



Cognigen | DILsym Services | Lixoft

# Building a machine learning model for tautomer prediction

Marvin Waldman  
Simulations Plus, Inc.  
March 23, 2022

# Overview

- Introduction
  - Motivation
  - Goals
- Dataset collection
  - Sources
  - Filters
  - Curation
- Training/Methodology
- Results
- Applications
- Interesting Examples
- Summary
- Acknowledgements

# Motivation

- Many drug molecules exhibit tautomerism
  - Internal estimates with drug-like dataset find ~30% molecules have 2 or more tautomers
- Tautomeric state affects many properties
  - logP
  - solubility
  - permeability
  - activity
- Choice of tautomer affects both QSAR model building and model predictions

# Prior Art

- Several rule-based or scoring methods have been proposed for standardizing the tautomeric form
  - Usually with the intent of producing the likely dominant tautomer
    - Oellien et al., J Chem Inf Model, **46** 2342 (2006)
    - Milletti et al., J Chem Inf Model, **49** 68 (2009)
    - Warr, W.A., J Comput Aided Mol Des, **24** 497 (2010)
    - Sitzmann et al., J Comput Aided Mol Des, **24** 521 (2010)
    - Urbaczek et al., J Chem Inf Model **54** 756
  - Tautomeric preference can sometimes result from a complicated interplay of multiple factors leading to limitations in “simple” rule-based/scoring methods
    - Taylor, P.J.; Kenny, P.W., Figshare (2019), <https://doi.org/10.6084/m9.figshare.8966276.v1>

# Goals

- Develop a machine learning model to predict tautomeric preference
  - Accuracy
  - Speed
- Uses
  - Tautomer standardization
  - Tautomer ranking

# Approach

- Collect a dataset of known tautomeric preferences from literature and public domain sources
- Leverage capabilities of ADMET Predictor<sup>®</sup> and ADMET Modeler<sup>™</sup> to build an Artificial Neural Network Ensemble (ANNE) model based on molecular descriptors
- Augment with special descriptors and modeling methodologies as needed

# Dataset Construction

TautoBase\* + Internal collection

Exclude non-aqueous/gas-phase

Include aq./solid-state/neat liquid

Exclude weakly dominant tautomers

e.g.,  $\log K_t \leq 0.1$

Remove duplicates

Correct Errors (consult lit.)

→ 1529 molecules

201 internal

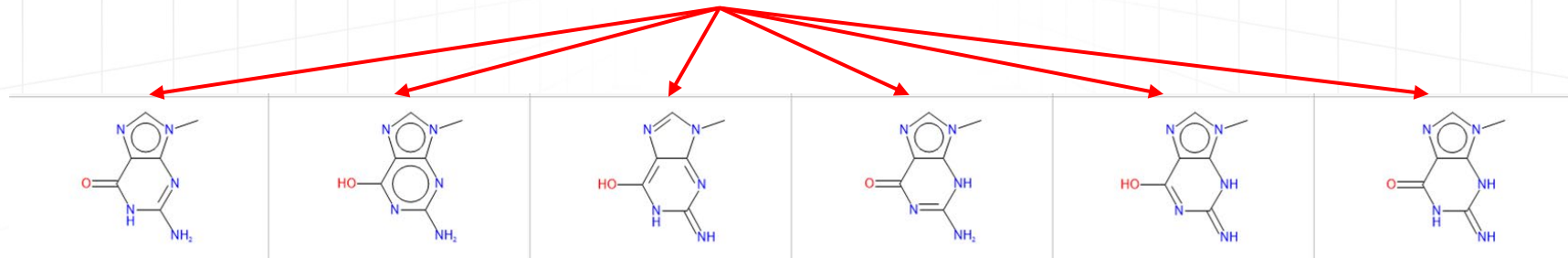
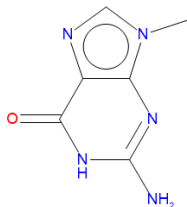
1127 TautoBase

201 Addtl from lit.

\*Wahl, O.; Sander, T., J Chem Inf Model **60** 1085 (2020)

# Dataset Creation

Use of ADMET Predictor to enumerate tautomers for the 1529 molecules led to 7251 tautomers



Preferred 1

0

0

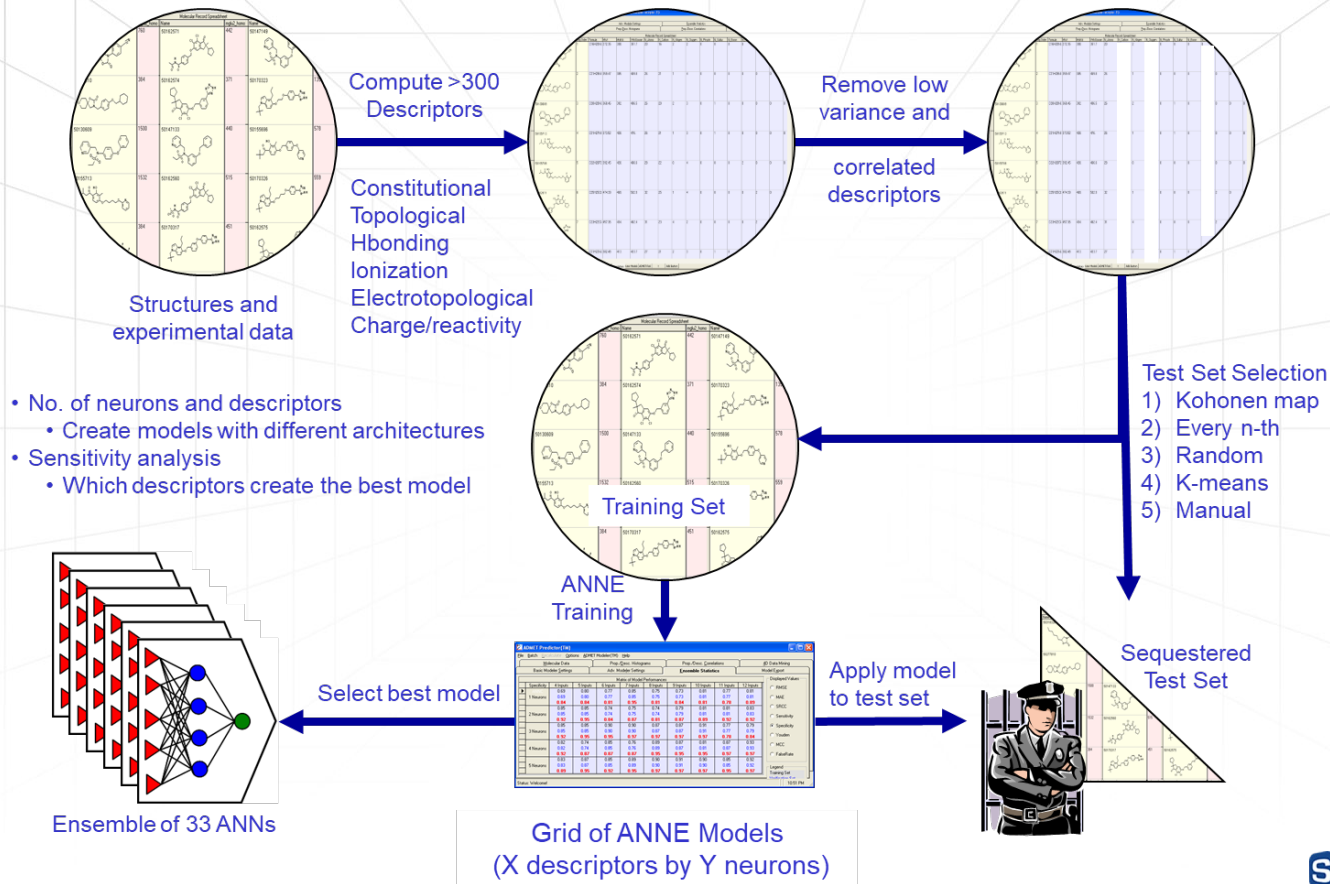
0

0

0



# Standard Model Building Overview



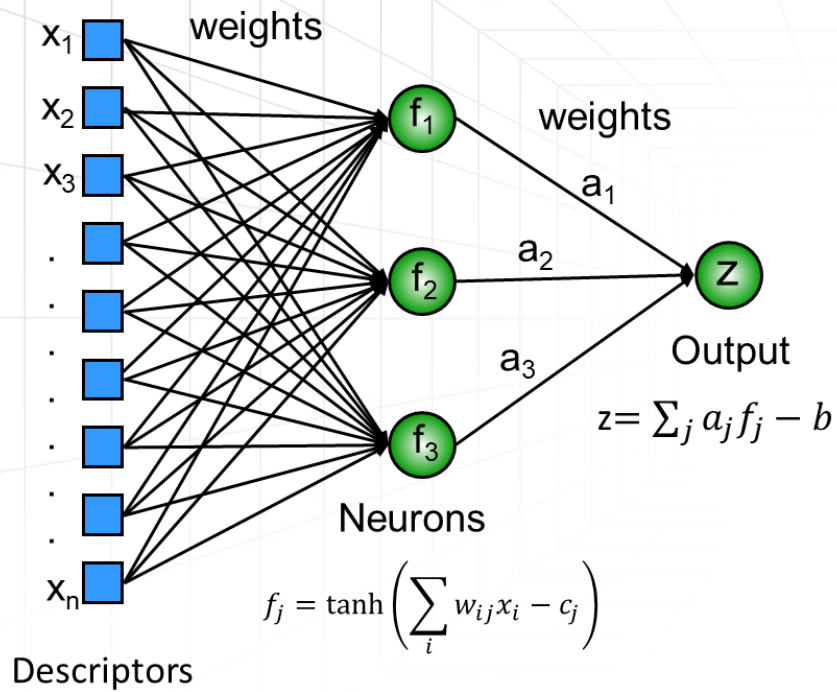
# Specialized Model Building Details

- Descriptors
  - Exclude tautomer-independent descriptors (e.g., N\_Carbons)
    - ➡ ~200 descriptors
    - Add/augment descriptors important for tautomer preference\*
      - Anti-aromatic rings
      - dipole/dipole and lone-pair repulsions
      - Extensions to internal Hydrogen bonds
- Train/Test set partition
  - All tautomers of a given molecule assigned to train or test set exclusively
  - Partition train/test set using dominant tautomers only and then assign all tautomers of a given molecule to the same partition

\*Taylor, P.J.; Kenny, P.W., Figshare (2019), <https://doi.org/10.6084/m9.figshare.8966276.v1>

# Training

- Neural Network Architecture



## Softmax Transformation

$$\sigma(z_k) = \frac{e^{z_k}}{\sum_{m=1}^M e^{z_m}}$$

Sum in denominator is over all tautomers of a given molecule  
 $z$  varies from  $-\infty$  to  $+\infty$   
 $\sigma$  varies from 0 to 1

# Training (cont'd)

## “Modified” Cross-entropy Loss

$$\mathcal{L} = - \sum_{n=1}^N \log(\sigma_n)$$

Sum is over preferred tautomers only.

Reasons:

- Labels are not independent.
  - Only one tautomer of a collection can be marked as “Preferred”.
- Reduces effect of imbalance in the data.
- Non-dominant tautomers are not a “hard” 0

Minimize  $\mathcal{L}$  with respect to weights and bias terms

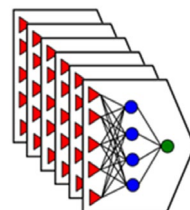
# Model Building

Grid of network architectures:  
neurons x descriptors

For each architecture:  
Train 165 networks  
Select best 33 (ensemble)  
Average the 33 scores  
Highest scoring tautomer is “Preferred”

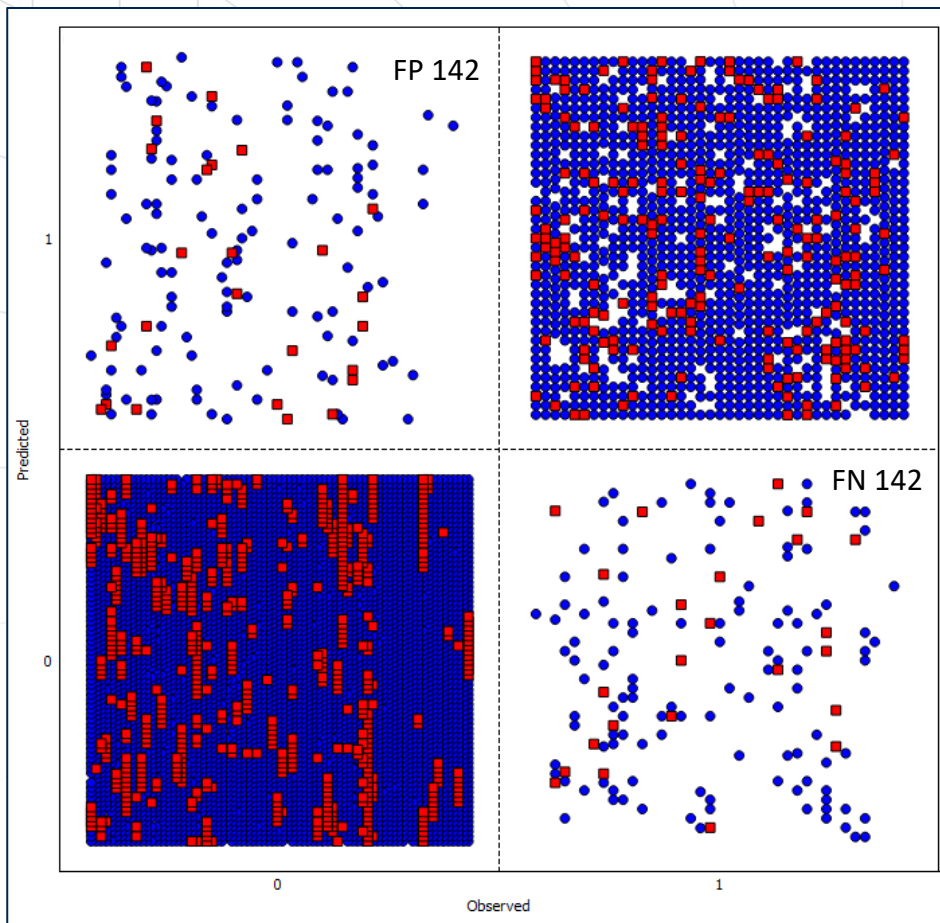
Youden	55 Inputs	60 Inputs	65 Inputs	70 Inputs	75 Inputs	80 Inputs	85 Inputs	90 Inputs	95 Inputs	100 Inputs
2 Neurons	0.79	0.79	0.79	0.79	0.81	0.81	0.81	0.82	0.81	0.81
	-	-	-	-	-	-	-	-	-	-
4 Neurons	0.80	0.80	0.81	0.81	0.81	0.81	0.82	0.82	0.83	0.81
	-	-	-	-	-	-	-	-	-	-
6 Neurons	0.80	0.81	0.82	0.82	0.83	0.83	0.83	0.84	0.82	0.83
	-	-	-	-	-	-	-	-	-	-
8 Neurons	0.80	0.82	0.83	0.82	0.83	0.83	0.82	0.83	0.83	0.83
	-	-	-	-	-	-	-	-	-	-
10 Neurons	0.82	0.83	0.84	0.83	0.84	0.83	0.84	0.84	0.85	0.84
	-	-	-	-	-	-	-	-	-	-

Select best ensemble model  
Fewest false negatives/positives  
for train and test sets



Ensemble of 33 ANNs

# Results

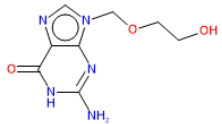
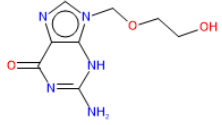
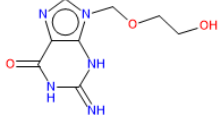
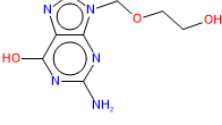
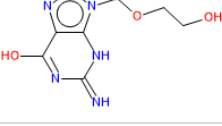
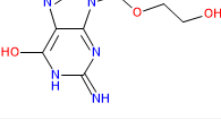


● Train

■ Test

FN=FP because only 1 tautomer in a collection is marked/predicted as Preferred

# Applications: Ranking/Scoring

Structure	Identifier	Tautomer_Score
	Aciclovir	0.792
	Aciclovir - T1	0.609
	Aciclovir - T2	0.242
	Aciclovir - T3	0.527
	Aciclovir - T4	0.166
	Aciclovir - T5	0.157

0.792

0.609

0.242

0.527

0.166

0.157



# Applications: Tautomer Standardization

**Tautomer settings**

**Tautomer enumeration method**

☐ Use legacy algorithm

**Tautomer standardization method**

☐ Rule based

☒ Model based

☐ Use standardization queries

Allows for user customization

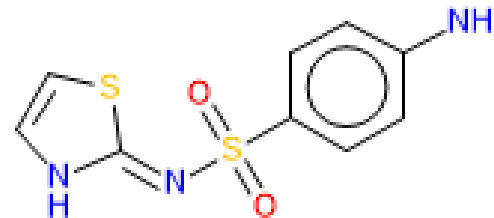
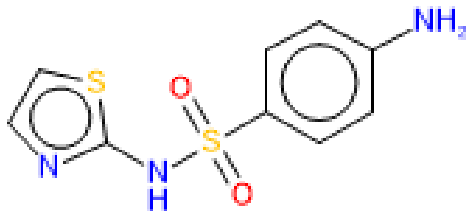
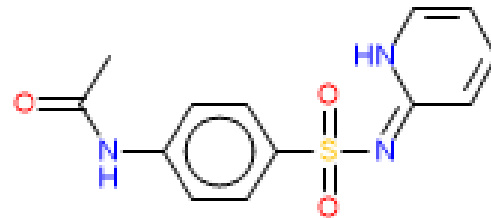
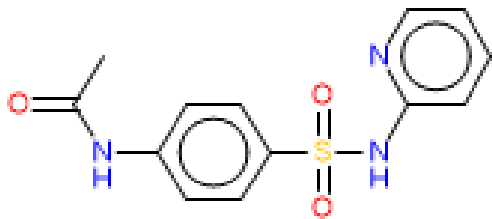
Legacy	Method	Queries	Incorrect #Pref = 1529
On	Rule	On	318
On	Rule	Off	363
Off	Rule	On	355
Off	Rule	Off	397
On	Model	On	141
On	Model	Off	119
Off	Model	On	145
Off	Model	Off	123

Model based  
~5 seconds  
8 core i-7 2.6 GHz



# Some Interesting Examples

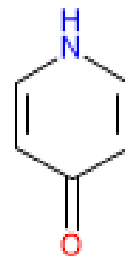
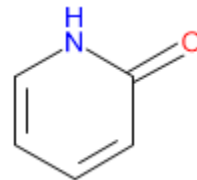
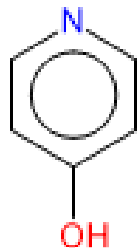
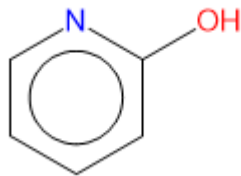
## Misses:



Preferred (model and rules)

Observed

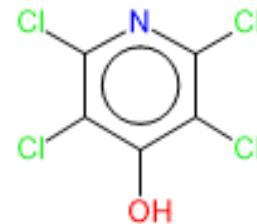
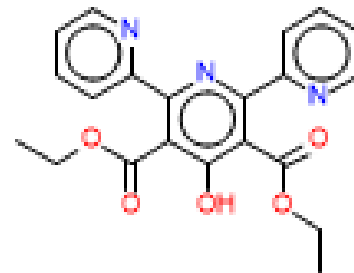
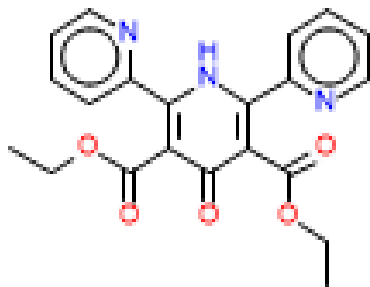
# Pyridone vs Hydroxypyridine



Preferred (usually)

# Pyridone vs Hydroxypyridine

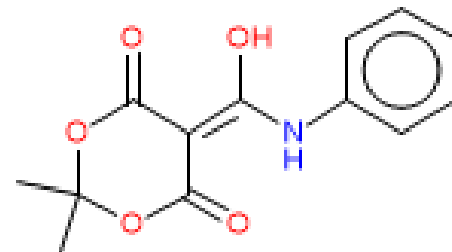
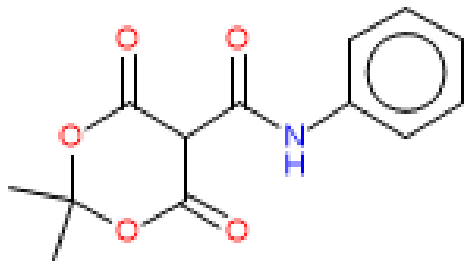
## Except when ...



Preferred by rules

Preferred by model  
Observed

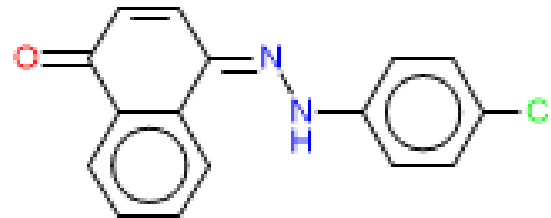
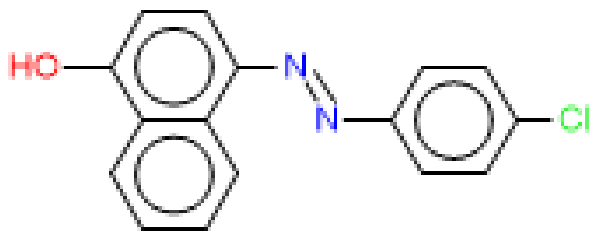
# Amide preferred over Enolamine Except when ...



Preferred by rules

Preferred by model  
Observed

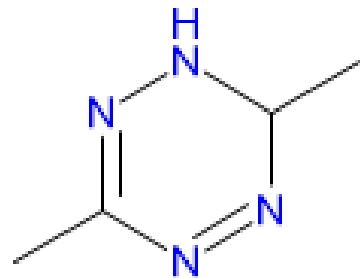
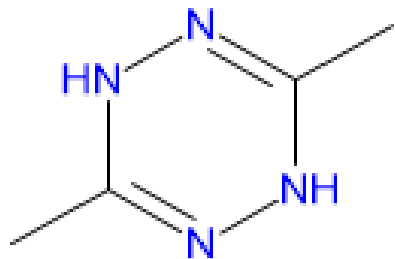
# Another Example



Preferred by rules

Preferred by model  
Observed

# Sometimes the rules win

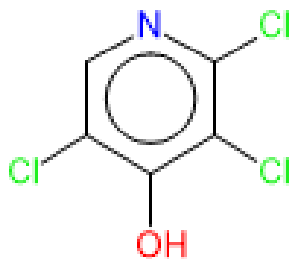


Preferred by model  
(even though anti-aromatic)

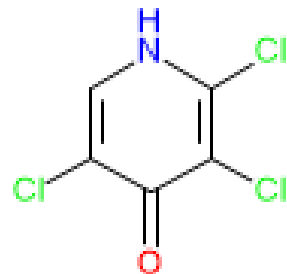
Preferred by rules  
Observed

# Sometimes the rules win

## Pyridone/Hydroxypyridine example



Preferred by model



Preferred by rules  
Observed

# Possible Future Directions

- Improved/customized descriptors
  - OO vs. NN lone pairs, OH...N vs C=O...HN H-bonds
- Deep learning/multi-layer networks
- More data



# Summary

- Machine Learning ANNE model for predicting tautomer preference has been built from a collection of literature data of ~1500 examples
- Model outperforms our rule-based method by better than a factor of 2 (based on no. incorrect)
- Can be used to standardize or rank tautomers for QSAR model building and other cheminformatics applications

# Acknowledgements

- David Miller
- Robert Fraczekiewicz
- Bob Clark
- Michael Lawless