout set. Additionally, a holdout set corresponding to the Cts L and kallikrein data was created in order to train a model which has predictability against cysteine and serine proteases as well as across different peptide sequences. Each holdout set was validated by the nine algorithms that were tuned using five-fold cross validation (CV5) in an exhaustive grid search in sklearn. Tuning parameters can be found in Tables S17-S32 in SI.

RESULTS AND DISCUSSION

PyRosetta Simulations

For each protease and all-amide peptide combination, the complexes of the P3-P3' thioamide peptides (see Fig. 1) as well as the complex of the all-amide peptide were analyzed following local relax to identify any potential differences in structure induced by the incorporation of thioamides. We hypothesized that positions which demonstrated proteolytic resistance would show greater structural change when compared to nonperturbing positions. However, the structures from both the thioamide and all-amide complexes were extremely similar for all positions regardless of the

observed impact on proteolysis. CαRMSD analysis showed that across different relax trajectories, every position converged to a single structure for both the all-amide and thioamide complexes, with a CαRMSD of the locally relaxed residues no greater than 0.5 Å (data shown in SI, Tables S5-S6), and that the complexes retained all major contacts. Figure 2 shows Cts S in complex with the all-amide and P1 thioamide forms of the μLLKAAAμ peptide and trypsin in complex with the all-amide and P1 thioamide forms of the μLLRAAAμ pepride. In the case of Cts S, there does not appear to be any discernable difference between the P1 thioamide (blue) and the all-amide (green) complexes, and for trypsin, the only visible distinction between the complexes is a small translation of the C-terminal methoxycoumarin amino acid. Further investigation, shown in the inset panels, reveals that both the thioamide and all-amide peptides are making the exact same sets of polar contacts in the Cts S and trypsin active sites.

In addition to the Wilcoxon testing, analysis of the minimum, maximum, mean, and standard deviation metrics of the population values for each feature were analyzed. Tables S8-S9 in the SI section Feature Analysis demonstrate that while feature values from decoy sets were convergent, many features showed great variance over the population, suggesting utility in machine learning. Features for machine learning were generated by subtraction of the thioamide complex scores from the all-amide complex scores for each Rosetta energy feature.

Feature Selection

We analyzed the simulations to determine whether thioamide modeling could directly explain differences between resistant (an apparent decrease in the cleavage rate) and nonresistant positions in the peptides. Interestingly, although the simulated complexes were structurally similar, clear differences between the score terms of the Rosetta full atom score function were observed between

thioamide complexes and their all-amide counterparts. We investigated whether the Rosetta total REU scores were able to linearly separate our response classes (SI, Fig. S5).[40] Given that total score alone was not sufficient for perfect classification, we decided to train a new custom score function by further examining the decomposed score terms of the Rosetta score function. A Wilcoxon signed rank test of the score terms between these populations demonstrated that 34 of the 75 score terms were statistically different using a p value of 0.05. This appeared to be a surprisingly high number given how similar the complexes were structurally, and potentially speaks to the sensitivity of Rosetta score functions. In addition to the Wilcoxon testing, analysis of the minimum, maximum, mean, and standard deviation metrics of the population values for each feature were analyzed. Tables S8-S9 in the SI section Feature Analysis demonstrate that while feature values from decoy sets were convergent, many features showed great variance over the population, suggesting utility in machine learning. Features for machine learning were generated by subtraction of the thioamide complex scores from the all-amide complex scores for each Rosetta energy feature.

In order to reduce the dimensionality of the system without losing direct information about our features through techniques like principal component analysis, we used the select KBest module to perform univariate statistical analysis. Feature selection coupled with untuned model prediction showed that when considering each of the nine models discussed in the methods section in a majority voting model, seven features provided the most accurate prediction of cleavability (SI, Table S15). Those seven features include: total residue scores and van der Waals repulsion terms for the residues which contain the sulfur atom and the nitrogen atom of the thioamide. Other terms include the change in total score for the entire protein/peptide complex, the long-range back bone hydrogen bonding term from the residue containing the sulfur of the thioamide, and the intra-

residue electrostatic interaction term for the residue containing the nitrogen. These terms correspond directly to the overall energy of the thioamide containing residues and parameters which are altered in thioamide substitution, such as repulsion, hydrogen bond ability, and atomic charges. These results may seem unsurprising because these features are directly affected by the modifications made to introduce the thioamide patches. It is important to note however, that other differences which may be expected to result from the altered properties of the thioamide (e.g., solvation) were identified by the hypothesis testing, although they did not appear to be the most important features by our univariate statistical testing.

Prediction of the holdout set

Following feature selection, exhaustive grid searching (CV5) was performed to tune the hyperparameters of our models. Our holdout set is comprised of the unpublished Cts L dataset and the Kallikrein dataset (selected because of an observed P3 thioamide effect that was not anticipated based on the structural similarity of trypsin and kallikrein) was created in order to validate that our trained model will have predictability against both cysteine and serine proteases as well as against potentially diverse peptide sequences. When we applied our tuned models to this holdout set, we found considerable success. Figure 3 shows the accuracy of each model in predicting the holdout set. Five of the nine models were able to completely recapitulate the thioamide effects in the holdout scanning data. At worst, two models predict the holdout data at 83%, corresponding to 10 out of 12 positions predicted correctly, and a 33% enrichment compared to random chance. Analysis of additional holdout sets can be found in SI. Given that we had five models capable of perfectly classifying the holdout set, we decided to create an ensemble majority voting model composed of these five individual models for practical use. Although we cannot test the efficacy of the ensemble model relative to any of the individual models in this work since they already

predict the holdout set with 100% accuracy, we will further investigate the majority voting model's ability to predict larger novel holdout datasets as additional thioamide peptides and proteases are studied in our laboratory.[41]

Following accurate prediction of the holdout set, we wanted to investigate which features the models found to be most important. Based on the kernels identified using exhaustive grid searching, extraction of this information from the LOG, SVM, KR, and LDA models is available in sklearn, while for models like GPC, it is not. Analysis of the normalized weights of the features shows that the LOG, SVM, and KR models all weight features extremely similarly (Fig. 4). In these models, the highest weighted features generally correspond to the total residue scores of the two residues constituting the thioamide, followed by the repulsive terms of these residues, and then finally the intra-electronic and hydrogen bonding terms. Interestingly, the LDA model achieves the same perfect accuracy, but has an entirely different weighting of the features. The LDA model barely considers the total residue scores of the residues constituting the thioamide and relies almost exclusively on the repulsive term of the residue containing the sulfur of the thioamide with some contribution from the change in global total score and the repulsive term of the residue containing the nitrogen of the thioamide.

Pseudo-atom Modeling

Accurate classification of our holdout set with a small number of features and their corresponding feature importance analysis demonstrated that in general, all model terms were important in predicting the thioamide scanning data. However, the LDA feature importance strongly suggested that thioamides may function in inducing proteolytic resistance through changes in steric repulsion as the repulsive van der Waals term was the dominant feature. In order to further test which features

made the largest contribution to predicting the thioamide effects, we generated alternative patch files which create pseudo-atoms which have mixed properties of oxygen and sulfur atoms that allow us to more deeply probe specific characteristics of the thioamide. These patches act as controls for analyzing the impact of individual properties of a thioamide, allowing us to test our hypotheses, which are experimentally intractable, via molecular modeling. Specifically, we hold all properties of sulfur constant and independently vary the van der Waals radius to that of oxygen, the bond length to that of an amide carbon-oxygen double bond, the charge to that of an amide oxygen, and the ability for the sulfur to act as a hydrogen bond acceptor. Additionally, we wrote a patch for the nitrogen, turning off its ability to act as a hydrogen bond donor. As before, the same local relax simulations at each P position were performed following modification of FlexPepDocked structures using these pseudo-atoms, and the Rosetta score features were extracted and score deltas were computed through subtraction of the corresponding values in the all-amide peptide simulations as in the full thioamide patch analysis. When using features from the alternative patch simulations as inputs into our tuned ensemble majority voting model, we found variable prediction accuracy. Figure 5 shows the prediction accuracy of the features from each new pseudo-atom. Ultimately, changing the ability of the nitrogen to hydrogen bond did not have any effect on accuracy at all. However, if the sulfur is not allowed to be a hydrogen bond acceptor, or the charge of the sulfur was set to that of oxygen, we saw a decrease in predictive accuracy to 91.67%, corresponding to 11 out of 12 positions classified correctly. Even more interestingly, if the carbon-sulfur bond length was converted to that of a carbon-oxygen bond, accuracy decreased to 83.33% corresponding to 10 out of 12 positions classified correctly. Finally, when the van der Waals radius of the sulfur is converted back to that of an oxygen, we see the most dramatic decrease to 67% accuracy, corresponding to 8 out of 12 positions classified correctly. These results along

with the feature importance analysis have led us to a heuristic understanding that sterics, charge, and hydrogen bonding ability play a role in how the thioamide alters proteolysis, however the effect can be primarily attributed to the steric effects associated with the increased size of sulfur's van der Waals radius and carbon-sulfur bond length based on these modeling efforts.

In order to further investigate why some of the patches were more detrimental to prediction accuracy than others, we performed hypothesis testing between the features from the full patch to the alternative patches. Based on statistical significance alone, there did not appear to be one term, or a set of terms, that recapitulated the prediction data. We then performed a set of multiple linear regressions in which we used two of the model features at a time to attempt to recreate the observed decreases in prediction accuracy. We found there were many combinations which yielded strong linear correlations, with the highest being a product of the fa_rep_S and residue_total_score_N terms displaying an R^2 value of 0.92 (Fig. 6). These data further indicated to us that the method with which thioamides function to imbue proteolytic resistance is primarily through steric repulsion of the sulfur.

It is important to remember that all potential differences identified by Rosetta and machine learning in this study are from modeling peptide binding to single structure that is part of a dynamic process involving substrate binding, conformational changes of the active site, the chemical steps of cleavage (both active site Ser or Cys nucleophilic attack and subsequent resolution of the peptide ester by hydrolysis), and dissociation of the cleavage products. The modeled structure does not necessarily represent any specific intermediate along this pathway, but it is reasonable to assume that the simulated structures are most similar to the initially bound conformation. Therefore, the high accuracy of our predictions, feature analyses, and pseudo-atom modeling imply that thioamide substitution primarily affects this initial binding step via steric repulsion and that it is

rate-limiting. For trypsin, we have directly measured the rates of binding and catalysis.[18] We found that catalysis was the rate-limiting step for the all-amide substrates, but that binding and catalysis steps became comparable for the thioamides at the μM substrate concentrations used in our assays, with a significant unbinding rate. This is consistent with our modeling results. While a more extensive analysis of our machine learning models with a broader kinetic data set is clearly necessary, the strong correlations that we see support the idea that steric interactions in the initial binding step allow thioamides at certain locations to perturb cleavage.

CONCLUSION

In summary, complexes from thioamide scanning experiments were simulated using Rosetta and energy terms of the simulated complexes were then used as inputs into machine learning models that accurately predict the cleavage propensity of thioamide-substituted peptides for a diverse holdout dataset. Feature analysis of the highest performing models as well as control simulations suggested that resistance to proteolysis imbued by the thioamide is primarily the result of steric interactions in the initial binding step. While five individual models already predict the challenging holdout dataset with perfect accuracy, we have combined them in a majority voting ensemble to provide a potentially more robust model for testing with future thioamide proteolysis data sets. The resulting custom score function, Thio_Class, can easily be incorporated into Rosetta design to create novel thioamide peptides with exhibit enhanced proteolytic stability. We will pursue such applications in our future studies of thioamide peptides for therapeutic and imaging applications and we will continue to expand our repertoire of custom score functions to study other aspects of protein structure and binding interactions.

Figures

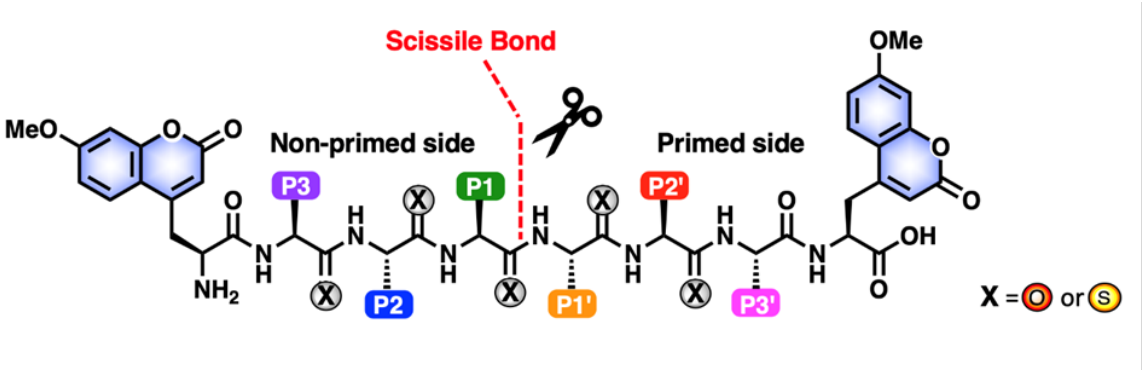


Figure 1. Thioamide positional definitions. Peptides contain methoxycoumarinyl alanine (μ) residues at both termini, and amide (X=O) or thioamide (X=S) residues between. The red line represents the cleavage site. Amino acid residues are denoted P3 to P1 from the N-terminus to the scissile bond, and P1' to P3' from the scissile bond to the C-terminus.

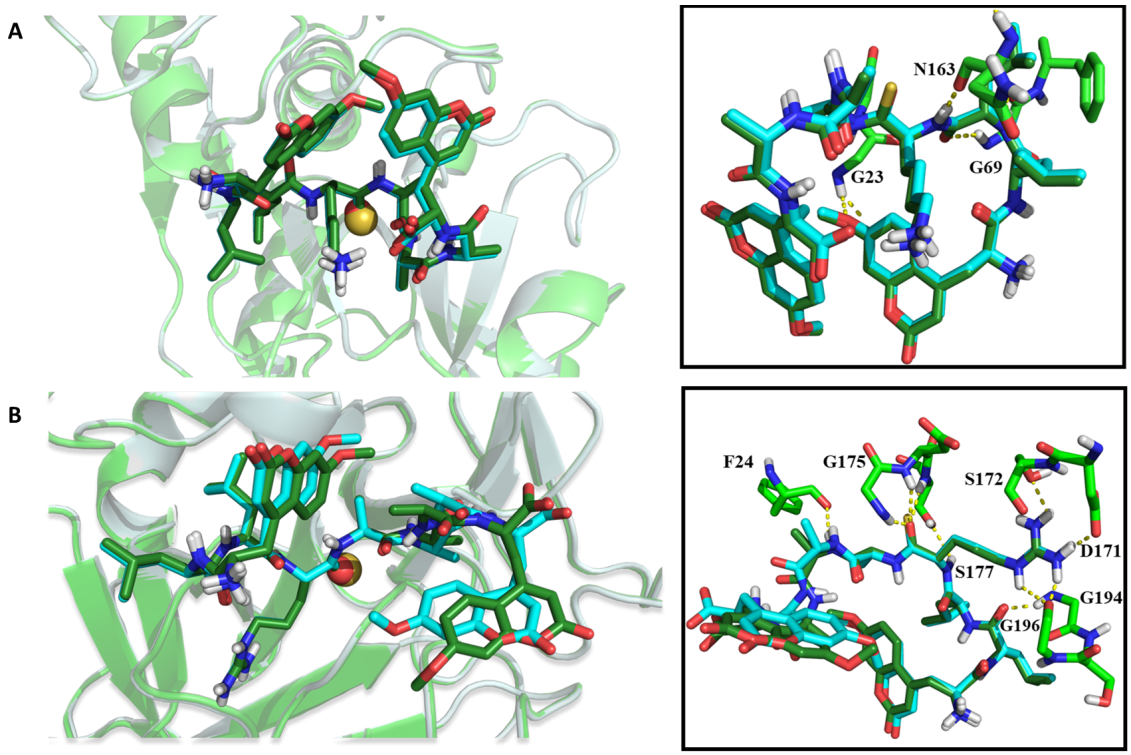


Figure 2. Comparison of thioamide complexes (cyan) versus all-amide control (green). (A) Cts S (1MS6) in complex with μLLKAAAμ or μLLK^{S}AAAμ thioamide peptide (P1 position). (B) Trypsin (2PTC) in complex with μLLRAAAμ or μLLR^{S}AAAμ thioamide peptide (P1 position).

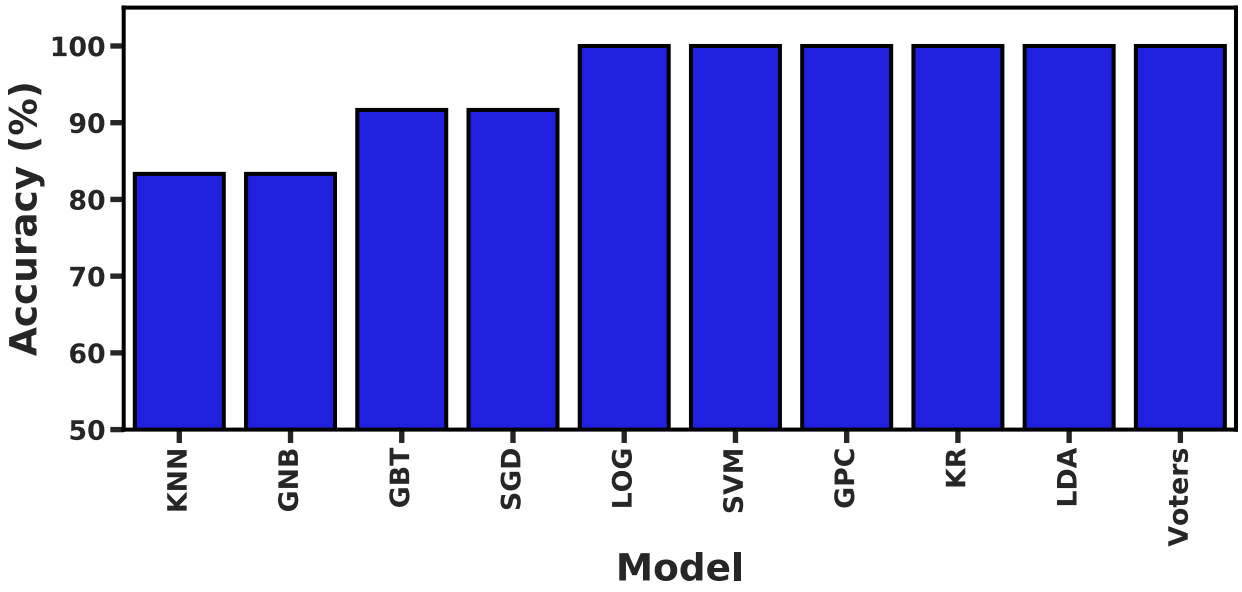


Figure 3. Bar chart demonstrating prediction accuracy of the Cts L and kallikrein holdout set by various models. KNN (KNeighbors Classifier), GNB (Gaussian Naïve Bayes Classifier), GBT (Gradient Boosted Random Forest Classifier), SGD (Stochastic Gradient Descent Classifier), LOG (Logistic Classifier), SVM (Support Vector Classifier), GPC (Gaussian Process Classifier), KR (Linear Kernel Ridge Classifier), LDA (Linear Discriminant Analysis Classifier), Voters (Majority Voters Model: LOG, SVM, GPC, KR, LDA).

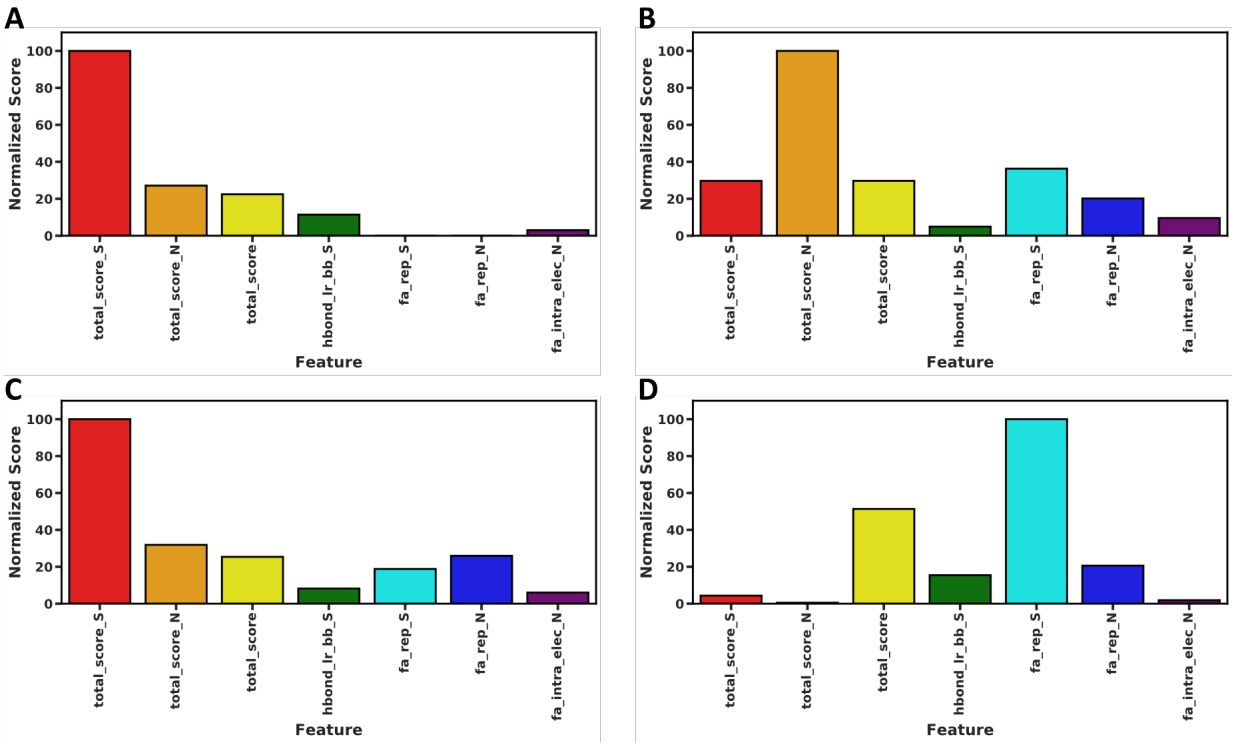


Figure 4. Normalized feature importance for models used in Voters classification. (A) Logistic (LOG) classification feature importance, (B) Linear SVM classification feature importance, (C) Ridge (KR) classification feature importance, and (D) LDA classification feature importance.

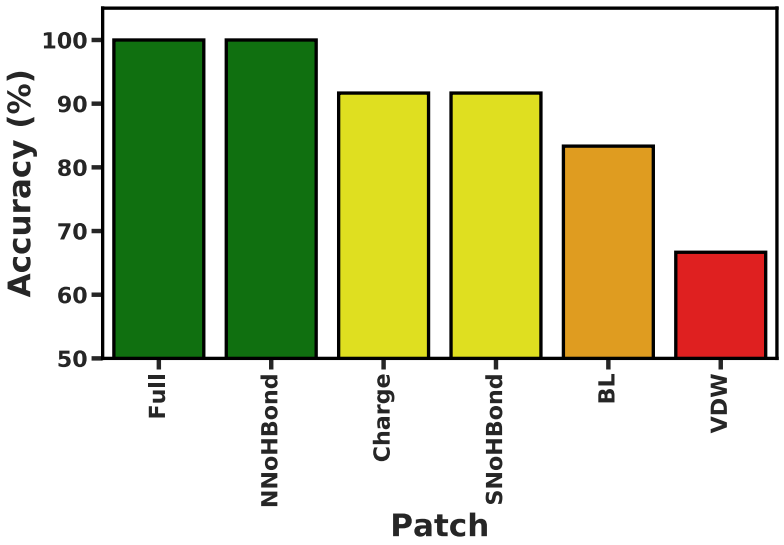


Figure 5. Bar chart demonstrating prediction accuracies of the features generated from simulation with alternative patch files which include pseudo-atoms that retain only some of the physical characteristics of thioamides. The Full, NNoHBond (without the ability for the nitrogen to hydrogen bond), Charge, SNoHBond (without the ability for the sulfur to hydrogen bond), bond length (BL), NNoHBond (without the ability for the nitrogen to hydrogen bond) and van der Waals (VDW) patches are defined in SI.

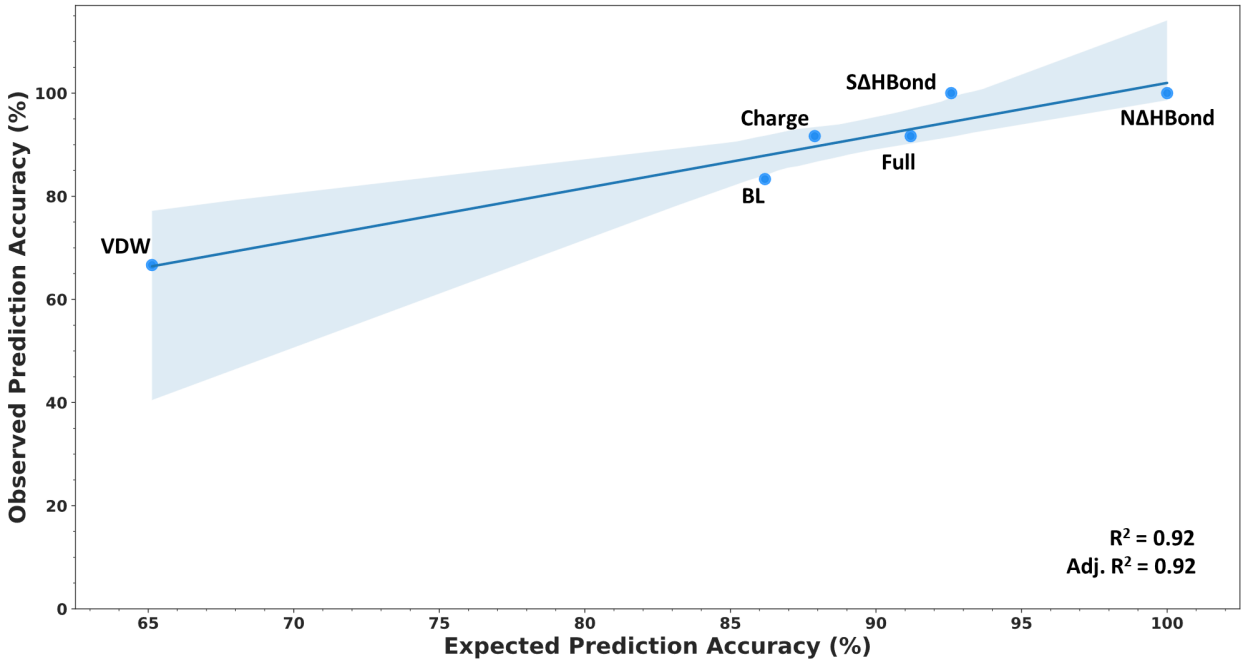


Figure 6. Scatter plot showing the relationship between a linear combination of model features (fa_rep_S and residue_total_score_N) and the prediction accuracy from the patch simulations.

Supporting Information.

The Supporting Information (SI) is available free of charge on the ACS Publications website at http://pubs.acs.org . SI includes experimental methods and proteolysis data, Rosetta simulation details, machine learning descriptions, feature analysis, and associated references (PDF).

AUTHOR INFORMATION

Corresponding Author

* JJF: jferrie@berkeley.edu; EJP: ejpetersson@sas.upenn.edu

Present Addresses

[b]Key Laboratory for Northern Urban Agriculture of Ministry of Agriculture and Rural Affairs, Beijing University of Agriculture, Beijing 102206, P. R. China

[c]Department of Molecular & Cell Biology, University of California, Berkeley, Berkeley, California, 94720, United States

Author Contributions

S.G.G. and S.R. performed Rosetta simulations and analysis, guided by J.J.F. C.L., H.A.P., and T.M.B. performed protease cleavage experiments. The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript.

ACKNOWLEDGMENT

This work was supported by the University of Pennsylvania and the National Science Foundation (NSF CHE-1708759 to E.J.P.). T.M.B. thanks the NIH for funding through the Chemistry Biology Interface Training Program (T32 GM071399). J.J.F. and S.G.G. thank the NSF for funding through

the NSF Graduate Research Fellowship Program (DGE-1321851 and DGE-1845298, respectively).

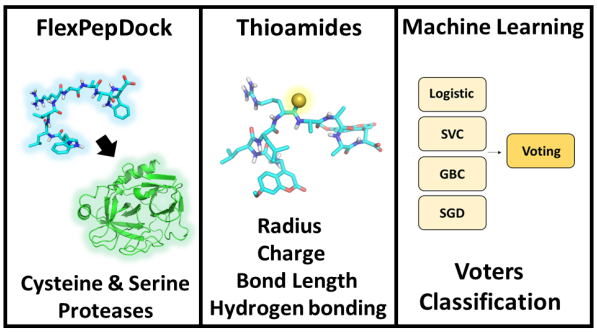
REFERENCES

1. Lee, A. C. L.; Harris, J. L.; Khanna, K. K.; Hong, J. H., A Comprehensive Review on Current Advances in Peptide Drug Development and Design. *Int. J. Mol. Sci.* **2019,** *20*, 21.
2. Cunningham, A. D.; Qvit, N.; Mochly-Rosen, D., Peptides and Peptidomimetics as Regulators of Protein-Protein Interactions. *Curr. Opin. Struct. Biol.* **2017,** *44*, 59-66.
3. Nauck, M. A., Glucagon-Like Peptide 1 (Glp-1): A Potent Gut Hormone with a Possible Therapeutic Perspective. *Acta Diabetol.* **1998,** *35*, 117-129.
4. Lau, J. L.; Dunn, M. K., Therapeutic Peptides: Historical Perspectives, Current Development Trends, and Future Directions. *Bioorg. Med. Chem.* **2018,** *26*, 2700-2707.
5. Otvos, L.; Wade, J. D., Current Challenges in Peptide-Based Drug Discovery. *Front. Chem.* **2014,** *2*, 4.
6. Craik, D. J.; Fairlie, D. P.; Liras, S.; Price, D., The Future of Peptide-Based Drugs. *Chem. Biol. Drug Des.* **2013,** *81*, 136-147.
7. Fosgerau, K.; Hoffmann, T., Peptide Therapeutics: Current Status and Future Directions. *Drug Discov. Today* **2015,** *20*, 122-128.
8. Henninot, A.; Collins, J. C.; Nuss, J. M., The Current State of Peptide Drug Discovery: Back to the Future? *J. Med. Chem.* **2018,** *61*, 1382-1414.
9. Ferrie, J. J.; Gruskos, J. J.; Goldwaser, A. L.; Decker, M. E.; Guarracino, D. A., A Comparative Protease Stability Study of Synthetic Macrocyclic Peptides That Mimic Two Endocrine Hormones. *Bioorg. Med. Chem. Lett.* **2013,** *23*, 989-995.
10. Gang, D.; Kim, D. W.; Park, H. S., Cyclic Peptides: Promising Scaffolds for Biopharmaceuticals. *Genes* **2018,** *9*, 15.
11. Bockus, A. T.; McEwen, C. M.; Lokey, R. S., Form and Function in Cyclic Peptide Natural Products: A Pharmacokinetic Perspective. *Curr. Top. Med. Chem.* **2013,** *13*, 821-836.
12. Guarracino, D. A.; Riordan, J. A.; Barreto, G. M.; Oldfield, A. L.; Kouba, C. M.; Agrinsoni, D., Macrocyclic Control in Helix Mimetics. *Chem. Rev.* **2019,** *119*, 9915-9949.
13. Hook, D. F.; Bindschadler, P.; Mahajan, Y. R.; Sebesta, R.; Kast, P.; Seebach, D., The Proteolytic Stability of 'Designed' Beta-Peptides Containing Alpha-Peptide-Bond Mimics and of Mixed Alpha,Beta-Peptides: Application to the Construction of Mhc-Binding Peptides. *Chem. Biodivers.* **2005,** *2*, 591-632.
14. Schwochert, J.; Turner, R.; Thang, M.; Berkeley, R. F.; Ponkey, A. R.; Rodriguez, K. M.; Leung, S. S. F.; Khunte, B.; Goetz, G.; Limberakis, C. *et al.*, Peptide to Peptoid Substitutions Increase Cell Permeability in Cyclic Hexapeptides. *Org. Lett.* **2015,** *17*, 2928-2931.
15. Bottger, R.; Hoffmann, R.; Knappe, D., Differential Stability of Therapeutic Peptides with Different Proteolytic Cleavage Sites in Blood, Plasma and Serum. *PLoS One* **2017,** *12*, 15.
16. Verma, H.; Khatri, B.; Chakraborti, S.; Chatterjee, J., Increasing the Bioactive Space of Peptide Macrocycles by Thioamide Substitution. *Chem. Sci.* **2018,** *9*, 2443-2451.
17. Chen, X.; Mietlicki-Baase, E. G.; Barrett, T. M.; McGrath, L. E.; Koch-Laskowski, K.; Ferrie, J. J.; Hayes, M. R.; Petersson, E. J., Thioamide Substitution Selectively Modulates Proteolysis and Receptor Activity of Therapeutic Peptide Hormones. *J. Am. Chem. Soc.* **2017,** *139*, 16688-16695.
18. Barrett, T. M.; Chen, X. S.; Liu, C. X.; Giannakoulias, S.; Phan, H. A. T.; Wang, J. L.; Keenan, E. K.; Karpowicz, R. J.; Petersson, E. J., Studies of Thioamide Effects on Serine Protease Activity Enable Two-Site Stabilization of Cancer Imaging Peptides. *ACS Chem. Biol.* **2020,** *15*, 774-779.

19. Liu, C. X.; Barrett, T. M.; Chen, X.; Ferrie, J. J.; Petersson, E. J., Fluorescent Probes for Studying Thioamide Positional Effects on Proteolysis Reveal Insight into Resistance to Cysteine Proteases. *ChemBioChem* **2019,** *20*, 2059-2062.
20. Mahanta, N.; Szantai-Kis, D. M.; Petersson, E. J.; Mitchell, D. A., Biosynthesis and Chemical Applications of Thioamides. *ACS Chem. Biol.* **2019,** *14*, 142-163.
21. Cho, K., Synthesis of Fluorescent Peptidyl Thioneamides and the Assay of Papain in the Presence of Trypsin. *Anal. Biochem.* **1987,** *164*, 248-253.
22. Asboth, B.; Polgar, L., Transition-State Stabilization at the Oxyanion Binding-Sites of Serine and Thiol Proteinases - Hydrolyses of Thiono and Oxygen Esters. *Biochemistry* **1983,** *22*, 117-122.
23. Kaufmann, K. W.; Lemmon, G. H.; DeLuca, S. L.; Sheehan, J. H.; Meiler, J., Practically Useful: What the Rosetta Protein Modeling Suite Can Do for You. *Biochemistry* **2010,** *49*, 2987-2998.
24. Chaudhury, S.; Lyskov, S.; Gray, J. J., Pyrosetta: A Script-Based Interface for Implementing Molecular Modeling Algorithms Using Rosetta. *Bioinformatics* **2010,** *26*, 689-691.
25. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, E., Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011,** *12*, 2825-2830.
26. Shringari, S. R.; Giannakoulias, S.; Ferrie, J. J.; Petersson, E. J., Rosetta Custom Score Functions Accurately Predict ΔΔG of Mutations at Protein–Protein Interfaces Using Machine Learning. *Chem. Commun.* **2020,** *56*, 6774-6777.
27. Raveh, B.; London, N.; Schueler-Furman, O., Sub-Angstrom Modeling of Complexes between Flexible Peptides and Globular Proteins. *Proteins* **2010,** *78*, 2029-2040.
28. Renfrew, P. D.; Choi, E. J.; Bonneau, R.; Kuhlman, B., Incorporation of Noncanonical Amino Acids into Rosetta and Use in Computational Protein-Peptide Interface Design. *PLoS One* **2012,** *7*, 15.
29. Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J., J. A.; Vreven, T.; Kudin, K. N.; Burant, J. C.; *et al. Gaussian 09*, Gaussian, Inc.: Pittsburgh PA, 2009.
30. Alford, R. F.; Leaver-Fay, A.; Jeliazkov, J. R.; O'Meara, M. J.; DiMaio, F. P.; Park, H.; Shapovalov, M. V.; Renfrew, P. D.; Mulligan, V. K.; Kappel, K.; Labonte, J. W.; Pacella, M. S.; Bonneau, R.; Bradley, P.; Dunbrack, R. L.; Das, R.; Baker, D.; Kuhlman, B.; Kortemme, T.; Gray, J. J., The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design. *J. Chem. Theory Comput.* **2017,** *13*, 3031-3048.
31. Cote, M. L.; Yoo, W.; Wenzlaff, A. S.; Prysak, G. M.; Santer, S. K.; Claeys, G. B.; Van Dyke, A. L.; Land, S. J.; Schwartz, A. G., Tobacco and Estrogen Metabolic Polymorphisms and Risk of Non-Small Cell Lung Cancer in Women. *Carcinogenesis* **2009,** *30*, 626-635.
32. Szymczak, S.; Biernacka, J. M.; Cordell, H. J.; Gonzalez-Recio, O.; Konig, I. R.; Zhang, H. P.; Sun, Y. V., Machine Learning in Genome-Wide Association Studies. *Genet. Epidemiol.* **2009,** *33*, S51-S57.
33. Fix, E.; Hodges, J. L., Discriminatory Analysis - Nonparametric Discrimination - Consistency Properties. *Int. Stat. Rev.* **1989,** *57*, 238-247.
34. Breiman, L., Random Forests. *Mach. Learn.* **2001,** *45*, 5-32.

35. Challis, E.; Hurley, P.; Serra, L.; Bozzali, M.; Oliver, S.; Cercignani, M., Gaussian Process Classification of Alzheimer's Disease and Mild Cognitive Impairment from Resting-State Fmri. *Neuroimage* **2015,** *112*, 232-243.
36. Misaki, M.; Kim, Y.; Bandettini, P. A.; Kriegeskorte, N., Comparison of Multivariate Classifiers and Response Normalizations for Pattern-Information Fmri. *Neuroimage* **2010,** *53*, 103-118.
37. Mitteroecker, P.; Bookstein, F., Linear Discrimination, Ordination, and the Visualization of Selection Gradients in Modern Morphometrics. *Evol. Biol.* **2011,** *38*, 100-114.
38. Whittaker, J. C.; Thompson, R.; Denham, M. C., Marker-Assisted Selection Using Ridge Regression. *Genet. Res.* **2000,** *75*, 249-252.
39. Huber, M. B.; Bunte, K.; Nagarajan, M. B.; Biehl, M.; Ray, L. A.; Wismuller, A., Texture Feature Ranking with Relevance Learning to Classify Interstitial Lung Disease Patterns. *Artif. Intell. Med.* **2012,** *56*, 91-97.
40. Alexander, N. S.; Preininger, A. M.; Kaya, A. I.; Stein, R. A.; Hamm, H. E.; Meiler, J., Energetic Analysis of the Rhodopsin-G-Protein Complex Links the Alpha 5 Helix to Gdp Release. *Nat. Struct. Mol. Biol.* **2014,** *21*, 56-+.
41. Dietterich T.G. Ensemble Methods in Machine Learning. *In: Multiple Classifier Systems*. **2000**. Lecture Notes in Computer Science, vol 1857. Springer, Berlin, Heidelberg.

TOC Graphic


FlexPepDock
Cysteine & Serine Proteases
Thioamides
Radius
Charge
Bond Length
Hydrogen bonding
Machine Learning
Logistic
SVC
GBC
SGD
Voting
Voters Classification