Machine Learning Models Identify Inhibitors of New Delhi Metallo-β-lactamase

Maria Weese

Miami University Center for Analytics and Data Science
Miami University Department of Information Systems & Analytics
Miami University Department of Chemistry and Biochemistry

Undergraduate Co-authors

Aidan Sturgill

Amy Hu

Mitch Fairweather

Antibiotic resistance

Antimicrobial resistance happens when germs like bacteria and fungi develop the ability to defeat the drugs designed to kill them.

Antimicrobial resistance is an urgent global public health threat, killing at least 1.27 million people worldwide.

In the U.S., more than 2.8 million antimicrobial-resistant infections occur each year.

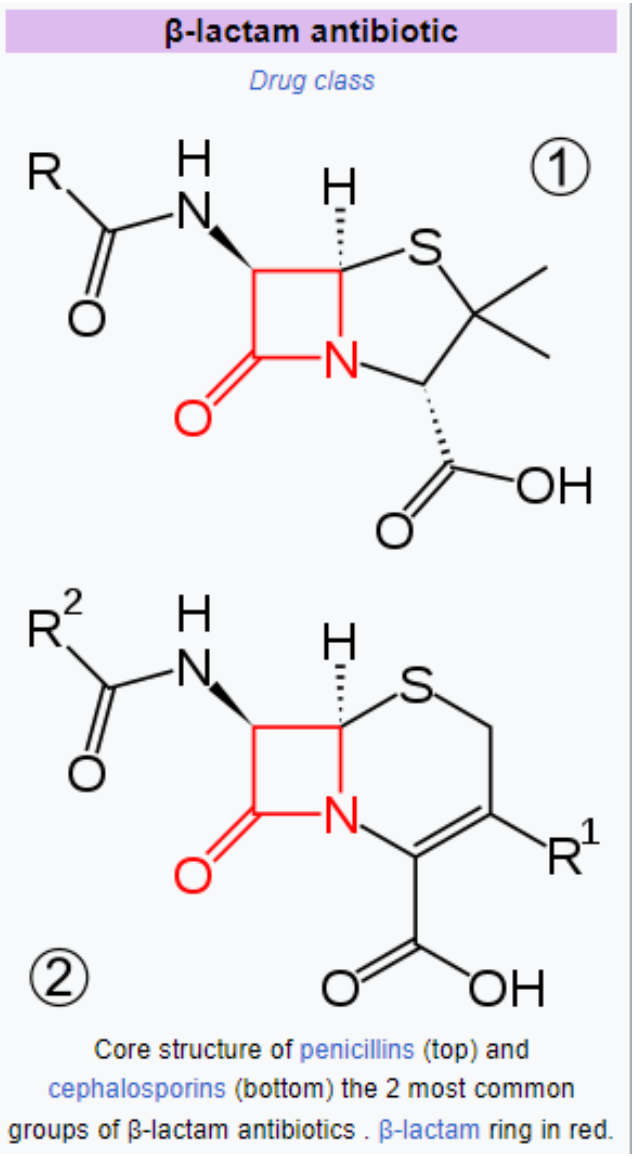
Source: https://www.cdc.gov/drugresistance/about.html

β-lactam antibiotics

Bacterial infections are most commonly treated by the use of β-lactam antibiotics.

A common mechanism for β-lactam resistance is the production of β-lactamases, which hydrolyze the β-lactam ring, thus rendering the drugs inactive.

Source: https://en.wikipedia.org/wiki/Beta-lactam_antibiotics



Core structure of penicillins (top) and cephalosporins (bottom) the 2 most common groups of β-lactam antibiotics . β-lactam ring in red.

β-Lactamases

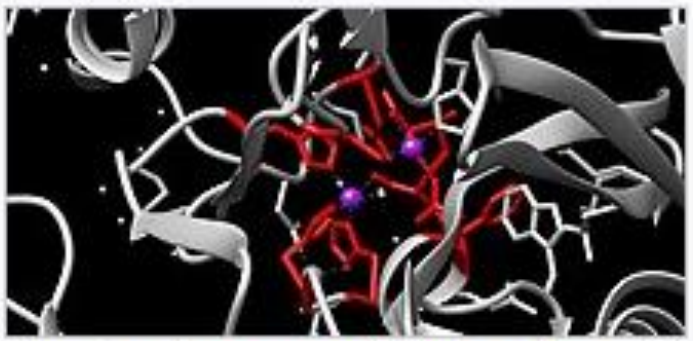
β-Lactamases can be categorized into serine-β-lactamases (SBLs) and metallo-β-lactamases (MBLs)

SBLs are more clinically-prevalent and there exist inhibitors, which given in combination with β-lactam containing antibiotics, combat bacteria that produce some of the SBLs.

There are no clinically-approved inhibitors available for MBLs, making infections from MBL-producing bacteria a serious challenge.

New Delhi metallo-β-lactamase (NDM-1),

- New Delhi metallo-β-lactamase (NDM-1) is an enzyme that makes bacteria resistant to a broad range of beta-lactam antibiotics
- NDM-1 is the most prevalent MBL worldwide.
- NDM-1 functions through two zinc ions present in the active site that cause hydrolysis of the beta-lactams, rendering them ineffective.
- Inhibitors either bind at the zinc site or rip the zinc off completely.



Two zinc ions present in the active binding site

Finding MBL Inhibitors

Current Techniques

- High-throughput screening (HTS) of large chemical libraries
- Fragment-based drug discovery (FBDD)
- Molecular docking

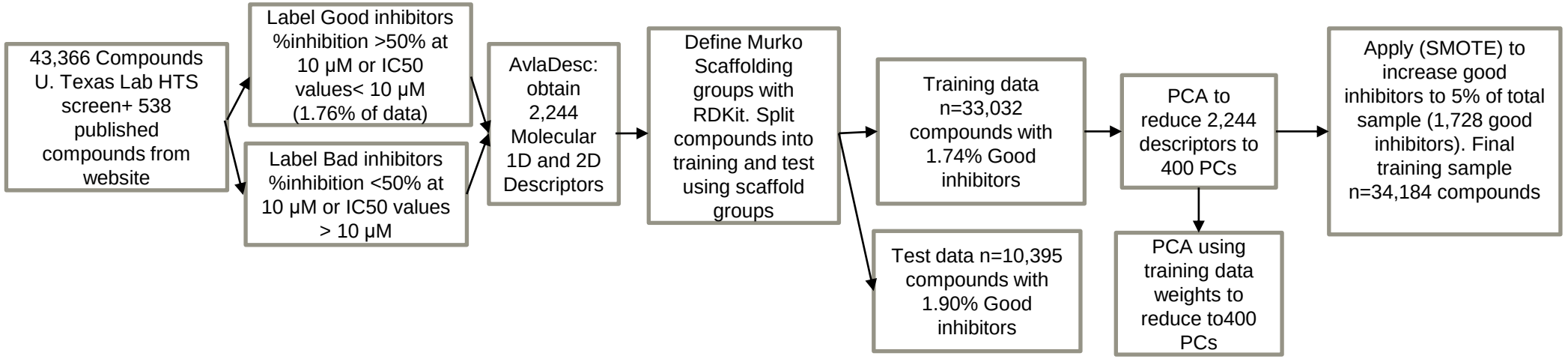
Drawbacks:

- The HTS and FBDD approaches are labor-intensive, costly, and time-consuming
- “Accurate” docking of compounds into existing MBLs (crystal structures) requires initial assumptions of how the compound(s) bind

Our approach

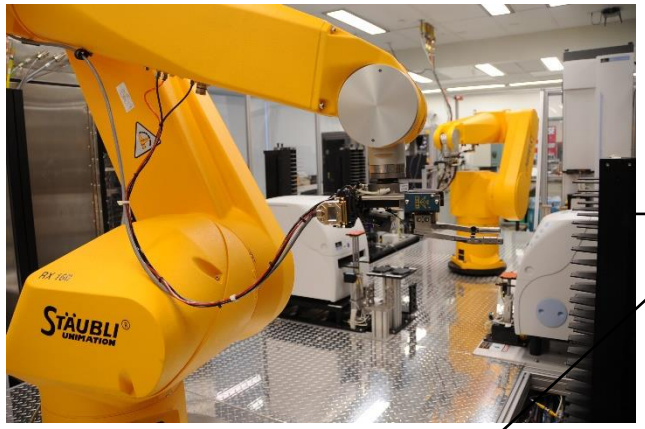
- Combines machine learning and HTS to identify inhibitors of NDM-1

Phase 1: Data and Pre-Processing



Phase 1: Data and Pre-Processing

43,366 Compounds U. Texas Lab HTS screen+ 538 published compounds from website



MβL INHIBITORS.com

https://mblinhibitors.miamioh.edu/

	A	B	C	D	E	F	G	H	I	J
		SMILE								
1		BrC1=CC2=C(NC(=O)\C2=N\NC(=O)C2=CC=C(C=C2)C2=NC=C(O2)C2=CC=CC=C2)C=C1								
2		OC1=C2N=CC=CC2=C(Cl)C=C1C(NC1=CC=CC=N1)C1=CC=CC(Cl)=C1								
3		COC1=CC(=CC(OC)=C1OC)C(=O)N\N=C\C1=CC(=CC=C1O)\N=N\C1=NON=C1C								
4		CCOC1=CC=C(C=C1)C(NC1=CC=CN=C1)C1=CC(Cl)=C2C=CC=NC2=C1O								
5		CC1=CC=NC(NC(C2=CC(C)=CC=C2C)C2=CC=C3C=CC=NC3=C2O)=C1								
6		OC1=C2N=CC=CC2=CC=C1C(NC1=NC=CC=C1)C1=CC(OC2=CC=CC=C2)=CC=C1								
7		CC1=CC=NC(NC(C2=CC=C(C=C2)N(=O)=O)C2=CC=C3C=CC=NC3=C2O)=C1								
8		OC1=C2N=CC=CC2=CC=C1C(NC1=CC=CC=N1)C1=CC=C(F)C=C1								
9		OC1=C2N=CC=CC2=CC=C1CN1CCN(CC2=CC=C3C=CC=NC3=C2O)CC1								
10		CC(C)C1=CC=C(C=C1)C(NC1=CC=CC=N1)C1=CC=C2C=CC(C)=NC2=C1O								
11		COC1=CC=C(C=C1C)C(NC1=CC(C)=CC=N1)C1=CC=C2C=CC=NC2=C1O								
12		COC1=C(OC)C=C(C=C1)C1=CSC(NC(=O)NC2=CC=C(C=C2)N(=O)=O)=N1								
13		CCC1=CC=C(C=C1)C(NC1=CC(C)=CC=N1)C1=CC=C2C=CC=NC2=C1O								
14		CC1=CC=C2C=CC(C(NC3=CC=CC=N3)C3=CC=CC=C3C)=C(O)C2=N1								
15		BrC1=CC=C(C=C1)C1=NC2=CC(NC(=O)COC3=CC=CC=C3N(=O)=O)=CC=C2O1								
16		CC1=CC=NC(NC(C2=CC=C3C=CC(C)=NC3=C2O)C2=CC=CC=C2N(=O)=O)=C1								
17		CC1=CC=C2C=CC(C(NC3=CC=CC=N3)C3=CC=CC(=C3)N(=O)=O)=C(O)C2=N1								
18		ClC1=CC=C(O1)\C=N\NC(=O)COC1=C2N=CC=CC2=C(Br)C=C1Br								
19		CC1=CC=C2C=CC(C(NC3=CC=CC=N3)C3=CC=CC=C3Cl)=C(O)C2=N1								
20		COC1=CC(=CC(I)=C1O)C(NC1=CC=CC=N1)C1=CC(Cl)=C2C=CC=NC2=C1O								

Phase 1: Data and Pre-Processing

Label Good inhibitors %inhibition >50% at 10 μM or IC50 values< 10 μM (1.76% of data)

Label Bad inhibitors %inhibition <50% at 10 μM or IC50 values > 10 μM

- IC_{50} is a quantitative measure indicating how much of a particular substance (e.g. drug) is needed to inhibit, *in vitro*, a given biological process or biological component by 50%.
- Published compounds were defined as "Good" if the published IC_{50} values were <10 μM.
- Compounds from HTS were defined as "Good" when more than 50% inhibition was observed at 10 μM

	A	B	C
1		SMILE	Response
2	1	BrC1=CC2=C(NC(=O)\C2=N\NC(=O)C2=CC=C(C=C2)C2=NC=C(O2)C2=CC=CC=C2)C=C1	Good
3	2	OC1=C2N=CC=CC2=C(Cl)C=C1C(NC1=CC=CC=N1)C1=CC=CC(Cl)=C1	Good
4	3	COC1=CC(=CC(OC)=C1OC)C(=O)N\N=C\C1=CC(=CC=C1O)\N=N\C1=NON=C1C	Good
5	4	CCOC1=CC=C(C=C1)C(NC1=CC=CN=C1)C1=CC(Cl)=C2C=CC=NC2=C1O	Good
6	5	CC1=CC=NC(NC(C2=CC(C)=CC=C2C)C2=CC=C3C=CC=NC3=C2O)=C1	Good
7	6	OC1=C2N=CC=CC2=CC=C1C(NC1=NC=CC=C1)C1=CC(OC2=CC=CC=C2)=CC=C1	Good
8	7	CC1=CC=NC(NC(C2=CC=C(C=C2)N(=O)=O)C2=CC=C3C=CC=NC3=C2O)=C1	Good
9	8	OC1=C2N=CC=CC2=CC=C1C(NC1=CC=CC=N1)C1=CC=C(F)C=C1	Good
10	9	OC1=C2N=CC=CC2=CC=C1CN1CCN(CC2=CC=C3C=CC=NC3=C2O)CC1	Good
11	10	CC(C)C1=CC=C(C=C1)C(NC1=CC=CC=N1)C1=CC=C2C=CC(C)=NC2=C1O	Good
12	11	COC1=CC=C(C=C1C)C(NC1=CC(C)=CC=N1)C1=CC=C2C=CC=NC2=C1O	Good
13	12	COC1=C(OC)C=C(C=C1)C1=CSC(NC(=O)NC2=CC=C(C=C2)N(=O)=O)=N1	Good
14	13	CCC1=CC=C(C=C1)C(NC1=CC(C)=CC=N1)C1=CC=C2C=CC=NC2=C1O	Good
15	14	CC1=CC=C2C=CC(C(NC2=CC=CC=N2)C2=CC=CC=C2)=C(O)C2=N1	Good

Phase 1: Data and Pre-Processing

SMILE: Simplified Molecular Input Entry System

AvlaDesc: obtain 2,244 Molecular 1D and 2D Descriptors

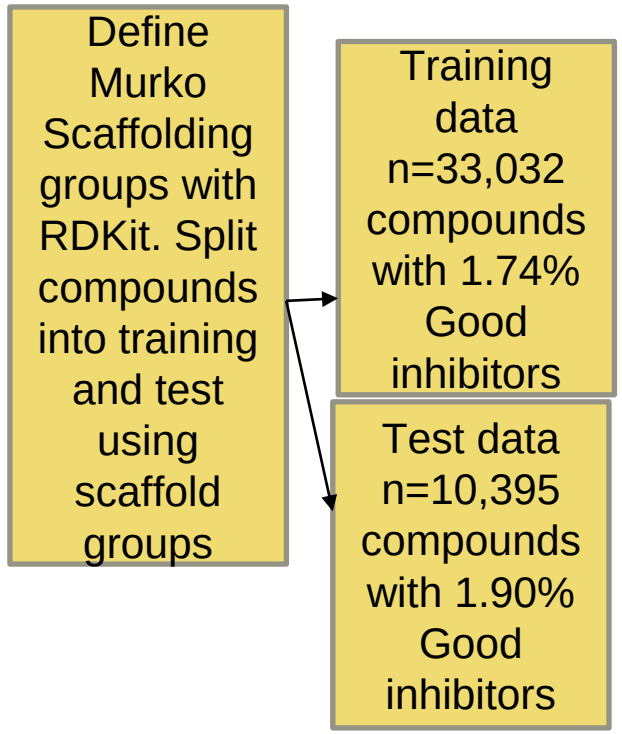
alvaDesc

A	B	C	D	E	F	G	H	I	J	K	L	M	N
	SMILE.x	Response	MW	AMW	Sv	Se	Sp	Si	Mv	Me	Mp	Mi	GD
1	BrC1=CC2=C(NC(=O)\C2=N\NC(=O)C2=CC=C(C=C2)C2=NC=C(O2)C2=CC=CC=C2)C=C1	Good	487.33	10.36872	34.416	47.9198	35.307	51.9542	0.732255	1.01957	0.751213	1.105409	0.072581
2	OC1=C2N=CC=CC2=C(Cl)C=C1C(NC1=CC=CC=N1)C1=CC=CC(Cl)=C1	Good	396.29	9.435476	30.1216	42.4653	31.5172	46.4987	0.717181	1.011079	0.75041	1.107112	0.08547
3	COC1=CC(=CC(OC)=C1OC)C(=O)N\N=C\C1=CC(=CC=C1O)\N=N\C1=NON=C1C	Good	440.46	8.470385	34.1048	53.7598	34.091	59.1526	0.655862	1.033842	0.655596	1.13755	0.068548
4	CCOC1=CC=C(C=C1)C(NC1=CC=CN=C1)C1=CC(Cl)=C2C=CC=NC2=C1O	Good	405.91	8.283878	33.0625	49.2361	34.6366	54.5945	0.674745	1.004818	0.706869	1.114173	0.078818
5	CC1=CC=NC(NC(C2=CC(C)=CC=C2C)C2=CC=C3C=CC=NC3=C2O)=C1	Good	369.5	7.245098	33.047	50.4687	35.0856	56.8563	0.64798	0.989582	0.687953	1.114829	0.082011
6	OC1=C2N=CC=CC2=CC=C1C(NC1=NC=CC=C1)C1=CC(OC2=CC=CC=C2)=CC=C1	Good	419.51	7.915283	36.235	52.9124	37.7787	58.6505	0.683679	0.998347	0.712806	1.106613	0.072581
7	CC1=CC=NC(NC(C2=CC=C(C=C2)N(=O)=O)C2=CC=C3C=CC=NC3=C2O)=C1	Good	386.44	8.222128	31.9176	47.5743	32.7161	52.5278	0.679098	1.012219	0.696087	1.117613	0.078818
8	OC1=C2N=CC=CC2=CC=C1C(NC1=CC=CC=N1)C1=CC=C(F)C=C1	Good	345.4	8.22381	28.8499	42.3306	29.7389	46.9504	0.686902	1.007871	0.708069	1.117867	0.089231
9	OC1=C2N=CC=CC2=CC=C1CN1CCN(CC2=CC=C3C=CC=NC3=C2O)CC1	Good	400.52	7.417037	34.7832	53.8978	36.5458	60.564	0.644133	0.998107	0.676774	1.121556	0.078161
10	CC(C)C1=CC=C(C=C1)C(NC1=CC=CC=N1)C1=CC=C2C=CC(C)=NC2=C1O	Good	383.53	7.102407	34.5738	53.3523	36.847	60.2715	0.640256	0.988006	0.682352	1.116139	0.078818

Phase 1: Data and Pre-Processing

There is a relationship between the chemical or 3D structure of a molecule and its biological activity.

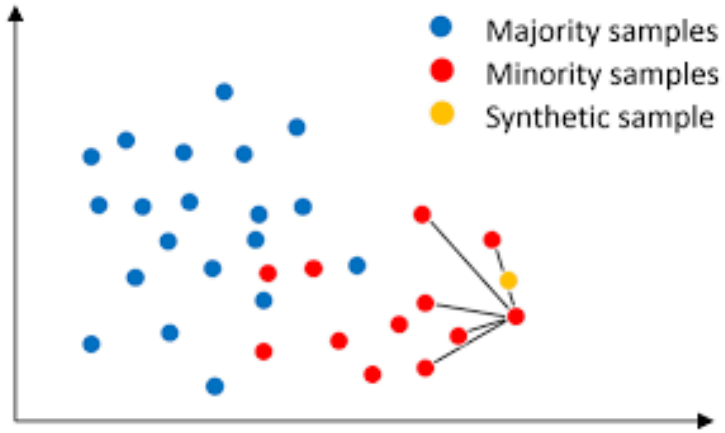
We used Python's RDKit package to break compounds into scaffolding groups.

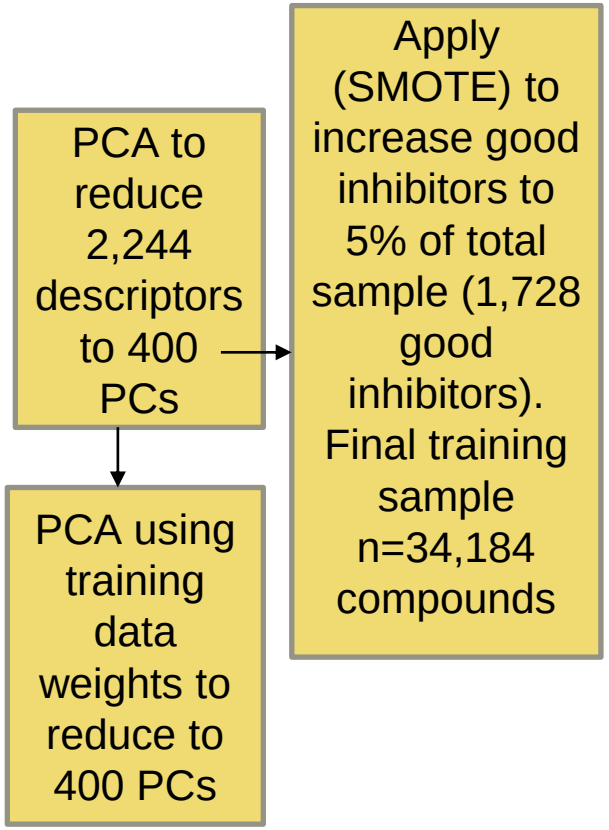


The use of scaffold-based sampling in the creation of the training and test data sets led to improved model performance when new compounds were encountered by the model (Yang et al. (2018))

Phase 1: Data and Pre-Processing

- To reduced computational burden and account for redundancy in the descriptors we used PCA.
- 400 PC accounted for 99.125% of the original variation
- Synthetic Minority Oversampling Technique (SMOTE) is a common method to deal with imbalanced data.





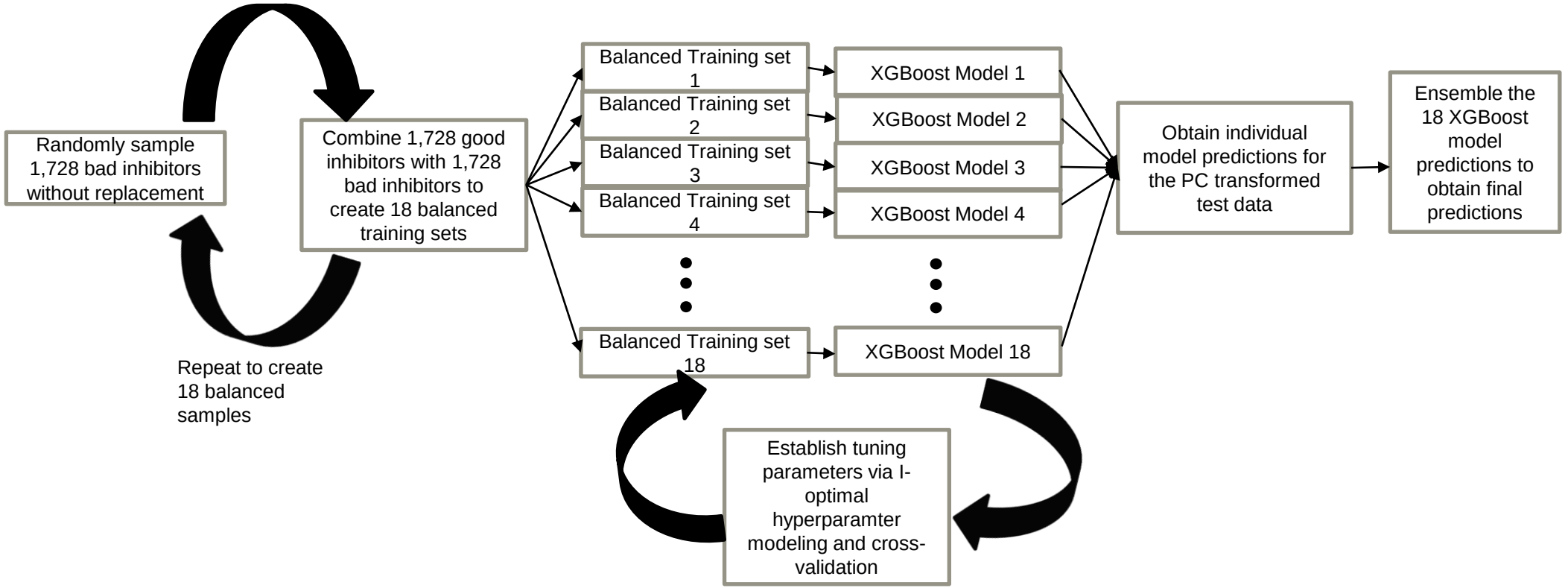
XGBoost

- XGBoost (eXtreme Gradient Boosting) is an implementation of gradient boosted decision trees designed for speed and performance.
- The objective function of XGBoost contains a regularization parameter that controls the complexity of the trees.

$$Obj = \sum_{i=1}^{n} l(y_i, \widehat{y_i}) + \sum_{k=1}^{K} \Omega(f_k)$$

- The regularization parameter encourages simple models which in turn have smaller variance in future predictions, making them stable.

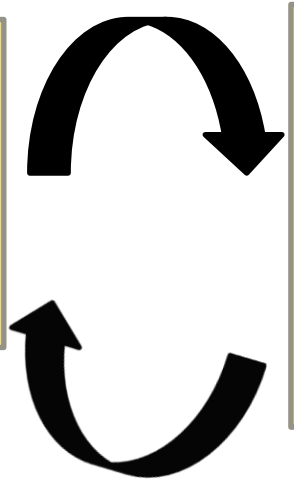
Phase II: Model Tuning



Phase II: Model Tuning

Randomly sample 1,728 bad inhibitors without replacement

Combine 1,728 good inhibitors with 1,728 bad inhibitors to create 18 balanced training sets



Repeat to create 18 balanced samples

IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS—PART B: CYBERNETICS, VOL. 39, NO. 2, APRIL 2009 539

Exploratory Undersampling for Class-Imbalance Learning

Xu-Ying Liu, Jianxin Wu, and Zhi-Hua Zhou, *Senior Member, IEEE*

***Abstract*—Undersampling is a popular method in dealing with class-imbalance problems, which uses only a subset of the majority class and thus is very efficient. The main deficiency is that many majority class examples are ignored. We propose two algorithms to overcome this deficiency. `EasyEnsemble` samples several subsets from the majority class, trains a learner using each of them, and combines the outputs of those learners. `BalanceCascade` trains the learners sequentially, where in each step, the majority class examples that are correctly classified by the current trained learners are removed from further consideration. Experimental results show that both methods have higher Area Under the ROC Curve, F-measure, and G-mean values than many existing class-imbalance learning methods. Moreover, they have approximately the same training time as that of undersampling when the same number of weak classifiers is used, which is significantly faster than other methods.**

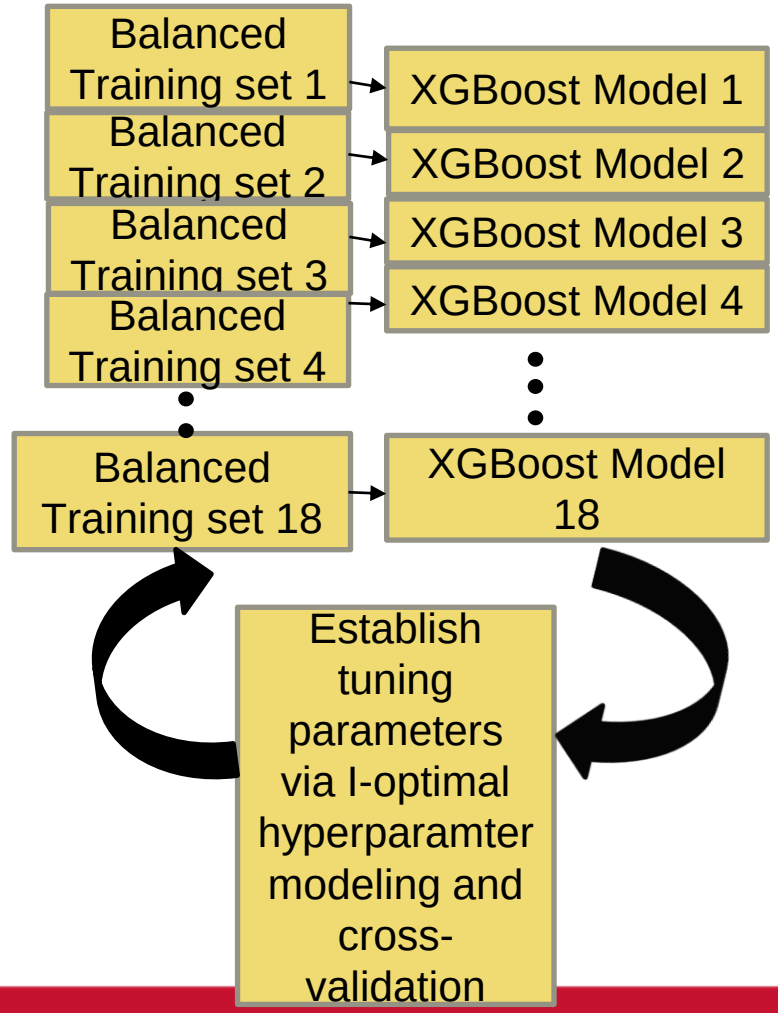
***Index Terms*—Class-imbalance learning, data mining, ensemble learning, machine learning, undersampling.**

fying a minority class instance is usually more serious than misclassifying a majority class one. For example, approving a fraudulent credit card application is more costly than declining a credible one. Breiman *et al.* [7] pointed out that training set size, class priors, cost of errors in different classes, and placement of decision boundaries are all closely connected. In fact, many existing methods for dealing with class imbalance rely on connections among these four components. Sampling methods handle class imbalance by varying the minority and majority class sizes in the training set. Cost-sensitive learning deals with class imbalance by incurring different costs for the two classes and is considered as an important class of methods to handle class imbalance [37]. More details about class-imbalance learning methods are presented in Section II.

In this paper, we examine only binary classification problems by ensembling classifiers built from multiple undersampled

Phase II: Model Tuning

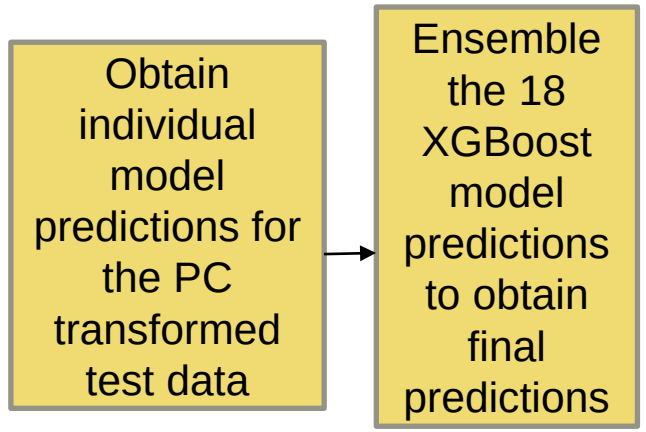
- XGBoost has 6 tuning parameters
- Model Tuning Methods
  - grid search
  - Bayesian optimization
- We used an I-optimal experimental design (n=34) and second order model to find optimum values for the tuning parameters for each of 18 samples
- Response is AUC



- Sample to sample variation is small
- We can assess overfitting, common to ensemble models
- Second-order model R^2=0.99 (training) R^2=0.94 (test)

Phase II: Model Tuning

- At this point we have 18 individual XGBoost models.
- We used each model to obtain predictions on full test set and then combined predictions (simple average).

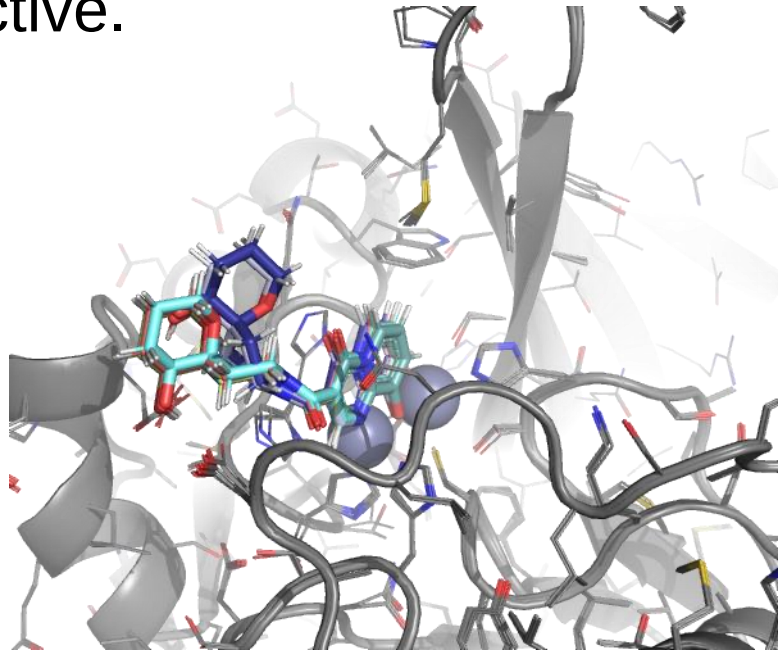


Model Validation

- We applied the model to the National Center for Advancing Translational Sciences (NCATS) Genesis library containing 76,369 unique compounds
- The model was used to score and rank compounds
- The top 2,816 compounds were then used in quantitative HTS
- 160 compounds were flagged as “hits” IC_{50} < 50μM
- 9 of those had IC_{50} < 10μM
- This translates to activity rate of 0.32% for the IC_{50} < 10μM compounds
- Most HTS studies have activity rate of 0.01-0.14%
- Model lift between 2.23 and 32!

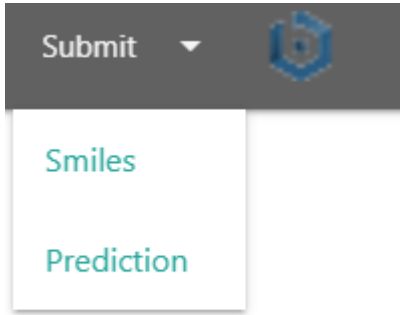
Inhibitor 72922413

- This one of the “Good” inhibitors flagged by the model in the Genesis library.
- This inhibitor binds to NDM-1 at the metal site the inhibitor does not "pull" the zinc out of NDM-1, which is good!
- This inhibitor potentially stops NDM-1 from hydrolyzing the beta-lactams by binding at the zinc ions, rendering them ineffective.



Conclusion

- The model is freely available for scientist to score potential new MBL on the https://mblinhibitors.miamioh.edu website.



- Next steps are to create training data for toxicity, solubility and other important properties needed for a compound to be a viable drug.

Research reported in this presentation was supported by the National Institutes of Health under award number: R15GM134454

References

Yang K, Swanson K, Jin W, Coley C, Eiden P, Gao H, Guzman-Perez A, Hopper T, Kelley B, Mathea M, Palmer A, Settels V, Jaakkola T, Jensen K, Barzilay R. 2019. Analyzing Learned Molecular Representations for Property Prediction. *J Chem Inf Model* 59:3370-3388.

Liu XY, Wu J, Zhou ZH. 2009. Exploratory undersampling for class-imbalance learning. *IEEE Trans Syst Man Cybern B Cybern* 39:539-50.