Pharmaceutical Research
https://doi.org/10.1007/s11095-024-03699-x

ORIGINAL RESEARCH ARTICLE

Antimicrobial Activity Classification of Imidazolium Derivatives Predicted by Artificial Neural Networks

Andżelika Lorenc[1] · Anna Badura[1] · Maciej Karolak[2] · Łukasz Pałkowski[2] · Łukasz Kubik[3] · Adam Buciński[1]

Received: 14 November 2023 / Accepted: 9 April 2024

Abstract

Purpose This study assesses the Multilayer Perceptron (MLP) neural network, complemented by other Machine Learning techniques (CART, PCA), in predicting the antimicrobial activity of 140 newly designed imidazolium chlorides against Klebsiella pneumoniae before synthesis. Emphasis is on leveraging molecular properties for predictive analysis.

Methods Classification and regression decision trees (CART) identified the top 200 predictive molecular descriptors. Principal Component Analysis (PCA) reduced these descriptors to 5 components, retaining 99.57% of raw data information. Antimicrobial activity, categorized as high or low, was based on experimentally proven minimal inhibitory concentration (MIC), with a cut-point at MIC = 0.856 mol/L. A 12-fold cross-validation trained the MLP (architecture 5-12-2 with 5 Principal Components).

Results The MLP exhibited commendable performance, achieving almost 90% correct classifications across learning, validation, and test sets, outperforming models without PCA dimension reduction. Key metrics, including accuracy (0.907), sensitivity (0.905), specificity (0.909), and precision (0.891), were notably high. These results highlight the MLP model's efficacy with PCA as a high-quality classifier for determining antimicrobial activity.

Conclusions The study concludes that the MLP neural network, along with CART and PCA, is a robust tool for predicting the antimicrobial activity class of imidazolium chlorides against Klebsiella pneumoniae. CART and PCA, used in this study, allowed input variable reduction without significant information loss. High classification accuracy and associated metrics affirm the method's potential utility in pre-synthesis assessments, offering valuable insights for antimicrobial compound design.

Keywords artificial neural networks · classification · imidazolium derivatives · *klebsiella pneumoniae* · principal component analysis

Introduction

Antibiotic resistance is a pressing concern as bacteria continue to develop resistance to antimicrobial substances employed in healthcare and industrial settings. This poses a significant threat to public health, as noted by the World Health Organization (WHO) and other reputable health organizations [1–4]. One of the most alarming microorganisms, with the ability to develop multi-resistance against antimicrobial agents, is *Klebsiella pneumoniae* [5]. This Gram-negative bacteria from the *Enterobacteriaceae* family was reported by the Centers for Disease Control and Prevention (CDC) as a high-risk pathogen due to its increasing rate of antibiotic resistance and potential to spread [6]. That trend of *K. pneumoniae* strains is especially dangerous in hospitals and other healthcare

✉ Andżelika Lorenc
andzelika.lorenc@cm.umk.pl

[1] Department of Biopharmacy, Faculty of Pharmacy, Collegium Medicum in Bydgoszcz, Nicolaus Copernicus University in Toruń, dr A. Jurasza 2, 85-089 Bydgoszcz, Poland

[2] Department of Pharmaceutical Technology, Faculty of Pharmacy, Collegium Medicum in Bydgoszcz, Nicolaus Copernicus University in Toruń, dr A. Jurasza 2, 85-089 Bydgoszcz, Poland

[3] Department of Biopharmaceutics and Pharmacodynamics, Medical University of Gdańsk, Gen. J. Hallera 107, 80-416 Gdańsk, Poland

Published online: 17 April 2024

institutions where asepsis and antiseptics have a huge impact on patients' outcomes and hospitalization time [7]. The problem increased due to large-scale disinfectant usage when the COVID-19 pandemic raised [8, 9]. The *K. pneumoniae's* propensity to develop resistance against popular antibiotics and disinfection agents and the ability to create biofilm forces scientists to find new ways of searching for new antimicrobial compounds [10–12].

The identification of potential drugs can be a time-consuming and resource-intensive process. One of the problems in the area of searching for new active agents is the multitude of compounds that can be synthesized, not all of which will exhibit desired properties. However, this process can be streamlined and made more efficient, by leveraging computational chemistry and machine learning (ML) techniques in the initial research stage. This approach reduces the need for significant resources such as time, money, and chemicals. The ultimate goal is to identify the most promising compounds, which can then be further developed [13–16].

Traditional QSAR models may struggle to capture complex and non-linear relationships between chemical structures and antimicrobial activity while the non-linear nature of ML algorithms can represent these intricate connections. Those methods are able to extract patterns and insights directly from data without relying on predefined equations, making them more adaptive to the nuances of bis-imidazolium compounds' structure–activity relationships. It can also learn relevant features from the data, potentially uncovering subtle structural elements that contribute to antimicrobial activity. This is in contrast to QSAR, where feature selection is often a manual and hypothesis-driven process [17, 18].

Among various ML techniques used in drug research and development, the Artificial Neural Networks (ANNs) are one of the most promising. At the initial stages of molecular development, both 2D and 3D structures of compounds can be modeled, and subsequent to these models, molecular descriptors can be calculated, The obtained data can be used as predictors of antimicrobial activity by utilizing computational intelligence methods like ANNs [19–21]. This approach allows the preselection of potential antimicrobial compounds and the synthesis of only the most promising ones [22].

ANNs, widely employed in bioinformatics operate in a manner inspired by biological neural systems. These networks consist of interconnected artificial neurons organized into layers, and they become active under specific conditions. One prominent type of supervised ANN utilized in this field is the Multilayer Perceptron (MLP). The MLP is structured as a feed-forward network comprising a minimum of three layers of artificial neurons – input, (at least one) hidden, and output. Information flows from one layer to the next when it surpasses a predefined activation threshold determined by the activation function within each layer's neurons [23, 24].

Since MLPs are supervised neural networks with known input–output pairs, the primary objective is to minimize the discrepancy between predicted and actual outcomes which is indicator of the predictive quality of the network. Achieving this with a single learning attempt is nearly impossible. Therefore, the training process of an MLP is based on epochs, which represent repeated cycles of presenting the entire dataset to the network in order to optimize its performance. During each epoch, the strengths of connections (weights) between neurons are recalibrated (starting from initial random values) based on the errors observed in previous epochs – the network learns on its own mistakes.. This iterative process continues until the network's error reaches its minimum, enhancing its predictive accuracy [25].

The problem which can occur in this process is overfitting, which means that the network overly adjusts to the data which deprives its ability to generalize the knowledge. In this case, the network will perform well on known data but the ability to correctly predict outcomes for the new data (generalization) is limited.

The reasons that the model overfits the data include sample size, input data dimensionality, or regularization techniques. A small number of cases taking part in the learning process can lead the model to treat the irrelevant information (noise) as important characteristics of presented data. As the prospects of to increase the number of cases are mostly limited, some kind of multiple sampling, for example, different types of cross-validation (CV), is suggested to avoid this problem and to optimize the model and enhance its performance by better hyperparameters tuning [26].

Another issue is the high dimensionality of the input data, known as the 'Curse of Dimensionality' [27]. This term, introduced by Bellman indicates that the rising number of features in the dataset should be followed by an exponentially rising number of samples to maintain the balance in the model. Similarly to the previous issue, as the number of samples cannot be enhanced, the way to avoid overfitting caused by excessive dimensionality is dimensional reduction. One of the methods applied as solution is Principal Component Analysis (PCA) which, based on correlations between features, calculates the Principal Components (PCs) describing as much information included in original features as possible [28].

An alternative method to mitigate overfitting involves the utilization of regularization techniques, with one such approach being early stopping. This strategy involves the introduction of a separate dataset, referred to as a validation set. The validation set containing new, never presented to the network cases predicts the solution based on the network architecture previously built using the learning set. The function of the validation set is to check the network's

generalization ability by comparing the errors made by ANN in each epoch. When the errors made in both sets are comparable, the learning process is continued, but if the error made in the learning set decreases and the validation set error increases, the network overfits the learning data and the process should be stopped [29].

The only set which doesn't take part in the learning process is the test set. As the cases presented for this set weren't revealed for it in the learning and test stages of network modeling, the test set role is to verify the ability of the developed model to generalize its knowledge for newly presented data [30].

In this paper, the authors would like to present a presynthesis approach for the classification of bis-imidazolium derivatives using Artificial Neural Networks in combination with other machine-learning techniques.. The mechanism of action of studied compounds is closely related to their chemical structure and physical properties. Compounds interact electrostatically with the negatively charged cell surfaces of microbes and surface active compounds easily penetrate through the protein–lipid biological membranes, causing disturbances in their structural and functional coherence. That make them potentialy active against considered in this research *Klebsiella pneumoniae* strains.

Materials and Methods

Structures, Synthesis, and Molecular Descriptors

The authors decided to investigate the 140 novel bis-imidazolium compounds (quaternary ammonium salts) as potential agents [31, 32].

The structures of 140 analyzed imidazolium homologs differed in the length of the linker chain (*n* value—from 2 to 12 CH_2 groups) and the substituent chain (Fig. 1, Table I) (full list of homologs available in Table S1 in supplementary material). Designed compounds were modeled into 2D and 3D structures as neutral compounds and the quantum mechanical Density Functional Theory (DFT), B3LYP method using Pople's 6-31G as basis set was applied to optimization. Optimization was performed in the implicit solvent model SCRF (Self-Consistent Reaction Field) with dielectric constant set to that of water, with the use of the PCM (Polarizable Continuum Model).

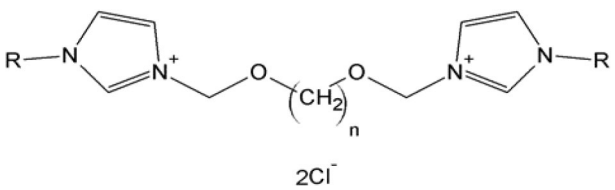

Fig. 1 General structure of analyzed imidazolium homologs.

Table I Informative System for 3,3'-(α, ω-Dioxaalkan)bis(1-Alkylimidazolium) Chlorides

Code	R
1	CH_3
2	C_2H_5
3	C_3H_7
4	C_4H_9
5	C_5H_{11}
6	C_6H_{13}
7	C_7H_{15}
8	C_8H_{17}
9	C_9H_{19}
10	$C_{10}H_{21}$
11	$C_{11}H_{23}$
12	$C_{12}H_{25}$
14	$C_{14}H_{29}$
16	$C_{16}H_{33}$

Calculations were performed employing Gaussian 09, Revision D.02 (Gaussian, Inc., Wallingford CT, USA), on a supercomputer cluster nods with 12-core Intel® Xeon® E5 v3 2.3 GHz processors. Geometry optimization calculations were carried out at the Centre of Informatics—Tricity Academic Supercomputer & Network. In further analysis, the 5270 molecular descriptors from 29 logical blocks were determined using Dragon 7.0 software (Talete, Milan, Italy) [33].

The synthesis, molecules determination using nuclear magnetic resonance spectroscopy (NMR Spectra), purity examination with thin-layer chromatography (TLC), and elemental analysis of studied imidazolium compounds is described in article by Pałkowski *et al*. [34].

Antimicrobial Activity

The minimal inhibitory concentration (MIC) value of each homolog was determined for the *K. pneumoniae* ATCC 27853 strain based on references of Clinical and Laboratory Institute (CLSI) which was also conducted and reported by mentioned above Pałkowski et.al. With reference to the standard, which was didecylmethylammonium chloride (DDAC) (a quaternary ammonium salt used for disinfection and approved by the European Economic Area as biocide [35]) with MIC = 0.856 mol/L, imidazolium compounds were categorized to high activity or low activity group. Imidazolium compounds with MICs below the MIC of the DDAC were classified in the high activity category, and those above—low activity category. This division resulted in the separation of 64 compounds classified as high activity class (45.71%) and 76 compounds classified as low activity class (54.29%) (Table S1 in supplementary material).

Data Preprocessing

In the preprocessing stage, from the 5270 descriptors obtained in previous steps with Dragon 7.0 software the ones that were constant for all the cases and/or incomplete were deleted. That left 2711 variables for further analysis. In the next step, the data was standardized to minimize the influence of variables scaling and to make them comparable.

All the calculations in this research were conducted using STATISTICA 13, provided by StatSoft Inc., Tulsa, USA.

Descriptors Selection

As each of the 140 examined molecules was characterized by a set of 2711 molecular descriptors, a volume of data deemed excessive for meaningful analysis within the framework of Artificial Neural Networks (ANN), so the authors opted to employ different ML techniques as a strategic approach to mitigate the challenges posed by highly probable overfitting issue.

1. Authors decided to use Classification and Regression Trees (CART) to select the best activity descriptive variables—descriptors. Based on chi-square statistics and p-value ($p < 0.001$) the trees have selected 200 best-fitted variables. Selected for this research 200 molecular descriptors belonged to 19 molecular blocs, and most of the descriptors were from the 2D matrix-based descriptors block (92 descriptors, 46%), 3D autocorrelations block (25 descriptors, 12.5%) and 2D autocorrelations (23 descriptors, 11.5%). (full list of selected descriptors available in Table 2 in supplementary material)
2. To limit the number of neurons and avoid data overload in the input layer, the principal component analysis (PCA) was applied. PCA is a linear method of dimension reduction, allowing to reduce the number of variables included in the network by searching the relationships between them. The correlated variables are transformed to save most of the variance of the reduced variables and create the new variable called the principal component (PC) [36] This way, 5 principal components describing 99.57% of the variability of the data contained in the 200 selected variables were extracted.

Artificial Neural Networks calculations

Obtained in previous steps 5 PCs were used as input variables to build ANN classification models and predict, based on previously presented conditions, whether the compound presents high activity or low activity against *K. pneumoniae*. Cases were randomly divided into 3 sets: learning—98 cases (70%), validation—28 cases (20%), and test—14 cases (10%). Using the Broyden-Fletcher-Goldfarb-Shanno (BFGS) [37] learning algorithm, the software automatically modeled 500 artificial neural networks from which the one with the most optimal architecture, the best predictive ability, and the lowest error (the number of correctly classified cases) was chosen as principle ANN hyperparameters setting for further analysis.

Results

Model selection, metrics and evaliation

The chosen MLP neural network was built of 5 neurons in the input layer, 12 neurons in the hidden layer, and 2 neurons in the output layer. The graph of the modeled MLP with inputs, outputs, and activation functions for hidden and input layers is shown in Fig. 2.

As the sample size was relatively small, the 12-fold cross-validation (CV) sampling technique has been applied. The percentage of correctly classified cases in every CV fold is shown in Fig. 3.

The quality of the learning, validation, and test set was considered as an average of models obtained in the cross-validation step and established at 88.61%, 90.48%, and 86.90% respectively, which means that the model correctly classified almost 90% of cases in each set. To evaluate the quality of modeled ANN the classification metrics and statistics were calculated (Table II).

To compare the performance of MLPs with and without PCA dimension reduction, the authors constructed various MLP models with different input configurations. The minimum number of inputs for MLPs without PCA was determined by the number of inputs used in MLPs with dimension reduction. The maximum number of inputs was constrained to not exceed 10% of the total cases (given the dataset size of 140 cases, the maximum number of inputs was set at 14).

The inputs were selected from the top of the list of the 200 most significant variables (according to p-value) derived initially with the CART method used for PCA dimension reduction. These variables were not subjected to further reduction and were directly used as the inputs for MLPs. The resulting models are presented in Table III.

Discussion

Current reports show that effectiveness in the search for new antimicrobial agents is estimated at around 5% [38]. Released in 2015 in 'Review of Antimicrobial Resistance' [39] pointed out that the number of drugs under development showing notable antimicrobial activity, especially against Gram-negative bacteria with broaden resistance is

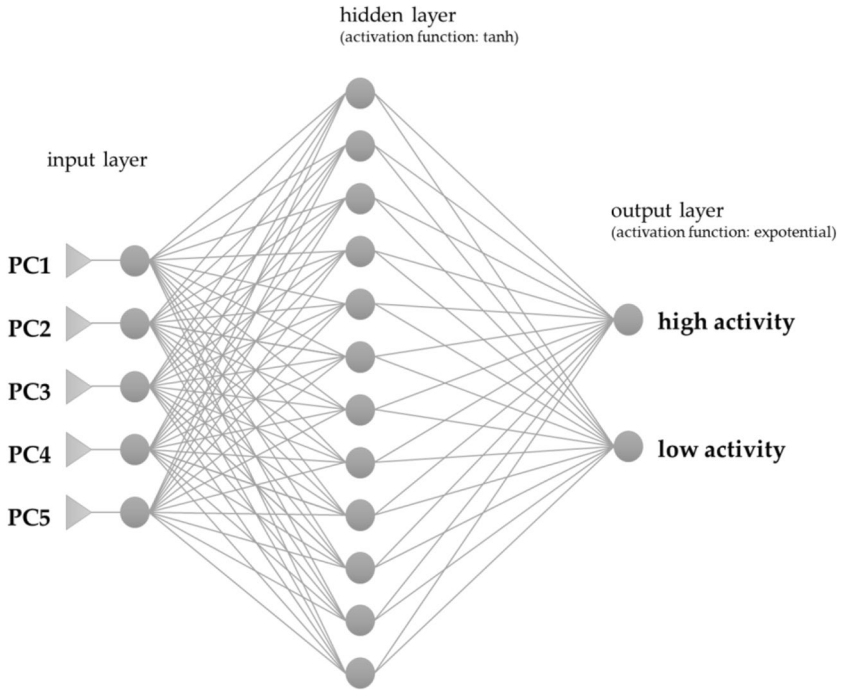

Fig. 2 MLP 5-12-2 graph

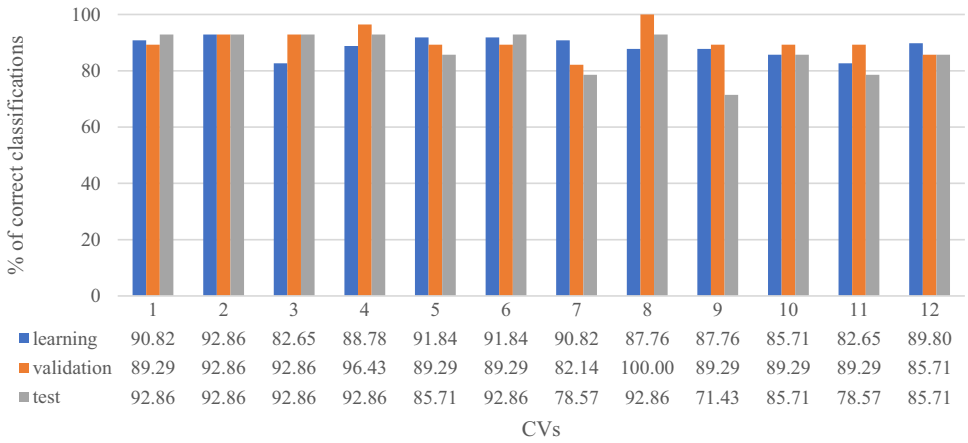

Fig. 3 Histogram and table of classification correctness (%) of learning, validation, and test sets

desperately low and it is estimated that among the vast of drugs under development only a few show potential against most resistant pathogens. In the mentioned review authors also pointed out that lowering the costs of antibiotics development is one of the major issues to focus on.

The majority of research is limited by financials, hence the most important problem seems to be the disproportion between the costs of the development of new drugs and the resulting benefits [40]. Implementing i*n silico* methods in the preliminary selection of modeled drugs allows excluding the compounds with low antibacterial properties, which improves the whole process by minimizing the number of synthesized compounds, reducing costs, and lowering chemical waste as the methods of searching for new antimicrobial agents are hindered by vast of financial and ecological restrictions nowadays. The combination of molecular modeling and ML methods seems to have a high potential for usage in the preclinical stage of drug research [41] limiting the needed resources. That also has been investigated on a large-scale dataset by Rahman *et al.* where they showed that the use of ML methods in the initial part of drug research can significantly increase the hit rate in searching for antimicrobial agents [42]. Research presented by Badura *et al.* shows that ANNs created to categorize 140 imidazolium compounds by their antimicrobial abilities have achieved over 90% accuracy for two bacterial strains – *E. coli* [43] and *S. aureus* [44] as well as for *C. albicans* fungus strain [45] using 20 molecular descriptors for each, but in contrast to the work presented in this paper, they worked on raw data.

Table II ANN Classification Model Metrics and Statistics

Model metrics			
Layer	Input	Hidden	Output
No. of neurons	5	12	2
Activation function		Tanh	Expotential
Model quality			
	Learning set	Validation set	Test set
Number of cases (%)	98 (70%)	28 (20%)	14 (10%)
Corectness (mean ± SD)	**88.61 ± 3.45**	**90.48 ± 4.65**	**86.90 ± 7.36**
Classification correctness			
	High activity	Low activity	All
General	64	76	140
Classified correctly	57 (89.06%)	70 (92.11%)	127 (90.71%)
Classified incorrectly	7 (10.94%)	6 (7.89%)	13 (9.29%)
Classification metrics			
	TPR	0.905	
	SPC	0.909	
	FPR	0.091	
	FDR	0.109	
	PPV	0.891	
	NPV	0.921	
	ACC	0.907	
	MCC	0.813	

The predictive ability of the selected in this research ANN allows classifying the compound to the group of high activity or low activity with almost 90% certainty. Used in the research MLP 5–12-2 neural network based on five Principal Components obtained good classification metrics with 0.905 sensitivity, 0.909 specificity, 0.891 precision, and 0.907 accuracy. The MCC which is considered as one of the most balanced classification metrics was established at 0.813. Also, FPR at the level of 0.091 shows a low probability of classifying compounds with low antimicrobial activity into the group of highly active homologs.

It should be kept in mind that appropriate feature selection is of great importance for developing accurate predictive models. The number of inputs should be adjusted to the analyzed dataset – with an increasing number of input features the risk of overfitting rises, but ANN with too small number of inputs could not gain the ability to generalize knowledge. Taking that into account, the authors used the PCA method for dimension reduction which allowed lowering the number of input variables to 5, maintaining over 99% of information carried by the 200 molecular descriptors selected with CART. This way of data condensation is widely used in computational intelligence problem-solving to avoid overfitting and increase algorithm performance speed. The concept of using PCA dimension reduction in combination with ML models is well known in medical sciences [46–49]. The paper by Chippalakatti *et al*., comparing different classification models with and without PCA dimension reduction shows that those with PC as inputs reached better performance than models based on raw data which was also shown in this paper [50]. The results depicted in Table III reveal that none of the MLP models trained on the original descriptors data outperformed the predictive quality of the MLP model trained with PCA inputs. Remarkably, the MLP model constructed with only 5 principal components encompassing the data from the 200 molecular descriptors exhibited better activity prediction than larger models based on the raw data. That result reinforces the decision to use PCA as a dimension-reduction method.

Conclusions

The supervised MLP NN used in this research reached high values in predicting the activity class of presented imidazole compounds. The combination of different ML techniques, including dimension reduction with PCA for multivariable datasets elevated the model performance in comparison to

Table III Predictive Qualities Comparison of Different MLP Models

	Set quality			Error function	Activation function	
	Learning	Validation	Test		Hidden layer	Output layer
PCs MLP 5-6-2	**88.61**	**90.48**	**86.90**	**SOS**	**Logistic**	**Tanh**
MLP 5-5-2	83.25	84.23	80.95	Entropy	Expotential	Softmax
MLP 6-5-2	83.59	88.99	83.93	Entropy	Tanh	Softmax
MLP 7-16-2	82.65	89.29	83.93	Entropy	Tanh	Softmax
MLP 8-13-2	84.10	88.99	80.36	SOS	Logistic	Linear
MLP 9-13-2	82.91	89.29	82.14	SOS	Exponential	Logistic
MLP 10-16-2	84.18	90.48	79.17	Entropy	Tanh	Softmax
MLP 11-10-2	85.80	90.77	83.33	Entropy	Logistic	Softmax
MLP 12-3-2	84.27	88.69	86.31	Entropy	Tanh	Softmax
MLP 13-15-2	84.01	86.61	85.71	SOS	Tanh	Logistic
MLP 14-2-2	81.89	87.80	80.36	SOS	Logistic	Logistic

the models based on raw data.. The application of ML methods gives scientists the opportunity to study the antimicrobial properties of compounds even in the earliest design and development stages.

Furthermore, use of several different methods such as PCA for dimension reduction to limit the number of ANN inputs, the learning process early stopping method supported by a validation set and use of additional test set to evaluate the generalization ability of the network allowed to avoid overfitting problem maintaining as best as possible performance.

We are aware that artificial neural network models are designed to predict, in this case, antimicrobial properties of certain homogenous groups of molecules, and for another group, there will be a need to build a different model. To optimize the accuracy of our models and to improve their predictive abilities, it is essential to conduct a comprehensive analysis of a wider range of compounds. The research, however, shows that the NN models demonstrate high potential as a preliminary approach to selecting designed with computational methods molecules. In our opinion, the use of ANN as a tool for the selection of compounds with antibacterial potential can have a significant impact on the performance of the initial phase of drug development.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11095-024-03699-x.

Author Contributions A.L. designed the project, performed the calculations, and wrote the manuscript. M.K, Ł.P, and Ł.K. provided data to analyze, supported and supervised the chemoinformatics part of the research. A.Badura and A.Buciński. supported and supervised the ML part of the research. All the authors reviewed the manuscript.

Funding Funding based on Elsevier and Nicolaus Copernicus University in Toruń (Elsevier SIS: 2055) agreement to provide Open Access publishing.

Data Availability The datasets generated during and/or analyzed during the current study are available from the corresponding author upon reasonable request.

Declarations

Conflict of Interest The authors declare no competing financial interest.

References

1. WHO Regional Office for Europe/European Centre for Disease Prevention and Control. Antimicrobial resistance surveillance in Europe 2022–2020 data. Copenhagen: WHO Regional Office for Europe; 2022.
2. Global antimicrobial resistance and use surveillance system (GLASS) report 2021. Geneva: World Health Organization; 2021. Licence: CC BY-NC-SA 3.0 IGO.
3. TalebiBezminAbadi A, Rizvanov AA, Haertlé T, Blatt NL. World Health Organization Report: Current Crisis of Antibiotic Resistance. Vol. 9, BioNanoScience. Springer New York LLC; 2019. p. 778–88.
4. Murray CJ, Ikuta KS, Sharara F, Swetschinski L, Robles Aguilar G, Gray A, *et al*. Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis. Lancet. 2022;399(10325):629–55.
5. Wang G, Zhao G, Chao X, Xie L, Wang H. The characteristic of virulence, biofilm and antibiotic resistance of klebsiella pneumoniae. Int J Environ Res Public Health. 2020;17:1–17.
6. CDC. Antibiotic Resistance Threats in the United States, 2019. Atlanta, GA: U.S. Department of Health and Human Services, CDC; 2019.
7. Naylor NR, Atun R, Zhu N, Kulasabanathan K, Silva S, Chatterjee A, *et al*. Estimating the burden of antimicrobial resistance: a systematic literature review. Antimicrob Resist Infect Control. 2018;7(58).
8. Getahun H, Smith I, Trivedi K, Paulin S, Balkhy HH. Tackling antimicrobial resistance in the COVID-19 pandemic. Vol. 98, Bulletin of the World Health Organization. World Health Organization; 2020.
9. Clancy CJ, Buehrle DJ, Nguyen MH. PRO: the COVID-19 pandemic will result in increased antimicrobial resistance rates. JAC Antimicrobial Resistance. 2020;2(3).
10. Navon-Venezia S, Kondratyeva K, Carattoli A. Klebsiella pneumoniae: a major worldwide source and shuttle for antibiotic resistance. FEMS Microbiol Rev. 2017;41(3):252–75.
11. Adler A, Katz DE, Marchaim D. The continuing Plague of extended-spectrum β-lactamase–producing Enterobacteriaceae Infections. Infect Dis Clin N Am. 2016;30(2):347–75.
12. Antoniadou A, Kontopidou F, Poulakou G, Koratzanis E, Galani I, Papadomichelakis E, *et al*. Colistin-resistant isolates of Klebsiella pneumoniae emerging in intensive care unit patients: first report of a multiclonal cluster. J Antimicrob Chemother. 2007;59(4):786–90.
13 Kore PP, Mutha MM, Antre RV, Oswal RJ, Kshirsagar SS. Computer-aided drug design: an innovative tool for modeling. Open J Med Chem. 2012;02(04):139–48.
14. Lionta E, Spyrou G, Vassilatis D, Cournia Z. Structure-based virtual screening for drug discovery: principles, applications and recent advances. Curr Top Med Chem. 2014;14(16):1923–38.
15. Liu G, Stokes JM. A brief guide to machine learning for antibiotic discovery. Curr Opin Microbiol. 2022;1(69): 102190.
16. Macalino SJY, Gosu V, Hong S, Choi S. Role of computer-aided drug design in modern drug discovery. Arch Pharm Res. 2015;38(9):1686–701.
17. Dara S, Dhamercherla S, Jadav SS, Babu CM, Ahsan MJ. Machine learning in drug discovery: a review. Artif Intell Rev. 2022;55(3):1947–99.
18. Tsou LK, Yeh SH, Ueng SH, Chang CP, Song JS, Wu MH, et al. Comparative study between deep learning and QSAR classifications for TNBC inhibitors and novel GPCR agonist discovery. Sci Rep. 2020;10:1–11.
19. Grisoni F, Consonni V, Todeschini R. Impact of molecular descriptors on computational models. Methods Mol Biol. 2018;1825:171–209.

20. Mauri A, Consonni V, Todeschini R. Molecular descriptors. Handbook of Computational Chemistry. 2017;2065–93.
21. Tropsha A. Predictive quantitative structure–activity relationship modeling. Compr Med Chem II. 2007;4:149–65.
22. Torres MDT, de la Fuente-Nunez C. Toward computer-made artificial antibiotics. Curr Opin Microbiol. 2019;1(51):30–8.
23. Murtagh F. Multilayer perceptrons for classification and regression. Neurocomputing. 1991;2(5–6):183–97.
24. Popescu MC, Balas VE, Perescu-Popescu L, Mastokaris N. Multilayer perceptron and neural networks. Wseas Transactions on Circuits and Systems. 2009;8(7):579–88.
25. Camacho Olmedo MT et al. (eds). Geomatic approaches for modeling land change scenarios. Lect Notes Geoinf Cartogr. https://doi.org/10.1007/978-3-319-60801-3_27.
26. Bates S, Hastie T, Tibshirani R. Cross-validation: what does it estimate and how well does it do it? J Am Stat Assoc. 2023.
27. Bellman R. Dynamic Programming. Dover Publications; 1957.
28. Hasan BMS, Abdulazeez AM. A review of principal component analysis algorithm for dimensionality reduction. J Soft Comput Data Min. 2021;2(1):20–30.
29. Lawrence S, Tsoi AC, Giles CL. Lessons in neural network training: overfitting may be harder than expected. In: Fourteenth National Conference on Artificial Intelligence. AAAI Press; 1997. pp. 540–5.
30. Tadeusiewicz R, Lula P. Wprowadzenie do sieci neuronowych. Kraków: StatSoft Polska Sp. z o.o.; 2001.
31. Walsh SE, Maillard JY, Russell AD, Catrenich CE, Charbonneau DL, Bartolo RG. Activity and mechanisms of action of selected biocidal agents on Gram-positive and -negative bacteria. J Appl Microbiol. 2003;94(2):240–7.
32. Bharate SB, Thompson CM. Antimicrobial, antimalarial, and antileishmanial activities of mono- and bis-quaternary pyridinium compounds. Chem Biol Drug Des. 2010;76(6):546–51.
33. Pałkowski Ł, Karolak M, Błaszczyński J, Krysiński J, Słowiński R. Structure-activity relationships of the imidazolium compounds as antibacterials of staphylococcus aureus and pseudomonas aeruginosa. Int J Mol Sci. 2021;22(15):7997.
34. Pałkowski Ł, Błaszczyński J, Skrzypczak A, Błaszczak J, Kozakowska K, Wróblewska J, et al. Antimicrobial Activity and SAR Study of New Gemini Imidazolium-Based Chlorides. Chem Biol Drug Res. 2013;(83):278–88.
35. European Chemicals Agency. Imidazole - Substance Infocard. https://echa.europa.eu/substance-information/-/substanceinfo/100.005.473. Accessed 30 Sep 2023.
36. Kherif F, Latypova A. Chapter 12 - Principal component analysis. In: Mechelli A, Vieira S, editors. Machine Learning. Academic Press; 2020. p. 209–25.
37. Kelley CT. The BFGS Method. In: Iterative methods for optimization. Society for Industrial and Applied Mathematics; 1999. p. 71–86.
38 Payne DJ, Gwynn MN, Holmes DJ, Pompliano DL. Drugs for bad bugs: confronting the challenges of antibacterial discovery. Nat Rev Drug Discov. 2007;6:29–40.
39. The Review on Antimicrobial Resistance. Securing new drugs for future generations: the pipeline of antibiotics. https://amr-review.org/sites/default/files/SECURING%20NEW%20DRUGS%20FOR%20FUTURE%20GENERATIONS%20FINAL%20WEB_0.pdf. Accessed 2 Feb 2023.
40. Årdal C, Balasegaram M, Laxminarayan R, McAdams D, Outterson K, Rex JH, *et al*. Antibiotic development — economic, regulatory and societal challenges. Nat Rev Microbiol. 2020;18(5):267–74.
41. Souza Leite M, Alles de Jesus A, Leite Araujo PJ, Ferreira dos Santos B. Chapter 4 - Advances in drug development with the application of artificial intelligence. In: Török M, editor. Contemporary Chemical Approaches for Green and Sustainable Drugs. Elsevier; 2022. p. 69–88. (Advances in Green and Sustainable Chemistry).
42. Zisanur Rahman ASM, Liu C, Sturm H, Hogan AM, Davis R, Hu P, *et al*. A machine learning model trained on a high-throughput antibacterial screen increases the hit rate of drug discovery. PLoS Comput Biol. 2022;18(10):e1010613.
43. Badura A, Krysiński J, Nowaczyk A, Buciński A. Application of artificial neural networks to prediction of new substances with antimicrobial activity against Escherichia coli. J Appl Microbiol. 2020;130:40–9.
44. Badura A, Krysiński J, Nowaczyk A, Buciński A. Prediction of the antimicrobial activity of quaternary ammonium salts against Staphylococcus aureus using artificial neural networks. Arab J Chem. 2021;14(7):103233.
45. Badura A, Krysiński J, Nowaczyk A, Buciński A. Application of artificial neural networks to the prediction of antifungal activity of imidazole derivatives against Candida albicans. Chemom Intell Lab Syst. 2022;15:222.
46. Sudharsan M, Thailambal G. Alzheimer's disease prediction using machine learning techniques and principal component analysis (PCA). Mater Today Proc. 2023;81:182–90.
47. Reddy KVV, Elamvazuthi I, Aziz AA, Paramasivam S, Chua HN. Heart disease risk prediction using machine learning with principal component analysis. 2020 8th International Conference on Intelligent and Advanced Systems (ICIAS), Kuching, Malaysia, 2021, pp. 1–6. https://doi.org/10.1109/ICIAS49414.2021.9642676.
48. Shimpi P, Shah S, Shroff M, Godbole A. A machine learning approach for the classification of cardiac arrhythmia. 2017 International Conference on Computing Methodologies and Communication (ICCMC), Erode, India; 2017. pp. 603–607. https://doi.org/10.1109/ICCMC.2017.8282537.
49. Jaques LE, Depoian AC, Xie D, Bailey CP, Guturu P. A machine learning approach to medical data identification through principal component analysis. Big Data III: Learning, Analytics, and Applications. SPIE. 2021;11730:7–15.
50. Chippalakatti S, Renumadhavi CH, Pallavi A. Comparative Review on the Machine Learning Algorithms for Medical Data. 2022 6th International Conference on Computation System and Information Technology for Sustainable Solutions (CSITSS), Bangalore, India; 2022. pp. 1-6. https://doi.org/10.1109/CSITSS57437.2022.10026396.