This is a pre-copyedited, author-produced version of an article accepted for publication in J Mol Model following peer review. The version of record Jagdev, R., Madsen, T.B. & Finn, P.W. On the ability of machine learning methods to discover novel scaffolds. J Mol Model 29, 22 (2023) is available online at: https://doi.org/10.1007/s00894-022-05359-6

On the ability of machine learning methods to discover novel scaffolds

Rishi Jagdev[1*], Thomas Bruun Madsen[2†] and Paul W. Finn[1,3†]

[1*]School of Computer Science, University of Buckingham, Hunter Street, Buckingham, MK18 1EG, UK.
[2]University of West London, St Mary's Road, Ealing, London, W5 5RF, UK.
[3]Oxford Drug Design Ltd., Oxford Centre for Innovation, New Road, Oxford, OX1 1BY, UK.

*Corresponding author(s). E-mail(s): rishi.jagdev@buckingham.ac.uk;
Contributing authors: thomas.Madsen@uwl.ac.uk; paul.finn@buckingham.ac.uk;
†These authors contributed equally to this work.

Abstract

The recent advances in the application of machine learning to drug discovery have made it a "hot topic" for research, with hundreds of academic groups and companies integrating machine learning into their drug discovery projects. Nevertheless, there remains great uncertainty regarding the most appropriate ways to evaluate the relative performance of these powerful methods against more traditional cheminformatics approaches, and many pitfalls remain for the unwary. In 2020, researchers at MIT [Stokes, J.M., Yang, K., Swanson, K., Jin, W., Cubillos-Ruiz, A., Donghia, N.M., MacNair, C.R., French, S., Carfrae, L.A., Bloom-Ackermann, Z., et al.: A deep learning approach to antibiotic discovery. Cell 180(4), 688–702(2020)] reported the discovery of a new compound with antibacterial activity, halicin, through the use of a neural network machine learning method. A robust ability to identify new active chemotypes through computational methods would be very useful.

In this study, we have used the Stokes et al. dataset to compare the performance of this method to two other approaches, Mapping of Activity Through Dichotomic Scores (MADS) by Todeschini et al. [Todeschini R, Consonni V, Ballabio D, et al (2018) Mapping of activity through dichotomic scores (mads): A new chemoinformatic approach to detect activity-rich structural regions. Journal of Chemometrics 32(4):e2994] and Random Matrix Theory (RMT) by Lee et al. [Lee AA, Yang Q, Bassyouni A, et al (2019) Ligand biological activity predicted by cleaning positive and negative chemical correlations. Proceedings of the National Academy of Sciences 116(9):3373–3378)]. Our results demonstrate that all three methods are capable of predicting halicin as an active antibacterial compound, but that this result is dependent on the dataset composition, pre-processing and the molecular fingerprint used. We have further assessed overall performance as determined by several performance metrics.

We also investigated the scaffold hopping potential of the methods by modifying the dataset by removal of the β-lactam and fluoroquinolone chemotypes. MADS and RMT are able to identify actives in the test set that contained these substructures. This ability arises because of high scoring fragments of the withheld chemotypes that are in common with other active antibiotic classes. Interestingly, MADS is relatively better compared to the other two methods based on general predictive performance.

Keywords: Ligand Based Virtual Screening, Antibiotics, Machine Learning Algorithms, Deep Neural Network

Introduction

Drugs to treat bacterial infections have made an enormous contribution to the improvements in human health and lifespan over the past 100 years [1]. However, to maintain effectiveness, the regular delivery of new antibacterial drugs is required, because bacteria evolve over time to develop resistance to the agents used. For the last several decades, few new antibacterial drugs have reached the market, and those that have are mostly members of existing antibiotic classes rather than having novel mechanisms of action [2]. This has led to the much publicized concern regarding anti-microbial resistance (AMR) and the real possibility of the widespread emergence of infections for which there are no effective treatments. The underlying reasons for the poor state of the antibacterial development pipeline are complex, but at a high level are a combination of lack of economic incentives to develop new antibacterials and a high degree of technical difficulty [3]. As a result, the need for new antibacterials is now urgent. There has been a recent resurgence in preclinical antibacterial discovery, with significant interest in applying the emerging

techniques of machine learning to this vital problem.

Deep learning (DL) uses artificial neural networks with multiple non-linear processing layers to learn from a given data representation. One major difference between the working of neural networks and conventional machine learning algorithms, apart from the scale and the complexity of the networks used, is their ability to learn directly from molecular structure instead of descriptor representations such as molecular fingerprints, 3D structure or pharmacophores [4]. For example, graph convolution methods [5] have facilitated prediction of drug-target interactions by discovering (learnable) chemical properties of compounds from their graphs. Graphs provide spatial information about higher dimensional molecular features, by modelling the structures of molecules and exploiting the interdependency of one feature on another. In principle, this enables the neural network to learn chemical properties themselves. As a result, deep learning continues to be a focus of research in cheminformatics, and other areas such as molecular dynamics, chemical physics, genomics studies and bioinformatics.

Although predecessors of deep learning have been applied to drug discovery since the 1990's [6], interest in their use has grown significantly in recent years, owing to their success in QSAR applications [7], [8], [9] and predictive toxicology [10]. Another stimulus has been the increase in publicly available bioactivity datasets upon which to tune these models and increased computational power. Previous studies have investigated the performance of deep learning techniques in cheminformatics, many including a comparison with conventional machine learning techniques. Lenselink *et al.* [11] evaluated the performance of diverse machine learning techniques with a deep learning method using a standardized ChEMBL [12] dataset. While the DL model gives the best overall performance, random forests (added with proteochemometrics descriptors) and support-vector machines are a close second with comparable Boltzmann-Enhanced Discrimination of ROC (BEDROC) [13] and higher Matthews Correlation Coefficient (MCC) values. Similarly, Koutsoukas *et al* [14] performed an extensive comparison of feed forward deep neural networks with classical machine learning techniques, where the DL technique performed better than ML techniques for a majority of diverse targets. Another study by Duvenaud *et al.* [15] describes a convolutional neural network graph for learning molecular fingerprints, where it takes a molecular graph as input and extracts molecular features. Even though the neural network model performs as well as, or better than, conventional approaches that use a circular fingerprint representation, various disadvantages are also discussed. These include computation cost, limited information propagation through graphs and inability to distinguish stereoisomers, among others. Recently, augmented message passing neural networks that also use graph-derived chemical properties for bioactivity prediction of ligands were introduced by Withnall *et al.* [16], who benchmarked their approach using eight different chemical datasets

taken from the literature. Comparison of classical machine learning and deep learning based methods demonstrated roughly equivalent performance.

However, in some studies, classical techniques perform better than neural networks. For instance, recently Jiang *et al.* [17] conducted a study on 11 diverse public datasets to compare the performance of graph neural networks [5] with descriptor (1-D, 2-D and fingerprints) based machine learning models. Considering both prediction accuracy and computational efficiency, the descriptor based ML models performed better for the most part. Although some studies investigate error frameworks for deep learning architectures, it remains a challenge to quantify the confidence intervals of activity prediction by these models [18]. A bigger concern, given the "black box" nature of the methods, is the limited ability to interpret the results, and therefore uncertainty about the reasons for assigning inactivity/activity for a given ligand. Other concerns include the vast number of chemical features available, which could lead to highly correlated features, and a potential for overfitting [19].

Given the unpredictable performance of machine learning and deep learning techniques in computational drug discovery, preference in future will be given to methods that have consistent scaffold hopping potential across multiple molecular classes [20]. 'Scaffold hopping' is the process of identifying compounds with different molecular backbones, but similar activity/property relations. Schneider *et. al* in [21] have used a pharmacophore based virtual screening technique for identifying novel scaffolds, while preserving the biological activity. Other techniques that have been suggested in this area are replacement of motifs in active compounds [22–24]. In this study, we have investigated the potential of computational methods for detecting novel scaffolds by removing their respective active chemotypes from the training datasets. We compare the performance of a deep learning method, with two machine learning techniques that are based on mathematical interpretation of the fingerprint matrix, to predict bioactivity of drug data and identify novel scaffolds. We use the dataset described in Stokes *et al.* [25] (see section 1.1). In their paper, they use a deep learning neural network model (Chemprop) to identify a compound with previously unknown antibiotic activity, which they name "halicin". This compound is structurally dissimilar to known antibiotics. Briefly, Chemprop learns molecular features directly from simplified molecular input line entry system (SMILES) of the molecules fed as the input layer, iteratively aggregating the features of atoms and bond paths by applying a directed bond-based message passing approach. The chemical structure data was augmented by RDKit molecular features. The dataset and deep neural network (DNN) method are publicly available, enabling comparison with other methods. Also, the biological data for the training set and for the predicted actives were generated using the same assay, removing some of the issues associated with training on literature data assembled from multiple sources. We compare the performance of Chemprop with two classification

algorithms for ligand-based virtual screening: Mapping of Activity through Dichotomic Scores (MADS) by Todeschini *et al.* [26] and Random Matrix Theory (RMT) by Lee *et al.* [27] respectively.

Methods

1.1 Datasets

Training Dataset

Stokes *et al* used a training dataset of 1760 pharmaceutical drugs approved by the FDA (Food and Drug Administration) and 800 natural products. They removed duplicates (methodology not specifically described) which resulted in 2335 compounds. The training data was classified as 'active' and 'inactive' using 80% growth inhibition against *E. coli* as a hit cut-off from the experimental assay. The dataset was downloaded from the publication website [25]. When used without any modification it is referred to as TD-Original.

However, visual inspection of data identified potential issues with the training data. These included lack of stereochemistry, presence of salts or other additional components and lack of consistency in representation. The above inconsistencies also lead to duplicates in the dataset, where the same parent (drug) compound is present in multiple entries. Therefore, compounds were standardized using the sdwash functionality of the MOE software [28]. Salts and other disconnected components were removed and ionizable groups were neutralized for consistency. Any duplicates arising from the standardization were removed, resulting in here are 2299 unique records in the processed file, TD-Cleaned.

To assess the effects of removing a structural class of compounds and thus explore the ability of the methods to identify activity in a chemical class not present in the training set, all training dataset molecules containing the β-lactam ring (which is the key component of penicillin and cephalosporin antibiotics, among others) were removed from the cleaned datasets. This removed 57 compounds from the training set, 28 of which were active in the *E. coli* growth inhibition assay. This dataset is named TD-NoBL. Similarly, in a different experiment, we removed all compounds containing the fluoroquinolone structure in the training set. This removed 20 compounds, all of which were active. We call this training set TD-NoFQ. Both datasets are provided in the supplementary information. Both β-lactam and fluoroquinolone compounds were removed from the training set by identifying their presence in the SMILES strings of compounds.

Test Dataset

The test dataset is derived from the Broad Institute's Drug Repurposing Hub library consisting of 4,496 molecules [29]. The dataset was downloaded from the publication website [25].

As with the training set, consistency issues for stereochemistry and salt representation were identified with the test set. However, these were different from those of the training set. Stereochemistry is present in this set and most, but not all, salts have been removed. For consistency with the training set, stereochemistry was removed by processing the MOE's SMILES structure representation using a sed script and then standardizing with sdwash as for the training set. The number of duplicates found were 52 in total, which were removed, leaving 4,444 unique structures.

To evaluate the performance of Chemprop in the original publication, the activity of 162 molecules was empirically tested in the *E.coli* assay. This allows comparison of a subset of molecules from the test set for which activity is experimentally known. In the subsequent experiments described in this paper, we use the 162 compounds for comparison of performance of Chemprop, MADS and RMT.

Molecular Descriptors

Molecular descriptors are an abstract representation of physicochemical and other properties of a molecule, generated by computational algorithms. They are widely used in ligand based virtual screening, clustering and similarity studies [30–33]. In this study, we use the 2048 bit RDKit-Morgan (radius 2) fingerprints, RDKit's implementation of the Extended-Connectivity Fingerprint (ECFP) that represent circular atom neighborhoods based on a user defined radius for the ML algorithms because of the ease of computation and performance [34–37]. We use 2048 bits to reduce bit collision.

Previous studies have suggested that the performance of a particular fingerprint for determining activity is largely dependent on the dataset, robustness of the model, and performance metrics used [33]. For comparison with other fingerprint types, we also discuss the performance of dictionary-based (MACCS keys [38]) and path-based (topological torsion fingerprint [39], Avalon fingerprint, RDK fingerprint [40]) fingerprints in addition to the RDKit Morgan fingerprints as representatives of different types.

Machine Learning Models

We investigated the performance of three machine learning methods, the DNN method Chemprop and two machine learning methods that focus on featurization of molecular structures using fingerprint incidence matrices to predict activity. All the methods described below have been implemented by us, to analyse and compare performance.

Chemprop DNN

To facilitate comparison between the methods and enable investigation of dataset composition, we implemented Chemprop internally and recreated model generation using the original dataset and the description provided in the original study. The experiments were conducted on an intel Core i5-7200U processor with 8 GB RAM. The models takes SMILES structures, augmented with 200 physicochemical RDKit descriptors [25, Supplementary Table S2A] as input. Bayesian optimization was used to choose the best set of hyperparameters. An ensemble of models was then trained on 20 folds to predict molecular activity of compounds in the test set. The model ranked the test set based on predicted scores, and the top 99 compounds and bottom 63 (162 in total) were tested empirically. 51 of the predicted actives and 2 of the predicted inactives showed activity in the *E. Coli* assay.

Mapping of Dichotomic Scores

Todeschini *et al* described a method, 'Mapping of Activity through Dichotomic Scores' [26] (MADS), that calculates an activity score for each compound on the basis of its substructures. Simultaneously, these substructures are given 'weights' on the basis of their contribution. An interesting aspect of this approach is that, unlike classical weighting schemes, the MADS approach considers the interactions between pairs of substructures, i.e. their frequencies of co-occurrence in the molecules. The basis of the method is as follows:

Given a sample of N fingerprints with p features, we record this as an $N \times p$ (binary) data matrix

$$\mathbf{X} = (x_{ij}) \in \mathbb{R}^{N \times p}$$

where entry x_{ij} equals 1 if fingerprint $i \in \{1, \ldots, N\}$ contains substructure $j \in \{1, \ldots, p\}$, and 0 otherwise. In the following, two data matrices will be constructed from either known active compounds only or known inactive compounds only.

The method used by MADS for extracting information from this data matrix is by forming a so-called scatter matrix:

$$\mathbf{S} = \frac{1}{N}\mathbf{X}^t\mathbf{X}. \tag{1}$$

The diagonal entries of the square matrix $\mathbf{S}$ are the relative frequencies of occurrences of a given substructure in the sample of fingerprints. The off-diagonal entries of $\mathbf{S}$ gives the relative frequencies of co-occurrences of a given pair of substructures.

Again following [26], we associate with the scatter matrix $\mathbf{S}$, the activity (or inactivity) score w_i associated with substructure i, given by summing over the ith row (or equivalently, column) $\mathbf{S}$:

$$w_i = \sum_{j=1}^{p} S_{ij} = \sum_{j=1}^{p} S_{ji},$$

for $1 \leqslant i \leqslant p$. Finally, given any molecule represented by the relevant type of fingerprint $\mathbf{x} \in \mathbb{R}^p$, we compute its activity (or inactivity) score $h_{\mathbf{x}}$ (where h_{AC} is activity score, h_{IN} is inactivity score) as :

$$h_{\mathbf{x}} = \sqrt{\mathbf{x}^t \mathbf{S} \mathbf{x}}$$

Note that as $\mathbf{S}$ is positive semi-definite, this score will be a non-negative (real) number. Ligands are predicted to be 'actives' if their activity score (calculated by the difference of h_{AC} and h_{IN} is beyond a certain threshold.

Random Matrix Theory (RMT)

From a mathematical viewpoint, following Lee et al. [27, 41], it might be more natural to allow for a mixing of features and decompose according to eigenspaces. In this case, we standardize our data matrix so that each column has zero mean and unit variance; the latter involves removing any columns of $\mathbf{X}$ with no variation. Retaining the notation $\mathbf{X}$ for this standardized data matrix, we form the sample covariance matrix:

$$\mathbf{C} = \frac{1}{N} \mathbf{X}^t \mathbf{X}. \tag{2}$$

For the covariance matrix $\mathbf{C}$, we allow for mixing of features. We begin by diagonalising $\mathbf{C}$. In terms of our chosen basis, $\{\mathbf{e}_i\}$, of eigenvectors corresponding to the eigenvalues, λ_i, of $\mathbf{C}$, this means the covariance matrix can be expressed in diagonal form:

$$\begin{pmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \lambda_p \end{pmatrix}.$$

We focus only on eigendirections $\mathbf{e}_i \in \mathbb{R}^p$ whose eigenvalues, λ_i, are larger than the upper threshhold distinguished by the Marchenko-Pastur distribution. Specifically this threshhold is given by λ_+, where

$$\lambda_{\pm} = (1 \pm \sqrt{p/N})^2.$$

The significant directions $\mathbf{e}_i$, corresponding to eigenvalues $\lambda_i > \lambda_+$, span a subspace Span$\{\mathbf{e}_i\}$ of $\mathbb{R}^p$ and if a molecule is close to this subspace, with respect to the Euclidean distance, it will be deemed as active (or inactive). The ratio $\lambda = p/N$ dictates the form of the Marchenko-Pastur distribution. In general, the probability density is given by the formula:

$$p(x) = \frac{1}{2\pi\lambda x}\sqrt{(\lambda_+ - x)(x - \lambda_-)}$$

for $x \in (\lambda_-, \lambda_+)$, zero otherwise. However, if $\lambda \in (0, 1)$, we also have point mass of $1 - \frac{1}{\lambda}$ at the origin $x = 0$.

When including both the active and inactive subspaces in computations for determining activity, [27] refers to the method as the Random Matrix Discriminant approach. We do not make this distinction: for our computations below, both subspaces have been used simultaneously for our computations. Concretely, if a compound has fingerprint $\mathbf{x}$ and $d_{ac}(\mathbf{x})$ is its Euclidean distance to the subspace of actives and $d_{inac}(\mathbf{x})$ is its Euclidean distance to the subspace of inactives, then it is deemed to be active provided:

$$d_{ac}(\mathbf{x}) < d_{inac}(\mathbf{x}) + \varepsilon.$$

Here, ε is a parameter, which can be interpreted as the trade-off between false positives and false negatives; larger values of ε will lead to more substances being classified as active.

Similarity analysis of training and test data:

The training and test data come from distinct datasets and this could lead to an overestimated performance, as a result of analogue or other biases. To determine any analogue bias present in the datasets, we calculated the Tanimoto similarity [42] of active compounds present in the training set with the active and inactive compounds in the test set. Compounds are represented using the 2048-bit RDKit Morgan (radius 2) fingerprints for similarity calculations. Each active compound in the training set generates a receiver operating characteristic-area under the curve (ROC-AUC) based on its similarity score with every compound in the test set. The reported AUC is the average AUC obtained across the actives.

Comparison Metrics

As activity is inherently rare, drug discovery datasets are usually imbalanced, with far fewer actives than inactives. Accuracy is, therefore, often a poor performance metric. There has been a considerable amount of research in analysing metrics suitable for handling imbalanced data, to combat overestimation of accuracy values [43].

Keeping in mind the difference of the number of actives and inactives in our experiments, we used additional metrics to evaluate the performance of the three models without any bias. We use $F1$-score, Matthews Correlation Coefficient and Balanced Accuracy, all of which can be statistically determined from the two-classed confusion matrix described in Figure 1. The figure demonstrates four basic outputs of a confusion matrix: true positives, false positives, true negatives and false negatives. All other metrics that we discuss subsequently are derived from these.

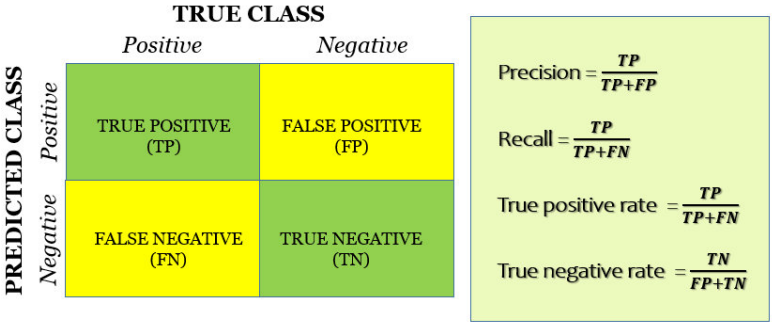


Fig. 1: Confusion matrix derived after classification by a predictive model

F1-score

By definition, this metric is the harmonic mean of the precision, and the recall (Fig 1). Hence, it can be computed through the formula:

$$\mathrm{F1} = \frac{2\mathrm{TP}}{2\mathrm{TP} + \mathrm{FP} + \mathrm{FN}}.$$

This metric can be generalised to multi-class problems, either through micro- or macro-averaging. [44].

Matthews Correlation Coefficient

This is a robust metric that gives equal importance to both positive and negative classes by taking true negatives into account. The value 1 corresponds to a perfect agreement, 0 denotes a random performance, and -1 denotes a perfect disagreement between the predicted and observed values. Matthews Correlation Coefficient is computed from the confusion matrix via the formula:

$$\mathrm{MCC} = \frac{\mathrm{TP} \cdot \mathrm{TN} - \mathrm{FP} \cdot \mathrm{FN}}{\sqrt{(\mathrm{TP} + \mathrm{FP})(\mathrm{TP} + \mathrm{FN})(\mathrm{TN} + \mathrm{FP})(\mathrm{TN} + \mathrm{FN})}}.$$

Balanced Accuracy

This is a metric that accounts for both the positive and negative outcome classes. It is often to used to assess the performance of ligand based virtual screening models [45]. The Balanced Accuracy is defined as the arithmetic mean of true positive rate (TPR), and the true negative rate (TNR), both defined in Figure 1. The range of BA is between 0 and 1, for worst-possible and the best-possible classifier, respectively. It is obtained by the formula:

$$\mathrm{BA} = \frac{1}{2}\left(\mathrm{TPR} + \mathrm{TNR}\right).$$

Results and Discussion

To provide a basis for our comparison, Chemprop, MADS and RMT were first applied to TD-Original. Thereafter, the experiments were conducted using the standardized version of the dataset (TD-Cleaned), followed by those where chemotypes from the datasets have been removed (TD-NoBL and TD-NoFQ).

Results for the Original dataset, TD-Original

Our in-house implementation of Chemprop was used to reproduce the results described in the original publication. Due to the inherent randomness of the method, it is not possible to reproduce the published results exactly. However we were able to obtain receiver operating characteristic -area under the curve (ROC-AUC) of 0.86, which compares well to the ROC-AUC of 0.89 reported by Stokes *et al.* The rank correlation coefficient of the probabilities of activity of our implementation with the original probabilities is 0.92, thereby providing a firm basis for the subsequent experiments.

MADS and RMT are both trained on the 2048 RDKit Morgan (radius 2) fingerprints of the compounds. The result of prediction is evaluated by the AUC score (Figure 2) using the empirically tested 162 compounds, described in section 1.1. MADS gives an AUC of 0.87, comparable to Chemprop. The AUC value obtained by RMT is lower, at 0.82. For other metrics, such as $F1$-score, MCC and BA, the difference in performance is even more significant as seen in Table 1. MADS clearly outperforms Chemprop and, by a larger margin, RMT.

To investigate the potential effect of analogue bias, we compared the performance of the methods with the simple similarity-based technique described in Section 1.1. Using each active molecule as a query, the average AUC returned by this method is 0.66, showing a modest ability to predict activity based on molecular similarity, and thus a limited analogue bias. It also implies that the three methods are indeed more effective in learning abstract molecular information than simple methods for activity prediction.

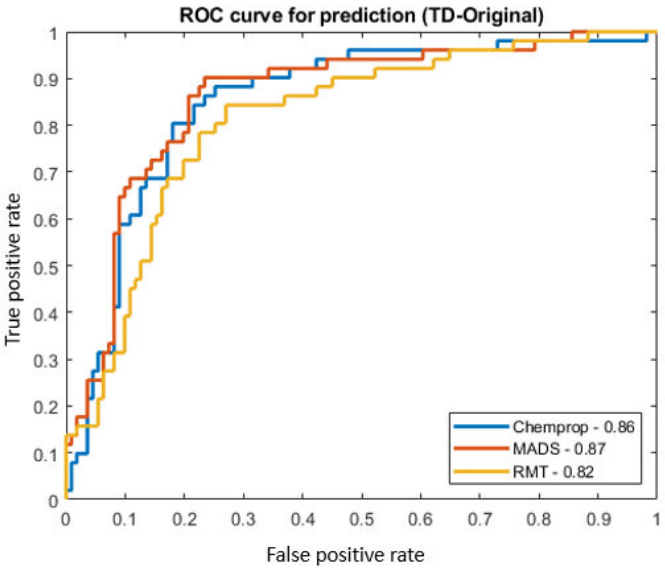


Fig. 2: ROC curve for prediction by Chemprop, MADS and RMT trained on TD-Original

	TD-Original			TD-Cleaned		
Method	MADS	Chemprop	RMT	MADS	Chemprop	RMT
Accuracy	0.80	0.78	0.62	0.80	0.79	0.76
$F1$-Score	0.71	0.68	0.59	0.73	0.68	0.68
MCC	0.53	0.46	0.39	0.60	0.48	0.58
BA	0.81	0.77	0.69	0.82	0.76	0.77
ROC-AUC	0.87	0.86	0.82	0.87	0.87	0.83

Table 1: Performance comparison of MADS, Chemprop and RMT when trained using the original dataset (TD-Original) and standardized dataset (TD-Cleaned)

The normalized density histograms of physicochemical properties of active (120) and inactive (2215) compounds in the training set are provided in the supplementary information (Figure S1). While some molecular property distributions are moderately similar (logP, molecular weight), there are evident differences in distributions of some (topological polar surface area, number of hydrogen donors and acceptors). It is to be noted that while MADS and RMT are trained using molecular fingerprints only, Chemprop is augmented with physicochemical descriptors for classification. These differences in molecular descriptors between the actives and inactives in the training set could

therefore contribute to prediction of activity in this case.

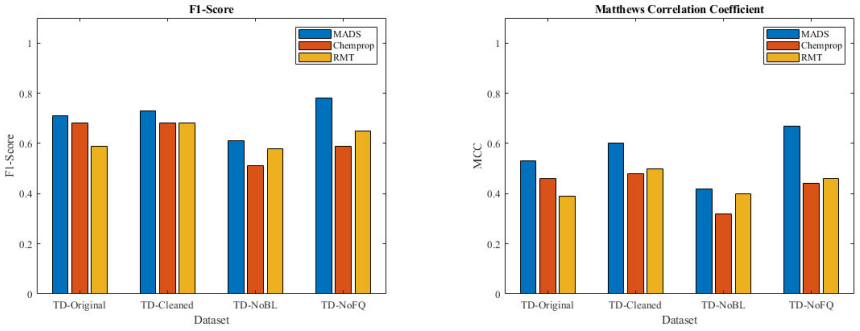
As seen in Table 1 and Figure 3, MADS marginally performs better than Chemprop for the original dataset. The Pearson's correlation of prediction of MADS and Chemprop for the experimentally verified dataset is 0.78, so overall there is moderately good consistency between the methods. The prediction correlation of Chemprop and RMT for the same data is 0.59, consistent with the relatively lower AUC of RMT. It is to be noted that the according to [25], a compound is classified as active against *E. coli*, if it displays growth inhibition based on a cut-off of $OD_{600} < 0.2$. All three methods predict cefminox, imipenem and cefmetazole as actives, all of which are false positives according to this cut-off. However, in previous studies these compounds have all been experimentally proven to have activity against *E. coli* [46–48], so this is a discrepancy in the data.

Results for Standardized dataset, TD-Cleaned

We next investigated the effect of data standardization on performance. The results are displayed in Table 1 and Figure 3. One might anticipate that inconsistencies in the data would impair the ability of the machine learning method to detect activity patterns in the data. However, those inconsistencies could also be used by the methods to improve performance, for example by ascribing activity to the presence of a counterion could boost performance for molecules otherwise difficult to distinguish. In fact, for this dataset, standardization leads to small absolute changes in prediction performance by MADS, RMT and Chemprop. Performance is always improved by standardization. So, in this case, standardization is making the detection of meaningful patterns in the data slightly easier, rather than allowing the machine learning methods to exploit inconsistencies to inflate performance artificially.

Impact of fingerprint type on performance metrics

Molecular fingerprints capture the substructural composition and, to a degree, some of the physicochemical properties of molecules which can be used to predict the activity of compounds. Fingerprints vary in the specific ways in which this information is captured. This can lead to different performance on a particular dataset, and between datasets. To investigate the impact of choice of fingerprint on performance, we trained MADS and RMT on the 5 different molecular fingerprints described in section 1.1 and compared the results. Figure S2 (Supplementary Information) displays the performance of each fingerprint as a box and whisker plot as evaluated by the four performance metrics, averaged across all of the datasets studied.



(a) F1-Score (b) Matthews Correlation Coefficient

Fig. 3: Performance comparison of Chemprop, RMT and MADS using (a) $F1$-Score and (b) Matthews Correlation Coefficient

Variability of fingerprint performance is observed with respect to the the fingerprint and performance measure, but there are also some common patterns.

Topological torsion and RDKit fingerprints perform the best for MADS and RMT respectively. MACCS keys, on the other hand, perform relatively poorly for both methods. It appears that the preselected substructures represented in the MACCS keys are not well suited to this dataset.

Prediction of chemotypes not present in the training set

An important question in the development of predictive models is whether they are able to generalize to predict the activity of chemotypes not present in the training data. We have investigated this in two additional experiments, through the removal of β-lactams and fluoroquinolones from the standardized datasets.

Removal of β-lactams, TD-NoBL

The results are shown in Table 2. As expected, the performance of all three methods declines when β-lactams are removed from the training set. In the 162 compound test set, of the 51 actives, 20 contain the β-lactam ring. The prediction performance of Chemprop declines markedly, as it is now able to correctly predict only 3 β-lactam containing actives: cefoxitin, faropenem and moxalactam in the test set. However, both MADS and RMT are able to correctly predict the activity of relatively more β-lactam containing drugs in the test set, with MADS detecting 15, and RMT predicting 16 of the 20 actives.

Interestingly, removing β-lactams also affects the prediction of other classes of drug by Chemprop. For example, Chemprop is unable to correctly detect any glycopeptide antibiotics in the active test set. Due to more false negatives being predicted in the test set, the values of both MCC and $F1$-score are considerably lower. Chemprop and MADS fail to detect halicin as an active in this experiment, but it remains correctly predicted by RMT.

	TD-NoBL			TD-NoFQ		
Method	MADS	Chemprop	RMT	MADS	Chemprop	RMT
Accuracy	0.74	0.70	0.70	0.85	0.77	0.75
$F1$-Score	0.61	0.51	0.58	0.78	0.59	0.65
MCC	0.42	0.32	0.40	0.67	0.44	0.46
BA	0.72	0.66	0.64	0.85	0.71	0.74
ROC-AUC	0.80	0.73	0.76	0.86	0.80	0.78

Table 2: Performance comparison of MADS, Chemprop and RMT when trained using TD-NoBL and TD-NoFQ)

Removal of fluoroquinolones, TD-NoFQ

To explore the generality of these findings, we performed another experiment; this time removing all fluoroquinolones from the training set. The results are displayed in Table 2.

Chemprop was not able to identify any true active fluoroquinolones in the test set. While MADS is able to predict 4 of the 6 true active fluoroquinolones correctly, RMT predicts all 6. While the prediction of RMT is better than the other methods for true positives, the considerably high number of false positives predicted lowers the overall performance value of RMT.

The similarity of the TD-NoBL and TD-NoFQ datasets with the test dataset can be quantified using the average ROC-AUC of Tanimoto similarities of each compound in the training set with the test set. The average ROC-AUC returned is 0.60 and 0.55 for TD-NoBL and TD-NoFQ respectively.

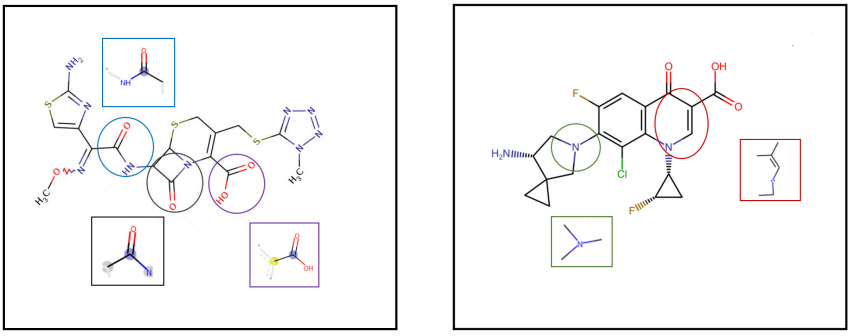
For completeness, we examined the effects of using the non-standardized (original) dataset in these experiments. Results for the original dataset are slightly poorer than for their standardized counterpart for Chemprop, as it fails to predict any β-lactam containing active in the test set when trained on TD-NoBL (Supplementary Information, Table S1). However, MADS and RMT exhibit a better performance when trained on non-standardized TD-NoBL, as compared to when trained on standardized TD-NOBL. We discuss these performance variations and underlying explanations in detail in the

following section.

Substructural contribution to activity analysis

As discussed in the previous section, both MADS and RMT were able to correctly predict compounds in the test set as actives even when their corresponding chemotype class was removed from the training set. Even though due to bit collision, multiple fragments can be assigned to the same motif, it is still possible to "unbox" the models by studying which fragments have higher activity weights or contribute to the activity eigen space. To investigate further, a detailed analysis of the fingerprint representation of both training (TD-NoBL and TD-NoFQ) and test sets was done.

Figure 4 shows some of the high scoring substructures for activity prediction using RDKit-Morgan fingerprint for MADS (Figure 4a), and RMT (Figure 4b). The circled substructures in Figure 4a are associated with bits having high activity-weights, and the circled substructures in Figure 4b are represented by bits contributing to the activity eigen-space. Investigation of these high-scoring substructures allows us to rationalize the ability of MADS and RMT to correctly predict the activity of β-lactams and fluoroquinolones, despite their absence from the training sets.



(a) (b)

Fig. 4: (a) Structure of cefmenoxime, a third generation β-lactam containing cephalosporin with highlighted substructures present in β-lactams (circle) and corresponding highlighted structures (square) that are predicted by MADS to have higher activity scores (when trained on TD-NoBL), and (b) Structure of sitafloxacin, a fluoroquinolone antibiotic with highlighted substructures present in fluoroquinolones (circle) and corresponding highlighted structures (square) that are predicted by RMT to contribute to the activity eigen-space (when trained on TD-NoFQ).

TD-NoBL

The β-lactam ring contains an amide group (highlighted in the black square in Figure 4a), a fingerprint bit with high weighted contribution towards activity (as predicted by MADS when trained on TD-NoBL). This bit is, of course, not exclusively present in β-lactam antibiotics, it is also present in other compounds in the training set (27% actives, and 19% inactives in TD-NoBL). Figure 5 displays the presence of the amide group in the glycopeptide drug bleomycin, an active in the *E-coli* assay present in the training dataset. The amide substructure is also present in polymyxins and nucleopeptides in the training set. Thus, the presence of these peptidic substructures in other active antibiotic classes contributes to β-lactams scoring highly, increasing their probability of being predicted as active.

TD-NoFQ

In a similar way, the core bicyclic ring present in fluoroquinolones is broken down to simpler constituting substructures (highlighted in the red and green squares in Figure 4b). These substructures, represented by bits that are closer to the linear subspace of activity (as predicted by RMT when trained on TD-NoFQ), are present in most cephalosporins (cefpiramide, cefuroxime, cefoperazone, cefepime etc) and also other compounds, eg. zidovudine, all of which are actives in TD-NoFQ. Figure 6 shows the structure of cefpramide, an active in the *E-coli* assay, containing this significant substructure. The ability of RMT to learn these structures from cephalosporins and other compounds present in the TD-NoFQ dataset and identify them as significant bits for determining activity, facilitates the correct prediction of all fluoroquinolones as active in the test set.

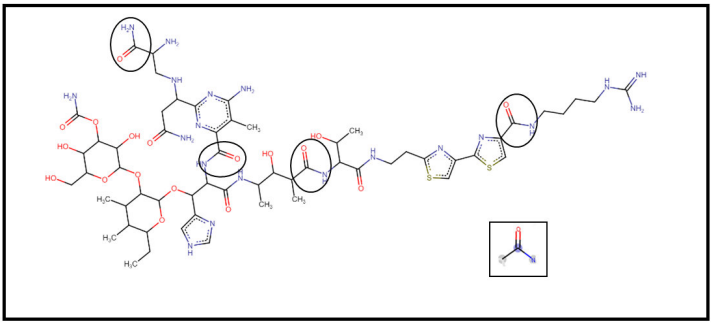


Fig. 5: Significant substructures present in bleomycin as determined by MADS

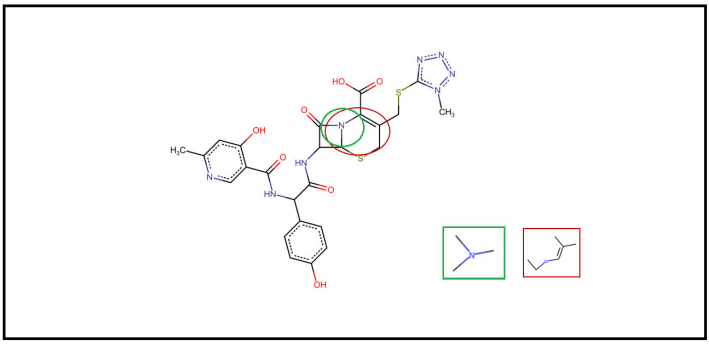


Fig. 6: Significant substructures present in cefpiramide, as determined by RMT

Thus, MADS and RMT appear able to identify β-lactams and fluoroquinolones as novel active chemotypes, which Chemprop does not.

Prediction of halicin

All three methods predict halicin as an active when trained on TD-Cleaned and TD-NoFQ. However, when trained on TD-NoBL, only RMT predicts it as an active. Halicin is a small molecule containing 3 sulphur atoms. Sulphur is present in 37 of the 119 active structures (31%) and 320 of the 2180 (14%) inactive structures in TD-Cleaned. After removing β-lactam compounds, sulphur is present in 9 of the 91 (10%) active structures, and 293 of the 2151 (13%) inactive structures. When removing the compounds with the fluoroquinolone structure, sulphur is present in 35 of the 98 (35%) active structures, and 320 of the 2180 (14%) inactive structures.

Thus, when the three methods are trained on TD-NoBL, prediction of halicin may become challenging due to the fact that sulphur-containing fragments are now less frequent in the actives than in the inactives, the opposite of the situation in the other datasets. The correct prediction by RMT is impressive, in this case. However, the reason for this result is not entirely clear, but we suspect it is related to the occurrence of bits (such as a nitro group and sulphur fragments) present as significant bits in the principal eiegenvectors, thus contributing to the eigendirections spanning the RMT 'active' subspace.

A related issue occurs while comparing performance when the methods are trained on the non-standardized version of TD-NoBL (Supplementary Information, Table S1). Unlike when trained on the full datasets, where standardization improved performance, both MADS and RMT exhibit a slightly better performance when trained on the non-standardized version of TD-NoBL. In addition to the presence of sulphur in many active β-lactam

antibiotics, many of the salts present in the non-standardized dataset also contain sulphur (sulphonate, mesylate, edisylate etc.) One of the underlying reasons of this difference in performance could be the activity signal produced by additional sulphur containing structures, in the form of salts, in the non-standardized active TD-NoBL dataset where sulphur based salts are present in 11% active and 3% inactive structures. Consequently, sulphur fragments are recognised as a high weighted substructure for activity prediction by MADS when trained on non-standardized TD-NoBL. For RMT, while sulphur is still recognised as one of the substructures in the activity eigenspectrum when trained on TD-NoBL, determining the difference of magnitude of its contribution from non-standardized TD-NoBL is not as straightforward as it is for MADS because a combination of multiple bits contribute to the principal eigenvectors in this method. This provides another example where performance of a method is dependent on the training dataset and again highlights the importance of standardization to minimize artefactual results.

Conclusions

In the study conducted by Stokes et al, the deep learning method Chemprop was able to identify halicin as a compound with inhibitory activity against E. coli growth. Further characterization showed it to also inhibit the growth of a wide variety of other pathogens including *Mycobacterium tuberculosis* [49] and carbapenem-resistant *Enterobacterales* [50]. We have compared the performance of Chemprop and two other methods, MADS and RMT, which are based on the interpretation of the fingerprint matrix. Using the original data, the performance of Chemprop and MADS was comparable, with MADS performing slightly better as assessed by confusion-matrix-based performance measures. There is also a high correspondence between the compounds predicted to be active. RMT performed not quite as well, but still generated a clear activity signal.

The ability of a machine learning model to predict chemical activity is dependent on the way in which the chemical data is represented [51, 52]. Employing a consistent molecular structure representation minimizes the potential for artefactual results. Therefore, standardization of input data is routinely employed to remove potential bias due to, for example, the presence of salts or duplicates. Standardizing the halicin dataset led to small, but consistent, improvements in the predictions. This is true for all three methods. The improvement is greatest for RMT, which implies that Chemprop and MADS are more robust to noise in the data. Standardization appears, on average, to remove artefacts that were degrading the predictive ability of models.

It is generally worthwhile to benchmark any method against a simpler approach. In that way, the added value of the more complex, and usually more

resource intensive, method can be understood. All three methods perform significantly better than using a simple similarity-based approach to activity prediction, implying that the underlying techniques are more accurately capturing the relationships between chemical structure and biological activity in the data.

It is also highly desirable if a method can correctly predict activity for compounds that are chemically dissimilar to any in the training set. Identifying a novel chemotype could take a project in a new direction and increases the chances of success. However, this is not easy to achieve. Chemprop has demonstrated ability to do this, as it identified the compound halicin, a kinase inhibitor whose antibacterial activity was unknown before the study. The machine learning methods MADS and RMT also have this potential, as halicin was predicted as an active by all three methods in most experiments.

We also performed experiments in which active chemotype classes were removed from the training set and investigated the performance of the resulting models. MADS and RMT are able to predict the activity of many β-lactam and fluoroquinolone antibiotics, even when there are no examples of these chemotypes in the training set. Chemprop was unable to predict the activity of either chemotype in these experiments.

A detailed analysis of the chemical substructures (fingerprint bits) that lead to this ability of MADS and RMT provides evidence of how this happens. For example, there are structural similarities between the bicyclic ring system of penicillin and cephalosporins, which can be considered to be highly modified dipeptides, and other peptidic antibiotics, such as glycopeptides and polymyxins. Thus, the amide-like substructures in β-lactams gain high weights because these are present in other antibiotic classes that remain in the dataset after β-lactam removal. Similar examples are present regarding similarities of the fluoroquinolone ring system and cephalosporins.

Do these findings represent genuine examples of scaffold-hopping or are they in some way artefactual or fortuitous? The mode of action of each of these different classes (β-lactams, fluoroquinolones, glycopeptides, polymyxins) are very distinct, interacting with very different targets in different cellular compartments. Similarly, the mode of action of halicin is distinct from any compound in the training and test sets. It is, therefore, hard to imagine a mechanistic connection leading to these "scaffold-hopping" predictions.

Antibiotic activity does not just depend on potent inhibition of the target protein, but many other factors such as accumulation at the site of action and metabolic stability. Thus, it is possible that the methods could be capturing contributions to activity from these other factors. The Gram-negative cell

wall is known to present a formidable barrier to compound entry. Some guidelines for physicochemical property ranges compatible with accumulation and rules for Gram-negative entry have been developed [53, 54] but the predictive models remain rudimentary. As a result, many antibacterial drug discovery projects have failed to optimize potent inhibitors of their molecular targets into compounds with clinically useful antibacterial activity. It is conceivable that the machine learning models are capturing contributions to generic factors such as penetration through porin channels or susceptibility to efflux mechanisms. Whilst it is difficult to be certain, it appears unlikely to be an explanation in this case. The most important bits are most clearly linked to substructures involved in the antibiotic mechanisms of action, and the fluoroquinolones and cephalosporins are active in different cellular compartments (the cytoplasm and periplasm respectively).

Thus, in our opinion, it seems most likely that these predictions by MADS and RMT, based on similarities to active compounds in the training set at the substructure level, but which are not linked by a common mechanism of action of the compounds, are fortuitous. The same is likely true for the prediction of halicin as active by all three methods. For MADS and RMT, the occurrence of high scoring bits containing sulphur was discussed above. For Chemprop, rationalizing the prediction with confidence is difficult because of the inherent lack of transparency of the method. However, the frequency of occurrence of sulphur and nitro groups in the ZINC15 virtual screening hits predicted by Chemprop [25] is perhaps noteworthy and may represent a preference for compounds containing those substructures. In one sense, this behaviour could be likened to the concept of discovering 'privileged scaffolds', i.e. molecular frameworks that can be incorporated into ligands binding to a diverse array of target proteins. The "scaffolds" detected by use of Morgan fingerprint descriptors may be too simple to be of much practical utility in drug discovery, but the potential for effective use with fingerprints incorporating more meaningful chemical substructures is worth further research. Further work will also explore capturing more detail of the relative disposition of 'active' substructure to increase the chances of mechanistically meaningful predictions.

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Authors' contribution

All authors contributed to the study conception and design. RJ did the experimental analysis and wrote the first draft of the main manuscript. TM contributed to the mathematical interpretation of all the ML methods and reviewed the manuscript. PF substantially contributed to the conception of the experiments, critically reviewed and revised the manuscript. All authors read and approved the final manuscript.

Funding

This research received no external funding.

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Data Availability Statement

The training and test datasets are available as supplementary information with the publication by Stokes *et. al.* [25, Supplementary Tables S2A, S2B]. The codes for all three methods discussed in the paper are publicly available. The links to the codes are provided as follows :
Chemprop [25] : https://github.com/swansonk14/chemprop
Mapping of Activity through Dichotomic Scores (MADS) [26]: https://michem.unimib.it/download/matlab-toolboxes/virtual-screening-toolbox-for-matlab/
Random Matrix theory (RMT) [27] : https://github.com/alphaleegroup/RandomMatrixDiscriminant

Conflicts of Interest

The authors declare that they have no known competing interests.

Acknowledgements

The authors would like to thank Dr. Jean Paul Ebejer, University of Malta, for his valuable suggestions to improve the manuscript.

References

[1] Yanling J, Xin L, Zhiyuan L (2013) The antibacterial drug discovery. Drug Discovery pp 289–307

[2] Aminov RI (2010) A brief history of the antibiotic era: lessons learned and challenges for the future. Frontiers in microbiology 1:134

[3] Laxminarayan R, Duse A, Wattal C, et al (2013) Antibiotic resistance—the need for global solutions. The Lancet infectious diseases 13(12):1057–1098

[4] Goh GB, Hodas NO, Vishnu A (2017) Deep learning for computational chemistry. Journal of computational chemistry 38(16):1291–1307

[5] Scarselli F, Gori M, Tsoi AC, et al (2008) The graph neural network model. IEEE transactions on neural networks 20(1):61–80

[6] Baskin II, Winkler D, Tetko IV (2016) A renaissance of neural networks in drug discovery. Expert opinion on drug discovery 11(8):785–795

[7] Salt DW, Yildiz N, Livingstone DJ, et al (1992) The use of artificial neural networks in qsar. Pesticide science 36(2):161–170

[8] Ghasemi F, Mehridehnavi A, Perez-Garrido A, et al (2018) Neural network and deep-learning algorithms used in qsar studies: merits and drawbacks. Drug Discov Today 23(10):1784–1790

[9] Staszak M, Staszak K, Wieszczycka K, et al (2021) Machine learning in drug design: Use of artificial intelligence to explore the chemical structure–biological activity relationship. Wiley Interdisciplinary Reviews: Computational Molecular Science p e1568

[10] Mayr A, Klambauer G, Unterthiner T, et al (2018) Large-scale comparison of machine learning methods for drug target prediction on chembl. Chemical science 9(24):5441–5451

[11] Lenselink EB, Ten Dijke N, Bongers B, et al (2017) Beyond the hype: deep neural networks outperform established methods using a chembl bioactivity benchmark set. Journal of cheminformatics 9(1):1–14

[12] Gaulton A, Bellis LJ, Bento AP, et al (2012) Chembl: a large-scale bioactivity database for drug discovery. Nucleic acids research 40(D1):D1100–D1107

[13] Truchon JF, Bayly CI (2007) Evaluating virtual screening methods: good and bad metrics for the "early recognition" problem. Journal of chemical information and modeling 47(2):488–508

[14] Koutsoukas A, Monaghan KJ, Li X, et al (2017) Deep-learning: investigating deep neural networks hyper-parameters and comparison of performance to shallow methods for modeling bioactivity data. Journal of

cheminformatics 9(1):1–13

[15] Duvenaud D, Maclaurin D, Aguilera-Iparraguirre J, et al (2015) Convolutional networks on graphs for learning molecular fingerprints. arXiv preprint arXiv:150909292

[16] Withnall M, Lindelöf E, Engkvist O, et al (2020) Building attention and edge message passing neural networks for bioactivity and physical–chemical property prediction. Journal of cheminformatics 12(1):1–18

[17] Jiang D, Wu Z, Hsieh CY, et al (2021) Could graph neural networks learn better molecular representation for drug discovery? a comparison study of descriptor-based and graph-based models. Journal of cheminformatics 13(1):1–23

[18] Robinson MC, Glen RC, et al (2020) Validating the validation: reanalyzing a large-scale comparison of deep learning and machine learning models for bioactivity prediction. Journal of computer-aided molecular design pp 1–14

[19] Pérez-Sianes J, Pérez-Sánchez H, Díaz F (2016) Virtual screening: a challenge for deep learning. In: International Conference on Practical Applications of Computational Biology & Bioinformatics, Springer, pp 13–22

[20] Bajorath J (2017) Computational scaffold hopping: cornerstone for the future of drug design?

[21] Schneider G, Neidhart W, Giller T, et al (1999) "scaffold-hopping" by topological pharmacophore search: a contribution to virtual screening. Angewandte Chemie International Edition 38(19):2894–2896

[22] Vainio MJ, Kogej T, Raubacher F, et al (2013) Scaffold hopping by fragment replacement

[23] Saluste G, Albarran MI, Alvarez RM, et al (2012) Fragment-hopping-based discovery of a novel chemical series of proto-oncogene pim-1 kinase inhibitors. PloS one 7(10):e45,964

[24] Ertl P (2012) Database of bioactive ring systems with calculated properties and its use in bioisosteric design and scaffold hopping. Bioorganic & medicinal chemistry 20(18):5436–5442

[25] Stokes JM, Yang K, Swanson K, et al (2020) A deep learning approach to antibiotic discovery. Cell 180(4):688–702

[26] Todeschini R, Consonni V, Ballabio D, et al (2018) Mapping of activity through dichotomic scores (mads): A new chemoinformatic approach to detect activity-rich structural regions. Journal of Chemometrics 32(4):e2994

[27] Lee AA, Yang Q, Bassyouni A, et al (2019) Ligand biological activity predicted by cleaning positive and negative chemical correlations. Proceedings of the National Academy of Sciences 116(9):3373–3378

[28] Inc CCG (2019) Molecular operating environment (moe)

[29] Corsello SM, Bittker JA, Liu Z, et al (2017) The drug repurposing hub: a next-generation drug library and information resource. Nature medicine 23(4):405–408

[30] Cereto-Massagué A, Ojeda MJ, Valls C, et al (2015) Molecular fingerprint similarity search in virtual screening. Methods 71:58–63

[31] Willett P (2006) Similarity-based virtual screening using 2d fingerprints. Drug discovery today 11(23-24):1046–1053

[32] Muegge I, Mukherjee P (2016) An overview of molecular fingerprint similarity search in virtual screening. Expert opinion on drug discovery 11(2):137–148

[33] Riniker S, Landrum GA (2013) Open-source platform to benchmark fingerprints for ligand-based virtual screening. Journal of cheminformatics 5(1):1–17

[34] Wale N, Watson IA, Karypis G (2008) Comparison of descriptor spaces for chemical compound retrieval and classification. Knowledge and Information Systems 14(3):347–375

[35] Russo DP, Zorn KM, Clark AM, et al (2018) Comparing multiple machine learning algorithms and metrics for estrogen receptor binding prediction. Molecular pharmaceutics 15(10):4361–4370

[36] Kensert A, Alvarsson J, Norinder U, et al (2018) Evaluating parameters for ligand-based modeling with random forest on sparse data sets. Journal of cheminformatics 10(1):1–10

[37] Chen B, Harrison RF, Papadatos G, et al (2007) Evaluation of machine-learning methods for ligand-based virtual screening. Journal of computer-aided molecular design 21(1):53–62

[38] (1984) Maccs keys, mdl information systems. Inc: San Leandro, CA

[39] Nilakantan R, Bauman N, Dixon JS, et al (1987) Topological torsion: a new molecular descriptor for sar applications. comparison with other descriptors. Journal of Chemical Information and Computer Sciences 27(2):82–85

[40] Landrum G (2013) Rdkit documentation. Release 1(1-79):4

[41] Lee AA, Brenner MP, Colwell LJ (2016) Predicting protein–ligand affinity with a random matrix framework. Proceedings of the National Academy of Sciences 113:13,564 – 13,569

[42] Bajusz D, Rácz A, Héberger K (2015) Why is tanimoto index an appropriate choice for fingerprint-based similarity calculations? Journal of cheminformatics 7(1):1–13

[43] Hussin SK, Abdelmageid SM, Alkhalil A, et al (2021) Handling imbalance classification virtual screening big data using machine learning algorithms. Complexity 2021

[44] Branco P, Torgo L, Ribeiro RP (2017) Relevance-based evaluation metrics for multi-class imbalanced domains. In: Pacific-Asia Conference on Knowledge Discovery and Data Mining, Springer, pp 698–710

[45] Ballabio D, Grisoni F, Todeschini R (2018) Multivariate comparison of classification performance measures. Chemometrics and Intelligent Laboratory Systems 174:33–44

[46] Schubert S, Dalhoff A (2012) Activity of moxifloxacin, imipenem, and ertapenem against escherichia coli, enterobacter cloacae, enterococcus faecalis, and bacteroides fragilis in monocultures and mixed cultures in an in vitro pharmacokinetic/pharmacodynamic model simulating concentrations in the human pancreas. Antimicrobial agents and chemotherapy 56(12):6434–6436

[47] Marie MAM, Krishnappa LG, Lory S (2016) In vitro activity and the efficacy of arbekacin, cefminox, fosfomycin, biapenem against gram-negative organisms: New treatment options? Proceedings of the National Academy of Sciences, India Section B: Biological Sciences 86(3):749–755

[48] Goto S, Sakamoto H, Ogawa M, et al (1982) Bactericidal activity of cefazolin, cefoxitin, and cefmetazole against escherichia coli and klebsiella pneumoniae. Chemotherapy 28(1):18–25

[49] Russell DG (2001) Mycobacterium tuberculosis: here today, and here tomorrow. Nature reviews Molecular cell biology 2(8):569–578

[50] Brenner DJ, Farmer III J (2015) Enterobacteriaceae. Bergey's manual of systematics of archaea and bacteria pp 1–24

[51] Fourches D, Muratov E, Tropsha A (2010) Trust, but verify: on the importance of chemical structure curation in cheminformatics and qsar modeling research. Journal of chemical information and modeling 50(7):1189

[52] Williams AJ, Ekins S, Tkachenko V (2012) Towards a gold standard: regarding quality in public domain chemistry databases and approaches to improving the situation. Drug discovery today 17(13-14):685–701

[53] Richter MF, Drown BS, Riley AP, et al (2017) Predictive compound accumulation rules yield a broad-spectrum antibiotic. Nature 545(7654):299–304

[54] Ebejer JP, Charlton MH, Finn PW (2016) Are the physicochemical properties of antibacterial compounds really different from other drugs? Journal of cheminformatics 8(1):1–9