Predicting Nitrosation of Secondary and Tertiary Amines Using Statistical (Q)SAR Models

Krystle Reiss[1], Roustem Saiakhov[1]

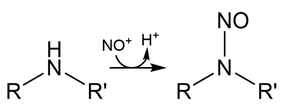
[1]MultiCASE, Inc. | 5885 Landerbrook Dr #210, Mayfield Heights, OH 44124 | USA

GTA 2024

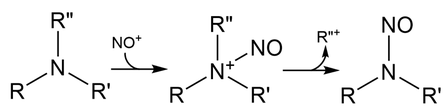
Introduction

Since the discovery of the carcinogenic nitrosamine NDMA in pharmaceuticals such as Valsartan in 2018, there has been a push to not only quantify the presence of N-nitrosamine (NA) impurities in existing APIs but to calculate what risk individual NA molecules pose. Ideally, reliable statistical (Q)SAR models can be created using machine learning algorithms trained on experimental nitrosation results to predict if a particular amine is likely to be nitrosated. However, published results can be limited with widely varying experimental conditions. Here, we limit our training set to amines evaluated with a reaction time of 3-4 hrs., at a pH of 3-4, and in the presence of a surplus of nitrite, which aligns with the WHO's NAP test. This yielded a dataset of 153 secondary and tertiary amine-containing molecules, which were used to train 9 machine learning classifiers, ranging from Nearest Neighbors to Neural Net.

Secondary Amine



Tertiary Amine



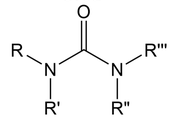
Methods and Data

Of 153 molecules, roughly half contained a secondary amine, while the rest had tertiary only. The dataset is detailed in the table on the right.

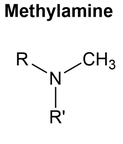
Amines	>0%	≥1%	≥10%
Secondary (77)	60	43	34
Tertiary Only (76)	50	25	10
All (153)	110	68	44

Nine machine learning classifier models from the scikit-learn[1] Python module were trained. (The complete list is given in the legends of the graphs under Results.) Fifteen functional groups were used to generate descriptors to train the models, in addition to pK_a. Most of the descriptors were computed using the RDKit[2] Fragments module. The pK_a of molecules containing secondary amines was predicted using an in-house tool.

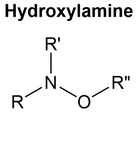
Urea



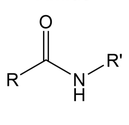
Methylamine



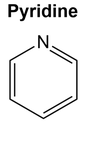
Hydroxylamine



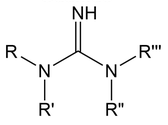
Amide



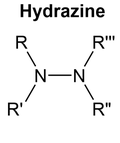
Pyridine



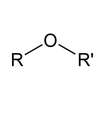
Guanido



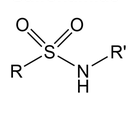
Hydrazine



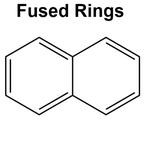
Ether



Sulfonamide



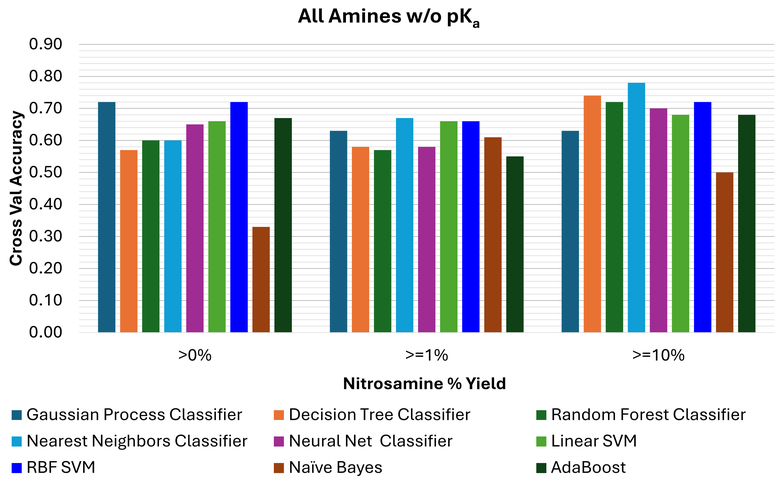
Fused Rings



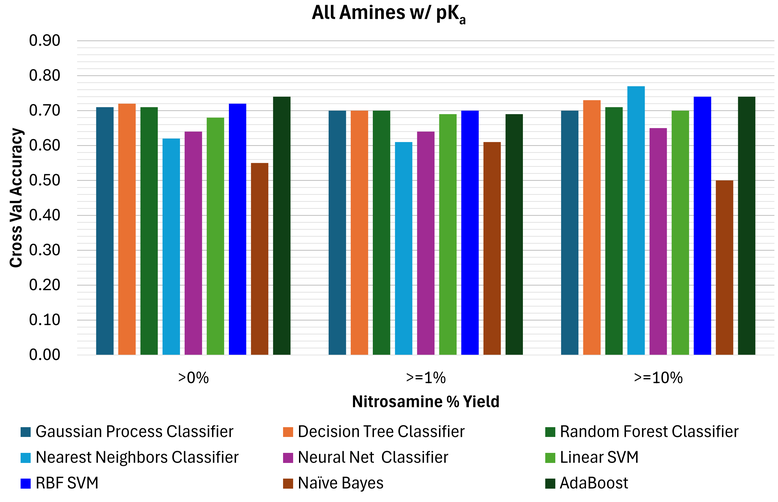
The descriptors were recorded as a count of the number of each functional group present in the molecule. In the case of fused rings, each ring joint was counted. For example, naphthalene (two benzene rings) would be counted as 1 pair of fused rings and anthracene (three benzene rings) would be counted as 2. The nine models were trained on several versions of the data: with and without pK_a as a descriptor, using only secondary or tertiary amines, and with the positive cutoff set to >0%, ≥1%, ≥10%.

Results

All Amines w/o pK_a



All Amines w/ pK_a



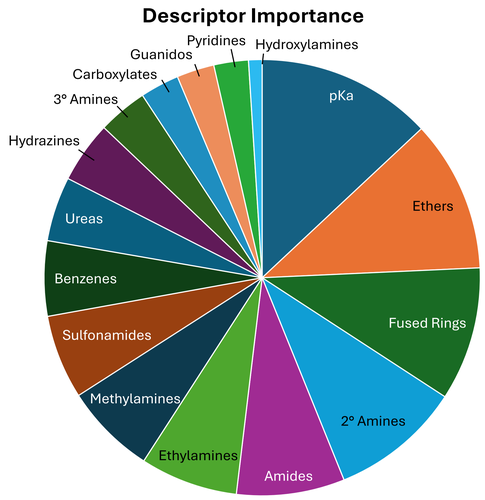
Equation 1-4: (1) Accuracy measures overall correctness. (2) Precision is the correctness of positive calls. (3) Recall is the fraction of possible positive calls that are retrieved. (4) F1 is a balance of Precision and Recall.

(1) **Accuracy** = (*TP* + *TN*) / *All*
(2) **Precision** = *TP* / (*TP* + *FP*)
(3) **Recall** = *TP* / (*TP* + *FN*)
(4) **F1** = 2 / (1/*Pre* + 1/*Rec*)

Classifier Model	Accuracy	Precision	Recall	F1
Gaussian Process	0.71±0.16	0.73±0.17	0.97±0.08	0.83±0.11
Decision Tree	0.72±0.17	0.72±.017	1.00±0.00	0.84±0.09
Random Forest	0.71±0.16	0.72±0.17	0.99±0.04	0.83±0.07
Nearest Neighbors	0.62±0.19	0.71±0.22	0.81±0.24	0.75±0.23
Neural Net	0.64±0.19	0.71±0.16	0.90±0.17	0.79±0.16
Linear SVM	0.68±0.16	0.76±0.19	0.87±.0.16	0.81±0.17
RBF SVM	0.72±0.17	0.72±0.17	1.00±0.00	0.84±0.09
Naïve Bayes	0.55±0.19	0.83±0.36	0.40±0.25	0.54±0.30
AdaBoost	0.74±0.19	0.80±0.20	0.90±0.16	0.85±0.18

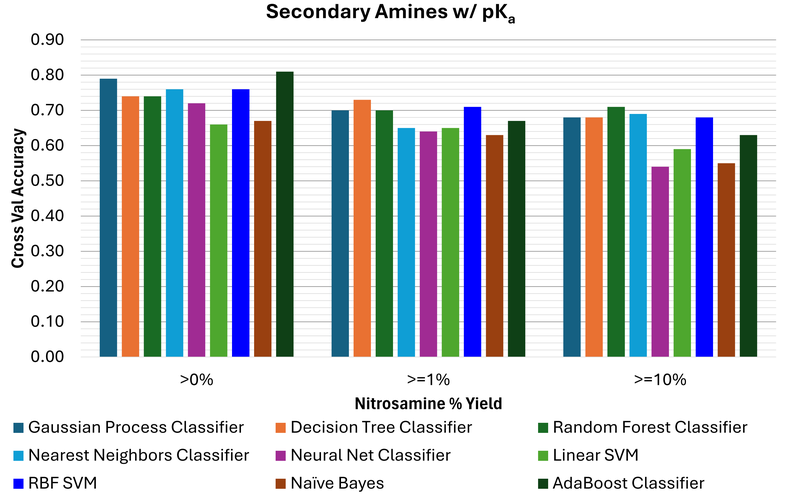
Above: Cross-validated statistics of nine classifier models when the nitrosation threshold is set to >0% and pK_a is included. For validation, data was split into 20 training sets, each having 5 molecules withheld as a test set.

Descriptor Importance



Across almost all models, pK_a was ranked as the most important descriptor, even though it was not available for all amines. Secondary amines were ranked 4th, likely because most (but not all) also had a pK_a given, meaning that a secondary amine was often already indicated by the pK_a. The remaining descriptors formed a continuous series, with no large jumps or drops in importance between them.

Secondary Amines w/ pK_a



Conclusions and Future Work

The inclusion of pK_a as a descriptor improved the accuracy of the models by 0.05 on average, having a greater effect at lower nitrosation thresholds. pK_a had the greatest effect in the Naïve Bayes model, which was overall the worst predictor, but had its accuracy improved by 0.22. No individual model emerged as the clear winner, but Neural Net, Linear SVM, and Naïve Bayes generally performed the worst of the nine.

With the importance of pK_a demonstrated as a critical descriptor for predicting nitrosation of secondary amines, we will be looking to improve the models by generating descriptors for each amine, as opposed to describing the molecule as a whole, as was done here. We will also evaluate related pK_a analogs for tertiary amines, such as pK_b or electrostatic potential.

1. Buitinck et al. *arXiv*. 2013, DOI: arXiv:1309.0238
2. RDKit: Open-source cheminformatics. https://www.rdkit.org
3. Gillatt et al. *Food Chem. Toxicol*. 1984, 22, 269-274.
4. Gillatt et al. *Food Chem. Toxicol*. 1985, 23, 249-255.
5. Takeda and Kanaya. *Cancer Lett*. 1981, 12, 81-86.
6. Schmidtsdorff et al. *Arch. Pharm*. 2022, 355.
7. Sakai et al. *Gan*. 1984, 75, 245-252.
8. Krishna and Rao. *J. Pharm. Sci*.1975, 64, 1579-1581.
9. Lijinsky et al. *Nature*. 1972, 239,165–167.
10. Lijinsky. *Cancer Res*. 1974 ,4, 255-258.
11. Kikugawa et al. *Mutat. Res*. 1987, 177, 35-43.