



OPEN

An interpretable machine learning approach to identify mechanism of action of antibiotics

Mihir Mongia^{1,2}, Mustafa Guler^{1,2} & Hosein Mohimani¹✉

As antibiotic resistance is becoming a major public health problem worldwide, one of the approaches for novel antibiotic discovery is re-purposing drugs available on the market for treating antibiotic resistant bacteria. The main economic advantage of this approach is that since these drugs have already passed all the safety tests, it vastly reduces the overall cost of clinical trials. Recently, several machine learning approaches have been developed for predicting promising antibiotics by training on bioactivity data collected on a set of small molecules. However, these methods report hundreds/thousands of bioactive molecules, and it remains unclear which of these molecules possess a novel mechanism of action. While the cost of high-throughput bioactivity testing has dropped dramatically in recent years, determining the mechanism of action of small molecules remains a costly and time-consuming step, and therefore computational methods for prioritizing molecules with novel mechanisms of action are needed. The existing approaches for predicting bioactivity of small molecules are based on uninterpretable machine learning, and therefore are not capable of determining known mechanism of action of small molecules and prioritizing novel mechanisms. We introduce InterPred, an interpretable technique for predicting bioactivity of small molecules and their mechanism of action. InterPred has the same accuracy as the state of the art in bioactivity prediction, and it enables assigning chemical moieties that are responsible for bioactivity. After analyzing bioactivity data of several thousand molecules against bacterial and fungal pathogens available from Community for Open Antimicrobial Drug Discovery and a US Food and Drug Association-approved drug library, InterPred identified five known links between moieties and mechanism of action.

Antibiotic resistance bacteria kills 700,000 worldwide each year¹. Overcoming the challenge of antibiotic resistance requires development of antibiotics that can kill bacteria using novel modes of action. While the need for novel antibiotics has become more urgent in recent years, the progress in development of novel therapeutics has slowed down due to hurdles in discovery pipelines and lack of economic incentives. With current trends it is forecast that by 2050, the mortality rate of antibiotic resistance bacteria will be over ten million worldwide², surpassing that of cancer.

With hundreds of millions of known molecular structures available in molecular libraries^{3–5}, methods for prediction of bioactivity solely based on chemical structure can aid in selecting promising molecules active against targets of interest for downstream bioactivity testing. The first techniques for predicting the relationship between chemical structure and activity were rule based⁶. The first machine learning approaches for prediction of structure-activity relationships from training data appeared in the 1990s⁷.

Neural networks were among the earliest methods used for learning structure-activity relationship from training data⁸. However until very recently, neural networks (and many of other machine learning methods) were not capable of directly taking the graph structure of molecules as input. Recent advances in graph-based machine learning have enabled representation of complex molecular structures with real-valued vectors, making it feasible to incorporate local/global structural information for the prediction of molecular properties⁹.

Recently, Stokes et al.¹⁰ applied a directed message passing neural network (D-MPNN) for prediction of the antibiotic activity of small molecules. In this approach, a vector representation is learned for each atom, and the molecular property is predicted as a learned non-linear transformation of these representations^{11,12}. Stokes et al. reported strong antimicrobial activity for Halicin, a drug chemical compound originally developed for treatment of diabetes¹³.

¹Computational Biology Department, School of Computer Science, Carnegie Mellon University, Pittsburgh, USA.

²These authors contributed equally: Mihir Mongia and Mustafa Guler. ✉email: hoseinm@andrew.cmu.edu

One of the main bottlenecks of the existing approaches is that they usually report hundreds/thousands of molecules, where the majority of them possess known mechanisms of action. Overcoming antibiotic resistant pathogens crucially depends on finding small molecules with novel mechanism of action, and currently determining the mechanism of action for small molecules remains an expensive and time-consuming effort. Therefore, it is crucial to develop computational methods for determining molecules with known mechanisms of action and prioritize molecules with novel mechanisms.

Mechanism of action of small molecules are usually linked to their bioactive moieties. One way to extract these moieties would be to find features of the molecule graph that correlate to bioactivity. Methods such as recursive feature elimination¹⁴, boruta¹⁵, and lasso¹⁶ have been developed for this purpose, but they are limited to cases where a feature set is available.

Another method to find bioactive moieties is to determine the portion of a molecular graph that a D-MPNN uses to make a prediction. Several heuristic approaches have been developed in order to interpret graph neural networks. One approach is to take the gradient of neural networks with respect to the atoms in the molecular graph^{17–20} and to attribute atoms with more importance if the gradient value for an atom is large. The set of atoms determined to be important by this approach, however, are not necessarily biologically relevant as often a large portion of the molecular graph is flagged as important. Furthermore although gradient methods have had some empirical success, the gradient only represents how the model changes with small perturbations, and high gradient values for atoms do not necessarily mean those atoms are important for classification by a neural network. Another approach for interpreting graph neural networks is to exhaustively search all subgraphs of a molecular graph and find those subgraphs that are either subsets of important nodes as determined by the gradient method or those that do not change the output of the neural network significantly^{17,21}. These methods again often fail in capturing reasonable bioactive moieties as they highlight subgraphs that are common in the molecular space.

Interpreting which substructures are responsible for bioactivity is a challenging problem for the existing algorithms, as there are an exponential number of substructures of molecular graphs, and it is impossible to correctly infer which of these millions of substructures are responsible for activity from a few thousand training points. One way to overcome this issue is to limit the candidate substructures to those that are biologically important, including simple ring structures and functional groups^{22,23}. This knowledge however has rarely been integrated in machine learning methods for drug discovery.

In this work we develop an interpretable machine learning model by first identifying the simple ring structures and functional groups in the training data and using them to create binary feature vectors for each molecule where zeros and ones indicate absence/presence of rings and functional groups. Using simple rings and structures as features is advantageous since it is easier to interpret the correlations between these features and mechanism of action in the downstream analysis. Then we train a logistic regression or extra trees model with balanced scoring on these features in order to create a low complexity model that accounts for imbalanced data. Our model achieves similar accuracy as the D-MPNN in Stokes et al. while being fully interpretable, and it clusters molecules based on their mechanism of action. Moreover, the method can associate a bioactive molecule with its bioactive moiety, providing a strategy for prioritizing molecules with novel mechanism of action. Application of our method to the Community for Open Antimicrobial Drug Discovery (CO-ADD) and a FDA-approved dataset of antibacterial and antifungal bioactives of several thousand molecules assigned five known mechanism of actions to their moieties.

Results

Overview of InterPred. InterPred predicts bioactivity of small molecules in the following steps (Fig. 1). Given (a) a collection of molecules (b) all unique simple ring structures and functional groups are extracted into binary vectors where 0/1 indicates absence/presence of a substructure. Then, (c) extra trees/logistic regression classifier with ℓ_1 regularization is trained on the extracted binary features using balanced scoring. Given (d) a query molecule, (e) binary features are extracted, and (f) the trained model is used for predicting bioactivity.

Overview of MOACluster. MOACluster groups molecules with similar mechanism of action (MOA) in the following steps (Fig. 2). Given (a) a collection of molecules, MOACluster extracts their binary features. Then (b) a logistic regression classifier with ℓ_1 regularization and balanced scoring is trained to predict bioactivity, and the model parameters are extracted. (c) MOACluster finds the indices of the top k coefficients and reduces molecule binary features to those k indices. (d) The molecules are clustered according to the reduced binary features.

Datasets. InterPred was trained on two datasets. The first dataset contains molecules from a US Food and Drug (FDA)—approved library, along with 800 natural products isolated from plant, animal, and microbial sources (total of 2335 unique compounds)¹⁰. Data on growth inhibition against *Escheria coli* is available for all the molecules. The corresponding test data contains growth inhibition of 162 molecules from the Drug Repurposing Hub²². Each molecule in the test data is annotated with a mechanism of action by which it fights the disease it was originally purposed for. The second data set, CO-ADD^{23,24}, contains bio-activity data from 4,803 molecules against seven bacterial and fungal pathogens, which include *Staphylococcus aureus*, *Pseudomonas aeruginosa*, *Acinetobacter baumannii*, *Candida albicans*, *Klebsiella pneumoniae*, *Cryptococcus neoformans*, and *Escheria coli*. For this dataset, 80% of the molecules are randomly selected for training and the rest are allocated for testing.

Bioactivity prediction. Figure 3 illustrates the receiver operating characteristic (ROC) curve of InterPred compared to the approach from Stokes et al. on predicting activity against *E. coli*. Here 2335 molecules have been used for training, 162 molecules have been used for testing. These test molecules correspond to the portion of

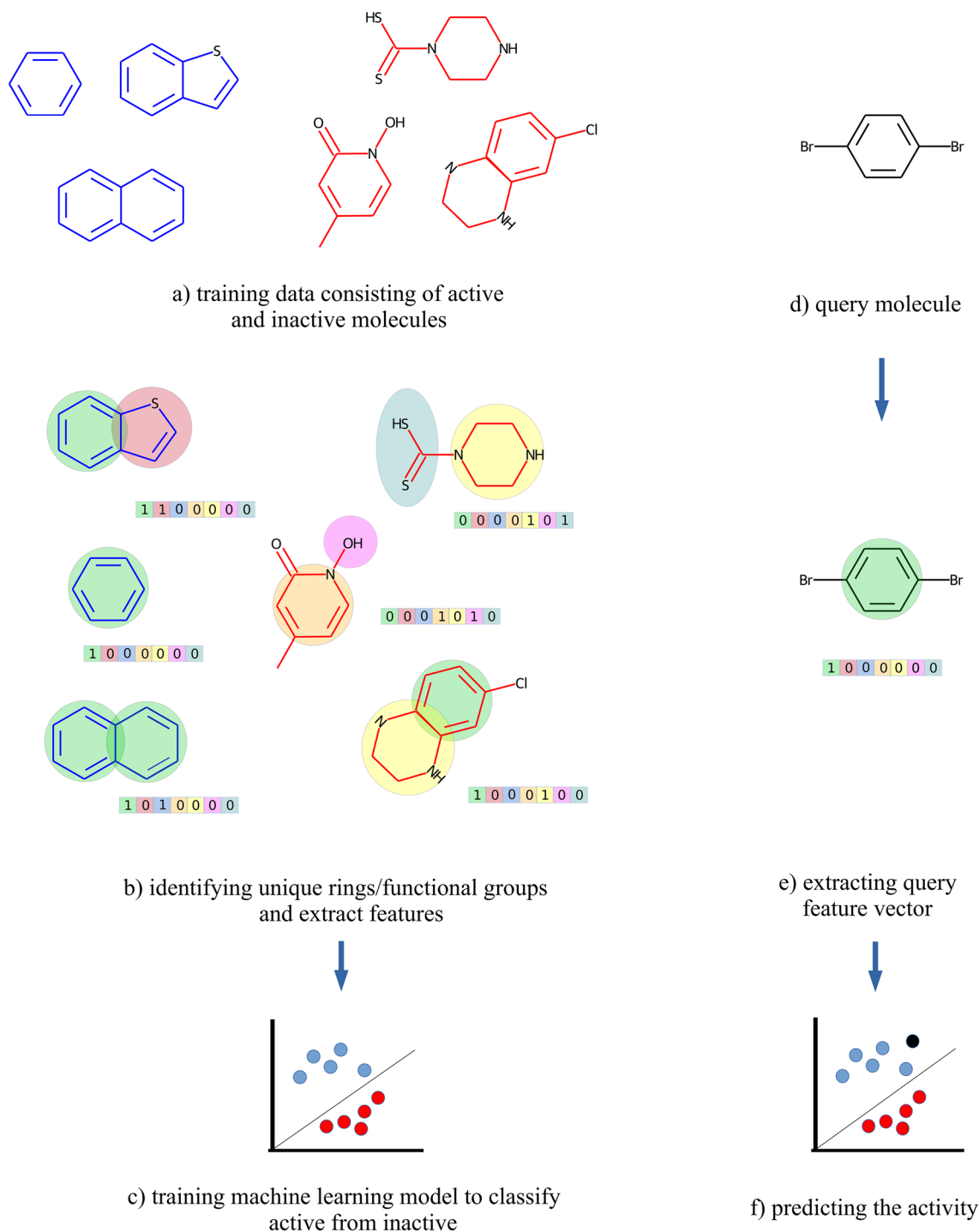


Figure 1. Interpretable prediction of antibiotic activity. Given (a) a collection of molecules, (b) InterPred finds all unique ring structures and functional groups and creates a binary vector for each molecule where zero/one indicates absence/presence of ring structure. Then (c) InterPred trains a logistic regression classifier with l_1 -regularization and balanced scoring/extra trees classifier on the resulting binary features. Given (d) a query molecule, (e) InterPred extracts the binary features, and (f) applies the logistic regression classifier/extra trees classifier to predict the activity.

the Drug Repurposing Hub for which screening data against *E. coli* is available (Stokes et al.). Figure 4 shows the distribution of the tanimoto similarity between each test data point and their closest neighbor in the training dataset. InterPred achieves nearly the same accuracy as Stokes et al. The area under the curve (AUC) for InterPred is 0.87 while the AUC for Stokes et al. is 0.88. Unlike Stokes et al., InterPred uses fully interpretable features.

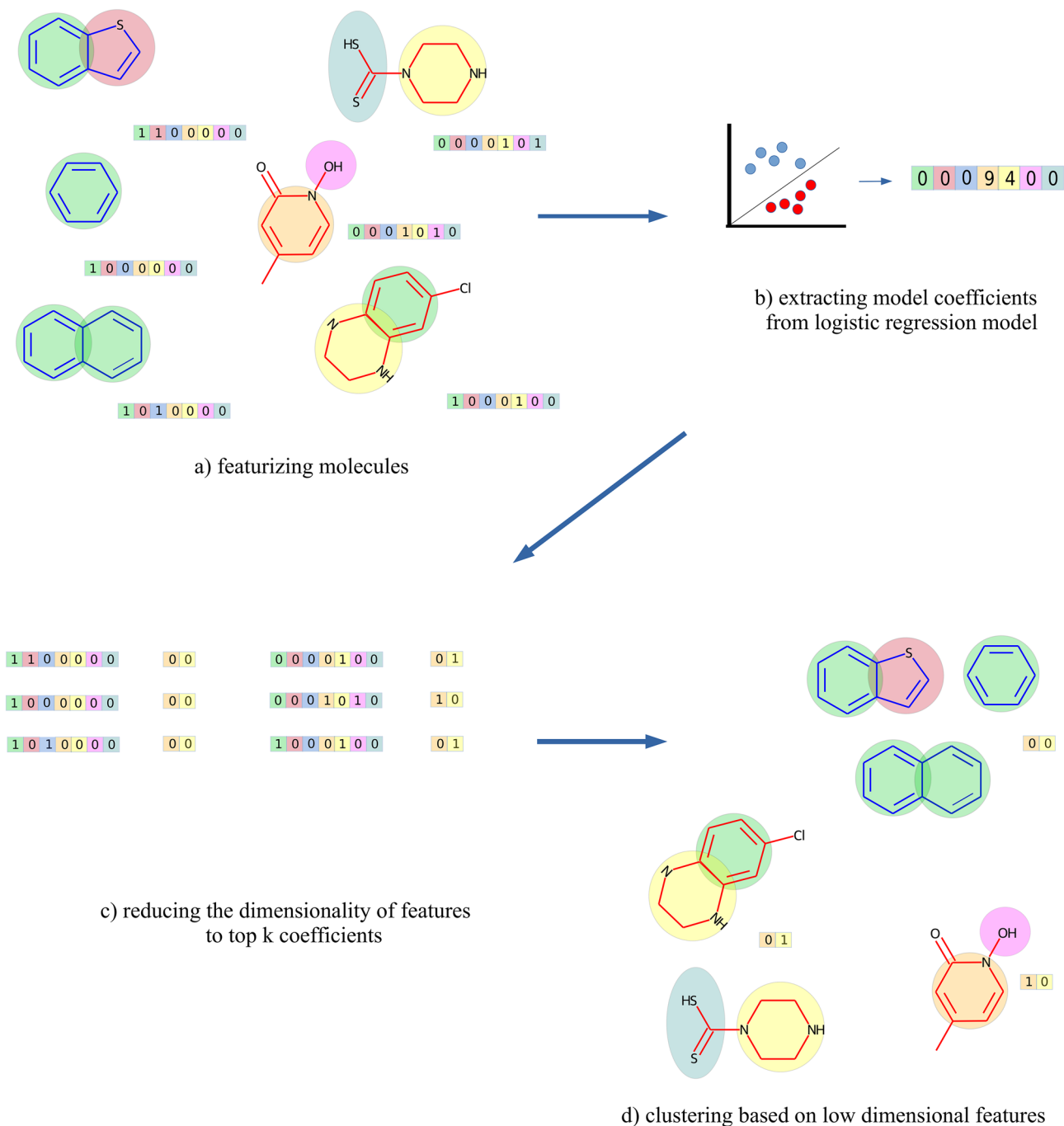


Figure 2. Clustering molecules based on their mechanism of action. Given (a) a collection of molecules, MOACluster extracts their binary features. Then (b) a logistic regression classifier is trained on the resulting binary features and the model coefficients are extracted. (c) MOACluster finds the indices of the top k coefficients and reduces molecule binary features to those k indices. (d) The molecules are clustered according to the reduced binary features.

Linking substructures to mechanism of action in Stokes et al. dataset. Figure 5 shows the mechanism of action of molecules containing at least one of the five most important simple rings according to the logistic regression model. For the majority of molecules with similar bioactive moiety, the mode of action is the same. For example beta-lactam rings (shown in blue), are present in antibiotics such as penicillins and cephalosporins, and they have been reported to prevent cell wall synthesis^{25,26}. The majority of the molecules with this ring are mapped to the cell wall inhibition (G1) mechanism of action. In cases when molecules with the same moiety are mapped to multiple mechanisms of action, those mechanisms of action are usually similar. For example, for cyclohexane (shown in purple) associated mechanisms of action are bacterial 30S ribosomal subunit inhibitor (G3) and protein synthesis inhibitor (G6), both related to inhibiting protein synthesis. For moiety 4-quinolone (shown in green), the associated modes of action are HDAC inhibitor (G18), DNA gyrase

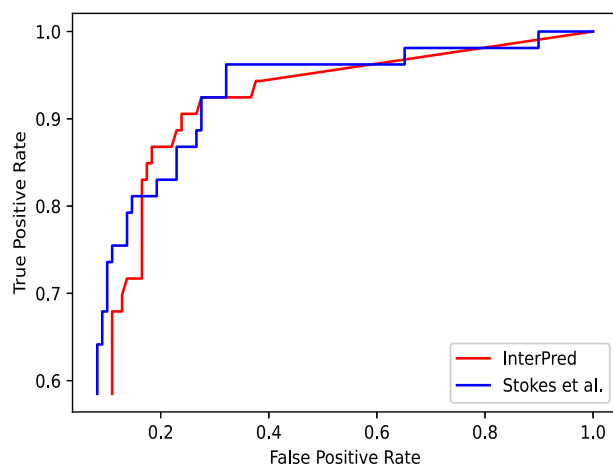


Figure 3. ROC curve for neural network model from Stokes et al. and InterPred. For false positive rates greater than 0.3, the models have nearly identical true positive rate.

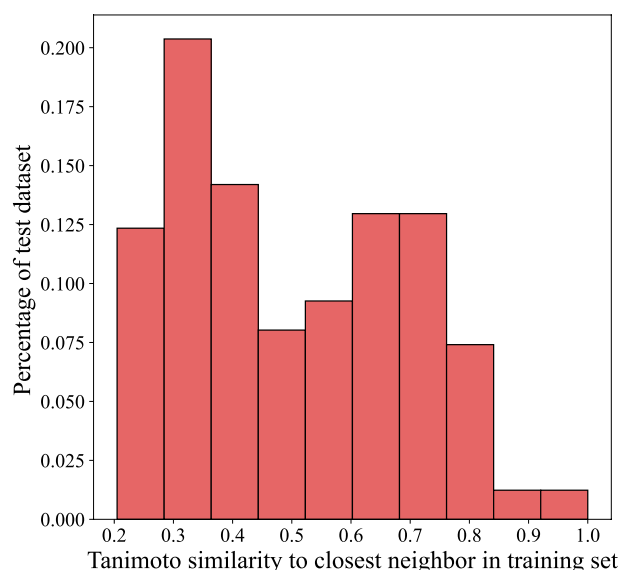


Figure 4. Distribution of the tanimoto similarity between each test data point and its nearest neighbor in the training dataset. The average tanimoto similarity between test data points and their closest neighbors is 0.5035 and the standard deviation is 0.18. Only 1.2% of test data points are more than 90% similar to a training data point.

inhibitor (G2), and topoisomerase inhibitor (G7), which are all related to inhibition of bacterial nucleic acid synthesis. Molecules containing 4-quinolone are known to inhibit bacterial nucleic acid synthesis by disrupting the enzymes topoisomerase IV and DNA gyrase²⁷. In cases where two molecules contain distinct bioactive moieties, they usually have distinct mechanisms of action. The only exceptions G13 and G17 can be explained by the fact that MAP kinases are a subset of Serine/Threonine Kinases²⁸. Among all the pairs of molecules with the same mechanism of action, 76% are clustered together by MOACluster, and among all the pairs of molecules clustered together by MOACluster, 67.6% have identical and 71% have similar mechanisms of action.

Linking substructures to mechanism of action in CO-ADD dataset. Figure 6 shows the ROC curve of InterPred for prediction for growth inhibition for seven different pathogens in the CO-ADD dataset^{23,24}. Figure 7 shows the most dominant bio-active moieties detected by InterPred. The pathogens that are predicted to be inhibited by each moiety are also shown. It has been reported that guanadine^{29–31} and nitro³² are the bioactive moiety in various antibacterial molecules. Moreover hydrazone/hydarazine have been reported to be potent against *S. aureus*, *A. baumannii*, and *C. albicans*^{33,34}.

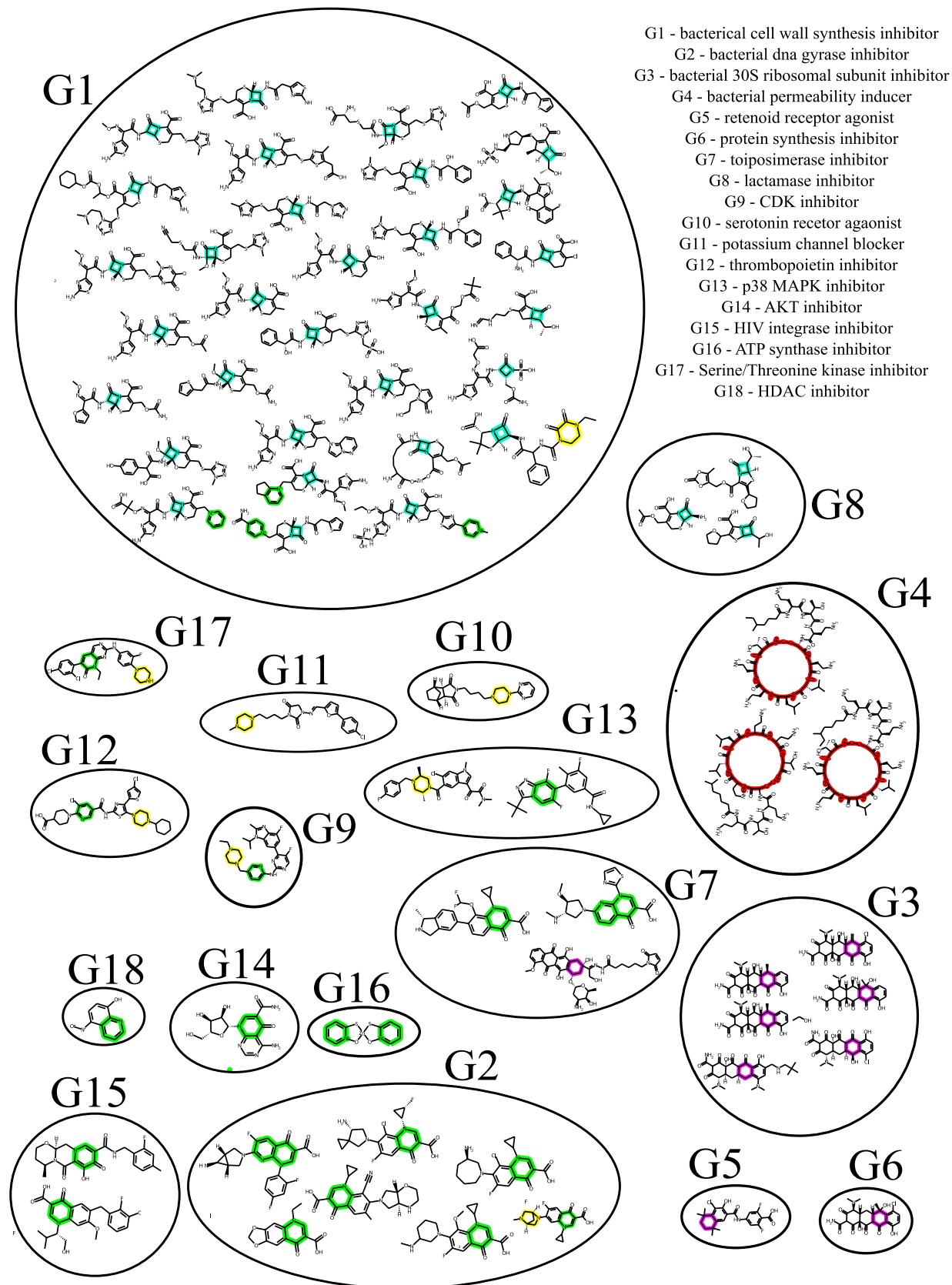


Figure 5. Mechanism of action of molecules containing at least one of the five most important simple rings according to the logistic regression model. Each of the five rings are highlighted with a different color. Molecules sharing the same mechanism of action, as reported in the Drug Repurposing Hub, are further circled together. For the majority of molecules with similar bioactive moiety, the mode of action is the same.

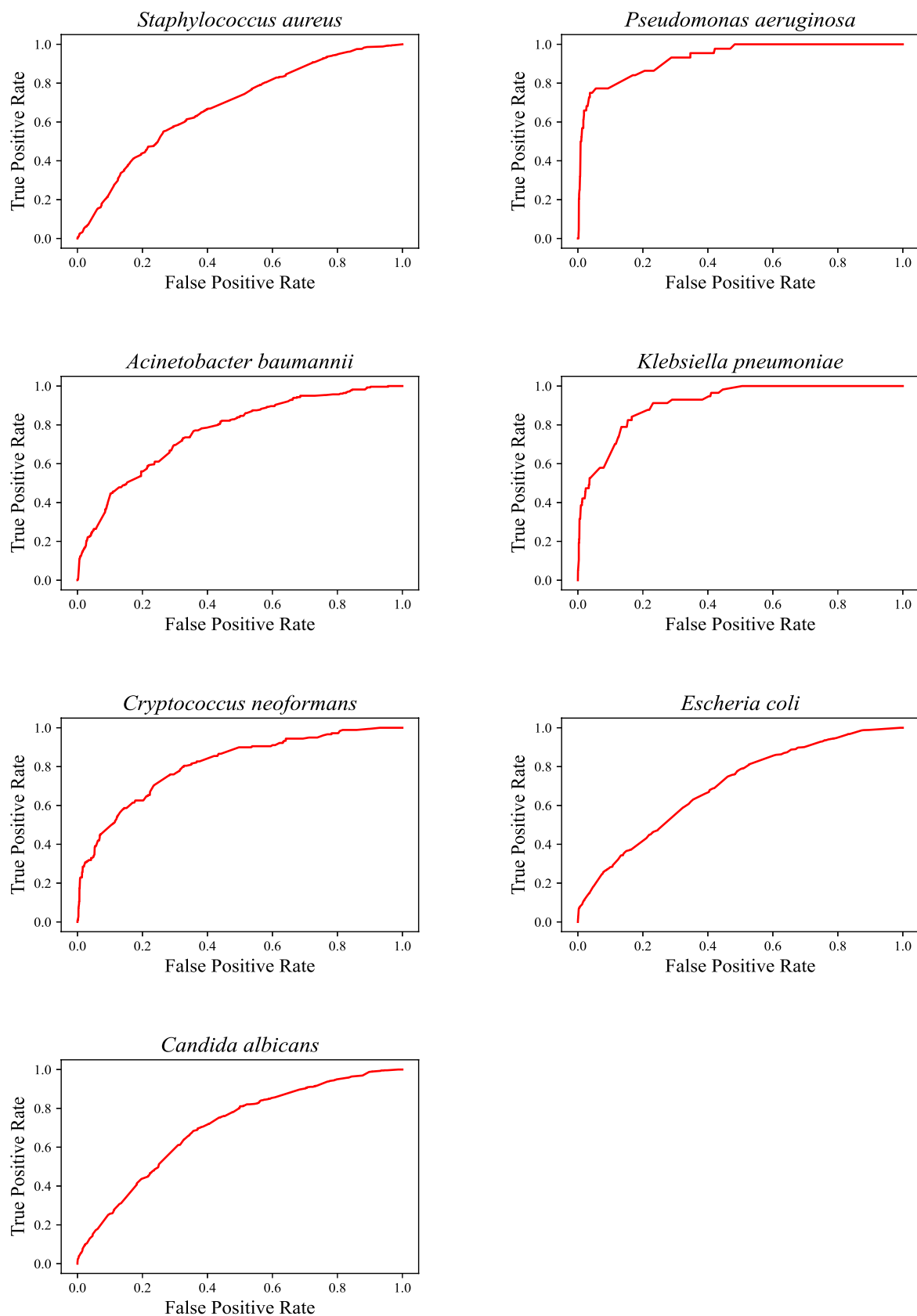


Figure 6. ROC curve of InterPred for prediction of growth inhibition for 7 different bacteria in the CO-ADD dataset.






















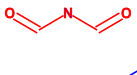





	<i>Staphylococcus aureus</i>	<i>Pseudomonas aeruginosa</i>	<i>Acinetobacter baumannii</i>	<i>Klebsiella pneumoniae</i>	<i>Cryptococcus neoformans</i>	<i>Escheria coli</i>	<i>Candida albicans</i>
 <chem>O=[N+]([O-])R</chem> nitro <chem>C1=NNC1</chem>	X		X				
 <chem>C1=CNCC1</chem>	X						
 <chem>C1=CNCC1</chem> amidine	X						
 <chem>C(=O)R</chem> carbonyl	X					X	
 <chem>Cl-CH2-R</chem> haloalkane	X	X				X	
 <chem>HO-R</chem> hydroxyl	X						X
 <chem>C1=CNC=CC1</chem> guanidine	X	X	X	X	X		
 <chem>C1=COCC1</chem>		X	X				
 <chem>C1=CNCCC1</chem>		X	X	X	X		
 <chem>C1CNCCN1</chem>		X					
 <chem>NR2</chem> amine		X		X			
 <chem>C1=CCNC=C1</chem> hydrazine		X					X
 <chem>C1=CCC=CC1</chem>			X				
 <chem>C1CCCCC1</chem>			X				
 <chem>C1CC[N+]C1</chem>				X			
 <chem>C1=CNCS1</chem>				X	X		
 <chem>C1CCNC1</chem>				X		X	
 <chem>C1=CCCC1</chem>				X			
 <chem>C1CNCC1</chem>					X		
 <chem>C1=CCCCC1</chem>					X		
 <chem>C1=CCCC=C1</chem>					X		
 <chem>O=C1NC(=O)CC1</chem> imide						X	X
 <chem>C1=CCOC1</chem> imine						X	X
 <chem>C1CCNCC1</chem>						X	X
 <chem>C1COCCN1</chem>						X	
 <chem>C1COCO1</chem>							X
 <chem>RC(=O)OR</chem> ester							X

Figure 7. Top 31 ring/functional group features predicted to govern the mechanism of action of molecules along with pathogens they inhibit.

Discussion

Fighting antibiotic resistance requires discovery of small molecules with novel mechanism of action. Recent advances in high-throughput screening have made it possible to collect large databases of small molecules and their activities against resistant pathogens. Moreover, neural network models have been developed to predict the antimicrobial activity of molecules by training on these large datasets. However, the existing models are largely uninterpretable, and do not provide any insight about the mechanism of action of small molecules, and their bioactive moieties. Therefore, it remains unclear which of the molecules among their predictions has novel activity, and should be pushed toward pre-clinical trials.

One of the reasons graph neural network methods remain uninterpretable is that there are millions of molecular substructures and it is impossible to correctly identify which substructures are correlated to bioactivity when there are only a few thousand training data. By focusing on simple ring structures and other substructures that are known to be biologically important, InterPred bypasses this issue. When trained on bioactivity data of several thousand molecules from an FDA-approved library against *E. coli*, InterPred identified five important ring structures, each mapping to a distinct set of modes of action.

As massive public datasets of small molecule bioactivities are becoming available, it has become impossible to investigate all the molecules in these datasets that show interesting activities. Therefore, methods for mining these datasets and prioritizing for molecules with unseen mechanisms of actions are highly desired. InterPred is a fully interpretable approach for predicting the bioactivity of small molecules from training data and retains the same accuracy as the state of the art neural network approach. Additionally, InterPred finds chemical moieties in small molecule datasets that are responsible for bioactivity. Since molecules with novel bioactive moieties usually possess novel modes of action, if a molecule contains a previously unreported moiety that InterPred determines to be bioactive, the molecule could be prioritized for follow up studies.

Methods

Outline of InterPred algorithm. InterPred is an interpretable machine learning algorithm for prediction of bioactivity, functional groups responsible for bioactivity, and mechanism of action by training on data. Below we describe various steps of the InterPred algorithm.

Extracting molecular features. Presence of simple rings are extracted using open source package rdkits³⁵ by finding symmetrized smallest set of smallest rings. Additionally the presence 32 functional groups are extracted by checking whether each molecule has a graph substructure matching the functional group using the descriptors module in RDKit. These substructures are deduplicated using kekulized canonical SMILES³⁶. Since small molecules have only a few simple rings, feature vectors for each molecules usually only have a few non-zero entries.

Training. Both the extra trees ensemble classifier and logistic regression model with ℓ_1 norm regularization are trained on training data and hyperparameters are optimized via five-fold cross validation. The number of trees in the extra trees model was cross-validated for the numbers 10, 40, 70, 100, 130, and 160. The lambda parameter for ℓ_1 -regularized logistic regression was cross-validated for values $\frac{1}{.01}$, $\frac{1}{.05}$, $\frac{1}{.1}$, $\frac{1}{.15}$, and $\frac{1}{.2}$. In logistic regression and extra trees, the loss function is of the form

$$\min \sum_{t=1}^T L(f(\mathbf{x}^t), y^t) \quad (1)$$

where t is used as an index for each training point, y^t represents the true label of each molecule in the training dataset, \mathbf{x}^t represents the features of each molecule, f is a function with range $[0,1]$, and L refers to a loss function that is low when $f(\mathbf{x}^t)$ is close to y^t and high otherwise. y^t takes on value 1 if molecule t inhibits bacterial growth and 0 otherwise. In logistic regression, $f(\mathbf{x}^t) = \text{Sigmoid}(\mathbf{c}^T \mathbf{x}^t)$. \mathbf{c} is the coefficient vector of logistic regression and $\text{Sigmoid}(z) = \frac{1}{1 + \exp(-z)}$. $L(f(\mathbf{x}^t), y^t) = \text{CrossEntropy}(f(\mathbf{x}^t), y^t) + \lambda \frac{\|\mathbf{c}\|_1}{T}$ where $\text{CrossEntropy}(y, \hat{y}) = -y \log(\hat{y}) - (1 - y) \log(1 - \hat{y})$ and λ is a regularization parameter optimized via cross validation. In extra trees $f(\mathbf{x}^t)$ is either 0 or 1 and is determined by the majority label produced by all the trees in the extra trees ensemble. $L(f(\mathbf{x}^t), y^t)$ is 0 if $f(\mathbf{x}^t)$ and y^t are not the same and 1 otherwise.

In the training set introduced by Stokes et al., nearly 95% of the molecules do not have antibacterial activity. Such an imbalance could result in misclassification of bioactive molecules as inactive. To avoid this, we use a “balanced” approach³⁷. We modify the objective function in (1) to the following:

$$\min \sum_{t=1}^T \frac{L(f(\mathbf{x}^t), y^t)}{b^t} \quad (2)$$

where b^t is the number of training points with label y^t . This way bioactive and inactive molecules will contribute to the training nearly equally.

Identifying bioactive moieties. Substructures corresponding to largest positive coefficients of the logistic regression model are reported as bioactive moieties.

Clustering mechanism of action. Since the logistic regression model is trained with ℓ_1 regularization, only a few coefficients are non-zero in the model. InterPred algorithm first reduces the feature vector for each

molecule to these non-zero features, and then molecules with identical reduced feature vectors are assigned to the same cluster.

Data availability

The results present in this study are available from <https://gitlab.com/mongolicious/interpretableml-for-mechanism-of-action/>. The datasets analyzed in Figs. 3, 4, and 5 are available at <https://www.sciencedirect.com/science/article/pii/S0092867420301021>. The datasets analyzed in Figs. 6 and 7 are available at <https://db.co-add.org/>.

Received: 29 January 2022; Accepted: 2 June 2022

Published online: 20 June 2022

References

- World Health Organization *et al.* Lack of new antibiotics threatens global efforts to contain drug-resistant infections. (2020).
- Neill, J. O. Antimicrobial resistance: tackling a crisis for the health and wealth of nations the review on antimicrobial resistance chaired. In *Review Paper-Tackling a Crisis for the Health and Wealth of Nations*, 1–20 (HM Government Wellcome Trust, 2014).
- Pence, H.E., & Williams A. Chempid: an online chemical information resource (2010).
- Kim, S. *et al.* Pubchem substance and compound databases. *Nucl. Acids Res.* **44**(D1), D1202–D1213 (2016).
- Irwin, J. J. & Shoichet, B. K. Zinc—A free database of commercially available compounds for virtual screening. *J. Chem. Inf. Model.* **45**(1), 177–182 (2005).
- Ashby, J. The value and limitations of short-term genotoxicity assays and the inadequacy of current cancer bioassay chemical selection criteria. *Ann. N. Y. Acad. Sci.* **534**(1), 133–138 (1988).
- King, R. D., Muggleton, S., Lewis, R. A. & Sternberg, M. Drug design by machine learning: The use of inductive logic programming to model the structure-activity relationships of trimethoprim analogues binding to dihydrofolate reductase. *Proc. Natl. Acad. Sci.* **89**(23), 11322–11326 (1992).
- Hirst, J. D., King, R. D. & Sternberg, M. J. Quantitative structure-activity relationships by neural networks and inductive logic programming. I. The inhibition of dihydrofolate reductase by pyrimidines. *J. Comput.-Aided Mol. Des.* **8**(4), 405–420 (1994).
- Bronstein, M. M., Bruna, J., LeCun, Y., Szlam, A. & Vandergheynst, P. Geometric deep learning: Going beyond Euclidean data. *IEEE Signal Process. Mag.* **34**(4), 18–42 (2017).
- Stokes, J. M. *et al.* A deep learning approach to antibiotic discovery. *Cell* **180**(4), 688–702 (2020).
- Dai, H., Dai, B., & Song, L. Discriminative embeddings of latent variable models for structured data. In *International Conference on Machine Learning*, 2702–2711 (PMLR, 2016).
- Yang, K. *et al.* Analyzing learned molecular representations for property prediction. *J. Chem. Inf. Model.* **59**(8), 3370–3388 (2019).
- De, S. K. *et al.* Design, synthesis, and structure-activity relationship of substrate competitive, selective, and in vivo active triazole and thiazole inhibitors of the c-jun n-terminal kinase. *J. Med. Chem.* **52**(7), 1943–1952 (2009).
- Chen, X.-W., & Jeong, J. C. Enhanced recursive feature elimination. In *Sixth International Conference on Machine Learning and Applications (ICMLA 2007)*, 429–435 (IEEE, 2007).
- Kursa, M. B. & Rudnicki, W. R. Feature selection with the Boruta package. *J. Stat. Softw.* **36**, 1–13 (2010).
- Tibshirani, R. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B Methodol.* **58**(1), 267–288 (1996).
- Pope, P.E., Kolouri, S., Rostami, M., Martin, C. E., & Hoffmann, H. Explainability methods for graph convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10772–10781 (2019).
- Kojima, R. *et al.* kgcn: A graph-based deep learning framework for chemical structures. *J. Cheminform.* **12**(1), 1–10 (2020).
- Baldassarre, F., & Azizpour, H. Explainability techniques for graph convolutional networks. *CoRR*, Vol. abs/1905.13686, (2019).
- Montavon, G., Lapuschkin, S., Binder, A., Samek, W. & Müller, K.-R. Explaining nonlinear classification decisions with deep Taylor decomposition. *Pattern Recogn.* **65**, 211–222 (2017).
- Ying, R., Bourgeois, D., You, J., Zitnik, M., & Leskovec, J. Gnn explainer: A tool for post-hoc explanation of graph neural networks. arXiv preprint [arXiv:1903.03894](https://arxiv.org/abs/1903.03894), (2019).
- Corsello, S. M. *et al.* The drug repurposing hub: A next-generation drug library and information resource. *Nat. Med.* **23**(4), 405–408 (2017).
- Zuegg, J., Hansford, K. A., Elliott, A. G., Cooper, M. A. & Blaskovich, M. A. How to stimulate and facilitate early stage antibiotic discovery. *ACS Infect. Dis.* **6**(6), 1302–1304 (2020).
- Blaskovich, M. A., Zuegg, J., Elliott, A. G. & Cooper, M. A. Helping chemists discover new antibiotics. *ACS Infect. Dis.* **1**(7), 285–287 (2015).
- Reeve, S. M., Lombardo, M. N. & Anderson, A. C. Understanding the structural mechanisms of antibiotic resistance sets the platform for new discovery. *Future Microbiol.* **10**(11), 1727–1733 (2015).
- Cho, H., Uehara, T. & Bernhardt, T. G. Beta-lactam antibiotics induce a lethal malfunctioning of the bacterial cell wall synthesis machinery. *Cell* **159**(6), 1300–1311 (2014).
- Pham, T. D., Ziora, Z. M. & Blaskovich, M. A. Quinolone antibiotics. *MedChemComm* **10**(10), 1719–1739 (2019).
- Pearson, G. *et al.* Mitogen-activated protein (MAP) kinase pathways: Regulation and physiological functions. *Endocr. Rev.* **22**(2), 153–183 (2001).
- Fleeman, R. *et al.* Combinatorial libraries as a tool for the discovery of novel, broad-spectrum antibacterial agents targeting the ESKAPE pathogens. *J. Med. Chem.* **58**(8), 3340–3355 (2015).
- Wang, B. *et al.* Antibacterial diamines targeting bacterial membranes. *J. Med. Chem.* **59**(7), 3140–3151 (2016).
- Qian, L., Guan, Y., He, B. & Xiao, H. Modified guanidine polymers: Synthesis and antimicrobial mechanism revealed by AFM. *Polymer* **49**(10), 2471–2475 (2008).
- Olender, D., Żwawiak, J. & Zaprutko, L. Multidirectional efficacy of biologically active nitro compounds included in medicines. *Pharmaceuticals* **11**(2), 54 (2018).
- Whitt, J. *et al.* Synthesis of hydrazone derivatives of 4-[4-formyl-3-(2-oxochromen-3-yl) pyrazol-1-yl] benzoic acid as potent growth inhibitors of antibiotic-resistant *Staphylococcus aureus* and *Acinetobacter baumannii*. *Molecules* **24**(11), 2051 (2019).
- Lv, Q.-Z. *et al.* A new antifungal agent (4-phenyl-1, 3-thiazol-2-yl) hydrazine induces oxidative damage in *Candida albicans*. *Front. Cell. Infect. Microbiol.* **10**, 557 (2020).
- Landrum, G. *et al.* Rdkit: Open-source cheminformatics. (2006).
- Weininger, D. Smiles, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **28**(1), 31–36 (1988).
- He, H. & Ma, Y. *Imbalanced Learning: Foundations, Algorithms, and Applications* (Wiley, 2013).

Acknowledgements

The work of M.M., M.G., and H.M was supported by a National Institute of Health Award DP2GM137413.

Author contributions

M.M. and M.G. designed and implemented the model. M.M. performed benchmarking experiments and analyzed the data. H.M. supervised the project. All authors contributed in writing the manuscript.

Competing Interests

H.M. is a co-founder and has equity interest from Chemia Biosciences. The other authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to H.M.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2022