Machine Learning Methods for Peptide Toxicity Prediction

Andrzej Stanisławczyk[1], Marcin Król[2], Łukasz Proszek[2] and Mariusz Milik[2]
[1]AGH University of Science and Technology, Kraków, Poland; [2]Selvita S.A., Kraków, Poland
(corresponding author: mariusz.milik@selvita.com)

Selvita

Abstract

Sets of about 2000 toxic and 70000 non-toxic peptides 8 to 70 amino acid long were selected from the Uniprot[1] database, and used for training and performance tests of several machine learning methods.

Two variants of logistic regression[2] models, a multilayer perceptron[3] (MLP) neural network, and two variants of classification tree[4] models were tested. In these tests, best performance was obtained for the neural network model; however, authors suggest that this approach is less attractive for practical application because of the problem with biochemical interpretation of the obtained rules. This issue is often observed for applications of the neural network methodology. As the best compromise between classification performance and interpretability, authors propose the logistic regression model with interactions.

This work was partially supported by a grant from *PEPLASER*[5] collaborative project *(FP7-HEALTH-2007-B)*

Feature Extraction

The peptide sequences were processed using MEME[6] and SeqCode[7] programs to identify relevant motifs. Reversibility of sequence patterns was assumed (*ABCD=DCBA*). The presence of a motif in a given sequence was binary coded.
529 motifs were identified and divided into 4 subsets based on the sequence length.
$len \in \{[8, 37), [37, 44), [44, 60), [60, 70]\}$
For the modelling process 57 motifs were selected based on the frequency of observation.

Association rules

Association rules analysis was performed for the complete feature set. Most of the motifs associated with toxicity contain cysteine and/or *CC* dipeptide.
Most prominent rules of the *toxic* classification are: presence of *CC*, presence of *ILLLL*.

41% of sequences with *CC* are toxic.
3% of all sequences are toxic.

Conversely, several sequence motifs with no observed toxicity were identified: *WYH*, *HWH*, *AELG*, *TYQ*, *KPSV*. Motifs *KR*, *RF* and *FK* are frequently observed in the *nontoxic* class.

Peptide Toxicity Models

logit 1 - a logit model with all relevant variables
logit 2 - a logit model with selected variables and interactions
MLP - a multilayer 61-5-1 perceptron neural network with logistic activation functions
tree 1 - a binary decision tree model
tree 2 - a binary decision tree with 50× increased False Negative decision cost

model	TPR	TNR	FPR	FNR	ACC	MCC
logit 1	0.547	0.996	0.004	0.453	0.876	0.598
logit 2	0.566	0.996	0.004	0.434	0.985	0.672
MLP	0.645	0.995	0.005	0.355	0.986	0.707
tree 1	0.461	0.955	0.005	0.539	0.981	0.560
tree 2	0.623	0.755	0.245	0.377	0.759	0.236

References

[1] A. Bairoch, R. Apweiler, et al., "The universal protein resource (uniprot)", *Nucleic acids research*, vol. 33, no. suppl 1, pp. D154, 2005.

[2] M. Hossain et al., "A review on some alternative specifications of the logit model", *Journal of Business & Economics Research (JBER)*, vol. 7, no. 12, 2011.

[3] D.S. Modha, R. Ananthanarayanan, et al., "Cognitive computing", *Communications of the ACM*, vol. 54, no. 8, pp. 62–71, 2011.

[4] W.Y. Loh, "Classification and regression trees", *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 1, no. 1, pp. 14–23, 2011.

[5] "Peplaser - combinatorial synthesis of peptide arrays with a laser printer".

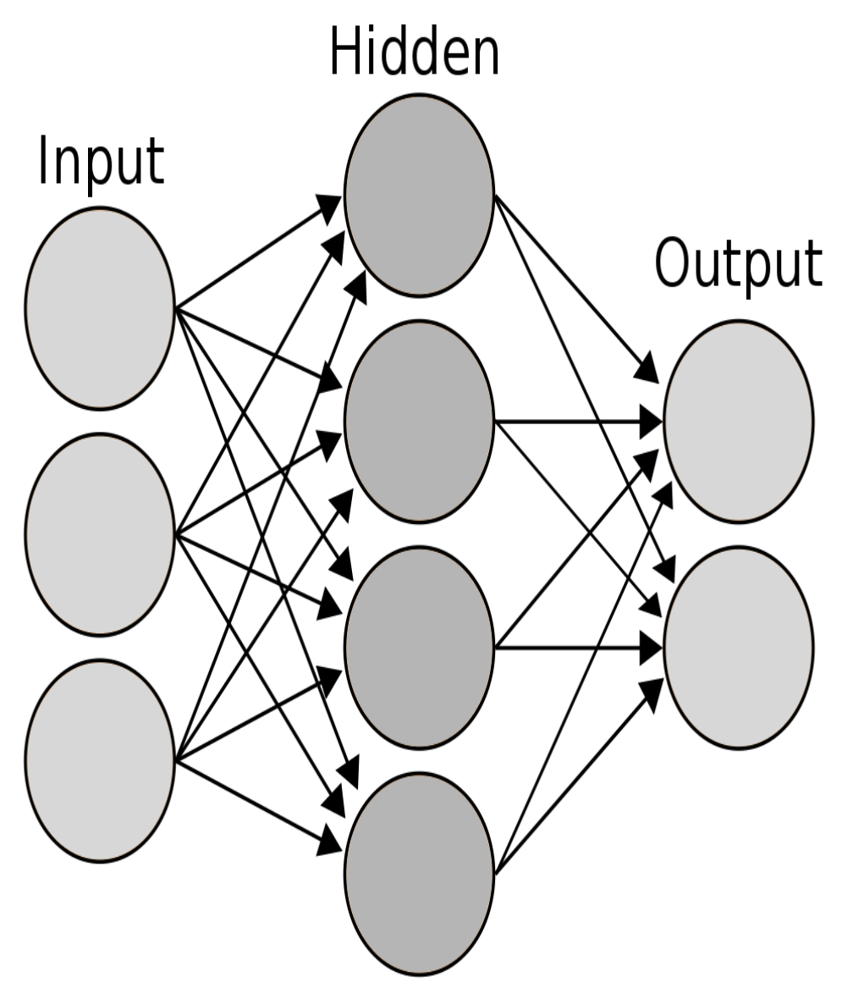
[6] T.L. Bailey and C. Elkan, "Fitting a mixture model by expectation maximization to discover motifs in biopolymers.", in *Proceedings/... International Conference on Intelligent Systems for Molecular Biology*, 1994, vol. 2, p. 28.

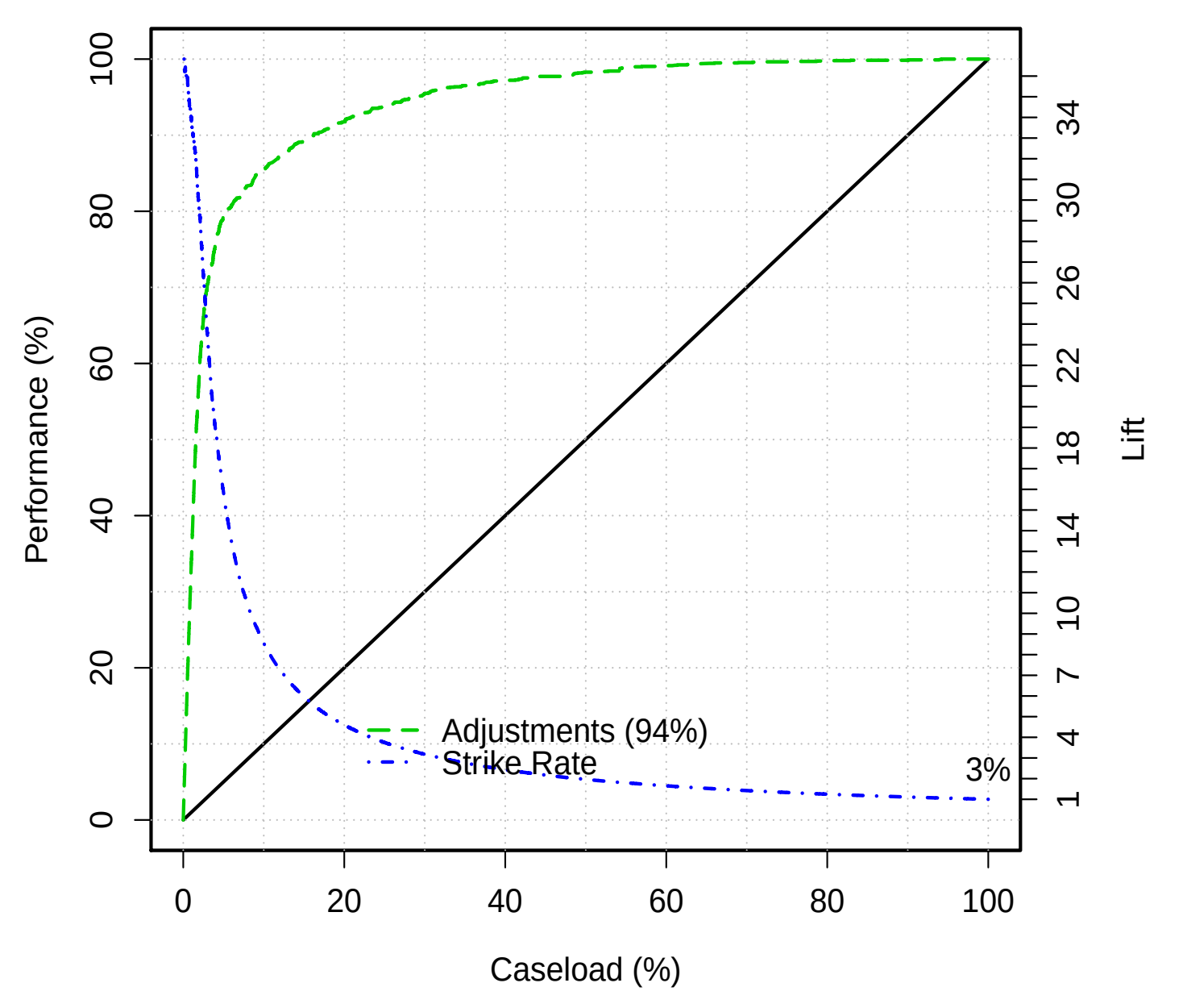
[7] R. Saidi, M. Maddouri, and E.M. Nguifo, "Protein sequences classification by means of feature extraction with substitution matrices", *BMC bioinformatics*, vol. 11, no. 1, pp. 175, 2010.

[8] X.F. Wang and M.S. Cynader, "Pyruvate released by astrocytes protects neurons from copper-catalyzed cysteine neurotoxicity", *J Neurosci.*, vol. 21, no. 10, pp. 3322–31, 2001.

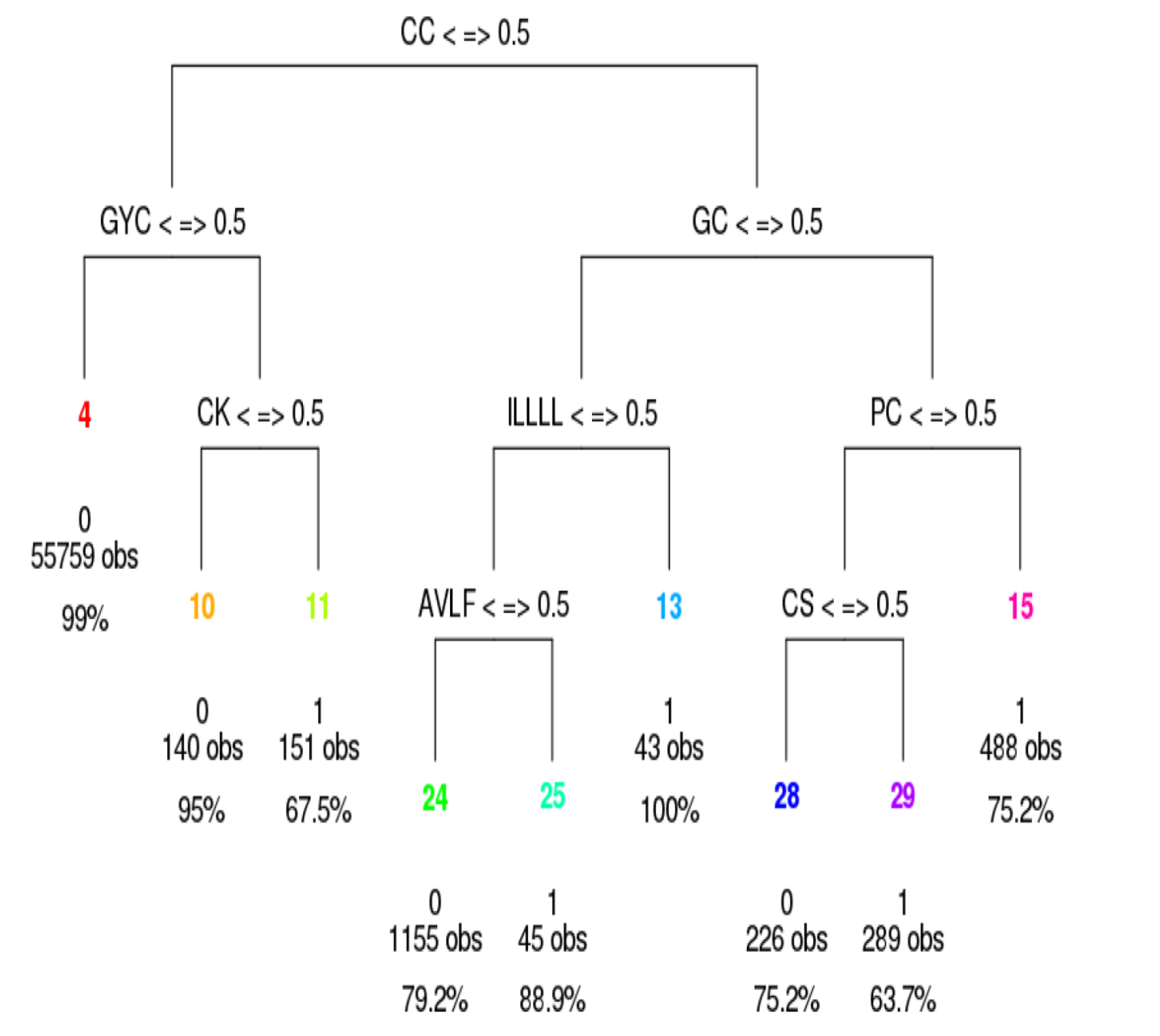
Results

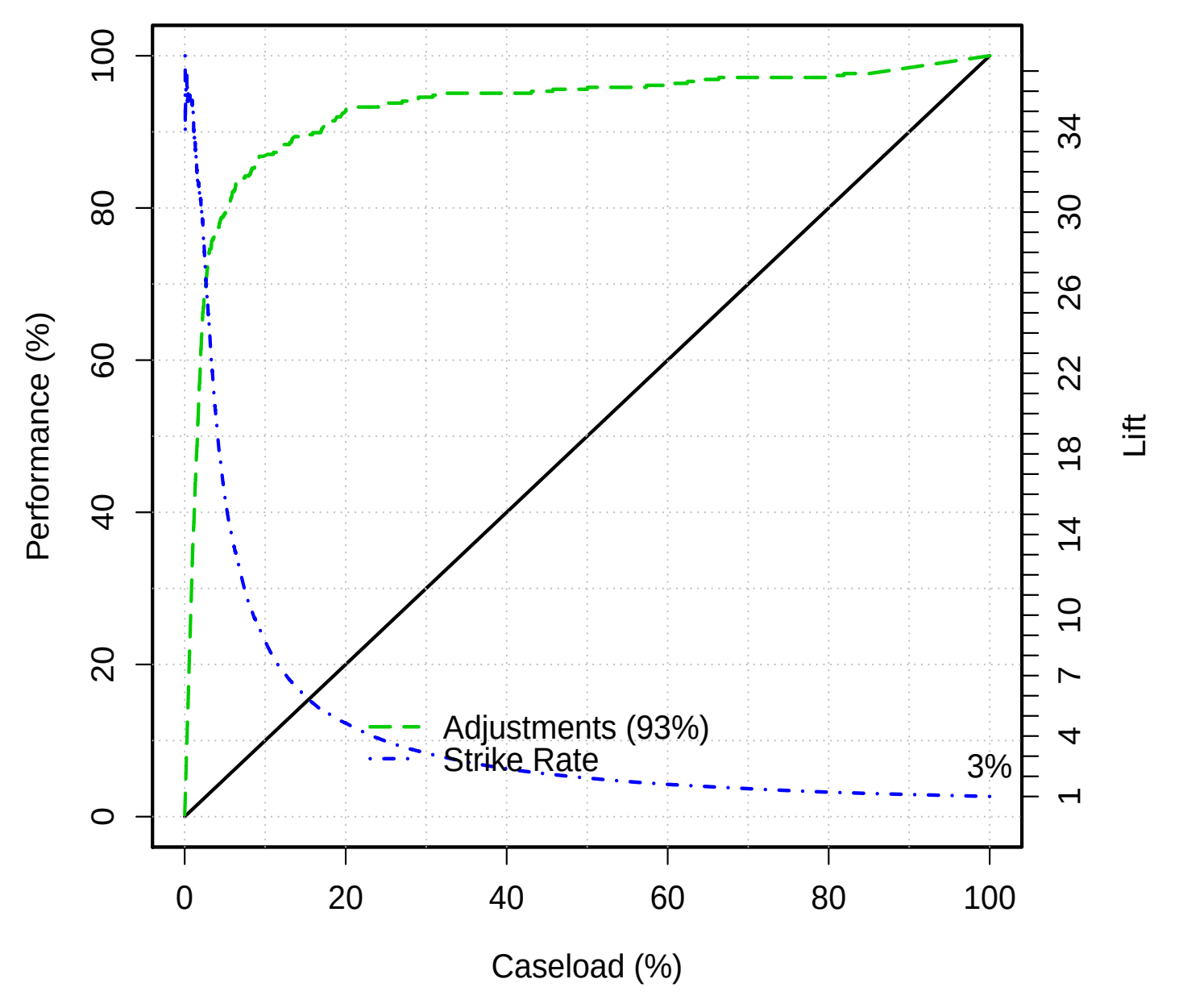
The association rules show that the cysteine-cysteine motif coincides with 15× increased probability of peptide toxicity. This motif is present in 58% of the toxic peptides, while 41% of all peptides containing the *CC* motif are toxic.
The observed toxicity of cysteine containing peptides may result from high neurotoxicity of this aminoacid[8].



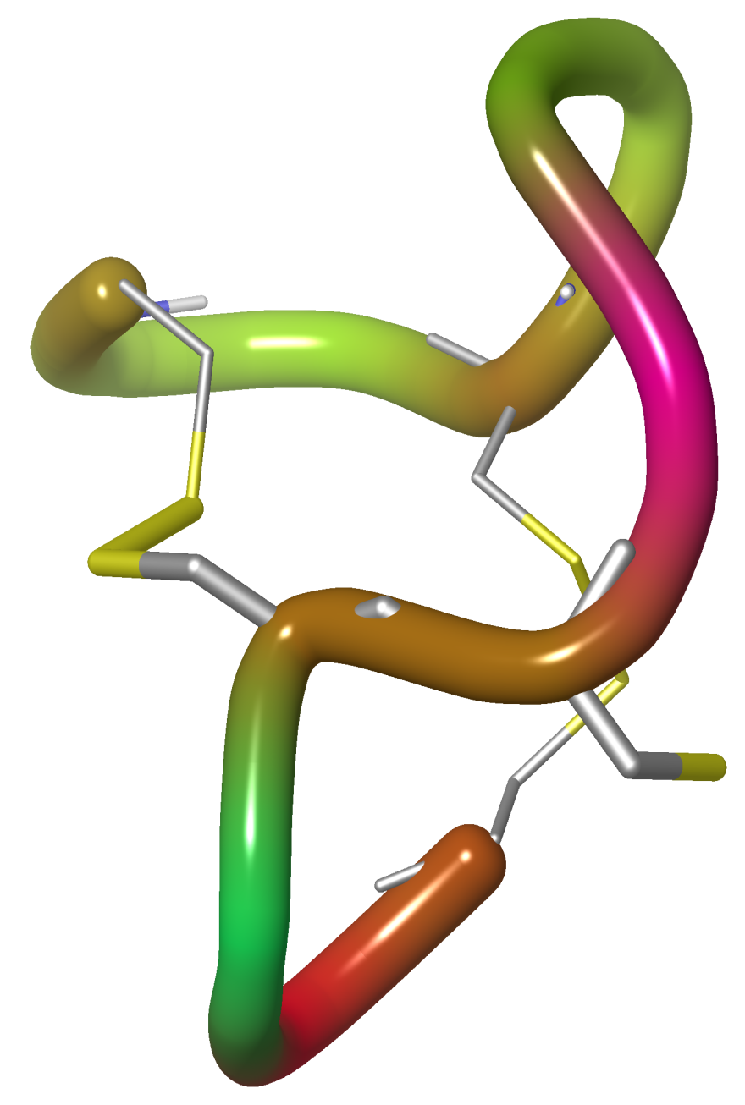


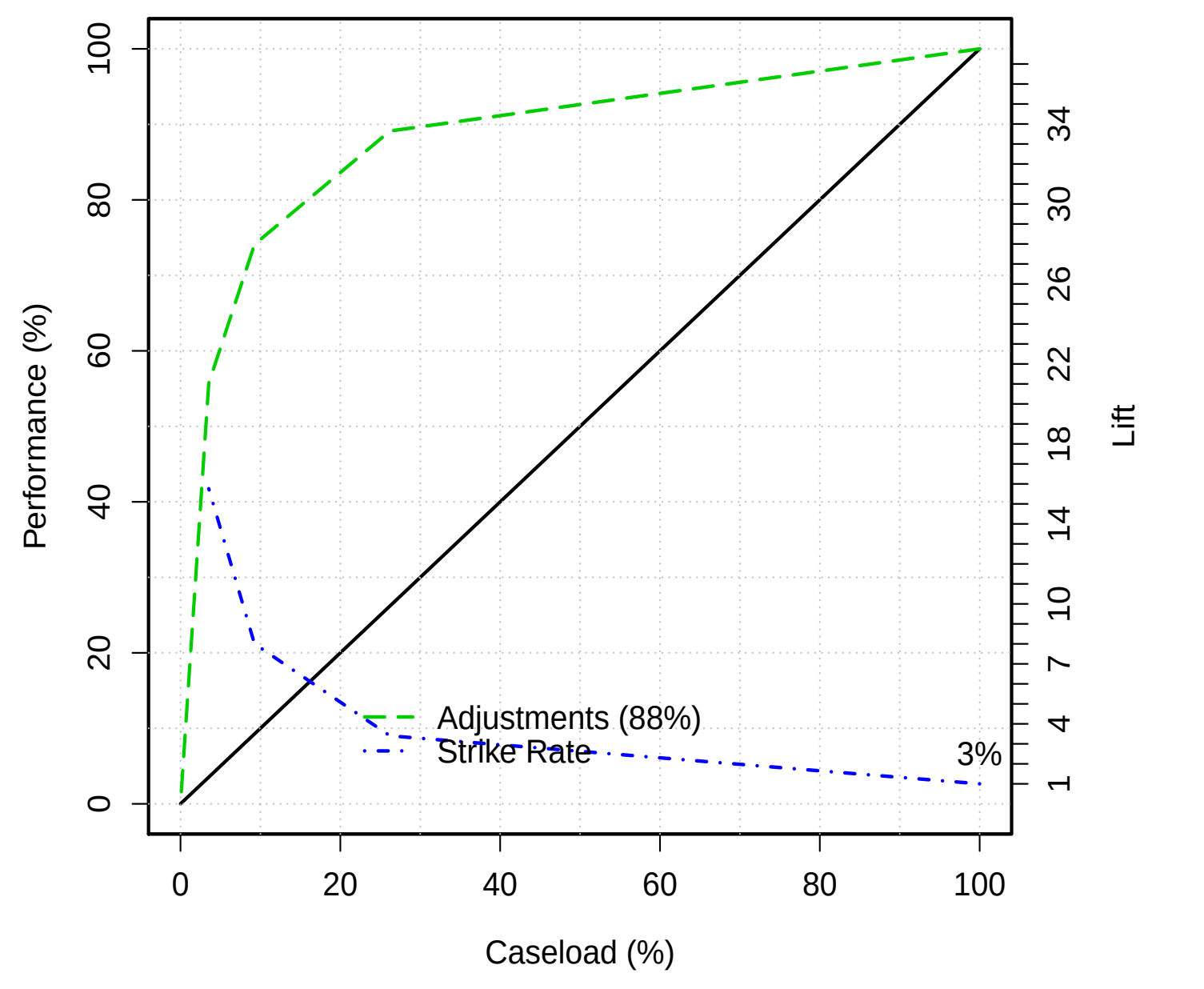
ROC curve for the Logit model with interactions.





ROC curve for the Multilayer Perceptron model with 61 inputs, 5 neurons in the input layer, and 1 in the hidden layer.





ROC curve for the Decision Tree model with 50× increased penalty for classifing a toxic peptide as nontoxic.

The Multilayer Perceptron model gives the most precise results, but it is difficult to explain the underlying biochemical properties that cause toxicity. The Logit model with interactions is the second best model. It classifies 56.6% of toxic peptides correctly with only 0.4% *nontoxic* peptides wrongly classified as *toxic*.

Conclusion

The results show noticeable relation between peptide toxicity, as defined in the UniProt database, and the presence of single and double cysteine motifs. In all the tested models this relation is one of the most relevant classification factors.
Among all the models, the Multilayer Perceptron model is the most precise; however, it is less preferred because of the problems with biochemical interpretation of the results.
Authors suggest the Logit model with interactions as the best compromise between performacne and interpretability.