JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

pubs.acs.org/jcim Article

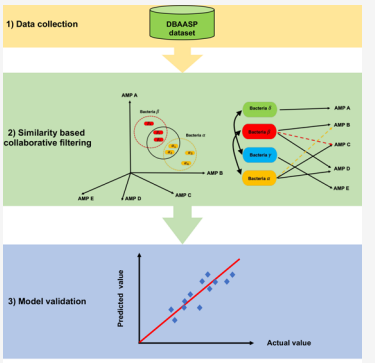
Predicting Antimicrobial Activity for Untested Peptide-Based Drugs Using Collaborative Filtering and Link Prediction

Angela Medvedeva, Hamid Teimouri, and Anatoly B. Kolomeisky*

Cite This: J. Chem. Inf. Model. 2023, 63, 3697−3704 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: The increase of bacterial resistance to currently available antibiotics has underlined the urgent need to develop new antibiotic drugs. Antimicrobial peptides (AMPs), alone or in combination with other peptides and/or existing antibiotics, have emerged as promising candidates for this task. However, given that there are thousands of known AMPs and an even larger number can be synthesized, it is impossible to comprehensively test all of them using standard wet lab experimental methods. These observations stimulated an application of machine-learning methods to identify promising AMPs. Currently, machine learning studies combine very different bacteria without considering bacteria-specific features or interactions with AMPs. In addition, the sparsity of current AMP data sets disqualifies the application of traditional machine-learning methods or makes the results unreliable. Here, we present a new approach, featuring neighborhood-based collaborative filtering, to predict with high accuracy a given bacteria's response to untested AMPs based on similarities between bacterial responses. Furthermore, we also developed a complementary bacteria-specific link prediction approach that can be used to visualize networks of AMP-antibiotic combinations, enabling us to propose new combinations that are likely to be effective.



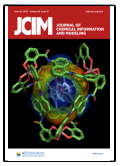
INTRODUCTION

Bacterial resistance to conventional antibiotics is becoming a major global health threat in the 21st century.[1] Specifically, there are several types of superbugs (bacterial strains resistant to all known antibiotics), including *A. baumannii*, *P. aeruginosa*, and *Enterobacteriaceae*, that have been listed as highest-priority targets for the future antibiotic drugs by the World Health Organization.[2] These bacteria are resistant to the most effective classes of conventional antibiotics available, including the broad-spectrum carbapanems: imipenem, meropenem, ertapenem, and doripenem.[3,4] It is known that these antibiotics target the intercellular environment and deactivate the enzymes that inhibit cell death, leading to the destruction of the bacteria cell (autolysis). They are meant to be administered as a last resort to preserve their efficacy, minimize toxicity, and avoid the development of bacteria resistance. However, frequent exposures of bacteria to these antibiotics have already led to the development of resistance via mutations that stimulated the production of different enzymes to replace the deactivated autolysis inhibitor enzymes. Other resistance mutations are associated with the creation of new efflux pumps to push the antibiotics out of the cell and also include the porin mutations that prevent the antibiotics from penetrating the cell walls.[3,4]

Antimicrobial peptides (AMPs), which include several classes of short and mostly positively charged peptides, have been suggested as possible alternatives or adjuncts to conventional antibiotics.[5] Their positively charged groups bind efficiently to negatively charged segments of bacterial membranes, after which they stimulate the formation of pores in the cell walls. These pores eventually lead to the cell deaths.[6] AMPs play a critical role in the innate immune system as the first line of defense against invading pathogens, but they can also potentially serve as the most effective last line of defense.[7] Furthermore, two or more types of AMPs in combination, and also one type of AMPs in combination with an antibiotic, have shown the potential to be highly effective drugs with lower toxicity than the individual AMPs or antibiotics.[8] Remarkably, a single AMP type that effectively halts the bacterial growth can continue to exert significant antimicrobial activity when used in lower concentrations in combination with an ineffective AMP or resistant antibiotic, essentially sensitizing the bacteria to the antibiotic to which they were previously resistant.[9] Moreover, combinations of AMPs can remain effective in spite of bacteria resistance to individual AMPs.[10]

While decades of experimental studies have tested multiple AMP-antibiotic combinations against the antibiotic-resistant bacteria, it is realistically not feasible to test quickly all possible combinations against all possible bacteria. The process from drug production and discovery to testing and releasing into the

Received: January 27, 2023
Published: June 12, 2023



https://doi.org/10.1021/acs.jcim.3c00137
J. Chem. Inf. Model. 2023, 63, 3697−3704

market is resource-costly, which is why various machine-learning methods, including recommendation systems, have been utilized to accelerate the identification of promising drug candidates.[11−13] With machine-learning methods, the existing data on AMP antimicrobial activity can be leveraged to narrow the pool of AMP combinations that need to be tested.[14,15] However, the main challenge is that when considering individual bacteria the data on AMP and AMP-antibiotic efficacy are highly sparse. The data are even sparser for AMP-AMP combinations, which may contribute to the lack of investigations on elucidating molecular mechanisms of antibacterial action for existing AMP-AMP combinations and predicting activity of new ones.[16] Previous studies focused on the efficacy of single AMP types alone and were limited in combining very different bacteria for their analyses [e.g., broad categories of Gram-positive vs Gram-negative bacteria or active against one type vs inactive against all tested types of bacteria[17,18]]. Although certain AMP types and AMP combinations have shown broad-spectrum antimicrobial activity, there are key differences between bacteria, for example in membrane composition[19] or morphology,[20] that can affect the efficacy of a single AMP type or AMPs in combination and render some bacteria but not others resistant to the antimicrobial agents. It is important to develop new AMP-based antibiotic drugs by taking into account the specificity of the target bacterial species.

There are different methods for predicting functions of different proteins including sequence-based methods,[21,22] embedding-based methods,[23,24] KEGG and GO enrichment score-based methods,[25] and network-based methods.[26] Specifically, in recent years, several computational frameworks have been developed for prediction of minimal inhibitory concentration (MIC) values of AMPs targeting different bacteria.[17,27] However, such approaches have two main disadvantages. First, they are typically based on establishing a structure−activity relationship, aiming at predicting the MIC values from the peptide sequences. However, the MIC values, which quantify the bacteria's response to AMPs, can depend on a combination of several complex phenomena such as interactions of AMPs with membranes or interactions of AMPs with intracellular components.[28] Therefore, the MIC values cannot be predicted solely based on the structural features of peptides. Another disadvantage of current MIC-prediction machine-learning models is that they were trained by data sets which included several different types of bacteria. This selection of collapsed data sets ignores the role of bacterial membranes since bacteria have different membrane architectures. Thus, the predictions of such machine-learning methods cannot be reliable because one peptide can be effective against one bacterium (low MIC), but it might not be effective against another bacterium (high MIC). However, more advanced machine-learning-based methods were developed recently that take into account the differences in AMP antimicrobial activity between different strains and how physicochemical features of AMPs are related to these differences.[29,30]

One method that takes into account such specificity is the user-based recommendation system, which was originally developed for predicting unknown movie ratings for one user based on similar users who watched and rated those movies.[31] In particular, a recommendation-system method called neighborhood-based collaborative filtering (NBCF) has been successfully applied in biomedical research to predict protein−protein interactions,[32] new anticancer drugs,[33] drug side effects,[34] and for treatments of diseases for which existing drugs have not been tested and could potentially be effective.[35] NBCF seemed to be an ideal method also for investigating AMP antimicrobial activity across numerous bacteria because similarities in responses of different bacteria to the same AMPs can be used to predict responses of bacteria to an untested AMP type or AMP combination. One critical advantage of NBCF over other methods is that the minimum sparsity requirement for the data set is quite low, particularly as the number of users (bacteria) and items (AMP types or AMP combinations) increases. Moreover, the accuracy of predictions is user-independent and data-dependent, so as the available amount of data on AMP efficacy increases, NBCF will be more useful and can provide reliable predictions that could not be made previously for a greater number of bacteria. The predictions of NBCF can then be fed to additional machine-learning models to augment the size of the training data and to increase the prediction accuracy for aspects in which the NCBF method is limited.[36] However, a limitation of NBCF is that the method cannot lead to the generation of predictions for new AMP types or AMP combinations that have not yet been tested on any bacteria. Thus, NBCF is optimal as a data-augmentation method to be used with other models, for example, the link prediction method, which can recommend new combinations to be tested.

In this work, we adopted the user-based NBCF method and applied it to predict antimicrobial activity of individual AMP types and AMP-antibiotic combinations. We found high predictive accuracy for known activity ratings for numerous bacteria. Our predictions can be tested since specific quantitative estimates that can be measured in experiments are obtained. Furthermore, we visualized AMPs and antibiotics as nodes in a network with edges representing synergistic or nonsynergistic combined activity and predicted future edges representing AMP-antibiotic and antibiotic-antibiotic combinations using a link prediction algorithm.

MATERIALS AND METHODS

Fractional Inhibitory Concentration (FIC). Antimicrobial activity can be evaluated through a parameter known as minimum inhibitory concentration (MIC), which is defined as the minimum concentration of an antimicrobial required for a population of bacteria to stop growing.[37] Synergy between antimicrobials, including AMPs and antibiotics, is commonly tested through a checkerboard assay in which the concentration of one antimicrobial is kept constant while the other is increased until the combination is effective in inhibiting bacteria growth, as schematically presented in Figure 1.[38]

One of the ways to quantify the synergy between antimicrobial peptides and antibiotics is through evaluation of fractional inhibitory concentration (FIC) coefficients. For two antimicrobial agents, including an AMP and an antibiotic, labeled as 1 and 2, acting individually or in combination, the FIC indexes are defined as

$$FIC_1 = \frac{MIC_{(1\text{ in presence of }2)}}{MIC_{(1\text{ alone})}} = \frac{C_1}{C_{1,MIC}}$$
$$FIC_2 = \frac{MIC_{(2\text{ in presence of }1)}}{MIC_{(2\text{ alone})}} = \frac{C_2}{C_{2,MIC}} \quad (1)$$

The combined FIC index is then given by $FIC = FIC_1 + FIC_2$ The antibiotic efficiency of the AMP combination is quantified

https://doi.org/10.1021/acs.jcim.3c00137
J. Chem. Inf. Model. 2023, 63, 3697−3704

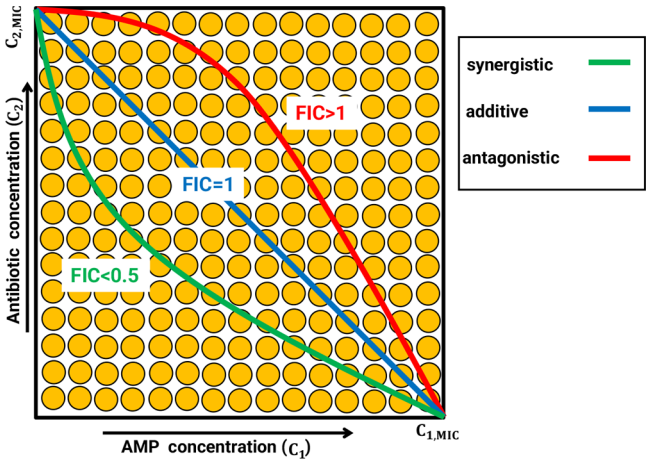


Figure 1. Schematic representation of a synergy assay checkerboard for measuring cooperativity of antimicrobial combinations. The concentrations of antimicrobial peptide (AMP) and antibiotic increase in x and y directions, respectively. The minimum inhibitory concentration of AMP (antibiotic) in the absence of antibiotic (AMP) is denoted as $C_{1,MIC}$ ($C_{2,MIC}$).

then by the value of the combined FIC index as follows: $FIC < 1$ indicates the synergism, $FIC = 1$ corresponds to additivity, and $FIC > 1$ indicates the antagonism. Note, however, that for practical applications a stricter condition with $FIC \leq 0.5$ is typically applied to identify the strongest synergistic combinations of antibiotic drugs.[39,40] This is a common procedure in the field due to some variability in measuring of antimicrobial properties.

Data Set Structure. The data were extracted from the DBAASP database.[41] We did not remove AMPs with nonstandard amino acids, so our model provides an advantage of being able to predict antimicrobial activity for these peptides. The selected data for each strain of bacteria were collected under the same experimental conditions. Multiple MIC values for the same strain and AMP collected under the same experimental conditions were averaged into one value if the values were less than 2 standard deviations from their mean. The MIC values were converted to the same units, μg/ml. It was ensured that all bacteria strains in the data set were tested on at least ten of the same AMPs. Moreover, all included MIC values were below 100 to exclude cases in which bacteria were resistant or insensitive to AMPs.[18] In other words, the included bacteria strains were sensitive to various degrees to the included AMPs or components of AMP-based combinations. To ensure that there was sufficient data for analysis,[42] we only selected the bacteria strains that had MIC values for at least 12 AMPs. The final data set for single AMPs included 111 bacteria and 1045 AMPs. For AMP-antibiotic combinations, we only considered *E. coli* for the link method but not collaborative filtering. For collaborative filtering, there was not a sufficient number bacteria with responses for at least 12 AMP-antibiotic combinations (there were only 4 bacteria that met this criterion). Combinations were considered synergistic when reported FIC values were ≤ 0.5, and combinations were considered nonsynergistic if reported FIC values were >0.5. For the 192 synergistic AMP-antibiotic pairs, there were 98 unique AMPs and 36 unique antibiotics, and there were 173 nonsynergistic AMP-antibiotic pairs with 87 unique AMPs and 35 unique antibiotics. There was sufficient data to make predictions but not sufficient data to evaluate the model performance, so the results are included in the Supporting Information.

Nearest-Neighbor Based Collaborative Filtering. For a given bacterium, depending on their patterns of ratings for individual AMPs, we can consider several nearest neighbors in the d-dimensional space (as visualized in Figure 2a). The nearest neighbors for each bacterium can be determined using a proper similarity measure, which is a function that quantifies the similarity between two objects. Then, a threshold can be selected such that only pairs with similarity at or exceeding the threshold are considered nearest neighbors. For example, a correlation greater than 0.90 is implemented in our analysis,

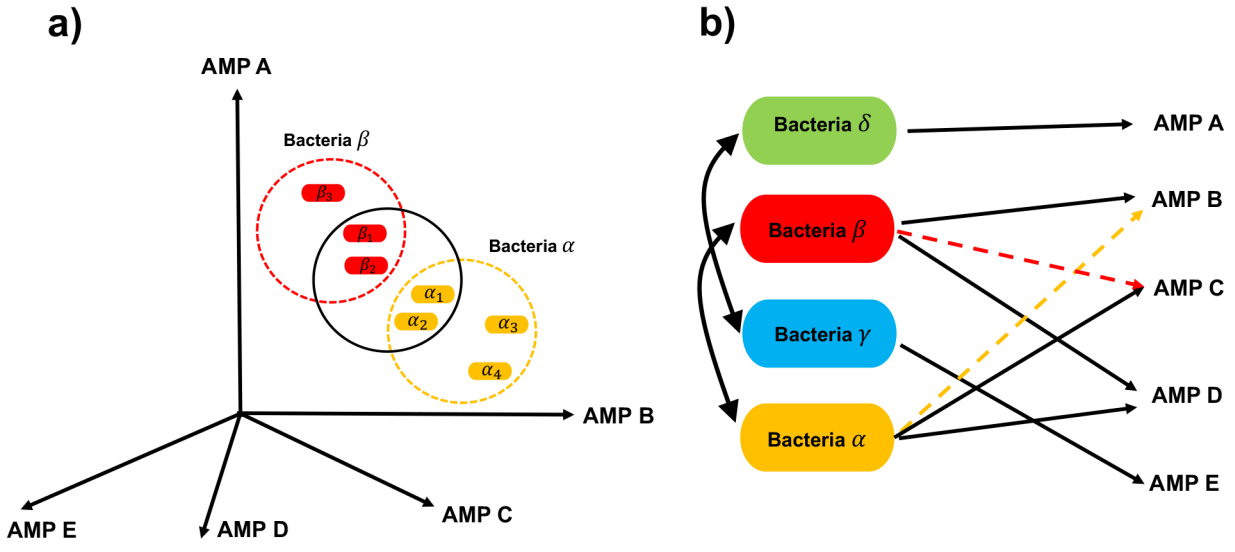


Figure 2. A schematic view of the nearest-neighbor based collaborative filtering approach. a) Each bacterium can be viewed as a point in the d-dimensional space where each dimension represents an AMP or AMP-antibiotic combination. Different strains of a certain species of bacteria α and β are shown in the dashed circles. b) Bacteria-based collaborative filtering. Bacterium α and Bacterium β are nearest neighbors because they are both inhibited by AMP D. Thus, based on similarity between Bacteria α and β, we can predict the rating for Bacterium β in response to AMP combination C (red dashed line) and the rating for Bacterium α in response to AMP combination B (orange dashed line).

https://doi.org/10.1021/acs.jcim.3c00137
J. Chem. Inf. Model. 2023, 63, 3697−3704

but in principle, any value of the threshold can be used. We opted to use the Pearson correlation coefficient (r), which is an important measure that captures the similarity between the properties of two bacteria α and β by quantifying the correlations.[31] It is given by

$$Sim(\alpha, \beta) = r = \frac{\sum_{i=1}^{N_a}(R_{\alpha,i} - M_\alpha)(R_{\beta,i} - M_\beta)}{\sqrt{\sum_{i=1}^{N_a}(R_{\alpha,i} - M_\alpha)^2}\sqrt{\sum_{i=1}^{N_a}(R_{\beta,i} - M_\beta)^2}} \quad (2)$$

where $R_{j,i}$ is the rating of AMP i for bacteria j, M_j is the average rating for bacteria j $\left(M_j = \frac{1}{N_a}\sum_{i=1}^{N_a} R_{j,i}\right)$, and N_a is the total number of available ratings. Using this similarity function, we can predict the unknown ratings for a given bacterium, α,[31] and AMP, i

$$\hat{R}_{\alpha,i} = M_\alpha + \frac{\sum_{k=1}^{N_n} Sim(\alpha, k)(R_{k,i} - M_k)}{\sum_\beta |Sim(\alpha, k)|} \quad (3)$$

where N_n is the number of nearest neighbors for bacterium α with a Pearson's correlation coefficient greater than 0.9. We divided the data into an 80/20 training/test set to evaluate model performance. There were 2585 MIC values in the training set, and there were 641 in the test set. To test the quality of our predictions, several quantitative measures can be utilized. The one that we used is the coefficient of determination, denoted as R^2, which is defined as

$$R^2 = 1 - \frac{\sum_{i=1}^{N_{cv}}(R_i^{(predicted)} - R_i^{(actual)})^2}{\sum_{i=1}^{N_{cv}}(R_i^{(predicted)} - \bar{R})^2} \quad (4)$$

where $R_i^{(actual)}$ and $R_i^{(predicted)}$ are actual and predicted ratings for AMP i, respectively, for a given bacterium α, and N_{cv} is the number of actual vs predicted values in the cross-validation fold that can be compared. Also, $\bar{R} = \frac{1}{N_a}\sum_{i=1}^{N_a} R_i^{(actual)}$ is the average of the actual ratings for a given bacterium. The closer R^2 is to unity, the better the prediction. R^2 is a standard method of evaluating statistical models, especially regression.[43] Higher R^2 reflects greater predictive accuracy. This is particularly important because the values of MIC (describing individual AMP) have different ranges. In addition, we also utilized p-values statistical analysis to evaluate the quality of our predictions.

There are several advantages of applying the neighborhood-based collaborative filtering for our data set. First, the properties of synthetic peptides cannot be analyzed by established AMP properties extractors, and therefore, these properties cannot be used to predict antimicrobial activity. Because the user-based recommendation systems rely on similarities in activity between users (bacteria), synthetic peptides can be included in the analysis, and their antimicrobial activity values can be predicted. Second, the antimicrobial activity values may differ according to species and strains of bacteria. While AMP antimicrobial activity has been determined for a wide range of bacteria, the AMP antimicrobial activity values for individual bacteria, when considered separately, are limited. Previous studies in their analysis frequently combined across the species of bacteria without taking into account the differences in membranes and other bacterial characteristics that could affect their antimicrobial activity. Our theoretical method allows us to estimate the antimicrobial activity for *specific* bacterial species. Our collaborative filtering approach is summarized in Figure 2.

Link Prediction. The neighborhood-based collaborative filtering approach is limited in that predictions can only be made for existing AMPs or AMP-based combinations that have been tested on a sufficiently large number of bacteria strains. For example, there is a case in which for a target bacterium strain the response to a specific, target AMP or AMP-based combination is unknown. If the target bacterium shows similar responses to the same AMPs as the bacteria that have been tested with the target AMP, a prediction can be made for the response of the target bacterium to the target AMP. In the case of data for AMP-antibiotic combinations from the DBAASP data set, there was not sufficient data to evaluate predictions because most AMP-antibiotic combinations were tested on a limited number of bacteria, and each combination was tested with different bacteria. Thus, the collaborative filtering approach is useful only for specific cases. In this section, we propose to utilize a complementary link method for predicting the possible synergistic AMP-antibiotic combinations targeting specific bacteria.[44] This approach is particularly useful because N unique AMPs can be paired into $N(N-1)/2$ different combinations. To proceed further, let us visualize the rating pattern of a given bacteria for different AMP-antibiotic combinations as a network, in which each AMP can be represented as a node and an edge between two nodes would indicate their synergistic combination. A snapshot of the corresponding network is visualized in Figure 3. The full networks for synergistic and nonsynergistic AMP-antibiotic combinations are shown in Figures S1 and S2, respectively. In our data set, only a small subset of possible AMP-antibiotic combinations has been tested, and we would like to predict the potential synergistic and nonsynergistic combinations. We

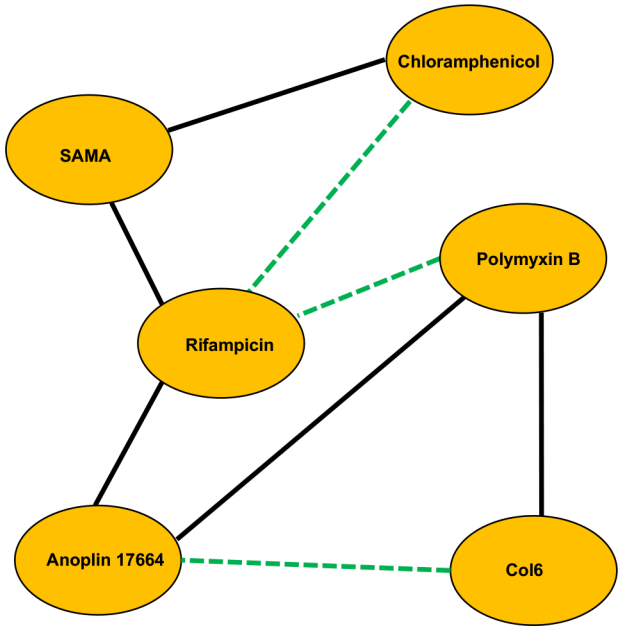


Figure 3. A schematic representation of the network of synergistic AMP-antibiotic combinations for a specific bacterium. Solid black lines connecting two nodes indicate that the corresponding AMPs synergize, while dashed green lines suggest potential synergistic AMP-antibiotic pairs. It is more likely that two AMPs will synergize if they have a common neighbor.

https://doi.org/10.1021/acs.jcim.3c00137
J. Chem. Inf. Model. 2023, 63, 3697−3704

focused on the *E. coli* bacterium since the link prediction must be conducted for one bacterium at a time.

The link prediction method aims to determine the possibility of probable links in an incomplete graph. There are different methods for prediction of most probable links. Two widely used methods for link prediction are common neighbors and preferential attachment. Let us consider an arbitrary network in which two nodes x and y are not linked. We define $N(x)$ and $N(y)$ as a set of neighbors of nodes x and y, respectively. Then, common neighbors of node x and node y are given by the intersection of the two sets:

$$\text{common neighbors} = N(x) \cap N(y) \tag{5}$$

An alternative method for the link prediction is preferential attachment, which is defined as the multiplication of the number of neighbors of node x by the number of neighbors of node y:

$$\text{preferential attachment} = N(x)N(y) \tag{6}$$

Therefore, each possible link between any two nodes can be characterized by the above characteristics. The link with the highest score would be considered as the most probable link in the complete network.

RESULTS

Predicting Individual AMP Activity. In our analysis, the predictions are minimum inhibitory concentration (MIC) values of antimicrobial activity of AMPs against each bacterium strain. We selected the strains for which training accuracy was sufficiently high, specifically training data sets with R^2 values above 0.8 and significant p-values < 0.01 (Figure 4). We

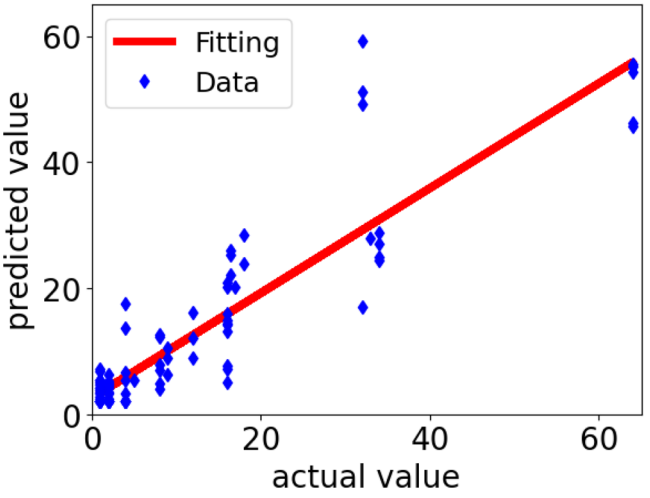


Figure 4. Collaborative filtering prediction validation for individual AMP antimicrobial activity training data: scatterplot of 95 actual vs predicted values, training subset for model validation, with the line of best fit, $R^2 = 0.82$ and $p = 8.24 \times 10^{-37}$. The fitted regression equation is $R_i^{(predicted)} = aR_i^{(actual)} + b$. The estimated fitting parameters are $a = 0.83$ and $b = 2.54$.

evaluated the corresponding test data sets for these strains to estimate model performance and found comparable accuracy to the training data sets, $R^2 = 0.82$ and $p = 8.24 \times 10^{-37}$ (see Figure 5).

However, Tables S1 and S2, which show bacteria-strain-specific training and test accuracy, indicate that while for most strains test accuracy was comparable or greater than training

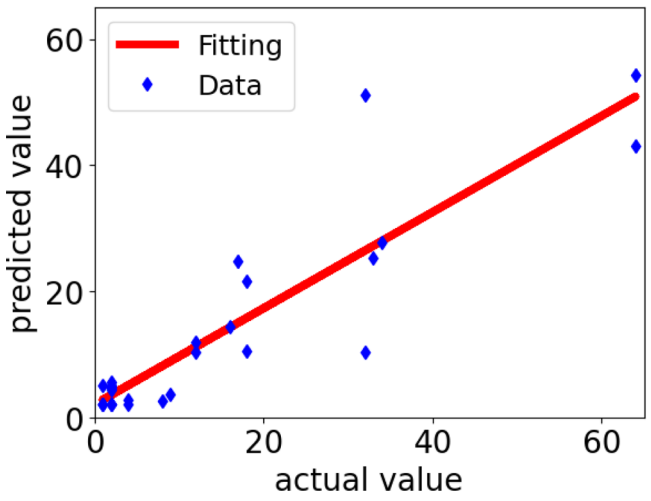


Figure 5. Collaborative filtering prediction validation for the individual AMP antimicrobial activity test set: scatterplot of 29 actual vs predicted values with the line of best fit, $R^2 = 0.81$ and $p = 3.52 \times 10^{-11}$. The fitted regression equation is $R_i^{(predicted)} = aR_i^{(actual)} + b$. The estimated fitting parameters are $a = 0.76$ and $b = 2.0$.

accuracy, for *Escherichia coli ATCC 25922 del-waaE* and *Listeria ivanovii CIP 12510* test accuracy was significantly lower than training accuracy and the test prediction accuracy was poor ($R^2 = 0.0$ and 0.5, respectively). It is unclear why for these strains predictive accuracy was lower, but notably, for the strains with training $R^2 > 0.9$, test accuracy was equally high or higher, so it is likely that a higher threshold of R^2 for training accuracy is required for reliable test accuracy. Moreover, the test accuracy could be more reliable with larger test data sizes.

DISCUSSION

Using the method of nearest-neighbor collaborative filtering, we predicted the antimicrobial activity, quantified in the MIC values, for bacteria in response to individual AMPs. The high predictive accuracy of the approach was shown in terms of the coefficient R^2. NBCF can be used as a data-augmentation step before performing feature selection, which can specify which features of AMPs lead to efficacy against different bacteria. Thus, the model can provide important physical-chemical insights into the mechanisms of AMPs, particularly the ability of AMPs to effectively inhibit the growth of multiple pathogens. Identifying specific AMPs that can kill the given bacteria can help to understand what properties are critically important for this to occur.

AMPs are a diverse group of molecules that may target more than one bacterium effectively. They have been shown to be effective against a variety of pathogens and even against some diseases, including diabetes[45] and cancer.[46] Because some AMPs in the data sets that we utilized in our analysis show this multitargeting ability and others do not, the bacteria-based collaborative filtering may have shown high predictive accuracy for some bacteria but not for others. Predictions will be more accurate for items in which nearest neighbors continue to have similar response patterns; however, for some AMPs, the response patterns of two bacteria might diverge. For example, Table S2 for individual AMP collaborative filtering results for the test data set showed that the predictive accuracy was not equally high across bacteria with similar nearest neighbors, potentially because for some AMPs predictions were less accurate. Collaborative filtering is data-dependent, and thus,

https://doi.org/10.1021/acs.jcim.3c00137
J. Chem. Inf. Model. 2023, 63, 3697−3704

the predictions will be more reliable and accurate for larger and more homogeneous data sets. In the current study, the bacteria with a greater number of data points for the calculations of R^2 are likely to have more reliable estimates, and we recommend that these predictions be tested first experimentally.

Nonetheless, NBCF may not necessarily select the nearest neighbors according to the structural similarity but rather by the functional similarity, specifically in response to AMPs. Two bacteria might be very different from a structural or evolutionary point of view, but if their responses to a set of AMPs are similar because the AMPs can target both, they will be grouped by the NBCF as nearest neighbors. This might present a problem for the accuracy of predictions because AMPs are very diverse and may target multiple species of bacteria or only one, and this might happen via different microscopic mechanisms. To probe the validity of our theoretical approach, it will be crucial to test our predictions with the quantitative experimental measurements for the same properties. The identification of effective AMP-antibiotic combinations is particularly important because our approach suggests that antibiotics that have been expensive to develop can continue to be used in lower concentrations and with less toxicity and side effects when combined with the proper AMPs. Bacteria show reduced resistance to AMPs and AMP-antibiotic combinations in contrast to antibiotic-antibiotic combinations.[10,47]

The benefit of AMP-antibiotic combinations is that even if a bacterium develops resistance to a certain combination, the same antibiotic can be used together with a different type of AMP, and the new combination may be effective with a lower risk of the development of the resistance. Accordingly, we utilized the link prediction method to identify potential synergistic AMP-antibiotic pairs for the specific bacterial species, *E. coli*. This method can be considered as a complementary approach to the nearest-neighbor collaborative filtering method since new combinations that are not yet established in experiments can be predicted. For example, pairs that have not yet been tested on any bacteria and thus have no FIC values from which to predict the unknown FIC values can be identified. While the bacterial-based collaborative filtering can extend the results from one or more bacteria to predictions for another bacterium, the link predictions can suggest new combinations that are likely to be effective but have not yet been tested on any bacteria. Moreover, if the collaborative filtering cannot be used to make predictions for a certain bacterium because there are no similar bacteria that can serve as nearest neighbors in the data set, the link prediction still can be used to make binary predictions (synergistic or not synergistic) and can potentially predict new AMP-AMP combinations and antibiotic-antibiotic combinations in addition to AMP-antibiotic combinations from a network that contains only AMP-antibiotic edges. This method is especially important because it allows us to analyze the synergistic activities of AMP-antibiotic combinations for a specific bacterium. However, the validity of the method remains to be tested with experimental studies that measure synergy of the suggested combinations.

It is important also to discuss the limitations of our study. One of the biggest limitations is the size and the nature of the data sets. Although collaborative filtering can be used to make predictions about sparse data sets, excessive sparsity is an issue for model evaluation, particularly when using the coefficient of determination R^2. Moreover, sufficient data is required to perform cross-validation and robustly assess the model performance. The nature of the data set may also pose an issue. Because a small amount of data may be used, collaborative filtering is sensitive to specific values of ratings, which might be influenced by many factors. The MIC values are subject to experimental error, as indicated by large ranges reported in individual experiments, though we accounted for this as best as possible by selecting data where there were multiple MIC values reported with low variance. We also selected data collected under the same experimental conditions and considered each strain separately. The MIC measurements may not be comparable within the same strain for antimicrobial agents that are resistant, so future studies should investigate collaborative filtering for predicting responses of resistant antimicrobial agents. It is expected that link prediction and the collaborative filtering will become more relevant as more AMP-based drugs are tested on more bacteria, particularly on similar and overlapping bacteria. Specifically, it would be helpful for the same measure (e.g., FIC) to be reported, although other measures have been also proposed for AMP-antibiotic and AMP-AMP combinations tested in the same bacteria.[48] Another limitation of our study is that we applied the models in their most simplified forms. Based on the demonstrated utility of the basic models for predicting AMP antimicrobial activity, more advanced models can be developed. For example, we can extract the latent features from the recommendation systems using matrix factorization.[49] It would be also crucial to analyze the role of the evolutionary relationship among bacteria in predicting the antimicrobial activity. In this case, one can represent the bacterial similarity by the phylogenetic distance.[50] We are planning to investigate all these possibilities in future studies.

■ ASSOCIATED CONTENT

Data Availability Statement

All source codes and data files are available online at https://figshare.com/s/1f9125b2be6e31292eef. The data (machine learning predictions) obtained in this work and the in-house scripts are available on figshare at the following URL https://figshare.com/s/1f9125b2be6e31292eef.

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.3c00137.

Additional figures, tables, and data (prediction of missing values, evaluation of the model, link prediction results) (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Anatoly B. Kolomeisky − *Department of Chemistry, Rice University, Houston, Texas 77005, United States; Center for Theoretical Biological Physics, Department of Chemical and Biomolecular Engineering, and Department of Physics and Astronomy, Rice University, Houston, Texas 77005, United States;* orcid.org/0000-0001-5677-6690; Email: tolya@rice.edu

Authors

Angela Medvedeva − *Department of Chemistry, Rice University, Houston, Texas 77005, United States; Center for Theoretical Biological Physics, Rice University, Houston, Texas 77005, United States*

https://doi.org/10.1021/acs.jcim.3c00137
J. Chem. Inf. Model. 2023, 63, 3697−3704

Hamid Teimouri − *Department of Chemistry, Rice University, Houston, Texas 77005, United States; Center for Theoretical Biological Physics, Rice University, Houston, Texas 77005, United States*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jcim.3c00137

Author Contributions

A.M. and H.T. designed the research. A.M. and H.T. performed the research. A.M., H.T., and A.B.K. wrote the article.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The work was supported by the Welch Foundation (C-1559), by the NIH (R01 HL157714-02), by the NSF (CHE-1953453 and MCB-1941106), and by the Center for Theoretical Biological Physics sponsored by the NSF (PHY-2019745).

■ REFERENCES

(1) O'Neill, J. *Tackling Drug-Resistant Infections Globally: Final Report and Recommendations-The Review on Antimicrobial Resistance*; Government of the United Kingdom: 2016.

(2) Tacconelli, E.; Carrara, E.; Savoldi, A.; Harbarth, S.; Mendelson, M.; Monnet, D. L.; Pulcini, C.; Kahlmeter, G.; Kluytmans, J.; Carmeli, Y.; et al. Discovery, research, and development of new antibiotics: the WHO priority list of antibiotic-resistant bacteria and tuberculosis. *Lancet Infectious Diseases* **2018**, *18*, 318−327.

(3) Suay-García, B.; Pérez-Gracia, M. T. Present and Future of Carbapenem-Resistant Enterobacteriaceae Infections. *Advances in Clinical Immunology, Medical Microbiology, COVID-19, and Big Data*; 2021; pp 435−456.

(4) Aslam, B.; Rasool, M.; Muzammil, S.; Siddique, A. B.; Nawaz, Z.; Shafique, M.; Zahoor, M. A.; Binyamin, R.; Waseem, M.; Khurshid, M., et al. Carbapenem resistance: Mechanisms and drivers of global menace. *Pathogenic Bacteria*; IntechOpen: 2020.

(5) Magana, M.; Pushpanathan, M.; Santos, A. L.; Leanse, L.; Fernandez, M.; Ioannidis, A.; Giulianotti, M. A.; Apidianakis, Y.; Bradfute, S.; Ferguson, A. L.; et al. The value of antimicrobial peptides in the age of resistance. *Lancet Infectious Diseases* **2020**, *20*, e216−e230.

(6) Zhu, Y.; Hao, W.; Wang, X.; Ouyang, J.; Deng, X.; Yu, H.; Wang, Y. Antimicrobial peptides, conventional antibiotics, and their synergistic utility for the treatment of drug-resistant infections. *Medicinal Research Reviews* **2022**, *42*, 1377.

(7) Pasupuleti, M.; Schmidtchen, A.; Malmsten, M. Antimicrobial peptides: key components of the innate immune system. *Critical reviews in biotechnology* **2012**, *32*, 143−171.

(8) Sierra, J. M.; Viñas, M. Future prospects for Antimicrobial peptide development: Peptidomimetics and antimicrobial combinations. *Expert Opinion on Drug Discovery* **2021**, *16*, 601−604.

(9) Hyun, S.; Choi, Y.; Jo, D.; Choo, S.; Park, T. W.; Park, S.-J.; Kim, S.; Lee, S.; Park, S.; Jin, S. M.; et al. Proline hinged amphipathic α-helical peptide sensitizes gram-negative bacteria to various gram-positive antibiotics. *J. Med. Chem.* **2020**, *63*, 14937−14950.

(10) Maron, B.; Rolff, J.; Friedman, J.; Hayouka, Z. Antimicrobial Peptide Combination Can Hinder Resistance Evolution. *Microbiology spectrum* **2022**, *10*, e0097322−22.

(11) Güvenç Paltun, B.; Kaski, S.; Mamitsuka, H. Machine learning approaches for drug combination therapies. *Briefings in Bioinformatics* **2021**, *22*, bbab293.

(12) Jin, W.; Stokes, J. M.; Eastman, R. T.; Itkin, Z.; Zakharov, A. V.; Collins, J. J.; Jaakkola, T. S.; Barzilay, R. Deep learning identifies synergistic drug combinations for treating COVID-19. *Proc. Natl. Acad. Sci. U. S. A.* **2021**, *118*, No. e2105070118.

(13) Li, J.; Tong, X.-Y.; Zhu, L.-D.; Zhang, H.-Y. A machine learning method for drug combination prediction. *Frontiers in Genetics* **2020**, *11*, 1000.

(14) Wu, L.; Wen, Y.; Leng, D.; Zhang, Q.; Dai, C.; Wang, Z.; Liu, Z.; Yan, B.; Zhang, Y.; Wang, J.; et al. Machine learning methods, databases and tools for drug combination prediction. *Briefings in bioinformatics* **2022**, *23*, bbab355.

(15) Lee, E. Y.; Lee, M. W.; Fulan, B. M.; Ferguson, A. L.; Wong, G. C. What can machine learning do for antimicrobial peptides, and what can antimicrobial peptides do for machine learning? *Interface focus* **2017**, *7*, 20160153.

(16) Nguyen, T. N.; Teimouri, H.; Medvedeva, A.; Kolomeisky, A. B. Cooperativity in Bacterial Membrane Association Controls the Synergistic Activities of Antimicrobial Peptides. *J. Phys. Chem. B* **2022**, *126*, 7365.

(17) Lee, E. Y.; Fulan, B. M.; Wong, G. C.; Ferguson, A. L. Mapping membrane activity in undiscovered peptide sequence space using machine learning. *Proc. Natl. Acad. Sci. U. S. A.* **2016**, *113*, 13588−13593.

(18) Söylemez, Ü. G.; Yousef, M.; Kesmen, Z.; Büyükkiraz, M. E.; Bakir-Gungor, B. Prediction of Linear Cationic Antimicrobial Peptides Active against Gram-Negative and Gram-Positive Bacteria Based on Machine Learning Models. *Applied Sciences* **2022**, *12*, 3631.

(19) Oliveira, J.; Reygaert, W. C. *Gram negative bacteria*; 2019.

(20) Yang, D. C.; Blair, K. M.; Salama, N. R. Staying in shape: the impact of cell shape on bacterial survival in diverse environments. *Microbiology and Molecular Biology Reviews* **2016**, *80*, 187−203.

(21) Chen, L.; Chu, C.; Huang, T.; Kong, X.; Cai, Y.-D. Prediction and analysis of cell-penetrating peptides using pseudo-amino acid composition and random forest models. *Amino acids* **2015**, *47*, 1485−1493.

(22) Zhang, Y.-H.; Li, Z.; Lu, L.; Zeng, T.; Chen, L.; Li, H.; Huang, T.; Cai, Y.-D. Analysis of the Sequence Characteristics of Antifreeze Protein. *Life* **2021**, *11*, 520.

(23) Li, H.; Zhang, S.; Chen, L.; Pan, X.; Li, Z.; Huang, T.; Cai, Y.-D. Identifying Functions of Proteins in Mice With Functional Embedding Features. *Frontiers in Genetics* **2022**, *13*, 909040.

(24) Pan, X.; Li, H.; Zeng, T.; Li, Z.; Chen, L.; Huang, T.; Cai, Y.-D. Identification of protein subcellular localization with network and functional embeddings. *Frontiers in Genetics* **2021**, *11*, 626500.

(25) Ren, X.; Wang, S.; Huang, T. Decipher the connections between proteins and phenotypes. *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics* **2020**, *1868*, 140503.

(26) Chen, L.; Zhang, Y.-H.; Huang, T.; Cai, Y.-D. Identifying novel protein phenotype annotations by hybridizing protein−protein interactions and protein sequence similarities. *Molecular Genetics and Genomics* **2016**, *291*, 913−934.

(27) Dean, S. N.; Alvarez, J. A. E.; Zabetakis, D.; Walper, S. A.; Malanoski, A. P. PepVAE: variational autoencoder framework for antimicrobial peptide generation and activity prediction. *Frontiers in microbiology* **2021**, *12*, 725727.

(28) Benfield, A. H.; Henriques, S. T. Mode-of-Action of Antimicrobial Peptides: Membrane Disruption vs Intracellular Mechanisms. *Front. Med. Technol.* **2020**, *2*, 610997.

(29) Kleandrova, V. V.; Ruso, J. M.; Speck-Planche, A.; Dias Soeiro Cordeiro, M. N. Enabling the discovery and virtual screening of potent and safe antimicrobial peptides. Simultaneous prediction of antibacterial activity and cytotoxicity. *ACS combinatorial science* **2016**, *18*, 490−498.

(30) Speck-Planche, A.; Kleandrova, V. V.; Ruso, J. M.; D S Cordeiro, M. First multitarget chemo-bioinformatic model to enable the discovery of antibacterial peptides against multiple Gram-positive pathogens. *J. Chem. Inf. Model.* **2016**, *56*, 588−598.

(31) Aggarwal, C. C. Neighborhood-Based Collaborative Filtering. *Recommender Systems: The Textbook*; Springer International Publishing: Cham, 2016; pp 29−70.

(32) Luo, X.; You, Z.; Zhou, M.; Li, S.; Leung, H.; Xia, Y.; Zhu, Q. A highly efficient approach to protein interactome mapping based on collaborative filtering framework. *Sci. Rep.* **2015**, *5*, 7702.

(33) Zhang, L.; Chen, X.; Guan, N.-N.; Liu, H.; Li, J.-Q. A hybrid interpolation weighted collaborative filtering method for anti-cancer drug response prediction. *Frontiers in pharmacology* **2018**, *9*, 1017.
(34) Zhang, W.; Zou, H.; Luo, L.; Liu, Q.; Wu, W.; Xiao, W. Predicting potential side effects of drugs by recommender methods and ensemble learning. *Neurocomputing* **2016**, *173*, 979−987.
(35) Newman, R.; Pietras, C. M.; Qu, F.; Sapashnik, D.; Kofman, L.; Butze, S.; Ocitti, F.; Fried, I.; Slonim, D. Improving cell-specific drug connectivity mapping with collaborative filtering. *bioRxiv* **2020**, DOI: 10.1101/2020.08.10.231720.
(36) Aggarwal, C. C. Mining Web Data. *Data Mining*. Springer, Cham: 2015; pp 589−617.
(37) Andrews, J. M. Determination of minimum inhibitory concentrations. *Journal of antimicrobial Chemotherapy* **2001**, *48*, 5−16.
(38) Laishram, S.; Pragasam, A. K.; Bakthavatchalam, Y. D.; Veeraraghavan, B. An update on technical, interpretative and clinical relevance of antimicrobial synergy testing methodologies. *Indian Journal of Medical Microbiology* **2017**, *35*, 445−468.
(39) Zhang, L.; Benz, R.; Hancock, R. E. Influence of proline residues on the antibacterial and synergistic activities of α-helical peptides. *Biochemistry* **1999**, *38*, 8102−8111.
(40) Ruden, S.; Rieder, A.; Chis Ster, I.; Schwartz, T.; Mikut, R.; Hilpert, K. Synergy pattern of short cationic antimicrobial peptides against multidrug-resistant Pseudomonas aeruginosa. *Frontiers in microbiology* **2019**, *10*, 2740.
(41) Pirtskhalava, M.; Amstrong, A. A.; Grigolava, M.; Chubinidze, M.; Alimbarashvili, E.; Vishnepolsky, B.; Gabrielian, A.; Rosenthal, A.; Hurt, D. E.; Tartakovsky, M. DBAASP v3: database of antimicrobial/cytotoxic activity and structure of peptides as a resource for development of new therapeutics. *Nucleic acids research* **2021**, *49*, D288−D297.
(42) Vozalis, M.; Margaritis, K. G. On the combination of user-based and item-based collaborative filtering. *International Journal of Computer Mathematics* **2004**, *81*, 1077−1096.
(43) Chicco, D.; Warrens, M. J.; Jurman, G. The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. *PeerJ. Computer Science* **2021**, *7*, No. e623.
(44) Daud, N. N.; Ab Hamid, S. H.; Saadoon, M.; Sahran, F.; Anuar, N. B. Applications of link prediction in social networks: A review. *Journal of Network and Computer Applications* **2020**, *166*, 102716.
(45) Soltaninejad, H.; Zare-Zardini, H.; Ordooei, M.; Ghelmani, Y.; Ghadiri-Anari, A.; Mojahedi, S.; Hamidieh, A. A. Antimicrobial peptides from amphibian innate immune system as potent antidiabetic agents: a literature review and bioinformatics analysis. *Journal of Diabetes Research* **2021**, *2021*, 2894722.
(46) Jin, G.; Weinberg, A. Human antimicrobial peptides and cancer. *Seminars in cell & developmental biology* **2019**, *88*, 156−162.
(47) Sheard, D. E.; O'Brien-Simpson, N. M.; Wade, J. D.; Separovic, F. Combating bacterial resistance by combination of antibiotics with antimicrobial peptides. *Pure Appl. Chem.* **2019**, *91*, 199−209.
(48) Yu, G.; Baeder, D. Y.; Regoes, R. R.; Rolff, J. Combination effects of antimicrobial peptides. *Antimicrob. Agents Chemother.* **2016**, *60*, 1717−1724.
(49) Zhang, W.; Yue, X.; Lin, W.; Wu, W.; Liu, R.; Huang, F.; Liu, F. Predicting drug-disease associations by using similarity constrained matrix factorization. *BMC bioinformatics* **2018**, *19*, 233.
(50) Lalucat, J.; Mulet, M.; Gomila, M.; García-Valdés, E. Genomics in bacterial taxonomy: impact on the genus Pseudomonas. *Genes* **2020**, *11*, 139.

https://doi.org/10.1021/acs.jcim.3c00137
J. Chem. Inf. Model. 2023, 63, 3697−3704