Modeling tautomer preference

Marvin Waldman

Simulations Plus, Inc.

marvin.waldman@simulations-plus.com

SimulationsPlus

MIDD+22

Model Informed Drug Development

INTRODUCTION

Many drug molecules exhibit tautomerism (internal estimates find ~30% of drug-like molecules are tautomeric). The tautomeric state of a molecule determines many of its properties such as lipophilicity, solubility, permeability, binding, toxicity, etc. In addition, choice of the tautomeric state affects the results of most QSAR/machine learning models for property prediction. Consequently, several rule-based or scoring methods have been developed[1-6] for predicting the preferred or dominant tautomer of a molecule, but they have limitations as often the preferred tautomer results from a complicated interplay of multiple factors.[7]

OBJECTIVE

To build an Artificial Neural Network Ensemble (ANNE) machine learning model that predicts the preferred tautomer from a list of candidate tautomers by leveraging our ADMET Predictor® and ADMET Modeler™ methodologies.

Performance Goals:

- Accuracy
- Speed

Uses:

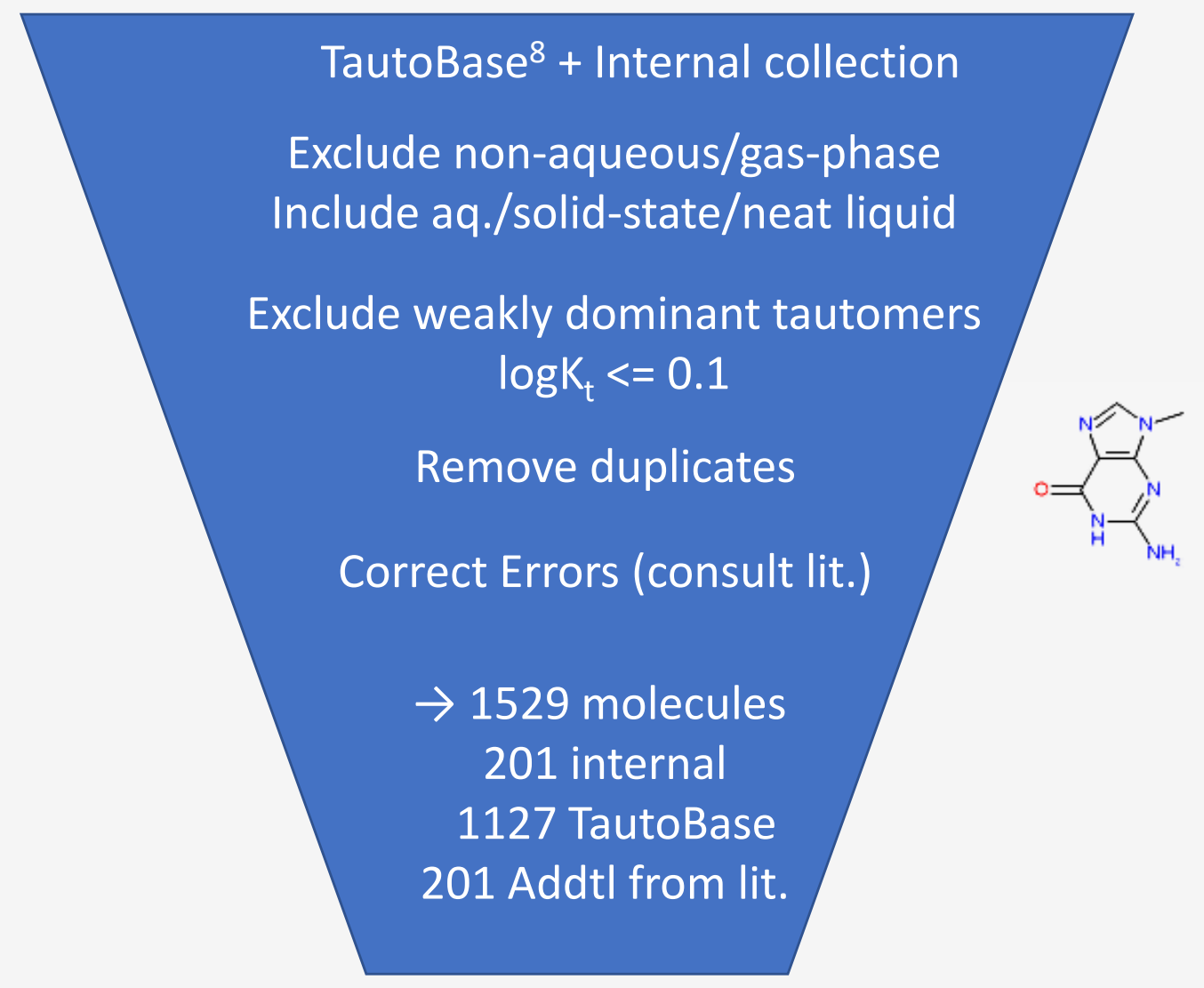
- Tautomer Standardization
- Rank candidate tautomers

CONCLUSIONS

We have built a machine learning model capable of accurately selecting the preferred or dominant tautomer among a series of candidate tautomers. Its accuracy outperforms our rule-based approach by better than a factor of 2 on over 1500 examples. The model may be used to standardize tautomers for building QSAR models and other cheminformatics applications.

METHODS

Filtering/Curation



TautoBase[8] + Internal collection

Exclude non-aqueous/gas-phase
Include aq./solid-state/neat liquid

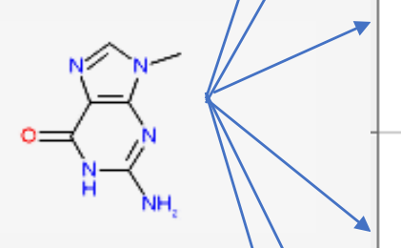
Exclude weakly dominant tautomers
$\log K_t <= 0.1$

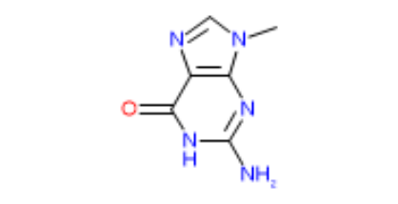
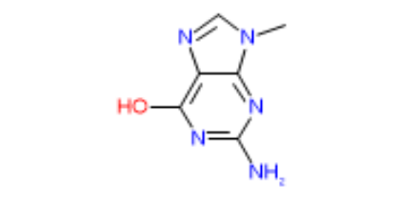
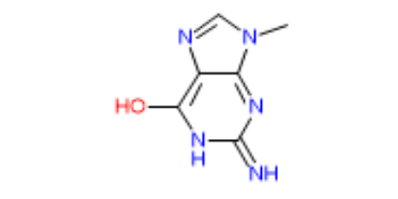
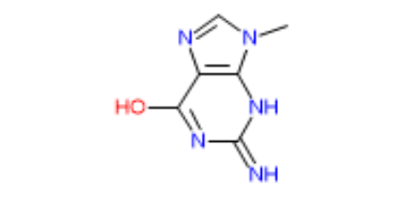
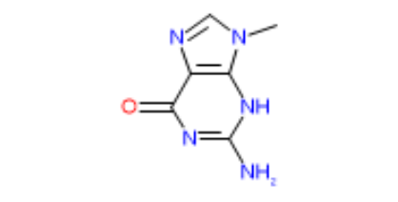
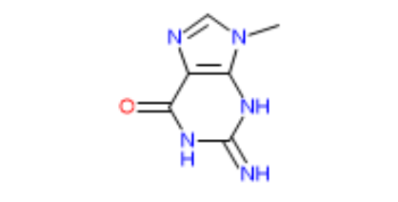
Remove duplicates

Correct Errors (consult lit.)

→ 1529 molecules
201 internal
1127 TautoBase
201 Addtl from lit.

Enumerate Tautomers using ADMET Predictor



Structure	Preferred
	1
	0
	0
	0
	0
	0

7251 Tautomers

Calculate Descriptors:
~200 Tautomer-dependent descriptors:
E-states
Functional Groups
Aromatic/Aliphatic/Anti-aromatic rings
dipole-dipole/lone pair repulsions
Partial Charges
Internal H-bonds

Filter/Rank Descriptors:
Low variance
Correlated
Sensitivity Analysis

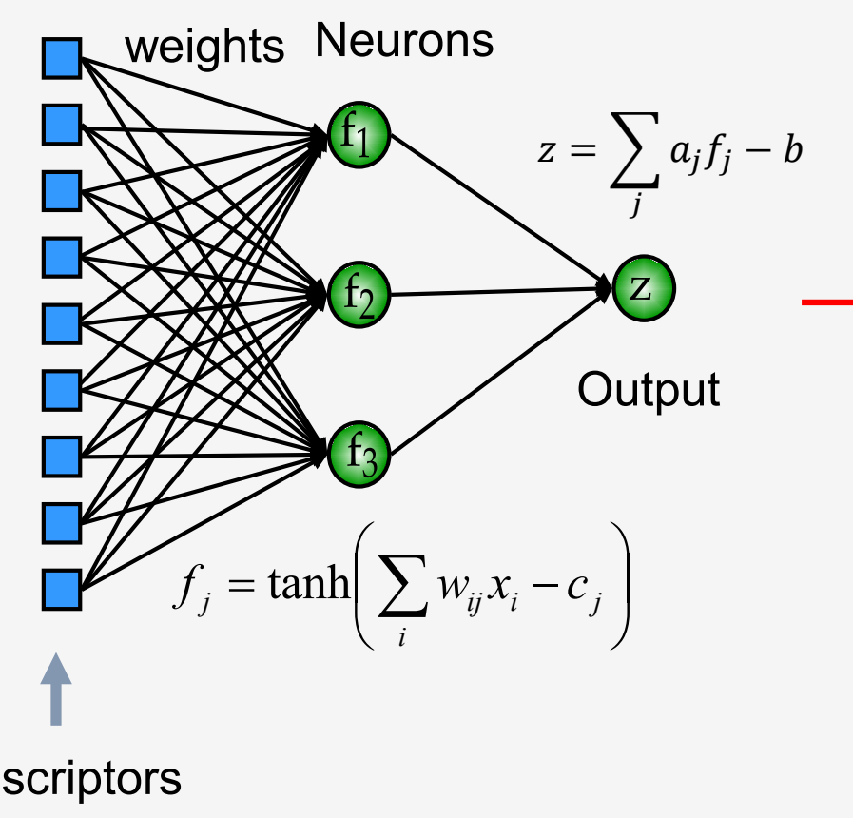
Partition Data into Training/Test Sets:
Train/Test sets partitioned per molecule, not per tautomer
All tautomers of a given molecule belong exclusively to train or test set.
Use Kohonen map to perform initial partitioning for dominant tautomers only. Then, add all tautomers of each test set molecule to the test set.

Train:
Grid of network architectures:
neurons x descriptors

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 - 0.84	0.79 - 0.83	0.79 - 0.84	0.79 - 0.83	0.81 - 0.82	0.81 - 0.83	0.81 - 0.85	0.82 - 0.84	0.81 - 0.84	0.81 - 0.83
4 Neurons	0.80 - 0.83	0.80 - 0.84	0.81 - 0.85	0.81 - 0.87	0.81 - 0.87	0.81 - 0.87	0.82 - 0.86	0.82 - 0.86	0.83 - 0.87	0.81 - 0.86
6 Neurons	0.80 - 0.85	0.81 - 0.87	0.82 - 0.88	0.82 - 0.86	0.83 - 0.86	0.83 - 0.86	0.83 - 0.86	0.84 - 0.87	0.82 - 0.87	0.83 - 0.86
8 Neurons	0.80 - 0.86	0.82 - 0.87	0.83 - 0.86	0.82 - 0.86	0.83 - 0.86	0.83 - 0.85	0.82 - 0.87	0.83 - 0.86	0.83 - 0.85	0.83 - 0.86
10 Neurons	0.82 - 0.87	0.83 - 0.87	0.84 - 0.87	0.83 - 0.87	0.84 - 0.87	0.83 - 0.87	0.84 - 0.89	0.84 - 0.87	0.85 - 0.88	0.84 - 0.86

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

Training:



Artificial Neural Network (ANN)

Softmax Function

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

Sum in denominator is over all tautomers of a given molecule
z varies from –∞ to +∞
σ varies from 0 to 1

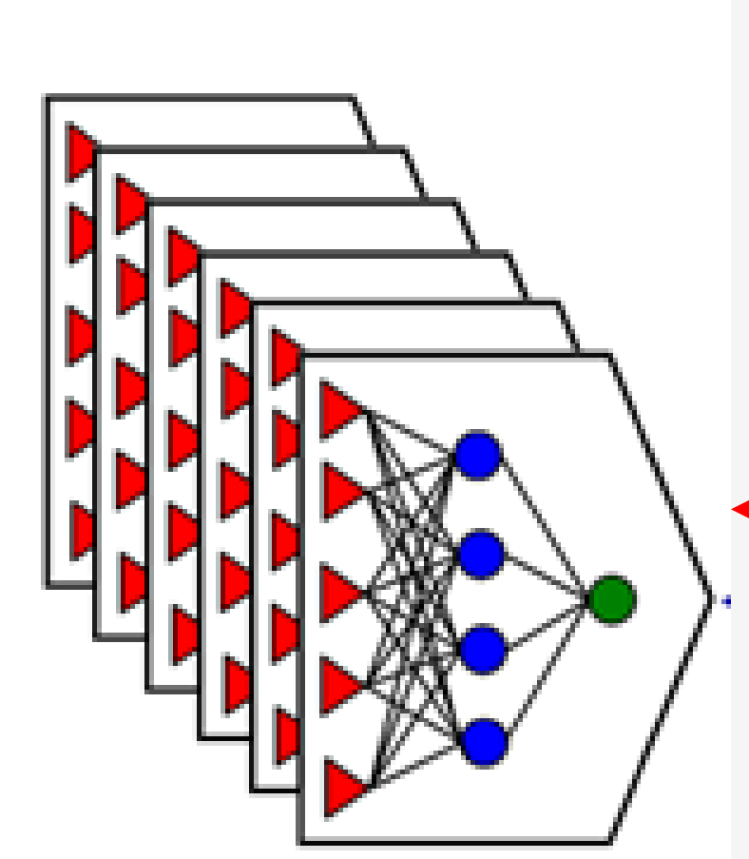
"Modified" Cross-entropy Loss

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

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

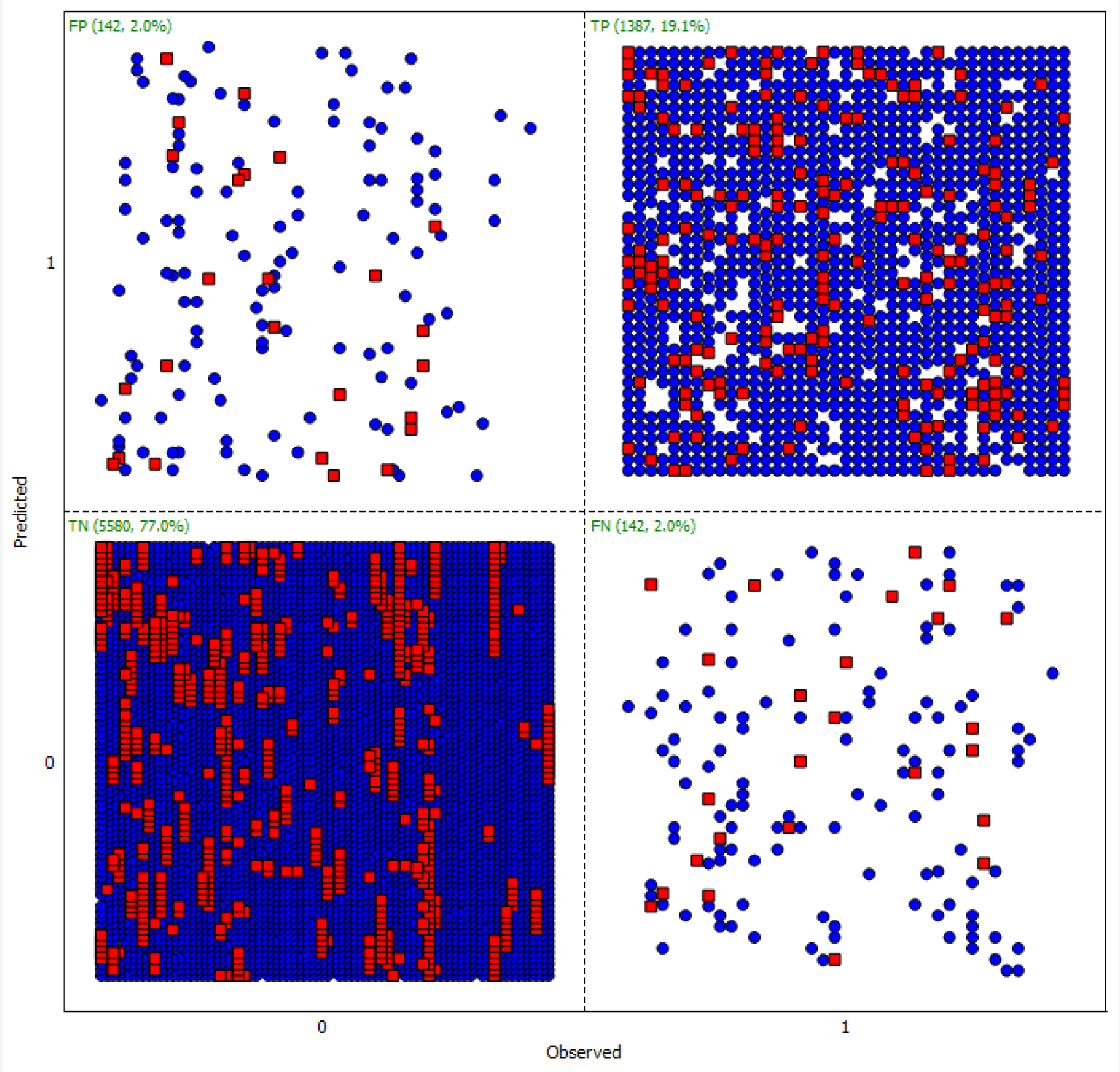
Train 165 networks (adjust weights to minimize $\mathcal{L}$).
Select best 33 (ANN ensemble) for each architecture.
Average scores over the 33 networks.
Highest scoring tautomer is classified as "Preferred".



Ensemble of 33 ANNs

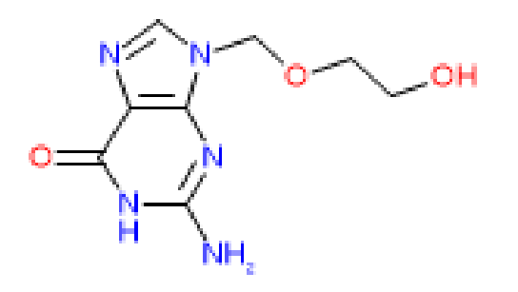
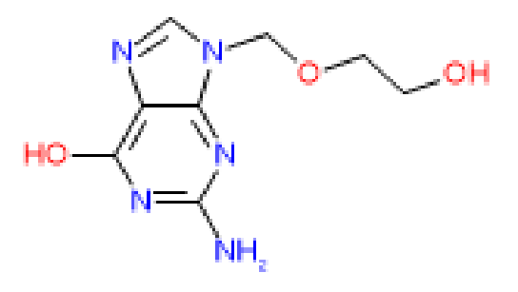
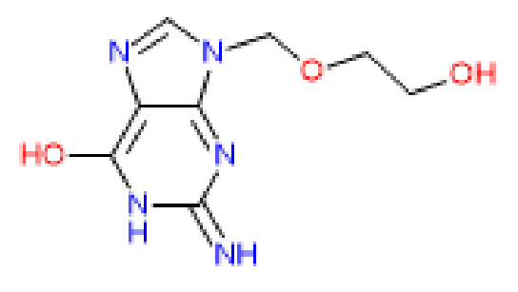
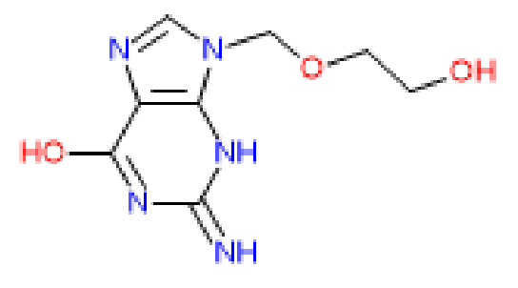
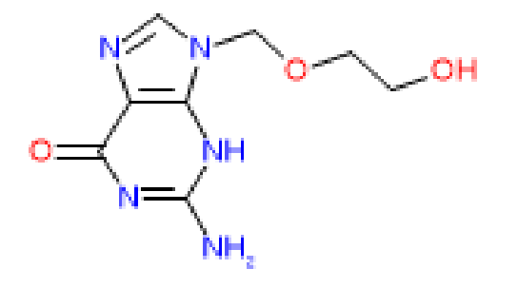
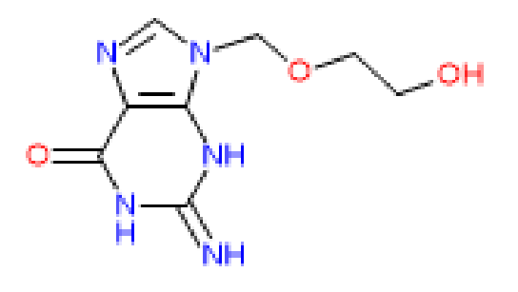
RESULTS

Model Performance

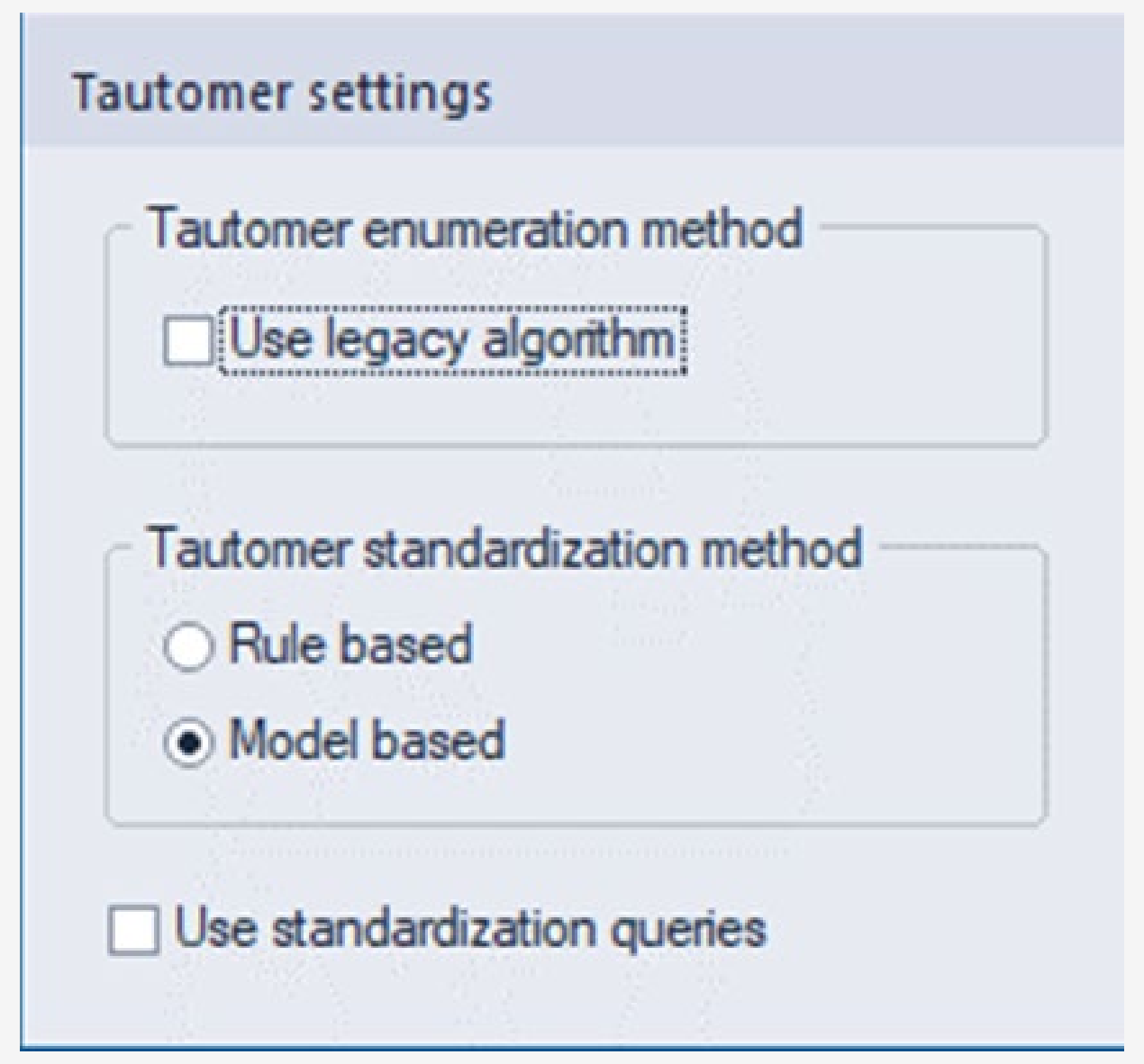


ALL: Sensitivity=0.907 Specificity=0.975 MCC=0.882 False Rate=0.039
TRAIN: Sensitivity=0.910 Specificity=0.977 MCC=0.887 False Rate=0.037
TEST: Sensitivity=0.893 Specificity=0.963 MCC=0.856 False Rate=0.055

Ranking Tautomers

Structure	Identifier	Tautomer_Score
	Aciclovir	0.995
	Aciclovir - T1	0.607
	Aciclovir - T2	0.001
	Aciclovir - T3	0.002
	Aciclovir - T4	0.856
	Aciclovir - T5	0.010

Tautomer Standardization



Tautomer settings

Tautomer enumeration method

☐ Use legacy algorithm

Tautomer standardization method

○ Rule based
● Model based

☐ Use standardization queries

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

REFERENCES

1. Oellien et al., J Chem Inf Model, **46** 2342 (2006)
2. Milletti et al., J Chem Inf Model, **49** 68 (2009)
3. Warr, W.A., J Comput Aided Mol Des, **24** 497 (2010)
4. Sitzmann et al., J Comput Aided Mol Des, **24** 521 (2010)
5. Urbaczek et al., J Chem Inf Model **54** 756 (2014)
6. ADMET Predictor®, Simulations Plus, Inc.
7. Taylor, P.J.; Kenny, P.W., Figshare (2019), https://doi.org/10.6084/m9.figshare.8966276.v1
8. Wahl, O.; Sander, T., J Chem Inf Model **60** 1085 (2020)

SimulationsPlus
Cognigen | DILIsym Services | Lixoft