W207 Applied Machine Learning
(Summer 2021)

Final Project Presentation:
Tox21 Structure-Activity Relationship Models

Tony Angell, Elaine Chang, Lea Cleary

Berkeley
UNIVERSITY OF CALIFORNIA

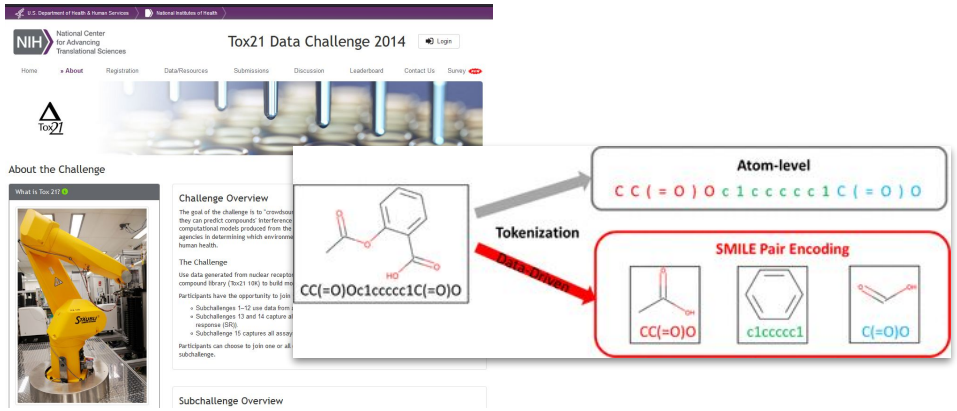
Tox21 Structure-Activity Relationship Models

- Federal collaboration between the NIH, EPA, and FDA
- Data challenge to "crowdsource" data analysis
- Potential to disrupt biological pathways leading to drug discovery,

Problem Statement

Prediction of compound toxicity using only chemical structure data

Dataset



Project Pipeline Overview

1 Data Wrangling

- Explore and select dataset for project scope
- Remove duplicates and invalid compounds
- Standardize SMILES representations for compounds using MolVS

2 Feature Generation

- Tokenize SMILES strings using different approaches
- Generate a matrix of feature vectors (vectorize)

Approaches: atom-wise, k-mers, SMILES pair encoding, ECFP, molecular descriptors

3 Model Matrix Experiment

- Selected primary performance metric (ROC-AUC vs. Recall)
- Used GridSearchcv to iterate and identify best performing model

Approaches: Similarity-based (KNN, SVM); Features-based (NB, RF, logistic regression)

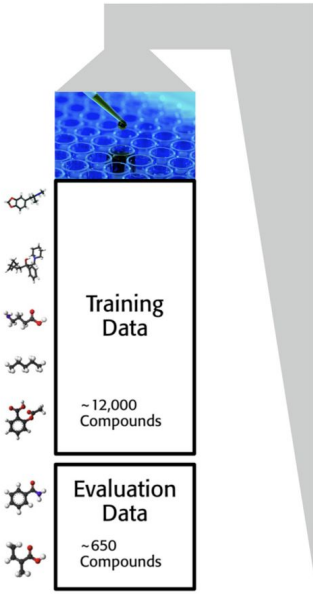
4 Model Oversample

- Create new oversampled dataset
- Compare results to random oversampling

5 Model Exploration

- Explore differences in model performance
- Explore alternative and automated approaches to modeling

Data Selection: NR-AhR

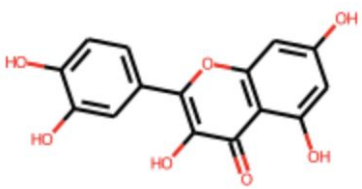


	Train			Test			Score		
	Active	Inactive	Ratio I:A	Active	Inactive	Ratio I:A	Active	Inactive	Ratio I:A
NR-AR	380	8982	24	3	289	96	12	574	48
NR-AR-LBD	303	8296	27	4	249	62	8	574	72
NR-ER	937	6760	7	27	238	9	51	465	9
NR-ER-LBD	446	8307	19	10	277	28	20	580	29
NR-AhR	950	7219	8	31	241	8	73	537	7
NR-Aromatase	360	6866	19	18	196	11	39	49	1
NR-PPAR	222	7962	36	15	252	17	31	571	18
SR-ARE	1098	6069	6	48	186	4	93	462	5
SR-ATAD5	338	8753	26	25	247	10	38	584	15
SR-HSE	248	7722	31	10	257	26	22	588	27
SR-MMP	1142	6178	5	38	200	5	60	483	8
SR-p53	537	8097	15	28	241	9	41	575	14

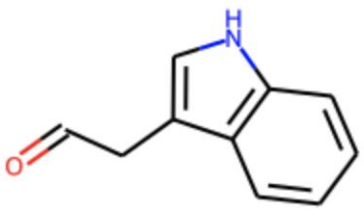
Aryl Hydrocarbon Receptor

- Ligand-activated protein that functions primarily as a sensor of xenobiotic chemicals (e.g. natural plant flavonoids, indoles, synthetic polycyclic aromatic hydrocarbons, dioxin-like compounds)
- Also regulates enzymes such as cytochrome P450s that metabolize these chemicals

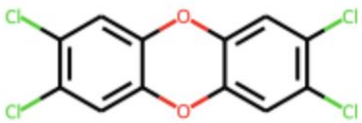
Examples of known ligands:



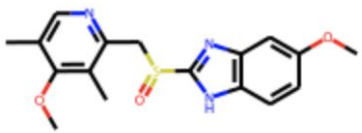
Quercetin



Indole-3-acetaldehyde

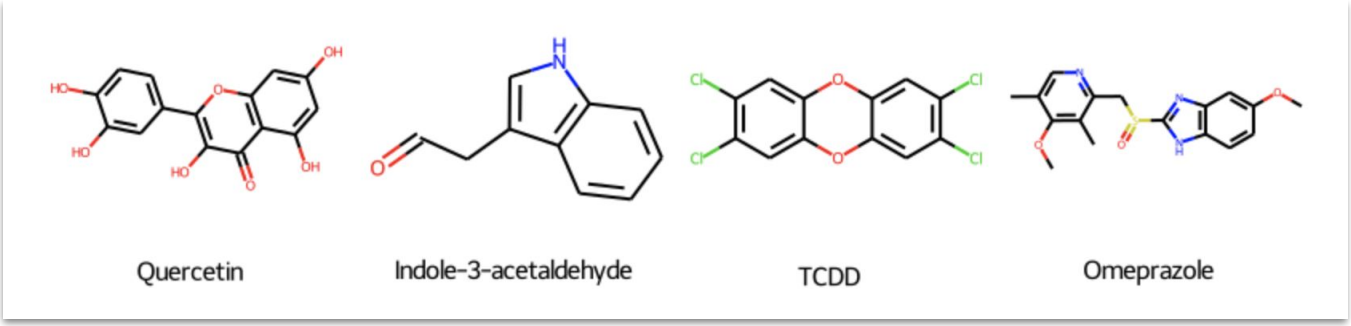


TCDD

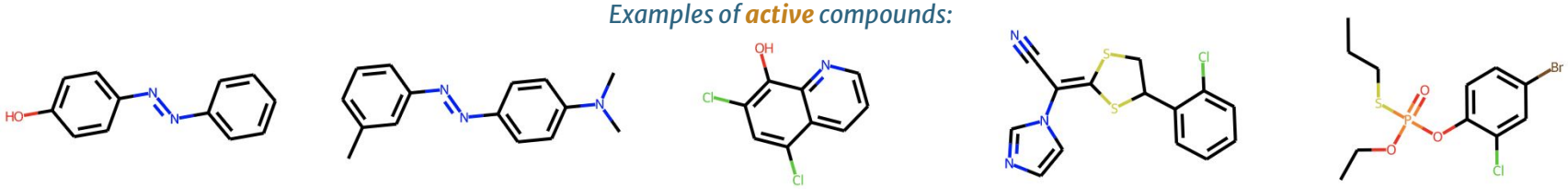


Omeprazole

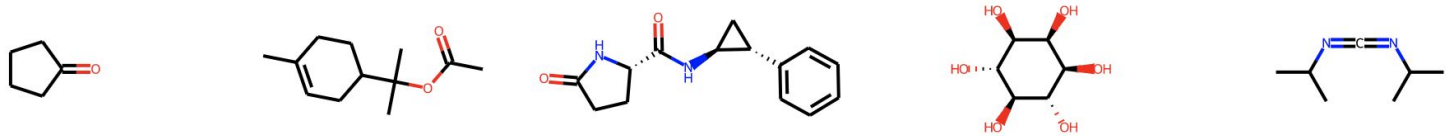
EDA: Visualizing the Data



*Examples of **active** compounds:*



*Examples of **inactive** compounds:*



Data Standardization: Tox21 NR-AhR Dataset

Removed duplicate and invalid compounds…

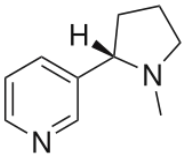
	Train	Test
Size (after cleaning)	6,709	607
Inactive compounds	5,948	536
Active compounds	761	71
Inactive : Active	~7	~7

Used MolVS to standardize SMILES representations for compounds…

	compounds	id	label	std_compounds
0	CC(O)=O.[H][C@@]12CCC3=CC(=CC=C3[C@@]1(C)CCC[C…	NCGC00255644-01	0	CC(=O)O.CC(C)c1ccc2c(c1)CC[C@@H]1[C@]2(C)CCC[C…
1	Cl.C[C@@H](NCCCC1=CC=CC(=C1)C(F)(F)F)C2=CC=CC3…	NCGC00181002-01	0	C[C@@H](NCCCc1cccc(C(F)(F)F)c1)c1cccc2ccccc12.Cl
2	CC(C)OC(=O)C1=C(C)NC(N)=C(C1C2=CC(=CC=C2)[N+](…	NCGC00167436-01	0	CC1=C(C(=O)OC(C)C)C(c2cccc([N+](=O)[O-])c2)C(C…
3	Cl.CN(C)C(=O)C1(CCN(CCC2(CN(CCO2)C(=O)C3=CC=CC…	NCGC00254013-01	0	CN(C)C(=O)C1(N2CCCCC2)CCN(CCC2(c3ccc(Cl)c(Cl)c…
4	Cl.CCOC(=O)O[C@H](C)OC(=O)C1=CC=C2N(CC3=NOC(=C…	NCGC00254071-01	0	CCOC(=O)O[C@H](C)OC(=O)c1ccc2c(c1)cc(C(=O)NC1C…

Featurization Techniques

Tokenizer	NR-AhR Features	Description
Atom-wise	131	Tokenizes SMILES based on individual atoms
K-mer (n-gram)	7,831	Tokenizes SMILES based on sequences of *k* overlapping characters in a string (4-mers)
SMILES Pair Encoding (SPE)	2,378	Tokenizes SMILES based on vocabulary learned from high frequency substrings of a large chemical dataset (ChEMBL)
Extended Circular Fingerprints (ECFP)	100	Set of all atom identifiers for each radius of perception up to a limit *n* (default = 4)
Molecular Descriptors (1D)	9	One-dimensional properties of a molecule (e.g., number of atoms, molecular weight)

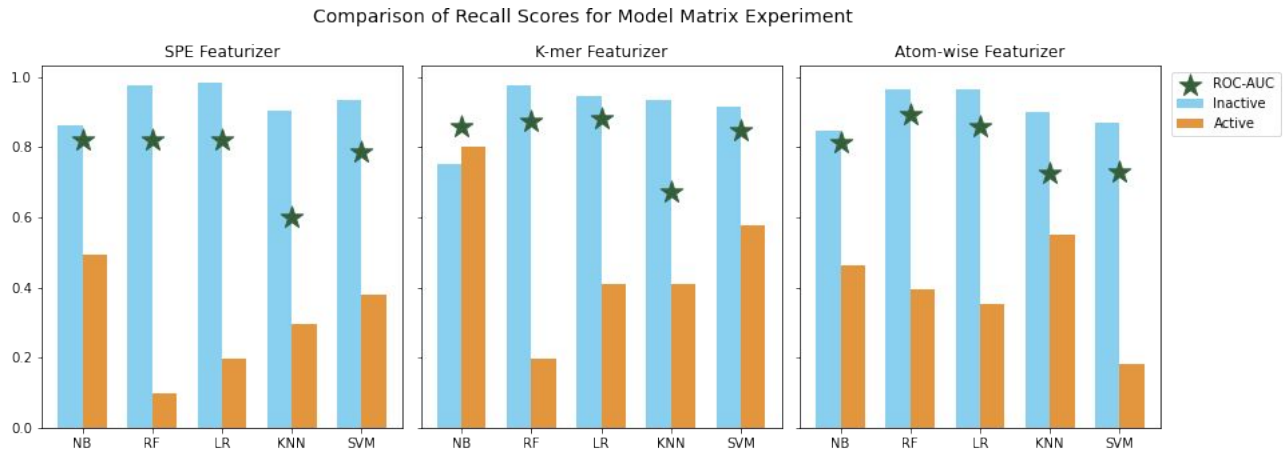


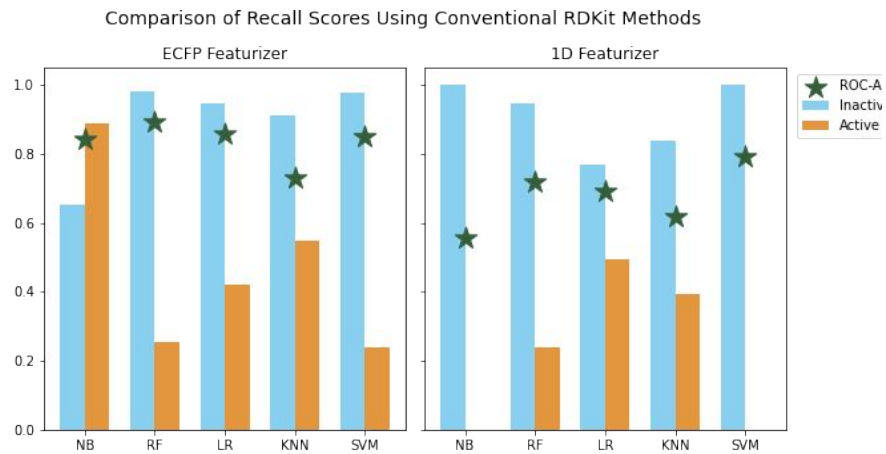
SMILES format:

`CN1CCC[C@H]1c2cccnc2`

Model Matrix Experiment Results

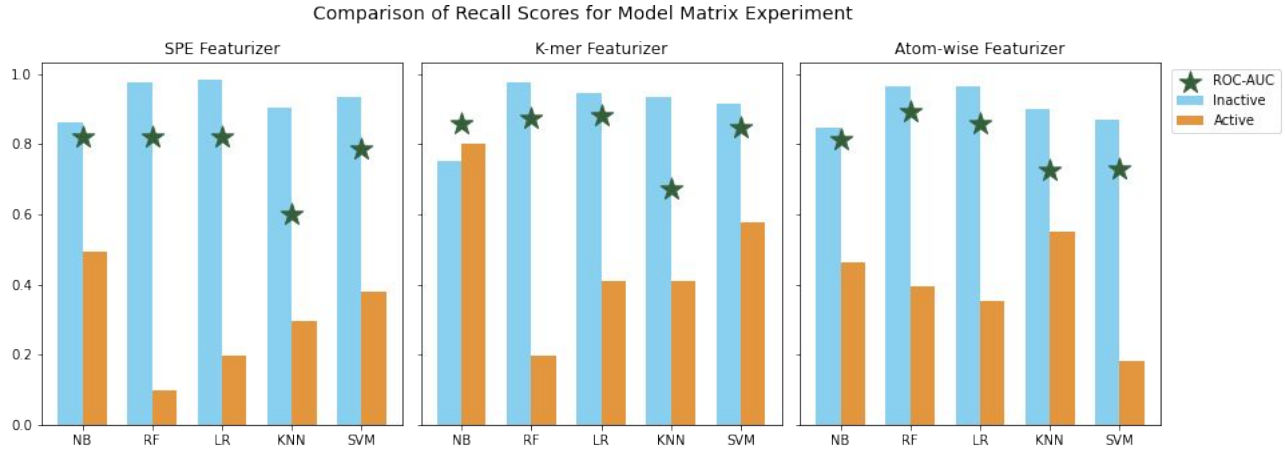
# Features	
SPE	2378
K-mer	7831
Atom	131
ECFP	100
1D	9

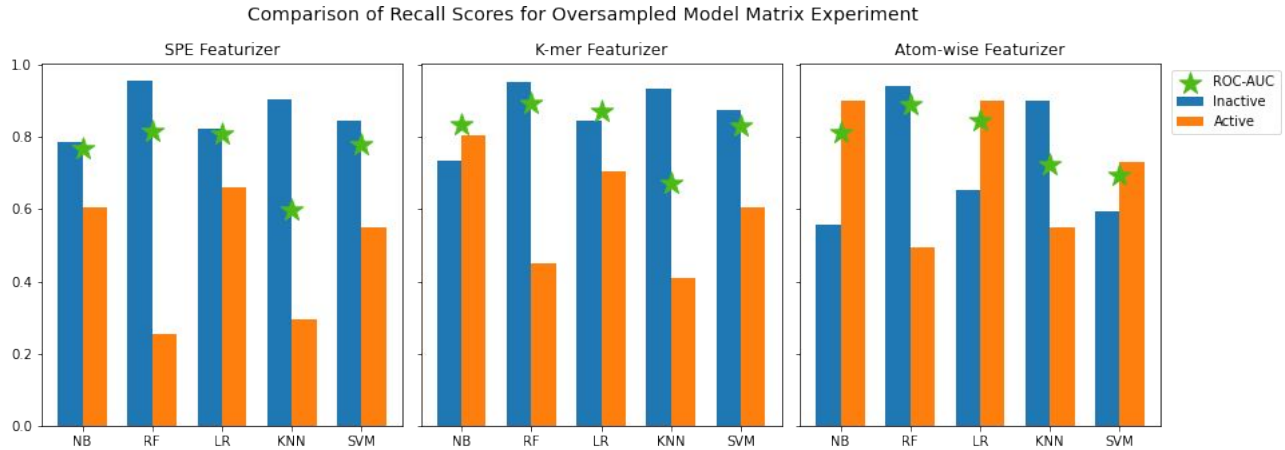




Oversampling Results

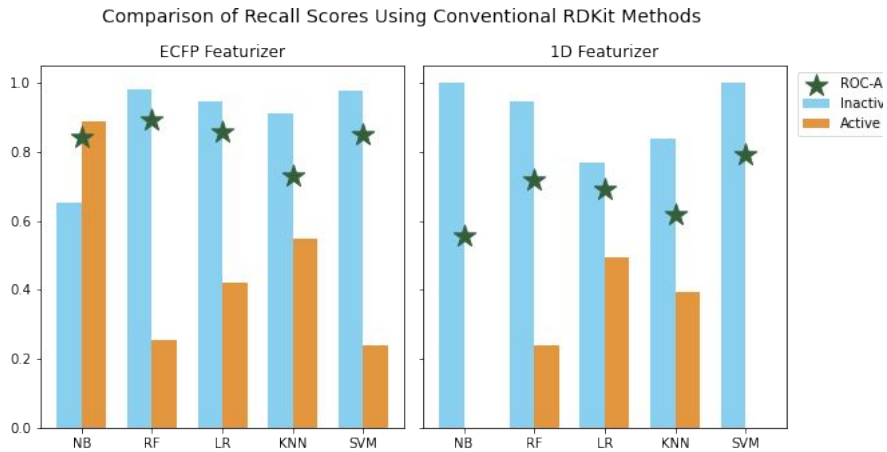
# Features	
SPE	2378
K-mer	7831
Atom	131
ECFP	100
1D	9

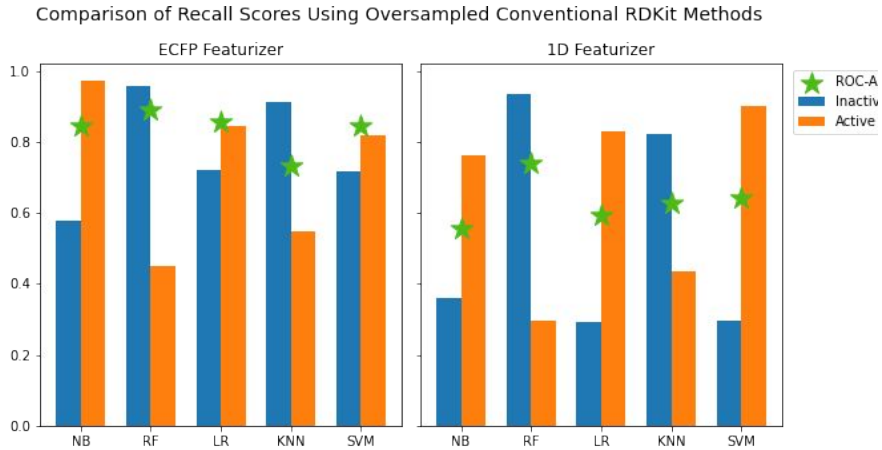




Oversampling Results, con't.

# Features	
SPE	2378
K-mer	7831
Atom	131
ECFP	100
1D	9



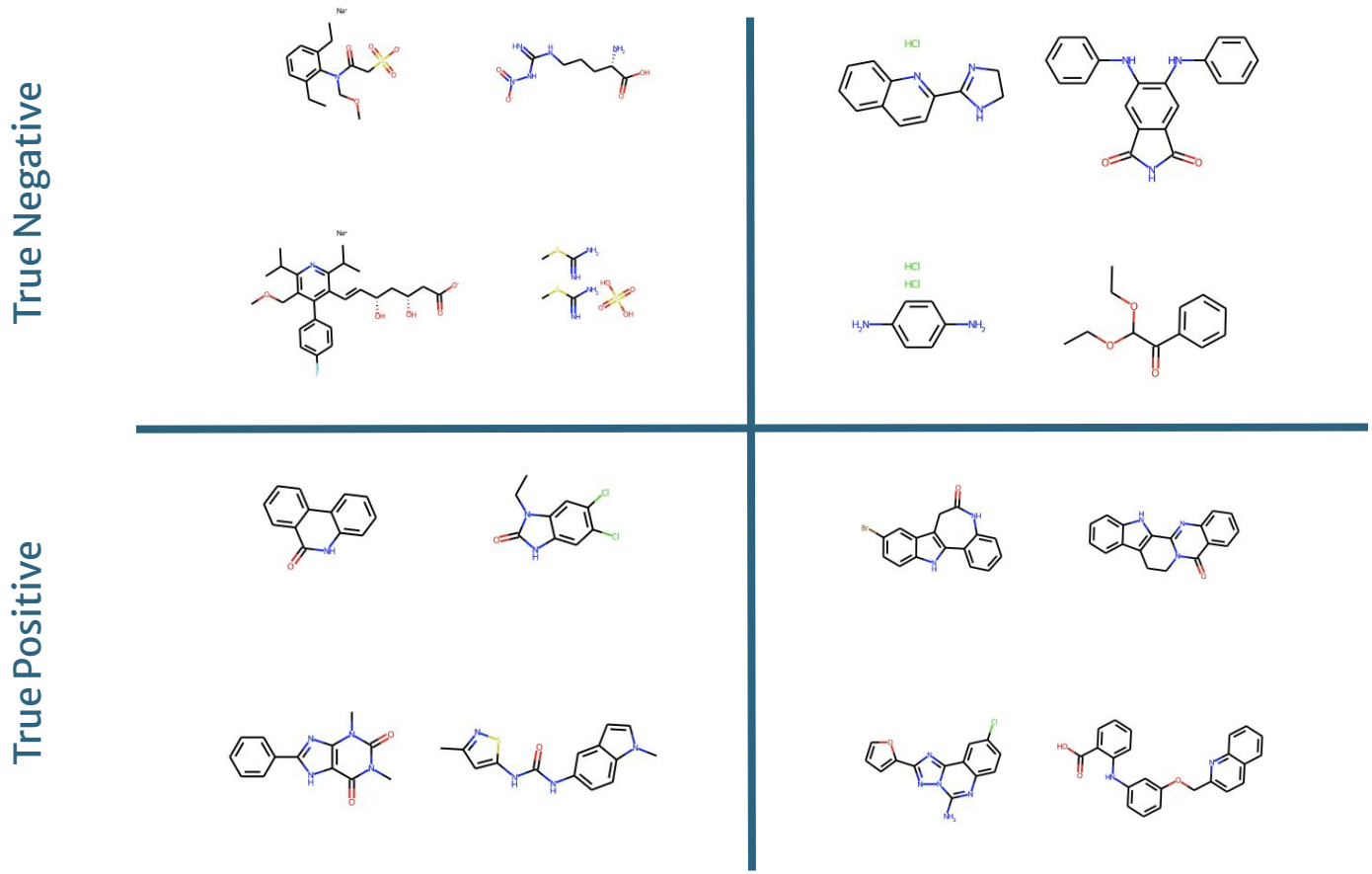


Top Features in SPE NB Model

Top 5 **Inactive** features

Top 5 **Active** features

Visual Confusion Matrix for SPE NB Model



Thank You!

Questions?