# AI-Driven Antimicrobial Discovery: Harnessing Artificial Intelligence to Combat Antimicrobial Resistance

*Megha Puri[1], Shikha Mittal[1], Nishant Jain[2], Jitendraa Vashistt[1]**

*[1]Department of Biotechnology & Bioinformatics, Jaypee University of Information Technology, Waknaghat, Solan, Himachal Pradesh, India*

*[2]Department of Electronics and Communication Engineering, Jaypee University of Information Technology, Waknaghat, Solan, Himachal Pradesh, India*

**Corresponding author: jitendraa.vashistt@juitsolan.in*

**Abstract:** In recent years, a large surge in antimicrobial resistance is one of the prominent problems in tackling nosocomial infections. The current arsenal of antimicrobials seems to be insufficient for treating these infections, as pathogens are evolving and manifesting various resistance mechanisms to overcome the action of existing drugs. The search for novel antimicrobial compounds is of utmost priority. However, the traditional approach of drug discovery is tedious, time-consuming and labour-intensive. To address this problem, study was conducted as a computational framework that could help to screen various compounds in order to narrow down the most probable active compounds from the inactive ones. Machine Learning approach was utilized for curating a dataset from the ChEMBL database, and three machine learning classifiers (Random Forest, Support Vector Machine (SVM) and Logistic Regression) were trained to distinguish compounds based on their structural and physicochemical features. Random Forest outperformed the other two classifiers with accuracy of 95.28% and AUC-ROC of 0.986. This suggests that the model has strong ability to discriminate between antimicrobial and non-antimicrobial compounds. This study demonstrates that machine learning can be integrated into early steps of antimicrobial drug discovery to narrow down the search for novel compounds virtually from large databases.

**Keywords:** Antimicrobial resistance, Machine learning, ESKAPE pathogens, Drug discovery, Random Forest, ChEMBL

## 1 Introduction

Antimicrobials are an essential part of modern medicine and play an important role in the treatment of various pathogenic infections. However, the effectiveness of current antimicrobials is decreasing due to the rise in multidrug resistance. As a result, there are a limited number of compounds left to manage the infections that were earlier easy to treat. Thus, there is an urgent need to find novel compounds to reduce rapidly rising rates of multidrug resistance [1]. Resistance to antimicrobials is an outcome of genetic mutations in pathogens. These alterations may result in the modification of the drug target, augmented activity of efflux pumps, or promoted enzymatic degradation of antimicrobials. Resistance-related genes further spread because of horizontal gene transfer, which primarily includes three mechanisms, i.e., transformation, conjugation, and transduction [2]. Consequently, Antimicrobial Resistance (AMR) is estimated to cause an alarmingly high number of deaths in the future if no effective strategies are implemented in time. According to projections, AMR is expected to be responsible for about 1.91 million deaths worldwide by 2050 [3].

Currently, the majority of available antimicrobials have either been discovered years ago or are available as modified versions of molecules that already exist, resulting in a cycle of repetitive compounds. This has been seen in the case of beta lactams, tetracyclines, macrolides, and various other classes of antibiotics, where the newer generations are mostly modified derivatives of existing antibiotics [4]. Moreover, the process of antimicrobial discovery is laborious and it may take several years of preclinical evaluation before a potential compound can even be submitted for clinical trials. From the beginning of screening to optimization and regulatory approval, the process requires a significant amount of resources and includes a high risk of failure. To overcome this limitation, Artificial Intelligence (AI) has appeared as a promising solution that allows performing extensive search across a large number of compounds from numerous available repositories. It could be used to virtually screen out potential novel compounds with potent antimicrobial activity [5]. Machine learning (ML), in particular, can help to decipher complex biological patterns by training models on available datasets. These algorithms can be utilized to predict biological activity and repurpose existing compounds from available databases as well as design novel molecules with potent antimicrobial properties. Furthermore, the interaction of potential molecules with various drug targets can be computationally analyzed to obtain more accurate predictions [6].

In this study, we address this limitation by developing a machine learning framework based specifically on ESKAPE pathogens (*Enterococcus faecium*, *Staphylococcus aureus*, *Klebsiella pneumoniae*, *Acinetobacter baumannii*, *Pseudomonas aeruginosa*, and *Enterobacter* species). Our approach integrates structural fingerprints with physicochemical characteristics of the

compounds to classify the compounds into potential antimicrobial and non-antimicrobial compounds. We utilized three classifiers, i.e., Random Forest, SVM and Logistic Regression to develop a prediction framework. We also identified substructures that could be linked with the antimicrobial properties of these compounds. Overall, ML can hasten the process of discovery of new compounds and at the same time also improve the efficiency of compound selection.

## 2 Methodology

### *2.1 Data Acquisition*

An open-access repository of bioactive molecules, the ChEMBL database [7], was utilized for mining and creation of positive and negative datasets. The positive dataset here refers to those compounds that have known antimicrobial activity. The database was utilized to retrieve compounds with established activity against the WHO priority ESKAPE pathogens. Stringent filtering parameters were applied to ensure that the data utilized for ML training was of optimum quality. Only those compounds were retrieved from the database that had Minimum Inhibitory Concentration (MIC) values reported from inhibitory experiments. All the inhibitory values were normalized in μg/ml in order to ensure uniformity. The dataset was further improved by excluding any ambiguous data points. Compounds were separated on the basis of their MIC threshold. Compounds with MIC >128 μg/ml were utilized to create the negative dataset and the compounds with <10 μg/ml MIC were chosen for the creation of positive dataset. RDKit (https://www.rdkit.org) was utilized for data standardization. The compounds with missing canonical residues, molecular weight or other descriptions were also excluded. Duplicates were removed and the compounds were sorted into positives and negatives on the basis of MIC thresholds. The final dataset comprised of 15,793 compounds with 12,110 positives and 3,683 negatives.

### *2.2 Feature Extraction and Feature Importance Analysis*

The feature extraction used a combination strategy where structural features were encoded as binary vectors, while physicochemical properties were represented as numerical values suitable for machine learning. To achieve this, the initial chemical structures were represented as Simplified Molecular-Input Line-Entry System (SMILES) strings and converted into feature representations using the RDKit library. Morgan fingerprints of radius 2 and 2048-bits were computed using RDKit. These fingerprints make it possible for the model to identify various structural patterns including specific atoms and functional groups that have an impact on the antimicrobial potency of compounds. Morgan fingerprints encode the circular substructures around each atom which can help to capture local structural information of compounds. A radius of two includes both the immediate neighbors and the second-order atomic surroundings. Furthermore, physicochemical descriptors like molecular weight, LogP (lipophilicity), Hydrogen Bond Acceptor (HBA) and Hydrogen Bond Donors (HBD), rotatable bonds and Topological Polar Surface Area (TPSA) were computed using RDKit to achieve additional information about the properties of the compounds. Molecular weight and LogP represent the compound's size and lipophilicity, respectively. These two features were required to determine the ability of a molecule to interact and move across the biological membrane. HBA and HBD were utilized to determine the intermolecular interactions. TPSA and rotatable bonds were included to analyse the molecular polarity and flexibility of the compounds, which are known to govern their membrane permeability. These parameters can help to understand the bioavailability and membrane interaction of the compounds. The resultant hybrid vector captured both local structural patterns and global chemical properties. These features provide a complementary descriptor space together. The combination approach determines the ability of the compound to interact with the specific bacterial targets based on its topology and its physical properties helps to determine its ability to penetrate the bacterial membrane. Furthermore, all the features were standardized using Z-score normalization before model training. This was done to ensure that features with larger value ranges would not significantly impact the process of training.

### *2.3 Machine Learning and Model Development*

Three classifiers, Random Forest, SVM, and Logistic Regression were trained on the curated datasets using the Scikit-learn Python package [8]. Random Forest builds multiple regression trees and then uses the outcome by averaging the results from each tree for prediction. SVM turns the input features into higher dimensional space and the two classes are separated by a hyperplane [9]. Furthermore, Logistic Regression predicts the probability of an outcome on the basis of individual characteristics to give a binary response [10]. The data was split into a training set (80%) for the development of the model and an independent testing set (20%) for evaluation of model performance. To avoid model biasness, the data was split using stratified sampling. This avoids the model from being biased in its learning. In addition, 5-fold cross-validation was used to prevent overfitting and ensure generalization to unseen chemical spaces.

### *2.4 Performance Evaluation*

To determine the performance of models, we utilized several key evaluation metrics. Accuracy was measured to assess the overall proportion of correct predictions by the model; precision and recall were employed to balance specificity and sensitivity. F1 score was computed to analyse the class balance for confirmation of the model performance on the negative class. In addition, Area Under the Receiver Operating Characteristic Curve (AUC-ROC) was used to determine the discriminating ability of the model independent of decision thresholds. The closer the score is to one, the greater is the ability of the model to separate active compounds from inactive ones. Furthermore, feature importance evaluation was carried out to understand the contribution of individual features in model prediction.

## 3 Results and Discussion

### *3.1 Model Evaluation*

The comparative analysis showed that Random Forest performed significantly better than SVM and Logistic Regression in terms of performance across all evaluation metrics (Accuracy, Precision, Recall, F1-Score and AUC-ROC).The summary of the performance of classifiers is presented in Table 1. We evaluated all the models on the independent testing dataset, and RF gave an accuracy of 95.28% with an AUC-ROC of 0.986. The AUC-ROC curve for all the models is shown in Fig.2. In contrast, SVM and Logistic Regression achieved an accuracy of 94.14% and 91.01%, respectively. The comparison of model performance is depicted in Fig. 1. Various studies confirmed the better performance of Random Forest as compared to linear models as it can learn complex non-linear relations between chemical features and biological activity, resulting in better predictive ability [11]. To reduce the risk of inflation of results because of class imbalance in datasets, stratified sampling was applied during the splitting of datasets into testing and training sets to maintain the same class proportions. This helps to ensure that the evaluation is fair and not biased due to the differences in class distribution between the two sets. Furthermore, F1-score was utilized to evaluate model performance instead of just relying on accuracy. The F1-score combines precision and recall which makes it more suitable for imbalanced datasets. It reflects how well the model performs on the minority class and reduces the chance of reporting inflated results that can occur when the majority class dominates.

**Table 1. Comparative Performance Metrics of Classifiers. Evaluation of Random Forest, SVM, and Logistic Regression classifiers.**

| Classifier | Accuracy (%) | Precision (%) | Recall / Sensitivity (%) | F1-Score (%) | AUC-ROC |
|---|---|---|---|---|---|
| Random Forest | 95.28 | 96.52 | 97.36 | 96.94 | 0.986 |
| SVM | 94.14 | 97.33 | 94.96 | 96.13 | 0.977 |
| Logistic Regression | 91.01 | 95.18 | 92.98 | 94.07 | 0.952 |

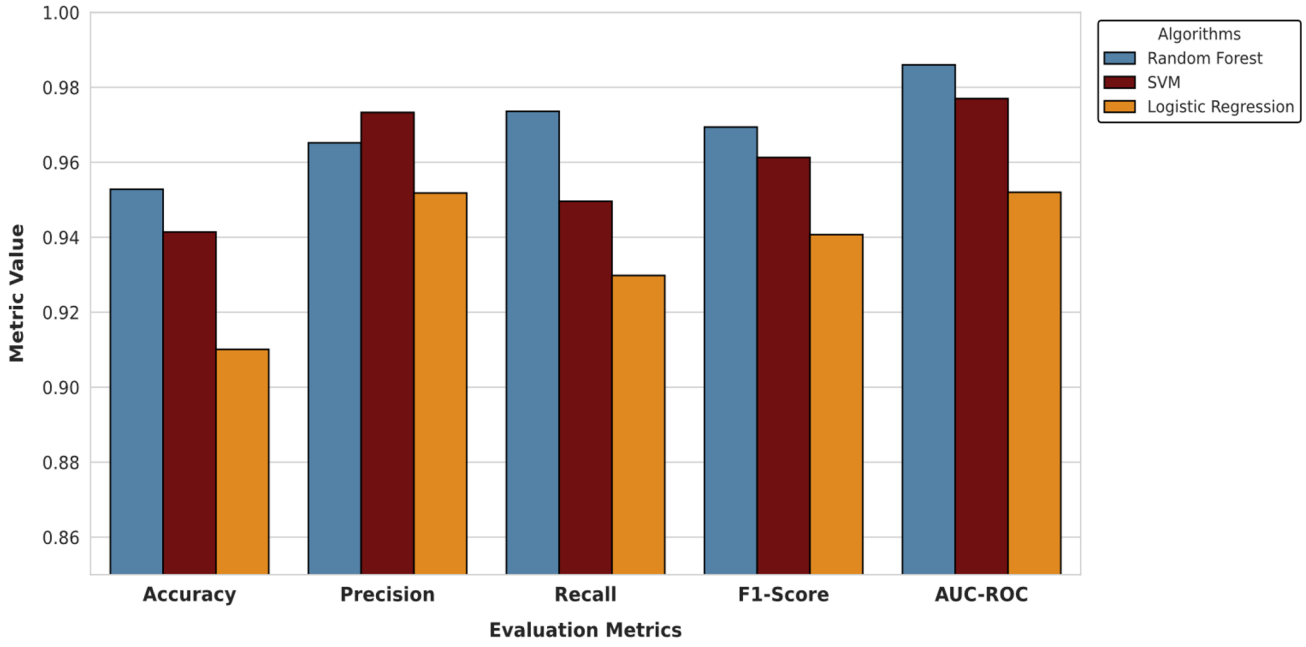


*Fig. 1. Comparative Model Performance. Performance metrics across classifiers.*

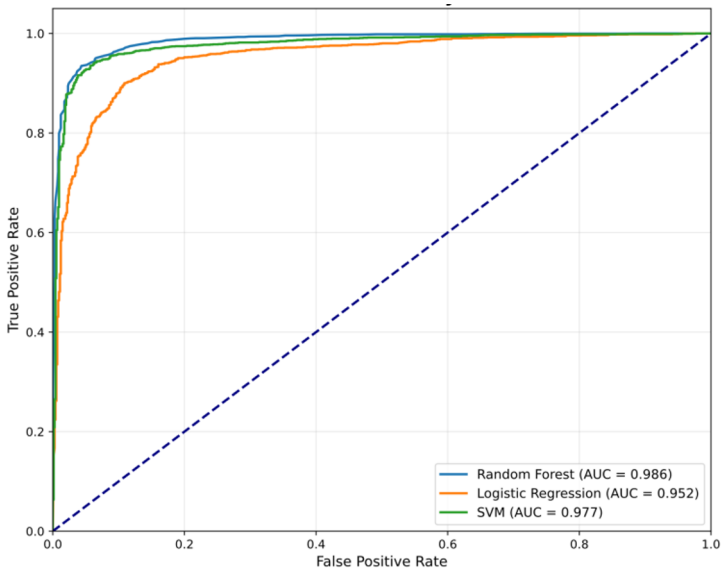


*Fig. 2. AUC-ROC Analysis. The Random Forest model achieved the highest Area Under the Curve.*

### *3.2 Feature Importance Analysis*

The results revealed that the antimicrobial activity of compounds depends on both physicochemical properties and substructures. As shown in Fig.3, molecular weight, Topological Polar Surface Area (TPSA), and LogP (lipophilicity) were amongst the top predictors of antimicrobial activity. This indicates that molecular size and the molecule's ability to interact and partition across lipid membranes are of primary importance for a molecule to penetrate through the microbial cell membrane. While size and lipophilicity are important factors that determine the ability of the compound to enter the pathogen, specific structural motifs also govern the molecule's ability to bind and interact with intracellular targets.

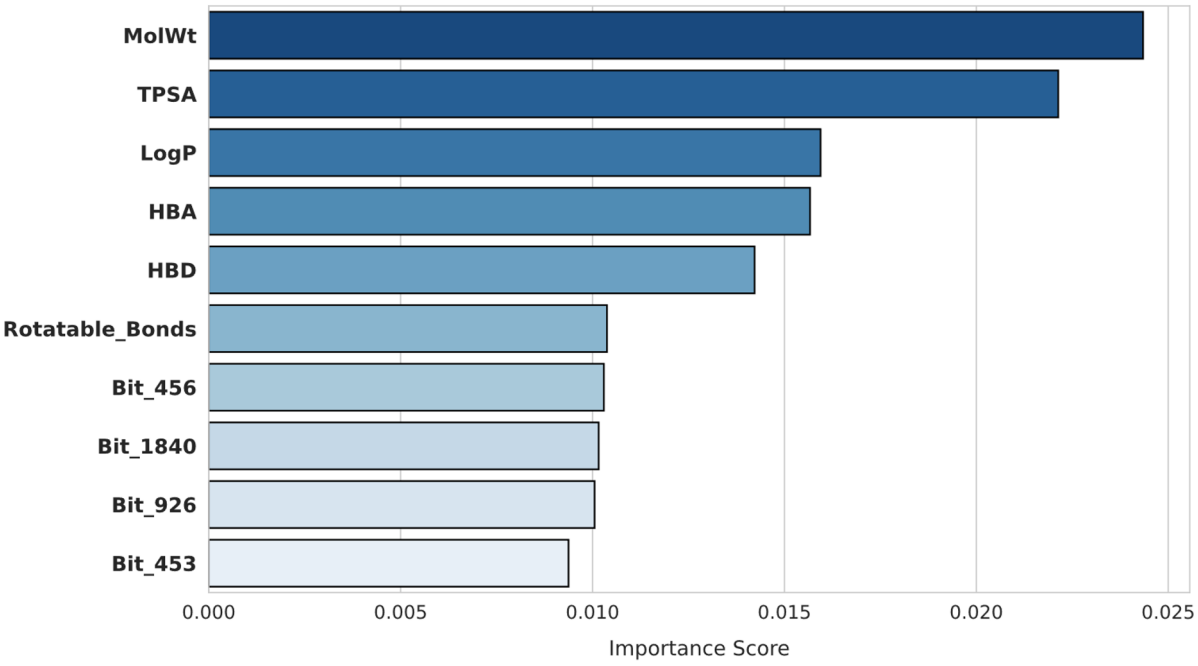


*Fig. 3. Feature Importance Analysis. Molecular weight (MolWt), Topological Polar Surface Area (TPSA), LogP (lipophilicity) are physicochemical features that primarily determine the antimicrobial activity of a compound. Hydrogen Bond Acceptor (HBA) and Donors (HBD) are known to affect how strongly a chemical binds to other compounds. Rotatable bonds control how flexible the molecule is at the molecular level. Bits (456, 1840, 926, 453) are the structural scaffolds that the compounds require to inhibit the pathogens.*

Structural analysis showed that halogenated motifs and nitrogen-containing heterocycles are key substructures that contribute to antimicrobial activity. These contributions are illustrated in Fig.4. These findings can be supported by the fact that halogenation of compounds is associated with their lipophilicity [12]. The higher affinity of compounds towards lipid groups allows the molecules to move across the bacterial membranes. These features were ranked at a higher position in the feature importance analysis, which shows that the model is learning chemically relevant features and not just relying on superficial factors.

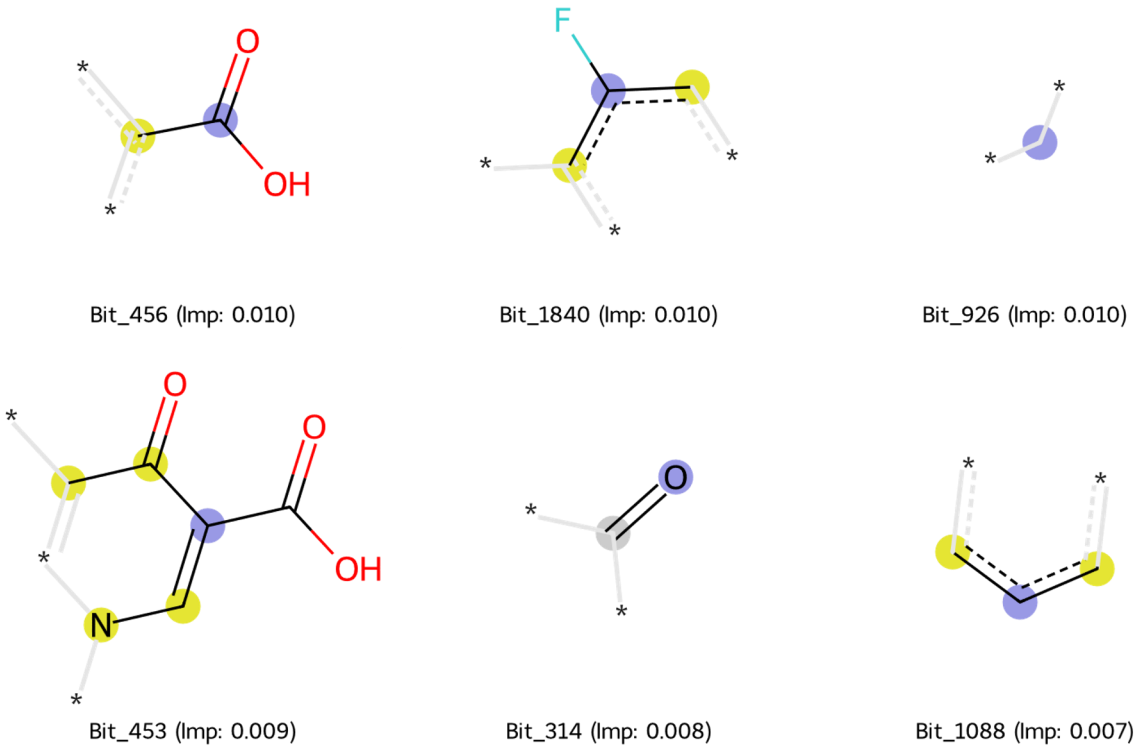


*Fig. 4. Chemical Substructures. Key molecular fragments contributing to antimicrobial activity as identified by the model.*

## 4 Conclusion

This study evaluated multiple machine learning methods to identify possible antimicrobial compounds against multidrug-resistant ESKAPE pathogens. The Random Forest model had superior predictive performance compared to the others, demonstrating enhanced accuracy, stability, and generalization capability. Its capacity to handle high-dimensional molecular descriptors and identify intricate non-linear correlations renders it an effective tool for screening antimicrobial compounds. In future research, our proposed work can be incorporated into the early stages of computational filtering of antimicrobial compounds, where it could provide a predictive framework to help researchers obtain a refined subset of lead compounds. This approach can help to narrow down the search of large chemical repositories into a smaller, more manageable set of candidate compounds. Thus, the suggested pipeline can efficiently streamline the process from computational screening to clinical evaluation. Upcoming research may also focus on validating the model using independent external datasets and integrating more omics data to improve predictive accuracy. Employing sophisticated deep learning techniques and evaluating optimal compounds in the laboratory will enhance this computational framework for the discovery of novel antimicrobial compounds applicable in real-world scenarios.

## Acknowledgements

We would like to express our sincere gratitude to Jaypee University of Information Technology (JUIT) for providing the necessary infrastructure and laboratory facilities to carry out this study.

## Funding

No funding received.

## Data Availability Statement

Data available upon reasonable request from the corresponding author.

## Author Contribution Statement

Jitendraa Vashistt: Conceptualization, Supervision, Review. Megha Puri: Data curation, Methodology, Manuscript writing. Shikha Mittal: Supervision, Review, Validation. Nishant Jain: Supervision, Review.

## References

[1] H. J. Lau, C. H. Lim, S. C. Foo, and H. S. Tan, The role of artificial intelligence in the battle against antimicrobial-resistant bacteria, Curr Genet 67, 421 (2021). https://doi.org/10.1007/s00294-021-01156-5

[2] W. C. Reygaert, An overview of the antimicrobial resistance mechanisms of bacteria, AIMS Microbiol 4, 482 (2018). https://doi.org/10.3934/microbiol.2018.3.482

[3] M. Naghavi et al., Global burden of bacterial antimicrobial resistance 1990–2021: a systematic analysis with forecasts to 2050, The Lancet 404, 1199 (2024). https://doi.org/10.1016/S0140-6736(24)01867-1

[4] A. Coates, Y. Hu, R. Bax, and C. Page, The future challenges facing the development of new antimicrobial drugs, Nat Rev Drug Discov 1, 895 (2002). https://doi.org/10.1038/nrd940

[5] J. M. Stokes et al., A Deep Learning Approach to Antibiotic Discovery, Cell 180, 688 (2020). https://doi.org/10.1016/j.cell.2020.01.021

[6] F. Wong, C. de la Fuente-Nunez, and J. J. Collins, Leveraging artificial intelligence in the fight against infectious diseases, Science 381, 164 (2023). https://doi.org/10.1126/science.adh1114

[7] B. Zdrazil et al., The ChEMBL Database in 2023: a drug discovery platform spanning multiple bioactivity data types and time periods, Nucleic Acids Res 52, D1180 (2024). https://doi.org/10.1093/nar/gkad1004

[8] F. Pedregosa et al., Scikit-learn: Machine Learning in Python, J. Mach. Learn. Res. 12, 2825 (2011).

[9] V. Rodriguez-Galiano, M. Sanchez-Castillo, M. Chica-Olmo, and M. Chica-Rivas, Machine learning predictive models for mineral prospectivity: An evaluation of neural networks, random forest, regression trees and support vector machines, Ore Geology Reviews 71, 804 (2015). https://doi.org/10.1016/j.oregeorev.2015.01.001

[10] S. Sperandei, Understanding logistic regression analysis, Biochem Med (Zagreb) 24, 12 (2014). https://doi.org/10.11613/bm.2014.003

[11] R. L. Marchese Robinson, A. Palczewska, J. Palczewski, and N. Kidley, Comparison of the Predictive Performance and Interpretability of Random Forest and Linear Models on Benchmark Data Sets, J. Chem. Inf. Model. 57, 1773 (2017). https://doi.org/10.1021/acs.jcim.6b00753

[12] A. Priimagi, G. Cavallo, P. Metrangolo, and G. Resnati, The Halogen Bond in the Design of Functional Supramolecular Materials: Recent Advances, Acc Chem Res 46, 2686 (2013). https://doi.org/10.1021/ar400103r